UNIVERSITAT DE BARCELONA

Structural Modeling and Characterization of Protein Interactions of Biomedical Interest: The Challenge of Molecular Flexibility

Chiara Pallara

UNIVERSITAT DE BARCELONA
Facultat de Fàrmacia

Programa de Doctorat en Biomedicina
RD 778/1998

Structural Modeling and Characterization of Protein Interactions of Biomedical Interest: The Challenge of Molecular Flexibility

Director de tesis
Dr. Juan Fernández-Recio
Barcelona Supercomputing Center

Doctorand
Chiara Pallara

Chiara Pallara

Barcelona, 2015

To my parents for their
endless support and understanding.

Acknowledgements

“The only true voyage of discovery, the only fountain of Eternal Youth, would be not to visit strange lands but to possess other eyes, to behold the universe through the eyes of another, of a hundred others, to behold the hundred universes that each of them beholds, that each of them is.”

Marcel Proust

El barco ya ha llegado a puerto y mi viaje termina aquí. Ha sido una aventura estimulante, quizá dura a veces, pero sin duda gratificante y sobretodo constructiva. Los retos perseguido me han apasionado, las metas logradas entusiasmado y los obstáculos encontrados fortalecido. No sé cual será mi próximo destino pero este viaje no puede acabar sin dar las gracias a todas las personas sin las cuales todo esto no hubiera sido posible.

Primero, quiero agradecer al capitán Juan, por haberme acogido en su barco y haberme dado el privilegio de estar a su lado. Por confiar en mí desde el principio, por toda su paciencia, por su disponibilidad y dedicación, gracias.

Quiero también agradecer el trabajo y el esfuerzo realizado por todas las personas involucradas en los diferentes proyectos de colaboración llevados a cabo durante la tesis, asi como a los miembros de la commisión de seguimiento y del tribunal por el tiempo dedicado a este trabajo.

Quiero dar las gracias también a mis compañeros de viaje, los que ya se fueron, lo que desde le principio compartieron conmigo esta aventura y los que la revitalizaron con oxigeno nuevo. Mil gracias a Laura y a Solène para acogerme con infinito cariño desde el primer día y gracias también a Carles y Albert por sus ilimitada disponibilidad. Gracias a Brian por estar siempre, por compartir conmigo penas y alegrías y ofrecer su más sincero apoyo cada día. Gracias a Miguel por mitigar mi entropía con su inviolable coherencia y racionalidad. Gracias a Didier porque vencer los sempiternos miedos de uno es más fácil con una plácida sonrisa a tu lado. Gracias a Mireia por su alegría y su simpatía inagotable y a Lucía por su infinita dulzura. Gracias a Sergio por la tranquilad que su mirada trasmite y a Luis por su vivacidad y impetuoso entusiasmo.

Un gracias especial a Laia por su apoyo incondicional y por ayudarme a creer que era posible. Gracias a Dmitry por su inalterable tranquilidad, aún en los momentos más estresantes, y a Montse porque las mañanas son más amenas si estan acompañadas de su sonrisa.

Un infinito gracias también a Francesco por darme la oportunidad de colaborar con él, acogerme en su grupo, ofrecer su disponibilidad y compartir conmigo sus ideas. Muchas gracias también a su encantador grupo, Silvia, Giorgio, Ludovico y Kristen, por haberme hecho sentir en familia en tan poco tiempo juntos.

Un gracias a todos los compañeros del departamento de Life Science por todo el apoyo que he recibido en estos años vividos juntos, y también a todos los que indirectamente han hecho posible llevar a cabo esta aventura, el equipo de suporte de Mare Nostrum y todo el personal del BSC.

Finalmente un inmenso gracias a todas mi familia y mis amigos por creer en mi, por animarme, alentarme y apoyarme cada día.

Muchas gracias a todos

Contents

1. Introduction

“*Men love to wonder,*
and that is the seed of science.”

Ralph Waldo Emerson

1.1. Biological and biomedical importance of proteins

Accounting for about half of the total dry mass of cells, proteins play a major role in nature (Alberts, 1998). Often described as the factories of the cell, proteins are large biomolecules that play essential functional and structural roles within cells.

The building blocks of proteins are the amino acids, which are small molecules composed of an amine and a carboxylic group and differ in the side chain they carry on the alpha carbon (Cα) atom. Amino acids polymerize by forming a peptide bond between the carboxyl group of one amino acid and the amide group of another one, yielding large polypeptide chains. Through transcription and translation, the information carried by DNA is transformed into a polypeptide chain that might eventually be chemically modified by post-translational modifications. A total of 20 amino acids are encoded in the genome.

Once formed, proteins only exist for a certain period of time (typically ranging from minutes to years) and are then degraded and recycled by the cell's machinery through a process referred to as protein turnover.

1.1.1. Physiological function of proteins

According to their functions, different protein types have been identified. Enzymes are known to catalyze more than 5,000 biochemical reactions (e.g., pepsin is a digestive enzyme in the stomach that degrades food proteins into peptides). Antibodies, produced by the immune system, identify and neutralize pathogens such as bacteria and viruses. DNA-associated protein, like histones or cohesin proteins, arrange chromosome structure during cell division and play a role in regulating gene expression. Contractile proteins, such as actin and myosin, are involved in muscle contraction and movement. Hormone proteins coordinate bodily functions (e.g., insulin controls our blood sugar concentration by regulating the uptake of glucose into cells). Finally transport proteins move molecules within our bodies (e.g., hemoglobin transports oxygen through the blood).

Given their central role as executive machinery of a cell, proteins cover nearly every task in all the biological processes that occur within and between cells. Examples of such important processes are signal transduction (Furge, 2008), cellular energy metabolism (Atkinson, 1977), transcriptional regulation (Lee and Young, 2013) or membrane trafficking (Cheung and de Vries, 2008), some of which are relevant for the work of this thesis.

Signal transduction

Signal transduction generally refers to any basic cellular process involving the conversion of a signal from outside to a functional change within the cell. It generally starts when an extracellular signaling molecule (usually hormones, neurotransmitters, cytokines, growth factors or cell recognition molecules) activates a specific receptor located on the cell surface or inside the cell. In turn, this receptor triggers a signaling relay inside the cell, eventually ending to the modulation of DNA-related processes in the nucleus, which finally provokes a response (e.g., altering cell's metabolism, shape or ability to divide). The signal is amplified at any step so that one signaling molecule can cause many responses (Furge, 2008).

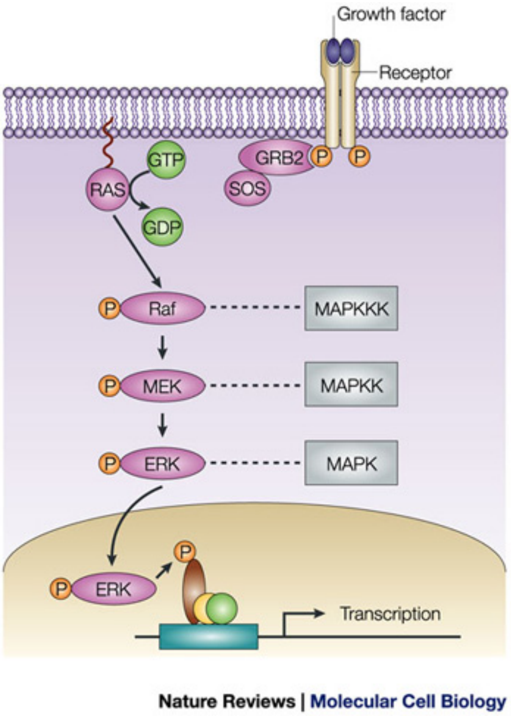


Figure 1. Example of signaling transduction process: MAPK/ERK signaling pathway. (From Kim and Bar-Sagi, 2004)

MAPK/ERK signaling pathway (also known as the RAS-RAF-MEK-ERK cascade) (Wortzel and Seger, 2011) is one of the principal and better-known signal transduction processes in cells. Indeed, it is involved in the tight regulation of many biological events, such as meiosis, gastrulation, embryogenesis, cell fate determination, angiogenesis and immune response. As shown in **Figure 1**, the starting point of the cascade consists in the binding of a ligand (e.g., a growth factor, cytokine, or hormone) to the extracellular portion of two subunits of a transmembrane protein (i.e., a receptor tyrosine kinase

(RTK)). This interaction, in turn, leads to RTK dimerization, phosphorylation of its cytoplasmic domains and the consequent binding of a cytoplasmic adaptor protein (CAP), such as GRB2. The newly formed complex attracts SOS protein, a guanine-nucleotide exchange factor (GEF), to the plasma membrane, which activates a small G proteins belonging to RAS superfamily. During the time it is active, RAS stimulates BRAF, a mitogen-activated protein kinase kinase kinase, which in turn binds and activates MEK dual-specificity protein kinase (Roskoski, 2012). MEK in turn prompts the stimulation of ERK, the third and final kinase in the cascade, which is responsible for the activation of a huge rooster of substrates, at least 160 proteins, including several transcriptional factors (e.g., ELK1, ETS, and c-FOS).

Cellular energy metabolism

Cellular metabolism is the set of life-sustaining chemical transformations within the cells. It includes all the reactions involved in the breakdown of molecules to obtain energy (catabolism) as well as in synthesizing macromolecules or small precursors (e.g., amino acids) needed by the cell (anabolism). Metabolic pathways can be simple linear sequences of a few reactions, but they can also be extensively branched with reactions converging on or diverging from a central main pathway. Alternatively, they can be cyclic, with a precursor of an early reaction regenerated at the end of a pathway (Atkinson, 1977).

One of the more complex and widely studied anabolic pathways is photosynthesis (Whitmarsh and Govindjee, 1999). This physicochemical process is carried out in all autotrophic organisms, such like green plants, algae and photosynthetic bacteria. It mainly includes two steps: (i) sunlight absorption and its conversion into chemical energy through Photosystem I and II (Caffarri et al., 2014); (ii) usage of the previously stored energy to assemble carbohydrates from carbon dioxide molecules by means of the so-called Calvin cycle.

Transcriptional regulation

Transcriptional regulation is the mechanism by which a cell regulates the conversion of DNA to RNA, thereby orchestrating gene activity modulation. This basic process, shared within all the living organisms, is tightly coordinated by transcription factors and other proteins, which work in concert to finely tune the amount of RNA being produced. (Lee and Young, 2013).

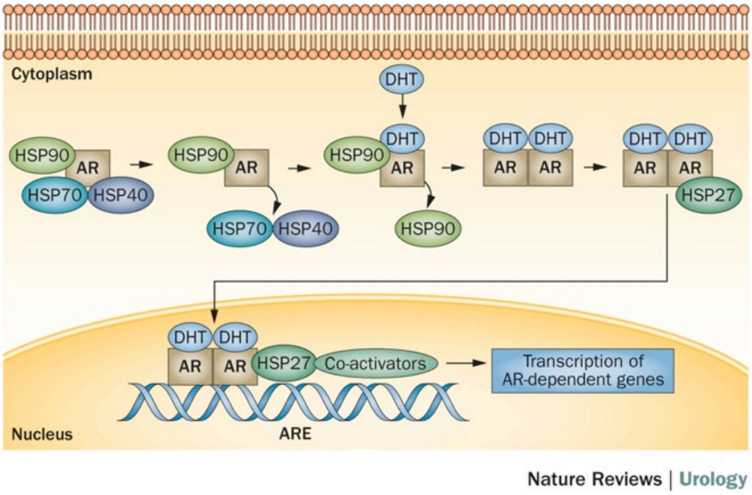


Figure 2. Example of transcriptional regulation: Androgen receptor (AR) signaling pathway. (From Azad et al., 2015)

An example of transcription factor is the androgen receptor (AR) (Gao et al., 2005), a member of the steroid hormone receptor. As shown in **Figure 2**, AR signaling pathway involves (i) the direct binding of the receptor to either androgenic hormone (i.e., testosterone or dihydrotestosterone) in the cytosol; (ii) the consequent conformational change in the AR that triggers dissociation of heat shock protein 90 (HSP90), dimerization and binding to HSP27; (iii) AR translocation into the cell nucleus and its final binding to specific DNA sequences (i.e, androgen response elements (ARE)) that eventually regulates cellular transcriptional activity.

Membrane vesicle trafficking

Membrane trafficking refers to a fundamental activity in eukaryotic cells which supports different basic processes (e.g., intercellular communication, extracellular matrix building through secretion, biomolecules import or export by endocytosis or exocytosis, periodic turnover of cellular organelles and pathogen phagocytosis). These tasks are typically mediated by membrane-bounded carriers serving as shuttles that link specialized cellular compartments with the cell surface in a highly organized and dynamic network (Cheung and de Vries, 2008). Although each specific pathway is governed by its own set of controlling factors, they all contain Rab GTPase proteins (Hutagalung and Novick, 2011) that serve as master regulators, modulating many steps of membrane trafficking, including vesicle formation, vesicle movement along actin and tubulin networks and membrane fusion. Therefore, it is not surprising that several human intracellular bacterial pathogens (e.g., *Chlamydiae, Coxiella burnetii, Helicobacter pylori, Legionella pneumophila, Listeria monocytogenes, Pseudomonas aeruginosa and Salmonella enterica* and several Mycobacteria) target Rab through post-translational modifications to precisely manipulate host cell functions and colonize its vacuolar compartments (Muller et al., 2010).

1.1.2. Protein dysfunction and disease

DNA carries and transmits the genetic information by specifying the amino acid sequence template for protein synthesis. Therefore, pathological mutations in genes can affect the folding and stability (Ode et al., 2007; Alfalah et al., 2009), the function (Yamada et al., 2006), the interactions (Jones et al., 2007) as well as the expression and subcellular localization of the proteins they encode (Krumbholz et al., 2006). Moreover, since proteins have a variety of functions and many of them are active as multimeric complexes (e.g., interacting with small molecules, other proteins or cellular membranes), the molecular mechanisms underlying even the simplest of genetic disorders are typically composite and heterogeneous.

An example of such complexity is the MAPK/ERK signaling pathway. Given its involvement in a large variety of cellular activities (see **section 1.1.1**), deviation from the strict control of this pathway has been implicated in the development of many human diseases including Alzheimer's disease (AD), Parkinson's disease (PD), amyotrophic lateral sclerosis (ALS), various types of cancers (e.g., pancreas, colon, lung, ovary (Shields et al., 2000; Davies et al., 2002; Rajagopalan et al., 2002; Hingorani et al., 2003; Mercer and Pritchard, 2003; Singer et al., 2003; Vos et al., 2003; Sieben et al., 2004; Sharma et al., 2005; Hoeflich et al., 2006; Sumimoto et al., 2006; Dhillon et al., 2007) as well as the Ras/MAPK syndromes (the “RASopathies”). The latter are a group of rare germline developmental disorders, e.g., Noonan, cardio-facio-cutaneous (CFC), Costello and LEOPARD syndromes (Tartaglia and Gelb, 2005; Bentires-Alj et al., 2006; Schubbert et al., 2007; Aoki et al., 2008; Tartaglia et al., 2010), sharing phenotypic features that include postnatal reduced growth, facial dysmorphism, cardiac defects, mental retardation, skin defects, musculo-skeletal defects, short stature and cryptorchidism (Rauen, 2013).

Interestingly, both cancer and RASopathies-related mutations share the same 15 genes of this pathway (i.e., PTPN11, SOS1, RASA1, NF1, KRAS, HRAS, NRA S, BRAF, RAF1, MAP2K1, MAP2K2, SPRED1, RIT1, SHOC2 and CBL (Aoki et al., 2013; Rauen, 2013). The phenotypic fate of a given mutation seems to be related to the structural and energetic effect at molecular level, although all the details of the mechanisms underlying each disorder are not yet fully understood (Kiel and Serrano, 2014).

Therefore, the knowledge of the structural details of a protein is fundamental not only to characterize its biological function in physiological conditions, but also to understand its role in pathological situations. Mapping disease-related mutations on the 3D structure of a protein can provide invaluable insights on the disease-causing mechanism and can help to explain the phenotypic outcome associated to a specific mutation.

1.1.3. Fundaments of protein structure

The sequence of a polypeptide chain determines how it folds into one or several specific spatial conformations, which in turn define its function. Therefore, the study of the 3D structure of a given protein can provide invaluable details about its functional role at molecular level (Shakhnovich et al., 2003).

Protein folds

Nowadays, protein structure is generally referred to in terms of four aspects: (i) the primary structure consisting of the amino acids sequence; (ii) the secondary structure, which contains regularly repeating arrangements (i.e., alpha-(α)-helices and beta-(β)-sheets) mainly stabilized by hydrogen bonds; (iii) the tertiary structure, which defines the final folding pattern incorporating various secondary structures; and finally (iv) a quaternary structure involving the clustering of several individual protein chains into a final specific configuration. An example of a protein with quaternary structure is the photosystem I (PSI), an integral membrane protein complex that uses light energy to mediate electron transfer from plastocyanin (or cytochrome c_6) to ferredoxin metalloproteins during the photosynthesis. It is composed of a reaction center of up to 14 subunits and a membrane-associated antenna complex (LHCI) that captures light and guides its energy to the reaction center (Golbeck, 1987).

According to their overall structural features, proteins represent a highly heterogeneous class of biological macromolecules, differing both in topology, shape and size. As defined by the two main structure classification databases, namely SCOP (Murzin et al., 1995) and CATH (Orengo et al., 1997) four large fold classes have been established in order to describe all existing protein topologies: the all-α and all-β proteins, as well as the α+β proteins and α/β proteins classes in which the secondary structure is composed of α-helices and β-strands that occur separately or alternatively along the backbone, respectively. A more generic classification consists in the overall shape of proteins, labeling them as globular, fibrous, disordered and membrane proteins. Such protein classes mainly differ in the secondary and tertiary structure as well as in several physico-chemical properties (e.g., thermal stability, solubility and inter-residue interactions types). Apart from in topology and shape, proteins significantly differ in size. Indeed, they range from 100 residues, as found in some short ribosomal proteins, to several thousand residues (Brocchieri and Karlin, 2005), as observed in titin, a giant multifunctional protein involved in the contraction of human striated muscle tissues, which is composed by ~33000 residues and reaches 1 μm in length.

Experimental determination of protein structure

The tertiary and quaternary structures of a large number of proteins have become available in the World Wide Protein Data Bank (WWPDB; (Berman et al., 2007)), a single worldwide repository of information about the 3D structures of large biological molecules (including proteins and nucleic acids), which originated from a collaborative effort of RCSB Protein Data Bank (RCSB-PDB, (Berman et al., 2000)), Protein Data Bank Europe (PDBe, (Velankar and Kleywegt, 2011)), Protein Data Bank Japan (PDBj; (Standley et al., 2008), and Biological Magnetic Resonance Databank (BMRB; BioMagResBank, (Ulrich et al., 2008)). Established in 1971, WWPDB recently archived its 100,000th molecule structure, doubling its size in just six years and reaching a releasing rate of 200 structures per week.

Accounting in July 2015 for 89.8% of the biomolecules deposited in the WWPDB, the most known and widely used experimental method for the structural resolution of proteins is X-ray crystallography (Smyth and Martin, 2000). This technique allows the 3D structural description at atomic resolution of crystallized macromolecules, based on the scattering produced by an X-ray beam after contacting the electrons of a protein in a crystal. The diffraction produced contains information about the electron density of the macromolecule, from which atom positions and chemical bonds can be calculated. In spite of its unquestionable success, X-ray crystallography certainly shows intrinsic limitations that affect its applicability in some protein systems: (i) not all the proteins can easily crystallize (e.g., membrane proteins); (ii) flexible parts of the proteins sometimes cannot be solved (e.g., loops); (iii) intrinsically disordered proteins and some proteins that may adopt many different conformations in solution represent a serious challenge for this technique; and finally, (iv) not all the contacts reported in the crystal are biologically relevant and the experimental conditions may not represent accurately those of the *in vivo* environment.

Some of these problems are solved by Nuclear Magnetic Resonance (NMR) spectroscopy (Wuthrich, 1990), which represents the second most widely used technique after X-ray crystallography and accounts for roughly 9.4% of the proteins structures deposited in the WWPDB. By this method, the protein is placed under a strong magnetic field and short radio frequency pulses are aimed at the sample. This allows the detection of distinct chemical shift produced by each of the nuclei of the macromolecule, which depend on their chemical environment. Using different radio frequency pulses and analyzing the chemical shift of the different nuclei, it is possible to determine the distance between the different atoms in the protein and therefore obtain its 3D structure. The main advantage over X-ray crystallography is the description of the dynamics of the protein under study: thus, mobile loops, different conformations of a protein and intrinsically disordered structures can be efficiently described by

NMR. Moreover, the sample is studied in solution, which represents more realistic conditions as compared to X-ray crystallography. However, NMR appears scarcely suitable for large proteins of more than 40kDa (Krishnan and Rupp, 2001) unlike X-ray crystallography whose applicability is not significantly affected by protein size.

The remaining protein structures deposited in the WWPDB are mainly solved by electron microscopy (EM) techniques (Bernadó, 2011). For many years, EM has been limited to large complexes or low-resolution models (typically around 15 Å) and thus typically used in combination with complementary tools (e.g., computational modeling) (Petoukhov and Svergun, 2005)). Indeed, thanks to recent advances in electron detection and image processing, the technique has experienced a dramatic improvement in resolution, reaching roughly 5 Å by cryo-EM (Alushin et al., 2014), and thus beginning to rival NMR and X-ray crystallography.

Despite impressive progress in automating experimental structure determination techniques, they still remain highly time-consuming and with no guaranteed success. On the contrary, the advances in DNA sequencing techniques are giving rise to an unprecedented avalanche of new sequences (UniProt, 2013), dramatically widening the gap between protein solved structures and annotated sequences. This is reflected by the fact that the number of structurally characterized proteins deposited in Protein Data Bank is about two orders of magnitude smaller than the number of known protein sequences in the SwissProt and TrEMBL (recently exceeding 50 million) (Boeckmann et al., 2003) (**Figure 3**).

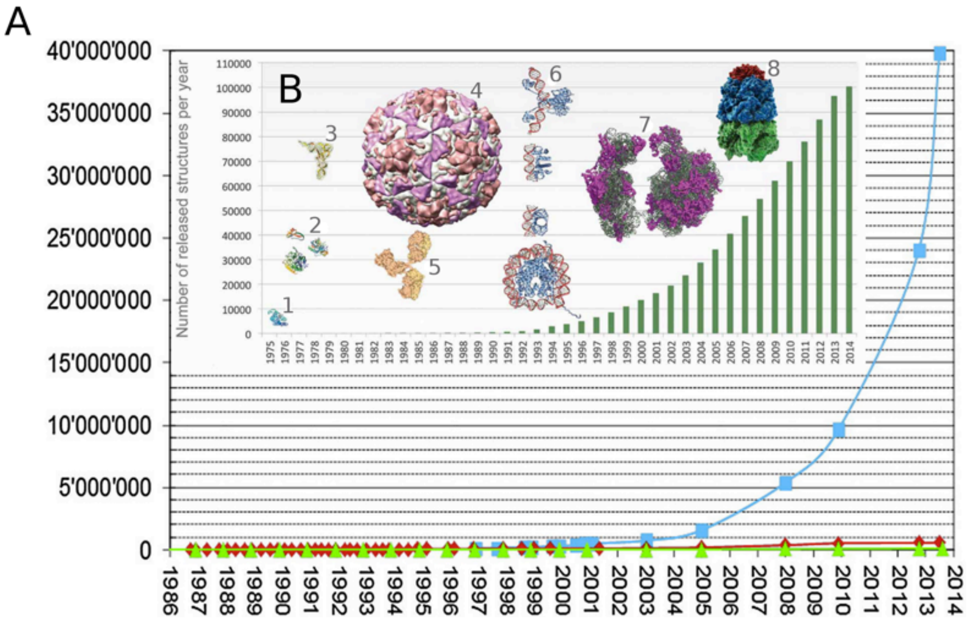


Figure 3. (A) Comparison between the number of entries in the SwissProt (in red), TrEMBL (in blue) and PDB (in green) from 1986 to 2014. (From Schwede, 2013) (B) Number of structures available in the PDB per year, as of May 14, 2014. Highlighted examples include: 1) myoglobin, one of the first structures solved by X-ray crystallography; 2) small enzymes; 3) examples of tRNA; 4) viruses; 5) antibodies; 6) protein-DNA complexes; 7) ribosomes and 8) chaperones. (Image courtesy of wwPDB)

Computational modeling of protein structure

The above cited findings make obvious that it will be impossible to determine experimentally the structure of every protein of interest with current techniques, despite the huge efforts of the ongoing PSI (Protein Structure Initiative) worldwide project and similar structural genomics efforts. However, based on the observation that homologous proteins sharing detectable sequence similarity have similar 3D structures and that their structural diversity is increasing with evolutionary distance (Chothia and Lesk, 1986), during the last two decades several comparative modeling techniques, also known as template-based modeling (or TBM), have been developed. More recently, TBM techniques have been extended to model tertiary structure of remote homologs through threading methods (Bowie et al., 1991), which aim to recognize the template structures without evolutionary relation to the target by incorporating structure information into sequence alignments.

Thanks to recent computational advances in large-scale data management, several different TBM methods have been developed (Ginalski, 2006; Zhang, 2008b), based on fully automated stable pipelines which typically include: (i) finding one or more appropriate templates; (ii) aligning the target sequence with the templates using sequence alignment, profile-based alignment, or threading; (iii) building an initial model for the target by copying the structural fragments from the aligned regions of the template(s); (iv) replacing the side chains to match the sequence of the target; (v) constructing missing loops and termini; and finally (vi) realigning the model to obtain a full-length atomic structure.

Moreover, the notable advance in the homology modeling tools led to the development of different web servers which allow (i) the interactive extrapolation of the available experimental structure information of homologous proteins and (ii) the supply of reliable three-dimensional (3D) models, starting from the uncharacterized protein sequences. Additionally extended databases of annotated 3D comparative protein structure models have been recently compiled. Some well-known comparative modeling servers and databases are listed in Table 1.

Table 1. (A) CASP-cited comparative modeling servers and (B) protein most-cited databases of comparative protein structure models.

A. Comparative modeling servers	
I-TASSER	http://zhanglab.ccmb.med.umich.edu/I-TASSER/ (Yang et al., 2015)
ROBETTA	http://robetta.bakerlab.org/ (Raman et al., 2009)
HHpred	http://toolkit.tuebingen.mpg.de/hhpred (Soding et al., 2005)
METATASSER	http://cssb.biology.gatech.edu/skolnick/webservice/MetaTASSER/index.html (Zhou et al., 2009)
MULTICOM	http://sysbio.rnet.missouri.edu/multicom_cluster/ (Cao et al., 2014)
Pcons	http://pcons.net/ (Larsson et al., 2011)
SAM-T08	http://compbio.soe.ucsc.edu/SAM_T08/T08-query.html (Karplus et al., 1998)
3D-Jury	http://BioInfo.PL/Meta/ (Ginalski et al., 2003)
RaptorX	http://raptorx.uchicago.edu/ (Kallberg et al., 2012)
THREADER	http://bioinf.cs.ucl.ac.uk/?id=747 (Jones et al., 1992; Jones et al., 1995)
SwissModel	http://swissmodel.expasy.org/ (Biasini et al., 2014)
ModWeb	http://modbase.compbio.ucsf.edu/modweb/ (Pieper et al., 2011)
B. Comparative protein structure models databases	
Swiss-Model	http://swissmodel.expasy.org/repository/ (Kiefer et al., 2009)
ModBase	http://modbase.compbio.ucsf.edu/modbase-cgi/index.cgi (Pieper et al., 2011)
Protein Model Portal (PMP)	http://www.proteinmodelportal.org/ (Arnold et al., 2009)
PMDB Protein Model DataBase	https://bioinformatics.cineca.it/PMDB/ (Castrignano et al., 2006)
Swiss-Model	http://swissmodel.expasy.org/repository/ (Kiefer et al., 2009)

The current state of homology modeling field is periodically assessed in a biennial large-scale experiment known as the Critical Assessment of Techniques for Protein Structure Prediction, or CASP

(Moult et al., 1995). As arisen from the last CASP editions, the major inaccuracies in homology modeling (which typically worsen with lower sequence identity) derive from errors in the initial sequence alignment and from improper template selection (Joo et al., 2014).

The three dimensional structure of a protein defines not only its size and shape, but also its function. Nevertheless, the structural charachterization of an isolated protein is often not enough to understand its function. Indeed, proteins act by forming complexes with other molecules. Moreover, proteins in solution are not static objects but rather ensembles of varying heterogeneous conformations constantly interconverting from one to another. Thus, consideration of molecular recognition phenomena as well as the dynamic nature of proteins cannot be neglected for a complete understanding of protein function at molecular level. Given the importance of these protein features, they will be revised more in details in the following two sections (**section 1.2** and **1.3**, respectively).

1.2. Protein-protein interactions: a broad overview

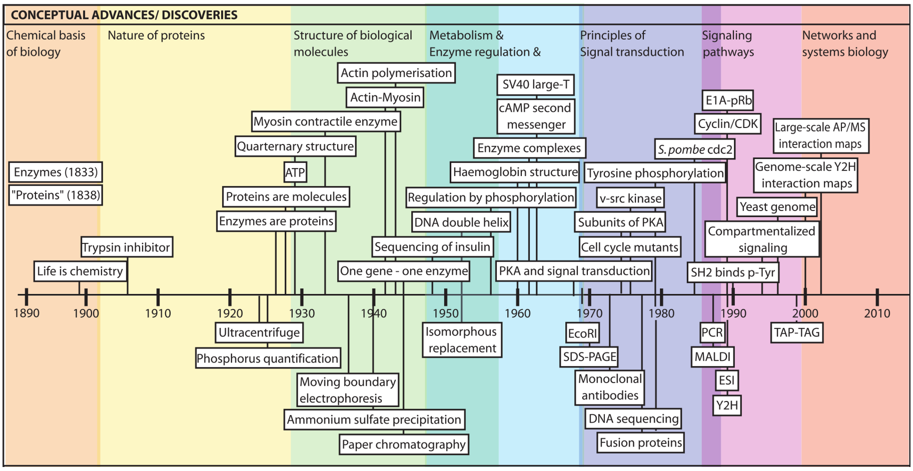


Figure 4. Time-line of protein research. In the top part conceptual advances and discoveries are indicated, in the lower part technological advances and inventions are indicated. From Braun and Gingras, 2012.

Until the late 1990's, protein function analyses had been mainly focused on single proteins. However more recent conceptual and technological advances in biochemistry and molecular biology as well as the several ongoing projects on protein-protein interaction mapping for many model species and humans (Rolland et al., 2014) confirmed that the majority of proteins mediate their functions by physically interacting with different biomolecules (i.e., other proteins, lipids, nucleic acids or small molecules) and thus forming intricate, highly organized and dynamic interaction networks (Rual et al., 2005; Stelzl et al., 2005) (**Figure 4**). These findings definitely suggested the necessity of exhaustively studying each protein in its proper biological context in order to fully comprehend their functions within the cell and thus paved the way for today's system-wide approaches to protein-protein interaction (PPI) analysis (Braun and Gingras, 2012).

1.2.1. Large-scale identification of protein-protein interactions (PPIs)

The observation of the involvement of protein interactions in almost every cellular process as well as the implication of aberrant PPIs in an increasingly number of diseases have clearly shown the necessity to identify and characterize such interactions. For this reason, protein-protein interactions are currently the object of intense research in many biological fields.

High-throughput experimental methods for PPIs detection

Different experimental techniques have been developed to measure physical interactions between proteins; these methods vary considerably in terms of time, costs, resources and their applicability to proteome-scale mapping. Two widely used methods adapted for high-throughput approaches are yeast two-hybrid (Y2H) system (Fields and Song, 1989) and tandem affinity purification followed by mass spectroscopy (TAP-MS) (Rigaut et al., 1999) (**Figure 5**).

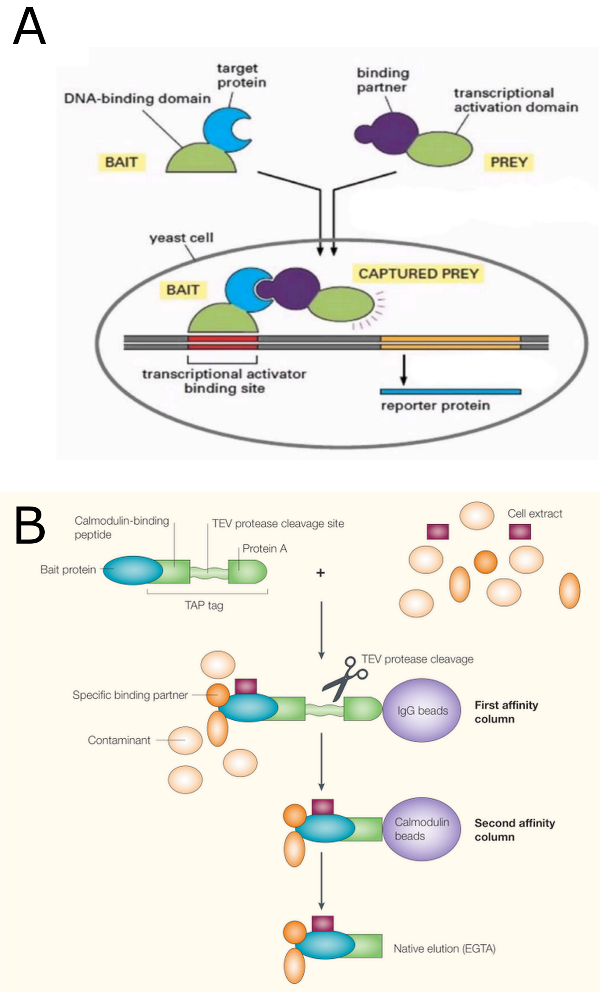


Figure 5. Schematic diagrams describing the key steps of (A) yeast two-hybrid (Y2H) and (B) tandem-affinity-purification (TAP) techniques. After http://technologyinscience.blogspot.com.es/ and Huber, 2003

The Y2H screening assays whether two proteins physically interact with each other: a *bait* and a *prey* protein are thus expressed using genetically modified yeast triggering the expression of a reporter gene as consequence of their interaction, if it happens. Y2H techniques have been used for many large-scale screening studies providing an extraordinary amount of protein-protein interaction data for a variety of model organisms including yeast (Uetz et al., 2000; Ito et al., 2001), fly (Giot et al., 2003), worm (Li et al., 2004) and human (Rual et al., 2005; Stelzl et al., 2005). Nevertheless, this technique is known to report false positives, including interactions of proteins that will never physically meet *in vivo* because of being expressed in incompatible cellular states or being present in different cellular compartments.

In contrast to Y2H approach, TAP-MS experiments allow high-throughput identification of protein interactions under near-physiological conditions. The protein of interest is firstly fused to a large protein (i.e., the *tag*), which easily allow its isolation; the resulting tagged protein is then expressed in the host cell, allowing the binding with its native partners, and consequentially purified from the cell extract using the *tag* (e.g., by specific antibodies). Tagged protein binders are finally co-purified and subsequently identified by MS. Large-scale TAP-MS experiments have been performed for yeast (Gavin et al., 2002; Ho et al., 2002; Krogan et al., 2006), bacteria (Butland et al., 2005) or human (Ewing et al., 2007) proteins.

Interestingly, data extracted from both Y2H and TAP-MS techniques weakly overlap and result highly complementary (Aloy and Russell, 2002b; von Mering et al., 2002; Titz et al., 2004). Indeed, Y2H experiments usually reveals more transient and binary interactions, whether tandem affinity purification screenings typically detect more stable complexes, involving two or more proteins (Aloy and Russell, 2002b). However, in spite of the great and valuable amount of data provided, high-throughput experimental methods could show lack of acceptable reproducibility in the results (Collins et al., 2007).

Protein interaction networks databases

Thanks to the considerable advances in high-throughput methods gained over the past few years, massive PPIs data of various organisms have became available and currently stored in several databases. Indeed, more than 100 related repositories have been published and are now available online (Orchard et al., 2012) showing high diversity in their overall features. Firstly, their size can range from less than 100 (like in HUGE (Nakayama et al., 2002)) to millions of interactions (such as in Prolinks (Bowers et al., 2004) or STRING (Szklarczyk et al., 2015) databases). Secondly, they can gather data of thousands of organisms (like in BIND (Bader et al., 2003) which contains interactions of more than 1500 different species) or focus on a specific class (such as MPPI with only mammalian data (Pagel et al., 2005)), single organisms (like HPRD database (Keshava Prasad et al., 2009) containing only human data) or even converge on a specific type of interaction, such as human cancer associated protein interactions (i.e., HCPIN (Huang et al., 2008)) or interactions between HIV-1 and human proteins (i.e., HIV PI Database (Fu et al., 2009)). Moreover, although almost all the databases archive interactions detected by different experimental methodologies, they can also collect only data obtained using one specific technique (e.g., Yeast Interaction Protein Database with only yeast two-hybrid analysis data). Finally, despite the majority of the databases exclusively collect protein-protein interactions, a small fraction of them also includes many other types of interactions involving RNA, DNA or small molecules (Bader et al., 2003; Kerrien et al., 2012). The main overall features of the most popular repositories are summarized in **Table 2**.

Table 2. A general overview on the most popular PPIs repositories*.

Database	#Inter	#Organisms	Last Update	URL
BioGrid[a ,c]	~750000	30	2015	http://thebiogrid.org/ (Breitkreutz et al., 2008)
BIND/ BOND[a,c]	~200000	~1500	2014	http://download.baderlab.org/BINDTranslation/ (Isserlin et al., 2011)
DIP[b,c]	79646	749	2014	http://dip.doe-mbi.ucla.edu/dip (Salwinski et al., 2004)
HPRD[b]	41327	Human	2010	http://www.hprd.org/ (Keshava Prasad et al., 2009)
I2D[b,c]	900529	6	2013	http://ophid.utoronto.ca/ophidv2.204/ (Brown and Jurisica, 2007)
IntAct[a,c]	351397	> 8	2015	http://www.ebi.ac.uk/intact/ (Orchard et al., 2014)
MINT[a,c]	241458	> 30	2012	http://mint.bio.uniroma2.it/mint/ (Licata et al., 2012)
STRING[a]	> 200 MM (predicted)	2031	2015	http://string-db.org/ (Szklarczyk et al., 2015)

[*]databases with more than 500 citations by June 2015 according to Google Scholar; [a]Free to all users; [b]Free only to academic users; [c]iMEX partner; MM=million;

Since many resources are independently funded, use different identifiers and often contain redundant data from overlapping sets of publications, accessing all publicly available data (even on a specific biological or biomedical topic) is often a challenging and time-consuming task that requires the user to query multiple resources, each with a different interface (De Las Rivas and Fontanillo, 2010). Therefore, efforts to address this problem and thus integrate data from PPIs disparate repositories have recently given rise to (i) the definition of the MIMIX (Minimum Information about a Molecular Interaction eXperiment) guidelines (Orchard et al., 2007), (ii) the development of the PSI-MI XML format (i.e., a unified file format for representing PPIs data) and finally (iii) the establishment of the IMEx (International Molecular Exchange) consortium (Orchard et al., 2012). IMEx consists in an international collaboration between the major public protein interaction data providers (e.g., DIP, IntAct, MINT, I2D) cooperating in the creation of a single non-redundant set of homogeneously curated protein-interaction data available in a single search interface on a common website (http://www.imexconsortium.org/)

However, in spite of the huge amount of data available nowadays, a large fraction of them lack reliability and suffer from the integration of a large number of spurious interactions. Indeed, the estimated size of the human interactome ranges from about 130,000 (Venkatesan et al., 2009) to around 650,000 PPIs (Stumpf et al., 2008), but only around 50,000 of them have been annotated with high confidence (Mosca et al., 2013).

In silico prediction of PPIs

The considerable amount of information provided by the genomic sequencing projects and high throughput screening techniques during the last years has fostered the development of several new methods for the prediction of PPIs.

Genomic context-based methods, such as gene-neighboring (or co-localization), phylogenic profile, gene fusion, phylogenetic tree, correlated mutation or *in silico* two-hybrid are highly successful methods in which genetic information are used to derive network of protein interactions (Valencia and Pazos, 2002). Moreover, PPIs can be also predicted by integrating evidence of known interactions with information regarding sequential homology, such as in ortholog- or domain-pairs-based approaches (Lee et al., 2008; Lo et al., 2015). A different approach consists in structural similarity-based methods in which the likeliness of the interaction is determined through homology modeling of the complex structure and consequent scoring of the modeled interface using empirical (Aloy and Russell, 2002a) or statistical (Lu et al., 2002) potentials.

A novel and highly promising alternative to the above cited techniques are text-mining algorithms, based on the data screening of both scientific articles and extensive databases. One of the first examples of such tools was PubGene (Jenssen et al., 2001) followed by iHOP (Hoffmann and Valencia, 2004), iProLINK (Hu et al., 2004), GoPubMed (Doms and Schroeder, 2005), CbioC (Baral et al., 2007), Chilibot (Chen and Sharp, 2004) and lastly Whatizit (Rebholz-Schuhmann et al., 2008).

Finally, an additional option in the detection of PPIs consists in the integration of experimental or computational data in machine learning algorithms, such as support vector machines (SVM), Naïve Bayes, K-Nearest neighbors, Decision tree or Random Forest (Zahiri et al., 2013).

1.2.2. Experimental characterization of protein-protein complexes

In the last few decades, our knowledge about PPIs has grown exponentially (Ceol et al., 2008). Nevertheless, the need of integrating the binary information provided by the interactions with detailed structural data of the interacting proteins has become increasingly evident. Indeed, this synergic approach appears compulsory to extract the mechanistic basis of protein association and design new therapies to modulate these interactions.

Protein-protein complex structure determination

The significant efforts in traditional structural biology and the structural genomics projects (Montelione, 2012) as well as the important technical advances in the last few decades have produced a consistent increase in the amount of high-resolution experimental structures available in the PDB (see **section 1.1.3**). Nevertheless, X-ray crystallography and NMR spectroscopy still remain labor-intensive and time-consuming techniques especially in the determination of multi-monomer assemblies. Indeed, despite a complete or even partial experimental structure is already available for roughly 30% of human proteins, only 8% of the high-confidence identified PPIs in the human interactome have an associated complex structure (either experimentally solved or built by homology modeling) (Mosca et al., 2013; Szilagyi and Zhang, 2014).

Available structural data of protein-protein interactions is compiled in several existing databases that collect large sets of protein-protein complex structures. Some of the most renowned are DOCKGROUND (Douguet et al., 2006), PDBePISA (Krissinel and Henrick, 2007), Interactome3D (Mosca et al., 2013) and 3DCOMPLEX (Levy et al., 2006). In addition, 3did (Levy et al., 2006), PIBASE (Davis and Sali, 2005), InterPare (Gong et al., 2005), SCOWLP (Teyra et al., 2006), SCOPPI (Winter et al., 2006) and PRINT (Tuncbag et al., 2008) provide collections of high-resolution

3D structures of protein-protein interfaces classified at different levels of definition.

Characterization of protein-protein interfaces

Given the rather limited applicability and the high costs of atomic-resolution structural techniques, many approaches (e.g., cross-linking, site-directed mutagenesis or NMR chemical shift perturbation) result attractive options for a faster characterization of protein interfaces often suitable for high-throughput application.

Especially in the case of low affinity complexes, cross-linking represents a highly successful technique, which allows to freeze two proteins together while they are interacting via covalent attachment of a small cross-linker (such as carboxyl, amine, sulfhydryl or hydroxyl). This allows the creation of a stable protein pair that can be studied by gel electrophoresis, Western Blotting or mass spectrometry (MS) (Sato et al., 2011; Holding, 2015).

Site-directed mutagenesis (SDM), consisting on the exchange of a single amino acid in the protein sequence for another with different chemical properties, enables to assess the function of a single residue side chain at a specific site in the protein. Although being commonly used in functional studies on enzymes (Ahn et al., 2014), SDM has been also proved to be remarkably effective in identifying key residues in protein-protein interactions (Liu et al., 2000). Moreover, combinatorial libraries of alanine-substituted proteins can be used to rapidly identify residues important for protein function, stability and shape (Morrison and Weiss, 2001). Indeed, thanks to the non-bulky and chemically inert alanine side chain methyl group, each substitution can easily examine the contribution of an individual amino acid side chain to the functionality of the protein.

Finally, one of the most widely used approaches for probing protein-protein interfaces by NMR spectroscopy consists in the chemical shift perturbation (CSP) analysis (Hall et al., 2001). Its utility and popularity are due to the straightforward nature and its high sensitivity in mapping putative sites of interaction on a protein surface by detecting perturbations caused by the addition of the protein partner (O'Connell et al., 2009).

Many of the above mentioned analyses aim to identify hot-spot residues, which those that contribute the most to the protein-protein binding affinity, and are important for mechanistic reasons as well as for being putative targets for drug discovery. Information on experimentally determined hot-spot residues has been collected in the last few decades and is now freely available for many complexes of interest. Some well-known databases are ASEdb (Thorn and Bogan, 2001), which was the first alanine mutation database, and BID (Fischer et al., 2003), which gathers the majority of the experimentally verified hot-spots located in protein interfaces and collected from literature.

Structural features of protein-protein interfaces

Protein-protein interfaces are generally planar, although sometimes they can be protruding or concave (as in the case of enzyme/inhibitor complexes (Jones and Thornton, 1997)). They may cover a wide range of the all monomer surface area (from 5 to 30% (Stites, 1997)), spanning from less than one thousand to several thousand Å^2 (Lo Conte et al., 1999). Deeper analysis of interface surfaces reveals a rather high degree of complementarity between the complex partners (Jones and Thornton, 1996), whose extent varies depending on the type of protein interaction: permanent complexes exhibit highest complementarity, while non-obligate complexes and protein-inhibitor complexes are characterized by lower complementarity that results even worst in antigen-antibody complexes.

Although with some discrepancy, several studies provided strong evidence for a significant enrichment of aromatic (i.e., His, Phe, Tyr and Trp) and aliphatic (e.g., Leu, Val, Ile and Met) residues (Jones and Thornton, 1996) within the interfaces as well as a scarcity of charged residues (except for Arg) (Bahadur and Zacharias, 2008). Thus, it seems clear that hydrophobicity plays an important role as a stabilizing factor in protein-protein interactions.

However, no secondary structure types resulted to be essential for protein interactions, although a higher propensity to be involved in protein-protein interfaces has been observed for random coils and α-helices with respect to β-sheets (Jones and Thornton, 1996). On the other side, several structural domains involved in protein-protein interactions have been defined (e.g., Src homology, phosphotyrosine-binding (PTB), LIM domain and Sterile Alpha Motif (SAM) domain) and stored in several freely available databases such as 3did (Stein et al., 2005), CBM (Shoemaker et al., 2006), iPfam (Finn et al., 2005), PIBASE (Davis and Sali, 2005), PSIbase (Gong et al., 2005b) and SNAPPI (Jefferson et al., 2007).

Finally, the important issue of the degree of conformational change upon protein binding has received relatively low attention in literature. Despite the scarcity of proteins whose structure has been structurally determined before and after the binding and assuming the unbound structure as representative of the solvated state, various levels of conformational changes has been distinguished: (i) side chain movements alone, (ii) secondary structure segments (e.g., hinged loop), (iii) entire domains movements (e.g., enzyme/substrate complexes) or, (iv) in some extreme cases, disorder-to-order transitions (Janin et al., 2008).

Energetic details of protein-protein complexes

As expected from the significant enrichment in apolar residues found in the interfaces, the hydrophobic effect provides a significant contribution to the protein-protein interaction (Tsai et al., 1996, 1997), due to desolvation energy, associated to the removal of the solvent from the interface upon binding, as well as because of tight, unspecific and short-distance van der Waals contacts created between non-polar residues. These are generally clustered in several patches whose size ranges from 200 to 400 Å, reaching even 3000 Å in some cases (Lijnzaad and Argos, 1997).

Besides hydrophobicity, electrostatics is the other significant force involved in protein-protein interactions (Xu et al., 1997; Sheinerman et al., 2000). Indeed, apart from their influence on protein-protein affinity and specificity, long-range electrostatic forces have been proposed to have an influence on the binding process (Sheinerman et al., 2000) (i.e., pre-orienting protein partners, promoting encounter complexes formation and therefore accelerating the rate of association) as well as on its lifetime.

Together with their involvement in electrostatic interactions, polar groups at interfaces have been found to be regularly involved in hydrogen bonds (one per about 200 Å^2 of buried surface area (Jones and Thornton, 1996)), either interacting with protein groups of the complex partner or with water molecules located at the interface. Indeed, water molecules are often found specifically located at the protein-protein interfaces, and play a major role in polar interactions that stabilize the complexes (Rodier et al., 2005).

1.2.3. Theoretical models for protein binding

During the last two centuries considerable research effort has been focused on understanding the mechanism of association between proteins. This led to an increasingly comprehensive knowledge of the physicochemical properties of the binding (e.g., thermodynamics and kinetics) and the consequent postulation of various theoretical models aimed to accurately describe such process.

Lock-and-key paradigm

The first attempt to explain the protein complex binding consisted in the *lock-and-key* paradigm that was formulated in 1894 even before any structural knowledge of proteins. Developed by H.E. Fisher, it offered a schematic and static representation of protein interactions emphasizing the importance of steric complementarity between the partners to achieve affinity and specificity upon the binding (Fischer, 1894).

The lock-and-key paradigm was later found to provide an excessively simplistic description of binding, which is not adequate to describe the expected conformational flexibility of the interacting proteins in most of the cases.

Induced-fit mechanism

During the last century, the *lock-and-key* paradigm was replaced with a more dynamic representation of protein binding, expressed in the so-called *induced-fit* mechanism (Koshland, 1958). According to this model, proteins initially interact in an unbound conformation, while the bound state is induced by the physical-chemical environment provided by the protein partner. Indeed, *induced-fit* paradigm is consistent with the conformational flexibility frequently observed during binding (Echols et al., 2003) as well as the promiscuity of some proteins in their interactions (Tidow and Nissen, 2013).

Nevertheless, the *induced-fit* mechanism does not explain the intrinsic plasticity of several systems, existing as an ensemble of conformations dynamically fluctuating between them, as supported by some X-ray and cryo-electron microscope images, kinetics studies and, above all, NMR data showing a repertoire of conformational states of unbound protein, including conformations similar to the bound state (Boehr et al., 2009; Esteban-Martín et al., 2012).

Conformational-selection and population-shift model

Firstly suggested in 1964 by Straub (Straub and Szabolcsi, 1964) and experimentally supported by Zavodszky et al. in 1966 (Závodszky et al., 1966), the *conformational-selection* model initially postulated that the unbound proteins naturally sample a variety of conformational states, a subset of which are suitable to bind the other protein.

This original formulation was partially reassessed by Frauenfelder, Sligar and Wolynes over 25 years later (Frauenfelder et al., 1991) and finally formalized in 1999 (Ma et al., 1999; Tsai et al., 1999; Kumar et al., 2000): in solution proteins exist in a range of conformations which regularly interchange between high populated lowest-energy conformations and low populated higher-energy ones that are more suitable to bind the bound state. However, given their optimal geometry and physical-chemical complementarity, the high-energy bound-like conformations get preferentially selected and stabilized by the interacting partner, thus shifting the population of the protein microstates in favor of the bound state.

Extended Conformational-selection model

The recent growth in volume and precision of data related to protein dynamics suggested that the distinction between the original *conformational-selection* and *induced-fit* models is not absolute (Grunberg et al., 2004; Wlodarski and Zagrovic, 2009).

These findings provided support for an extended version of the original *conformational-selection* model where both selection and adjustment-type steps would follow each other. Thus, proteins in solution would contain an ensemble of conformational states, not necessarily structurally similar to the bound state, available for the mutual selection and adjustment. As binding proceeds, the partners' conformations change, as well as their position on the energy landscape, whose shape results in turn altered by increasing adjustments of the binding environment related to emerging electrostatic and water-mediated hydrogen-bonding signals between the protein partners (Kovacs et al., 2005; Antal et al., 2009). Upon this mutual adaptation, although converging to a common end-state, protein partners can follow alternative 'binding trajectories' (Tsai et al., 2008; Antal et al., 2009) (i.e., sequences of conformational selection and adjustment steps), where every step of the encounter by each subunit depends on the proceeding conformational change by the protein partner, generating a kind of 'interdependent protein dance' (Antal et al., 2009).

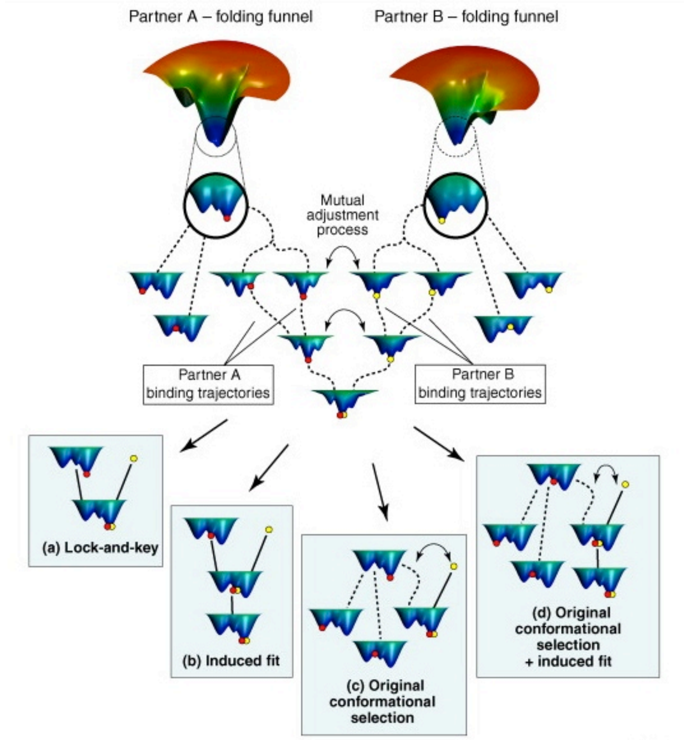

Partner A – folding funnel
Partner B – folding funnel
Mutual adjustment process
Partner A binding trajectories
Partner B binding trajectories
(a) Lock-and-key
(b) Induced fit
(c) Original conformational selection
(d) Original conformational selection + induced fit

Figure 6. Schematic representation of the extended conformational-selection model. (a) the classical lock-and-key model (b) the classical induced-fit (c) the classical conformational-selection model (d) the conformational-selection-plus-induce-fit model. From Csermely et al., 2010

All in all, protein binding process would be triggered by the formation of transient encounter complexes, mainly stabilized by electrostatic forces (Tang et al., 2006; Bashir et al., 2010), whereas its completion would involve *induced-fit*-based events including (i) anchor residues rearrangement (Rajamani et al., 2004), (ii) hinge and hinge-like motions (Ma et al., 2002), (iii) rearrangements of crucial nodes located between communities of amino acids networks responsible for the structural reorganization of each subunit upon the binding (Bode et al., 2007; Del Sol et al., 2007; Csermely, 2008; Sethi et al., 2009).

Within the above described binding scenario, the mechanisms previously proposed (i.e., *lock-and-key* and *induced-fit*) are not rejected but reinterpreted as special cases of a unique binding paradigm (**Figure 6**). Thus, the *lock-and-key* model would represent the case in which both the partners are either rigid or have exactly matching binding surfaces. Moreover, the *induced-fit* mechanism would be interpreted as an evolution of the extended *conformational-selection* type binding scenarios triggered by some specific binding conditions: (i) the occurrence of the strong and long-range or directed interaction, such as ionic forces or hydrogen bonding (Csermely et al., 2010); (ii) high partner's concentration (Junker et al., 2009; Weikl and von Deuster, 2009) or (iii) large difference in size or cooperativity between the complex partners (Pereira-Leal et al., 2006).

1.2.4. Computational methods for protein-protein complex structure prediction

The number of experimentally determined protein structures accounts only for a tiny fraction of the massive amount of protein known and sequenced proteins (Anishchenko et al., 2014) (see **section 1.1.3**), and this discrepancy between sequence and structural data results even more evident when considering protein-protein complexes (see **section 1.2.2**).

Nevertheless, computational approaches provide useful resources to bridge both these gaps. Indeed, they not only succeed in modeling isolated protein structures, using experimentally determined structures as templates (see **section 1.1.3**), but also, clearly more challengingly, provide the structural characterization of unknown multimeric protein structures either using template-based modeling (TBM) or *ab initio* docking (**Figure 7**).

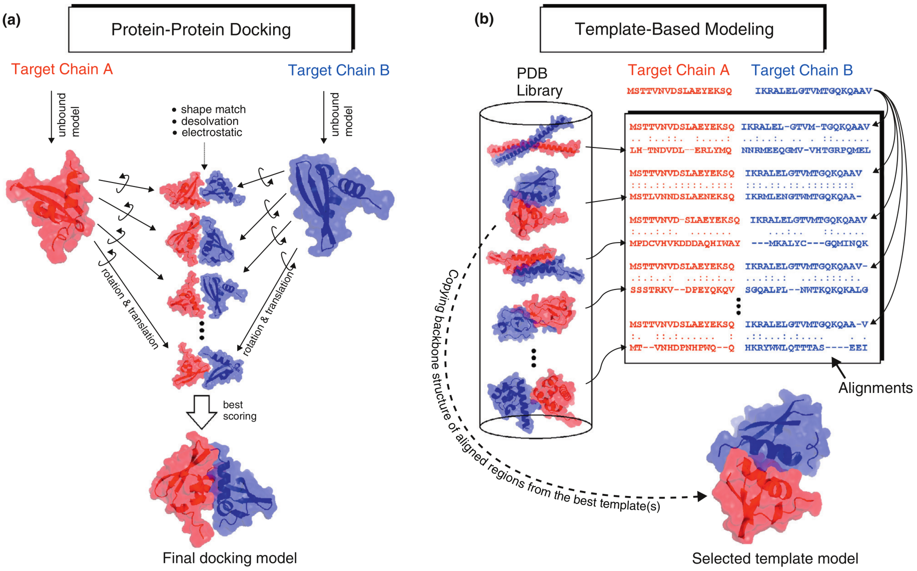

(a)
Protein-Protein Docking
Target Chain A
Target Chain B
unbound model
unbound model
• shape match
• desolvation
• electrostatic
rotation & translation
rotation & translation
best scoring
Final docking model
(b)
Template-Based Modeling
PDB Library
Target Chain A
Target Chain B
MSTTVNVDSLAEYEKSQ
IKRALELGTVMTGQKQAAV
MSTTVNVDSLAEYEKSQ IKRALEL-GTVM-TGQKQAAV
LH-TNDVDL--ERLYMQ NNRMEEQGMV-VHTGRPQMEL
MSTTVNVDSLAEYEKSQ IKRALELGTVMTGQKQAAV
MSTLVNNDSLAENEKSQ IKRMLENGTWMTGQKQAA-
MSTTVNVD-SLAEYEKSQ IKRALELGTVMTGQKQAAV
MPDCVHVKDDDAQHIWAY ---MKALYC---GQMINQK
MSTTVNVDSLAEYEKSQ IKRALELGTVMTGQKQAAV-
SSSTRKV--DPEYQKQV SGQALPL--NWTKQKQKALG
MSTTVNVDSLAEYEKSQ IKRALELGTVMTGQKQAA-V
MT--VNHDPNHPWQ--Q HKRYWWLQTTTAS----EEI
Alignments
Copying backbone structure of aligned regions from the best template(s)
Selected template model

Figure 7. Two principal protocols for protein complex structure prediction. Red and blue represent sequences and structures of two individual chains. (a) *Ab initio* docking and (b) Template-based modeling (TBM) methods. From Szilagyi and Zhang, 2014.

Template-based modeling of protein complexes

Homology modeling of protein-protein complexes appears as an extension of TBM of isolated proteins and consists in building the structure of a protein-protein complex by using as template other related protein-protein complexes whose structure has been experimentally solved. This approach have become increasingly popular over the last few years (Vakser, 2013), mainly supported by the awareness that accurate PPI models can be yielded using proper templates (Aloy and Russell, 2002a) as well as the potential availability of templates suitable to model nearly all PPIs (Kundrotas et al., 2012).

In the majority of TBM protocols, the complex structure templates are generally detected by homology-based sequence alignments, applying the technique of threading (Bowie et al., 1991), as observed in MULTIPROSPECTOR (Lu et al., 2002) or COTH (Mukherjee and Zhang, 2011) pipeline. However, since the components isolated structures are typically known, a growing number of approaches (e.g., ISEARCH (Gunther et al., 2007), iAlign (Gao and Skolnick, 2010), KBDOCK (Ghoorah et al., 2011), PrISE (Jordan et al., 2012), SCPC (Koike and Ota, 2012), ProBiS (Konc et al., 2012), PRISM (Tuncbag et al., 2012), TrixP (von Behren et al., 2013)), typically referred as template-based structure-comparison approaches (Zhang et al., 2012), exploit structural alignment techniques for the alignment of backbone, secondary structure, and/or coarse-grained elements of the overall structure or the interface alone.

A standard procedure of conventional template-based complex modeling, starting from the sequence of the complex components, consists of four steps which are essentially identical to those used in TBM of isolated protein: (i) finding known structures related to the sequence to be modeled; (ii) aligning the target sequences to the template structure; (iii) constructing structural frameworks by coping the aligned regions of the template structures;

(iv) constructing the unaligned loop regions and adding side chain atoms.

As the quality of the TBM models essentially depends on the accuracy of the template identification whereas the full-length complex structure construction and refinement are in general more complicated and time-consuming, most current TBM algorithms (i.e., COTH (Mukherjee and Zhang, 2011), SPRING (Guerler et al., 2013), MULTIPROSPECTOR (Lu et al., 2002), HOMBACOP (Kundrotas et al., 2008), Struct2Net (Singh et al., 2010) and iWRAP (Hosur et al., 2011)) focus on the identification of templates, while only a few methods, such as M-TASSER (Chen and Skolnick, 2008) perform a full pipeline. All these algorithms mainly differ in the strategies applied during the complex template identification and structure combination step, which typically are (i) dimeric threading, (ii) monomer threading and oligomer mapping or (iii) template based docking, as summarized in **Figure 8**.

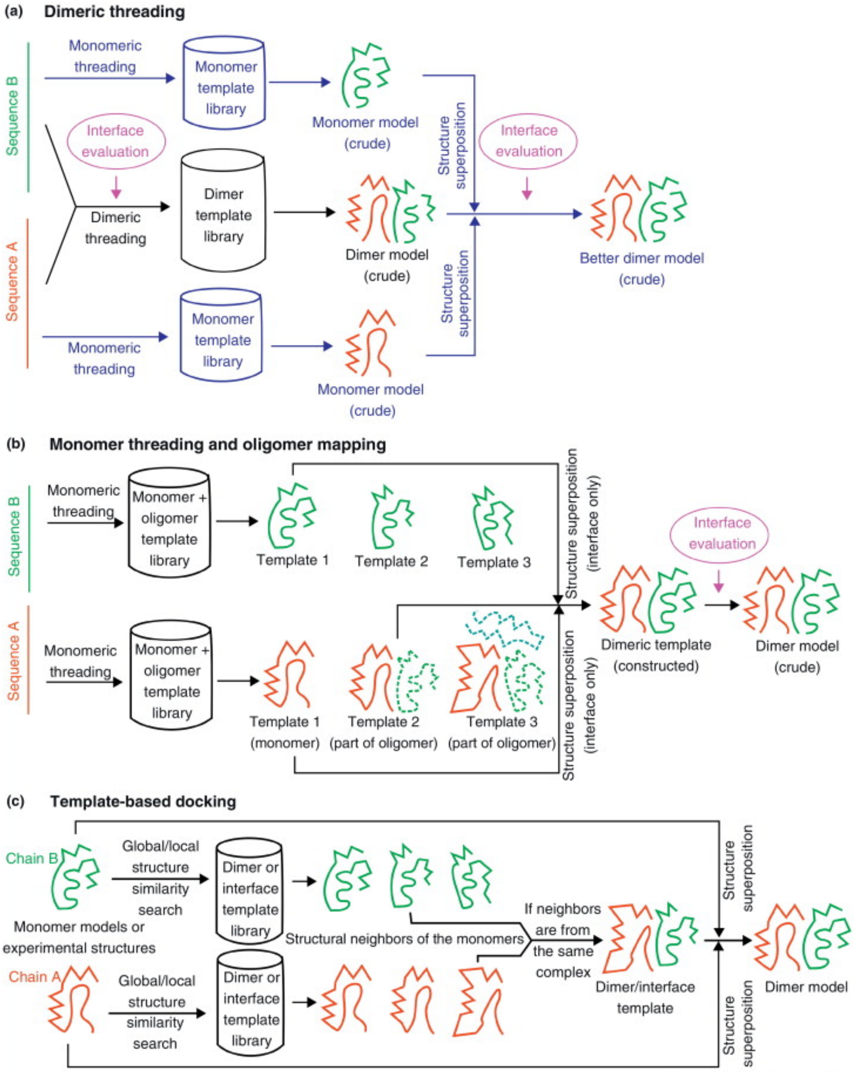


Figure 8. Flowcharts for the three representative template-based complex structure prediction strategies. (a) Dimeric threading method. (b) Monomer threading and oligomer mapping. (c) Templatebased docking. From Szilagyi and Zhang, 2014.

Ab initio modeling of protein complexes

Protein-protein docking methods aim to build complex models by assembling known structures of the interacting components, previously predicted or solved in the unbound state. They basically consist in an initial exhaustive search and consequent selection of various binding orientations and thus ideally provide a realistic description of the association process and the complex energy landscape.

Since the first attempt, performed by Wodak and Janin in the late 1970s (Wodak and Janin, 1978), the number of protein-protein docking programs is continuously increasing. All the currently used docking frameworks address the modeling task usually through two (or three) independent and consecutive processes: (i) sampling of the rotational and translational space of the two interacting proteins; (ii) scoring of the generated docking orientations; and finally (iii) an optional refinement and minimization of the complexes. Nevertheless, they typically differ in (i) the sampling method implemented, (ii) the scoring function used to rank the docking models, and (iii) the strategy applied for the treatment of protein flexibility. According to these key features, all the docking methodologies developed so far can be divided onto two main categories, namely the geometry and energy minimization-based docking methods.

In the heart of the geometry-based docking methods is the steric complementarity at the protein-protein interface. Thus, several simplified protein models and approximate functions have been devised in order to find the best fitting between interacting surfaces. A wide-used approach consists in the discretization of the interacting protein into grids and the consequent exploration of the rotational and translational space at a certain resolution by searching for the best correlation between both grids. However, although this approach clearly reduces the computational costs of the sampling with respect to a full-atom representation of the interacting proteins, an exhaustive conformational exploration remains still prohibitive even for standard size complexes.

Nevertheless, the sampling of both the translational or rotational space can be dramatically speed up (around four orders of magnitude) by computing the correlation function between the two discrete grids, thanks to the application of Fast Fourier Transform (FFT) algorithms (Katchalski-Katzir et al., 1992). Since its first implementation in MolFit, this technique has been applied in different docking methods, where geometry has been combined in different ways with electrostatics and physico-chemical terms during the sampling. For instance FTDock (Gabb et al., 1997) added an electrostatic grid; PIPER (Kozakov et al., 2006), GRAMM-X (Tovchigrechko and Vakser, 2006) and BIGGER (Palma et al., 2000) introduced pairwise interaction potentials; ZDOCK (Chen et al., 2003a) implemented (in successive versions) geometry-based complementarity, electrostatics, desolvatation and statistical potentials terms (Pierce et al., 2011). Other successful shape-based methods use Fourier Transform (FT) on the rotational instead of the translational space as previously described. Among such approaches, Hex (Ritchie and Kemp, 2000) uses 2D spherical harmonics to represent the surface of the interacting proteins whereas FRODOCK (Garzon et al., 2009) is based on fast rotational matching (FRM), where for each translational point, the rotational search is accelerated by Fourier Transform (FT) using radial spherical harmonics.

Another well-known geometric-based method is PatchDock (Schneidman-Duhovny et al., 2005) based on the extraction of geometric features from the interacting proteins, and the use of geometric hashing algorithms to compute the complementarity between surfaces (which are described by shape representations like knobs, hole and flat areas).

Finally, the relatively low computational cost associated to geometric-based docking algorithms makes them suitable for the application to multi-molecular docking for the prediction of small multi-protein assemblies. Noteworthy examples are CombDock (Inbar et al., 2003), where the combinatorial problem during the sampling is solved by the application of graph-based algorithm, or SymmDock (Schneidman-Duhovny et al., 2005), which allows the prediction of multimeric complexes with a given rotational symmetric starting from its asymmetric unit.

Despite the remarkable speed and exhaustive sampling, a major drawback of geometry-based docking methods arises from the approximations made in protein shape and energy description as well as the null or limited consideration of protein flexibility during the binding. Alternative solutions to this limitation, represented by energy minimization-based docking methods, will be described in **section 1.3.3.**

In parallel to the above mentioned docking protocols, in which scoring is implicitly considered within the sampling procedure, many other programs, exclusively specialized on independent scoring of docking poses generated in a previous rigid-body step, has been developed. One of the most successful scoring schemes is pyDock (Cheng et al., 2007), which uses an energy function composed of van der Waals, electrostatics and atomic solvation parameter (ASP)-based desolvation energy. Moreover, several sets of residue-based potentials have been recently reported, such as SIPPER (Pons et al., 2011) or PIE (Ravikant and Elber, 2010), whose major benefit is the speed of the calculation, especially those defined at residue level.

Ab initio docking versus TBM methods

Compared to a*b initio* docking, the main advantage of TBM lies in the fact that only the sequence and not the structure of the monomer components are pre-required. Moreover, as the models are built based on the complex templates that are in a bound form (in contrast to the unbound structures used in *ab initio* docking), TBM methods are not sensitive to the type of the complex (large or small interface area, permanent or transient interaction) and to the extent of conformational changes upon binding.

Nevertheless, the most crucial limitation of TBM consists in the complete neglect of mutations or post-translational modifications effects, which might seriously perturb monomer components folding, modulate the interaction or create new aberrant interactions. In addition, as happens with isolated proteins (see **section 1.1.3**), the quality of a model is strikingly subjected to the target-template sequence identity (Aloy et al., 2003; Launay and Simonson, 2008; Kundrotas and Vakser, 2013) and thus modeling of interactions in the absence of close homologous templates is still a challenging task. Indeed, although it was reported to be ideally possible to find templates for nearly all known interactions (Kundrotas et al., 2012), it was recently revealed that the quality of the resulting models appears to be quite poor and significantly worse than those obtained by *ab initio* docking in cases where the available template shares a low sequence identity with the target (i.e., below 30%) (Negroni et al., 2014).

1.2.5. Evaluation of protein complex structural prediction

In order to be properly evaluated, any docking approach should generally be tested on a statistically significant, non-redundant and representative subset of all the complexes with known structure (i.e., docking benchmark sets) and objectively compared to other existing approaches on such benchmark sets or even better in blind community-wide assessments, such as CAPRI (Critical Assessment of Predicted Interactions) (Janin et al., 2003).

Protein-protein docking benchmarks

The widely used benchmark sets of protein-protein complexes were developed in Weng's (Chen et al., 2003b) and Vakser's (Gao et al., 2007) groups. After several updates, the current versions of the two benchmarks (version 5.0 (Vreven et al., 2015) with 230 entries and version 2.0 (Anishchenko et al., 2015) with 165 entries, respectively) contain more than one hundred complexes of co-crystallized proteins and either their isolated components (unbound structures) or arrays of low sequence identity homology-based models.

Moreover, several groups compiled decoy sets of docking models containing false positive matches of proteins that result useful in the optimization of potentials and scoring functions for the discrimination of false positive predictions. The ones generated by ZDOCK, FTDock and Rosetta are publicly available at the web pages of the respective groups, while another one was recently provided by Vakser lab (Liu et al., 2008).

In the last few years, great efforts have been devoted to assembly datasets combining structural and energetic information. Indeed, a non-redundant set of 144 protein-protein complexes for which not only the unbound and bound structures but also their dissociation constants are available was recently published (Kastritis et al., 2011) and consequently updated (Vreven et al., 2015) with 35 additional cases. Moreover, in 2012 Fernández-Recio's group published SKEMPI, a database containing data on the changes in thermodynamic parameters and/or kinetic rate constants upon more than 3000 mutations for protein-protein interactions of which at least one co-crystallized complex structure has been solved and is available in the PDB (Moal and Fernandez-Recio, 2012).

All together these datasets offer remarkable tools for the development, assessment, optimization and comparison of new docking algorithms.

CAPRI (Critical Assessment of Predicted Interactions)

CAPRI, established in 2001 (Vajda et al., 2002), currently consists in a community-wide scientific experiment, conducted on a discretionary basis, which allows the comparison of different docking methods on a set of targets (i.e., experimentally determined complex structures unknown to the participants) based on two different prediction assessments, namely *predictors* and *scorers*.

Thus, in a given CAPRI Round, *predictors* are asked to generate, score and finally submit a total of ten own complex models starting from the separately crystallized structures of the complex components, or their homologous supplied by the CAPRI organizers. In a second step, the *scorers* are invited to evaluate a common pool of docking models made up from the contributions of different participating groups (*uploaders*) and finally submit their own ten best ranking ones. At the end of each Round, the 10-model sets submitted by each group of the *predictors* and *scorers* community are evaluated by the organizers by comparison with the corresponding complex structure which is still unpublished and made known only to the organizers. The regular criteria for the evaluation of protein-protein interaction models are described in Figure 1 of Lensink et al., 2007 and Table II of Lensink and Wodak, 2010. More recently, they have been slightly adapted to the assessment of protein-peptide interaction (http://www.ebi.ac.uk/msd-srv/capri/round28/round28.html).

Since its inception, five CAPRI editions were completed corresponding to 34 prediction Rounds and a total of more than 100 targets. Moreover, in the last CAPRI edition, taken place in the years 2010-2012, in addition to the standard protein assemblies predictions, several different assessments were proposed (including binding affinity, sugar binding and interface water molecule prediction) (Lensink and Wodak, 2013; Moretti et al., 2013; Lensink et al., 2014). The analysis of the docking results obtained in all the previous CAPRI editions offer a useful resource to track the evolution of the protein docking field (Mendez et al., 2003; Mendez et al., 2005; Lensink et al., 2007; Lensink and Wodak, 2010, 2013), as well as to identify its main challenges and its major ventures for the years to come.

1.2.6. Interface and hot-spot residues prediction

The docking methods described above aim to model the binding mode of two interacting proteins at atomic resolution. However, given the accuracy limitations of these methods, especially in some difficult cases, sometimes it may be easier and more reliable trying first to identify the residues that are involved in the interaction, which in turn could be also helpful in the target characterization step during a drug discovery program.

Identification of potential protein binding sites

Taking into account specific properties, which distinguish protein-protein interfaces from the rest of the protein surface (Jones and Thornton, 1996, 1997), diverse binding site prediction methods have been developed in the last few decades.

Some of the better-known are InterproSurf (Negi et al., 2007), based on solvent accessibility and statistical potential; PINUP (Liang et al., 2006), using an empirical scoring function; ProMate (Neuvirth et al., 2004), combining residues types, secondary structure and sequence conservation; WHISCY (de Vries et al., 2006), related to conservation and surface properties; ISIS (Ofran and Rost, 2007a), identifying interacting residues from protein sequence only; and finally ODA (Fernandez-Recio et al., 2005), based on pyDock (Cheng et al., 2007) desolvation energy.

A more specific interface analysis is the one supplied by PRISM server (Ogmen et al., 2005; Keskin et al., 2008), which detect the specific interaction between two given proteins. Finally, although phylogenetic conservation alone is often insufficient to reliably predict protein binding sites, it can be successfully combined with other interface properties. ConSurf (Glaser et al., 2003) and SCORECONS (Valdar, 2002) are web servers that can provide these data.

Detection of protein hot-spot residues

Despite protein-protein interfaces are often large, flat and do not have clear binding cavities (Jones and Thornton, 1997; Chakrabarti and Janin, 2002), it has been reported that just a small number of residues, typically referred to as *hot-spots*, are responsible for the stabilization of the complex (i.e., contributing in more than 1-2 kcal to the overall complex binding energy) (Clackson and Wells, 1995), and thus are interesting targets for drug discovery or for a better understanding of the mechanism of association between proteins. These findings have inspired the development of a large number of computational tools focused on the prediction of such *hot-spot* residues as well as the compilation of different databases.

The vast majority of the predictive methods reported until now strongly relied on the availability of the complex structure. Some renowned examples are energy-based tools, such as ROBETTA (Kortemme and Baker, 2002), FoldX (Schymkowitz et al., 2005), HSPred (Lise et al., 2011) or Molecular Dynamics (MD) with generalized Born model in a continuum medium (Moreira et al., 2007), supported in several MD platforms (e.g., AMBER (Salomon-Ferrer et al., 2013) and GROMACS (Pronk et al., 2013)), which are based on computational alanine scanning of protein-protein interfaces and subsequent evaluation of the change in binding affinity.

Other valuable approaches are machine learning-based tools. Some of the most recently reported methods are KFC2 (Zhu and Mitchell, 2011), based on interface solvation, atomic density and plasticity features; PCRPi (Assi et al., 2010), combining sequence conservation, energy score and contact number information; PPI-Pred (Bradford and Westhead, 2005), considering surface shape and electrostatics; MINERVA, which weights atomic packing density and hydrophobicity (Cho et al., 2009) or a recent neural network-based protocol (an adaptation of ISIS), which combines several interface features such as sequence profiles, solvent accessibility and evolutionary conservation (Ofran and Rost, 2007b). Another well-known machine learning-based tool is PocketQuery web-server (Koes and Camacho, 2012), which provides an assortment of metrics (including changes in solvent accessible surface area, energy-based scores, and sequence conservation) extremely useful for *hot-spots*, anchor residues and hot regions prediction.

Empirical formula-based methods are also used instead of machine learning algorithms. Some example are MAPPIS (Shulman-Peleg et al., 2007), whose prediction relies on the evolutionary conservation of hot-spots in the interface along the members of a given family; HotSpot Wizard (Pavelka et al., 2009), based on the integration of structural, functional and evolutionary information provided by several databases; DrugScore^{PPI} (Kruger and Gohlke, 2010), performing fast and accurate alanine scanning calculation derived from experimental alanine scanning results; iPRED (Geppert et al., 2011), using pairwise potentials atom types and residue properties; APIS (Xia et al., 2010), where the *hot-spots* identification is performed by combining residue physical/biochemical features, such as protrusion index and solvent accessibility; and finally

HotPoint (Tuncbag et al., 2010) that incorporates a few simple rules consisting of occlusion from solvent and total knowledge-based pair potentials of residues. Very recently, ECMIS (Shingate et al., 2014) has been reported, using a new algorithm combining energetic, evolutionary and structural features.

In spite of their high accuracy in the identification of *hot-spot* residues, a major limitation of all the above cited tools lies in the mandatory requirement of the protein-protein complex structure (or that of a homologous one). By the contrary, in cases with no available complex structure, very few *hot-spot* prediction methods have been reported until now. One of them, pyDock (Cheng et al., 2007) module pyDockNIP (Grosdidier and Fernandez-Recio, 2008) is based on protein-protein docking simulations and computes the propensity of a given residue to be located at the interface in the 100 lowest-energy rigid body docking solutions. A novel computational tool, laying in the same category, is SIM (Agrawal et al., 2014), which consists in predicting *hot-spot* residues involved in evolutionarily conserved protein-protein interactions starting from the unbound protein structure.

Besides the huge number of predictive methods and web-servers available, in the last few decades the undisputed biological relevance of *hot-spot* residues has also inspired the creation of the several *hot-spots* databases based on computational prediction, including HotRegion (Cukuroglu et al., 2012), HotSprint (Guney et al., 2008) and PCRPi-DB (Segura and Fernandez-Fuentes, 2011).

1.3. Protein conformational plasticity

Proteins are not static objects. Their structure in solution can be described as ensembles of variously heterogeneous conformations, whose transitions between one to another are mainly related to environmental changes (e.g., temperature (Caldwell, 1989), pH (Di Russo et al., 2012), voltage (Navarro-Polanco et al., 2011), ion concentration (Negi, 2014)) or post-translational modifications (Karunatilaka and Rueda, 2014) (e.g., phosphorylation) and occur on a variety of length scales (typically from tenths of Å to nm) and time scales (ranging from ns to s). This new dynamic perspective has been conceptually synthesized in an *energy landscape* paradigm, in which highly populated protein states and the transitions between them can be described by the depths of energy wells and the heights of energy barriers, respectively (Frauenfelder et al., 1991).

However, although the dynamic nature of proteins is absolutely unquestionable, its description and incorporation into an intuitive perception of protein function remain challenging. Indeed, this status results further exacerbate by the fact that although conformational sub-states (located in energy well) and their rates of interconversion can be detected experimentally (i.e., from the relaxation of the nuclei after excitation through NMR data), a description of the transition pathway on an atomic-scale is out of the reach for any currently available experimental technique because of the extremely low probability and short lifetime of the high-energy conformers. On the contrary, computational modeling has the unbeatable advantage to offer an exhaustive description of protein plasticity.

1.3.1. Computational exploration of protein plasticity

Despite being theoretically accessible with computational techniques, in-depth characterization of proteins in action is not trivial. Indeed, their various dynamics processes cover an extensive spectrum of amplitudes and energies as well as a huge time-scale range spanning 13 orders of magnitude, from femtoseconds to hours. Thus, from the fastest to the slowest motions one can find covalent bond vibrations occurring in femtoseconds; side chain rotations and loop flips usually on the pico- to nanosecond timescale; large domain motions, macromolecular associations and protein folding that might take several minutes or even hours (**Figure 9**).

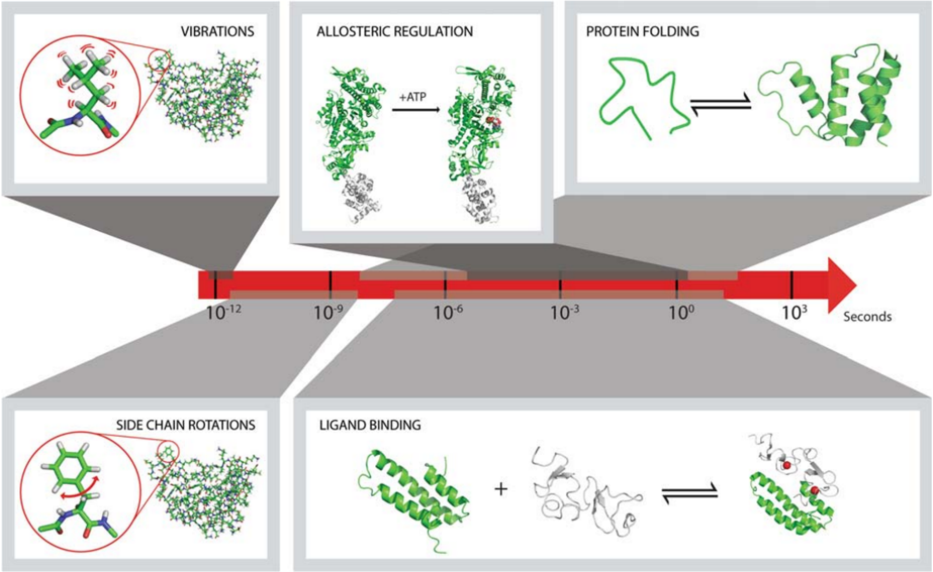


Figure 9. The timescale of the conformational events that underlie protein flexibility: from the fast vibrations of covalent bonds to slow protein (un)folding events. After Teilum et al., 2009.

Exploring protein plasticity by Molecular Dynamics (MD)

Since the publication of the first Molecular Dynamics simulation of a protein in 1977 (McCammon et al., 1977), specific aspects of biomolecular structure, kinetics and thermodynamics has been investigated via MD (such as macromolecular stability (Tiana et al., 2004), conformational and allosteric properties (Kim et al., 1994), enzyme activity (Warshel, 2003), molecular recognition (Wang et al., 2001), ion and small molecule transport (Roux, 2002), protein association (Abriata and Dal Peraro, 2015) and folding (Day and Daggett, 2003)). These finding provided significant advances in several research fields ranging from drug (Kerrigan, 2013) and protein design (Kiss et al., 2013) to material sciences and biophysics.

MD simulations can provide a detailed description of the thermodynamic properties and time-dependent phenomena of proteins through discrete integration of Newton's equation of motions (Lindahl, 2008). Each simulation requires only three items: (i) the initial coordinates of the system, (ii) a *force field* and (iii) a solvent model.

The initial coordinates are generally obtained from experimental structures (e.g., NMR or X-ray) or from homology-based models. The *force field* model consists in sets of *ab initio* and empirical parameters combined with detailed mathematical functions, which basically provide the parametrization of the energy surface of the protein. Although each force field uses own parameters sets and slightly different energy terms to calculate a system potential energy, globally all consider that the potential energy of the system is additive and composed of a potential from bonded (or covalent) and non-bonded (non covalent) interactions. Several force field models have been developed so far, including the popular latest CHARMM (MacKerell et al., 1998), AMBER (Ponder and Case, 2003) and GROMACS (Oostenbrink et al., 2004) force field versions. Although giving quite consisting results among each others (Price and Brooks, 2002), some of the existent force fields can lack agreement with experimental measurements (Beauchamp et al., 2012).

The next crucial step in MD simulations is the decision upon the solvent model (Xia et al., 2002). The simplest and commonly used is explicit solvation, in which water molecules and ions are explicitly represented in the force field (Bizzarri and Cannistraro, 2002), such as in TIP3P, TIP4P, TIP5P, SPC and SPC/E models. However, given the high computational costs of such models, sometimes an implicit consideration of the solvent is preferred (Orozco and Luque, 2000; Tsui and Case, 2000; Simonson, 2001; Hassan and Mehler, 2002; Lee et al., 2002). Here the solvent is treated as a continuous medium having the average properties of the real solvent. Much longer trajectories are thus accessible, although with lower accuracy, especially in protein complexes and conformational analysis (Roe et al., 2007; Yeh and Wallqvist, 2009).

All this variety of force fields and solvent models is implemented in a considerable amount of available software packages, such as CHARMM (Brooks et al., 2009), AMBER (Case et al., 2005), GROMACS (Pronk et al., 2013) and NAMD (Phillips et al., 2005). They typically share common basic features but also bear peculiar strengths and weaknesses, regarding force field, flexibility, licensing models, functionality and scalability (Salsbury, 2010).

1.3.2. Beyond standard Molecular Dynamics

The introduction of special-purpose machines such as Anton (Shaw et al., 2010), the adaptation of MD codes to specialized graphics-processing-units (GPUs) (Friedrichs et al., 2009) and the evolution of parallel codes (Pronk et al., 2013) have enormously increased the time scales accessible by fully atomistic MD. Current MD simulations can perform trajectories lasting up to a few ms, which enables the description of protein folding and unfolding processes (Shaw et al., 2010) as well as simulations of entire molecular machines composed by large multiple subunits (e.g., the nicotinic acetylcholine receptor (Kraszewski et al., 2015), ATP synthase (Bockmann and Grubmuller, 2002), virus capsids (Zhao et al., 2013) or the entire ribosome (Sothiselvam et al., 2014)).

However, although these and future techniques are likely to make great progress in the applicability of MD, its routine applicability is still limited by the intense computational demands that are required for atomic-detailed simulations longer than microsecond scale in medium-sized systems. This makes it virtually prohibitive the exploration of slow molecular motions that occur at the scale of the whole protein using conventional MD simulations and thus has fostered the development of alternative sampling methods, which lead to a more exhaustive exploration of the conformational space at lower time and computational costs.

Noteworthy solutions to leverage the present-day power of atomistic MD simulations consisted on the application of novel enhanced sampling algorithms, such as Umbrella Sambling (Patey and Valleau, 1975), Replica-Exchange Molecular Dynamics (REMD) (Sugita and Okamoto, 1999), Metadynamics (Laio and Parrinello, 2002), steered MD (Isralewitz et al., 2001), milestoning (Faradjian and Elber, 2004), accelerated MD (Hamelberg et al., 2004), transition path sampling (Bolhuis et al., 2002) and their many combinations and derivatives in which large energy barriers are artificially reduced, allowing proteins to shift between conformations that would not be accessible within the time scales of conventional MD.

Alternative widespread strategies are based on Normal Mode Analysis (NMA) approach, which allow to extract large-amplitude macromolecular motions (expected to be involved in functionally important transition pathways) by approximating the complex dynamical behavior of a macromolecule to a simple set of harmonic oscillators vibrating around a given equilibrium conformation (Brooks and Karplus, 1985).

Metadynamics (MetaD)

During the past decade, Metadynamics (MetaD) (Laio and Parrinello, 2002), especially in the well-tempered formulation (Barducci et al., 2008), has become increasingly popular as a powerful approach to accelerate rare events (i.e., those which occur infrequently in a simulation trajectory, regardless of the trajectory timescale) in macromolecular systems, by biasing specific degrees of freedom (generally referred as collective variables, CVs) and computing multidimensional free energy surfaces (FESs) as a function of such CVs. Thus, the diffusion in the CVs space is enhanced by disfavoring already visited regions through the cumulative addition of a repulsive Gaussian potential to the physical force field potential, which flatters the FES and thus prevents the system from being trapped in local free energy minima (**Figure 10**). This framework successfully produces the exploration of new reaction pathway without *a priori* knowledge of the landscape and the following estimation of the FES without any CVs bias. Nevertheless, given the pivotal relevance of the CVs during all the procedure, the accuracy of the results is dramatically dependent on their appropriate choice. Generally speaking, they should (i) enable to distinguish between the initial and final state of the transition studied, (ii) include all the slow modes of the system and (iii) be limited in number. However, the identification of the correct CVs is usually far from being trivial, and hidden degrees of freedom, which may not be accurately described by the chosen CVs, often frustrate the sampling and thus limit the extent of convergence and the accuracy of results (Sutto et al., 2012).

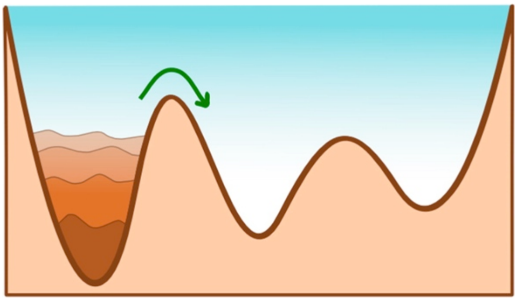


Figure 10. Pictorial representation of the way the MetaD algorithm fills the free energy landscapes. From Cavalli et al., 2015

A more efficient approach recently developed, generally referred as PTMetaD, consists in manipulating all degrees of freedom in a more general way (e.g., by increasing system temperature) by combining MetaD with parallel tempering (PT) protocols. Here, a series of replicas of a system are simulated at different temperatures, and periodical exchanges between adjacent replicas are performed using the Metropolis criterion of acceptance (Sugita and Okamoto, 1999; Hansmann, Dic 1997).

Although PTMetaD protocol succeeds in overcoming hidden energy barriers and comprehensively explores the CV space of a system, its applicability is dramatically limited by high computational costs given by the dramatic increase in replicas required to guarantee an efficient exchange between the energy distributions of neighboring temperatures. However, the combination between PTMetaD with the well-tempered ensemble (WTE), a novel alternative sampling framework lately proposed (Bonomi and Parrinello, 2010), enable to amplify the potential energy fluctuations of each replica, dramatically reducing the number of trajectories required and thus the consequent computational costs of the overall simulation (Deighan et al., 2012). Thus PTMetaD-WTE represents a bridge toward different enhanced sampling protocols, definitely extending the applicability and the performance of MetaD method.

Finally, the development of PLUMED (Bonomi et al., 2009), an open source plug-in implementation working with many widely used MD suites (e.g., Amber, NAMD, GROMACS, ACEDM) has further enlarged the notoriety of MetaD frameworks.

Normal Mode Analysis (NMA)

Since its first application in structural biology in the early 80s (Brooks and Karplus, 1983; Go et al., 1983), normal mode analysis (NMA) has proved to be a useful and reliable approach to study collective and large amplitude motions of either single small proteins or large molecular machines (e.g., lysozyme (Brooks and Karplus, 1983), HIV1-protease (Zoete et al., 2002), myosin (Adamovic et al., 2008), integrins (Gaillard et al., 2007)) apart from being much less demanding than MD in term of computer resources required.

During the last decades several algorithms, based on either coarse-grained or all-atoms models, have been developed (Skjaerven et al., 2009). A novel example is represented by eNMA (enhanced NMA), a Anisotropic Network Model (Atilgan et al., 2001) based framework, recently developed by Rueda et al. (upcoming publication). Indeed, it enables to create enriched structurally diverse ensembles by performing an iterative exploratory search among the NMA models created at each sampling step.

The versatility and simplicity of NMA-based methods in calculating and storing data have also supported the development of several web servers performing NMA calculations or gathering large databases of pre-calculated protein motions. ProMode (Atilgan et al., 2001), MoViES (Cao et al., 2004), MolMovDB (Flores et al., 2006) and iGNM (Yang et al., 2005) are some of the most currently used databases, while ElNémo web server (Suhre and Sanejouand, 2004) quickly performs all-atom calculations starting from a given protein structure, and provides a comprehensive set of post-processing tools to analyze and display results.

Apart from capturing functional movements of proteins, NMA can be used in a wide variety of applications to (i) automatically predict hinge residues in protein structures (as performed in HingeProt server (Emekli et al., 2008)), (ii) refine low-resolution structures (experimental or predicted) or (iii) calculate the transition path between two conformations (as in MinActionPath (Franklin et al., 2007)).

1.3.3. Integration of molecular flexibility into protein-protein docking

As mentioned in **section 1.2.4**, one of the main challenges in *ab initio* prediction of protein-protein complex structure is properly dealing with molecular flexibility. During the last decades, several approaches have been proposed to address this issue. The easier and simpler approaches consist in an implicit treatment of flexibility by using soft potentials. Basically inspired by *induce-fit* and *conformational-selection* protein binding mechanisms, more complex and innovative strategies have been recently developed. They basically consist in implementing a final refinement step (Bonvin, 2006; Zacharias, 2010) or performing multiple docking runs from various precomputed conformations, respectively.

Soft-docking methods

Although relatively fast, one of the main limitations of the FFT-based docking methods is their incompatibility with an explicit treatment of protein flexibility. Alternative strategies to overcome this limitation consist in implementing a soft surface layer that allow overlapping of the proteins in the models (i.e., soft-core approach (Palma et al., 2000), or trimming long side chains (Heifetz and Eisenstein, 2003)).

Soft potentials are successfully applied in pyDock scoring function (Cheng et al., 2007), where the van der Waals and the electrostatic energies are truncated (to a maximum of 1 kcal/mol and between -1 kcal/mol and +1 kcal/mol, respectively) in order to avoid excessive penalization for the clashes generated during the rigid-body docking phase.

Flexible refinement

The majority of strategies to include flexibility in docking mimic the *induced-fit* association model, by involving a first exploration of the docking space using simplified/course grain and rigid-body protein representation, followed by a local refinement to a higher resolution, where a limited degree of flexibility is introduced by using specific energy optimization protocols involving only side chains or including also backbone atoms.

In ICM-DISCO (Fernandez-Recio et al., 2002) a Monte-Carlo (MC) optimization of the ligand side chains is performed after a soft-grid docking, while HADDOCK (Dominguez et al., 2003) explicitly provides backbone and side chain flexibility of both the docking partners during an MD simulated annealing refinement step. Finally, in RosettaDock (Lyskov and Gray, 2008) an initial low-resolution search is followed by a repacking and further MC optimization of the side chains, combined with small backbone deviations and rigid-body displacements. Finally, other methods involved a more exhaustive consideration of the protein plasticity by integrating small deformations of the global structures along soft harmonic modes during the initial sampling step, as implemented in ATTRACT (Zacharias, 2003, 2004) or SwarmDock (Moal and Bates, 2010) programs.

Docking of conformational ensembles

Based on the *conformational-selection* association model, a strategy to include flexibility in docking would consist in integrating precomputed conformational ensembles of the interacting proteins into a rigid-body framework, by repeating the docking process through various combinations of the docking partners. Such structural ensemble can be obtained experimentally (e.g., from NMR experiments) or be generated computationally by any sampling method (e.g., MD, NMA or homology modeling) thus spanning various degrees of flexibility, from small local rearrangement to large-scale global motions.

Although potentially promising, to date this strategy has not been really used for practical docking predictions, and very few systematic studies has been published so far on exploring the use of either conformational ensembles derived from theoretical simulation (Grunberg et al., 2004; Smith et al., 2005; Chaudhury and Gray, 2008) or experimental data (i.e., NMR spectroscopy) (Chaudhury and Gray, 2008). Indeed, both the studies from Grümberg et al. and Smith et al. agreed that the ensemble docking failed to improve structure prediction of protein complexes, although leading to an increase in the number of native solutions generated. Intriguingly, no clear correlation was found between success rate and RMSD from the bound structure (Grunberg et al., 2004). A more successful approach was reported by Chaudhury et al., where flexibility was restrained to the smaller protein in the complex. Indeed, a real improvement of the docking results was observed using MD structures, while the performance dramatically dropped with NMR structures (Chaudhury and Gray, 2008). However, the ensembles used in all these studies do not really represent the population of unbound state, as only few conformers were used in the docking procedure.

A more recent example of ensemble docking was reported, consisting in the successful integration of large RDC-based ensembles of free ubiquitin into a rigid-body docking protocol (Pons et al., 2013). Experimental limitations precluded the application of RDC-based ensembles at a large-scale, and therefore more research work was needed in finding practical ways for generating successful ensembles of unbound proteins in solution and their optimal use in docking protocols, which is one of the goals of this thesis.

2. Objectives

*“The impossible of today
will become the possible of tomorrow.”*

Konstantin Tsiolkovsky

Proteins function through their interaction with other proteins and biomolecules, forming specific complexes that are determined by the 3D structure and energetics of the interacting subunits. Computational methods can successfully contribute to predict and characterize these mechanistic aspects of protein function at atomic level, in which conformational flexibility plays a major role. However, an accurate consideration of protein plasticity within computational modeling of protein function at molecular level is still far from trivial, mostly because of both technical and methodological limitations. In this context the main purpose of this PhD thesis has been the assessment, development and application of computational tools for the structural, energetic and dynamic characterization of protein molecules and their interactions. This general purpose englobes several specific objectives:

1. Analysis of advances and new challenges of methods for the energetic characterization of protein-protein interfaces;
2. Assessment of current *in silico* techniques for the structural prediction of protein interactions;
3. Systematic study on the role of conformational heterogeneity in protein-protein association process;
4. Development and benchmarking of a novel protocol to integrate unbound conformational ensembles in protein-protein docking;
5. Application of computational methods for the prediction and characterization of protein interactions in cases of biological interest;
6. Application of computational methods to elucidate the dynamic basis of protein dysfunction for biomedical applications.

3. Articles

“If you are out to describe the truth,
leave elegance to the tailor.”

Albert Einstein

3.1. Advances and new challenges in modeling of protein interactions

Given the growing interest in protein-protein interactions and the technical advances in computational field, an increasing number of *in silico* tools have been developed with the aim of (i) identifying residues that significantly contribute to binding, and (ii) modeling protein complexes starting from the isolated component structures (*docking problem*). Testing and comparing these computational methodologies is fundamental in order to assess their performance, identify their limitations, and finally guide new developments in the field. In this context, CAPRI experiment provides a common ground for testing the predictive capability of currently available docking methods.

Firstly, this section will be focused on the analysis of several existing computational protocols for the characterization of protein-protein interfaces. Secondly, the performance of our pyDock protocol (Cheng et al., 2007) on the last CAPRI round (Lensink and Wodak, 2013) will be evaluated and discussed.

Manuscripts presented in this section:

1. Romero-Durana M, Pallara C, Glaser F, Fernández-Recio J. **Modeling Binding Affinity of Pathological Mutations for Computational Protein Design.** *Methods Mol Biol.* (accepted)

2. Pallara C, Jiménez-García B, Pérez-Cano L, Romero-Durana M, Solernou A, Grosdidier S, Pons C, Moal IH, Fernandez-Recio J. (2013) **Expanding the frontiers of protein-protein modeling: From docking and scoring to binding affinity predictions and other challenges.** *Proteins* 81(12), 2192-200

3.1.1. Modeling binding affinity of pathological mutations for computational protein design.

Miguel Romero-Durana,[1#] Chiara Pallara,[1#] Fabian Glaser,[2] Juan Fernández-Recio[1*]

[1]Joint BSC-CRG-IRB Research Program in Computational Biology, Barcelona Supercomputing Center, Barcelona, Spain.

[2]Bioinformatics Knowledge Unit, The Lorry I. Lokey Interdisciplinary Center for Life Sciences and Engineering Technion, Israel

# Joined first authors

* Corresponding author

Modeling Binding Affinity of Pathological Mutations for Computational Protein Design

Miguel Romero-Durana,[1#] Chiara Pallara,[1#] Fabian Glaser,[2] Juan Fernández-Recio[1*]

[1]Joint BSC-CRG-IRB Research Program in Computational Biology, Barcelona Supercomputing Center, Barcelona, Spain.

[2]Bioinformatics Knowledge Unit, The Lorry I. Lokey Interdisciplinary Center for Life Sciences and Engineering Technion, Israel

# Joined first authors

* Corresponding author

Abstract

An important aspect of protein functionality is the formation of specific complexes with other proteins, which are involved in the majority of biological processes. The functional characterization of such interactions at molecular level is necessary, not only to understand biological and pathological phenomena, but also to design improved, or even new interfaces, or to develop new therapeutic approaches. X-ray crystallography and NMR spectroscopy have increased the number of 3D protein complex structures deposited in the Protein Data Bank (PDB). However, one of the more challenging objectives in biological research is to functionally characterize protein interactions and thus, identify residues that significantly contribute to the binding. Considering that the experimental characterization of protein interfaces remains expensive, time-consuming and labor-intensive, computational approaches represent a significant breakthrough in proteomics, assisting or even replacing experimental efforts. Thanks to the technological advances in computing and data processing, these techniques now cover a vast range of protocols, from the estimation of the evolutionary conservation of amino acid positions in a protein, to the energetic contribution of each residue to the binding affinity. In this chapter, we will review several existing computational protocols to model the phylogenetic, structural and energetic properties of residues within protein-protein interfaces.

Key Words

Protein-protein interactions, hot-spots identification, interface prediction, evolutionary conservation, protein protein docking, biomolecular dynamics simulation, *in silico* alanine scanning, pyDock, AMBER package, ConSurf.

Introduction

One of the current goals of proteomics is to predict and characterize protein-protein complex interfaces. Access to such information is highly valuable as it helps to elucidate large protein interaction networks, increases the current knowledge on biochemical pathways, improves comprehensive description of disease pathogenesis and finally suggests putative new therapeutic targets [1-3]. Moreover, the use of computational approaches offers faster and more cost-efficient procedures in comparison to experimental methods such as co-immunoprecipitation, affinity chromatography, yeast two-hybrid or mass spectroscopy.

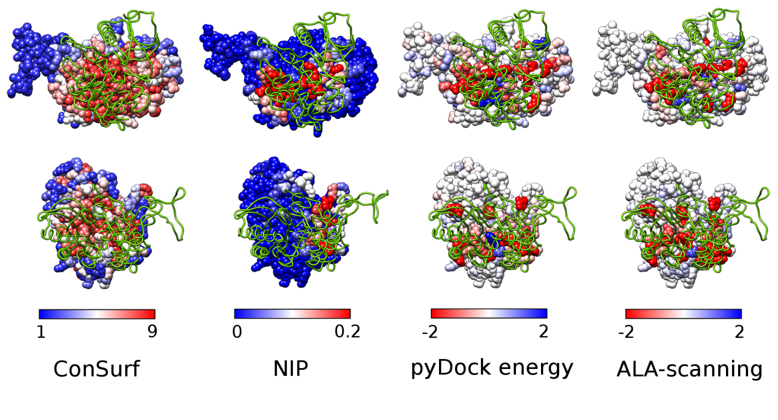


Fig. 1 MEK1-BRAF interface characterization. MEK1 and BRAF interface characterization using different computational techniques (first and second line respectively): ConSurf evolutionary conservation, pyDockNIP calculation, pyDock binding energy decomposition, binding free energy change (ΔΔG) estimated by *in silico* alanine scanning.

In this chapter, we will review several computational methods that exploit phylogenetic, structural and energetic properties of interface residues for the computational design of protein complexes, or the characterization of pathological mutations involved in protein-protein interfaces. First, we will describe two methods that do not need the structure of the protein-protein complex, namely ConSurf [4-7] and Normalized Interface Propensity (NIP) [8]. ConSurf identifies functionally and structurally important residues (e.g., involved in enzymatic activity, in ligand binding or protein-protein interactions [9]) on a protein by estimating the degree of conservation of each amino acid site among their close sequence homologues. NIP computes the tendency of a given residue to be located at the interface, based on rigid-body docking poses evaluated by pyDock scoring function [10] (based on accessible surface area-based desolvation, coulombic electrostatics and van der Waals energy). Then, we will describe two other protocols which require previous knowledge of the complex structure: residue contribution to binding energy computed with pyDock, and *in silico* Alanine (Ala) scanning, based on Molecular Dynamics simulations with AMBER14 package [11] and binding energy calculations using the MM-GBSA method [12]. The use of these methods will be illustrated on one example, the MEK1-BRAF complex (PDB ID 4MNE) [13], in which several pathological mutations are annotated [14].

Materials

ConSurf Server

1. ConSurf Server is a bioinformatics tool that estimates the evolutionary conservation of amino acid positions in protein molecules based on the phylogenetic relations among close homologous sequences. It can be found at http://consurf.tau.ac.il.

PyDock

1. PyDock is docking package freely available to academic users. Go to pyDock download web page http://life.bsc.es/pid/pydock/get_pydock.html [15] and fill in the form with the requested information. pyDock team will quickly send you a copy of the application and instructions to install it.

FTDock

1. From the FTDock [16] web page http://www.sbg.bio.ic.ac.uk/docking/download.html, download file *gnu_licensed_3D_Dock.tar.gz* to the folder of your choice.
2. From the FFTW web page http://www.fftw.org/download.html, download file f*ftw-2.1.5.tar.gz*.
3. Move to the folder where you have downloaded the file *fftw-2.1.5.tar.gz* and unpack the package with the following

commands:
> *cd folder-where-fftw-2.1.5.tar.gz-has-been-downloaded*
> *gunzip fftw-2.1.5.tar.gz*
> *tar xvf fftw-2.1.5.tar*

4. Move into directory *fftw-2.1.5* and compile the library:
   > *cd fftw-2.1.5;*
   > *./configure;*
   > *make*
5. Move to the folder where you have downloaded *gnu_licensed_3D_Dock.tar.gz* and unpack FTDock package.
6. Move to the unpacked folder *3D_Dock/progs.* Edit file *Makefile* and set the correct complete path to the *fftw-2.1.5* directory. This is done by setting the variable FFTW_DIR on line 15. You should also check the value of the CC_FLAGS variable, and make it fit to your system (e.g: for a x86_64 Linux system, CC_FLAGS variable has been modified and set to '-O -m64'.
7. Type the following command:
   > *make*
8. You should now have the executable files *ftdock, build* and *randomspin* available. Optional: Edit your *.bashrc* file to include *3D_Dock/progs* folder in your system path (PATH variable).

UCSF CHIMERA molecular viewer

UCSF Chimera [17] is a highly extensible program for interactive visualization, molecular structure analysis and high-quality images generation. Here are the instructions to install UCSF Chimera Molecular viewer:

1. Go to UCSF Chimera Molecular viewer web page at http://www.cgl.ucsf.edu/chimera.
2. Go to the download session, clicking on Download in the menu on the top-left of the web page and selecting the UCSF Chimera Molecular viewer installer appropriate for you platform.
3. Install UCSF Chimera Molecular viewer on your computer following the platform specific installation instructions available on the same page.

AMBER package

AMBER is a package of programs for Molecular Dynamics simulations of proteins and nucleic acids. It is distributed in two parts: AmberTools14 and Amber14. Here are the instructions to install AMBER package:

1. Go to the AMBER web page at http://ambermd.org/#Amber14.
2. After filling the registration form located on its own section at http://ambermd.org/AmberTools14-get.html, download AmberTools14 clicking on the *Download* button.
3. Download the Amber 14 License Agreement, print this form, fill it in, sign and return it to the address given at the bottom of the license agreement. Once the order is processed, download the AMBER program package following the download information you will receive via email.
4. Install AMBER on your machine and compile the source code format using Fortran 95, C or C++ compilers following the instructions in the Amber Reference Manual at http://ambermd.org/doc12/Amber14.pdf.

Methods

Analysis of residue conservation by ConSurf

1. Go to ConSurf web server page at http://consurf.tau.ac.il. Then, ConSurf web server will ask you several questions regarding the computation you want to run.
2. To the question *Analyze Nucleotides or Amino Acids*? select *Amino-Acids* option.
3. To the question *Is there a known protein structure*? select *Yes* option.
4. Provide the PDB ID (e.g., 4MNE) of the structure you want to analyze or upload your own PDB file, browsing to its location. Press Next button.
5. Select the chain identifier of the molecule to be analyzed.
6. Indicate whether there is a multiple sequence alignment (MSA) to upload. If there is not, ConSurf server will generate it. You may set the parameters ConSurf server will use to generate the MSA. For this work, ConSurf server has been run with default parameters.

7. At the bottom of the page, fill the *Job title* field to identify the job.
8. Fill the *User E-Mail* field, check the *Send a link to the results by e-mail* check-box and click the submit button. Thus, ConSurf server will send you an e-mail with a link to the results when it has finished.
9. Open the e-mail sent by ConSurf and go to the results page link.
10. Click on the *Download all Consurf outputs in a click!* link, save the ConSurf results file and unzip it.
11. Open *consurf.grades* file. From all the columns of the file, focus on three: *3LATOM*, *SCORE* and *COLOR*. The *3LATOM* column contains an id code of the analyzed residues. The *SCORE* column contains the computed normalized conservation score. Lower scores (more negative) correspond to more conserved residues, while higher scores (more positive) correspond to less conserved residues. A similar information is shown in column *COLOR* where, in order to ease visualization of the results, the continuous conservation scores have been partitioned into nine different bins, with bin 9 representing the most conserved positions and bin 1 the most variable positions. It is important to remark that neither the *SCORE* values nor the *COLOR* values indicate absolute magnitudes of conservation, but rather the relative degree of conservation of a given residue in the specific protein under study (i.e., neither *SCORE* nor *COLOR* values of residues of different proteins are generally comparable).
12. ConSurf provides two PDB files where the *SCORE* and *COLOR* values are assigned to the bfactor field. This is quite useful in order to get a picture of which residues are more conserved. With your favorite molecular visualization application open **.pdb_With_Conservation_Scores.pdb* and **.pdb_ATOMS_section_With_Consurf* files for displaying SCORE and COLOR values respectively (**Fig. 1**).

Prediction of binding hot-spots by NIP

NIP computation can be divided in four different steps: 1) initial setup, where the receptor and ligand PDB files of the complex are preprocessed in order to generate the input files that FTDock and pyDock require, 2) sampling phase, where FTDock generates a set of docking poses, 3) scoring phase, where pyDock dockser module scores and ranks the poses generated by FTDock and 4) NIP computation, where the first 100 ranked docking poses (those with lower binding energy) are selected from the whole set of generated docking poses, and pyDock patch module is used to compute the NIP values.

Next, we describe each one of these phases in more detail.

1. Initial setup

   a) Create a project folder and move to it.

   b) From the PDB web site, download the receptor and ligand structures: e.g. download the PDB files of receptor (3EQI) and ligand (4MNE) into the project_folder (see **Note 1**).

   c) Create pyDock ini file: open your favorite text editor and create the file 4mne.ini as shown in **Fig. 2**.

   d) Run pyDock setup module:

      > *pydock3 4mne setup*

   e) pyDock setup module should have generated several new files (see **Table 1**).

```
[receptor]
pdb     =       3eqi.pdb
mol     =       A
newmol  =       A

[ligand]
pdb     =       4mne.pdb
mol     =       B
newmol  =       B
```

Fig. 2 Example of pyDock input file. The input file is typically divided into two sections, *[receptor]* and *[ligand]*, designed to specify the variables related to the receptor and ligand, respectively. The pdb line defines the PDB file name. The *mol* line specifies the original chain name in each PDB file, whereas the *newmol* indicates the new one in the pyDock output files. Please, be aware that the *newmol* chain names must be different for the receptor and the ligand.

Table 1. pyDock modules input and output files.

Module name	Input files	Output files
setup	docking_name.ini	docking_name_rec.pdb docking_name_lig.pdb docking_name_rec.pdb.H docking_name_lig.pdb.H docking_name_rec.pdb.amber

		docking_name_lig.pdb.amber
rotftdock	docking_name_rec.pdb docking_name_lig.pdb	docking_name.rot
rotzdock	docking_name_rec.pdb docking_name_lig.pdb	docking_name.rot
dockser	docking_name_rec.pdb docking_name_lig.pdb docking_name_rec.pdb.H docking_name_lig.pdb.H docking_name_rec.pdb.amber docking_name_lig.pdb.amber docking_name.rot	docking_name.ene
patch	docking_name_rec.pdb docking_name_lig.pdb docking_name.rot docking_name.ene	docking_name.recNIP docking_name.rec.pdb.nip docking_name.ligNIP docking_name.lig.pdb.nip
BindEy	docking_name.ini	docking_name_rec.pdb docking_name_lig.pdb docking_name_rec.pdb.H docking_name_lig.pdb.H docking_name_rec.pdb.amber docking_name_lig.pdb.amber docking_name.rot docking_name.ene
resEnergy	docking_name_rec.pdb docking_name_lig.pdb docking_name_rec.pdb.H docking_name_lig.pdb.H docking_name_rec.pdb.amber docking_name_lig.pdb.amber docking_name.rot	docking_name.receptor.residueEne docking_name.ligand.residueEne docking_name.receptor.atomEne docking_name.ligand.atomEne

2. FTDock sampling

   a) Run FTDock:

   ```
   > ftdock -static 4mne_rec.pdb -mobile 4mne_lig.pdb -calculate_grid 0.7 -angle_step 12 -internal -15 -surface 1.3 -keep 3 -out 4mne.ftdock
   ```

   b) When FTDock is finished, you should have a new file named *4mne.ftdock* in the folder.

3. Scoring

   In this phase, the docking poses generated in the sampling phase are scored and ranked with pyDock dockser module.

   a) Run pyDock rotftdock module:

   ```
   > pydock3 4mne rotftdock
   ```

   b) There should now be a new file 4mne.rot. Each line in this file represents a rotation and translation matrix. FTDock *4mne.rot* file should have 10000 different lines.

c) Score and rank FTDock poses by running pyDock dockser module:

```
> pydock3 4mne dockser
```

d) Once dockser module has finished, it should have created file *4mne.ene* with 10002 different lines (see **Note 2** for a detailed description of this file).

4. NIP computation

a) Run pyDock patch module:

```
> pydock3 4mne patch
```

b) *4mne.recNIP* and *4mne.ligNIP* files should have been created. These files show the computed NIP value for each residue of receptor and ligand respectively. Those residues with NIP values greater than 0.2 are predicted to be *hot-spots*.

c) For visualization proposes, patch module output includes two PDB files, with extension **.pdb.nip*, where the NIP values have been assigned to the bfactor field. With your favorite molecular visualization application open **_rec.pdb.nip* or **_lig.pdb.nip* files for displaying the NIP values of receptor and ligand respectively (**Fig. 1**).

Computation of binding energy per residue with pyDock

1. Create a folder for computing residue binding energy.
2. From the PDB web site, download the structure of a protein-protein complex, e.g. BRAF/MEK1 (PDB ID 4MNE).
3. Create pyDock *ini* file. Open your favorite text editor and create the *4mne.ini* file specifying receptor and ligand subunits.
4. Compute pyDock binding energy by running the following command:

```
> pydock3 4mne bindEy
```

5. pyDock should have generated several new files. Please, see **Table 1** to confirm.
6. Run pyDock residue energy module:

```
> pydock3 4mne resEnergy
```

7. The module should have created for ligand and receptor **.atomEne* and **.residueEne* files with the contribution to the binding energy of each individual atom and residue respectively.
8. You may get a graphical representation of the residue binding energy (**Fig. 1**), by assigning the binding energy values given in **.residueEne* files to the bfactor field of the corresponding PDB file of the target molecules.

In-silico alanine scanning with AMBER

The Alanine scanning workflow can be divided into three different steps: 1) the preparation of the PDB files for both the wild type and the mutated structures, 2) the Molecular Dynamics simulation of the wild type complex and 3) the binding free energy calculation on both the wild type and the mutated structures.

1. Wild type and mutated structures PDB files preparation

   a) Start a new session of UCSF Chimera Molecular viewer and open 4MNE PDB file clicking on *File → Fetch by ID* entering *4mne* as PDB ID in the new window and then clicking on the Fetch button. Delete all chains but A and B, and all existing water molecules from the system.

   b) Build missing segments starting the Chimera interface to MODELLER. Click on *Tools → Structure Editing → Model/Refine Loops*. In the new window, select *all missing structure* as *model/remodel* option and *one* as both *number of residues adjacent to missing region allowed to move* and *number of models to generate*. Write the MODELLER license key and start the rebuilding by clicking on *OK*. The MODELLER license key is freely available only for academic use and can be requested at the MODELLER web page https://salilab.org/modeller/registration.html, filling up the license agreement and clicking on *agreed and accepted* buttom.

   c) Save the PDB files of the complex and each subunit in the wild type form. Go to *File → SavePDB*. In the new window enter MEK1-BRAF.pdb as file name of the refined complex structure and finally click on *Save*. Select each subunit of the complex by its chain name from *Select → Chain*. Go to *File → SavePDB*, specify the subunit new file name (i.e., *MEK1.pdb* for chain A and

BRAF.pdb for chain B), pick the save selected atom only option and finally click on *Save*.

d) Save the complex and the subunit PDB files for each mutant. Start a new session of UCSF Chimera Molecular viewer, open *MEK1-BRAF.pdb* file, select only one residue to be mutated then go to *Tools* → *Structure Editing* → *Rotamers*, choose ALA as rotamer type and click on *OK*. Save the resulting mutated complex structure going to *File* → *Save PDB* and specifying the mutation in the new file name (e.g., *MEK1-BRAF_F468A.pdb*). Finally, select the mutated subunit structure only and save it in a separate file (e.g., *BRAF_F468A.pdb*). Repeat the same protocol for each BRAF and MEK1 residue to be mutated.

e) Edit all *MEK1-BRAF.pdb* and *MEK1.pdb* files (both wild type and mutated). Rename MG residue to MG2 and convert ACP molecule to ATP.

```
source leaprc.ff99SB
source leaprc.gaff

#Load ATP parameters
loadamberprep ATP.prep
loadamberparams ATP.frcmod

#Check ATP parameters
check ATP

#Load pdb file
4mne=loadpdb MEK1-BRAF.pdb

#Check pdb structure
check 4mne

#Compute total charge
charge 4mne

#Put an 12A-buffer of TIP3P water around the system
solvateoct 4mne TIP3PBOX 12.0

#Neutralize the system
addions 4mne Na+ 4

#Save topology and coordinate files
saveamberparm 4mne MEK1-BRAF_solv.prmtop MEK1-BRAF_solv.inpcrd

quit
```

Fig. 3 Example of AMBER LEaP input file to build topology and coordinates files of wild type solvated system. The *source* command tells LEaP AMBER tool to execute the start-up script for ff99SB and GAFF force fields. First, ATP parameters are loaded and checked, then *MEK1-BRAF.pdb* file is loaded into a new unit called *4mne*, the structure is checked (i.e., close contacts and bond distances, bond and angle parameters) and the total charge is computed. Then, the system is solvated by adding a truncated octahedral 12 Å-box of TIP3P water molecules around the protein, and neutralized by adding 4 Na+ ions. Finally, the topology and coordinate files are saved in the *prmtop* and *inpcrd* AMBER format respectively.

2. Molecular Dynamics simulation

   a) Download the ATP molecule parameters from the AMBER parameter database (see **Note 3**). Go to the AMBER parameter database web page at http://www.pharmacy.manchester.ac.uk/bryce/amber/. Search the row ATP (revised phosphate parameters) in the Cofactors table and save the PREP and FRCMOD files as *ATP.prep* and *ATP.frcmod*, respectively.

   b) Modify the ATP atom names in your PDB file to match the atom names in the *ATP.prep* file so that LEaP AMBER tool will be able to match them up.

   c) Create the input files for the MD simulation (topology and coordinate files) using LEaP AMBER tool. Run the input script *tleap-solv.in* (**Fig. 3**, see **Note 4**) using the following command:

> $AMBERHOME/bin/tleap -f tleap-solv.in > tleap-solv.out

Flag *-f* tells tleap to execute the start-up script after-specified.

```
#Solvent minimization

&cntrl
imin=1,
maxcyc=1000,
ncyc=500,
ntb=1,
cut=12,
ntr=1,
restraintmask='!:WAT,Na+,Cl-',
restraint_wt=50,
drms=0.01
/
```

Fig. 4 Example of AMBER pmemd input file for solvent minimization. In the input file, *imin=1* specifies that minimization instead of Molecular Dynamics will be performed, the parameter *maxcyc* specifies the total number of minimization cycles to be run while *ncyc* specify the number of steepest descent minimization followed by *maxcyc-ncyc* steps of conjugate gradient minimization, drms sets the convergence criterion for the energy gradient (in Å). The parameter *ntb=1* means that a period boundary will be set around the system to maintain a constant volume while cut sets the cutoff value (in Å) applied for non-bonded interactions. The flag *ntr=1* indicates that the positional restraint method is applied for the energy minimization, *restraintmask* specifies the atoms to be restrained (in this cases all but water and ions molecule) and finally *restraint_wt* defines the restraints strength in terms of force constant in kcal mol^{-1} Å^{-2} applied on each restrained atom.

d) Run a short solvent minimization step using AMBER pmemd input script *min_solv.in* (**Fig. 4**) and the following input command:

> $AMBERHOME/bin/pmemd -i min_solv.in -o min_solv.out -c MEK1-BRAF_solv.inpcrd -p MEK1-BRAF_solv.prmtop -r MEK1-BRAF_min.rst -ref MEK1-BRAF_solv.inpcrd

Flag *-i* specifies the input file, *-o* the output file, *-c* the coordinate file, *-p* the parameter and topology file, *-r* the output restart file with coordinates and velocities, and *-ref* the reference coordinates file for positional restraints, if this option is specified in the input file.

```
#Equilibration (I)

&cntrl
imin=0,
irest=0,
ntx=1,
ntb=1,
cut=12,
ntc=2,
ntf=2,
tempi=0.0,
temp0=300.0,
ntt=3,
gamma_ln=1.0,
nstlim=20000,
dt=0.002,
ntwx=5000,
ntwr=5000,
ntpr=5000,
ntr=1,
restraintmask='!:WAT,Na+,Cl-',
restraint_wt=25,
ig=-1,
/
```

Fig. 5 Example of AMBER pmemd input file for first step equilibration. In the input file, *imin=0* specifies that Molecular Dynamics instead of minimization will be performed, the parameters *irest=0* and *ntx =1* indicate that only coordinates but no velocity information will be taken from the previous restart file, the flag *ntc=2* indicates that all bonds involving hydrogen bonds are constrained by the SHAKE algorithm to eliminate high frequency oscillations in the system while *ntf=2* means that all types of forces in the force filed are being calculated except bond interaction involving H-atoms. The parameters *temp0* and *tempi* define the initial and the temperature at which the system is to be kept respectively, *ntt=3* indicates that the temperature Langevin thermostat will be used while *gamma_ln=1.0* sets the collision frequency to 1fs. The flag *nstlim* defines the number of simulation steps, *dt* defines the length of each frame (set at 2 fs, here) while *ntwx*, *ntwr*, *ntpr* define the frequency of data deposition (coordinates, energy and restart respectively). Finally *ig=-1* sets the random seed based on the current date and time and hence will be different for every run. The meaning of the rest of the parameters listed in the input file was previously explained.

e) Run a 5-step equilibration by which the system temperature is raised from 0 to 300K, and a gradual relaxation is performed by progressively releasing positional restraints, initially set. The following protocol should be used:

- As a first equilibration step, run a 40-ps simulation in isovolume condition applying harmonic restraints to all the protein atoms and heating the system to 300K. Run *equil1.in* input script (**Fig. 5**) using the following command:

> $AMBERHOME/bin/pmemd -i equil1.in -o equil1.out -c MEK1-BRAF_min.rst -p MEK1-BRAF_solv.prmtop -r MEK1-BRAF_eq1.rst -ref MEK1-BRAF_min.rst -x MEK1-BRAF_eq1.mdcrd

- Perform an additional 20-ps step in isothermal-isovolume condition reducing the harmonic restraints to all the protein atoms from 25 to 10 kcal/(mol·Å^2). Run *equil2.in* input script (**Fig. 6**) using the following command:

  > $AMBERHOME/bin/pmemd -i equil2.in -o equil2.out -c MEK1-BRAF_eq1.rst -p MEK1-BRAF_solv.prmtop -r MEK1-BRAF_eq2.rst -ref MEK1-BRAF_eq1.rst -x MEK1-BRAF_eq2.mdcrd

```
#Equilibration (II)

&cntrl
imin=0,
irest=1,
ntx=5,
ntb=1,
cut=12,
ntc=2,
ntf=2,
tempi=300.0,
temp0=300.0,
ntt=3,
gamma_ln=1.0,
nstlim=10000,
dt=0.002,
ntwx=5000,
ntwr=5000
ntpr=5000,
ntr=1,
restraintmask='!:WAT,Na+,Cl-',
restraint_wt=10,
ig=-1,
/
```

Fig. 6 Example of AMBER pmemd input file for second step equilibration. In the input file, the flags *ntx=5* and *irest=1* mean that velocity and coordinate information will be taken from the previous restart file. The meaning of the rest of the parameters listed in the input file was previously explained.

- Run another 20-ps step applying the harmonic restraints only to the backbone atoms. Run *equil3.in* input script (**Fig. 7**) using the following command:

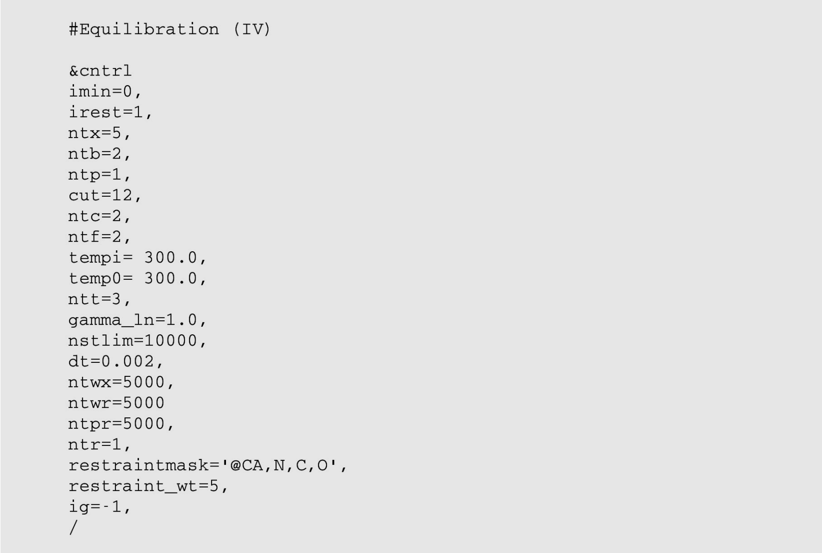
  > $AMBERHOME/bin/pmemd -i equil3.in -o equil3.out -c MEK1-BRAF_eq2.rst -p MEK1-BRAF_solv.prmtop -r MEK1-BRAF_eq3.rst -ref MEK1-BRAF_eq2.rst -x MEK1-BRAF_eq3.mdcrd

```
#Equilibration (III)

&cntrl
imin=0,
irest=1,
ntx=5,
ntb=2,
ntp=1,
cut=12,
ntc=2,
ntf=2,
tempi=300.0,
temp0=300.0,
ntt=3,
gamma_ln=1.0,
nstlim=10000,
dt=0.002,
ntwx=5000,
ntwr=5000
ntpr=5000,
ntr=1,
restraintmask='@CA,N,C,O',
restraint_wt=10,
ig=-1,
/
```

Fig. 7 Example of AMBER pmemd input file for third step equilibration. In the input file the flags *ntb=2* and *ntp=1* indicate that constant pressure instead of constant volume is applied. The meaning of the rest of the parameters listed in the input file was previously explained.

- Run further 20-ps step decreasing protein backbone restraints to 5 kcal/(mol·Å^2). Run *equil4.in* input script (**Fig. 8**) using the following command:

  > $AMBERHOME/bin/pmemd -i equil4.in -o equil4.out -c MEK1-BRAF_eq3.rst -p MEK1-BRAF_solv.prmtop -r MEK1-BRAF_eq4.rst -ref MEK1-BRAF_eq3.rst -x MEK1-BRAF_eq4.mdcrd



```
#Equilibration (IV)

&cntrl
imin=0,
irest=1,
ntx=5,
ntb=2,
ntp=1,
cut=12,
ntc=2,
ntf=2,
tempi= 300.0,
temp0= 300.0,
ntt=3,
gamma_ln=1.0,
nstlim=10000,
dt=0.002,
ntwx=5000,
ntwr=5000
ntpr=5000,
ntr=1,
restraintmask='@CA,N,C,O',
restraint_wt=5,
ig=-1,
/
```

Fig. 8 Example of AMBER pmemd input file for fourth step equilibration. The meaning of all the parameters listed in the input file was previously explained.

- Run the last step of the equilibration consisting on 100-ps unrestrained MD simulation in isothermal-isobaric condition. Run *equil5.in* input script (**Fig. 9**, see **Note 5**) using the following command:

  > $AMBERHOME/bin/pmemd -i equil5.in -o equil5.out -c MEK1-BRAF_eq4.rst -p MEK1-BRAF_solv.prmtop -r MEK1-BRAF_eq5.rst -ref MEK1-BRAF_eq4.rst -x MEK1-BRAF_eq5.mdcrd

```
#equilibration (V)

&cntrl
imin=0,
irest=1,
ntx=5,
ntb=2,
ntp=1,
cut=12,
ntc=2,
ntf=2,
tempi=300.0,
temp0=300.0,
ntt=3,
gamma_ln=1.0,
nstlim=50000,
dt=0.002,
ntwx=5000,
ntwr=5000
ntpr=5000,
ntr=0,
ig=-1,
/
```

Fig. 9 Example of AMBER pmemd input file for fifth step equilibration. In the input file, the flag *ntr=0* indicates that the positional restraint method is turned off. The meaning of the rest of the parameters listed in the input file was previously explained.

f) Finally, perform 5-ns MD unrestrained simulation keeping the same system condition as the last equilibration step. Run *prod.in* input script (**Fig. 10**, see **Note 6**) using the following command:

> $AMBERHOME/bin/pmemd -i prod.in -o prod.out -c MEK1-BRAF_eq5.rst -p MEK1-BRAF_solv.prmtop -r MEK1-BRAF_prod.rst -ref MEK1-BRAF_eq5.rst -x MEK1-BRAF_prod.mdcrd

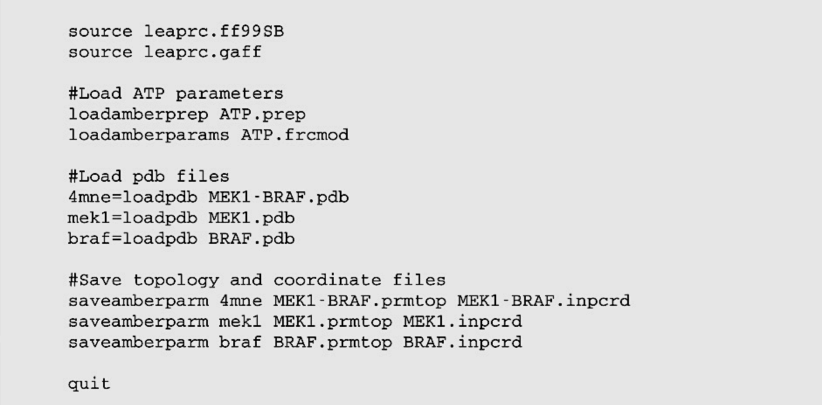
```
#5ns-MD simulation

&cntrl
imin=0,
irest=1,
ntx=5,
ntb=2,
ntp=1,
cut=12,
ntc=2,
ntf=2,
tempi=300.0,
temp0=300.0,
ntt=3,
gamma_ln=1.0,
nstlim=2500000,
dt=0.002,
ntwx=5000,
ntwr=5000
ntpr=5000,
ntr=0,
ig=-1,
/
```

Fig. 10 Example of AMBER pmemd input file for unrestrained MD. The meaning of all the parameters listed in the input file was previously explained.

3. Binding free energy calculation

   a) Build the topology and coordinate files of the unsolvated wild type (WT) structure for both the complex and its single subunits using *tleap-WT.in* input file (**Fig. 11**). Run LEaP AMBER tool using the following command:

      > $AMBERHOME/bin/tleap -f tleap-WT.in > tleap-WT.out

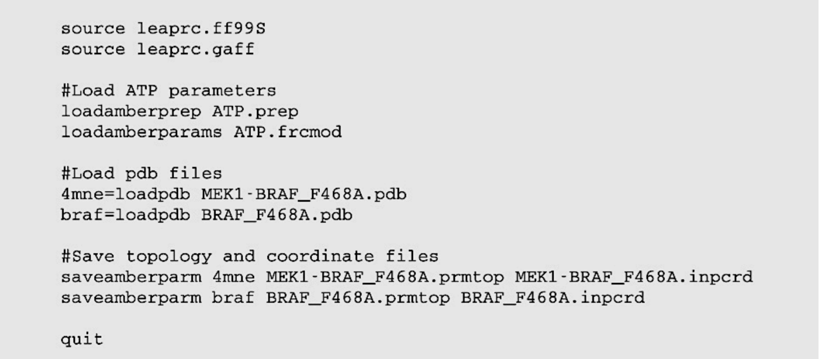


```
source leaprc.ff99SB
source leaprc.gaff

#Load ATP parameters
loadamberprep ATP.prep
loadamberparams ATP.frcmod

#Load pdb files
4mne=loadpdb MEK1-BRAF.pdb
mek1=loadpdb MEK1.pdb
braf=loadpdb BRAF.pdb

#Save topology and coordinate files
saveamberparm 4mne MEK1-BRAF.prmtop MEK1-BRAF.inpcrd
saveamberparm mek1 MEK1.prmtop MEK1.inpcrd
saveamberparm braf BRAF.prmtop BRAF.inpcrd

quit
```

Fig. 11 Example of AMBER LEaP input file to build topology and coordinates files of wild type dry systems.

b) For each mutation studied, build the topology and coordinate files of the mutated structure for both the complex and mutated subunit using *tleap-mut.in* input file (**Fig. 12**). Run LEaP AMBER tool using the following command:

> $AMBERHOME/bin/tleap -f tleap-mut.in > tleap-mut.out



```
source leaprc.ff99S
source leaprc.gaff

#Load ATP parameters
loadamberprep ATP.prep
loadamberparams ATP.frcmod

#Load pdb files
4mne=loadpdb MEK1-BRAF_F468A.pdb
braf=loadpdb BRAF_F468A.pdb

#Save topology and coordinate files
saveamberparm 4mne MEK1-BRAF_F468A.prmtop MEK1-BRAF_F468A.inpcrd
saveamberparm braf BRAF_F468A.prmtop BRAF_F468A.inpcrd

quit
```

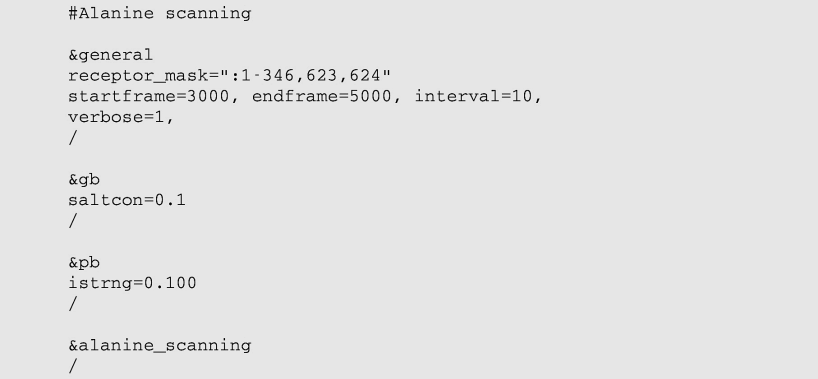
Fig. 12 Example of AMBER LEaP input file to build topology and coordinates files of mutated dry systems. Here, F468 BRAF residue is taken as example.

c) Perform alanine scanning calculation on 200 snapshots extracted from the last 2 ns of each MD trajectory. Run *mmpbsa.in* input file for *MMPBSA.py* script in AMBER14 (**Fig. 13**) using the following command:

> $AMBERHOME/bin/MMPBSA.py -i mmpbsa.in -sp MEK1-BRAF_solv.prmtop -cp MEK1-BRAF.prmtop -rp

MEK1-BRAF.prmtop -lp MEK1-BRAF.prmtop -y MEK1-BRAF_prod.mdcrd -mc MEK1-BRAF_F468A.prmtop -ml BRAF_F468A.prmtop

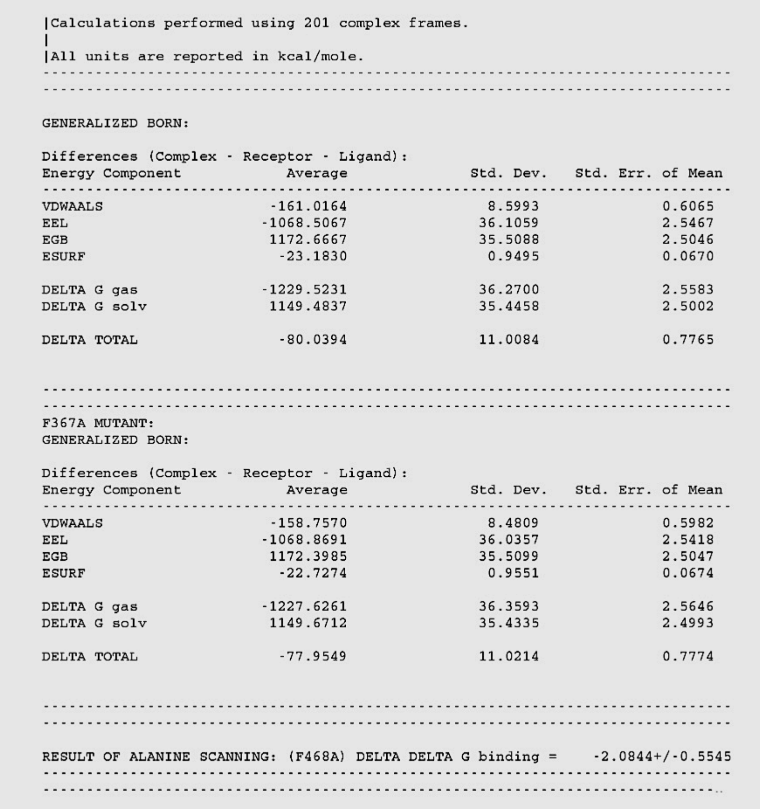
Flag *-i* specifies the input file, *-sp* the solvated WT complex topology file, *-cp* the unsolvated WT complex topology file, *-rp* the unsolvated WT receptor topology file, *-lp* the unsolvated WT ligand topology file, *-y* the complex trajectory file to analyze, *-mc* the unsolvent mutated complex topology file and *-ml* the unsolvated mutated subunit topology file. Please, be aware that as MEK1 is the first molecule in the complex, for alanine scanning calculations the unsolvated mutated subunit topology file will be specified with the flag *-mr*.



```
#Alanine scanning

&general
receptor_mask=":1-346,623,624"
startframe=3000, endframe=5000, interval=10,
verbose=1,
/

&gb
saltcon=0.1
/

&pb
istrng=0.100
/

&alanine_scanning
/
```

Fig. 13 Example of MMPBSA.py input file to perform alanine scanning calculation. The input file is typically divided into four sections (*&general*, *&gb*, *&pb*, *&alanine_scanning*). The *&general* section is designed to specify generic variables related to the overall calculation. For instance, the flag *startframe* and *endframe* specifies the frame from which to begin and to stop extracting snapshots respectively, the parameter interval indicates the offset from which to choose frames from the trajectory file, *verbose=1* means that complex, ligand and receptor energy terms will be printed in the output file. The *&gb* and *&pb* section markers tells the script to perform MM-GBSA and MM-PBSA calculations with the given values defined within those sections (i.e., the variables saltcon and istrng that specify the salt concentration and the ionic strength, respectively). Finally the *&alanine_scanning* section marker initializes alanine scanning in the script. Please be aware that given the higher computational costs of MM-PBSA calculation, only MM-GBSA calculation is performed in this work.

d) Extract the ΔΔG of binding related to the specific mutations estimated as the difference between the binding ΔG of the WT and that of the mutated complex. All these data are easily available in the final output file, *FINAL_RESULTS_MMPBSA.dat*, including all the wild type and mutated system average binding energies (reported as van der Waals, electrostatic and non polar energy contributions), as shown in **Fig. 14**.



```
|Calculations performed using 201 complex frames.
|
|All units are reported in kcal/mole.
-------------------------------------------------------------------------------
-------------------------------------------------------------------------------

GENERALIZED BORN:

Differences (Complex - Receptor - Ligand):
Energy Component            Average              Std. Dev.   Std. Err. of Mean
-------------------------------------------------------------------------------
VDWAALS                     -161.0164              8.5993              0.6065
EEL                        -1068.5067             36.1059              2.5467
EGB                         1172.6667             35.5088              2.5046
ESURF                        -23.1830              0.9495              0.0670

DELTA G gas                -1229.5231             36.2700              2.5583
DELTA G solv                1149.4837             35.4458              2.5002

DELTA TOTAL                  -80.0394             11.0084              0.7765


-------------------------------------------------------------------------------
-------------------------------------------------------------------------------
F367A MUTANT:
GENERALIZED BORN:

Differences (Complex - Receptor - Ligand):
Energy Component            Average              Std. Dev.   Std. Err. of Mean
-------------------------------------------------------------------------------
VDWAALS                     -158.7570              8.4809              0.5982
EEL                        -1068.8691             36.0357              2.5418
EGB                         1172.3985             35.5099              2.5047
ESURF                        -22.7274              0.9551              0.0674

DELTA G gas                -1227.6261             36.3593              2.5646
DELTA G solv                1149.6712             35.4335              2.4993

DELTA TOTAL                  -77.9549             11.0214              0.7774


-------------------------------------------------------------------------------
-------------------------------------------------------------------------------

RESULT OF ALANINE SCANNING: (F468A) DELTA DELTA G binding =     -2.0844+/-0.5545
-------------------------------------------------------------------------------
-------------------------------------------------------------------------------
```

Fig. 14 Extract from the MMPBSA.py FINAL_RESULTS_MMPBSA.dat output file. The file includes all the average energies, standard deviations, and standard error of the mean for GB followed by PB calculations (if calculated). After each section, the ΔG of binding is given along with the error values. After each method, the ΔΔG of binding is reported, corresponding to the relative effect the mutation has on the ΔG of binding for the complex. The specific mutation is also printed at the end of the file. Here, F468 residue alanine scanning is taken as example.

e) You may get a graphical representation of the ΔΔG of binding (**Fig. 1**), by assigning the values given in *FINAL_RESULTS_MMPBSA.dat* file to the bfactor field of the corresponding PDB file of the complex structure.

Notes

1) As there is no unbound structure for the ligand yet, the ligand structure contained on the complex PDB file (4MNE) is used here instead for illustration purposes. However, in a standard NIP computation, unbound structures should be used.

2) The principal columns of the *4mne.ene* file are:

    - *Conf*: Conformation number of the docking pose as in the last column of the rot file.
    - *Ele*: Electrostatic energy of the pose.
    - *Desolv*: Desolvation energy of the pose.
    - *VDW*: Van der Waals energy of the pose.
    - *Total*: Total docking energy of the pose, computed as ele + Desolv + 0.1 * VDW (note a 0.1 weight for VDW).
    - *RANK*: Pose rank according to its computed total binding energy.

3) Files from the PDB may contain bound ligands, cofactors or non-standard residues whose parameters are not available in the AMBER parameters database. In this case you should make use of the Antechamber tools, which ship with AmberTools, to create *PREP* and *FRCMOD* files. For more information, see the ANTECHAMBER tutorial (http://ambermd.org/tutorials/basic/tutorial4b/) and the AMBER manual.

4) LEaP AMBER tool renumbers PDB residues starting from 1. Thus, the original numeration of your PDB file may not be always kept.

5) Since your system may not start from an equilibrium state, additional time steps may be required during the minimization and equilibration steps of the MD simulation. One can check for equilibrium by verifying whether properties, such as potential energy, temperature or pressure, no longer change in any systematic fashion and are just fluctuating around a mean value.

6) To guarantee reliable results in the in silico Alanine scanning calculation, RMSD simulation should be highly equilibrated. Ideally one should probably run a much longer production run than 5ns (e.g., 100 ns).

References

1. Arkin M.R.,Wells J.A. (2004) Small-molecule inhibitors of protein-protein interactions: progressing towards the dream. *Nat Rev Drug Discov* 3, 301-317.

2. DeLano W.L. (2002) Unraveling hot spots in binding interfaces: progress and challenges. *Curr Opin Struct Biol* 12, 14-20.

3. Toogood P.L. (2002) Inhibition of protein-protein association by small molecules: approaches and progress. *J Med Chem* 45, 1543-1558.

4. Glaser F., Pupko T., Paz I. et al.(2003) ConSurf: identification of functional regions in proteins by surface-mapping of phylogenetic information. *Bioinformatics* 19, 163-164.

5. Ashkenazy H., Erez E., Martz E. et al. (2010) ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. *Nucleic Acids Res* 38, W529-533.

6. Landau M., Mayrose I., Rosenberg Y. et al. (2005) ConSurf 2005: the projection of evolutionary conservation scores of residues on protein structures. *Nucleic Acids Res* 33, W299-302.

7. Celniker G., Nimrod G., Ashkenazy H. et al. (2013) ConSurf: Using Evolutionary Data to Raise Testable Hypotheses about Protein Function. . *Isr. J. Chem.* 53, 199–206.

8. Grosdidier S., Fernandez-Recio J. (2008) Identification of hot-spot residues in protein-protein interactions by computational docking. *BMC Bioinformatics* 9, 447.

9. Branden C.I., Tooze J. (1999) Introduction to protein structure. 2 ed. Garland Pub., New York.

10. Cheng T.M., Blundell T.L., Fernandez-Recio J. (2007) pyDock: electrostatics and desolvation for effective scoring of rigid-body protein-protein docking. *Proteins* 68, 503-515.

11. Case D.A., Cheatham T.E., Darden T. et al. (2005) The Amber biomolecular simulation programs. *J Comput Chem* 26, 1668-1688.

12. Miller B.R.I., McGee D.T.J., Swails J.M., et al (2012) MMPBSA.py: An Efficient Program for End-State Free Energy Calculations. *J. Chem. Theory Comput.* 8, 3314–3321.

13. Haling J.R., Sudhamsu J., Yen I., et al. (2014) Structure of the BRAF-MEK complex reveals a kinase activity independent role for BRAF in MAPK signaling. *Cancer Cell* 26, 402-413.

14. Kiel C., Serrano L. (2014) Structure-energy-based predictions and network modelling of RASopathy and cancer missense mutations. *Mol Syst Biol* 10, 727.

15. Jimenez-Garcia B., Pons C., Fernandez-Recio J. (2013) pyDockWEB: a web server for rigid-body protein-protein docking using electrostatics and desolvation scoring. *Bioinformatics* 29, 1698-1699.

16. Gabb H.A., Jackson R.M., Sternberg M.J. (1997) Modelling protein docking using shape complementarity, electrostatics and biochemical information. *J Mol Biol* 272, 106-120.

17. Pettersen E.F., Goddard T.D., Huang C.C. et al. (2004) UCSF Chimera--a visualization system for exploratory research and analysis. *J Comput Chem* 25, 1605-1612.

3.1.2. Expanding the frontiers of protein-protein modeling: from docking and scoring to binding affinity predictions and other challenges.

Chiara Pallara,[1#] Brian Jiménez-García,[1#] Laura Perez-Cano,[1] Miguel Romero,[1] Albert Solernou,[1] Solène Grosdidier,[1] Carles Pons,[1,2] Iain H. Moal,[1] Juan Fernandez-Recio[1*]

[1]*Joint BSC-IRB Research Programme in Computational Biology, Barcelona Supercomputing Center, Barcelona, Spain*

[2]*Computational Bioinformatics, National Institute of Bioinformatics (INB), Barcelona, Spain*

# Equal contribution

* Corresponding author

PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Expanding the frontiers of protein–protein modeling: From docking and scoring to binding affinity predictions and other challenges

Chiara Pallara,[1] Brian Jiménez-García,[1] Laura Pérez-Cano,[1] Miguel Romero-Durana,[1] Albert Solernou,[1] Solène Grosdidier,[1] Carles Pons,[1,2] Iain H. Moal,[1] and Juan Fernandez-Recio[1]*

[1] Joint BSC-IRB Research Programme in Computational Biology, Barcelona Supercomputing Center, Barcelona, Spain
[2] Computational Bioinformatics, National Institute of Bioinformatics (INB), Barcelona, Spain

ABSTRACT

In addition to protein–protein docking, this CAPRI edition included new challenges, like protein–water and protein–sugar interactions, or the prediction of binding affinities and $\Delta\Delta G$ changes upon mutation. Regarding the standard protein–protein docking cases, our approach, mostly based on the *pyDock* scheme, submitted correct models as predictors and as scorers for 67% and 57% of the evaluated targets, respectively. In this edition, available information on known interface residues hardly made any difference for our predictions. In one of the targets, the inclusion of available experimental small-angle X-ray scattering (SAXS) data using our *pyDockSAXS* approach slightly improved the predictions. In addition to the standard protein–protein docking assessment, new challenges were proposed. One of the new problems was predicting the position of the interface water molecules, for which we submitted models with 20% and 43% of the water-mediated native contacts predicted as predictors and scorers, respectively. Another new problem was the prediction of protein–carbohydrate binding, where our submitted model was very close to being acceptable. A set of targets were related to the prediction of binding affinities, in which our pyDock scheme was able to discriminate between natural and designed complexes with area under the curve = 83%. It was also proposed to estimate the effect of point mutations on binding affinity. Our approach, based on machine learning methods, showed high rates of correctly classified mutations for all cases. The overall results were highly rewarding, and show that the field is ready to move forward and face new interesting challenges in interactomics.

Proteins 2013; 81:2192–2200.

Key words: complex structure; CAPRI; protein–protein docking; pyDock; protein–carbohydrate interactions.

INTRODUCTION

One of the major challenges in structural biology is to provide structural data for all complexes formed between proteins and other macromolecules. Current structural coverage of protein–protein interactions (i.e., available experimental structures plus potential models based on homologous complex structures) is below 4% of the estimated number of possible complexes formed between human proteins.[1,2] The pace of experimental determination of complex structures is still behind the determination of individual protein structures. In addition, many of these interactions will never be determined by X-ray crystallography because of their transient nature. For these reasons, computational docking methods aim to become a complementary approach to solve the structural interactome. The field of protein docking has experienced an explosion in recent years, partially propelled by the CAPRI experiment. Past editions showed an increasing amount of participant groups and computational approaches, and a

Grant sponsor: Spanish Ministry of Science; Grant number: BIO2010-22324; Grant sponsor: FPU (to LP-C) and FPI (to BJ-G) fellowships from the Spanish Ministry of Science; Grant sponsor: People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme (FP7/2007-2013) under REA (to IHM); Grant number: PIEF-GA-2012-327899.

Pallara and Jiménez-García contributed equally to this work.

*Correspondence to: Juan Fernandez-Recio, Barcelona Supercomputing Center (BSC), Jordi Girona 29, 08034 Barcelona, Spain. E-mail: juanf@bsc.es

Received 14 June 2013; Revised 31 July 2013; Accepted 1 August 2013
Published online 12 August 2013 in Wiley Online Library (wileyonlinelibrary.com).
DOI: 10.1002/prot.24387

large variety of targets. We have participated in all targets of this fifth CAPRI edition. In addition to the standard prediction of protein–protein targets, this edition has entered into related areas including binding affinity predictions and free energy changes upon mutation, as well as prediction of sugar binding and interface water molecules. Our overall experience has been highly rewarding and we describe here the details of our participation and the key factors of our success.

MATERIALS AND METHODS

Generation of rigid-body docking poses for the predicting experiment

In all targets, we used FTDock[3] with electrostatics and 0.7 Å grid resolution and ZDOCK 2.1[4] to generate 10,000 and 2000 rigid-body docking poses, respectively, as previously described.[5] For the final four targets of this edition (T53, T54, T57, and T58) we generated an additional pool of flexible docking poses using SwarmDock. For these runs, the standard protocol was employed,[6–8] with the Dcomplex score used as the objective function,[9] but without the final clustering and rescoring phase. In T46 we generated an additional pool of 10,000 solutions using FTDock without electrostatics and at lower resolution (1.2 Å), as part of an old protocol used with previous targets, but these conditions were not applied for the rest of the targets since we saw in the past that this step was not increasing the chances of correct predictions. In T46 and T47, we used RotBUS[10] to generate 59,112 and 41,021 additional docking poses, respectively, but this method was not used for the rest of the targets since we previously checked that this procedure did not improve the results. In Target T50, given the large size of 1918 H1N1 influenza virus hemagglutinin protein, we generated a total of 92,432 FTDock docking poses, increasing the number of translations selected from each rotation from 3 (default) to 10. Cofactors, water molecules and solvent ions were not included in our docking calculations.

Scoring of rigid-body docking poses for both the predicting and the scoring experiments

We scored the docking models generated by the above described methods with our pyDock protocol,[11] based on energy terms previously optimized for rigid-body docking. The binding energy is basically composed of accessible surface area-based desolvation, Coulombic electrostatics and van der Waals energy (with a weighting factor of 0.1 to reduce the noise of the scoring function). Electrostatics and van der Waals were limited to ±1.0 and 1.0 kcal/mol for each interatomic energy value, respectively, to avoid excessive penalization from possible clashes in the structures generated by the rigid-body approach. The same protocol was used in the scoring experiment to score all the docking models that were proposed. We did not include van der Waals in the T46 scoring experiment, although this did not affect the results. Cofactors, water molecules and solvent ions were not considered for scoring.

Removal of redundant docking poses

After scoring, we eliminated redundant predictions to increase the variability of the predictions and maximize the success chances using a simple clustering algorithm with a distance cutoff of 4.0 Å, as previously described.[12] In target T47, since the resulting solutions looked correct [according to the available structure of a highly homologous complex with protein data bank (PDB) code 2WPT], we reduced this cutoff to 0.5 Å.

Minimization of final models

The final 10 selected docking poses were minimized to improve the quality of the docking models and reduce the number of interatomic clashes. In the majority of the targets we used TINKER[13] as previously described.[12,14] In targets T53 and T54 we used CHARMM (50 steps conjugate gradient, 500 steps adopted-basis Newton–Raphson and 50 steps steepest decent, with the CHARMM19 force field).[15] In target T58 we used AMBER10 with AMBER parm99 force field.[16,17] The minimization protocol consisted of a 500-cycle steepest descent minimization with harmonic restraints applied at a force constant of 25 kcal/(mol·Å^2) to all the backbone atoms to optimize the side chains, followed by another 500-cycle conjugate gradient minimization without restraints. This minimization step was performed after ranking, solely to remove clashes.

Modeling of subunits with no available structure

For several targets, the structures of the subunits were not available and needed to be modeled. In most of the targets, we used Modeler 9v6 with default parameters[18] based on the template/s suggested by the organizers or on other homologue proteins found by BLAST[19] search. The final selected model was that with the lowest DOPE score.[20] For targets T53 and T54 we used POPULUS (http://bmm.cancerresearchuk.org/~populus/) with default template selection and model building settings.[21]

RESULTS AND DISCUSSION

In this CAPRI edition we submitted predictions for all the proposed targets. Our results for the standard protein–protein docking assessment are summarized in Table I. In addition, there were new challenges like the

Table I
Results of Our pyDock Protocol for All Protein–Protein Targets of the Last CAPRI Edition

		Predictors			Scorers		
Target	Type	Submission rank[a]	Quality[b]	Successful groups[c]	Submission rank[a]	Quality[b]	Successful groups[c]
T46	HH	—	—	2 (40)	—	—	8 (16)
T47	HU	1	***	25 (29)	2[d]	***	13 (14)
T48	UU	3	*	14 (32)	No scorers	No scorers	No scorers
T49	UU	4	*	14 (33)	6	*	7 (13)
T50	UH	1	**	18 (40)	4	**	12 (17)
T51	DHD	—	—	3 (46)	—	—	5 (13)
T53	UH	3[e]	**	20 (42)	1	**	11 (13)
T54	UH	—	—	4 (41)	—	—	0 (13)
T58	UU	5	**	11 (23)	No scorers	No scorers	No scorers

U, unbound; H, homology-based model; D, domain.
[a]Rank of the best model within our submission to CAPRI.
[b]Quality of our best model according to CAPRI criteria.
[c]Number of successful groups for each target; in brackets, total number of participants.
[d]Model Rank 1 had medium accuracy (**).
[e]Model Rank 1 had acceptable accuracy (*).

prediction of protein–water and protein–sugar interactions, as well as the estimation of binding affinities and energy changes upon mutation. Hereinafter, we describe in detail our submissions for each of the targets.

Standard protein-protein docking assessment: successful predictions

Target T47 (model/pseudounbound)

Target T47 was the structural prediction of the complex between the DNase domain of colicin E2 and the immunity protein Im2. The real challenge in this target was the prediction of interface water molecules, however, the protein–protein docking predictions were already assessed, and therefore we have included them in this section. The colicin E2 was modeled based on the structure of colicin E9 (85% sequence identity) in complex with Im9 immunity protein (PDB 1EMV).[22] The coordinates of the immunity protein Im2 were extracted from its structure in complex with colicin E9 (PDB 2WPT). Given the existence of this homologous colicin E9/Im2 complex structure (PDB 2WPT),[23] the binding mode for target T47 was easy to determine by template-based docking. However, we performed the template-free docking calculations to assess the automatic docking protocol. We only applied distance restraints after pyDock protocol by selecting those docking poses in which two key contacting residues, Im2 Y54 and colicin E2 F85 (equivalent to colicin E9 F86 in 2WPT),[23] were within an arbitrary distance of 6 Å (same distance used in standard restraints with pyDockRST module.[24] We submitted five correct models (one high accuracy, one medium accuracy, and three acceptable). Our first submitted model (Rank 1 according to pyDock energy, and generated by ZDOCK), was a high-quality model (Table I), with 75% native contacts, 2.48 Å ligand root mean square deviation (RMSD), and 0.75 Å interface RMSD with respect to the crystal structure (Fig. 1; PDB 3U43).[25] This docking model had the lowest ligand RMSD with respect to the homologous colicin E9/Im2 complex (PDB 2WPT) amongst all solutions (although we did not use this homologous structure for docking), and even more interestingly, we would have obtained exactly the same result without applying the above-mentioned distance restraint filter.

For the scoring experiment, we evaluated the provided 1051 models with our pyDock scoring function, and applied the same distance filter that we used as predictors (see above). All our submitted predictions resulted to be successful, consisting of six medium and four high-quality models. We had a high-accuracy model ranked second after pyDock scoring and distance filter (uploaded by Weng), with 77% native contacts, 0.9 Å ligand RMSD, and 0.4 Å interface RMSD with respect to the crystal structure (PDB 3U43[25]; Table I; Fig. 1). Interestingly, our Rank 5 model was the best model submitted among all 14 participants, with 79% native contacts, 0.7 Å ligand RMSD, and 0.5 Å interface RMSD. Two better models uploaded by Weng were not found by any of the participants. Remarkably, as in predictors, our results would not have changed had we not applied the distance restraints filter.

Target T48 (unbound/unbound)

Target T48 was the structural prediction of the complex between the diiron-hydroxylase toluene 4-monooxygenase and the Rieske-type ferredoxin T4moC protein (PDB 1VM9).[26] As suggested by the organizers, the heterohexameric biological unit of the diiron-hydroxylase was built by applying crystal symmetry operations to its trimeric structure in complex with the

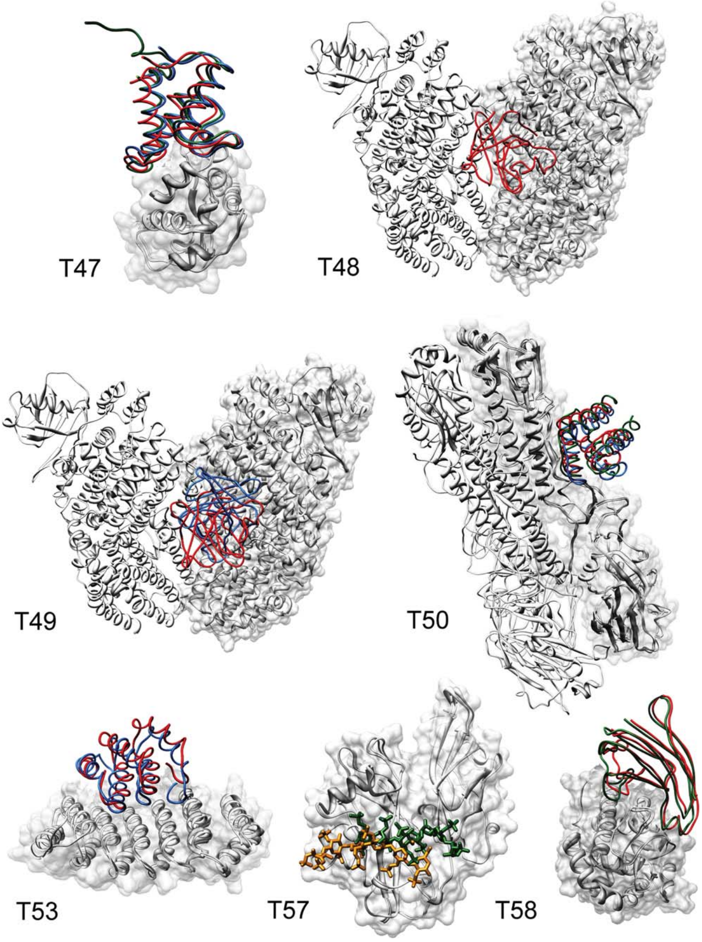


Figure 1

Representation of our best models for targets T47, T48, T49, T50, T53, T57, and T58. For each target, receptors are superimposed and shown in white. Ligand in our best model as predictors is shown in red, and as scorers in blue. For comparison, the structure of the experimental complex (if available) is represented in green.

T4moD effector protein (PDB 3DHH).[27] We used the hexameric construct for the generation of docking poses, which were scored by pyDock. Then, we selected those docking poses that had any of the diiron-hydroxylase Fe^{2+} and ferredoxin S_2Fe_2 atoms within 23 Å distance to allow for the electron transfer between these groups[27] (the distance cutoff we used was arbitrary, based on the expected distance of 16 Å in 3DHH plus a margin to allow the inclusion of some low-energy solutions). For the submission, we removed chains D, E, and F from the hexamer as we misinterpreted some of the organizers' instructions, but this did not affect the quality of the submitted models. The analysis of the results showed that we submitted three models of acceptable quality. Our prediction ranked third after pyDock scoring and electron transfer distance filtering (generated by FTDock) had 14% native contacts, 8.4 Å ligand RMSD, and 3.6 Å interface RMSD with respect to the complex crystal structure (not yet available). We found another acceptable model (ranked 10th in our submission set) that had 49% native contacts, 6.3 Å ligand RMSD, and 2.2 Å interface RMSD with respect to the complex crystal structure.

Target 49 (unbound/unbound)

Target T49 was the same complex as T48 but with a different hexameric conformation for diiron-hydroxylase toluene 4-monooxygenase (unbound coordinates not released). We applied the same protocol as for target T48 (pyDock scoring and electron transfer distance filtering). We submitted four acceptable quality models. The model ranked fourth of our submission set had acceptable quality, with 26% native contacts, 12.4 Å ligand RMSD, and 3.5 Å interface RMSD with respect to the complex crystal structure (not yet available). We also submitted another model with 11% native contacts, 6.9 Å ligand RMSD, and 2.7 Å interface RMSD.

For the scoring experiment, the 1085 solutions were scored by the same protocol, based on pyDock scoring and electron transfer distance filtering. In some models, the monooxygenase was uploaded as a trimer, therefore we reconstructed the biological hexamer (based on symmetry) to calculate the electron transfer distance filter. Since it was not clear whether in these cases the hexamer was going to be rebuilt for the assessment, our submission set was formed by the top five solutions obtained after rebuilding the hexamer, and by the top five solutions obtained by just considering the structure submitted by uploaders (i.e., without rebuilding the hexamer in cases of uploaded trimer). Our ranked sixth submission was an acceptable model (uploaded by Nakamura), with 11% native contacts, 7.9 Å ligand RMSD, and 2.9Å interface RMSD with respect to the complex crystal structure (not yet available).

Target 50 (unbound/model)

Target T50 was the structural prediction of the complex between the 1918 H1N1 influenza virus hemagglutinin and the HB36.3 *de novo* designed protein. The coordinates of the hemagglutinin were taken from its structure in complex with an antibody (PDB 3GBN)[28] and the biological hexamer was rebuilt by applying symmetry operations. We modeled the HB36.3 based on the crystal structure of the homologous (83% sequence identity) protein APC36109 from *Bacillus stearothermophilus* (PDB 1U84), using the target-template protein alignment offered by the organizers. Given the size of the system, we increased the number of rigid-body docking solutions generated by FTDock (see Materials and Methods section). Our submission as predictors contained nine successful models (five acceptable and four medium-quality solutions). Our Rank 1 submission (generated by FTDock) was a medium-quality model with 47% native contacts, 6.1 Å ligand RMSD, and 1.8 Å interface RMSD with respect to the complex crystal structure (Fig. 1; PDB 3R2X).[29] Interestingly, our Rank 4 submission, with 41% native contacts, 3.4 Å ligand RMSD, and 1.6 Å interface RMSD, was the best model submitted among all participants as predictors.

For the scoring experiment, we evaluated the 1451 models in the same conditions as in predictors. We found five acceptable and one medium-quality solutions. Our Rank 4 submission was a medium-quality model, with 44.9% native contacts, 4.71 Å ligand RMSD, and 1.93 Å interface RMSD with respect to the complex crystal structure (PDB 3R2X[29]; Fig. 1).

Target T53 (unbound/model)

Target T53 was a complex between two artificial alpha helicoidal repeat proteins (alpha-Rep), alpha-rep4 (PDB 3LTJ)[30] and alpha-rep2, both designed on the basis of thermostable HEAT-like repeats. The ligand alpha-rep2 was built using as template alpha-rep4 (PDB 3LTJ), with 77% sequence identity. All the docking poses, generated using Zdock, Ftdock, and SwarmDock, were scored by pyDock. We submitted four successful predictions (three acceptable and one medium-quality models). Our Rank 3 submission, a medium accuracy model generated by SwarmDock, had 44% native contacts, ligand RMSD 4.4 Å, and interface RMSD 1.8 Å with respect to the crystal structure (not yet available).

For the scoring experiment, we evaluated 1400 alpha-rep4/alpha-rep2 complex models applying the same protocol as in predictors in a completely automated fashion. We found three acceptable and a medium-quality models. Our Rank 1 submission, a medium-quality model (uploaded by Yan Shen), had 62% native contacts, 3.6 Å ligand RMSD, and 1.3 Å interface RMSD with respect to the complex crystal structure (not yet available).

Target T58 (unbound/unbound)

This target was a complex between the unbound G-Type Lysozyme (PDB 3MGW)[31] and the unbound *Escherichia coli* Plig lysozyme inhibitor (PDB 4DY3).[32] There was available small-angle X-ray scattering (SAXS) data for this complex, which we used for scoring with our module pyDockSAXS, previously developed to combine pyDock scoring and fitting to SAXS data.[33] In addition, there was some available information indicating a central role of the G-type lysozyme E73, D86, and D97 residues and the *E. coli* Plig lysozyme inhibitor R119 and Y47 residues.[34] Based on these residues, we imposed ambiguous distance restraints with our module pyDockRST.[35] We submitted one medium-accuracy and two acceptable models. Our Rank 5 model, generated by SwarmDock, was a medium-quality model, and resulted to be the fourth best model submitted among all the 23 participants, with 43% native contacts, 4.9 Å ligand RMSD, and 1.8 Å interface RMSD with respect to the complex crystal structure (PDB 4G9S).[36] Interestingly, although the distance restraints proved to be essential for this target, we would have obtained only slightly worse results without using the SAXS data (Rank 10 medium accuracy model). This is probably due to the shape of the complex, classified as spherical according to the anisotropy value (1.4) computed from the ratio between the size of the largest axis and the smallest ones. Indeed, we previously showed that SAXS data does not provide much beneficial information in this type of cases.[33]

Protein-protein docking: unsuccessful cases

In three of the protein–protein cases (T46, T51, and T54) we were not able to submit any correct model, either as predictors or as scorers. These cases seemed to be highly difficult for the majority of participants, since in all of them there were no more than three successful groups as predictors or as scorers or both (Table I). In target T46 (*model/model*), the interacting subunits Mtq2 and Trm112 were modeled based on the homologue templates with low sequence identity (Mtq2 was based on template with PDB code 1T43, 28% sequence identity; Trm112 was based on template with PDB code 2J6A, 36% sequence identity). The inaccuracies in the modeling added too much error and the docking was not successful. Target T51 (*bound/model/unbound*) was a difficult case of a multidomain protein, with interactions between GH5-CBM6/CBM13/Fn3 domains. This could be divided in two different docking cases both involving CBM13 domain, which needed to be modeled based on template with PDB code 1KNL (38% sequence identity). Again, a model based on a template with that level of homology can deteriorate docking results. Target 54 (*unbound/model*) was in principle easy, involving the modeling of Rep16 based on the template with PDB code 3LTJ (88% sequence identity), but the submitted solutions were incorrect for us as well as for the majority of participants. Indeed, despite the scoring set contained several acceptable models, no group was able to identify them (Table I).

Prediction of protein-water interactions

Target T47 was the prediction of a protein–protein complex structure, as described in above sections, but the real challenge was to predict the location of water molecules. After generating the protein–protein docking poses as above described, we predicted the water positions in each docking model using DOWSER[37] with default parameters (with a probe radius of 0.2Å and the default atoms dictionary). Our Rank 1 submitted model (generated by ZDOCK) had 20% of water native contacts, and was classified as fair (+). If we consider only the prediction of the buried water molecules, our success rates do not significantly change.

For the scoring experiment, we just applied our standard pyDock scoring function, plus distance restraints as described in above sections. The water molecules proposed in the different docking poses were not included in the scoring. Our Rank 8 submitted model (uploaded by Bates) had 43% of water native contacts and was classified as good (++). More details can be found in an upcoming publication.

Prediction of protein-carbohydrate complex structure

Target 57 (*unbound/model*) was a challenging target consisting in the prediction of the interaction between BT4661 protein and heparin. The structure of heparin in the complex was not known, so we modeled it using molecular dynamics starting with the provided conformation. We ran 10 ns using the force field AMBER parm99 of the Amber10 package[16,17] and extracted 1000 representative snapshots. Since our pyDock protocol was not intended for protein–sugar interactions, we had to devise a new *ad hoc* docking procedure. For that, we used FTDock to dock each of the 1000 heparin conformations to BT4661 protein. We selected the top 10,000 docking poses as scored by FTDock (no electrostatics). Then we applied different scoring functions to this set of docking poses: (i) PScore without minimization; (ii) PScore with minimization; and (iii) AMBER after minimization. We selected the 1000 best-scoring solutions from each method and finally we removed redundant solutions within 6.5 Å ligand RMSD. No correct submission was submitted. However, our Rank 4 submission was almost acceptable, with 65% native contacts, 11.2 Å ligand RMSD, and 4.3 Å interface RMSD with respect to the complex crystal structure (PDB 4AK2; Fig. 1). We checked *a posteriori* that there were several correct models within our docking sets, but our scoring approach failed to place them in the lowest scoring positions.

Other challenges: binding affinity and $\Delta\Delta G$ predictions

This CAPRI edition also involved the challenging problem of predicting binding affinities and energy changes upon mutations. Round 21 was the discrimination between 87 designed protein–protein interactions involving three proteins of interest (Spanish influenza HA; *Mt* ACP-2; Fc region of human IgG1) and 120 naturally occurring complexes. The pyDock function, although initially developed for the scoring of docking poses, was previously shown to have some correlation with the binding affinity data collected by Kastritis and Bonvin.[38] This was later confirmed on a subset of complexes with high-confidence affinity data, where pyDock ranked among the best performing scoring functions with a correlation of 0.63.[39] For round 21 predictions, we evaluated the correlation of each of the different pyDock individual terms with the binding affinities on the provided set of 120 naturally occurring complexes. We found that desolvation correlation with binding affinity data was not clear, showing even negative correlation with data obtained by ITC experiments. It seems that, although desolvation is essential for rigid-body docking (perhaps to compensate inaccurate calculation of electrostatics and van der Waals), it is not the most important factor for binding affinity predictions from the complex structure (in which electrostatics can be more accurately calculated). Based on these results, we devised a binding affinity descriptor (*pyDockAFF* = electrostatics $-1.0 \times$ desolvation), with confidence thresholds for the discrimination of complexes according to their binding affinities. Our predictions had area under the curve 83%, with good discrimination between designed and native interfaces. More details can be found in a recent publication.[40] It remains to be seen whether the *pyDockAFF* binding affinity predictor is suitable only for the cases in this CAPRI round, or it has more general applicability (further details in an upcoming publication).

Targets T55-56 aimed to predict the binding affinity changes upon mutations on two designed influenza hemagglutinin protein binders. We applied a multiparametric predictive model with 85 descriptors using an ensemble of models which were combined to produce consensus predictions. The models were trained upon a data set of 930 changes in affinity upon mutation which were taken from the literature. Due to the fairly low cases to descriptors ratio (10.9), we preferentially employed models with inherent overfitting avoidance bias, such as prepruning or feature selection using the Akaike information criterion, methods which construct multiple models using subsets of the descriptors and the training data, and by rejecting learners that performed poorly using leave-complex-out cross-validation.[41] To further avoid overfitting, we did not combine the selected learners together using stacking, instead opting for the unweighted mean for our consensus predictions. This approach provided an excellent ability to predict the effect of mutation, more details of which can be found in a recent publication.[42] We have since expanded this data set to form the SKEMPI database, which now includes 3047 $\Delta\Delta G$ values, as well as kinetic and thermodynamic data,[43] and have used the data to derive contact potentials that can circumvent some of the approximations associated with statistical potentials.[44]

CONCLUSIONS

We have continued our participation in CAPRI with pyDock, submitting models for all the predicting, scoring, and binding affinity prediction experiments. For the generation of docking poses, the better grid resolution used for FTDock and the use of flexible SwarmDock for the last targets were key for the success. This produced docking poses of sufficient quality to be identified by the pyDockSER scoring scheme. In selected targets, distance restraints were introduced by pyDockRST, but in most cases this did not make a difference. In one target, SAXS data was used for complementary scoring with pyDock-SAXS, which slightly improved the scoring. We obtained consistently good models for all nondifficult cases, although they were far from being trivial, since their subunits were unbound or needed to be modeled based on homology templates. In all cases but one our successful models were ranked within our first five submitted solutions, being ranked first in several cases. In this CAPRI edition we learned that our automated protocol is useful to provide correct models in easy-to-medium difficulty protein–protein docking cases, but we need further methodological development for difficult cases, especially when subunits need to be modeled based on homologues with low sequence identity. On the other side, interface water placement and sugar-binding proved to be highly challenging for our protein–protein methodology, but the results have encouraged us to develop new methods for these problems. Finally, prediction of binding affinity based on the pyDockSER scoring, and energy changes upon mutation based on multiparametric regression models showed excellent results. The overall experience has been highly rewarding and has shown once again the importance of community-wide assessment of prediction methods.

REFERENCES

1. Venkatesan K, Rual JF, Vazquez A, Stelzl U, Lemmens I, Hirozane-Kishikawa T, Hao T, Zenkner M, Xin X, Goh KI, Yildirim MA, Simonis N, Heinzmann K, Gebreab F, Sahalie JM, Cevik S, Simon C, de Smet AS, Dann E, Smolyar A, Vinayagam A, Yu H, Szeto D, Borick H, Dricot A, Klitgord N, Murray RR, Lin C, Lalowski M, Timm J, Rau K, Boone C, Braun P, Cusick ME, Roth FP, Hill DE, Tavernier J, Wanker EE, Barabasi AL, Vidal M. An empirical framework for binary interactome mapping. Nat Methods 2009;6:83-90.
2. Stumpf MP, Thorne T, de Silva E, Stewart R, An HJ, Lappe M, Wiuf C. Estimating the size of the human interactome. Proc Natl Acad Sci USA 2008;105:6959-3964.
3. Gabb HA, Jackson RM, Sternberg MJ. Modelling protein docking using shape complementarity, electrostatics and biochemical information. J Mol Biol 1997;272:106-120.
4. Chen R, Weng Z. A novel shape complementarity scoring function for protein–protein docking. Proteins 2003;51:397-408.
5. Grosdidier S, Pons C, Solernou A, Fernandez-Recio J. Prediction and scoring of docking poses with pyDock. Proteins 2007;69:852-858.
6. Moal IH, Bates PA. SwarmDock and the use of normal modes in protein–protein docking. Int J Mol Sci 2010;11:3623-3648.
7. Torchala M, Moal IH, Chaleil RA, Fernandez-Recio J, Bates PA. SwarmDock: a server for flexible protein–protein docking. Bioinformatics 2013;29:807-809.
8. Li X, Moal IH, Bates PA. Detection and refinement of encounter complexes for protein–protein docking: taking account of macromolecular crowding. Proteins 2010;78:3189-3196.

9. Liu S, Zhang C, Zhou H, Zhou Y. A physical reference state unifies the structure-derived potential of mean force for protein folding and binding. Proteins 2004;56:93-101.
10. Solernou A, Fernandez-Recio J. Protein docking by Rotation-Based Uniform Sampling (RotBUS) with fast computing of intermolecular contact distance and residue desolvation. BMC Bioinformatics 2010; 11:352.
11. Cheng TM, Blundell TL, Fernandez-Recio J. pyDock: electrostatics and desolvation for effective scoring of rigid-body protein–protein docking. Proteins 2007;68:503-515.
12. Pons C, Solernou A, Perez-Cano L, Grosdidier S, Fernandez-Recio J. Optimization of pyDock for the new CAPRI challenges: docking of homology-based models, domain–domain assembly and protein–RNA binding. Proteins 2010;78:3182-3188.
13. Ponder JW, Richards FM. An efficient Newton-like method for molecular mechanics energy minimization of large molecules. J Comput Chem 1987;8:1016-1024.
14. Pons C, Grosdidier S, Solernou A, Perez-Cano L, Fernandez-Recio J. Present and future challenges and limitations in protein–protein docking. Proteins 2010;78:95-108.
15. Brooks BR, Brooks CL, 3rd, Mackerell AD, Jr., Nilsson L, Petrella RJ, Roux B, Won Y, Archontis G, Bartels C, Boresch S, Caflisch A, Caves L, Cui Q, Dinner AR, Feig M, Fischer S, Gao J, Hodoscek M, Im W, Kuczera K, Lazaridis T, Ma J, Ovchinnikov V, Paci E, Pastor RW, Post CB, Pu JZ, Schaefer M, Tidor B, Venable RM, Woodcock HL, Wu X, Yang W, York DM, Karplus M. CHARMM: the biomolecular simulation program. J Comput Chem 2009;30:1545-1614.
16. Case DA, Cheatham TE, III, Darden T, Gohlke H, Luo R, Merz KM, Jr., Onufriev A, Simmerling C, Wang B, Woods RJ. The Amber biomolecular simulation programs. J Comput Chem 2005;26:1668-1688.
17. Wang J, Cieplak P, Kollman PA. How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules? J Comput Chem 2000; 21:1049-1074.
18. Sali A, Blundell TL. Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol 1993;234:779-815.
19. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol 1990;215:403-410.
20. Shen MY, Sali A. Statistical potential for assessment and prediction of protein structures. Protein Sci 2006;15:2507-2524.
21. Offman MN, Fitzjohn PW, Bates PA. Developing a move-set for protein model refinement. Bioinformatics 2006;22:1838-1845.
22. Kuhlmann UC, Pommer AJ, Moore GR, James R, Kleanthous C. Specificity in protein–protein interactions: the structural basis for dual recognition in endonuclease colicin-immunity protein complexes. J Mol Biol 2000;301:1163–1178.
23. Meenan NA, Sharma A, Fleishman SJ, Macdonald CJ, Morel B, Boetzel R, Moore GR, Baker D, Kleanthous C. The structural and energetic basis for high selectivity in a high-affinity protein–protein interaction. Proc Natl Acad Sci USA 2010;107:10080-10085.
24. Cheng TM, Blundell TL, Fernandez-Recio J. Structural assembly of two-domain proteins by rigid-body docking. BMC Bioinformatics 2008;9:441.
25. Wojdyla JA, Fleishman SJ, Baker D, Kleanthous C. Structure of the ultra-high-affinity colicin E2 DNase–Im2 complex. J Mol Biol 2012; 417:79-94.
26. Moe LA, Bingman CA, Wesenberg GE, Phillips GN, Jr., Fox BG. Structure of T4moC, the Rieske-type ferredoxin component of toluene 4-monooxygenase. Acta Crystallogr D Biol Crystallogr 2006;62: 476-482.
27. Bailey LJ, McCoy JG, Phillips GN, Jr, Fox BG. Structural consequences of effector protein complex formation in a diiron hydroxylase. Proc Natl Acad Sci USA 2008;105:19194-19198.
28. Ekiert DC, Bhabha G, Elsliger MA, Friesen RH, Jongeneelen M, Throsby M, Goudsmit J, Wilson IA. Antibody recognition of a highly conserved influenza virus epitope. Science 2009;324:246-251.
29. Fleishman SJ, Whitehead TA, Ekiert DC, Dreyfus C, Corn JE, Strauch EM, Wilson IA, Baker D. Computational design of proteins targeting the conserved stem region of influenza hemagglutinin. Science 2011;332:816-821.
30. Urvoas A, Guellouz A, Valerio-Lepiniec M, Graille M, Durand D, Desravines DC, van Tilbeurgh H, Desmadril M, Minard P. Design, production and molecular structure of a new family of artificial alpha-helicoidal repeat proteins (alphaRep) based on thermostable HEAT-like repeats. J Mol Biol 2010;404:307-327.
31. Kyomuhendo P, Myrnes B, Brandsdal BO, Smalas AO, Nilsen IW, Helland R. Thermodynamics and structure of a salmon cold active goose-type lysozyme. Comp Biochem Physiol B Biochem Mol Biol 2010;156:254-263.
32. Leysen S, Vanderkelen L, Van Asten K, Vanheuverzwijn S, Theuwis V, Michiels CW, Strelkov SV. Structural characterization of the PliG lysozyme inhibitor family. J Struct Biol 2012;180:235-242.
33. Pons C, D'Abramo M, Svergun DI, Orozco M, Bernado P, Fernandez-Recio J. Structural characterization of protein–protein complexes by integrating computational docking with small-angle scattering data. J Mol Biol 2010;403:217-230.
34. Helland R, Larsen RL, Finstad S, Kyomuhendo P, Larsen AN. Crystal structures of g-type lysozyme from Atlantic cod shed new light on substrate binding and the catalytic mechanism. Cell Mol Life Sci 2009;66:2585-2598.
35. Chelliah V, Blundell TL, Fernandez-Recio J. Efficient restraints for protein–protein docking by comparison of observed amino acid substitution patterns with those predicted from local environment. J Mol Biol 2006;357:1669-1682.
36. Leysen S, Vanderkelen L, Weeks SD, Michiels CW, Strelkov SV. Structural basis of bacterial defense against g-type lysozyme-based innate immunity. Cell Mol Life Sci 2013;70:1113-1122.
37. Zhang L, Hermans J. Hydrophilicity of cavities in proteins. Proteins 1996;24:433-438.
38. Kastritis PL, Bonvin AM. Are scoring functions in protein–protein docking ready to predict interactomes? Clues from a novel binding affinity benchmark. J Proteome Res 2010;9:2216-2225.
39. Moal IH, Agius R, Bates PA. Protein–protein binding affinity prediction on a diverse set of structures. Bioinformatics 2011;27:3002-3009.
40. Fleishman SJ, Whitehead TA, Strauch EM, Corn JE, Qin S, Zhou HX, Mitchell JC, Demerdash ON, Takeda-Shitaka M, Terashi G, Moal IH, Li X, Bates PA, Zacharias M, Park H, Ko JS, Lee H, Seok C, Bourquard T, Bernauer J, Poupon A, Aze J, Soner S, Ovali SK, Ozbek P, Tal NB, Haliloglu T, Hwang H, Vreven T, Pierce BG, Weng Z, Perez-Cano L, Pons C, Fernandez-Recio J, Jiang F, Yang F, Gong X, Cao L, Xu X, Liu B, Wang P, Li C, Wang C, Robert CH, Guharoy M, Liu S, Huang Y, Li L, Guo D, Chen Y, Xiao Y, London N, Itzhaki Z, Schueler-Furman O, Inbar Y, Potapov V, Cohen M, Schreiber G, Tsuchiya Y, Kanamori E, Standley DM, Nakamura H, Kinoshita K, Driggers CM, Hall RG, Morgan JL, Hsu VL, Zhan J, Yang Y, Zhou Y, Kastritis PL, Bonvin AM, Zhang W, Camacho CJ, Kilambi KP, Sircar A, Gray JJ, Ohue M, Uchikoga N, Matsuzaki Y, Ishida T, Akiyama Y, Khashan R, Bush S, Fouches D, Tropsha A, Esquivel-Rodriguez J, Kihara D, Stranges PB, Jacak R, Kuhlman B, Huang SY, Zou X, Wodak SJ, Janin J, Baker D. Community-wide assessment of protein–interface modeling suggests improvements to design methodology. J Mol Biol 2011;414: 289-302.
41. Witten IH, Frank E, Hall MA. Data mining: practical machine learning tools and techniques, 3rd ed. San Francisco, CA: Morgan Kaufmann; 2011.
42. Moretti R, Fleishman SJ, Agius R, Torchala M, Bates PA, Kastritis PL, Rodrigues JPGLM, Trellet M, Bonvin AMJJ, Cui M, Rooman M, Gillis D, Dehouck Y, Moal IH, Romero M, Perez-Cano L, Pallara C, Jimenez B, Fernandez-Recio J, Samuel Flores S, Pacella

M, Kilambi KP, Gray JJ, Grudinin S, Umeyama H, Iwadate M, Esquivel-Rodríguez J, Kihara D, Zhao N, Korkin D, Zhu X, Demerdash ON, Mitchell JC, Nakamura H, Lee H, Park H, Seok C, Standley D, Shimoyama H, Terashi G, Takeda-Shitaka M, Beglov D, Hall DR, Kozakov D, Vajda S, Pierce BG, Hwang H, Vreven T, Weng Z, Huang Y, Li H, Yang X, Ji X, Liu S, Xiao Y, Zacharias M, Qin S, Zhou H-X, Huang S-Y, Zou X, Velankar S, Janin J, Wodak SJ, Baker D. Community-wide evaluation of methods for predicting the effect of mutations on protein–protein interactions. Proteins, in press.

43. Moal IH, Fernandez-Recio J. SKEMPI: a structural kinetic and energetic database of mutant protein interactions and its use in empirical models. Bioinformatics 2012;28:2600-2607.
44. Moal IH, Fernandez-Recio J. Intermolecular contact potentials for protein–protein interactions extracted from binding free energy changes upon mutation. J Chem Theory Comput 2013;9:3715-3727.

3.2. Protein plasticity improves protein-protein docking

Despite recent methodological advances in currently used docking protocols, as shown by CAPRI (Critical Assessment of PRediction of Interactions) experiment, dealing with protein plasticity is still a crucial bottle-neck (see **section 3.1.2**). The development of efficient flexible docking algorithms is mostly hampered by our limited theoretical knowledge about the protein-protein association mechanism.

Firstly, this section will report a systematic study on the role of conformational heterogeneity in protein-protein recognition. Then, a novel protocol to integrate unbound conformational ensembles in protein-protein docking will be presented.

Manuscripts presented in this section:

1. Pallara C, Rueda M, Abagyan R, Fernández-Recio J. **Conformational heterogeneity in unbound state enhances recognition in protein-protein encounters.** *PLoS Comput Biol.* (submitted)

2. Pallara C, Fernández-Recio J. **Protein-protein ensemble docking at low-cost: improving predictive performance for medium-flexible cases.** (in preparation)

3.2.1. Conformational heterogeneity in unbound state enhances recognition in protein-protein encounters

Chiara Pallara,[1] Manuel Rueda,[2] Ruben Abagyan,[2] Juan Fernández-Recio[1*]

[1] *Joint BSC-CRG-IRB Research Program in Computational Biology, Life Sciences Department, Barcelona Supercomputing Center, Barcelona, Spain*

[2] *Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093, USA*

*Corresponding author

Conformational Heterogeneity in Unbound State Enhances Recognition in Protein-Protein Encounters

Chiara Pallara[1], Manuel Rueda[2], Ruben Abagyan[2], Juan Fernández-Recio[1]*

[1]Joint BSC-CRG-IRB Research Program in Computational Biology, Life Sciences Department, Barcelona Supercomputing Center, Barcelona, Spain

[2]Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093, USA

*Corresponding author

E-mail: juanf@bsc.es (JF-R)

Abstract

To understand cellular processes at the molecular level we need to improve our knowledge of protein-protein interactions, from a structural, mechanistic and energetic point of view. Current theoretical studies and computational docking simulations show that protein dynamics plays a key role in protein association, and support the need for including protein flexibility in modeling protein interactions. A strategy to include flexibility in docking predictions would be using conformational ensembles originated from unbound protein structures. This strategy assumes the conformational selection binding mechanism, in which the unbound state can sample bound conformers. Here we present an exhaustive computational study about the use of precomputed unbound ensembles in the context of protein docking, performed on a set of 124 cases of the Protein-Protein Docking Benchmark 3.0. Conformational ensembles were initially generated by modeling minimization with MODELLER. We identified those conformers providing optimal binding and investigated the role of protein conformational heterogeneity in protein-protein recognition. Our results show that a simple molecular mechanics minimization approach can generate conformers with better binding properties as well as improve docking encounters in medium-flexible cases. For comparison, we analyzed ensembles generated by short Molecular Dynamics trajectories with AMBER, which did not provide significantly better conformers for docking. For more flexible cases, a more extended conformational sampling based on Normal

Mode Analysis was proven helpful. We found that successful conformers provide better energetic complementarity to the docking partners but not necessarily higher similarity with respect to the bound state, which is compatible with recent views of binding association. In addition to the mechanistic considerations, these findings could be exploited for practical docking predictions of improved efficiency.

Author Summary

Proteins act as building blocks of the cells, forming complex interaction networks that are essential for almost any biological process. The comprehension of such interactions at the molecular level is necessary to improve our understanding of basic cell processes as well as for advances in biomedical and biotechnological applications. In that regard, computational methods complement experimental efforts by helping to structurally model and characterize protein interactions. However, still a major crusade is how to deal with the intrinsically flexible nature of proteins. A largely unexplored strategy to overcome this limitation is the use of precomputed conformational ensembles. Here we present a systematic study about the role of protein plasticity in protein association, based on docking simulations of unbound structural models derived from ensembles generated by different conformational sampling approaches. The results show that the description of the conformational heterogeneity of the unbound states improves their binding capabilities towards their partners, especially in cases of moderate unbound-to-bound mobility. This improvement is not necessarily related to better structural similarity to the bound state, which is consistent with an extended conformational selection mechanism.

Introduction

Proteins are key components in the cell and function through intricate networks of interactions [1] that are involved in virtually all relevant biological processes, such as gene expression and regulation, enzyme catalysis, immune response, or signal transduction [2-3]. Understanding such interactions at the molecular level is essential to target them for therapeutic or biotechnological purposes. X-ray crystallography and NMR techniques have produced a wealth of structural data on protein-protein complexes, which has largely extended our knowledge on molecular recognition and protein association mechanism and has fostered drug discovery. However, such structural data covers only a tiny fraction of the estimated number of protein-protein complexes formed in cell [4-6], and therefore, computational approaches that can complement such experimental efforts are strongly needed. One approach is template-based modeling [7], which could be potentially used to provide models at interactomics scale [4, 8-10]. However, its applicability is currently limited by the relative low number of available structures of protein complexes that can be used as accurate templates, and the difficulties in the identification of the correct templates in cases of remote homology [11-13]. On the other side, *ab initio* modeling of protein-protein complexes by computational docking shows higher applicability. The idea is to explore thousands or millions of possible orientations between two interacting proteins in order to identify the native orientation(s), based on different criteria ranging from simple geometrical considerations to a complete energy description of the interaction. In recent years, a variety of protein-protein docking methods have been reported. Geometry-based methods try to find the best surface complementarity between interacting proteins, using simplified structural models and approximate scoring functions. A popular strategy is to discretize the proteins into grids and use Fast Fourier Transform (FFT) algorithms [14] to accelerate search on the translational space, such as in FTDock [15], PIPER [16], GRAMM-X [17], ZDOCK [18], or on the rotational space, as in Hex [19] or FRODOCK [20]. Another strategy to explore surface complementarity is geometric hashing, as used in PatchDock [21]. Docking methods based on energy optimization use a variety of sampling strategies based on molecular mechanics, such as Molecular Dynamics in HADDOCK [22], or Monte-Carlo minimization in RosettaDock [23] or ICM-DISCO [24]. The function used to identify the best orientations is an important aspect of docking, and dedicated scoring schemes have been developed, based on energy terms, such as in pyDock [25], or on statistical potentials as in SIPPER [26] or PIE [27]. The Critical Assessment of PRediction of Interactions (CAPRI; http://www.ebi.ac.uk/msd-srv/capri/) experiment has indeed shown that accurate models can be produced by docking in most of the cases [28]. However, there are other cases in which all docking methods systematically fail, typically the most flexible ones [28].

Thus, one of the major challenges in docking is how to deal with molecular flexibility and conformational changes that happen upon association [29-30]. A major hurdle is the computational cost of integrating docking and conformational search, aggravated by our limited knowledge of the protein-protein association mechanism. Different mechanisms for flexible protein-protein binding have been proposed. Perhaps the most widespread view is the induced-fit mechanism, in which the interacting partners are involved in initial encounters that evolve towards the final specific complex by adjusting their interfaces. Most of the reported methods for flexible docking try to mimic this mechanism, typically using an initial rigid-body search followed by a final refinement of the interfaces as in ICM-DISCO [31], HADDOCK [22], RosettaDock [23] or FiberDock [32], or by integrating small deformations of the global structures during the sampling based on normal modes as in ATTRACT [33-34] or SwarmDock [35].

An alternative mechanism is conformational selection, which was initially proposed for systems in which the ligand selectively bound one of the conformers of the dynamically fluctuating receptor protein [36-37]. This was generalized to the "conformational selection and population shift" concept, which postulated that flexible proteins in solution naturally sample a variety of conformational states, and the ligand protein preferentially binds to a pre-existing subpopulation of such conformers, thus adjusting the equilibrium in favor of them [38-40]. Recently, the conformational selection model has been extended to include different mutual conformational selection and adjustment steps [41], so that the unbound conformational states that are available for mutual selection and adjustment might not be initially in the bound conformation. The conformational selection model has been largely supported by several structural studies including MD, NMA, X-ray crystallography and NMR experiments [41-45] and later strongly confirmed by theoretical analysis based on the correlation between complex association and dissociation rates and several molecular descriptors detailing specific features of both protein intrinsic flexibility and complex formation [42, 46]. This mechanism can be implemented in a computational docking strategy by using precomputed ensembles of unbound proteins, which hopefully contain conformers that are suitable for binding the partner. However, to date this strategy has not been really used for practical docking predictions. Most of prior studies were limited to the use of a few selected conformers and applied to specific cases of interest [47-49]. Unexpectedly, the few systematic analyses published so far [50-52] failed to improve structure prediction of protein complexes with respect to the unbound structure. This could be related to unrealistic representation of the motions occurring in the time scale of molecular association [50-52]. Indeed, for small proteins like ubiquitin it is possible to obtain more representative ensembles, based on RDC data, which are definitely useful in docking predictions [53]. However, this approach is difficult to generalize for large scale predictions due to experimental limitations. Therefore, it would be important to find practical ways of generating ensembles that include conformers that improve binding. This could help not only to improve docking predictions but also to advance towards a better understanding of flexible protein-protein association mechanism. With this purpose in mind, here we used three different computational approaches to represent the conformational heterogeneity of unbound proteins, and tested them on a standard protein-protein docking benchmark. Our analysis clearly shows that a simple molecular mechanics minimization approach provides sufficient conformational heterogeneity to improve docking predictions in medium-flexible cases, which are the most likely to follow the conformational selection mechanism.

Results

Unbound conformational ensembles by energy optimization and Molecular Dynamics contain conformers with better binding capabilities than the unbound structure

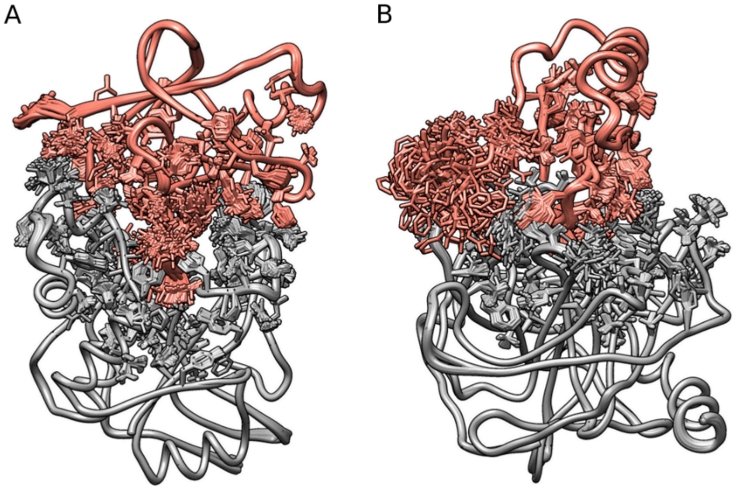


Fig 1. Representative conformational ensembles generated by MODELLER minimization. 100 conformers independently generated by MODELLER for receptor and ligand are shown for two benchmark cases: (A) 1PXV and (B) 1ACB. Conformers were superimposed onto the corresponding molecules in the reference complexes for visualization. Only interface side chains are shown for the sake of clarity.

Here we explored in a systematic way whether a minimal description of the conformational heterogeneity of the interacting proteins could significantly improve their binding capabilities. For that purpose, we created conformational ensembles from the unbound structures (for both the receptor and the ligand) of complexes from the protein-protein benchmark 3.0 [54]. Ensembles of 100 conformers were initially generated by using two distinct conformational sampling procedures, one being a fast energy optimization as implemented in MODELLER, and the other being the much more computationally demanding Molecular Dynamics method, as implemented in AMBER package (see Methods). **Fig 1** shows examples of the typical conformational heterogeneity (at backbone and side chain level) generated by MODELLER minimization (MM). The deviation of the interface atoms from the initial unbound structure was 1.2 Å RMSD on average (ranging from 0.6 Å for 1R0R receptor to 7.7 Å RMSD for 2QFW receptor).

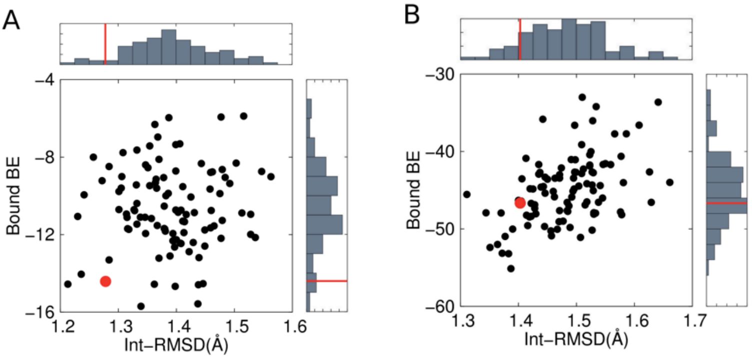


Fig 2. Distribution of geometrical and energetic values for ensemble conformers. Correlation between the full atom interface RMSD (Int-RMSD) with respect to the bound state and the binding energy towards the bound partner in the native orientation (bound BE) for all conformers in MODELLER ensembles are shown for two benchmark cases: (A) 2F0R (1S1Q receptor) and (B) 1MKF (1ML0 receptor). Distribution of Int-RMSD and bound BE values are shown as histograms. Data for the unbound x-ray structure are shown in red.

We compared the unbound models with respect to their native poses in the complex to structurally characterize these conformers and to estimate their capabilities for binding. In order to do that, we first superimposed each model into the native conformation and then computed the following parameters i) the RMSD for all Cα atoms (Cα-RMSD) with respect to the native structure; ii) the RMSD for all interface atoms (Int-RMSD) after superimposing only those interface atoms; iii) the binding energy with the bound partner.; iv) the binding energy with the unbound partner; and v) the number of clashes with the bound partner. The values for these parameters in the different conformers generated by MODELLER are randomly distributed following a Gaussian function (**S1 Fig**). Except for a few cases, like the viral chemokine binding protein M3 (1ML0 receptor), there is no significant correlation between the binding energy of the different conformers in the native orientation and their similarity with respect to the bound structure (**Fig 2**). Perhaps the main reason for this is that, in general, these conformers are not exploring the vicinity of the bound state. Indeed, only 20% of the benchmark proteins contain conformers within 1.0 Å Int-RMSD from the bound state (actually, in virtually all of these cases the unbound state already had Int-RMSD < 1.0 Å from bound).

Ensembles generated by MD showed larger conformational variability, but in general they were not closer to the bound state (**S1** and **S2 Fig**). Increasing the number of conformers to 1,000 (**S3 Fig**) did not significantly modify the range of conformational variability for either sampling method.

We aimed to identify which conformers of the ensemble seemed more promising for binding. Thus, we selected the best conformers of the ensemble according to the criteria analyzed in the previous section. **Fig 3** shows the best conformer according to each parameter as compared to that of the unbound structure for all benchmark cases. Regarding the RMSD with respect to the complex structure, only in a few cases (21% and 6%, according to Cα-RMSD and Int-RMSD, respectively) the best pair of conformers were significantly better (i.e., more than 10% change) than the unbound X-ray structure (and were not particularly enriched in conformers with Int-RMSD < 1.0 Å).

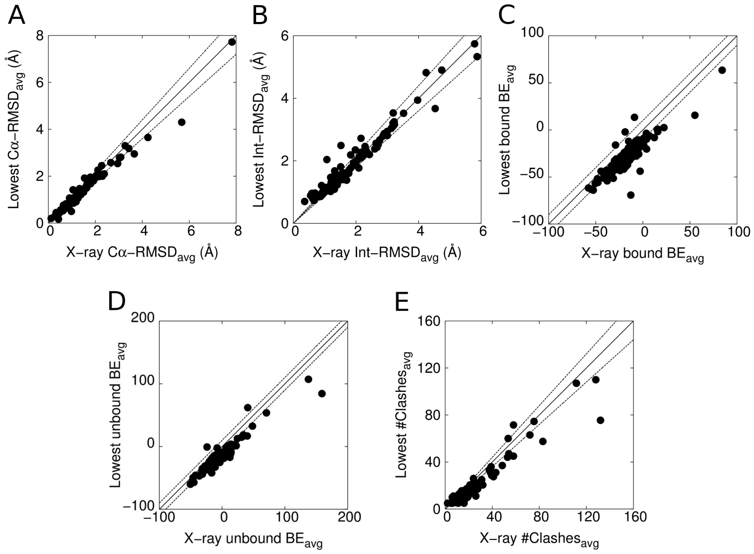


Fig 3. Best ensemble conformers according to quality criteria based on the complex native orientation. For each benchmark case, it is shown the best pairs of receptor and ligand conformers in the conformational ensemble according to the following criteria: (A) Cα-RMSD, (B) Int-RMSD, (C) binding energy with the bound partner, (D) binding energy with the unbound partner, (E) number of clashes with respect to the bound partner. The above described descriptors were calculated independently for the best receptor and ligand conformers and then averaged. These are compared to those of the unbound X-ray structures. Dashed lines represent the (arbitrary) range of variation that we used to consider a change as significant, and it was defined as 10% in the RMSD- and clash-based criteria, or 10 a.u. in the energy-based criteria.

Interestingly, we found a much higher number of cases in which the best conformers showed significantly better binding energy (in 46% and 51% of cases, when considering the bound or unbound structure as partner, respectively), or fewer clashes (in 69% of cases) than the unbound X-ray structure. It is remarkable that the improvement in binding energy was independent of the structural similarity to the bound. Again, the reason can be that in the majority of cases there is no real sampling around the bound state, and therefore, in such unbound minima any small improvement towards the bound state is not relevant in binding energy terms.

Although MD ensembles showed larger conformational variability (**S1** and **S2 Figs**), the percentage of cases with conformers that became significantly better than the unbound state according to each of the above mentioned criteria (12%, 7%, 37%, 62%, and 69%, respectively) was very similar to those observed for the MODELLER ensembles. Surprisingly, MD showed worse conformational sampling around the bound state, since less than 4% of the cases had conformers with Int-RMSD < 1.0 Å with respect to the bound state (as compared to 20% in MODELLER).

Selected conformers can yield significantly better docking results than unbound subunits

The fact that in the majority of cases the conformational ensembles contained conformers that showed better binding energy capabilities than the unbound X-ray structure encouraged us to evaluate their use for docking. Since the systematic cross-docking of all conformers for receptor and ligand would be impractical, we tried instead to estimate the expected performance of the unbound ensembles for docking in the best-case scenario. Therefore, based on the native orientation we selected conformers that seemed the best candidates to improve docking predictions, that is, those with: i) the lowest Cα-RMSD with respect to the bound state, ii) the lowest Int-RMSD, iii) the best binding energy with the bound partner, iv) the best binding energy with the unbound partner, and v) the smallest number of clashes with the bound partner. These conformers were used in protein-protein docking as described in the Methods section.

Fig 4A shows the docking success rates for the top 10 predictions when using these selected conformers, with all the details in **Table 1**. Interestingly, the results do not significantly change when using a larger number of conformers (1,000) generated by MODELLER (and applying the same procedure for selecting the best expected conformers), or when conformers were generated by Molecular Dynamics, either using 100 or 1,000 conformers (**S4 Fig**). Strikingly, when we used the best conformers based on Cα- or Int- RMSD with respect to the complex structure, the docking results were slightly worse than those of unbound docking, as can be seen in **Fig 4A** (the results did not significantly change when selecting only those cases in which the best conformer had significantly better Cα- or Int-RMSD than that of the unbound structure). This can be due to the fact that either MODELLER minimization or a short MD trajectory cannot generally sample too far from the unbound structure, and therefore cannot reach the vicinity of the bound state in most of the cases. However, when using the conformers that would give the best binding energy or the smallest number of clashes when in the native orientation, the docking results significantly improved with respect to those of the unbound structures, as can be seen in **Fig 4A**. Again, this did not correspond to an improvement in geometrical terms (e.g., in 99% of the cases in which the best-energy conformer improved the docking predictions, such conformer did not have significantly better Int-RMSD than the unbound structure). For comparison, we show the success rates that we would obtain when using the bound structures, which establish the upper limit for docking with this approach. The success rates of the binding energy-based selected conformers are more than half of the maximum expected success rates when using the bound structures.

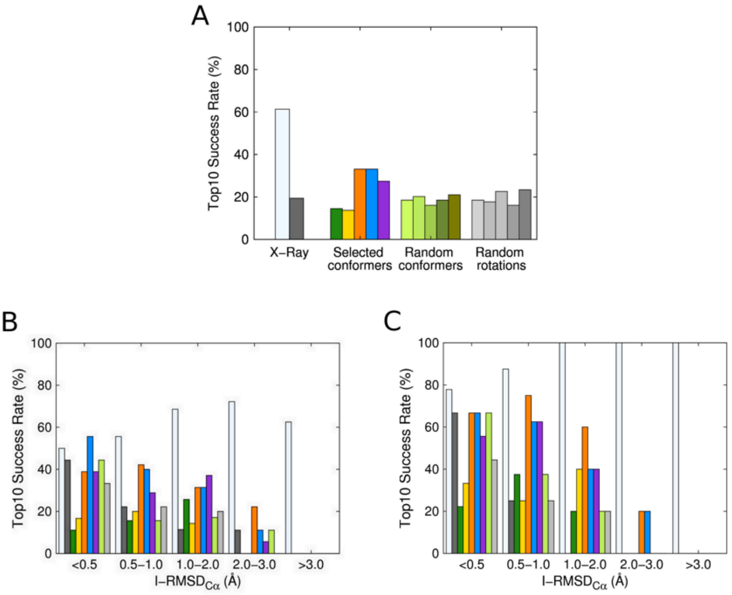


Fig 4. Docking performance for selected conformers. (A) Docking success rates for the top 10 predicted models on the protein-protein docking benchmark when using selected conformers according to specific criteria: Cα-RMSD (green), Int-RMSD (yellow), binding energy towards the bound partner (orange), binding energy towards the unbound partner (blue), number of clashes with respect to the bound partner (magenta). For comparison, the docking success rates for bound (white) and unbound (dark gray) X-ray structures are also shown. To show the significance, docking rates for five random conformers pairs (green gradations) and five random initial rotations of the unbound docking partners (gray gradations) are also shown. (B) Docking success rates according to conformational variability between unbound and bound structures for selected conformers (same color code as above). For comparison, docking success rates for bound and unbound X-ray structures, as well as for one random conformers pair (light green) and one random initial rotation of the unbound docking partners (light gray) are also shown. (C) Docking success

rates according to unbound-bound conformational variability on the 28 cases of the benchmark with reported high affinity (ΔG < -12.0 kcal/mol) when using selected conformers, as well as bound and unbound X-ray structures, one random conformers pair and one random initial rotation of the unbound docking partners (same color code as above).

Table 1. Docking performance of conformers selected from MODELLER ensembles.

PDB	*Bound*	*Unb.*	*Cα-RMSD*	*Int-RMSD*	*Bound BE*	*Unb. BE*	*Clashes*
Rigid (I-RMSD$_{C\alpha}$ < 0.5 Å) (18 cases)							
1AVX	**2**	**102**	**-**	**33**	**40**	**1**	**231**
1FSK	**3**	**3**	**39**	**34**	**1**	**1**	**514**
1GHQ	7455	-	-	6528	-	1878	-
1IQD	**1**	**8**	**64**	**3**	**6**	**6**	**3**
1KLU	18	1246	6002	4468	2587	6498	1647
1KTZ	48	3725	6333	-	-	-	309
1NCA	14	7	1269	-	1332	7	1
1NSN	**405**	**500**	**254**	**5587**	**33**	**33**	**1085**
1PPE	**28**	**6**	**12**	**2**	**5**	**1**	**4**
1R0R	**1**	**3**	**258**	**230**	**9**	**17**	**37**
1SBB	161	298	73	-	-	-	-
1WEJ	**1**	**274**	**5**	**456**	**64**	**2**	**9**
2JEL	1	42	16	25	12	2	1
2MTA	2	78	61	187	48	3	554
2PCC	12	6	91	12	6	4	11
2SIC	**1**	**8**	**3378**	**1**	**2**	**249**	**1**
2SNI	**1**	**3**	**1**	**16**	**1**	**1**	**1**
2UUY	69	4472	4801	64	159	11	1997
Low-flexible (I-RMSD$_{C\alpha}$ 0.5-1.0 Å) (45 cases)							
1AHW	1043	4049	6796	838	431	836	2974
1AY7	**1**	**24**	**130**	**118**	**4**	**2**	**7**
1AZS	1	30	-	-	6	6	-
1BJ1	9	-	-	-	18	9	25
1BUH	71	66	209	426	36	24	119
1BVN	**1**	**2**	**2**	**1**	**1**	**1**	**687**
1DQJ	216	**604**	**261**	**3363**	**75**	**25**	**223**
1E96	113	1	59	168	73	5	130
1EAW	**8**	**622**	**297**	**86**	**42**	**25**	**1**
1EFN	6	166	197	1684	203	97	172
1EWY	4	8	200	5	10	10	1
1F34	**1**	**139**	**174**	**226**	**52**	**280**	**2**
1F51	2	7	13	375	1505	130	8
1FQJ	14	309	396	482	218	438	101
1GCQ	274	1091	574	1540	5	5	364
1GLA	61	50	-	12	6	21	131
1GPW	1	1	1	1	1	1	1
1HE1	1	3958	102	4506	2425	523	2629

1HE8	138	2917	2612	1503	277	242	3437
1IJK	16	1309	69	61	493	487	388
1J2J	46	19	303	18	2	3	5
1JPS	**709**	**481**	**-**	**2135**	**1**	**2**	**-**
1K4C	-	-	3036	3369	2275	-	2379
1K74	150	14	172	82	1	1	24
1KAC	4737	1286	3545	990	107	19	917
1KXQ	1	250	8	4	4	1	1
1MAH	**1**	**19**	**2**	**4**	**4**	**1**	**1**
1MLC	2	37	50	10	1	97	144
1N8O	3	53	-	-	5	-	90
1QA9	3253	7378	5902	6152	1546	37	7973
1QFW	81	239	234	-	26	21	72
1RLB	1319	4094	-	7917	-	-	-
1S1Q	147	1211	2994	541	164	175	87
1T6B	**3**	**56**	**802**	**1464**	**2**	**11**	**3**
1TMQ	1	1	27	4	54	-	4
1UDI	1	1	2	47	1	1	420
1YVB	1	19	1	2	3	21	7
1Z0K	2	8	523	57	42	11	44
1ZHI	5	3	7450	-	196	5	5
2AJF	5	1788	-	311	562	2268	2122
2B42	**1**	**1**	**2**	**37**	**1**	**2**	**21**
2BTF	1	33	120	26	60	9	250
2OOB	588	112	131	217	106	547	432
2VIS	64	-	-	-	-	-	-
7CEI	1	19	11	1	1	1	20
Medium-flexible (I-RMSD$_{C\alpha}$ 1.0-2.0 Å) (35 cases)							
1A2K	36	114	5641	284	-	-	782
1AK4	2420	2040	3983	3619	-	2721	1055
1AKJ	89	656	345	261	204	162	1168
1B6C	1	3	6	11	1	1	21
1BGX	1	-	-	-	-	-	-
1BVK	7	18	4	146	87	85	2
1D6R	1050	2128	227	888	669	785	102
1DFJ	**6**	**557**	**2**	**1**	**1**	**1**	**4**
1E6E	1	3	2	1	1	8	1
1E6J	-	33	34	3	1	2	5
1EZU	**1**	**2048**	**3633**	**-**	**1449**	**1547**	**102**
1FC2	127	-	233	326	1256	683	171
1GP2	1	-	-	842	85	-	87
1GRN	2	858	184	1184	450	23	2909
1HIA	99	40	415	42	23	166	7
1I4D	1	-	-	642	-	44	132
1I9R	15	846	568	212	-	99	-
1K5D	**1**	**360**	**85**	**-**	**2**	**610**	**-**
1KXP	**1**	**16**	**14**	**1**	**1**	**1**	**1**
1ML0	1	173	80	140	1	1	9
1NW9	1	9	181	36	43	39	181
1OPH	59	14	-	469	-	-	2584

1VFB	37	59	86	59	128	31	95
1WQ1	4	2448	5	1077	16	6	6
1XD3	1	1	3	13	2	1	1
1XQS	1	14	55	628	1	8564	7
1Z5Y	1	16	320	-	4	39	17
2CFH	1	1904	202	1394	4066	43	5
2FD6	68	31	-	-	-	81	1
2H7V	1	-	734	-	1091	-	-
2HLE	1	13	1	1	2	1	3
2HQS	1	30	2	30	146	146	129
2I25	**1**	**40**	**443**	**1520**	**15**	**948**	**3599**
2O8V	1	60	5	186	220	1	-
2QFW	1	-	19	-	-	7	73
Flexible (I-RMSD$_{C\alpha}$ 2.0-3.0 Å) (18 cases)							
1ACB	**1**	**361**	**144**	**668**	**6**	**4**	**15**
1BKD	2	522	157	1050	99	114	646
1CGI	1	19	98	13	1	12	5
1DE4	1	-	-	366	-	-	-
1E4K	104	1215	722	148	200	4249	74
1EER	**3**	**1821**	**91**	**21**	**81**	**37**	**675**
1I2M	**1**	**-**	**683**	**632**	**50**	**149**	**247**
1IB1	34	-	2116	7028	255	2775	1626
1IBR	**1**	**-**	**-**	**-**	**-**	**-**	**-**
1KKL	88	49	271	176	1	2	289
1M10	1	81	5742	574	-	21	2873
1N2C	1	-	-	-	-	16	-
1PXV	**1**	**2073**	**100**	**429**	**673**	**1498**	**2375**
2C0L	83	3958	1024	1589	-	5105	3834
2HMI	2	-	-	-	-	-	-
2HRK	49	16	23	47	83	83	241
2NZ8	1	10	5509	247	2	168	5848
2OT3	1	5	212	14	91	-	131
Highly-flexible (I-RMSD$_{C\alpha}$ > 3.0 Å) (8 cases)							
1ATN	**7**	**2568**	**-**	**-**	**-**	**665**	**-**
1FAK	41	5327	-	-	43	41	-
1FQ1	6	3865	4315	7901	927	-	4833
1H1V	537	-	-	-	-	-	-
1IRA	1	-	-	-	-	-	-
1JMO	1	5325	-	5398	2969	5510	5547
1R8S	1	-	-	-	-	4043	-
1Y64	1420	-	-	-	-	1329	-

In bold: high affinity cases

As mentioned above, the results of bound docking are not optimal mostly due to the FFT-based discrete searching algorithm. This would be particularly critical in low-affinity cases, in which the small number of interactions would make them less tolerant to small errors in the atomic positions. To minimize the impact of this limitation in our evaluation, we have performed the same analysis as above but focusing only on the 28 cases of the benchmark that have been experimentally defined as high-affinity (DG < -12.0 kcal/mol), for which the results of bound docking are close to optimal. Under these conditions, we can observe even more clearly that the selected conformers improved the docking success rates in the low-flexible cases (**Fig 4C**).

In order to provide a statistical significance for these results, we randomly selected five conformers from the conformational ensemble. The results for each random conformer were similar (within experimental error) to those of the unbound structure (**Fig 4A**), which shows that the conformers selected according to the optimal binding energy improved significantly the docking results with respect to the randomly chosen conformers. An alternative possible explanation for the docking improvement when using ensemble conformers might be related to the limited sampling of FTDock discrete searching algorithm derived from the fix number of ligand rotations (which makes coarser surface sampling for large proteins) and the grid resolution of 0.7 Å (which introduces inaccuracies in the atomic coordinates). This creates a stochastic dependence of the FTDock docking algorithm on the initial rotation of the interacting subunits, and is indeed the cause of the suboptimal results shown even for bound docking, given that the exact complex orientation is very unlikely to be sampled. This is a limitation of any FFT-based algorithm, and it was shown before that performing parallel docking runs using several initial rotations provided more consistent docking results than using just a single one [20]. To evaluate the possibility that the extensive sampling in the atomic positions provided by the use of conformational ensembles prior to docking could compensate the suboptimal grid-based sampling of FTDock, we performed five different docking runs with random initial rotations for the unbound receptor and ligand molecules. The results from the individual random rotations were similar, within experimental error, to the unbound docking results (**Fig 4A**).

These results suggest that the selected conformers according to specific criteria (i.e., optimal energy, number of clashes) were more beneficial for docking than just a random selection of conformers or initial rotations. Overall, this clearly shows that conformational heterogeneity in the interacting subunits improves the binding capabilities of the unbound X-ray structures.

Conformational heterogeneity is particularly beneficial for low- and medium-flexible cases

We have analyzed whether the docking improvement when using ensembles depends on the conformational rearrangement of the interacting proteins upon binding (see Methods). The largest docking improvement when using the selected conformers is observed in the low- and medium-flexible cases, i.e. those with I-RMSD$_{C\alpha}$ between 0.5 and 2.0 Å (**Fig 4B**). The ensemble success rates are particularly good in the low-flexible cases, for which they reach predictive docking values similar to the optimal ones when using the bound structures. This could be related to the limited sampling used here, which did not explore too far from the unbound (1.2 Å of Int-RMSD as average) and therefore they can only sample in the vicinity of the bound state in low-flexible or rigid cases. Indeed, in the rigid cases (I-RMSD$_{C\alpha}$ < 0.5 Å), the selected conformers yield similar results to the unbound structures. In these cases, unbound structures already produced optimal results, similar to the optimal success rates obtained when using the bound structures. In flexible or highly-flexible cases (I-RMSD$_{C\alpha}$ > 2.0 Å), the docking results for the ensembles are as poor as those for the unbound structures, very far from the optimal success rates when using the bound structures. Using MD or more conformers does not significantly change the results (**S5 Fig**).

Discussion

Conformers providing better binding energy in the native orientation are more likely to improve docking

We have shown that set of discrete conformers representing the conformational heterogeneity of the unbound structure yielded better docking results than the unbound structures themselves. It would be important to analyze the reasons for the success of such conformers. Surprisingly, the conformers that were structurally more similar to the reference structure did not yield better docking results than the unbound structures. On the other side, selected conformers with the best binding energy in the native orientation yielded better docking results than the unbound structures. Thus, the capacity to provide favorable binding energy in the native orientation seems to be a major determinant for the success of docking, as opposed to the criterion of structural similarity to the native conformation. This might be due to the fact that in the majority of cases, ensembles are not exploring the conformational space close to the bound state, because sampling is limited to a region in the vicinity of the unbound.

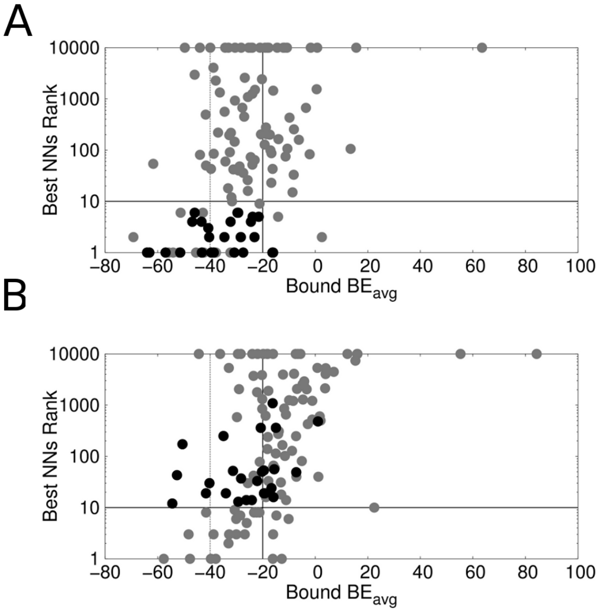


Fig 5. Docking performance dependence on energetic complementarity of the docking partners. Best rank of any near-native docking solution vs. average native-oriented binding energy towards the bound partner calculated for (A) best pair of conformers according to binding energy towards the bound state, and (B) unbound X-ray structures. Highlighted in black are the cases that largely improve docking performance (from near-native rank > 10 to rank ≤ 10) using the energy-based selected conformers.

Fig 5A shows, for each case, the best ranked near-native solution obtained when docking the conformers that had the best native-oriented energy with the bound partner (i.e., best near-native rank in ordinates; average native-oriented energy of best pair of conformers in abscissas). As we can see in **Fig 5A**, 90% of the successful cases (i.e., near-native solution ranked within top 10) have average conformer binding energy < -20.0 a.u. in the native orientation. Actually, 71% of the docking cases with conformers with binding energy in the native orientation < -40.0 a.u. were successful. This confirms that the existence of conformers with good optimal energy in the native orientation is determining the success of docking. **Fig 5A** highlights the cases that significantly improved, i.e. which had a near-native ranked ≤ 10 when using the energy-based selected conformers but not when using the unbound structures. In many of these cases, the unbound structures in the native orientation had binding energy < -20.0 a.u. (**Fig 5B**) but were not successful in unbound docking. In these cases, a little bit of conformational sampling seems to be sufficient to generate conformers that significantly improve docking results.

Ensembles in docking: does size really matter?

For a practical use in docking, the conformational ensembles should provide a reasonable coverage of the conformational space, using a minimal number of conformers. We have shown here that the selected conformers (based on the reference complex structure) from the 1000-member ensembles generated by MODELLER or MD yielded similar results to those selected from the 100-member ensembles (**S4 Fig**). Especially in the more rigid cases, a larger conformational ensemble does not seem to help to find better conformers to improve docking results. However, we can observe a small improvement in the flexible cases when using the larger ensembles (**S5 Fig**). Perhaps, in addition to larger ensembles, higher conformational variability would be needed in order to see further improvement in the flexible cases. In this sense, we have performed extended MD simulations (100 ns), at different temperatures (300K and 340K), on a random selection of 11 cases with no missing long loops (comprising all ranges of flexibility values). The 1000-member ensembles from these extended MD simulations showed larger conformational variability as compared to the shorter simulations. However, these larger ensembles did not increase the number of cases with conformers significantly closer to the bound structure, neither provided better docking success rates (S1 Table). Given the known convergence issues in MD [55], it seems that much longer MD trajectories would be needed in order to achieve exhaustive sampling of the unbound conformational space.

Binding mechanism: What can we learn from docking?

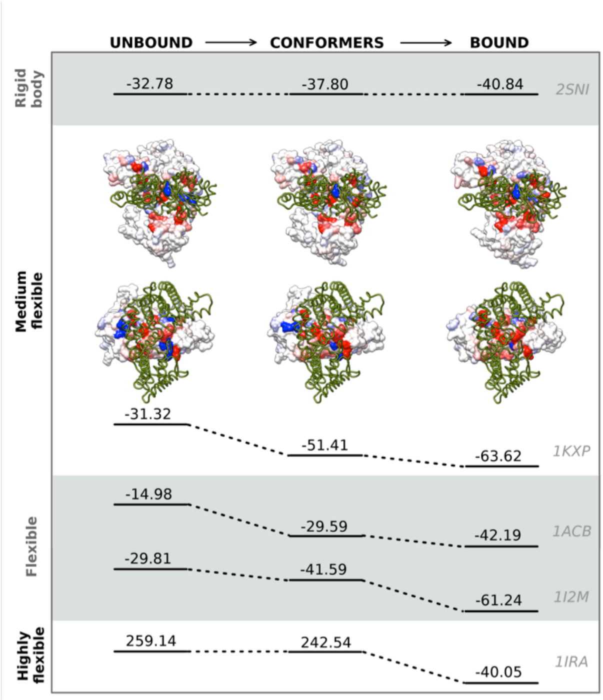


Fig 6. Binding energy towards the bound of unbound, best conformer and bound. Average binding energy towards the bound in the native orientation for the unbound proteins, best conformer according to native-oriented bound energy, and bound state, computed for examples of cases with different degrees of unbound-to-bound conformational changes. Similar unbound, conformer and bound binding energies suggest lock-and-key binding mechanism (as in 2SNI). Conformer binding energy better than unbound and similar to bound suggest conformational selection model (as in 1KXP, 1ACB). Conformer binding energy similar to unbound and worse than bound could be compatible with conformational selection (1I2M; see main text) or induced fit mechanism (1IRA).

The different possible mechanisms that have been proposed for protein-protein association could be described by existing computational approaches. In this context, we can consider several possible scenarios. For protein complexes following rigid association (similar to "lock-and-key" mechanism), the use of rigid-body docking with the unbound subunits could be a suitable approach to describe the binding process and obtain good predictive models. Indeed, this seems to be the case for complexes with small conformational changes between unbound and bound states (I-RMSD$_{C\alpha}$ < 0.5 Å), in which unbound docking already gives as similar success rates as bound docking (**Figs 4B** and **4C**). In these cases, the use of energy-based selected conformers from unbound ensembles gives also similarly good docking rates as unbound and bound docking. Indeed, **Fig 6** shows one example of rigid-body docking (2SNI) in which the unbound proteins in the native orientation showed good average binding energy towards the bound partner (-32.8 a.u.), not far from that of the bound structures (-40.8 a.u.). Consistently, the average binding energy of the best conformers were similar to that of the unbound or bound pairs (-37.8 a.u.). However, when conformers were selected by criteria of structural similarity to bound state, docking success rates were much worse than unbound or bound docking, because in these cases conformational heterogeneity is more likely to produce conformers that are further from the bound state than the unbound one (given that the unbound was already close to the bound state). Indeed, in none of these cases there were a single conformer that was significantly closer (in terms of Int-RMSD) to the bound state than the unbound structure.

On the other side, we know that in complexes involving flexible association rigid-body docking with the unbound structures is not going to produce correct models. For such cases, different binding mechanisms have been proposed, such as conformational selection or induced fit. For cases following the conformational selection mechanism, the hypothesis is that the unbound proteins naturally sample a variety of conformational states, a subset of which are suitable to bind the other protein. Therefore, for these cases the use of precomputed unbound ensembles describing conformational variability of free proteins in solution should generate conformers that would improve rigid-body docking predictions with respect to those with the unbound structures. Indeed, this is the case for the complexes undergoing unbound to bound transitions between 0.5 and 1.0 Å I-RMSD$_{C\alpha}$. In these cases, selected conformers from the unbound ensembles yielded much better docking predictions than the unbound structures, virtually achieving the success of bound docking (**Fig 4B**). For cases undergoing unbound-to-bound transition between 1.0 and 2.0 Å I-RMSD$_{C\alpha}$, the use of unbound ensembles also improved the predictions with respect to the unbound docking results, although to a lesser extent (**Fig 4B**). **Fig 6** shows one of these cases, 1KXP, in which binding energy of the selected pair of conformers in the native orientation (-51.4 a.u.) is better than the unbound structures (-31.3 a.u.) and similar to the bound structures (-63.6 a.u.). Some residues in the best pair of conformers show better energy contribution than in the unbound state, which explains why this specific pair of conformers improves docking results. In these cases, the existence of a subpopulation of "active" conformers, i.e. with good binding capabilities towards the bound partner, would be consistent with a conformational selection mechanism. The fact that these conformers with improved binding capabilities are not geometrically closer to the bound state seems counterintuitive. However, recent views of binding mechanism show that active conformers that are selected by partner (initial encounters) do not necessarily need to be in the bound state, as they can adjust their conformations during the association process [41]. Our docking poses are likely to represent these initial encounters between the most populated conformational states of the interacting proteins and would be compatible with this extended conformational selection view [41]. However, in other cases the limited conformational sampling used here might not be sufficient to explore all conformational states available in solution and therefore the specific binding mechanism cannot be easily identified.

As for the other extreme, in cases following an "induced-fit" mechanism the bound complexes would only be obtained after rearrangement of the interfaces when interacting proteins are approaching to each other, in which case the use of precomputed conformational ensembles in docking (even if generated by exhaustive sampling) would not produce favorable encounters around the native complex structure. This seems the case for complexes undergoing unbound to bound transitions above 3.0 Å I-RMSD$_{C\alpha}$. In all these cases, rigid-body docking, either with unbound structures or with selected conformers, fails to reproduce the experimental complex structure. **Fig 6** shows one of these highly-flexible cases, 1IRA, in which binding energy of the selected pair of conformers is similar to the unbound structures and much worse than the bound conformation. For these complexes, the use of precomputed unbound ensembles does not seem to see advantageous, and they would probably need to include flexibility during docking search, mimicking the induced fit mechanism. However, in the flexible category (i.e., unbound to bound transitions between 2.0 and 3.0 Å I-RMSD$_{C\alpha}$,), there are cases like 1ACB, which seem to follow the (extended) conformational selection mechanism, since the use of conformers helps to improve the docking results, and the conformers show better energy than the unbound structures (**Fig 6**). Again, there might be other complexes under this category that could still follow the conformational selection mechanism, but our conformational search was not sufficient to sample conformations that may exist in solution and could be productive for docking. This seems to be the case for 1I2M, in which ensembles based on MODELLER did not produce pairs of conformers with sufficiently good binding energy in the native orientation (**Fig 6**), but the docking rates improved when using extended sampling based on NMA (see later).

Of course the use of docking calculations to learn about the binding mechanism has some limitations, in addition to the ones already mentioned. The timescale of transitions between inactive and active conformers can play an important role in controlling the binding mechanism [56]. In the present work, we can only assume that our ensembles are formed by conformers that are the most accessible in solution, so the existence of active conformers that can be preferentially selected by the bound partner would be compatible (but not exclusively) with a mainly conformational selection mechanism. However, in a situation in which the active conformers are not easily accessible, as those that can only be generated with extended sampling, we could not identify the type of mechanism unless transition rates between conformers were considered.

Future perspectives: improving sampling with normal mode analysis for flexible cases

We have shown here that cases with large deformation after binding (I-RMSD$_{C\alpha}$ > 2 Å) do not generally benefit from the use of conformers from unbound ensembles generated by MODELLER or MD. This suggests that these complexes could follow the induced fit binding mechanism, and therefore, the use of precomputed unbound ensembles would not be appropriate to describe their association. However, we should not disregard that some of these complexes could still follow a conformational selection mechanism, but for some reason a dramatically larger conformational sampling would be needed to find suitable conformers.

One way to extend conformational sampling is by using Normal Mode Analysis (NMA). When generating 100 conformers for this group of cases (strong and flexible I-RMSD$_{C\alpha}$ > 2.0 Å) with an ad-hoc Monte-Carlo sampling method based on Cα NMA and full-atom rebuilding with MODELLER (**S6 Fig**; see Methods), the results were not better than those obtained with the conformers directly generated by MODELLER (**Tables 2** and **S2**). However, when generating 1000 conformers based on NMA (either 1000 NMA-based conformers rebuilt by MODELLER, or 100 NMA-based conformers with 10 models rebuilt by MODELLER for each of them), the success rates largely improved with respect to those obtained when generating conformers with only MODELLER (either 100 or 1000 conformers). It is interesting to comment on the flexible case 1I2M, which showed failing docking rates with the unbound structure and also with the best conformers from MODELLER or MD ensembles, but yielded successful docking results with 1000-member NMA-based ensembles. This shows that new sampling approaches based on NMA could produce the type of enhanced sampling needed for the most flexible cases following a conformational selection mechanism.

Table 2. Docking performance of conformers selected from NMA-based ensembles.

	Bound	*Unb.*	*MM (100 confs)*	*MM (1000 confs)*	*NMA (100 conf) * MM (1conf)*	*NMA (100 conf) * MM (10conf)*	*NMA (1000 conf) * MM (1conf)*
1ACB	1	361	4[d]	34[d]	46[c]	3[d]	1[d]
1ATN	7	2568	665[d]	-	292[c]	3245[a]	1788[a]
1EER	3	1821	21[b]	13[c]	17[d]	3[c]	3[b]
1I2M	1	-	50[c]	13[c]	23[c]	1[c,d]	1[d]
1IBR	1	-	-	1108[e]	87[a]	146[e]	88[a]
1PXV	1	2073	100[a]	822[c]	168[d]	168[d]	232[d]

[a] *Cα global RMSD .*
[b] *Full-atom interface RMSD.*
[c] *Native-oriented binding energy with bound partner.*
[d] *Native-oriented binding energy with unbound partner.*
[e] *Number of clashes with bound partner in the native orientation.*

We present here the most complete systematic study so far about the use of precomputed unbound ensembles in docking. The results show that considering conformational heterogeneity in the unbound state of the interacting proteins can improve their binding capabilities in cases of moderate unbound-to-bound mobility. In these cases, the existence of conformers with better binding energy in the native orientation is associated to a significantly improvement in the docking predictions. It seems that protein plasticity increases chances of finding conformations with better binding energy, not necessarily related to bound geometries. This is compatible with the extended conformational selection mechanism, since successful conformers are not necessarily more similar to the bound conformation in structural terms. Other moderately flexible cases have conformers that look promising from a binding energy perspective but did not provide good docking predictions. These cases could also follow a conformational selection mechanism, but they would need extensive sampling to find suitable conformers for binding. The most flexible cases would show larger induced fit effects and therefore would not be well described by ensemble binding. This work helps to set guidelines for future strategies in practical docking predictions based on unbound ensembles generated by molecular mechanics minimization.

Methods

Generation of protein conformational ensembles

We applied three different computational techniques to generate conformational ensembles starting from the unbound protein structures: modelling minimization (MM), Molecular Dynamics simulations (MD), and Normal Modes Analysis (NMA).

Conformational search based on modelling minimization (MM) was performed with the comparative modelling program MODELLER version 9v10 [57], using as template the unbound X-ray structure of the same protein. Cofactors and ligands, if present in the template structure, were taken into account during the modelling procedure.

Conformational search based on Molecular Dynamics (MD) was performed by a 10-ns-long explicit solvent unrestrained MD simulation on the unbound structure using the force field AMBER parm99 and the AMBER8 package [58]. As a first preparation step, all the missing loops in the protein structures, if any, were modeled using MODELLER program, in order to avoid an over-estimation of the protein flexibility during the simulations. The parameterization of each system was performed using AMBER's module LEAP, whereas the cofactor and ligand libraries, when needed, were written with the AMBER modules ANTECHAMBER and LEAP. Each system was then minimized, solvated and equilibrated at the same conditions as previously described for the MoDEL database [59]. Then, a 10-ns MD simulation was performed in isothermal-isobaric ensemble, setting pressure to 1 atm and temperature to 300K. Finally, two conformational ensembles were created by extracting trajectory snapshots every 10ps or 100ps. Additionally, a random subset of 11 benchmark cases (1ACB, 1AY7, 1D6R, 1E6J, 1GCQ, 1IRA, 1JMO, 1PXV, 2HRK, 2CFH, 2C0L) was selected for longer simulations. Each protein underwent two 100-ns-long explicit solvent unrestrained NPT-MD simulations, at the temperatures of 300K and 340K, respectively, using the same force field as above.

Conformational search based on Normal Mode Analysis (NMA) was performed by an in-house protocol on a small subset of 6 high-affinity and flexible benchmark cases. NMA is a powerful modeling technique that allows for a fast and accurate description of the intrinsic movements of biomolecules. Modern interpretations of the procedure use the elastic network model (ENM), first described by Tirion as an all-atom version [60], and later re-formulated as coarse-grained [61-63]. In the ENM, the biomolecule is represented as a network of connected atoms, where each node is connected to all the atoms within a cutoff, and the springs represent the interactions between the nodes. Here we used the Anisotropic Network Model [62] that describes the protein as a Ca model, and we assigned the spring constants by a continuum distance function that assumes an inverse exponential relationship with the distance [64]. We tried to enhance the conformational space by introducing an iterative exploratory search. The proposed method is called eNMA (enhanced NMA) and creates enriched structurally diverse ensembles. The algorithm works as follows:

Step 1 – Starting from the unbound Ca atoms, we created 100 discrete cartesian conformers from random combinations of displacements along the first 10 Normal Modes (as described elsewhere [65]). The average Ca displacement with respect to the original structure was set to ~1 Å.

Step 2 – The resulting conformers were then clustered hierarchically via *average linkage* method (as implemented in ptraj10 [58]) to obtain 100 diverse conformations.

Step 3 – Each conformer from the cluster was sent to Step 1, and the whole cycle was started.

Step 4 – The process was ended up after 8 iterations.

In total, a maximum of 70100 intermediate structures were created per protein, but we only kept the ones resulting from the clustering (i.e., 100 x 10 = 1000 discrete conformers). The final structures underwent a last modeling step with MODELLER 9.10. All-atom models were rebuilt by adding missing atoms and side chains and were atomically refined with MODELLER (using the original Ca model as template) to fix incorrect bond distances [57, 66]. In addition, 100 discrete conformers were randomly selected and for each of them 10 MODELLER models were built. The whole procedure took around 4 hours per protein on 1 CPU of a standard Linux workstation. Note that our conformational search was unguided, but it could be also guided in future applications (i.e., selecting the combination of models that provides the best score on a given fitness function).

Docking Simulations

For all the dockings experiments, FTDock docking program [15] was used to generate 10,000 rigid-docking poses based on surface complementary and electrostatics at 0.7 Å grid resolution, and then, each docking solution was evaluated by the energy-based pyDock scoring scheme [25], based on desolvatation, electrostatics and Van der Waals energy contributions. Cofactors and ions were excluded during the sampling and the scoring calculations.

Benchmark

In order to validate the approach proposed here, we used protein-protein docking benchmark 3.0 [54], comprising a total of 124 test cases in which the structure of both the free components and the complex are known. We have classified these cases according to the conformational variation of the proteins from the unbound to the bound state (based on the RMSD of Cα atoms of the interface residues as defined in the mentioned protein-protein benchmark 3.0), which resulted in the following categories: "rigid" (I-RMSD$_{C\alpha}$ < 0.5 Å), "low-flexible" (0.5 Å < I-RMSD$_{C\alpha}$ < 1.0 Å), "medium-flexible" (1.0 Å < I-RMSD$_{C\alpha}$ < 2.0 Å), "flexible" (2.0 Å < I-RMSD$_{C\alpha}$ < 3.0 Å), and "highly-flexible" (I-RMSD$_{C\alpha}$ > 3.0 Å). The quality of the docking predictions was evaluated according to the ligand Cα-RMSD with respect to the complex crystal structure (after superimposing the receptor molecules). A docking experiment was considered successful if a near native solution (a docking pose with ligand Cα-RMSD < 10 Å) was ranked among the top 10 predictions according to the pyDock scoring function. Structural analyses of proteins, including RMSD and clashes calculations, were performed using ICM program [67] (www.molsoft.com).

Acknowledgments

We thank Dr. Iain Moal for the extended Molecular Dynamics trajectories.

References

1. Alberts B. The cell as a collection of protein machines: preparing the next generation of molecular biologists. Cell. 1998 Feb 6;92(3):291-4.

2. Rual JF, Venkatesan K, Hao T, Hirozane-Kishikawa T, Dricot A, Li N, et al. Towards a proteome-scale map of the human protein-protein interaction network. Nature. 2005 Oct 20;437(7062):1173-8.

3. Stelzl U, Worm U, Lalowski M, Haenig C, Brembeck FH, Goehler H, et al. A human protein-protein interaction network: a resource for annotating the proteome. Cell. 2005 Sep 23;122(6):957-68.

4. Mosca R, Ceol A, Aloy P. Interactome3D: adding structural details to protein networks. Nat Methods. 2013 Jan;10(1):47-53.

5. Venkatesan K, Rual JF, Vazquez A, Stelzl U, Lemmens I, Hirozane-Kishikawa T, et al. An empirical framework for binary interactome mapping. Nat Methods. 2009 Jan;6(1):83-90.

6. Stumpf MP, Thorne T, de Silva E, Stewart R, An HJ, Lappe M, et al. Estimating the size of the human interactome. Proc Natl Acad Sci U S A. 2008 May 13;105(19):6959-64.

7. Szilagyi A, Zhang Y. Template-based structure modeling of protein-protein interactions. Curr Opin Struct Biol. 2014 Feb;24:10-23.

8. Kundrotas PJ, Zhu Z, Janin J, Vakser IA. Templates are available to model nearly all complexes of structurally characterized proteins. Proc Natl Acad Sci U S A. 2012 Jun 12;109(24):9438-41.

9. Aloy P, Ciccarelli FD, Leutwein C, Gavin AC, Superti-Furga G, Bork P, et al. A complex prediction: three-dimensional model of the yeast exosome. EMBO Rep. 2002 Jul;3(7):628-35.

10. Sinha R, Kundrotas PJ, Vakser IA. Docking by structural similarity at protein-protein interfaces. Proteins. 2010 Nov 15;78(15):3235-41.

11. Aloy P, Ceulemans H, Stark A, Russell RB. The relationship between sequence and interaction divergence in proteins. J Mol Biol. 2003 Oct 3;332(5):989-98.

12. Launay G, Simonson T. Homology modelling of protein-protein complexes: a simple method and its possibilities and limitations. BMC Bioinformatics. 2008;9:427.

13. Kundrotas PJ, Vakser IA. Global and local structural similarity in protein-protein complexes: implications for template-based docking. Proteins. 2013 Dec;81(12):2137-42.

14. Katchalski-Katzir E, Shariv I, Eisenstein M, Friesem AA, Aflalo C, Vakser IA. Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques. Proc Natl Acad Sci U S A. 1992 Mar 15;89(6):2195-9.

15. Gabb HA, Jackson RM, Sternberg MJ. Modelling protein docking using shape complementarity, electrostatics and biochemical information. J Mol Biol. 1997 Sep 12;272(1):106-20.

16. Kozakov D, Brenke R, Comeau SR, Vajda S. PIPER: an FFT-based protein docking program with pairwise potentials. Proteins. 2006 Nov 1;65(2):392-406.

17. Tovchigrechko A, Vakser IA. GRAMM-X public web server for protein-protein docking. Nucleic Acids Res. 2006 Jul 1;34(Web Server issue):W310-4.

18. Chen R, Li L, Weng Z. ZDOCK: an initial-stage protein-docking algorithm. Proteins. 2003 Jul 1;52(1):80-7.

19. Ritchie DW, Kemp GJ. Protein docking using spherical polar Fourier correlations. Proteins. 2000 May 1;39(2):178-94.

20. Garzon JI, Lopez-Blanco JR, Pons C, Kovacs J, Abagyan R, Fernandez-Recio J, et al. FRODOCK: a new approach for fast rotational protein-protein docking. Bioinformatics. 2009 Oct 1;25(19):2544-51.

21. Schneidman-Duhovny D, Inbar Y, Nussinov R, Wolfson HJ. PatchDock and SymmDock: servers for rigid and symmetric docking. Nucleic Acids Res. 2005 Jul 1;33(Web Server issue):W363-7.

22. Dominguez C, Boelens R, Bonvin AM. HADDOCK: a protein-protein docking approach based on biochemical or biophysical information. J Am Chem Soc. 2003 Feb 19;125(7):1731-7.

23. Lyskov S, Gray JJ. The RosettaDock server for local protein-protein docking. Nucleic Acids Res. 2008 Jul 1;36(Web Server issue):W233-8.

24. Fernandez-Recio J, Totrov M, Abagyan R. Soft protein-protein docking in internal coordinates. Protein Sci. 2002 Feb;11(2):280-91.

25. Cheng TM, Blundell TL, Fernandez-Recio J. pyDock: electrostatics and desolvation for effective scoring of rigid-body protein-protein docking. Proteins. 2007 Aug 1;68(2):503-15.

26. Pons C, Talavera D, de la Cruz X, Orozco M, Fernandez-Recio J. Scoring by intermolecular pairwise propensities of exposed residues (SIPPER): a new efficient potential for protein-protein docking. J Chem Inf Model. 2011 Feb 28;51(2):370-7.

27. Ravikant DV, Elber R. PIE-efficient filters and coarse grained potentials for unbound protein-protein docking. Proteins. 2010 Feb 1;78(2):400-19.

28. Lensink MF, Wodak SJ. Docking, scoring, and affinity prediction in CAPRI. Proteins. 2013 Dec;81(12):2082-95.

29. Pons C, Grosdidier S, Solernou A, Perez-Cano L, Fernandez-Recio J. Present and future challenges and limitations in protein-protein docking. Proteins. 2010 Jan;78(1):95-108.

30. Pallara C, Jimenez-Garcia B, Perez-Cano L, Romero-Durana M, Solernou A, Grosdidier S, et al. Expanding the frontiers of protein-protein modeling: from docking and scoring to binding affinity predictions and other challenges. Proteins. 2013 Dec;81(12):2192-200.

31. Fernandez-Recio J, Totrov M, Abagyan R. ICM-DISCO docking by global energy optimization with fully flexible side chains. Proteins. 2003 Jul 1;52(1):113-7.

32. Mashiach E, Nussinov R, Wolfson HJ. FiberDock: Flexible induced-fit backbone refinement in molecular docking. Proteins. 2010 May 1;78(6):1503-19.

33. Zacharias M. Protein-protein docking with a reduced protein model accounting for side chain flexibility. Protein Sci. 2003 Jun;12(6):1271-82.

34. Zacharias M. Rapid protein-ligand docking using soft modes from Molecular Dynamics simulations to account for protein deformability: binding of FK506 to FKBP. Proteins. 2004 Mar 1;54(4):759-67.

35. Moal IH, Bates PA. SwarmDock and the use of normal modes in protein-protein docking. Int J Mol Sci. 2010;11(10):3623-48.

36. Straub FB, Szabolcsi G. O dinamicseszkij aszpektah sztukturü fermentov (On the dynamic aspects of protein structure). In: Braunstein AE, editor. Molecular Biology, Problems and Perspectives. Moscow: Izdat. Nauka; 1964. p. 182-7.

37. Závodszky P, Abaturov LV, Varshavsky YM. Structure of glyceraldehyde-3-phosphate dehydrogenase and its alteration by coenzyme binding. . Acta Biochim Biophys Acad Sci Hung. 1966;1:389–403.

38. Foote J, Milstein C. Conformational isomerism and the diversity of antibodies. Proc Natl Acad Sci U S A. 1994 Oct 25;91(22):10370-4.

39. Tsai CJ, Ma B, Nussinov R. Folding and binding cascades: shifts in energy landscapes. Proc Natl Acad Sci U S A. 1999 Aug 31;96(18):9970-2.

40. Ma B, Kumar S, Tsai CJ, Nussinov R. Folding funnels and binding mechanisms. Protein Eng. 1999 Sep;12(9):713-20.

41. Csermely P, Palotai R, Nussinov R. Induced fit, conformational selection and independent dynamic segments: an extended view of binding events. Trends Biochem Sci. 2010 Oct;35(10):539-46.

42. Stein A, Rueda M, Panjkovich A, Orozco M, Aloy P. A systematic study of the energetics involved in structural changes upon association and connectivity in protein interaction networks. Structure. 2011 Jun 8;19(6):881-9.

43. Boehr DD, Nussinov R, Wright PE. The role of dynamic conformational ensembles in biomolecular recognition. Nat Chem Biol. 2009 Nov;5(11):789-96.

44. Fraser JS, Clarkson MW, Degnan SC, Erion R, Kern D, Alber T. Hidden alternative structures of proline isomerase essential for catalysis. Nature. 2009 Dec 3;462(7273):669-73.

45. Volkman BF, Lipson D, Wemmer DE, Kern D. Two-state allosteric behavior in a single-domain signaling protein. Science. 2001 Mar 23;291(5512):2429-33.

46. Moal IH, Bates PA. Kinetic rate constant prediction supports the conformational selection mechanism of protein binding. PLoS Comput Biol. 2012 Jan;8(1):e1002351.

47. Bastard K, Thureau A, Lavery R, Prevost C. Docking macromolecules with flexible segments. J Comput Chem. 2003 Nov 30;24(15):1910-20.

48. Bastard K, Prevost C, Zacharias M. Accounting for loop flexibility during protein-protein docking. Proteins. 2006 Mar 1;62(4):956-69.

49. van Dijk M, van Dijk AD, Hsu V, Boelens R, Bonvin AM. Information-driven protein-DNA docking using HADDOCK: it is a matter of flexibility. Nucleic Acids Res. 2006;34(11):3317-25.

50. Grunberg R, Leckner J, Nilges M. Complementarity of structure ensembles in protein-protein binding. Structure. 2004 Dec;12(12):2125-36.

51. Smith GR, Sternberg MJ, Bates PA. The relationship between the flexibility of proteins and their conformational states on forming protein-protein complexes with an application to protein-protein docking. J Mol Biol. 2005 Apr 15;347(5):1077-101.

52. Chaudhury S, Gray JJ. Conformer selection and induced fit in flexible backbone protein-protein docking using computational and NMR ensembles. J Mol Biol. 2008 Sep 12;381(4):1068-87.

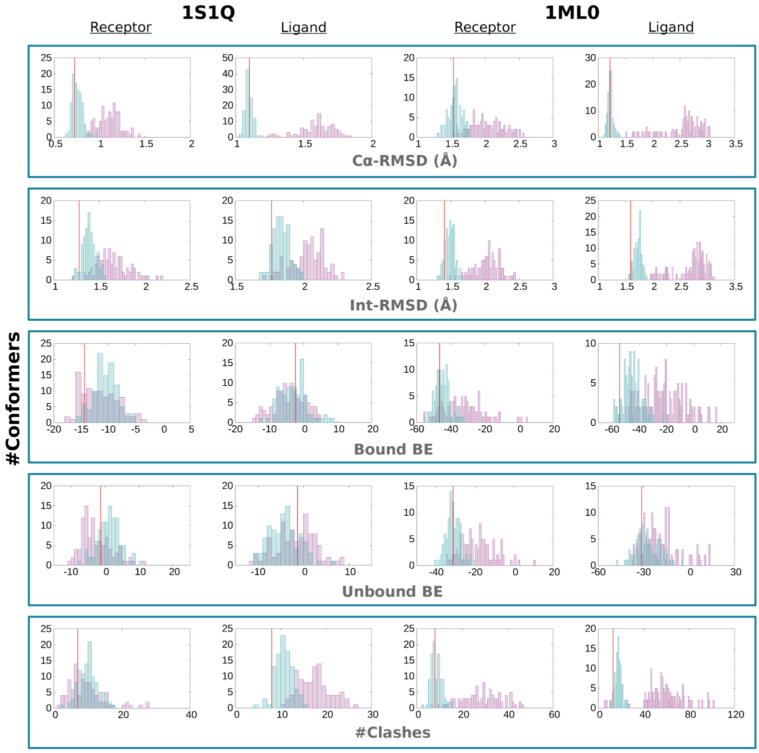
53. Pons C, Fenwick RB, Esteban-Martín S, Salvatella X, Fernandez-Recio J. Validated Conformational Ensembles Are Key for the Successful Prediction of Protein Complexes. Journal of Chemical Theory and Computation. 2013;9(3):1830-7.

54. Hwang H, Pierce B, Mintseris J, Janin J, Weng Z. Protein-protein docking benchmark version 3.0. Proteins. 2008 Nov 15;73(3):705-9.

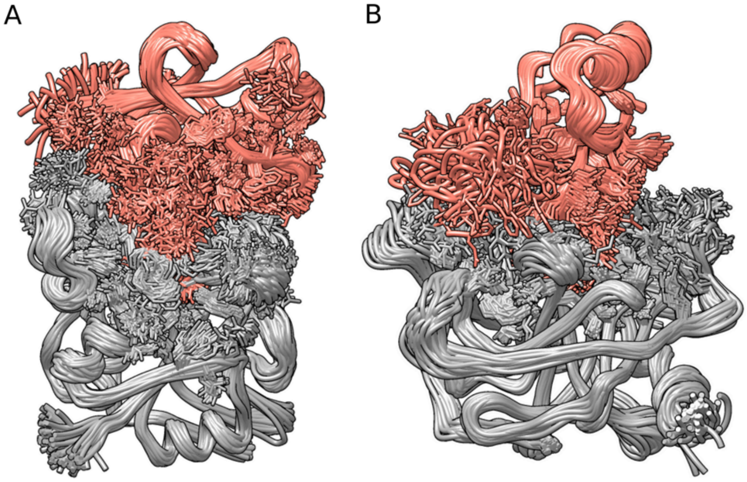
55. Lyman E, Zuckerman DM. On the structural convergence of biomolecular simulations by determination of the effective sample size. J Phys Chem B. 2007 Nov 8;111(44):12876-82.

56. Zhou HX. From induced fit to conformational selection: a continuum of binding mechanism controlled by the timescale of conformational transitions. Biophys J. 2010 Mar 17;98(6):L15-7.
57. Eswar N, Webb B, Marti-Renom MA, Madhusudhan MS, Eramian D, Shen MY, et al. Comparative protein structure modeling using Modeller. Curr Protoc Bioinformatics. 2006 Oct;Chapter 5:Unit 5 6.
58. Case DA, Cheatham TE, 3rd, Darden T, Gohlke H, Luo R, Merz KM, Jr., et al. The Amber biomolecular simulation programs. J Comput Chem. 2005 Dec;26(16):1668-88.
59. Meyer T, D'Abramo M, Hospital A, Rueda M, Ferrer-Costa C, Perez A, et al. MoDEL (Molecular Dynamics Extended Library): a database of atomistic Molecular Dynamics trajectories. Structure. 2010 Nov 10;18(11):1399-409.
60. Tirion MM. Large Amplitude Elastic Motions in Proteins from a Single-Parameter, Atomic Analysis. Phys Rev Lett. 1996 Aug 26;77(9):1905-8.
61. Hinsen K. Analysis of domain motions by approximate normal mode calculations. Proteins. 1998 Nov 15;33(3):417-29.
62. Atilgan AR, Durell SR, Jernigan RL, Demirel MC, Keskin O, Bahar I. Anisotropy of fluctuation dynamics of proteins with an elastic network model. Biophys J. 2001 Jan;80(1):505-15.
63. Doruker P, Atilgan AR, Bahar I. Dynamics of proteins predicted by Molecular Dynamics simulations and analytical approaches: application to alpha-amylase inhibitor. Proteins. 2000 Aug 15;40(3):512-24.
64. Kovacs JA, Chacon P, Abagyan R. Predictions of protein flexibility: first-order measures. Proteins. 2004 Sep 1;56(4):661-8.
65. Rueda M, Chacon P, Orozco M. Thorough validation of protein normal mode analysis: a comparative study with essential dynamics. Structure. 2007 May;15(5):565-75.
66. Sali A, Blundell TL. Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol. 1993 Dec 5;234(3):779-815.
67. Abagyan R, Lee WH, Raush E, Budagyan L, Totrov M, Sundstrom M, et al. Disseminating structural genomics data to the public: from a data dump to an animated story. Trends Biochem Sci. 2006 Feb;31(2):76-8.

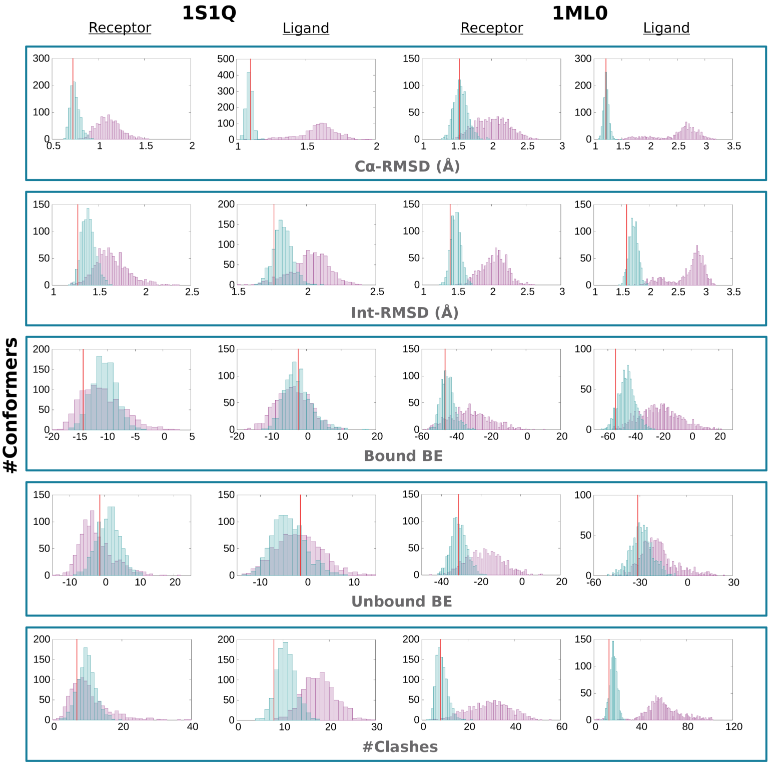
Supporting information



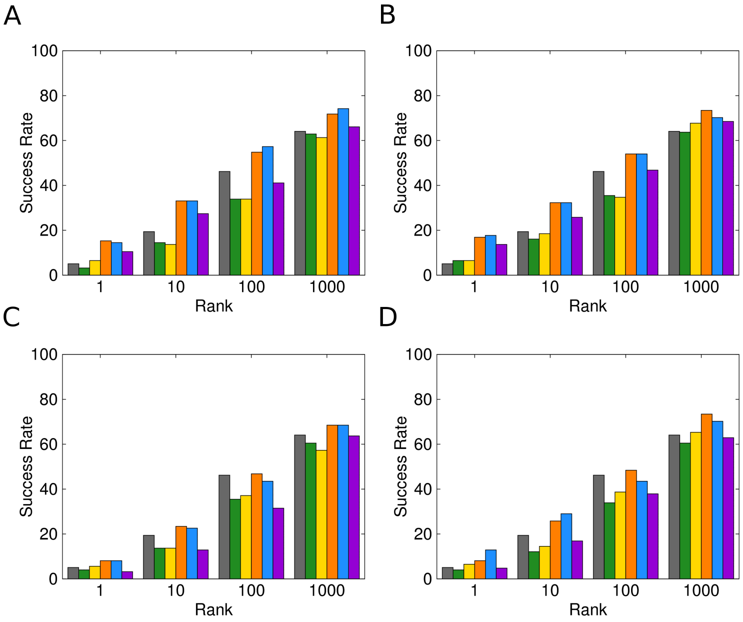
S1 Fig. Distribution of conformers according to different quality criteria in the 100-member ensembles. Distribution of conformers for different benchmark cases according to specific criteria based on the complex native orientation: Cα-RMSD, Int-RMSD, binding energy with the bound partner, binding energy with the unbound partner, number of clashes with respect to the bound partner (from top to bottom). Ensembles were generated by MODELLER (blue) and MD (magenta). Values for unbound X-ray structures are shown as red lines.



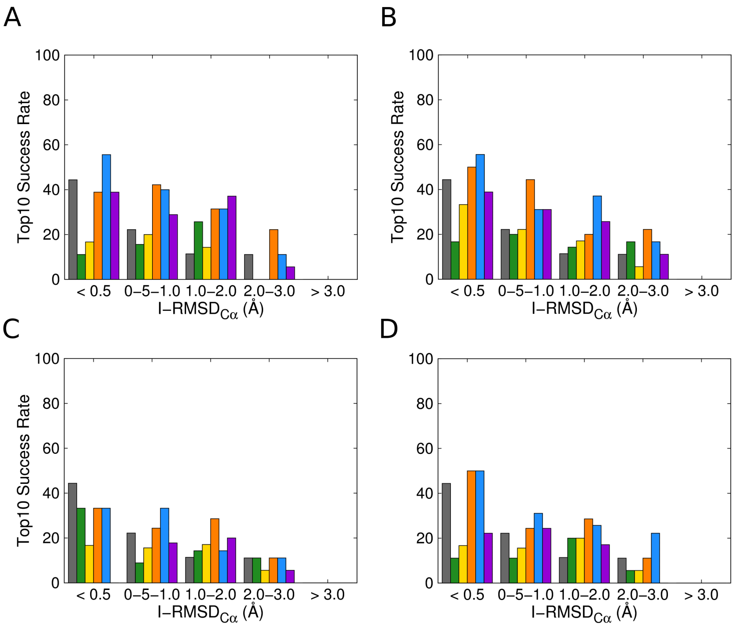
S2 Fig. Representative conformational ensembles generated by MD. 100 conformers independently generated by MD for receptor and ligand are shown for two benchmark cases: (A) 1PXV and (B) 1ACB. Conformers were superimposed onto the corresponding molecules in the reference complexes for visualization. Only interface side chains are shown for the sake of clarity.



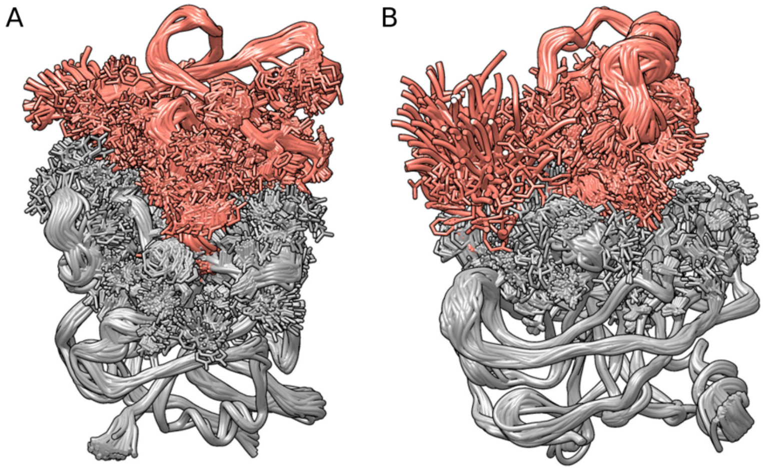
S3 Fig. Distribution of conformers according to different quality criteria in the 1000-member ensembles. Distribution of conformers for different benchmark cases according to specific criteria based on the complex native orientation: Cα-RMSD, Int-RMSD, binding energy with the bound partner, binding energy with the unbound partner, number of clashes with respect to the bound partner (from top to bottom). Ensembles were generated by MODELLER (blue) and MD (magenta). Values for unbound X-ray structures are shown as red lines.



S4 Fig. Docking performance for the best conformers of different ensembles. Docking success rates for the top 1 to 1000 predicted models on the protein-protein docking benchmark when using conformers selected by specific criteria based on the complex native orientation: Cα-RMSD (green), Int-RMSD (yellow), binding energy with the bound partner (orange), binding energy with the unbound partner (blue), number of clashes with respect to the bound partner (violet). Ensembles were composed of (A) 100 conformers generated by MODELLER, (B) 1000 conformers generated by MODELLER, (C) 100 conformers generated by MD, and (D) 1000 conformers generated by MD. For comparison, docking success rates for unbound X-ray structures are also shown (dark gray).



S5 Fig. Docking performance for the different ensembles according to unbound-to-bound variability. Docking success rates for the top 10 predicted models on the protein-protein docking benchmark with cases classified according to unbound-to-bound conformational variability, when using conformers selected by specific criteria based on the complex native orientation: Cα-RMSD (green), Int-RMSD (yellow), binding energy with the bound partner (orange), binding energy with the unbound partner (blue), number of clashes with respect to the bound partner (violet). Ensembles were composed of (A) 100 conformers generated by MODELLER, (B) 1000 conformers generated by MODELLER, (C) 100 conformers generated by MD, and (D) 1000 conformers generated by MD. For comparison, docking success rates for unbound X-ray structures are also shown (dark gray).



S6 Fig. Representative conformational ensembles generated by NMA-based sampling. 100 conformers independently generated by NMA-based sampling for receptor and ligand are shown for two benchmark cases: (A) 1PXV and (B) 1ACB. Conformers were superimposed onto the corresponding molecules in the reference complexes for visualization. Only interface side chains are shown for the sake of clarity.

S1 Table. Docking performance with conformers selected from extended MD ensembles (100ns trajectories, at 300K or 340K temperature).

	100ns-MD 300K (1000 confs)		*100ns-MD 340K (1000 confs)*	
	Cα-RMSD	*Int-RMSD*	*Cα-RMSD*	*Int-RMSD*
1AY7	19	1	4	155
1D6R	2181	2181	2018	2018
2HRK	162	161	314	1285
1GCQ	33	76	348	1025
1E6J	-	5	9	3
1ACB	48	2	99	434
1PXV	4118	182	3209	2027
2CFH	134	134	440	3647
1JMO	-	387	153	-
2C0L	1493	-	4121	974
1IRA	-	-	-	-

S2 Table. Docking performance of conformers selected from NMA-based ensembles. For each sampling method, docking results of all selected conformers are shown. For comparison, the docking results with the 1000-member ensembles generated by MODELLER are also shown, as well as those with the unbound and bound X-ray structures.

(A) MM1000

PDB	*Bound*	*Unbound*	*MM1000*				
			Cα-RMSD	*Int-RMSD*	*Bound BE*	*Bound BE*	*#Clashes*
1ACB	1	361	429	651	36	34	145
1ATN	7	2568	-	-	-	-	-
1EER	3	1821	371	132	13	48	24
1I2M	1	-	184	70	13	57	18
1IBR	1	-	681	-	-	-	1108
1PXV	1	2073	2400	1393	970	822	2332

*(B) NMA (100 Conf) *MM(1 conf)*

PDB	*Bound*	*Unbound*	*NMA (100 conf) * MM (1 conf)*				
			Cα-RMSD	*Int-RMSD*	*Bound BE*	*Bound BE*	*#Clashes*
1ACB	1	361	47	366	46	19	2614
1ATN	7	2568	2615	478	292	544	5818
1EER	3	1821	233	25	-	17	89
1I2M	1	-	5117	4566	23	310	506
1IBR	1	-	87	-	-	-	91
1PXV	1	2073	1712	-	315	168	1098

*(C) NMA (100 Conf) *MM(10 conf)*

PDB	*Bound*	*Unbound*	*NMA (100 conf) * MM (1 conf)*				
			Cα-RMSD	*Int-RMSD*	*Bound BE*	*Bound BE*	*#Clashes*
1ACB	1	361	122	-	9	3	-
1ATN	7	2568	3245	-	-	-	-
1EER	3	1821	-	-	3	16	-
1I2M	1	-	383	45	1	1	-
1IBR	1	-	326	-	-	-	146
1PXV	1	2073	687	-	168	2210	5868

*(D) NMA (1000 Conf) *MM(1 conf)*

PDB	*Bound*	*Unbound*	*NMA (1000 conf) * MM (1conf)*				
			Cα-RMSD	*Int-RMSD*	*Bound BE*	*Bound BE*	*#Clashes*
1ACB	1	361	759	589	11	1	229
1ATN	7	2568	1788	-	-	4332	-
1EER	3	1821	35	3	-	4	-
1I2M	1	-	2320	154	173	1	1344
1IBR	1	-	88	-	135	-	984
1PXV	1	2073	1717	1342	1183	232	866

3.2.2. Protein-protein ensemble docking at low-cost: improving predictive performance for medium-flexible cases

Chiara Pallara and Juan Fernández-Recio*

Joint BSC-CRG-IRB Research Program in Computational Biology, Life Sciences Department, Barcelona Supercomputing Center, Barcelona, Spain

*Corresponding author

Protein-protein ensemble docking at low-cost: improving predictive performance for medium-flexible cases

Chiara Pallara and Juan Fernández-Recio*

Joint BSC-CRG-IRB Research Program in Computational Biology, Life Sciences Department, Barcelona Supercomputing Center, Barcelona, Spain

*Corresponding author

Abstract

Protein-protein interactions are essential to understand cellular processes at molecular level. However, determining the atomic structure of the immense majority of protein-protein complexes is still highly challenging and constitutes one of the major goals of computational biology. Despite methodological advances in docking protocols, dealing with molecular flexibility is still a crucial bottleneck. Indeed, state-of-the-art rigid-body docking approaches, like pyDock, have difficulties in cases with large conformational changes upon binding. Although several protocols have been proposed to include flexibility as a refinement step, the approach of using precomputed conformational ensembles generated from unbound protein structures has been less explored. In the past we used ensembles derived from residue dipolar coupling (RDC) data to significantly improve docking predictions for ubiquitin complexes. More recently, we found that a simple molecular mechanics minimization method can generate conformers that improve energetic complementarity of the docking partners. Based on these studies, we have devised here a protocol to integrate unbound conformational ensembles within a rigid-body docking framework and systematically tested it on a data set of 124 protein-protein docking cases. For that, we docked every conformer from the receptor with a random one from the ligand, ranked all the resulting docking poses, and removed redundant solutions. Conformational ensembles generated here at low computational cost significantly improved docking predictions for cases in which interacting proteins showed moderate conformational changes upon binding. Future works on increasing the size and quality of these ensembles will expectedly extend the applicability of this docking strategy to more flexible cases.

Introduction

Proteins act as building blocks and functional components of a cell [1], and their interactions are crucially important for the virtual totality of biological processes. In this context, understanding the structural and functional details of the hundreds of thousands of protein-protein interactions that are formed in a living organism is essential to advance in biological knowledge and biomedical applications. However, current structural coverage of protein–protein interactions in human is below 4% of the estimated number of possible complexes [2-3].

Computational protein-protein docking aims to complement experimental efforts and provide structural models for protein complexes starting from the isolated component structures [4-5]. However, despite general methodological advances in docking, properly dealing with molecular flexibility is still a major bottleneck. Indeed, protein dynamics plays a key role in protein association and the need of integrating protein flexibility in docking simulations is now evident [4]. Several methodologies have been proposed to address this issue. The easier and simpler approaches consist in implicit treatment of flexibility by implementing soft potentials within FFT-based docking protocols [6], thus allowing a certain degree of inter-penetration between the interacting protein atoms. A more realistic and accurate description of the association process, although at higher computational cost, is the inclusion of conformational flexibility after a first rigid-body docking step, somehow mimicking the *induced-fit* binding model [7]. The majority of current docking methods that include flexibility follow this approach: ICM-DISCO [8], HADDOCK [9] or RosettaDock [10] protocols. A more recent strategy to include plasticity during the sampling step is by small deformation of the global structures along soft harmonic modes, such as in ATTRACT [11-12] or SwarmDock [13] programs.

An alternative approach, which has been largely unexplored, is to mimic the *conformational-selection* binding model [14-16] by docking separately a number of conformers selected from precomputed conformational ensembles of the docking partners. These conformational ensembles can be obtained experimentally (e.g., from NMR experiments) or computationally by conformational sampling methods (e.g., MD, NMA or homology modeling), ideally spanning various degrees of flexibility, from small local rearrangements to large-scale global motions. Although highly promising, to date this strategy has not been really used for practical docking predictions. Actually, very few systematic studies have been reported on the use of either conformational ensembles derived from theoretical simulation [17-19] or experimental data (i.e., NMR spectroscopy) [19]. Besides, the ensembles used in the majority of these studies do not really represent the population of the unbound state, as only few conformers were used in the docking procedure.

In this context, we recently showed that the use of ensembles generated with residual dipolar coupling (RDC) data in docking significantly improved the predictive rates in ubiquitin complexes [20]. However, the experimental limitations in the ensemble generation make it difficult to extend this protocol to large-scale application. More recently, we found that a simple molecular mechanics minimization approach using MODELLER (MM) can rapidly generate conformers with better binding properties, thus can improve docking predictions thanks to better energetic complementarity of the docking partners (Pallara et al., submitted).

Based on these findings, we have devised here a novel prococol for ensemble docking, which has been systematically tested on a large dataset of 124 protein-protein docking cases. For that, we docked every conformer from the receptor with another one randomly selected from the ligand. All the resulting docking poses were merged and clustered to remove redundant ones, and finally ranked according to an energy-based scoring function. The global docking predictions significantly improved the results with respect to the unbound docking, especially for medium-flexibility complexes.

Methods

Generation of protein conformational ensembles

MODELLER comparative protein structure modeling program [21] was used to obtain an ensemble of 100 conformations for the unbound interacting subunits of each docking case, using as template the unbound X-ray structure of the same protein. Ions and cofactors molecules, if any, were included during the modeling in order to preserve a reasonable accuracy in the structures obtained.

Ensemble docking

For each docking case, every conformer from the receptor ensemble was docked with another one randomly selected from the ligand ensemble, as described below. First, the Fast Fourier Trasform (FFT) program FTDock [22] was used to generate a set of 10,000 rigid-docking poses using surface complementary and electrostatics, at 0.7 Å grid resolution. All the models resulting from docking each pair of conformers from receptor and ligand ensembles were merged together, and sorted according to pyDock scoring scheme [23], based on desolvatation, electrostatics and Van der Waals energy contributions. Both ions and cofactors molecules were excluded during the sampling and the scoring calculations. Finally, the redundant predictions were eliminated by using a clustering algorithm starting from the top-ranked docking solution and removing all the following models with ligand RMSD lower than 10 Å.

Benchmark

In order to validate the protocol proposed here, we used the Weng's protein-protein docking benchmark version 3.0 [24], composed of 124 cases in which the structures of both the free components (unbound) and the complex (bound) are known.

For all the docking experiments, the predictive performance was evaluated by comparing the coordinates of each docking pose with the corresponding X-ray structure of the complex. A near-native solution (NNs) was defined as a docking pose with ligand RMSD lower than 10 Å (RMSD was calculates for the ligand c-alpha atoms with respect to the equivalent one in the X-ray structure of the complex after optimal superimposition between bound and unbound receptor molecules). The success rate is defined as the percentage of cases in which a near-native solution is found within the top N docking poses, as sorted by pyDock. For the evaluation and comparison of the docking results special attention was taken for the top 10 success rate. For the completeness of the analysis, additional docking simulations were performed using 100 random initial rotations of the unbound X-ray structures of the docking partners, followed by the same merging and clustering protocols as with the ensemble conformers.

As previously described in Pallara et al., submitted, all the 124 benchmark cases were classified according to the conformational variation of the proteins from the unbound to the bound state (based on the RMSD of Cα atoms of the interface residues as defined in the mentioned protein-protein benchmark 3.0), which resulted in the following categories: "rigid" (I-RMSD$_{C\alpha}$ < 0.5 Å), "low-flexible" (0.5 Å < I-RMSD$_{C\alpha}$ < 1.0 Å), "medium-flexible" (1.0 Å < I-RMSD$_{C\alpha}$ < 2.0 Å), "flexible" (2.0 Å < I-RMSD$_{C\alpha}$ < 3.0 Å), and "highly-flexible" (I-RMSD$_{C\alpha}$ > 3.0 Å).

Anchor residues were defined by the ANCHOR web server, (http://structure.pitt.edu/anchor/ [25-26]), as those with a predicted contribution to binding energy of more than 2.0 kcal/mol. The RMSD calculations were executed with ICM program (www.molsoft.com).

Results

Ensemble docking significantly improves complex structure predictions

We previously observed that conformational ensembles generated by a simple molecular mechanics approach contain specific conformers that can be highly useful for docking predictions. (Pallara et al., submitted) However, in a realistic situation it would be virtually impossible to identify which conformers are the best for docking, and thus it would be necessary to use all ensemble members within a docking protocol framework. The problem with this approach is that docking all 100 conformers from the receptor ensemble against all 100 conformers from the ligand ensemble would involve performing 10,000 individual docking runs for each case. This would be clearly impractical for the majority of cases and would need using high-performance computing facilities.

Here we have used an alternative strategy to dramatically reduce the computational costs, by docking each conformer of the receptor with a randomly selected conformer of the ligand, and thus running 100 docking jobs per case (see Methods). In this way, all conformers from receptor and ligand are used in docking, although obviously not all combinations of conformers are considered.

Fig 1A shows the docking success rates for the top 10 predictions on the overall benchmark when using this protocol, with all the details in **Table 1** and **Table S1**. Interestingly, ensemble docking (using 100 random receptor-ligand pairs of conformers) clearly improved the success rates (32.3%) with respect to the unbound subunits (19.4%). When only five conformers were used for the ensemble docking protocol, the prediction success rates (24.2%) significantly dropped with respect to those of the larger ensembles. For comparison, we also show the success rates that would be obtained when using the bound structures, which establish the upper limit for the expected docking results with this approach (61.3%). In order to provide a statistical significance for these results, we performed 100 different docking runs with random initial rotations of the unbound receptor and ligand molecule. The docking performance obtained (26.6% success rate) stood halfway between the ensemble docking and the unbound results.

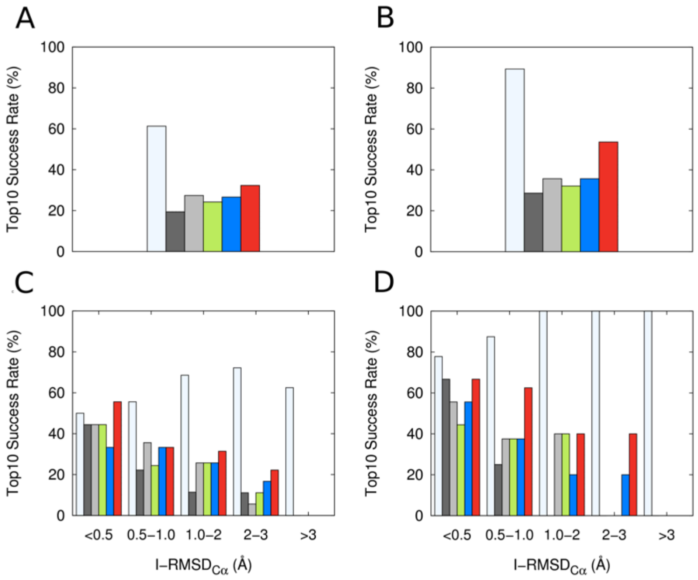


Fig 1. Ensemble docking predictive performance. (A) Predictive success rate obtained on the overall benchmark for the top 10 ranked docking poses when using the ensemble docking protocol described here (red). For comparison, the docking results for the bound (white) and unbound (dark gray) X-ray structures are shown. To assess the significance of the predictions, the figure also shows the docking results for random ensembles consisting on: five random initial rotations of the unbound docking partners (light gray), five random MM-derived conformers pairs (light green), and 100 random initial rotations of the unbound docking partners (blue). These random ensembles were used with the same protocol as the conformational ensembles. (B) Predictive success rates obtained on the 28 with experimental high binding affinity (ΔG < -12.0 kcal/mol) (same color code as above). (C) Predictive success rate obtained on the benchmark cases classified according to unbound-to-bound conformational motion (same color code as above). (D) Predictive success rate obtained on the high-affinity cases (ΔG < -12.0 kcal/mol) classified according to unbound-to-bound conformational motion (same color code as above).

These results suggest that conformational heterogeneity in the interacting subunits improves the binding capabilities of the unbound X-ray structures. Nevertheless, part of such improvement might be associated with the larger sampling in the atomic positions provided by the set of random initial rotations, which could somehow compensate the suboptimal grid-based sampling of FTDock as previously discussed (Pallara et al., submitted). Interestingly, lowering the number of random initial rotations (i.e., five rotations) did not substantially change the docking performance (27.4%), showing the importance of performing a small change in position of the starting molecules instead of using only the unbound X-ray structure.

The results of bound docking are far from optimal, probably due to the FFT-based discrete searching algorithm that makes it difficult to sample the exact native orientation. This would be particularly critical in low-affinity cases, in which the small number of interactions would make them less tolerant to small errors in the atomic positions. To minimize the impact of this limitation in our evaluation, we performed the same analysis as above but focusing only on the 28 cases of the benchmark that have been experimentally defined as high-affinity cases (with ΔG lower than -12 Kcal/mol). For these cases, the results of bound docking are much closer to optimal (89.3%). For these cases, we can observe more clearly that ensemble docking improved the success rates (53.6%) with respect to the unbound docking (28.6%). This improvement was clearly above that observed for the set of 100 random rotations (35.7%). (**Fig 1B**).

Table 1. Predictive performance of ensemble docking. Best rank of a near-native docking pose

Complex[a]	*Bound*	*Unbound*	*Ensemble docking*[b]
Rigid (I-RMSD$_{C\alpha}$ < 0.5 Å) (18 cases)			
1AVX	2	102	(29)
1FSK	3	3	1
1GHQ	7455	-	(107172)
1IQD	1	8	3
1KLU	18	1246	(59167)
1KTZ	48	3725	(7981)
1NCA	14	7	1
1NSN	405	500	(303)
1PPE	28	6	4
1R0R	1	3	(222)
1SBB	161	298	(2491)
1WEJ	1	274	3
2JEL	1	42	2
2MTA	2	78	(105)

2PCC	12	6	3
2SIC	1	8	2
2SNI	1	3	3
2UUY	69	4472	(6203)
Low-flexible (I-RMSD$_{C\alpha}$ 0.5-1.0 Å) (45 cases)			
1AHW	1043	4049	(430)
1AY7	1	24	4
1AZS	1	30	(363)
1BJ1	9	-	(388)
1BUH	71	66	(1004)
1BVN	1	2	1
1DQJ	216	604	(19)
1E96	113	1	(8)
1EAW	8	622	8
1EFN	6	166	(1014)
1EWY	4	8	(39)
1F34	1	139	(1713)
1F51	2	7	(12)
1FQJ	14	309	(1390)
1GCQ	274	1091	(3479)
1GLA	61	50	(150)
1GPW	1	1	1
1HE1	1	3958	(672)
1HE8	138	2917	(9893)
1IJK	16	1309	(1122)
1J2J	46	19	-
1JPS	709	481	(248)
1K4C	-	-	(6742)
1K74	150	14	3
1KAC	4737	1286	(3610)
1KXQ	1	250	1
1MAH	1	19	3
1MLC	2	37	2
1N8O	3	53	-
1QA9	3253	7378	(18552)
1QFW	81	239	(3338)
1RLB	1319	4094	(2803)
1S1Q	147	1211	(1465)
1T6B	3	56	(713)
1TMQ	1	1	7
1UDI	1	1	1
1YVB	1	19	5
1Z0K	2	8	(105)
1ZHI	5	3	2
2AJF	5	1788	(44037)
2B42	1	1	2
2BTF	1	33	7
2OOB	588	112	(3994)
2VIS	64	-	-
7CEI	1	19	5

Medium-flexible (I-RMSD$_{C\alpha}$ 1.0-2.0 Å) (35 cases)			
1A2K	36	114	(886)
1AK4	2420	2040	(5697)
1AKJ	89	656	(635)
1B6C	1	3	3
1BGX	1	-	(108647)
1BVK	7	18	(17)
1D6R	1050	2128	(2088)
1DFJ	6	557	1
1E6E	1	3	1
1E6J	-	33	6
1EZU	1	2048	(2107)
1FC2	127	-	(2869)
1GP2	1	-	(2278)
1GRN	2	858	(364)
1HIA	99	40	4
1I4D	1	-	(184)
1I9R	15	846	(871)
1K5D	1	360	(194)
1KXP	1	16	1
1ML0	1	173	10
1NW9	1	9	(140)
1OPH	59	14	(4056)
1VFB	37	59	(168)
1WQ1	4	2448	(76)
1XD3	1	1	3
1XQS	1	14	(49)
1Z5Y	1	16	(47)
2CFH	1	1904	(34)
2FD6	68	31	(367)
2H7V	1	-	(1828)
2HLE	1	13	1
2HQS	1	30	(13)
2I25	1	40	(774)
2O8V	1	60	2
2QFW	1	-	1
Flexible (I-RMSD$_{C\alpha}$ 2.0-3.0 Å) (18 cases)			
1ACB	1	361	1
1BKD	2	522	(915)
1CGI	1	19	1
1DE4	1	-	(189)
1E4K	104	1215	(7940)
1EER	3	1821	(177)
1I2M	1	-	1
1IB1	34	-	8447
1IBR	1	-	(5453)
1KKL	88	49	(15)
1M10	1	81	(14)
1N2C	1	-	(49)
1PXV	1	2073	(5089)

2C0L	83	3958	(34623)
2HMI	2	-	-
2HRK	49	16	(76)
2NZ8	1	10	(50)
2OT3	1	5	2
Highly-flexible (I-RMSD$_{C\alpha}$ > 3.0 Å) (8 cases)			
1ATN	7	2568	(1071)
1FAK	41	5327	(1252)
1FQ1	6	3865	(1349)
1H1V	537	-	(70803)
1IRA	1	-	-
1JMO	1	5325	-
1R8S	1	-	(1572)
1Y64	1420	-	-

[a] PDB of the complex; [b] In brackets, rank before clustering
(in bold: high affinity cases)

Low- and medium-flexible cases are the ones most benefited by precomputed ensembles

We expected that the docking improvement when using conformational sampling would depend on the flexibility of the interacting proteins upon binding. Therefore, we explored for which level of molecular flexibility our ensembles could be more beneficial. We classified the docking cases according to the conformational movement of the proteins from the unbound to the bound state (based on interface RMSD, I-RMSD$_{C\alpha}$). As show in **Fig 1C**, we found that the largest improvement when using the ensemble docking occurred in the medium-flexible cases, i.e. those with I-RMSD$_{C\alpha}$ between 1.0 and 2.0 Å. In the rigid cases (I-RMSD$_{C\alpha}$ lower than 0.5 Å), the ensemble docking results were as good as when using the unbound structures (close to optimal), whereas for the most flexible cases (I-RMSD$_{C\alpha}$ higher than 3.0 Å) the ensemble docking results were as poor as those for the unbound structures (very far from optimal). The improvement for the low-, medium- and flexible cases is even more evident when the analysis is focused on the high-affinity cases (**Fig 1D**).

Ensemble size in docking: the higher the better

We explored the question of which is the minimal number of conformers that would be needed in order to observe a significant improvement in the docking results. To avoid the complexity of the interpretation of the results in the weak affinity cases (see above), we focused our analysis on the high-affinity ones. **Fig 2** shows the ensemble docking results after randomly selecting a different number of receptor-ligand pairs of conformers (each sub-set is repeated five times and the results are averaged). We found that the docking performance increases linearly with the size of the ensemble (in grey). Interestingly, when only the high-affinity, low-flexible cases are considered (in red), the docking success improves dramatically with just a few conformer pairs, so that 30 conformers provide similar same success rates as 100 conformers. All these data suggest that the results might be further improved by using a higher number of conformer pairs, i.e. increasing the number of receptor-ligand conformer combinations, but the computational cost would be impractical and beyond this work.

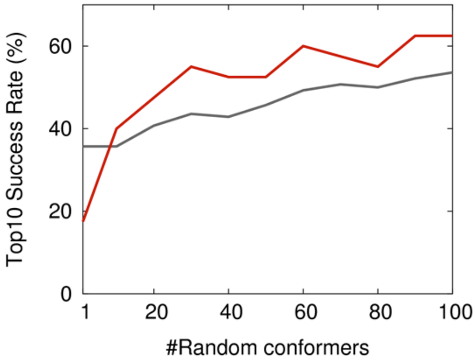


Fig 2. Ensemble docking success rates as a function of the ensemble size. Ensemble docking performance for the top 10 predictions when different number of randomly chosen conformer pairs are considered. Results are shown for the 28 high-affinity benchmark cases (in grey) and the 8 high-affinity and low-flexible cases (in red**).**

Discussion

Ensemble docking provides more near-native poses and as a consequence better predictive rates

As emerged from the previous analysis, a minimal structural heterogeneity provided by the ensembles generated by MODELLER minimization (MM) improves docking results with respect to the unbound X-ray structures. We have further studied the reasons of such improvement. We first explored for each receptor-ligand pair of conformers whether the docking energy of the best near-native solution (determinant for the docking success) depended on the number of near-native solutions obtained for such conformer pair.

As can be seen in **Fig S1**, for the majority of cases it can be observed that the higher the number of near-native solutions generated by a given conformer pair, the higher the probability of obtaining good docking energies by such near-native solutions. In this line, the conformers that generated more near-native solutions than the unbound structure provided in general better near-native docking energies than those generated by the unbound structure. We also observed that the bound X-ray structure typically yields more near-native solutions and with better docking energy than the unbound. Interestingly, for many cases (e.g., 1NSN, 1DFJ, 1I2M), there are a few conformers that generated even more near-native solutions than the bound structure. This is consistent with the previously observed correlation between the number of near-native solutions generated by docking and the predictive success rates [27]. Therefore, increasing the ratio between near-native solutions and false positives is the main reason for the beneficial effect of some of the conformers found in the precomputed unbound ensembles.

For each case, the percentage of conformers that produced more near-native solutions than the unbound structure is also a determinant of the ensemble docking success. Cases with more than 70% of the conformers producing more near-native solutions than the unbound structure show much higher success rate (72.7%) than the unbound docking (36.4%), almost reaching the optimal bound docking results. On the contrary, in those benchmark cases in which there were less than 70% of conformers that produced more near-native solutions, ensemble docking had similar success rate (41.2%) to when using unbound X-ray structures (35.3%) and far from the bound docking results.

Successful conformers are not necessarily more similar to bound state

As we have just seen, consideration of conformational heterogeneity in docking can increase the number of near-native solutions generated by FTDock, as well as their docking energy, which is a key determinant for the docking success of each conformer pair. However, neither the number of near-native solutions found for each conformer pair or their best binding energy (and as a consequence the success rate) depended on the similarity of such conformer pair to the bound state (**Fig S2** and **S3**). This is in agreement with previous findings (Pallara et al., submitted) and is consistent with an extended conformational selection mechanism [28].

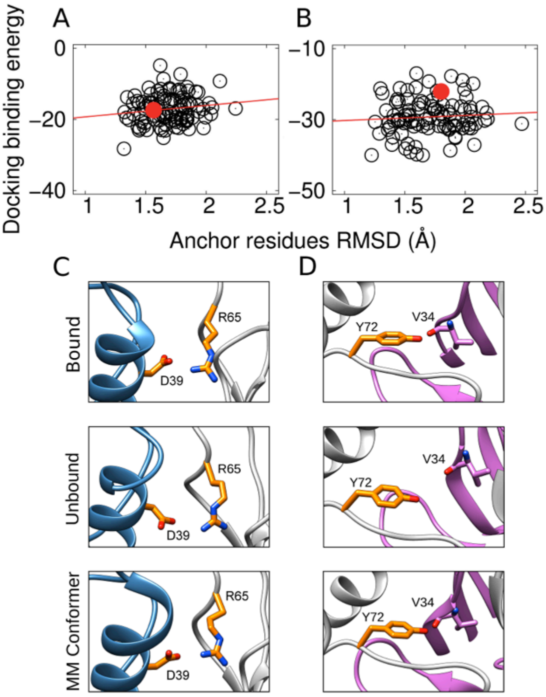


Fig 3. Ensemble docking performance as dependent on the structural similarity of anchor residues to the bound structure. Binding energy of the best ranked near-native solution for each conformer pair versus the full-atom RMSD of predicted anchor residues in such conformers with respect to those in the bound state, for (A) 1AY7 and (B) 1MAH benchmark cases. For comparison, unbound X-ray structure is shown in red. (C, D) Selected native key interface contacting pairs that are lost in the unbound docking near-native solutions and found in the ensemble docking near-native models: $R65_A$ with $D39_B$ from 1AY7; $Y72_A$ with $V34_B$ from 1MAH. Anchor residues are shown in orange.

In a few cases (e.g., 1AY7, 1MAH) we could observe that the most successful pair of conformers were the most similar to the bound state in terms of the RMSD of the predicted anchor residues [25-26] (**Fig 3A** and **3B**). In these cases, unbound ensembles can explore bound-like orientations of specific interface key residues that can improve the interacting capability of such conformers upon the docking and thus yield better docking results with respect to the unbound. **Fig 3C** and **3D** show two examples of such cases in which ensembles generated successful conformers that would reproduce key interface contacting pairs in the native orientation and that could not be obtained using the unbound docking partners. Additional key interface contacting pair involving anchor residues are shown in **Fig S4**.

Conformer pairs providing better binding energy in the native orientation are more likely to improve docking results

We reported that the structural similarity of the docking partners to the native conformation is not determinant for the docking success in general. However, what we found is that the better the binding energy of a conformer pair in the native orientation (i.e, after optimal superimposition on complex structure), the better the docking energy of the produced near-native solutions (and therefore the success rates) (**Fig S5**). This is in agreement with our previous findings (Pallara et al., submitted). Thus, capacity to provide favorable binding energy in the native orientation seems to be a major determinant for the success of a given conformer pair.

In this regard, **Fig 4A** shows that for a given case the predictive success (i.e., best ranked near-native solution) of unbound docking strongly depends on the expected optimal binding energy of the unbound subunits as calculated in the native orientation. All successful docking cases (i.e., best near-native rank ≤ 10) had optimal binding energy of the unbound subunits in the native orientation < 0.0 a.u. A number of unsuccessful cases had also optimal unbound binding energy < 0.0 a.u., but the majority of them (70%) significantly improved in the ensemble docking (highlighted as black circles). Only two cases out of 7 having a pair binding energy < -20.0 a.u. were unsuccessful in docking (1DFJ and 1MAH). Interestingly, the cases with pair binding energy between -20.0 and 0 a.u. seem the ones more benefited by the ensemble docking, since 62% of the successful cases had an optimal binding energy within such range. On the contrary, for cases with worse unbound optimal binding energy (> 0.0 a.u.), the docking results for the ensemble were as poor as those when using the unbound X-ray structures.

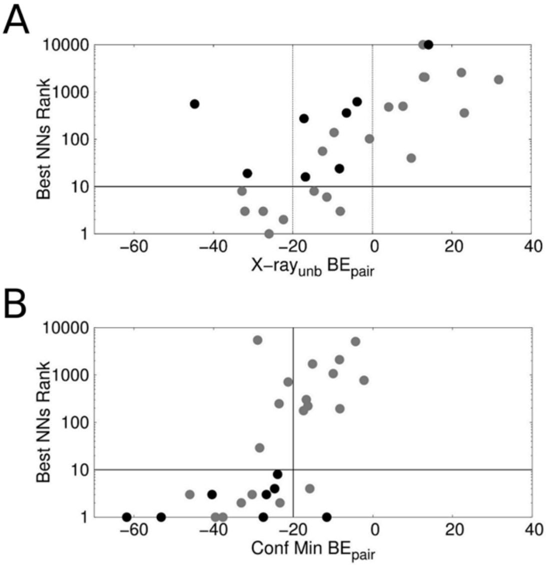


Fig 4. Dependence of docking performance on the energetic complementary of the docking partners. Distribution of the (A) best rank of any near-native solution for unbound docking versus the binding energy of the unbound partners in the native orientation; (B) best rank of any near-native solution from all conformer pairs vs. the optimal binding energy of the best pair of conformers in the native orientation. Cases that significantly improve docking performance after ensemble docking are highlighted as black circles.

Fig 4B shows the docking success for each case as compared with the best binding energy of all docked pairs of conformers in the native orientation. After ensemble docking, 87% of the successful cases had optimal conformer pair binding energy < -20.0 a.u.. The majority of cases with optimal conformer pair binding energy > -20 a.u. were unsuccessful after ensemble docking. All this confirms that the existence of conformers capable of providing favorable binding energy in the native orientation is a major determinant for the success of the ensemble docking.

Conclusions

We reported here the first systematic study about the unbiased use of precomputed unbound ensemble in docking. A novel docking strategy was devised consisting on the use of conformational ensembles of the interacting subunits derived from molecular mechanics minimization. Randomly selected pairs of receptor and ligand conformers were docked with an FFT-based method, and all the non-redundant resulting docking poses were scored by an energy-based function. The results showed improved predictive rates as compared with those of the unbound structures, especially in those cases with medium-flexible conformational changes between unbound and bound states. Docking success is not linked to an improvement in the structural similarity of the conformers with respect to the bound state, but rather to the better binding energy capabilities of the conformers in the native orientation. We have shown here that a minimal conformational heterogeneity can be used in a practical docking protocol to improve the results of unbound docking. This has the potential of further improving the predictive results by extending conformational sampling and/or considering larger ensembles, although this would involve an enormous computational cost. In this line, much more efficient algorithms to use larger ensembles in practical docking protocols will be needed.

References

1. Alberts B. The cell as a collection of protein machines: preparing the next generation of molecular biologists. Cell. 1998 Feb 6;92(3):291-4.
2. Venkatesan K, Rual JF, Vazquez A, Stelzl U, Lemmens I, Hirozane-Kishikawa T, et al. An empirical framework for binary interactome mapping. Nat Methods. 2009 Jan;6(1):83-90.
3. Stumpf MP, Thorne T, de Silva E, Stewart R, An HJ, Lappe M, et al. Estimating the size of the human interactome. Proc Natl Acad Sci U S A. 2008 May 13;105(19):6959-64.
4. Lensink MF, Wodak SJ. Docking, scoring, and affinity prediction in CAPRI. Proteins. 2013 Dec;81(12):2082-95.
5. Ritchie DW. Recent progress and future directions in protein-protein docking. Curr Protein Pept Sci. 2008 Feb;9(1):1-15.
6. Palma PN, Krippahl L, Wampler JE, Moura JJ. BiGGER: a new (soft) docking algorithm for predicting protein interactions. Proteins. 2000 Jun 1;39(4):372-84.
7. Koshland DE. Application of a Theory of Enzyme Specificity to Protein Synthesis. Proceedings of the National Academy of Sciences of the United States of America. 1958;44(2):98-104.
8. Fernandez-Recio J, Totrov M, Abagyan R. Soft protein-protein docking in internal coordinates. Protein Sci. 2002 Feb;11(2):280-91.
9. Dominguez C, Boelens R, Bonvin AM. HADDOCK: a protein-protein docking approach based on biochemical or biophysical information. J Am Chem Soc. 2003 Feb 19;125(7):1731-7.
10. Lyskov S, Gray JJ. The RosettaDock server for local protein-protein docking. Nucleic Acids Res. 2008 Jul 1;36(Web Server issue):W233-8.
11. Zacharias M. Protein-protein docking with a reduced protein model accounting for side chain flexibility. Protein Sci. 2003 Jun;12(6):1271-82.

12. Zacharias M. Rapid protein-ligand docking using soft modes from Molecular Dynamics simulations to account for protein deformability: binding of FK506 to FKBP. Proteins. 2004 Mar 1;54(4):759-67.

13. Moal IH, Bates PA. SwarmDock and the use of normal modes in protein-protein docking. Int J Mol Sci. 2010;11(10):3623-48.

14. Kumar S, Ma B, Tsai CJ, Sinha N, Nussinov R. Folding and binding cascades: dynamic landscapes and population shifts. Protein Sci. 2000 Jan;9(1):10-9.

15. Foote J, Milstein C. Conformational isomerism and the diversity of antibodies. Proc Natl Acad Sci U S A. 1994 Oct 25;91(22):10370-4.

16. Tsai CJ, Ma B, Nussinov R. Folding and binding cascades: shifts in energy landscapes. Proc Natl Acad Sci U S A. 1999 Aug 31;96(18):9970-2.

17. Grunberg R, Leckner J, Nilges M. Complementarity of structure ensembles in protein-protein binding. Structure. 2004 Dec;12(12):2125-36.

18. Smith GR, Sternberg MJ, Bates PA. The relationship between the flexibility of proteins and their conformational states on forming protein-protein complexes with an application to protein-protein docking. J Mol Biol. 2005 Apr 15;347(5):1077-101.

19. Chaudhury S, Gray JJ. Conformer selection and induced fit in flexible backbone protein-protein docking using computational and NMR ensembles. J Mol Biol. 2008 Sep 12;381(4):1068-87.

20. Pons C, Fenwick RB, Esteban-Martín S, Salvatella X, Fernandez-Recio J. Validated Conformational Ensembles Are Key for the Successful Prediction of Protein Complexes. Journal of Chemical Theory and Computation. [doi: 10.1021/ct300990h]. 2013 2013/03/12;9(3):1830-7.

21. Eswar N, Webb B, Marti-Renom MA, Madhusudhan MS, Eramian D, Shen MY, et al. Comparative protein structure modeling using Modeller. Curr Protoc Bioinformatics. 2006 Oct;Chapter 5:Unit 5 6.

22. Gabb HA, Jackson RM, Sternberg MJ. Modelling protein docking using shape complementarity, electrostatics and biochemical information. J Mol Biol. 1997 Sep 12;272(1):106-20.

23. Cheng TM, Blundell TL, Fernandez-Recio J. pyDock: electrostatics and desolvation for effective scoring of rigid-body protein-protein docking. Proteins. 2007 Aug 1;68(2):503-15.

24. Hwang H, Pierce B, Mintseris J, Janin J, Weng Z. Protein-protein docking benchmark version 3.0. Proteins. 2008 Nov 15;73(3):705-9.

25. Meireles LM, Domling AS, Camacho CJ. ANCHOR: a web server and database for analysis of protein-protein interaction binding pockets for drug discovery. Nucleic Acids Res. 2010 Jul;38(Web Server issue):W407-11.

26. Rajamani D, Thiel S, Vajda S, Camacho CJ. Anchor residues in protein-protein interactions. Proc Natl Acad Sci U S A. 2004 Aug 3;101(31):11287-92.

27. Pons C, Grosdidier S, Solernou A, Perez-Cano L, Fernandez-Recio J. Present and future challenges and limitations in protein-protein docking. Proteins. 2010 Jan;78(1):95-108.

28. Csermely P, Palotai R, Nussinov R. Induced fit, conformational selection and independent dynamic segments: an extended view of binding events. Trends Biochem Sci. 2010 Oct;35(10):539-46.

Supporting information

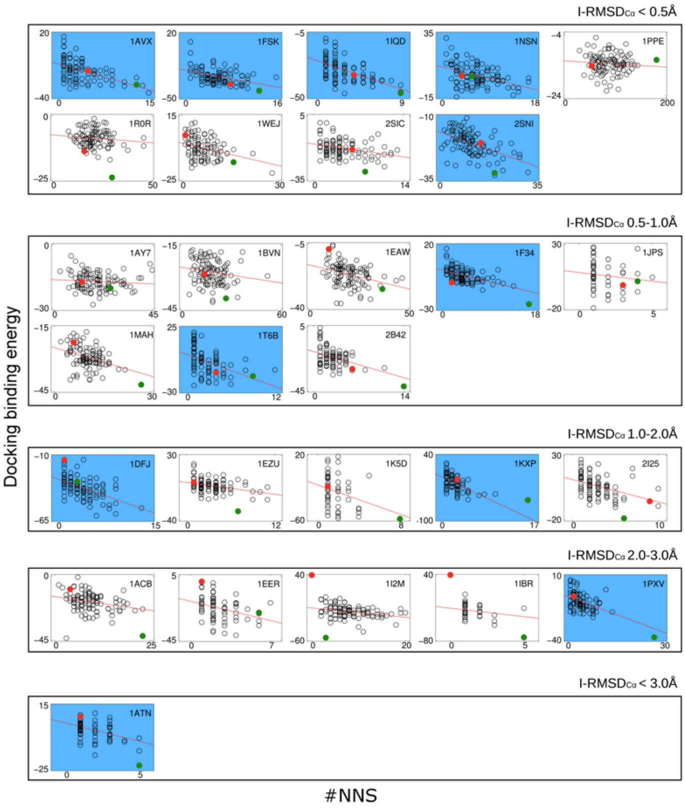


Fig S1. Relation between docking performance and sampling efficiency. Distribution of the best near-native solution binding energy for each receptor-ligand conformer pair versus the number of near-native solutions generated by FTDock for such conformer pair. Only high-affinity cases are shown, and they are classified according to conformational motion between unbound and bound states. For comparison, the results for the docking of unbound and bound X-ray structures are shown in red and green, respectively. Cases with correlation coefficient < -0.4 are shown in blue background.

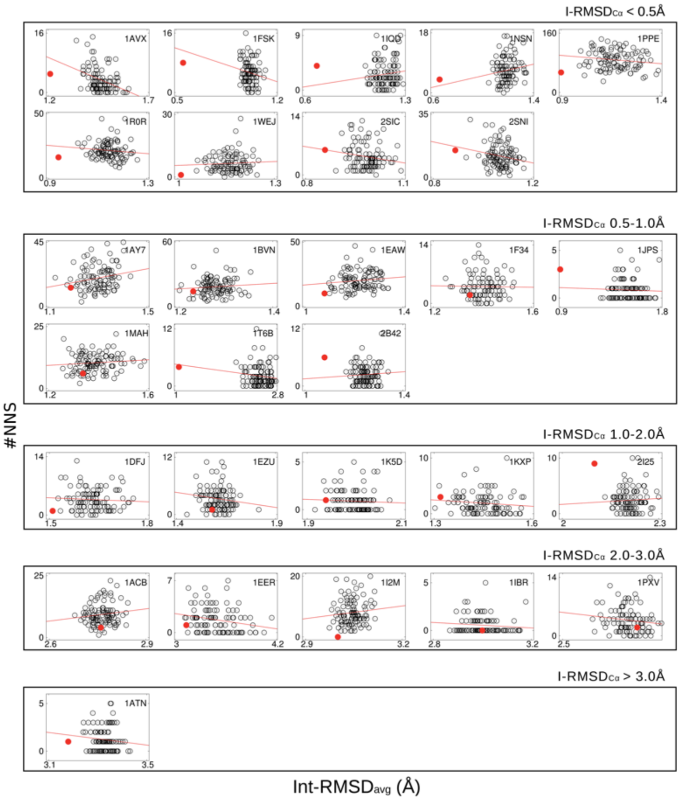


Fig S2. Relation between sampling efficiency and structural similarity to the bound. Distribution of the number of near-native solutions generated by FTDock for each conformer pair versus the average full-atom interface RMSD for receptor and ligand with respect to the bound structures. Only high-affinity cases are shown, and they are classified according to conformational motion between unbound and bound states. For comparison, the docking results for the unbound X-ray structures are shown in red.

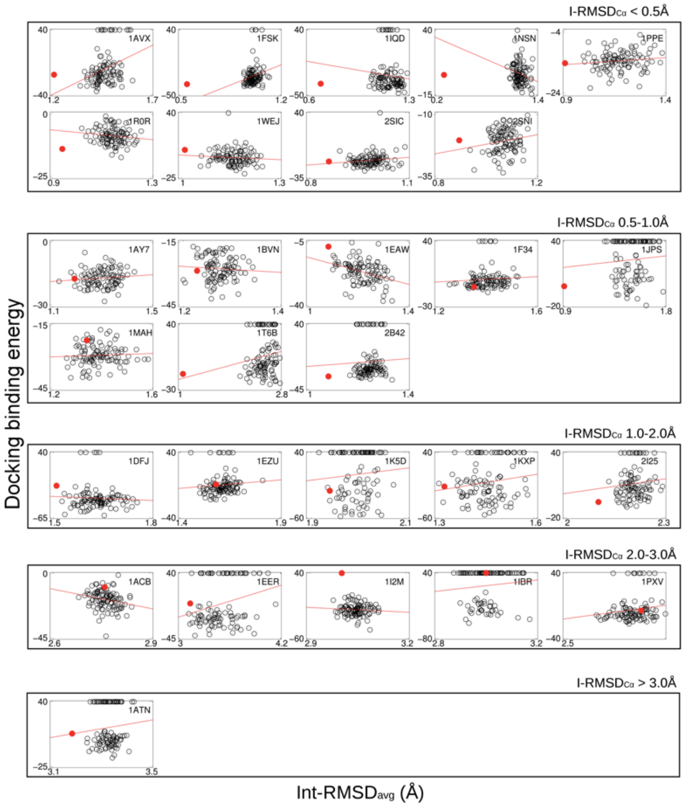


Fig S3. Relation between docking performance and structural similarity to the bound. Distribution of the best near-native solution binding energy for each conformer pair versus the average full-atom interface RMSD for receptor and ligand with respect to the bound structures. Only high-affinity cases are shown, and they are classified according to conformational motion between unbound and bound states. For comparison, docking results for the unbound X-ray structures are shown in red.

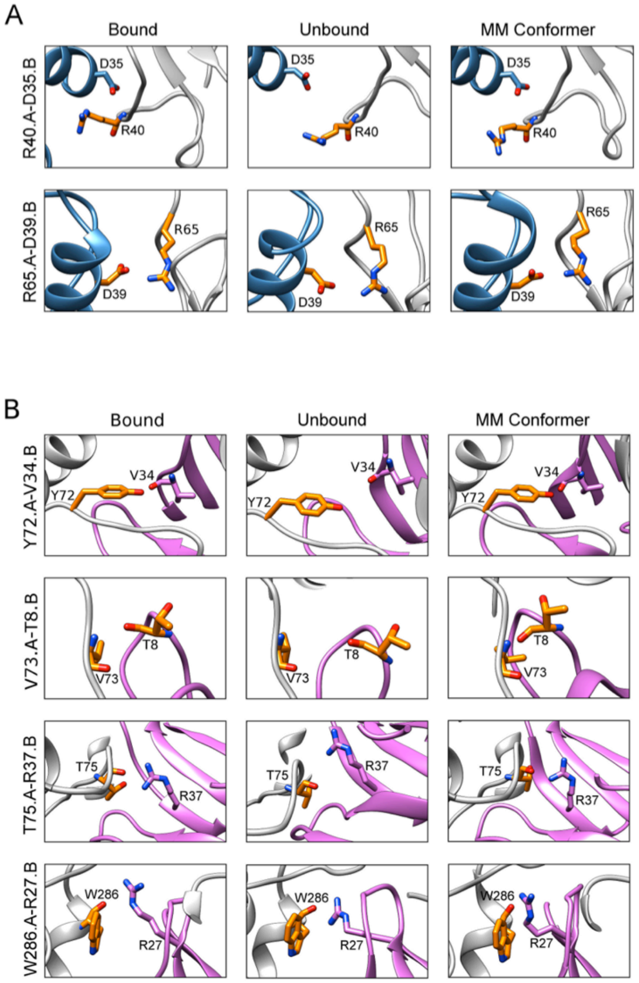


Fig S4. Native key interface contacting pairs improved in the ensemble docking. (A) 1AY7 and (B) 1MAH benchmark cases. Anchor residues are in orange.

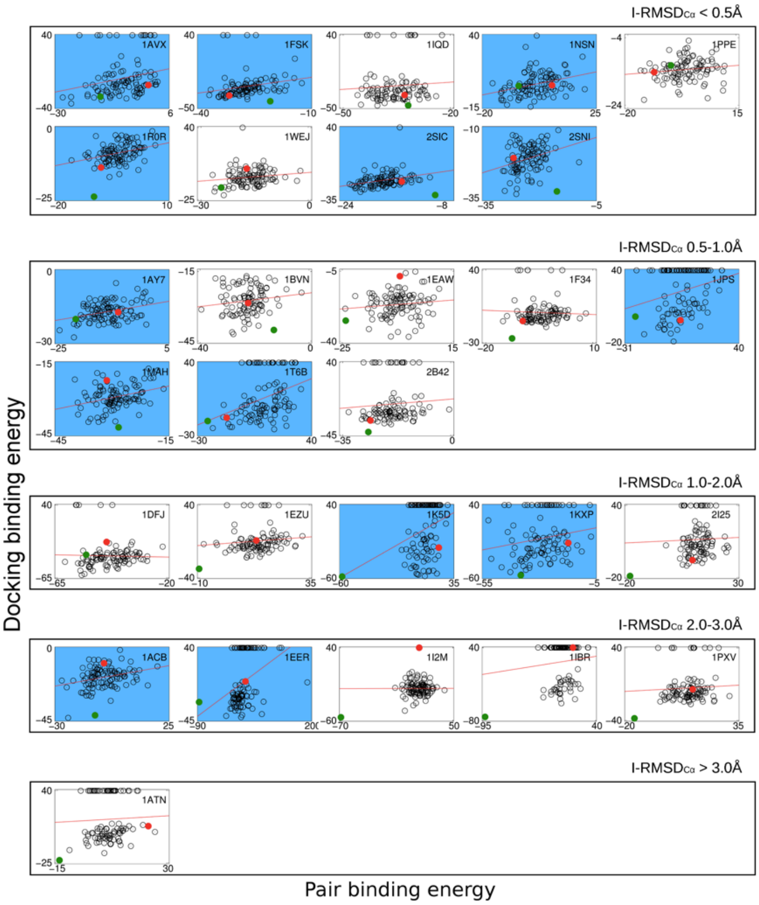


Fig S5. Relation between docking performance and energetic complementary of the docking partners. Distribution of the best near-native solution binding energy for each conformer pair versus the binding energy of the conformers in the complex native orientation. Only high-affinity cases are shown, and they are classified according to conformational motion between unbound and bound states. For comparison, the results for the docking of unbound and bound X-ray structures are shown in red and green, respectively. Cases with correlation coefficient < -0.2 are shown in blue background.

Table S1. Docking with random ensembles: Best rank of a near-native docking pose

Complex[a]	*Random Conf. (5)*	*Random Unb. Rot. (5)*	*Random Unb. Rot. (100)*
Rigid (I-RMSD$_{C\alpha}$ < 0.5 Å) (18 cases)			
1AVX	**(64)**	**(211)**	**(976)**
1FSK	**3**	**3**	**4**
1GHQ	(8001)	(3290)	(23384)
1IQD	**(7)**	**4**	**2**
1KLU	(3807)	(19418)	(98251)
1KTZ	(1562)	(795)	(9270)
1NCA	10	(50)	(420)
1NSN	**(65)**	**(580)**	**(6396)**
1PPE	**(8)**	**2**	**(21)**
1R0R	**(91)**	**(360)**	**(92)**
1SBB	(512)	(1205)	(5424)
1WEJ	**2**	**(29)**	**9**
2JEL	1	3	(41)
2MTA	5	2	(23)
2PCC	1	5	9
2SIC	**6**	**6**	**10**
2SNI	**5**	**4**	**4**
2UUY	(3166)	(892)	(13165)
Low-flexible (I-RMSD$_{C\alpha}$ 0.5-1.0 Å) (45 cases)			
1AHW	(13454)	(9777)	(360248)
1AY7	**(46)**	**(111)**	**(784)**
1AZS	(-)	8	(254)
1BJ1	(364)	(293)	(3180)
1BUH	(266)	(189)	(2301)
1BVN	**1**	**1**	**1**
1DQJ	(1118)	(328)	(1037)
1E96	7	5	(16)
1EAW	**1**	**(197)**	**(250)**
1EFN	(1001)	(490)	(6233)
1EWY	1	5	6
1F34	**(3200)**	**(670)**	**(4025)**
1F51	(43)	3	9
1FQJ	(668)	(1576)	(6501)
1GCQ	(48)	(62)	(2788)
1GLA	(38)	1	9
1GPW	1	1	1
1HE1	(616)	(4542)	(118475)
1HE8	(1046)	(4301)	(44116)
1IJK	(1121)	(966)	(777290)
1J2J	(16)	2	(5)
1JPS	**(24)**	**(1420)**	**(12959)**
1K4C	(1890)	(-)	(28657)
1K74	(61)	(54)	10
1KAC	(4359)	(8321)	(18603)

1KXQ	(29)	1	6
1MAH	**3**	**1**	**1**
1MLC	1	2	1
1N8O	(121)	(119)	(-)
1QA9	(15092)	(11140)	(103447)
1QFW	(675)	(614)	(5876)
1RLB	(42005)	(1801)	(4773)
1S1Q	(99)	(756)	(9589)
1T6B	**(205)**	**(196)**	**(580)**
1TMQ	1	8	4
1UDI	2	1	4
1YVB	4	2	(22)
1Z0K	(319)	(103)	10
1ZHI	7	1	4
2AJF	(3914)	(2430)	(69360)
2B42	**(39)**	**1**	**1**
2BTF	(77)	(167)	(34)
2OOB	(1234)	(1571)	(2788)
2VIS	(-)	(-)	(-)
7CEI	(27)	(35)	6
Medium-flexible (I-RMSD$_{C\alpha}$ 1.0-2.0 Å) (35 cases)			
1A2K	(3415)	(1290)	(3966)
1AK4	(2229)	(21200)	(14322)
1AKJ	(438)	(372)	(-)
1B6C	3	1	1
1BGX	(17600)	(-)	(841301)
1BVK	(46)	(289)	(462)
1D6R	(8)	(331)	(2660)
1DFJ	**5**	**5**	**9**
1E6E	1	1	1
1E6J	3	2	3
1EZU	**(1337)**	**(4350)**	**(6654)**
1FC2	(205)	(4336)	0
1GP2	(129)	(352)	(172024)
1GRN	(71)	(2171)	(11864)
1HIA	(62)	(50)	(37506)
1I4D	(305)	(18465)	(2033)
1I9R	(-)	(101)	(243)
1K5D	**(1475)**	**(1776)**	**(912)**
1KXP	**1**	**2**	**(7)**
1ML0	(78)	(41)	(424)
1NW9	(53)	9	9
1OPH	(376)	(70)	(660)
1VFB	(11)	(88)	(202)
1WQ1	(96)	(1618)	(2081)
1XD3	2	1	1
1XQS	7	(17)	(266)
1Z5Y	10	(22)	(23)
2CFH	(892)	(1908)	(27689)
2FD6	6	(114)	(762)

2H7V	(2345)	(-)	(1413)
2HLE	(17)	4	2
2HQS	(7)	(30)	(15)
2I25	**(231)**	**(553)**	**(284)**
2O8V	(80)	(11820)	(74)
2QFW	(132)	1	2
Flexible (I-RMSD$_{C\alpha}$ 2.0-3.0 Å) (18 cases)			
1ACB	**(135)**	**(369)**	**(555)**
1BKD	(23)	(966)	(1937)
1CGI	(45)	5	(87)
1DE4	(-)	(-)	(1725)
1E4K	(1197)	(14392)	(79673)
1EER	**(43)**	**(7292)**	**(325)**
1I2M	**(199)**	**(247)**	**5**
1IB1	(28994)	(18819)	(70454)
1IBR	**(-)**	**(-)**	**(-)**
1KKL	(280)	(458)	(222)
1M10	9	(497)	(1411)
1N2C	1	(22)	1
1PXV	**(1837)**	**(318)**	**(2852)**
2C0L	(3449)	(7473)	(62804)
2HMI	(-)	(-)	(-)
2HRK	(73)	(79)	(580)
2NZ8	(40)	(28)	(100)
2OT3	(213)	(-)	(555)
Highly-flexible (I-RMSD$_{C\alpha}$ > 3.0 Å) (8 cases)			
1ATN	**(2402)**	**(2503)**	**(3597)**
1FAK	(25)	(1670)	(34273)
1FQ1	(239)	(16535)	(35535)
1H1V	(-)	(-)	(259770)
1IRA	(-)	(-)	(-)
1JMO	(2982)	(13867)	(121326)
1R8S	(1174)	(8370)	(5277)
1Y64	(-)	(-)	(-)

[a] *PDB of the complex*
(in bold: high affinity cases)

3.3. Modeling protein interactions: application to cases of interest

The expertise acquired during the first part of this PhD thesis on the modeling and characterization of structure and dynamics of protein interactions has facilitated the successful application of computational methods to the modeling of protein interactions within different real-life contexts.

This section will be mainly focused on (i) the energetic characterization of host-pathogen protein interactions, and (ii) the *ab initio* modeling of encounter complex ensembles of redox proteins.

Manuscripts presented in this section:

1. Lucas M, Gaspar A, Pallara C, Rojas A, Fernández-Recio J, Machner M, Hierro A. (2014) **Structural basis for the recruitment and activation of the Legionella phospholipase VipD by the host GTPase Rab5.** *Proc Natl Acad Sci U S A* 111(34):E3514-23.

2. Bernal-Bayard P, Pallara C, Castell MC, Molina-Heredia FP, Fernández-Recio J, Hervás, Navarro JA. (2015) **Interaction of photosystem I from Phaeodactylum tricornutum with plastocyanins as compared with its native cytochrome c_6: reunion with a lost donor.** BBA Bioenergetics 1847(12):1549-1559.)

3.3.1. Structural basis for the recruitment and activation of the Legionella phospholipase VipD by the host GTPase Rab5.

Maria Lucas[1], Andrew H. Gaspar[2], Chiara Pallara[3], Adriana L. Rojas[1], Juan Fernández-Recio[3], Matthias P. Machner[2,*], and Aitor Hierro[1,4,*]

[1]*Structural Biology Unit, CIC bioGUNE, Bizkaia Technology Park, 48160 Derio, Spain;*

[2]*Cell Biology and Metabolism Program, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health, Bethesda, Maryland 20892, USA;*

[3]*Joint BSC-IRB research program in Computational Biology, Barcelona Supercomputing Center, Barcelona, 08034, Spain;*

[4]*IKERBASQUE, Basque Foundation for Science, 48011 Bilbao, Spain.*

*Corresponding authors

Structural basis for the recruitment and activation of the *Legionella* phospholipase VipD by the host GTPase Rab5

María Lucas[a], Andrew H. Gaspar[b], Chiara Pallara[c], Adriana Lucely Rojas[a], Juan Fernández-Recio[c], Matthias P. Machner[b,1], and Aitor Hierro[a,d,1]

[a]Structural Biology Unit, Center for Cooperative Research in Biosciences, 48160 Derio, Spain; [b]Cell Biology and Metabolism Program, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health, Bethesda, MD 20892; [c]Joint Barcelona Supercomputing Center-Institute for Research in Biomedicine Research Program in Computational Biology, Barcelona Supercomputing Center, 08034 Barcelona, Spain; and [d]IKERBASQUE, Basque Foundation for Science, 48011 Bilbao, Spain

Edited by Ralph R. Isberg, Howard Hughes Medical Institute, Tufts University School of Medicine, Boston, MA, and approved July 15, 2014 (received for review March 22, 2014)

A challenge for microbial pathogens is to assure that their translocated effector proteins target only the correct host cell compartment during infection. The *Legionella pneumophila* effector vacuolar protein sorting inhibitor protein D (VipD) localizes to early endosomal membranes and alters their lipid and protein composition, thereby protecting the pathogen from endosomal fusion. This process requires the phospholipase A1 (PLA_1) activity of VipD that is triggered specifically on VipD binding to the host cell GTPase Rab5, a key regulator of endosomes. Here, we present the crystal structure of VipD in complex with constitutively active Rab5 and reveal the molecular mechanism underlying PLA_1 activation. An active site-obstructing loop that originates from the C-terminal domain of VipD is repositioned on Rab5 binding, thereby exposing the catalytic pocket within the N-terminal PLA_1 domain. Substitution of amino acid residues located within the VipD–Rab5 interface prevented Rab5 binding and PLA_1 activation and caused a failure of VipD mutant proteins to target to Rab5-enriched endosomal structures within cells. Experimental and computational analyses confirmed an extended VipD-binding interface on Rab5, explaining why this *L. pneumophila* effector can compete with cellular ligands for Rab5 binding. Together, our data explain how the catalytic activity of a microbial effector can be precisely linked to its subcellular localization.

pathogenic bacteria | allosteric modulation | membrane composition | X-ray crystallography

Microbial pathogens have evolved numerous ways to subvert and exploit normal host cell processes and to cause disease. Intravacuolar pathogens use specialized translocation devices such as type IV secretion systems (T4SS) to deliver virulence proteins, so-called effectors, across the bacterial and host cell membrane into the cytosol of the infected cell (1–3). Many of the translocated effectors studied to date alter cellular events such as vesicle trafficking, apoptosis, autophagy, protein ubiquitylation, or protein synthesis, among others, thereby creating conditions that support intracellular survival and replication of the microbe (4, 5). Bacteria with a nonfunctional T4SS are often avirulent and degraded along the endolysosomal pathway, thus underscoring the importance of translocated effectors for microbial pathogenesis.

Although T4SS-mediated effector translocation may be a convenient way for pathogens to manipulate host cells from within the safety of their membrane-enclosed compartment, it also creates a challenging dilemma: how can the bacteria ensure that their translocated effectors reach the correct host cell target for manipulation, and how can they prevent them from indiscriminately affecting bystander organelles or proteins that may otherwise be beneficial for intracellular survival and replication of the microbe? It is reasonable to expect that regulatory mechanisms have evolved that restrain the catalytic activity of effectors. Although detailed insight into these processes is scarce, an emerging theme among effectors is that their enzymatic activity is functionally coupled to their interaction with a particular host factor. For example, SseJ from *Salmonella enterica* serovar Typhimurium displays glycerophospholipid-cholesterol acyltransferase activity only on binding to the active GTPases RhoA, RhoB, or RhoC (6–8). Likewise, *Pseudomonas aeruginosa* ExoU requires mono- or poly-ubiquitinated proteins for the activation of its phospholipase A2 (PLA_2) domain (9), whereas *Yersinia* YpkA exhibits kinase activity only in the presence of host cell actin (10). Exactly how binding to host ligands results in the activation of these translocated effectors remains unclear because no structural information for these protein complexes is available.

VipD is a T4SS-translocated substrate of *Legionella pneumophila*, the causative agent of a potentially fatal pneumonia known as Legionnaires' disease, and another example of an effector whose catalytic activity depends on the presence of a host factor (11–14). Following uptake by human alveolar macrophages, *L. pneumophila* translocates VipD together with more than 250 other effector proteins through its Dot/Icm T4SS into the host cell cytoplasm (15). These effectors act on numerous host processes to

Significance

A long-standing question in the field of microbial pathogenesis is how virulence factors are regulated within host cells and how their activity is specifically directed toward a particular host cell compartment. *Legionella pneumophila* resolves this dilemma by tightly coupling the phospholipase A1 activity of one of its effectors, vacuolar protein sorting inhibitor protein D (VipD), to this protein's interaction with endosomal host GTPases. We now present the crystal structure of VipD in complex with host cell Rab5c, providing a detailed look into the ingenious molecular mechanisms underlying the allosteric activation of a virulence factor by a host protein and its spatiotemporal regulation. These results open the path for the development of novel therapeutics aimed at blocking the VipD activation process rather than the enzyme's active site.

Author contributions: M.P.M. and A.H. designed research; M.L., A.H.G., C.P., A.L.R., and J.F.-R. performed research; M.L., A.H.G., C.P., A.L.R., J.F.-R., M.P.M., and A.H. analyzed data; and M.P.M. and A.H. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Data deposition: The structural coordinates of the VipD19-564-Rab5c18-182(Q80L):GMP–PNP complex have been deposited in the Protein Data Bank, www.pdb.org (PDB ID code 4KYI).

[1]To whom correspondence may be addressed. Email: machnerm@mail.nih.gov or ahierro@cicbiogune.es.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1405391111/-/DCSupplemental.

mediate evasion of the endolysosomal compartment and to establish a *Legionella*-containing vacuole (LCV) that supports bacterial growth (16). Although the precise biological role of most *L. pneumophila* effectors remains unclear, we recently showed that VipD is important for endosomal avoidance by LCVs. The protein localizes to endosomes presumably by binding to the small GTPases Rab5 or Rab22, key regulators of endosomal function (13, 14). Rab GTPase binding to the C-terminal domain of VipD triggers robust phospholipase A1 (PLA_1) activity within the N-terminal domain, resulting in the removal of phosphatidylinositol 3-phosphate [PI(3)P] and potentially other lipids from endosomal membranes (14). Without PI(3)P, endosomal markers such as early endosomal antigen 1 (EEA1) are lost from these membranes, most likely rendering the endosomal compartment fusion incompetent (17). *L. pneumophila* mutants lacking *vipD* are attenuated in avoiding endosomal fusion, and their LCVs acquire the endosomal marker Rab5 more frequently than LCVs containing the parental strain producing VipD (14). Thus, by coupling PLA_1 activity to Rab5 binding, the catalytic activity of VipD is directed specifically against the endosomal compartment without visibly affecting neighboring cell organelles.

VipD was originally identified in a screen for *L. pneumophila* effectors that interfere with the vacuolar sorting pathway in yeast (11). The N-terminal half of VipD possesses high homology to patatin, a lipid acyl hydrolase present in the potato tuber (12, 13). Analogous to other patatin-like proteins, VipD harbors a conserved serine lipase motif Gly-x-Ser-x-Gly (x = any amino acid) as part of a Ser-Asp catalytic dyad that, together with two consecutive glycine residues (Asp-Gly-Gly motif), is expected to stabilize the oxyanion intermediate during the acyl chain cleavage (13). Mutation of these conserved catalytic residues in VipD results in loss of PLA_1 activity (14), confirming their role in substrate hydrolysis.

The recently reported crystal structure of VipD confirmed the predicted bimodular organization (13) and, in addition, revealed a surface loop, called "lid" in other phospholipases, that shields the entry to the catalytic site. The inhibitory lid may explain why purified recombinant VipD alone exhibits little or no PLA_1 activity in vitro. However, given that binding of Rab5 or Rab22 to VipD activates the PLA_1 activity within the N-terminal region (14), we wondered if and how this binding event causes the inhibitory lid to be removed to render the active site substrate accessible.

Using an integrative approach involving X-ray crystallography, molecular dynamics, biochemistry, and cellular imaging, we now deciphered at a molecular level the mechanism that stimulates the intrinsic PLA_1 activity of VipD and determined the underlying specificity for the VipD–Rab5 interaction and endosomal targeting.

Results

Overall Structure of the VipD–Rab5 Complex. To determine the molecular basis underlying VipD binding and activation by Rab5, we initiated a crystallographic analysis of this complex. For that, we used a truncated form (residues 18–182) of constitutively active Rab5c(Q80L) lacking the N- and C-terminal hypervariable regions, and a VipD fragment [amino acid (aa) 19–564; $VipD_{19–564}$] that was designed based on a previously solved structure of full-length $VipD_{FL}$–$Rab5c_{18–182}$ at lower resolution in which the terminal residues (1–18 and 565–621) of VipD were not structured. We obtained well-diffracting crystals of $VipD_{19–564}$ in complex with $Rab5c_{18–182}$(Q80L) bound to nonhydrolyzable guanosine 5′-[β,γ-imido]triphosphate (GppNHp) and solved the structure by molecular replacement (Fig. 1). Only the last seven C-terminal residues of $VipD_{19–564}$ and a connecting loop formed by residues 345–354 could not be modeled because of poor electron density in these regions. The final model for the $VipD_{19–564}$-$Rab5c_{18–182}$(Q80L)-GppNHp structure was refined at 3.1 Å, with values for R_{factor} and R_{free} of 0.23 and 0.28, respectively (Table 1 and Fig. S1*A*).

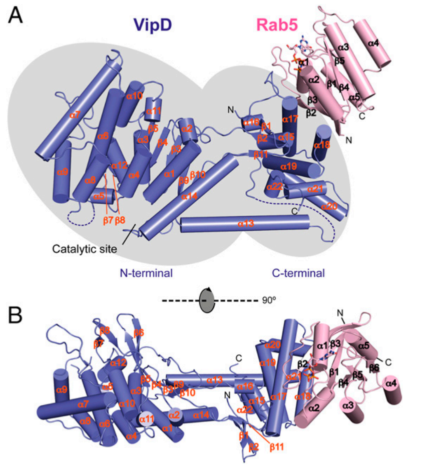


Fig. 1. The sites for substrate catalysis and Rab5 binding are situated at opposite ends of VipD. Two orthogonal tube drawing representations of the crystal structure of $VipD_{19–564}$ (slate) in complex with GppNHp-$Rab5c_{18–182}$ (pink). (*A*) Side view. (*B*) Top view. VipD comprises two distinguishable but interconnected domains highlighted by gray elliptical shadows. The N-terminal half of VipD comprises a patatin-like phospholipase domain, whereas the C-terminal domain interacts with Rab5c. Note that the catalytic site and the Rab5 binding interface are located at opposite ends of VipD.

The crystallographic asymmetric unit contained four $VipD_{19–564}$–$Rab5c_{18–182}$(Q80L) heterodimers with almost identical interaction modes (Fig. S1*B*). Superposition of the atomic coordinates showed a root mean square deviation (RMSD) of

Table 1. Data collection and refinement statistics for the $VipD_{19–564}$–$Rab5c_{18–182}$(Q80L):GppNHp complex

	$VipD_{19–564}$–$Rab5c_{18–182}$(Q80L)
Data collection	
Space group	P1
Cell dimensions	
a, b, c (Å)	94.3, 98.0, 109.9
α, β, γ (°)	76.6, 80.8, 78.9
Resolution (Å)	30–3.07 (3.26–3.07)*
R_{meas}	0.07 (0.74)
I/σ	17.0 (2.1)
Completeness (%)	97.4 (92.6)
Redundancy	3.5 (3.6)
Refinement	
Resolution (Å)	3.07
No. reflections	67,479
R_{work}/R_{free}	0.23/0.28
No. atoms	
Protein	21,659
Ligand/ion	132
B-factors	
Protein	54
Ligand/ion	74
RMSDs	
Bond lengths (Å)	0.002
Bond angles (°)	0.631

*Highest resolution shell is shown in parentheses.

0.65–0.69 Å among the four VipD$_{19–564}$–Rab5c$_{18–182}$(Q80L) complexes. Rab5c$_{18–182}$(Q80L) was in its active conformation and bound to one molecule of GppNHp and one Mg^{2+} ion (Fig. 1). It adopted the classical GTPase fold consisting of a central six-stranded β-sheet surrounded by five α-helices (18). The structure of VipD$_{19–564}$ exhibited two discernible but interconnected domains. Rab5c$_{18–182}$(Q80L) interacted extensively with a helical hairpin situated at the C-terminal domain of VipD$_{19–564}$, and, thus, at the distal end relative to the N-terminal catalytic site (Fig. 1). It is worth noting that, although the structure of active Rab5c$_{18–182}$(Q80L) remained essentially unaltered, VipD$_{19–564}$ exhibited several dramatic conformational rearrangements compared with the uncomplexed crystallographic model (13), as discussed next.

Rab5 Binding to VipD Induces Conformational Changes That Expose the Active Site. On Rab5, binding the largest RMSD in VipD$_{19–564}$ occurred in its C-terminal domain and in the structural elements that connect it to the N-terminal phospholipase domain (Fig. 2*A*). Residue Phe442 located in helix α17 of the C-terminal domain of VipD$_{19–564}$ undergoes a 90° rotation and enters a hydrophobic pocket in Rab5c$_{18–182}$(Q80L) formed by Arg82, Tyr83, and Leu86 (Fig. 2*B*). This rotation pulls the adjacent α16-α17 loop of VipD$_{19–564}$ toward Rab5c$_{18–182}$(Q80L), thereby facilitating the hydrophobic interaction of Ile433 of VipD$_{19–564}$ with Ile54 in the switch I region of Rab5c$_{18–182}$(Q80L) (Fig. 2*B*). The displacement of loop α16–α17 in VipD$_{19–564}$ induces a partial reorientation of the adjacent β-sheet formed by β1, β2, and β11, together with small shifts in helices of the C-terminal domain of VipD$_{19–564}$. These cumulative movements cause helices α13 and α14 of VipD$_{19–564}$ to swing out 14.5° and 6.6°, respectively, which is coupled with a coil-helix transition of the β9-α13 loop to adjoin helix α13 (Fig. 2*A*). This hinge motion of helices α13 and α14 ("chop-stick" mechanism) facilitates an outward displacement of the adjacent β10–α14 loop (subsequently named lid), resulting in the eventual opening of the active site (Fig. 2 *C*–*E*). Notably, there were no mayor crystallographic contacts in the areas corresponding to α13, α14, and the lid, making the displacement of the lid due to the proximity of neighboring protein molecules within the crystal lattice unlikely (Fig. S1*C*). The exposed cleft, with its catalytic residues and the oxyanion hole situated at one end, measures 16–18 Å in length and thus has the potential to accommodate a C$_{16}$–C$_{18}$ acyl chain from a lipid substrate within the adjacent hydrophobic ridge (Fig. 2*E* and Fig. S1*D*). Together, these findings provide evidence for an unprecedented heterotropic allosteric activation mechanism in which locally induced structural changes through Rab5c$_{18–182}$(Q80L) binding are transmitted from the C-terminal domain of VipD$_{19–564}$ to the N-terminal phospholipase domain, causing the displacement of the lid and exposure of the active site.

VipD–Rab5 Interface. Our complex structure revealed a single interaction path between VipD$_{19–564}$ and Rab5c$_{18–182}$(Q80L) that occluded ~722 Å^2 of solvent-accessible surfaces. Although Rab5c$_{18–182}$(Q80L) interacted with residues in the α16–α17 loop of VipD$_{19–564}$ and residues in an helical hairpin formed by helices α17 and α18 (Fig. 3*A*), the VipD binding surface in Rab5c$_{18–182}$(Q80L) included parts of the segment between α1 and β2 (the switch I region), the strands β2 and β3 (the interswitch region), and the β3–α2 segment (the switch II region) (Fig. 3*A*). The interface was composed of a core of hydrophobic contacts complemented by several polar interactions in the surrounding rim area (Fig. 3*B*). Specifically, the VipD binding epitope in Rab5c$_{18–182}$(Q80L) included nonpolar residues in the switch I (Ile54, Gly55, Ala56, and Phe58), the interswitch (Trp75), and the switch II element (Tyr83, Leu86, Met89, and Tyr90), as well as polar/charged residues in the interswitch

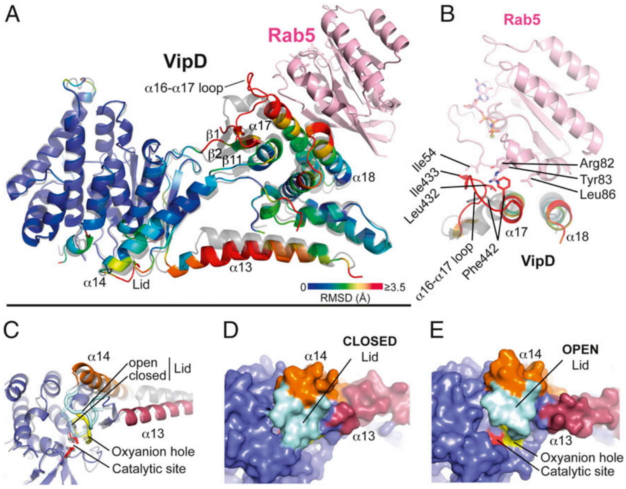


Fig. 2. Allosteric activation of VipD through Rab5 binding. (*A*) Structural changes in VipD on Rab5 binding. Rab5c$_{18–182}$ (colored in pink) is complexed to VipD$_{19–564}$, which is colored from slate to red based on the root mean square deviation (RMSD) of C-α atom pairs when superimposed with the unbound form of VipD$_{19–564}$ (PDB ID code 4AKF) shown in transparent gray. The black line represents the membrane plane. (*B*) Close-up of VipD–Rab5 interaction. The α17-α18 loop of VipD undergoes a Rab5-induced conformational rearrangement resulting in residue Phe442 of VipD being inserted into a hydrophobic pocket formed by Arg82, Tyr83, and Leu86 of Rab5. The displacement of the α16–17 loop favors the hydrophobic interaction between Leu432 and Ile433 of VipD with Ile54 of Rab5. Color code as in *A*. The remaining VipD structure has been omitted for clarity. (*C*) Close-up view of the catalytic site highlighting displacement of the lid (β10-α14 loop, light blue). (*D* and *E*) Surface representation of the unbound (*D*) and Rab5-bound (*E*) VipD molecule, respectively. Same view as in *C*.

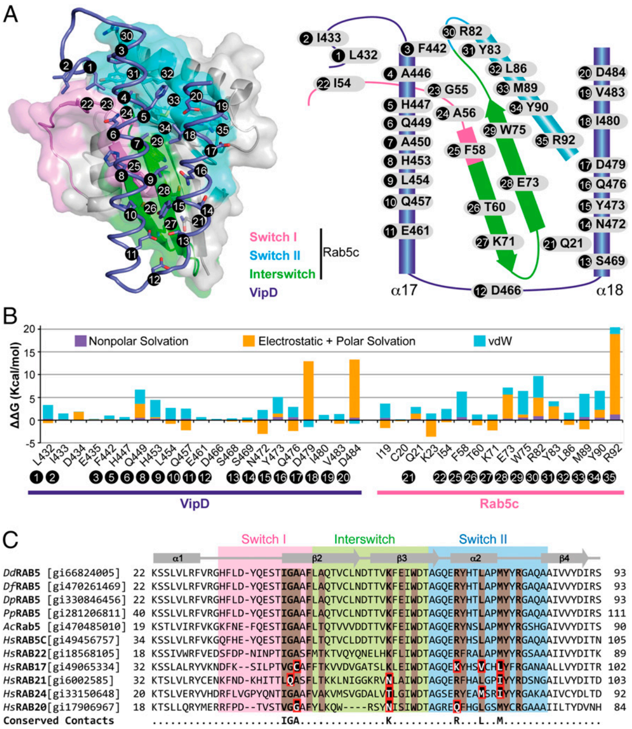

A
Switch I
Switch II
Interswitch
VipD
Rab5c
α17
α18
B
ΔΔG (Kcal/mol)
Nonpolar Solvation
Electrostatic + Polar Solvation
vdW
VipD
Rab5c
C
Switch I
Interswitch
Switch II
α1
β2
β3
α2
β4
DdRAB5 [gi66824005] 22 KSSLVLRFVRGHFLD-YQESTIGAAFLAQTVCLNDTTVKFEIWDTAGQERYHTLAPMYYRGAQAAIVVYDIRS 93
DfRAB5 [gi470261469] 22 KSSLVLRFVRGHFLD-YQESTIGAAFLAQTVCLNDTTVKFEIWDTAGQERYHTLAPMYYRGAQAAIVVYDIRS 93
DpRAB5 [gi330846456] 22 KSSLVLRFVRGHFLD-YQESTIGAAFLAQTVCLNDTTVKFEIWDTAGQERYHTLAPMYYRGAQAAIVVYDIRS 93
PpRAB5 [gi281206811] 40 KSSLVLRFVRGHFLD-YQESTIGAAFLAQTVCLNDTTVKFEIWDTAGQERYHTLAPMYYRGAQAAIVVYDIRS 111
AcRab5 [gi470485010] 19 KSTLVIRFVKGKFNE-FQESTIGAAFLTQTVVVDDTTVKFEIWDTAGQERYNSLAPMYYRGSHGAIVVYDITS 90
HsRAB5C[gi49456757] 34 KSSLVLRFVKGQFHE-YQESTIGAAFLTQTVCLDDTTVKFEIWDTAGQERYHSLAPMYYRGAQAAIVVYDITN 105
HsRAB22[gi18568105] 18 KSSIVWRFVEDSFDP-NINPTIGASFMTKTVQYQNELHKFLIWDTAGQERFRALAPMYYRGSAAAIIVYDITK 89
HsRAB17[gi49065334] 32 KSSLALRYVKNDFK--SILPTVGCAFFTKVVDVGATSLKLEIWDTAGQEKYHSVCHLYFRGANAALLVYDITR 102
HsRAB21[gi6002585] 32 KTSLVLRYCENKFND-KHITTLQASFLTKKLNIGGKRVNLAIWDTAGQERFHALGPIYYRDSNGAILVYDITD 103
HsRAB24[gi33150648] 20 KTSLVERYVHDRFLVGPYQNTIGAAFVAKVMSVGDALVTLGIWDTAGSERYEAMSRIYYRGAKAAIVCYDLTD 92
HsRAB20[gi17906967] 18 KTSLLQRYMERRFPD--TVSTVGGAFYLKQW----RSYNISIWDTAGREQFHGLGSMYCRGAAAIILTYDVNH 84
Conserved Contacts ....................IGA.............K..........R...L..M................

Fig. 3. Molecular interactions at the VipD–Rab5c interface. (*A*) (*Left*) Semitransparent surface of GppNHp-Rab5c$_{18–182}$ in complex with the minimal Rab binding domain of VipD (slate ribbon model) highlighting the interfacial residues below 4.0-Å distance. (*Right*) Schematic diagram of interfacial residues in the VipD–Rab5c complex. (*B*) Detailed description of per-residue contribution from van der Waals (vdW) energy (blue), nonpolar solvation energy (purple), and the sum of electrostatic and polar solvation energy (orange) calculated by computational alanine scanning for interfacial residues in the VipD–Rab5c complex. Existing glycines and alanines are excluded in the calculation. (*C*) Sequence conservation between Rab5 GTPases from amoebean species and human homologs. *Dd*, *Dictyostelium discoideum*; *Df*, *Dictyostelium fasciculatum*; *Dp*, *Dictyostelium purpureum*; *Pp*, *Polysphondylium pallidum*; *Ac*, *Acanthamoeba castellanii*; *Hs*, *Homo sapiens*. Rab5c residues contacting VipD at a distance less than 4 Å are colored in light brown. Amino acid substitutions within the equivalently aligned interfacial residues of other Rabs are highlighted in a red box. Interfacial residues strictly conserved between Rab5 and Rab22, but variable in any of the other Rabs, are depicted in the bottom line of the alignment. Protein accession numbers are in brackets.

(Thr60, Lys71, and Glu73) and switch II element (Arg82 and Arg92) (Fig. 3 *A* and *B*). A comparison of the primary sequence of human Rab5 and Rab22 with Rab5 from several natural amoebean hosts found conserved residues at equivalent contact sites in the switch I (Ile54, Gly55, and Ala56), interswitch (Lys71), and switch II region (Arg82, Leu86, and Met89) that were variable in other Rab proteins (Fig. 3*C*), suggesting these residues are involved in the specific recognition by VipD. The corresponding epitope in VipD included several hydrophobic residues in helix α17 (Phe442, Ala446, Ala450, and Leu454) and in helix α18 (Tyr473, Ile480, and Val483) that wrapped around an elongated hydrophobic path in Rab5 formed by the conserved triad (Phe58, Trp75, and Tyr90) and Leu86. Surrounding this hydrophobic core were additional hydrogen bonds that enhanced the interaction.

Like all GTPases, Rab5 exhibits structural changes within its switch regions dependent on its nucleotide-binding state (GDP vs. GTP), with the largest conformational variation in switch I (19). The structure of the VipD$_{19–564}$–Rab5c$_{18–182}$(Q80L) complex revealed that Leu432 and Ile433 of VipD$_{19–564}$ interacted with Ile54 in switch I of Rab5c$_{18–182}$(Q80L), therefore sensing its GTP-bound state (Fig. 2*B*). In fact, the conformation adopted by Rab5 in its GDP-bound state (19) resulted in a prominent steric clash between the switch I region and helix α17 of VipD$_{19–564}$ (Fig. S1*E*), thus explaining why this activation state of Rab5 is only a poor ligand for VipD (13, 14).

Validation of the VipD–Rab5 Interface Through Mutational Analysis. To experimentally validate the VipD–Rab5 binding interface seen in the crystal structure, we mutated several residues predicted to contribute to this protein–protein interaction and examined their role for complex formation in coprecipitation studies (Fig. 4*A*). Substitution of individual contact residues within VipD abrogated Rab5 binding either severely (F442A and H453D) or moderately (Q449A, E461R, Y473A, and D479H), whereas only a few of the tested substitutions in VipD were tolerated (Q476A and D484H). Similar results were observed for Rab5 interface mutants (Fig. 4*B*), with binding defects ranging from severe (F58A, Y83A, and R92E) to mild (E73R and Y90A). We also studied the mode of interaction between VipD

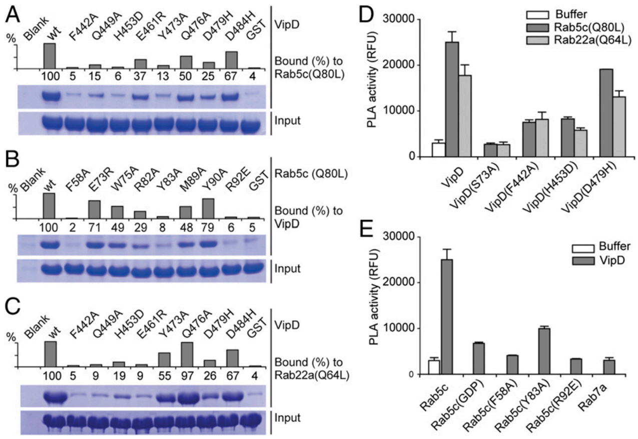


Fig. 4. Mutational analysis of interfacial recognition determinants. (*A*–*C*) Pulldown assays using the indicated VipD (*A* and *C*) or Rab5 (*B*) mutant proteins. The graphs are a densitometry-based quantification of the amount of query protein precipitated by the respective bait-coated beads. Input, total amount of query. (*D* and *E*) Fluorescence-based PLA activity assays using VipD (*D*) or Rab5c (*E*) variants. (*D*) The indicated VipD protein was incubated with Rab5(Q80L)$_{18-182}$:GppNHp or Rab22(Q64L)$_{16-181}$:GppNHp (molar ratio 1:2) or with buffer alone, and PLA$_1$-dependent cleavage of the substrate Bis-BODIPYFL C$_{11}$-PC was detected as an increase in fluorescence emission [relative fluorescence units (RFUs)]. (*E*) Same assay as in *D* using the indicated Rab protein variants.

and Rab22. As expected, substitution of individual contact residues of VipD required for Rab5c$_{18-182}$(Q80L) binding (Fig. 4*A*) also resulted in a failure to stably associate with Rab22(Q64L) (Fig. 4*C*), suggesting that Rab22 occupies an epitope in VipD very similar to that of Rab5. None of the amino acid substitutions significantly altered the overall structure of the mutant proteins as evaluated by circular dichroism (CD) (Fig. S2), indicating that a reduction in binding was most likely not a consequence of protein misfolding.

Given that the PLA$_1$ activity of VipD is triggered only in response to Rab5 binding, we analyzed how amino acid substitutions that attenuate VipD–Rab5 complex formation affect the PLA$_1$ activity of VipD. Using a generic fluorogenic substrate (bis-BODIPY FL C$_{11}$-PC), we found a tight correlation between loss of PLA$_1$ activity and the inability of VipD mutant proteins (F442A, H453D, and D479H) to enter a stable complex with Rab5c$_{18-182}$(Q80L) or Rab22$_{16-181}$(Q64L) (Fig. 4*D*). Similar results were observed for Rab5c$_{18-182}$(Q80L) mutant proteins (F58A, Y83A, and R92E) that had failed to stably associate with VipD and were hence unable to trigger its PLA$_1$ activity (Fig. 4*E*). The observed crystallographic interaction between VipD and Rab5 thus corresponded to their molecular association in solution, and failure to form a stable VipD–Rab5 or –Rab22 complex caused the PLA$_1$ domain to remain in its catalytically inactive state.

Disruption of the Interaction with Rab5 Precludes Endosomal Targeting of VipD. Within transiently transfected COS1 cells, fluorescently tagged VipD was enriched on Rab5-containing early endosomes, and this colocalization required the C-terminal Rab5 binding domain but not the N-terminal PLA$_1$ domain (13, 14). A recent study reported that depletion of Rab5 (isoforms a-c) and Rab22a from HeLa cells by RNA interference (RNAi) did not affect VipD targeting to endosomes, claiming that endosomal localization of VipD would not simply depend on the interaction with Rab proteins (20). Given that RNAi rarely depletes the entire pool of a given cellular target and that VipD recruitment to endosomes could have been mediated not only by Rab5 and/or Rab22 but by additional yet unidentified Rab GTPases, we set out to reevaluate VipD's endosomal targeting mechanism. For that, we analyzed the intracellular distribution pattern of four VipD mutant proteins that were either severely (F442A) or moderately (E461R and Q476A) attenuated for Rab5c binding in vitro (Fig. 4). Although WT VipD displayed robust colocalization with GFP-Rab5c$_{18-182}$(Q80L)-positive endosomes, VipD(F442A) was entirely cytosolic (Fig. 5), consistent with this mutant's inability to bind Rab5c. In contrast, VipD (E461R) and VipD(Q476A) showed no apparent difference in localization compared with WT VipD (Fig. 5). These findings strongly suggest that endosomal targeting of VipD is in fact dependent on the interaction with host cell Rab GTPases and that interference with the formation of these protein complexes results in the failure of VipD to properly localize to endosomes.

The N-Terminal Tail of VipD Is Crucial for PLA$_1$ Activity. In the uncomplexed structure of VipD, the N-terminal tail (residues 1–18; N18) contained a small amphipathic helix (H1) that was involved in an intermolecular crystal contact (13). The structure of full-length VipD$_{1-621}$ bound to Rab5c$_{18-182}$(Q80L), on the other hand, contained no clear electron density for N18, suggesting that this region of VipD possessed high flexibility. Small angle X-ray scattering (SAXS) and gel filtration chromatography analysis suggest a heterodimeric VipD$_{1-621}$–Rab5c$_{18-182}$(Q80L) complex in solution, indicating that N18 was not involved in any oligomer formation (Fig. 6 *A* and *B* and Fig. S3). Given that the complexes of Rab5c$_{18-182}$(Q80L) with either full-length VipD$_{1-621}$ or truncated VipD$_{19-564}$ exhibited nearly indistinguishable structures, we concluded that N18 was dispensable for the conformational changes induced by Rab5 binding. Consequently, we evaluated whether this short region was also dispensable for PLA$_1$ activity of VipD. Unexpectedly, we found that, unlike VipD$_{1-621}$, the truncated fragment VipD$_{19-564}$ lacking N18 was strongly attenuated for PLA$_1$ activity (Fig. 5*C*). To exclude the possibility that loss of PLA$_1$ activity in VipD$_{19-564}$ was caused by the lack of the C-terminal region (aa 565–621), we tested two additional constructs, VipD$_{1-564}$ and VipD$_{19-621}$, and detected robust PLA$_1$ activity only in VipD$_{1-564}$, indicating that N18 but not the

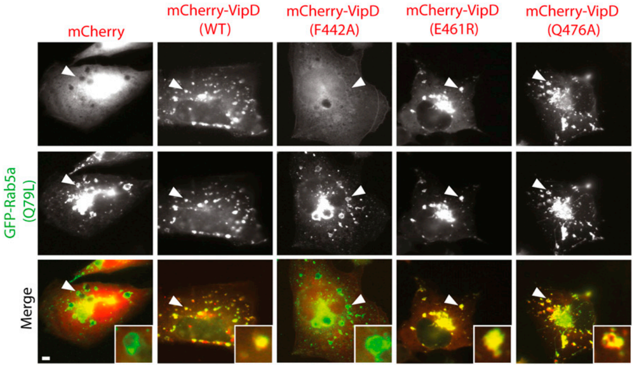


Fig. 5. VipD localization to endosomes requires a functional Rab5 binding interface. Transiently transfected COS-1 cells producing Rab5a(Q79L) and the indicated mCherry-tagged VipD variants were analyzed by fluorescence microscopy to determine protein localization. The merged images (bottom row) show Rab5a(Q79L) in green and VipD variants in red. (*Insets*) Magnified view of endosomes marked by an arrowhead. Control, mCherry. (Scale bar, 2 μm.)

C-terminal region critically contributed to the catalytic activity of this *L. pneumophila* effector (Fig. 5*C*). According to these observations, we propose that the flexible N18 with its amphipathic helix H1 and its close distance to the membrane plane may promote peripheral association of VipD with the lipid bilayer, possibly by orienting the catalytic site toward the membrane and/or assisting in substrate transfer.

Competitive Rab5 Binding Through Interface Expansion. To localize to and stably associate with endosomal membranes, VipD needs to outcompete cellular ligands for Rab5 binding. EEA1, Rabaptin-5, and Rabenosyn-5 are each bound by Rab5 through a surface that includes the switch and interswitch region and that significantly overlaps with the epitope for VipD binding (19, 21, 22). To determine if and how the distribution of interaction energies differs within each of these complexes, we extended the computational alanine scan to the EEA1, Rabaptin-5, and Rabenosyn-5 epitopes and calculated the free binding energy for each of their residues (Fig. 7*A* and Figs. S4 and S5). All four analyzed protein interfaces share a number of nonpolar interacting residues in Rab5, namely the conserved triad (Phe58, Trp75, and Tyr90), with relatively similar energetic contributions to binding (Fig. 7*A* and Fig. S5*B*). The polar interactions surrounding this hydrophobic triad, however, determine their differential affinity, with the contact of

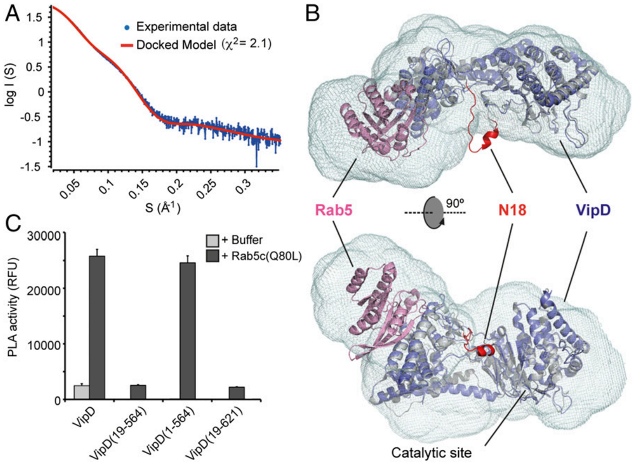


Fig. 6. The N-terminal 18 residues of VipD are essential for its PLA activity. (*A*) Fit of the optimized crystallographic $\text{VipD}_{1\text{–}564}\text{-Rab5c}_{18\text{–}182}$(Q80L) model (red line) to the experimental SAXS data of the complex (blue dots). (*B*) Fitting of the $\text{VipD}_{19\text{–}564}\text{-Rab5c}_{18\text{–}182}$(Q80L) crystallographic model (VipD in slate and Rab5 in pink) into the averaged ab initio envelope in two orthogonal views and superimposed with the unbound form of VipD (PDB 4AKF) in gray. Note the proximity of the N18 (residues 1–18 of VipD, PDB 4AKF) in red to the catalytic site. (*C*) Fluorescence-based PLA activity assays showing that the N18 segment of VipD is essential for its PLA_1 activity.

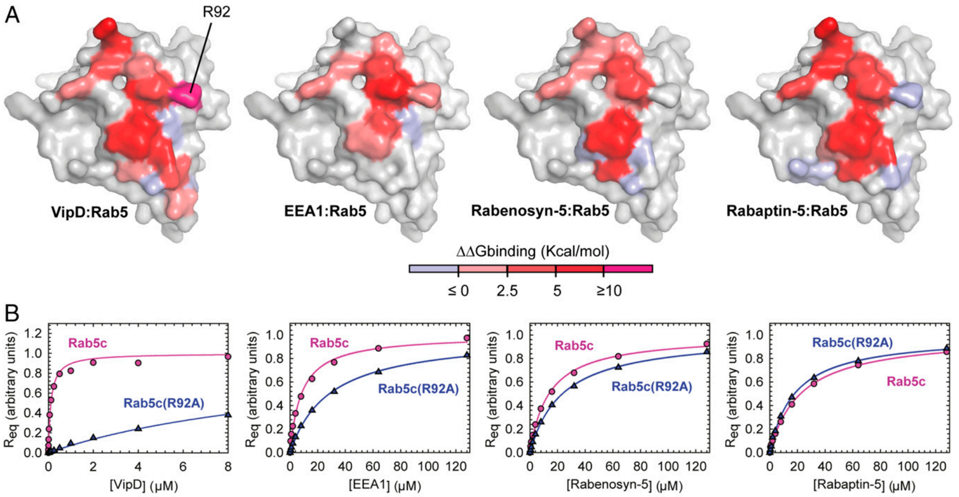
Arg92$_{\mathrm{Rab5}}$ with Asp479$_{\mathrm{VipD}}$ and Asp484$_{\mathrm{VipD}}$ providing a particularly large energetic contribution to the interaction of VipD with Rab5 compared with the other cellular ligands (Fig. 7*A* and Fig. S5*B*). To verify the importance of this predicted hot-spot for VipD binding, we analyzed the affinities of either Rab5c$_{18–182}$(Q80L) or Rab5c$_{18–182}$(Q80L, R92A) toward VipD by surface plasmon resonance (SPR) spectroscopy. R92A mutation in Rab5 severely decreases the binding for VipD 124-fold while affecting the interaction with EEA1, Rabenosyn-5, and Rabaptin-5 to a much lesser extent (3.2-, 1.7- and 1.0-fold, respectively) (Fig. 7 *B* and *C* and Fig. S6). These findings pinpoint a binding hotspot for the superior affinity of VipD over the endogenous Rab5 ligands and confirm a good qualitative correlation between the computational analysis and the experimentally observed results.

Discussion

VipD from *L. pneumophila* has long been predicted to function as a phospholipase during infection (11, 12), yet its catalytic activity had only recently been confirmed when it was shown that binding of host cell Rab GTPases (Rab5 and Rab22) is necessary for VipD to exhibit robust PLA$_1$ activity (14). The crystallographic analysis described here provides an in-depth view of the Rab5-mediated activation mechanism. Above all, it uncovered a complex cascade of structural rearrangements within the C-terminal domain of VipD that result in the relocation of an active site-occluding lid and the exposure of the substrate binding pocket within the N-terminal PLA$_1$ domain of VipD.

The structure of VipD in complex with Rab5c(Q80L) presented here is, to our knowledge, the first of a bacterial phospholipase bound to a host cell protein and the first of any translocated effector in complex with its allosteric activator molecule. Phospholipases constitute a common cellular tool to alter the lipid composition of membranes, and their activity must be carefully dosed and precisely directed toward the respective target membrane. There are more than 10,000 proteins (8,101 in Bacteria and 2,374 in Eukaryotes) containing potential patatin-like domains, most of them within a modular domain arrangement (23). Many members of the family of cytosolic phospholipases A$_2$ (cPLA$_2$), all of which share a patatin-like fold, contain a C2 domain crucial for membrane localization (24, 25). The patatin-like fold is also highly homologous to the group of calcium-independent phospholipases A$_2$ (iPLA$_2$), in which many members contain ankyrin repeats, a repetitive helix-turn-helix-loop structure considered to be a common platform for protein–protein interactions (24). Considering that Rab GTPases are key players in defining membrane identity and that many effectors from *L. pneumophila* have been acquired via horizontal gene transfer (26, 27), it is plausible that the scheme presented here for the concomitant localization and activation of VipD can be generalized across other microbial and eukaryotic phospholipases.

Human Rab5 interacted with VipD through a helix-turn-helix element that was similar to that used for Rabenosyn-5 binding (21), although the interface was slightly shifted toward the switch II region. Despite the observed overlapping contacts, the energy for



C

	VipD	EEA1$_{36\text{-}91}$	Rabenosyn-5$_{1\text{-}70}$	Rabaptin-5$_{739\text{-}862}$
Kd Rab5c(Q80L) (μM)	0.11 ± 0.02	7.4 ± 1.4	12.3 ± 1.6	19.7 ± 3.2
Kd Rab5c(Q80L,R92A) (μM)	13.6 ± 1.4	23.9 ± 4.7	20.5 ± 3.1	19.6 ± 1.9
Kd$_{\mathrm{Rab5c(R92A)}}$/Kd$_{\mathrm{Rab5c}}$	124.0	3.2	1.7	1.0

Fig. 7. Energy distribution between different effector binding epitopes in Rab5. (*A*) Space-filling model of Rab5c highlighting the epitopes for the interaction with VipD (*Far Left*), EEA1 C2H2 Zinc Finger (PDB ID code 3MJH) (*Center Left*), Rabenosyn-5 (rebuilt from PDB ID code 1Z0J) (*Center Right*), and Rabaptin-5 (PDB ID code 1TU3) (*Far Right*). Epitopes are colored as an intensity gradient according to the binding free energy change (ΔΔG) estimated as the difference between the binding ΔG of the WT and that of the alanine mutated complex. Existing glycines and alanines are excluded in the calculation. (*B*) Concentration dependence of the equilibrium surface plasmon resonance response for the binding of VipD WT, EEA1 C2H2 Zinc Finger (aa 36–91), Rabenosyn-5 (aa 1–70), and Rabaptin-5 (aa 739–862) to Rab5c$_{18–182}$(Q80L) or Rab5c$_{18–182}$(Q80L, R92A). R$_{\mathrm{eq}}$ represents the equilibrium SPR response normalized to the fitted maximum value for each dataset. (*C*) Table of the mean *K*$_{\mathrm{d}}$ values and SDs for at least two independent experiments showing the *K*$_{\mathrm{d}}$ ratio variation.

VipD binding was not distributed uniformly across the interface but instead concentrated into a combination of hotspots that provide superior binding affinity and specificity (Fig. 7 and Fig. S6). A conserved hydrophobic triad in Rab5 (Phe58, Trp75, and Tyr90) supplied the core binding energy that was complemented by more specific polar and nonpolar contacts. Notably, most of these residues were highly conserved among Rab5 homologs from amoebean species, the natural host of *L. pneumophila*, or from the surrogate host *Dictyostelium* sp. (Fig. 3*C*). The ability to discriminate between GDP- and GTP-bound Rab5 and to compete with endogenous ligands evidences a remarkable adaptation for directing and retaining VipD on endosomal membranes. Interference with VipD–Rab5 complex formation, for example, by substituting Phe442 or His453 of VipD, strongly reduced the capability of these mutant proteins to interact with Rab5 (Fig. 4*A*), to exhibit PLA$_1$ activity (Fig. 4*D*), and to efficiently localize to the endosomal compartment (Fig. 5), thus demonstrating that the function of VipD is intimately coupled to the presence of this host GTPase.

A hallmark feature of many phospholipases is to be minimally active on monomeric lipid substrates but undergo a substantial activation on binding to the surface of phospholipid membranes or micelles, a phenomenon known as interfacial activation (28–31). This behavior has been attributed to a flexible lid that at the lipid–water interface facilitates substrate diffusion to the catalytic site rather than being allosterically modulated through distant ligand binding (25). VipD does not display any interfacial activation despite having a short lid occluding the access to the catalytic site. Rather, when VipD is bound to Rab5, the lid is displaced through a chopstick-like activation mechanism in which the swing movement of two α-helices (α13 and α14) allosterically controls accessibility of the catalytic site. We cannot exclude the possibility that additional mechanisms contribute to the activity and/or specificity of the substrate catalysis by VipD. For example, the coil–helix transition of the β9–α13 loop to adjoin helix α13 relocates several charged residues closer to the catalytic groove, which might result in interactions with lipid head groups or other membrane components. Consistent with this notion, we discovered that the flexible N-terminal segment N18 of VipD is critical for the catalysis of a generic membrane-embedded substrate (Fig. 6). Deletion of N18 reduced the PLA$_1$ activity of VipD$_{19–564}$ but did not interfere with allosteric activation of the catalytic site. We hypothesize that N18 bearing a short amphipathic α-helix may facilitate the correct positioning of the PLA$_1$ domain toward the lipid layer, promote substrate diffusion from the lipid–water interface into the catalytic site, or a combination of both effects as has been previously described for secretory PLA$_2$ enzymes (32, 33). Interestingly, the N-terminal tail of VpdA (residues 11–54) and the N-terminal segment of *P. aeruginosa* ExoU (residues 57–96) showed structural similarity to the equivalent region of VipD despite lacking sequence homology (Fig. S7). Moreover, region 57–96 of ExoU, although not part of the phospholipase domain (34), was critical for cytotoxicity within transfected mammalian cells (35), suggesting that this segment contributes to the phospholipase activity of ExoU. We hypothesize that the equivalent region in VpdA may be equally important for the catalytic activity of this *L. pneumophila* effector and that appendixes, such as the N-terminal domains, may play important yet unresolved roles in membrane association and/or substrate transfer in other bacterial phospholipases.

In summary, our findings disclose an unexpected mode of long-range allosteric regulation of the PLA$_1$ activity of VipD and explain how endosomal targeting is accomplished through competitive Rab5 binding. Our study also provides the basis for the development of novel therapeutic approaches that, rather than directly targeting the enzyme's active site, specifically disturb the host factor-mediated activation process of VipD and related microbial phospholipases.

Materials and Methods

Plasmids and Cloning. The DNA sequences encoding VipD, VipD$_{19–564}$, VipD$_{1–564}$, VipD$_{19–621}$, Rab5c$_{18–182}$(Q80L), Rab22a$_{16–181}$(Q64L), EEA1$_{36–91}$, and Rabenosyn5$_{1–70}$ were cloned into the vector pGST-Parallel2 (36) using BamHI and XhoI restriction sites. Rabaptin5$_{739–862}$ and Rab7a(Q67L) were cloned between the NcoI and XhoI restriction sites of pGST-Parallel2. PCR was performed using Phusion polymerase (Thermo). The PCR product was purified with QIAquick Gel Extraction Kit (NewEngland) and ligated into the digested pGST-Parallel2 vector using Quickligase (BioLab). The ligation mixture was used to transform *Escherichia coli* XL1 Blue competent cells, and transformants were then selected on Luria-Bertani (LB) plates containing 100 μg/mL ampicillin. The presence of the insert in the plasmid was tested by colony PCR. Quickchange mutagenesis was used to make directed mutations. The correct transformants were grown to isolate the plasmids that were sequenced on both strands. Plasmids and oligonucleotides used in this study are listed in Tables S1 and S2, respectively.

Protein Expression and Purification. VipD was purified from *E. coli* BL21 (DE3) grown in LB medium and induced at an OD$_{600}$ = 0.8 by the addition of 0.5 mM isopropyl β-D-1-thiogalactopyranoside. Cells were harvested after 16 h of growth at 18 °C. The cell pellet was resuspended in buffer A (50 mM Tris·HCl, pH 8.0, 300 mM NaCl, and 1 mM DTT) supplemented with 0.1 mM phenylmethylsulfonyl fluoride, 1 mM benzamidine, and 1 mg/mL lysozyme and disrupted by sonication, and the lysate was cleared by centrifugation at 50,000 × *g* for 45 min. The supernatant was incubated for 2 h in batch with glutathione Sepharose beads (GE Healthcare) followed by extensive washing of the beads with buffer A in a gravity column. The N-terminal glutathione S-transferase (GST)-tag and linker were proteolytically removed by overnight incubation at 4 °C in the presence of tobacco etch virus (TEV) protease in 50 mM Tris·HCl, pH 8.0, 150 mM NaCl, and 1 mM DTT. The cleaved protein was eluted and further purified by ion exchange chromatography (HitrapQ; GE Healthcare) using a gradient of 50–1,000 mM NaCl, followed by size exclusion chromatography (Superdex200 16/60; GE Healthcare) in buffer B [25 mM Tris·HCl, pH 7.5, 150 mM NaCl, 5% (vol/vol) glycerol, and 1 mM DTT]. VipD mutants and truncated constructs were purified following the same procedure. The concentration of these proteins was calculated using the theoretical extinction coefficient.

Rab5c$_{18–182}$(Q80L) was purified as described for VipD, with the difference that the HitrapQ column gradient was 25–1,000 mM, and the size exclusion chromatography was performed in a Superdex75 16/60 (GE Healthcare). Nucleotide exchange was achieved by incubation of the purified protein with a 20-fold excess of GppNHp in 50 mM Tris, pH 7.5, 150 mM NaCl, and 5 mM EDTA for 12 h at 4 °C. The exchange reaction was stopped by addition of MgCl$_2$ (10 mM final concentration). Excess nucleotide was removed by gel filtration using a Superdex75 16/60 column in buffer C [25 mM Tris·HCl, pH 7.5, 150 mM NaCl, 5% (vol/vol) glycerol, 1 mM MgCl$_2$, and 1 mM DTT]. Rab5c$_{18–182}$, Rab5c$_{18–182}$(Q80L) mutants, Rab22a$_{16–181}$(Q64L), and Rab7a(Q67L) were purified as previously described. The concentration of the Rab proteins was determined by using Bradford's procedure with BSA as standard.

For complex formation, VipD was incubated with GppNHp-bound Rab5c$_{18–182}$(Q80L) in a 1:3 molar ratio for 2 h at 4 °C. The complex was further purified using a Superdex200 16/60 column equilibrated in buffer C and concentrated to 50 mg/mL using Amicon centrifugal concentrators (Millipore).

GST, GST-VipD, GST-Rabenosyn5$_{1–70}$, GST-EEA1$_{36–91}$, and GST-Rabaptin5$_{739–862}$ were purified with the same protocol as described for VipD with the only difference that no TEV protease cleavage was performed during the purification.

Crystallization and Structure Determination. Crystals were obtained by hanging-drop vapor diffusion at 18 °C by mixing 1 μL purified VipD$_{19–564}$–Rab5c$_{18–182}$(Q80L):GppNHp complex (50 mg/mL) and 1 μL precipitant solution (16% PEG6000, 0.1 M Tris·HCl, pH 8.0, and 0.2 M LiCl). Rod-shaped crystals grew within 2–3 d. Individual crystals were cryo-protected by a brief soak in well buffer supplemented with 25% (vol/vol) ethylene glycol and flash frozen in liquid nitrogen.

Diffraction data were collected at 100 K using radiation with a wavelength of 0.976 Å on beamline I04 at the Diamond Light Source (Didcot, UK). The data were integrated and scaled using XDS (37). The structure was solved by molecular replacement using the coordinates of VipD [Protein Data Bank (PDB) ID code 4AKF) and Rab5c (PDB ID code 1Z07) as a search model in Phaser (38). Subsequent rounds of refinement and interactive manual building were performed using Phenix (39) and Coot (40). For cross-validation, 5% of the original reflections was omitted from refinement and used to calculate the free R factor. The final model contained four complexes of VipD–Rab5c(Q80L). Only 10 residues (345–354) located in a connecting loop could not be modeled because of poor electron density in this region. Crystallographic data collection and model statistics are summarized in Table 1. The Ramachandran plot of the model calculated with the Rampage evaluation tool (41) shows 96.0% of the residues in the favored regions, whereas 3.9% fall in the allowed regions and 0.1% in disallowed regions. Graphics presented in this manuscript were generated using the program PyMOL (The PyMOL Molecular Graphics System; Schrödinger).

Pulldown Assays. In vitro pulldown assays involving VipD–Rab5 interface mutants included GST-VipD, GST-Rab5c$_{18–182}$(Q80L), VipD point mutants, and Rab5c$_{18–182}$(Q80L) point mutants. The pulldown between VipD and Rab22 proteins included Rab22a$_{16–181}$(Q64L) and VipD point mutants. No TEV protease cleavage was performed during the purification of GST-VipD, GST-Rab5c$_{18–182}$(Q80L), or GST-Rab22a$_{16–181}$(Q64L). For pulldowns involving VipD point mutants and GST-Rab5c$_{18–182}$(Q80L), 13 μM VipD (WT or mutants) was mixed with 10 μM of GST-Rab5c$_{18–182}$(Q80L) in binding buffer (50 mM Tris·HCl, pH 7.5, 150 mM NaCl, 1 mM MgCl$_2$, and 1 mM DTT). Then 10 μL of equilibrated glutathione Sepharose beads were added to 70 μL of the protein mixture and incubated for 2 h at 4 °C with gentle agitation. Beads were washed several times with 0.5 mL of binding buffer and resuspended in sample buffer. The samples were subjected to 4–12% SDS/PAGE analysis, and gels were stained with Coomassie brilliant blue. The pulldown experiment between GST-Rab22a$_{16–181}$(Q64L) and VipD point mutants was performed in a like manner. The binding between GST-VipD and Rab5c$_{18–182}$(Q80L) point mutants was analyzed similarly using 20 μM GST-VipD and 30 μM Rab5c$_{18–182}$(Q80L) (WT or mutants) in each individual reaction. Each pulldown was performed in triplicate with similar outcomes.

Phospholipase Assays. The phospholipase activity of VipD and its mutants was assayed using bis-BODIPY FL C11-PC (Invitrogen), a glycerophosphocholine with BODIPY FL dye-labeled sn-1 and sn-2 acyl chains, respectively. This fluoregenic substrate is selfquenched and release of the fluorophores by acyl chain cleavage by either PLA1 or PLA2 results in increased fluorescence. To prepare fluorescence-labeled liposomes, we mixed 30 μL 10 mM dioleoylphosphatidylcholine (DOPC), 30 μL 10 mM dieloylphosphatidylglycerol (DOPG), and 30 μL of 1 mM bis-BODIPY FL C11-PC. All these compounds were dissolved in ethanol. The substrate was incorporated into liposomes by a slow injection of this ethanolic lipid mix into 5 mL assay buffer (50 mM Tris, pH 7.5, 150 mM NaCl, and 1 mM MgCl$_2$) under continuous stirring. The mixture was pipetted into the side of the vortex using a narrow orifice gel-loading tip. Fifty microliters of each substrate solution was incubated with 50 μL bis-BODIPY FL C11-PC–labeled liposomes in 96-well plates for 2 h at 25 °C. The reaction mixtures contained VipD 2.5 μM and Rab 5 μM or the corresponding mutants in assay buffer. The fluorescence intensity was measured at 485-nm excitation and 530-nm emission in a multiwell reader (Biotek Synergy HT-1). All of the measurements were performed in triplicate. The assay buffer in the absence of enzyme was used as a blank.

Immunofluorescence Microscopy. VipD localization was analyzed in COS-1 cells grown on coverslips in 24-well plates in 5% CO$_2$ at 37 °C in DMEM media supplemented with 10% FBS. Semiconfluent monolayers were transiently transfected using Lipofectamine 2000 (Invitrogen) to produce fluorescently tagged (mCherry or GFP) proteins. Cells were fixed 10 h after transfection, and images were analyzed on a Zeiss Axio Observer.Z1 inverted light microscope using a Zeiss Plan-Apochromat 63×/ oil M27 objective and processed with Zeiss AxioVision 4.7.2 software.

SAXS. Synchrotron SAXS data were collected on beamline BM29 at ESRF (Grenoble, France) with a 2D detector (Pilatus 1M) over an angular range $q$min = 0.01 Å^{-1} to $q$max = 0.5 Å^{-1}. X-ray scattering patterns were recorded with the VipD$_{1–564}$–Rab5c$_{18–182}$(Q80L) complex at 2.2 and 6.4 mg/mL in 150 mM NaCl, 0.5 mM tris-(2-carboxyethyl)phosphine, and 25 mM Hepes, pH 7.5.

Data collection, processing, and initial analysis were performed using beamline software BsxCuBE. Further analyses were performed with the ATSAS suite. PRIMUS (42) was used for R_g determination with the Guinier method, and maximum distance (D_{max}) was evaluated using GNOM (43), which was also used to calculate the distance distribution functions. Fitting of the model of the VipD–Rab5c structure to the SAXS data was calculated with CRYSOL (44) with a χ^2 against raw data of 2.12 and 3.2 for samples at 2.2 and 6.4 mg/mL, respectively. To generate an ab initio model of the VipD–Rab5c complex, 20 runs of GASBOR (45) were performed using the merge of the two datasets (2 and 6.4 mg/mL) as raw data. Then, the most probable model was filtered with DAMSEL (46), and a 720 bead model was produced. Superposition of the bead model on the crystallographic VipD–Rab5c structure was carried out using the program SUPCOMB13 (47). The resulting bead model was converted to a mesh envelope and visualized using PYMOL (Schrödinger).

Molecular Dynamics Simulations. A total of four molecular dynamics (MD) simulations were performed starting from the Rab5–VipD crystallographic structure and from three different Rab5 complexes previously described (Rab5–EEA1 C2H2 Zinc Finger, Rab5-Rabaptin5, and Rab5-Rabenosyn5). The initial coordinates of Rab5–EEA1 C2H2 Zinc Finger (PDB ID code 3MJH) and Rab5–Rabaptin5 (PDB ID code 1TU3) were taken from the Protein Data Bank, whereas the Rab5–Rabenosyn5 complex was rebuilt using the Rab22–Rabenosyn5 crystal structure (PDB ID code 1Z0J) as a template. In Rab5–VipD, Rab5–Rabaptin5, and Rab5–Rabenosyn5 structures, GppNHp molecules were replaced by GTP. In case of incomplete chains, acetyl and amide capping groups were added to the N-term and C-term residues flanking the mission regions to avoid improper charges on them. The protonation state of the ionizable residues was estimated at pH 6.5 using the server H++ (http://biophysics.cs.vt.edu/H++) (48–50). The parameter files for the GTP molecule and Zn^{2+} ion were prepared with the AMBER (51) module ANTECHAMBER, and the topology files for the protein complexes were generated using LEAP. Before running the molecular dynamics simulations, a short minimization and a five-step equilibration protocol were performed on the solvated structure, as previously described (52). On Rab5–VipD, Rab5–Rabaptin5, and Rab5–Rabenosyn5 complexes, unrestrained 10-ns MD simulations were performed in an isothermal-isobaric ensemble, setting pressure to 1 atm and temperature to 300 K. In the Rab5–EEA1 complex, 2.5-Å distance restraints between Zn^{2+} ion and each EEA1 zinc finger motif residue (Cys43, Cys46, His59, and His64) were applied during all of the equilibration and MD simulation step, to keep the same coordination as in the initial structure. The RMSD for the Cα atoms of each complex along the MD trajectory were calculated with the ptraj AMBER12 tool (51).

Computational Alanine Scanning. We used the MMPBSA.py script in AMBER12 (51) to perform Computational Alanine Scanning calculations on 200 snapshots extracted from the last 2 ns of each complex MD trajectory (see above). All of the interface residues (defined as those located within 4-Å distance from the protein partner in the most representative structure along the last 2 ns of the trajectory) were mutated to alanine, and then the binding free energy change (ΔΔG) was estimated as the difference between the binding ΔG (MM-GBSA method) of the WT and that of the mutated complex. The contribution of conformational entropy was not included here, given the difficulty of computing it for a large protein– protein complex but that should not significantly affect the comparison of mutant and WT free energies.

Surface Plasmon Resonance Measurements. The binding affinity of VipD, Rabenosyn5$_{1–70}$, EEA1$_{36–91}$, and Rabaptin5$_{739–862}$ for Rab5c(Q80L) or Rab5c(Q80L, R92A) was calculated using SPR. SPR data were collected using a Biacore 3000 instrument (GE Healthcare) and a GST sensor chip. A research grade CM5 chip was first conditioned with three 5-μL injections of 100 mM glycine-NaOH, pH 12. Anti-GST antibody was covalently immobilized on the CM5 sensor chip injecting 45 μL at 30 μg/mL in 10 mM sodium acetate, pH 5.0, using the amine coupling kit [1-ethyl-3-(3-dimethylaminopropyl)carbodiimide hydrochloride and *N*-hydrosuccinimide] supplied by GE Healthcare. Nearly 30,000 resonance units (RUs) of the antibody were immobilized under these conditions in each flow cell, where 1 RU corresponds to immobilized protein concentration of ∼1 pg/mm^2. The unreacted moieties on the surface were blocked with ethanolamine. The immobilization procedure was done at 5 μL/min with the running buffer containing 10 mM Hepes, pH 7.5, 150 mM NaCl, and 0.005% Tween20. Binding experiments were performed with the same buffer supplemented with 2 mM MgCl$_2$. All of the proteins were dialyzed into this buffer. GST, GST-VipD, GST-Rabenosyn5$_{1–70}$, GST-EEA1$_{36–91}$, and GST-Rabaptin5$_{739–862}$ were captured on the sensor chip with a 5-μL injection of 100 nM ligand at 5 μL/min on a reference and sample flow cell. Rab5c(Q80L) and Rab5c(Q80L, R92A) incubated with GppNHp were injected at different concentrations for a contact time of 2 min. Binding experiments were carried out at a flow rate of 20 μL/min at 25 °C. The anti-GST sensor chip was regenerated after each analyte injection with a 2-min injection of 10 mM glycine-HCl, pH 2.1. This regeneration procedure did not alter to any measurable extent the ability of the immobilized antibody to bind protein in subsequent cycles. Analysis of the data was performed using the BIAevaluation software supplied with the instrument. The steady-state binding response was determined by averaging the response over 5 s at the end of the injection and was corrected for background binding. The saturation binding values were fitted according to a one-site binding model. Each experiment was repeated in triplicate. Values of K_D are reported as the means of independent experiments with corresponding SDs.

ACKNOWLEDGMENTS. We thank A. Vidaurrazaga for technical assistance. We thank all the staff from the facilities where work was done for technical and human support. This work was supported by Carlos III Health Institute

Grant PI11/00121 and Basque Government Grant PI2011-26 (to A.H.). This study made use of the Diamond Light Source (Oxfordshire, United Kingdom), partly funded by the BioStruct-X program (Proposal 2370), the European Synchrotron Radiation Facility (Grenoble, France) under the Block Allocation Group MX1283/MX1420, and ALBA synchrotron beamline BL13-XALOC under Proposal 2012100351.

1. Christie PJ, Whitaker N, González-Rivera C (2014) Mechanism and structure of the bacterial type IV secretion systems. *Biochim Biophys Acta* 1843(8):1578–1591.
2. Fronzes R, et al. (2009) Structure of a type IV secretion system core complex. *Science* 323(5911):266–268.
3. Voth DE, Broederdorf LJ, Graham JG (2012) Bacterial Type IV secretion systems: Versatile virulence machines. *Future Microbiol* 7(2):241–257.
4. Ribet D, Cossart P (2010) Pathogen-mediated posttranslational modifications: A re-emerging field. *Cell* 143(5):694–702.
5. Galán JE (2009) Common themes in the design and function of bacterial effectors. *Cell Host Microbe* 5(6):571–579.
6. Christen M, et al. (2009) Activation of a bacterial virulence protein by the GTPase RhoA. *Sci Signal* 2(95):ra71.
7. LaRock DL, Brzovic PS, Levin I, Blanc MP, Miller SI (2012) A Salmonella typhimurium-translocated glycerophospholipid:cholesterol acyltransferase promotes virulence by binding to the RhoA protein switch regions. *J Biol Chem* 287(35):29654–29663.
8. Ohlson MB, et al. (2008) Structure and function of Salmonella SifA indicate that its interactions with SKIP, SseJ, and RhoA family GTPases induce endosomal tubulation. *Cell Host Microbe* 4(5):434–446.
9. Anderson DM, et al. (2011) Ubiquitin and ubiquitin-modified proteins activate the Pseudomonas aeruginosa T3SS cytotoxin, ExoU. *Mol Microbiol* 82(6):1454–1467.
10. Juris SJ, Rudolph AE, Huddler D, Orth K, Dixon JE (2000) A distinctive role for the Yersinia protein kinase: Actin binding, kinase activation, and cytoskeleton disruption. *Proc Natl Acad Sci USA* 97(17):9431–9436.
11. Shohdy N, Efe JA, Emr SD, Shuman HA (2005) Pathogen effector protein screening in yeast identifies Legionella factors that interfere with membrane trafficking. *Proc Natl Acad Sci USA* 102(13):4866–4871.
12. VanRheenen SM, Luo ZQ, O'Connor T, Isberg RR (2006) Members of a Legionella pneumophila family of proteins with ExoU (phospholipase A) active sites are translocated to target cells. *Infect Immun* 74(6):3597–3606.
13. Ku B, et al. (2012) VipD of Legionella pneumophila targets activated Rab5 and Rab22 to interfere with endosomal trafficking in macrophages. *PLoS Pathog* 8(12):e1003082.
14. Gaspar AH, Machner MP (2014) VipD is a Rab5-activated phospholipase A1 that protects Legionella pneumophila from endosomal fusion. *Proc Natl Acad Sci USA* 111(12):4560–4565.
15. Ensminger AW, Isberg RR (2009) Legionella pneumophila Dot/Icm translocated substrates: A sum of parts. *Curr Opin Microbiol* 12(1):67–73.
16. Horwitz MA (1983) The Legionnaires' disease bacterium (Legionella pneumophila) inhibits phagosome-lysosome fusion in human monocytes. *J Exp Med* 158(6): 2108–2126.
17. Vieira OV, Botelho RJ, Grinstein S (2002) Phagosome maturation: Aging gracefully. *Biochem J* 366(Pt 3):689–704.
18. Merithew E, et al. (2001) Structural plasticity of an invariant hydrophobic triad in the switch regions of Rab GTPases is a determinant of effector recognition. *J Biol Chem* 276(17):13982–13988.
19. Zhu G, et al. (2004) Structural basis of Rab5-Rabaptin5 interaction in endocytosis. *Nat Struct Mol Biol* 11(10):975–983.
20. Zhu W, Hammad LA, Hsu F, Mao Y, Luo ZQ (2013) Induction of caspase 3 activation by multiple Legionella pneumophila Dot/Icm substrates. *Cell Microbiol* 15(11):1783–1795.
21. Eathiraj S, Pan X, Ritacco C, Lambright DG (2005) Structural basis of family-wide Rab GTPase recognition by rabenosyn-5. *Nature* 436(7049):415–419.
22. Mishra A, Eathiraj S, Corvera S, Lambright DG (2010) Structural basis for Rab GTPase recognition and endosome tethering by the C2H2 zinc finger of Early Endosomal Autoantigen 1 (EEA1). *Proc Natl Acad Sci USA* 107(24):10866–10871.
23. Finn RD, et al. (2014) Pfam: The protein families database. *Nucleic Acids Res* 42(Database issue):D222–D230.
24. Dennis EA, Cao J, Hsu YH, Magrioti V, Kokotos G (2011) Phospholipase A2 enzymes: Physical structure, biological function, disease implication, chemical inhibition, and therapeutic intervention. *Chem Rev* 111(10):6130–6185.
25. Dessen A, et al. (1999) Crystal structure of human cytosolic phospholipase A2 reveals a novel topology and catalytic mechanism. *Cell* 97(3):349–360.
26. de Felipe KS, et al. (2005) Evidence for acquisition of Legionella type IV secretion substrates via interdomain horizontal gene transfer. *J Bacteriol* 187(22):7716–7726.
27. Al-Quadan T, Price CT, Abu Kwaik Y (2012) Exploitation of evolutionarily conserved amoeba and mammalian processes by Legionella. *Trends Microbiol* 20(6):299–306.
28. Brzozowski AM, et al. (1991) A model for interfacial activation in lipases from the structure of a fungal lipase-inhibitor complex. *Nature* 351(6326):491–494.
29. van Tilbeurgh H, et al. (1993) Interfacial activation of the lipase-procolipase complex by mixed micelles revealed by X-ray crystallography. *Nature* 362(6423):814–820.
30. Roussel A, et al. (2002) Crystal structure of the open form of dog gastric lipase in complex with a phosphonate inhibitor. *J Biol Chem* 277(3):2266–2274.
31. Aloulou A, et al. (2006) Exploring the specific features of interfacial enzymology based on lipase studies. *Biochim Biophys Acta* 1761(9):995–1013.
32. Qin S, Pande AH, Nemec KN, Tatulian SA (2004) The N-terminal alpha-helix of pancreatic phospholipase A2 determines productive-mode orientation of the enzyme at the membrane surface. *J Mol Biol* 344(1):71–89.
33. Qin S, Pande AH, Nemec KN, He X, Tatulian SA (2005) Evidence for the regulatory role of the N-terminal helix of secretory phospholipase A(2) from studies on native and chimeric proteins. *J Biol Chem* 280(44):36773–36783.
34. Gendrin C, et al. (2012) Structural basis of cytotoxicity mediated by the type III secretion toxin ExoU from Pseudomonas aeruginosa. *PLoS Pathog* 8(4):e1002637.
35. Finck-Barbançon V, Frank DW (2001) Multiple domains are required for the toxic activity of Pseudomonas aeruginosa ExoU. *J Bacteriol* 183(14):4330–4344.
36. Sheffield P, Garrard S, Derewenda Z (1999) Overcoming expression and purification problems of RhoGDI using a family of "parallel" expression vectors. *Protein Expr Purif* 15(1):34–39.
37. Kabsch W (2010) Integration, scaling, space-group assignment and post-refinement. *Acta Crystallogr D Biol Crystallogr* 66(Pt 2):133–144.
38. McCoy AJ, et al. (2007) Phaser crystallographic software. *J Appl Cryst* 40(Pt 4):658–674.
39. Adams PD, et al. (2010) PHENIX: A comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr D Biol Crystallogr* 66(Pt 2):213–221.
40. Emsley P, Lohkamp B, Scott WG, Cowtan K (2010) Features and development of Coot. *Acta Crystallogr D Biol Crystallogr* 66(Pt 4):486–501.
41. Lovell SC, et al. (2003) Structure validation by Calpha geometry: Phi,psi and Cbeta deviation. *Proteins* 50(3):437–450.
42. Konarev PV, Volkov VV, Sokolova AV, Koch MHJ, Svergun DI (2003) PRIMUS: A Windows PC-based system for small-angle scattering data analysis. *J Appl Cryst* 36(5): 1277–1282.
43. Svergun D (1992) Determination of the regularization parameter in indirect-transform methods using perceptual criteria. *J Appl Cryst* 25(4):495–503.
44. Svergun D, Barberato C, Koch MHJ (1995) CRYSOL: A program to evaluate X-ray solution scattering of biological macromolecules from atomic coordinates. *J Appl Cryst* 28(6):768–773.
45. Svergun DI, Petoukhov MV, Koch MH (2001) Determination of domain structure of proteins from X-ray solution scattering. *Biophys J* 80(6):2946–2953.
46. Volkov VV, Svergun DI (2003) Uniqueness of ab initio shape determination in small-angle scattering. *J Appl Cryst* 36(3 Part 1):860–864.
47. Kozin MB, Svergun DI (2001) Automated matching of high- and low-resolution structural models. *J Appl Cryst* 34(1):33–41.
48. Anandakrishnan R, Aguilar B, Onufriev AV (2012) H++ 3.0: Automating pK prediction and the preparation of biomolecular structures for atomistic molecular modeling and simulations. *Nucleic Acids Res* 40(Web Server issue):W537–541.
49. Myers J, Grothaus G, Narayanan S, Onufriev A (2006) A simple clustering algorithm can be accurate enough for use in calculations of pKs in macromolecules. *Proteins* 63(4):928–938.
50. Gordon JC, et al. (2005) H++: A server for estimating pKas and adding missing hydrogens to macromolecules. *Nucleic Acids Res* 33(Web Server issue):W368–W371.
51. Case DA, et al. (2012) AMBER 12 (University of California, San Francisco).
52. Meyer T, et al. (2010) MoDEL (Molecular Dynamics Extended Library): A database of atomistic molecular dynamics trajectories. *Structure* 18(11):1399–1409.

Supporting Information

Lucas et al. 10.1073/pnas.1405391111

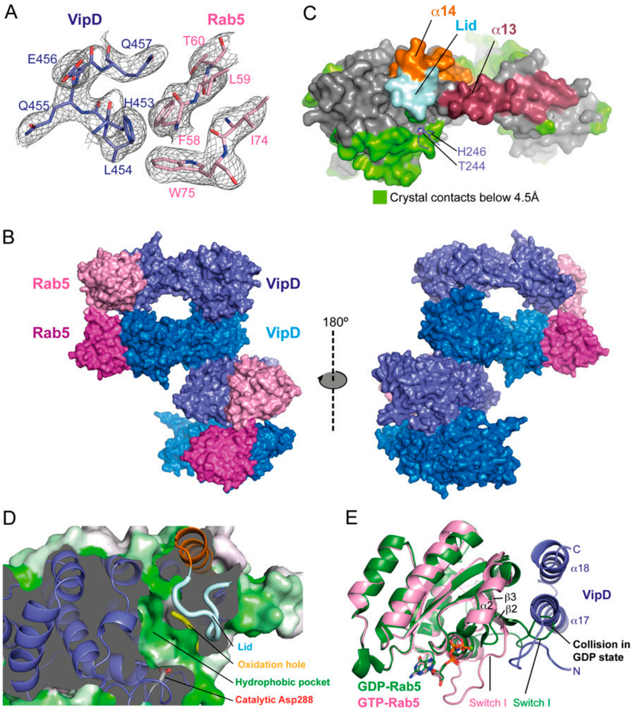


Fig. S1. (*A*) Electron density map (2Fo-Fc) calculated with phases derived from the final refined model and contoured at 1.5 sigma in the vicinity of the VipD-Rab5c interface. (*B*) Packing of four VipD–Rab5 complexes that form an asymmetric unit. VipD is colored in blue/slate; Rab5 is in light and dark pink. (*C*) Distribution of VipD surface areas with crystal lattice contacts below 4.5 Å (same orientation as in Fig. 2 *C–E*). Note that there are no mayor crystallographic contacts in areas corresponding to α13, α14, and the lid except for two residues (Thr244 and His246) from a symmetrically related VipD molecule that are partially inserted at the edge of the catalytic groove, most probably favored by the opening of the lid. The symmetrically related VipD molecule is omitted from the figure for reasons of clarity. (*D*) Close up view of the catalytic domain of VipD in ribbon diagram showing the surface clipped at the catalytic cleft. The surface is colored from white to green according to the Eisenberg hydrophobicity scale. (*E*) Superposition of the Rab5-GDP crystal structure [Protein Data Bank (PDB) ID code 1TU4] on the VipD:Rab5–GTP complex, illustrating the steric collision that prevents Rab-GDP from binding to VipD. The remainder of the VipD structure is omitted for clarity.

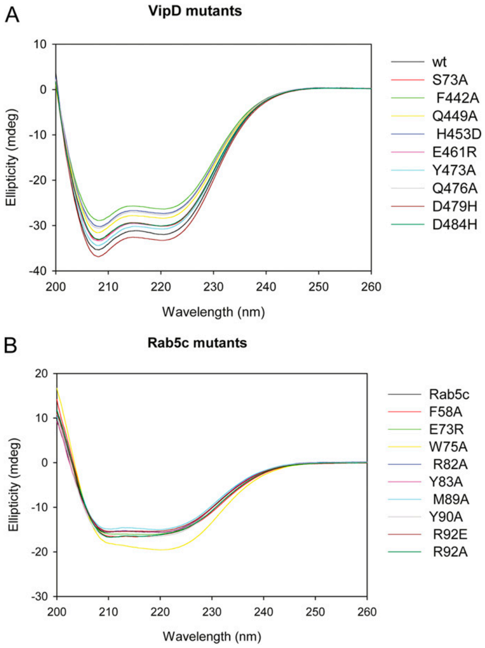


Fig. S2. Circular dichroism spectra of the the indicated wildtype and mutant proteins of (*A*) VipD and (*B*) Rab5c used in this study.

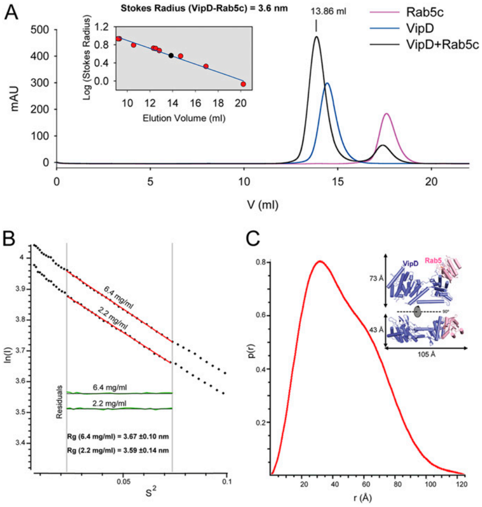


Fig. S3. Characterization of VipD–Rab5c complex in solution. (*A*) Analysis of the VipD$_{1–621}$–Rab5c$_{18–182}$ oligomeric state by gel filtration (Superdex 200 HR 10/30 column), leading to a Stokes radius estimation of ∼3.6 nm by comparison with the elution of model proteins (thyroglobulin, 8.5 nm; ferritin, 6.2 nm; γ-globulin, 5.3 nm; catalase, 5.2 nm; aldolase, 4.7 nm; BSA monomer, 3.5 nm; myoglobin, 2.1 nm; vitamin B12, 2.7 nm; *Inset*). Gel filtration of VipD (70 μM), Rab5c (140 μM), and VipD (70 μM) + Rab5c (140 μM). Note that the complex can be isolated using an excess of the Rab protein. (*B*) The guinier plot of the VipD$_{1–564}$–Rab5c$_{18–182}$ complex data at 2.2 and 6.4 mg/mL indicates a gyration radius of ∼3.6 nm. (*C*) This value is confirmed by the distance distribution function P(r), which suggests a bilobular structure with a radius of gyration of ∼36 Å and a maximum diameter of ∼120 Å fully compatible with the crystallographic structure of the VipD–Rab5c complex (*Inset*).

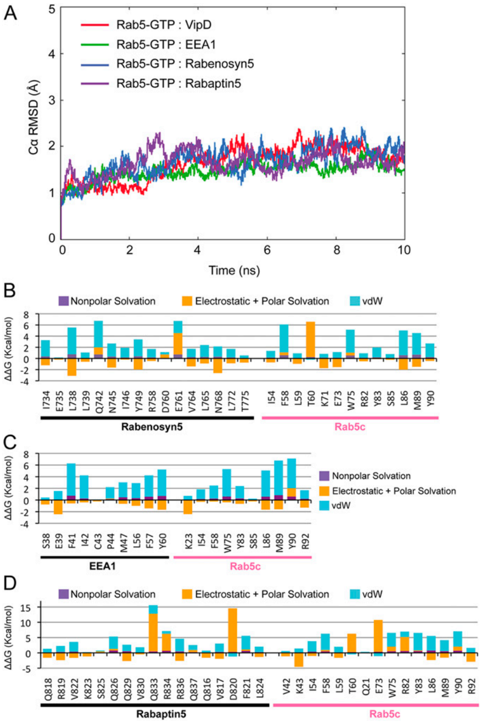


Fig. S4. (*A*) Plot of root mean square deviation (RMSD) relative to the coordinates of the initial (energy-minimized) structures during molecular dynamic simulations. (*B–D*) Per-residue contribution from van der Waals (vdW) energy (blue), nonpolar solvation energy (purple), and the sum of electrostatic and polar solvation energy (orange) calculated by computational alanine scanning for interfacial residues in the Rabenosyn5–Rab5c complex (*B*), EEA1–Rab5c complex (*C*), and Rabaptin5–Rab5c (*D*). Existing glycines and alanines are excluded in the calculation.

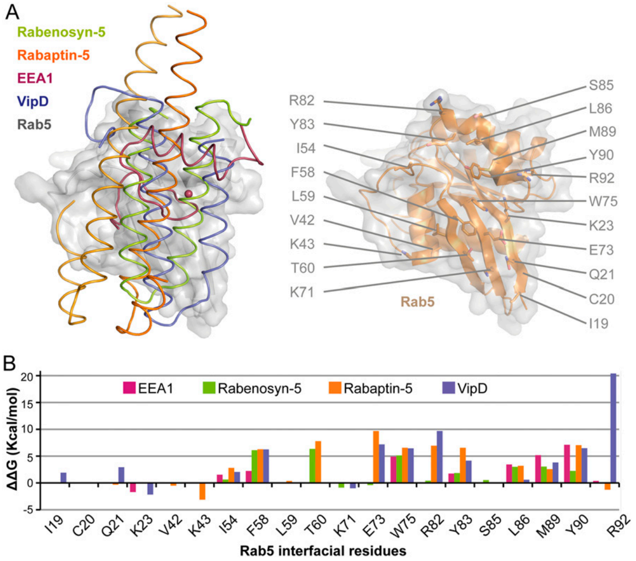


Fig. S5. (*A*) Superposition of effector binding epitopes in tube representation over Rab5c in transparent gray surface. EEA1 C2H2 Zinc Finger (PDB ID code 3MJH) in purple, Rabenosyn-5 (rebuilt from PDB ID code 1Z0J) in green, Rabaptin-5 (PDB ID code 1TU3) in orange, and VipD in slate. (*B*) Comparison of the ΔΔG values in the binding interface of Rab5c. Note that some residues at the preserved binding core show similar ΔΔG values, whereas other residues such as Arg92 constitute an effector-specific contact.

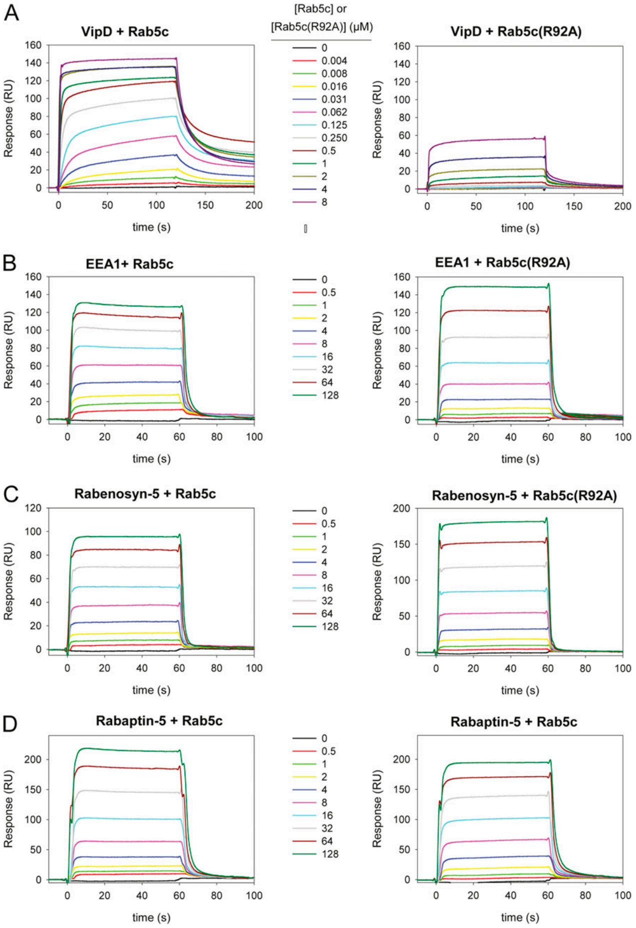
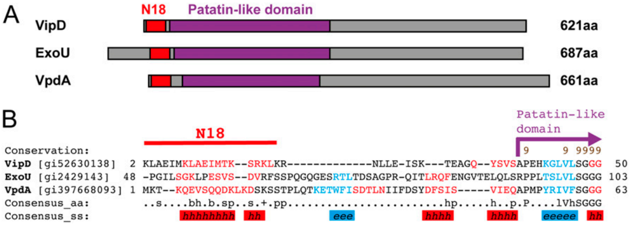


Fig. S6. Surface plasmon resonance (SPR) sensograms for binding of (*A*) VipD WT, (*B*) EEA1 C2H2 Zinc Finger [amino acids (aa) 36–91], (*C*) Rabenosyn-5 (aa 1–70), and (*D*) Rabaptin-5 (aa 739–862) to Rab5c$_{18–182}$(Q80L) (*Left*) or Rab5c$_{18–182}$(Q80L, R92A) (*Right*).



```
Conservation:                                                              9     9 9999
VipD [gi52630138]  2 KLAEIMKLAEIMTK--SRKLKR-------------NLLE-ISK---TEAGQ--YSVSAPEHKGLVLSGGG  50
ExoU [gi2429143]  48 -PGILSGKLPESVS--DVRFSSPQGQGESRTLTDSAGPR-QITLRQFENGVTELQLSRPPLTSLVLSGGG 103
VpdA [gi397668093] 1 MKT--KQEVSQQDKLKDSKSSTPLQTKETWFISDTLNIIFDSYDFSIS-----VIEQAPMPYRIVFSGGG  63
Consensus_aa:        ..s....bh.b.sp..s.+.pp.......................hp.....h..p.P....lVhSGGG
Consensus_ss:              hhhhhhhh  hh           eee           hhhh      hhhh    eeeee  hh
```

Fig. S7. The N-terminal tail of VipD shows structural similarity to equivalent regions in VpdA and ExoU. (*A*) Representative diagrams of VipD, VpdA, and ExoU highlighting the equivalent N-terminal tail. (*B*) PROMALS3D (PROfile Multiple Alignment with Predicted Local Structures and 3D Constraints) web server alignment (http://prodata.swmed.edu/promals3d/promals3d.php) of VipD, VpdA, and ExoU N-terminal regions.

Table S1. **Plasmids used in this study**

Plasmid	Properties	Reference
pGST-Parallel2-VipD(19-564)	Expression construct for N-terminal glutathione S-transferase (GST)-tagged *L. pneumophila* VipD domain covering residues 19–564 used for crystallization and PLA assays	This study
pGST-Parallel2-Rab5c$_{18–182}$ (Q80L)	Expression construct for N-terminal GST-tagged human Rab5c(Q80L) domain covering residues 18–182 used for crystallization, PLA assays and SPR assays	This study
pGST-Parallel2-VipD	Expression construct for N-terminal GST-tagged *L. pneumophila* VipD used for pull-down assays, PLA assays and SPR assays	This study
pGST-Parallel2-VipD(1-564)	Expression construct for N-terminal GST-tagged *L. pneumophila* VipD domain covering residues 1–564 used for PLA assays	This study
pGST-Parallel2-VipD(19-621)	Expression construct for N-terminal GST-tagged *L. pneumophila* VipD domain covering residues 19–621 used for PLA assays	This study
pGST-Parallel2-VipD(S73A)	Expression construct for N-terminal GST-tagged *L. pneumophila* VipD containing mutation S73A used for pull-down assays and PLA assays	This study
pGST-Parallel2-VipD(F442A)	Expression construct for N-terminal GST-tagged *L. pneumophila* VipD containing mutation F442A used for pull-down assays and PLA assays	This study
pGST-Parallel2-VipD(Q449A)	Expression construct for N-terminal GST-tagged *L. pneumophila* VipD containing mutation Q449A used for pull-down assays	This study
pGST-Parallel2-VipD(H453D)	Expression construct for N-terminal GST-tagged *L. pneumophila* VipD containing mutation H453D used for pull-down assays and PLA assays	This study
pGST-Parallel2-VipD(E461R)	Expression construct for N-terminal GST-tagged *L. pneumophila* VipD containing mutation E461R used for pull-down assays	This study
pGST-Parallel2-VipD(Y473A)	Expression construct for N-terminal GST-tagged *L. pneumophila* VipD containing mutation Y473A used for pull-down assays	This study
pGST-Parallel2-VipD(Q476A)	Expression construct for N-terminal GST-tagged *L. pneumophila* VipD containing mutation Q476A used for pull-down assays	This study
pGST-Parallel2-VipD(D479H)	Expression construct for N-terminal GST-tagged *L. pneumophila* VipD containing mutation D479H used for pull-down assays and PLA assays	This study
pGST-Parallel2-VipD(D484H)	Expression construct for N-terminal GST-tagged *L. pneumophila* VipD containing mutation D484H used for pull-down assays	This study
pGST-Parallel2-Rab5c$_{18–182}$	Expression construct for N-terminal GST-tagged human Rab5c(18-182) used for PLA assays	This study
pGST-Parallel2-Rab5c$_{18–182}$ (Q80L,F58A)	Expression construct for N-terminal GST-tagged human Rab5c(18-182) containing mutations Q80L and F58A used for pull-down assays and PLA assays	This study
pGST-Parallel2-Rab5c$_{18–182}$ (Q80L,E73R)	Expression construct for N-terminal GST-tagged human Rab5c(18-182) containing mutations Q80L and E73R used for pull-down assays	This study
pGST-Parallel2-Rab5c$_{18–182}$ (Q80L,W75A)	Expression construct for N-terminal GST-tagged human Rab5c(18-182) containing mutations Q80L and W75A used for pull-down assays	This study
pGST-Parallel2-Rab5c$_{18–182}$ (Q80L, R82A)	Expression construct for N-terminal GST-tagged human Rab5c(18-182) containing mutations Q80L and R82A used for pull-down assays	This study
pGST-Parallel2-Rab5c$_{18–182}$ (Q80L,Y83A)	Expression construct for N-terminal GST-tagged human Rab5c(18-182) containing mutations Q80L and Y83A used for pull-down assays	This study
pGST-Parallel2-Rab5c$_{18–182}$ (Q80L,M89A)	Expression construct for N-terminal GST-tagged human Rab5c(18-182) containing mutations Q80L and M89A used for pull-down assays	This study
pGST-Parallel2-Rab5c$_{18–182}$ (Q80L,Y90A)	Expression construct for N-terminal GST-tagged human Rab5c(18-182) containing mutations Q80L and Y90A used for pull-down assays	This study
pGST-Parallel2-Rab5c$_{18–182}$ (Q80L,R92E)	Expression construct for N-terminal GST-tagged human Rab5c(18-182) containing mutations Q80L and R92E used for pull-down assays	This study
pGST-Parallel2-Rab5c$_{18–182}$ (Q80L,R92A)	Expression construct for N-terminal GST-tagged human Rab5c(18-182) containing mutations Q80L and R92A used in SPR assays	This study

Table S1. Cont.

Plasmid	Properties	Reference
pGST-Parallel2-Rab22a$_{16–181}$ (Q64L)	Expression construct for N-terminal GST-tagged human Rab22a(16-181) containing mutation Q64L used in PLA and pull-down assays	This study
pGST-Parallel2-Rab7a(Q67L)	Expression construct for N-terminal GST-tagged human Rab7a containing mutation Q67L used in PLA assays	This study
pGST-Parallel2-EEA1$_{36–91}$	Expression construct for N-terminal GST-tagged human Rabenosyn5 domain covering residues 36–91 used in SPR assays	This study
pGST-Parallel2-Rabenosyn5$_{1–70}$	Expression construct for N-terminal GST-tagged human Rabenosyn5 domain covering residues 1–70 used in SPR assays	This study
pGST-Parallel2-Rabaptin5$_{739–862}$	Expression construct for N-terminal GST-tagged human Rabaptin5 domain covering residues 739–862 used in SPR assays	This study
pGST-Parallel2	Expression construct for GST used in SPR assays	(1)
pmCherry-C1	Mammalian expression vector generating an mCherry fusion to the N terminus of the protein of interest	Clontech (cat. 632524)
pmCherry-VipD	Expression construct generating an mCherry fusion to the N terminus of *L. pneumophila* full-length VipD	(2)
pmCherry-VipD(F442A)	pmCherry-VipD containing mutation F442A in the VipD-Rab5 interface	This study
pmCherry-VipD(H453D)	pmCherry-VipD containing mutation H453D in the VipD-Rab5 interface	This study
pmCherry-VipD(E461R)	pmCherry-VipD containing mutation E461R in the VipD-Rab5 interface	This study
pmCherry-VipD(Q476A)	pmCherry-VipD containing mutation Q476A in the VipD-Rab5 interface	This study
pEGFP-Rab5a(Q79L)	Expression construct generating a GFP fusion to the N terminus of human full-length Rab5a containing mutation Q79L generating constitutively active Rab5a	(3)

1. Sheffield P, Garrard S, Derewenda Z (1999) Overcoming expression and purification problems of RhoGDI using a family of “parallel” expression vectors. *Protein Expr Purif* 15(1):34–39.
2. Gaspar AH, Machner MP (2014) VipD is a Rab5-activated phospholipase A1 that protects Legionella pneumophila from endosomal fusion. *Proc Natl Acad Sci USA* 111(12):4560–4565.
3. Mattera R, Bonifacino JS (2008) Ubiquitin binding and conjugation regulate the recruitment of Rabex-5 to early endosomes. *EMBO J* 27(19):2484–2494.

Table S2. Oligonucleotides used in this study

Sequence	Plasmid	Cloning
TTTTTGGATCCGAAATATCAAAGACTGAGGCAGGACAATATTCTG TTTTTCTCGAGTCACGGTTCAGGTTGAACTTCAACTTTAAAGTCTTG	pGST-Parallel2-VipD(19-564)	BamHI/XhoI
TTTTTGGATCCAAGATCTGTCAATTTAAGCTGGTTCTGCTGGGGG TTTTTCTCGAGTCAAAGCTTCTTAGCTATTGCCATGAAGATTTCGTTCACG	pGST-Parallel2-Rab5c$_{18–182}$(Q80L)	BamHI/XhoI
TTTTTGGATCCATGAAACTTGCTGAAATTATGACAAAAAGCCGTAAATTAAAAAG TTTTTCTCGAGTCAATGGCCGCCAAATGTGGTTGAAAGAC	pGST-Parallel2-VipD	BamHI/XhoI
AATCTGACCCATGTTAGCGGAGCAGCAGCCGGAGCAATGACGGCGAGTAT ATACTCGCCGTCATTGCTCCGGCTGCTGCTCCGCTAACATGGGTCAGATT	pGST-Parallel2-VipD(S73A)	SD
GAGAAAGAGATTGCTGAGGCATCAGCGCATG CATGCGCTGATGCCTCAGCAATCTCTTTCTC	pGST-Parallel2-VipD(F442A)	SD
GAGATTTTTGAGGCATCAGCGCATGCAGCAGCTATTTTGCATCTTCAAGAACAAATCG CGATTTGTTCTTGAAGATGCAAAATAGCTGCTGCATGCGCTGATGCCTCAAAAATCTC	pGST-Parallel2-VipD(Q449A)	SD
GGCATCAGCGCATGCACAAGCTATTTTGGATCTTCAAGAACAAATCGTCAAAGAAATG CATTTCTTTGACGATTTGTTCTTGAAGATCCAAAATAGCTTGTGCATGCGCTGATGCC	pGST-Parallel2-VipD(H453D)	SD
GCTATTTTGCATCTTCAAGAACAAATCGTCAAACGAATGAATGATGGTGATTACAGTAG GCTACTGTAATCACCATCATTCATTCGTTTGACGATTTGTTCTTGAAGATGCAAAATAGC	pGST-Parallel2-VipD(E461R)	SD
GATGGTGATTACAGTAGCGTGCAAAATGCCTTAGATCAAATTGAAGACATTCTGACAG CTGTCAGAATGTCTTCAATTTGATCTAAGGCATTTTGCACGCTACTGTAATCACCATC	pGST-Parallel2-VipD(Y473A)	SD
GCGTGCAAAATTATTTAGATGCAATTGAAGACATTCTGAC GTCAGAATGTCTTCAATTGCATCTAAATAATTTTGCACGC	pGST-Parallel2-VipD(Q476A)	SD
CGTGCAAAATTATTTAGATCAAATTGAACACATTCTGACAGTCGATGCCAAAATGGATG CATCCATTTTGGCATCGACTGTCAGAATGTGTTCAATTTGATCTAAATAATTTTGCACG	pGST-Parallel2-VipD(D479H)	SD
AGATCAAATTGAAGACATTCTGACAGTCCATGCCAAAATGGATGACATCCAGAAAGAG CTCTTTCTGGATGTCATCCATTTTGGCATGGACTGTCAGAATGTCTTCAATTTGATCT	pGST-Parallel2-VipD(D484H)	SD
CAATTGGAGCGGCCGCACTCACACAGACTGTC GACAGTCTGTGTGAGTGCGGCCGCTCCAATTG	pGST-Parallel2-Rab5c$_{18–182}$(Q80L,F58A)	SD
CAACAGTCAAGTTTCGGATCTGGGACACAGC GCTGTGTCCCAGATCCGAAACTTGACTGTTG	pGST-Parallel2-Rab5c$_{18–182}$ (Q80L,E73R)	SD
CAAGTTTGAGATCGCGGACACAGCTGGACAG CTGTCCAGCTGTGTCCGCGATCTCAAACTTG	pGST-Parallel2-Rab5c$_{18–182}$(Q80L,W75A)	SD
GACACAGCTGGACTGGAGGCATATCACAGCCTGGC GCCAGGCTGTGATATGCCTCCAGTCCAGCTGTGTC	pGST-Parallel2-Rab5c$_{18–182}$(Q80L, R82A)	SD
GACACAGCTGGACTAGAGCGGGCTCACAGCCTGGCCCCCATG CATGGGGGCCAGGCTGTGAGCCCGCTCTAGTCCAGCTGTGTC	pGST-Parallel2-Rab5c$_{18–182}$(Q80L,Y83A)	SD
GAGCGGTATCACAGCCTGGCCCCCGCATACTATCGGGGGGCCCAGGC GCCTGGGCCCCCCGATAGTATGCGGGGGCCAGGCTGTGATACCGCTC	pGST-Parallel2-Rab5c$_{18–182}$(Q80L,M89A)	SD
GTATCACAGCCTGGCCCCCATGGCATATCGGGGGGCCCAGGCTGCC GGCAGCCTGGGCCCCCCGATATGCCATGGGGGCCAGGCTGTGATAC	pGST-Parallel2-Rab5c$_{18–182}$(Q80L,Y90A)	SD
TATCACAGCCTGGCCCCCATGTACTATGAAGGGGCCCAGGCTGCCATCGTGGTCTAT ATAGACCACGATGGCAGCCTGGGCCCCTTCATAGTACATGGGGGCCAGGCTGTGATA	pGST-Parallel2-Rab5c$_{18–182}$ (Q80L,R92E)	SD
TATCACAGCCTGGCCCCCATGTACTATGCGGGGGCCCAGGCTGCCATCGTGGTCTAT ATAGACCACGATGGCAGCCTGGGCCCCCGCATAGTACATGGGGGCCAGGCTGTGATA	pGST-Parallel2-Rab5c$_{18–182}$ (Q80L,R92A)	SD
TTTTTGGATCCGCGCTGAGGGAACTTAAAGTGTGCCTG TTTTTCTCGAGTCAGGATGGAATTCTTCGACTAATTTCTATAAAGAGTTC	pGST-Parallel2-Rab22a$_{16–181}$ (Q64L)	BamHI/XhoI
AAAAAACCATGGGAATGACCTCTAGGAAGAAAGTGTTGCTGAAGG AAAAAACTCGAGTTAGCAACTGCAGCTTTCTGCCGAGGCC	pGST-Parallel2-Rab7a(Q67L)	NcoI/XhoI
TTTTTGGATCCAGCTCTTCAGAGGGTTTCATATGTC TTTTTCTCGAGTTACTCTTGTCTGAGCAGTGTTACATC	pGST-Parallel2-EEA1$_{36–91}$	BamHI/XhoI
TTTTTGGATCCATGGCTTCTCTGGACGACCCAG TTTTTCTCGAGTTATGCTCGATCATCCCCTTCTCGT	pGST-Parallel2-Rabenosyn5$_{1–70}$	BamHI/XhoI
TTTTTCCATGGCTTCTATTTCTAGCCTAAAAGCTGAATTAG TTTTTCTCGAGTCATGTCTCAGGAAGCTGGTTAATG	pGST-Parallel2-Rabaptin5$_{739–862}$	NcoI/XhoI
AAAAGTCGACATGAAACTTGCTGAAATTATGACAAAAAGC AAGGATCCTTAATGGCCGCCAAATGTGGTTGAAAGAC	pmCherry-VipD	SalI/BamHI
GAAATCAGAGAAAGAGATTGCTGAGGCATCAGCGCATGCAC GTGCATGCGCTGATGCCTCAGCAATCTCTTTCTCTGATTTC	pmCherry-VipD(F442A)	SD
CATGCACAAGCTATTTTGGATCTTCAAGAACAAATCG CGATTTGTTCTTGAAGATCCAAAATAGCTTGTGCATG	pmCherry-VipD(H453D)	SD
CAAATCGTCAAAGAAATGAATCGTGGTGATTACAGTAGCGTG CACGCTACTGTAATCACCACGATTCATTTCTTTGACGATTTG	pmCherry-VipD(E461R)	SD
GTGCAAAATTATTTAGATGCAATTGAAGACATTCTGAC GTCAGAATGTCTTCAATTGCATCTAAATAATTTTGCAC	pmCherry-VipD(Q476A)	SD

SD, site-directed mutagenesis.

3.3.2. Interaction of photosystem I from *Phaeodactylum tricornutum* with plastocyanins as compared with its native cytochrome c_6: reunion with a lost donor.

Pilar Bernal-Bayard[1], Chiara Pallara[2], M. Carmen Castell[1], Fernando P. Molina-Heredia[1], Juan Fernández-Recio[2], [1]Manuel Hervás[1] and José A. Navarro[1]*

[1]Instituto de Bioquímica Vegetal y Fotosíntesis, Universidad de Sevilla and CSIC, Américo Vespucio 49, 41092-Sevilla, Spain

[2]Joint BSC-CRG-IRB Research Program in Computational Biology, Life Department, Barcelona Supercomputing Center, Barcelona 08034, Spain

*Corresponding author

Biochimica et Biophysica Acta 1847 (2015) 1549–1559

Contents lists available at ScienceDirect

Biochimica et Biophysica Acta

journal homepage: www.elsevier.com/locate/bbabio

Interaction of photosystem I from *Phaeodactylum tricornutum* with plastocyanins as compared with its native cytochrome c_6: Reunion with a lost donor

Pilar Bernal-Bayard [a], Chiara Pallara [b], M. Carmen Castell [a], Fernando P. Molina-Heredia [a],
Juan Fernández-Recio [b], Manuel Hervás [a], José A. Navarro [a],*

[a] *Instituto de Bioquímica Vegetal y Fotosíntesis, Universidad de Sevilla and CSIC, Américo Vespucio 49, 41092 Sevilla, Spain*
[b] *Joint BSC–CRG–IRB Research Program in Computational Biology, Life Sciences Department, Barcelona Supercomputing Center, Barcelona 08034, Spain*

ARTICLE INFO

Article history:
Received 16 June 2015
Received in revised form 2 September 2015
Accepted 20 September 2015
Available online 25 September 2015

Keywords:
Cytochrome c_6
Electron transfer
Computational docking
Laser flash photolysis
Phaeodactylum
Photosystem I
Plastocyanin

ABSTRACT

In the *Phaeodactylum tricornutum* alga, as in most diatoms, cytochrome c_6 is the only electron donor to photosystem I, and thus they lack plastocyanin as an alternative electron carrier. We have investigated, by using laser-flash absorption spectroscopy, the electron transfer to *Phaeodactylum* photosystem I from plastocyanins from cyanobacteria, green algae and plants, as compared with its own cytochrome c_6. Diatom photosystem I is able to effectively react with eukaryotic acidic plastocyanins, although with less efficiency than with *Phaeodactylum* cytochrome c_6. This efficiency, however, increases in some green alga plastocyanin mutants mimicking the electrostatics of the interaction site on the diatom cytochrome. In addition, the structure of the transient electron transfer complex between cytochrome c_6 and photosystem I from *Phaeodactylum* has been analyzed by computational docking and compared to that of green lineage and mixed systems. Taking together, the results explain why the *Phaeodactylum* system shows a lower efficiency than the green systems, both in the formation of the properly arranged [cytochrome c_6-photosystem I] complex and in the electron transfer itself.

1. Introduction

Diatoms are unicellular photosynthetic eukaryotes that are estimated to contribute to 30–40% of the global carbon fixation in the oceans, and thus can be considered major primary producers [1,2]. Diatoms have a complex evolutionary history, belonging to the red lineage of alga that diverged along evolution from the green lineage that led to plants [3,4]. Consequently, the photosynthetic machinery in diatoms possesses some singularities, arising from their evolutionary history and endosymbiotic origin [5–7]. In particular, while most cyanobacteria and unicellular green algae contain both the copper-protein plastocyanin (Pc) and the iron-containing cytochrome c_6 (Cyt) as alternative soluble electron carriers between the b_6f and photosystem I (PSI) membrane complexes, most diatoms lack Pc, thus containing Cyt as the only soluble carrier between these complexes [8], with the remarkable exception of the oceanic centric diatom *Thalassiosira oceanica*, for which the presence of an unusual Pc has been reported [9].

In green algae and plants (the green lineage of photosynthetic eukaryotes), the very acidic donor proteins to PSI interact, by means of strong attractive electrostatic interactions, with a well-conserved positively-charged docking site located in an extra loop extension of the PsaF subunit in PSI [10–12]. However, PSI from cyanobacteria, where both Pc and Cyt can be acidic, neutral or basic, lacks this extra loop, and thus the role of electrostatic forces in the interaction with PSI varies consequently [13–16]. By its turn, although the evolution of the electron transfer (ET) to PSI in diatoms has also led to complementary electrostatic interactions between acidic and basic patches in Cyt and the PsaF subunit, respectively, the electrostatic character of both partners is similarly reduced, the intensity of the interaction being accordingly weakened as compared with the strongest electrostatic properties of the Cyt(Pc)/PSI complex in the green lineage [17].

Although the ET from Pc and Cyt to PSI usually follows similar mechanisms in the same organism, electron donation to PSI has increased in complexity and efficiency in eukaryotic cells as compared with prokaryotic cyanobacteria [13,18–20]. Recently, the ET reaction mechanism from Cyt to PSI from the diatom *Phaeodactylum tricornutum* has been first analyzed by laser-flash absorption spectroscopy [17], indicating that ET occurs within a Cyt/PSI transient complex that undergoes a

Abbreviations: β-DM, β-dodecyl-maltoside; Cyt, cytochrome c_6; ET, electron transfer; k_2, second-order rate constant; k_2^{HI}, second-order rate constant at high ionic strength; K_A, equilibrium constant for the complex formation reaction; k_{ET}, first-order electron transfer rate constant; k_{ON} and k_{OFF}, association and dissociation rate constants, respectively, for complex formation; k_{OBS}, observed pseudo first-order rate constant; k_{SAT}, first-order limiting rate constant at infinite protein concentration; Pc, plastocyanin; PSI, photosystem I

* Corresponding author at: Instituto de Bioquímica Vegetal y Fotosíntesis, Centro de Investigaciones Científicas Isla de la Cartuja, Universidad de Sevilla & CSIC, Américo Vespucio 49, 41092 Sevilla, Spain.

E-mail address: jnavarro@ibvf.csic.es (J.A. Navarro).

http://dx.doi.org/10.1016/j.bbabio.2015.09.006
0005-2728/© 2015 Elsevier B.V. All rights reserved.

reorganization process from the initial encounter complex to the optimized final configuration, as already described in "green" PSI systems. However, the results also demonstrated that the "red" *Phaeodactylum* system possesses a lower efficiency than "green" systems, both in the formation of the properly arranged [Cyt–PSI] complex and in the electron transfer itself [17]. In addition, the relatively weak electrostatically attractive nature of the Cyt/PSI interaction seems to represent a compromise between the efficiency in the ET process and the need for a fast exchange of the protein donor [17].

In this work, the structure of the transient electron transfer complex between Cyt and PSI from the diatom *P. tricornutum* has been modeled by computational docking, and compared with that from green systems. Moreover, we have studied the cross-reactions between *Phaeodactylum* PSI and different prokaryotic and eukaryotic Pcs — including some mutant variants — in order to obtain relevant data on the differences and similarities of the diatom couple with respect to other well characterized systems, and on the evolution of the reaction mechanism in the different branches of photosynthetic organisms.

2. Experimental procedures

2.1. Protein purification

PSI particles from the diatom *P. tricornutum* were obtained by β-dodecyl-maltoside (β-DM) solubilization as previously described [17]. Pcs from the cyanobacterium *Nostoc* sp. PCC 7119, the green alga *Monoraphidium braunii* and the plants *Arabidopsis thaliana* and spinach were purified as described elsewhere [21,22]. Fern *Dryopteris crassirhizoma* Pc was generously provided by Prof. Marcellus Ubbink (Leiden University, The Netherlands). *Chlamydomonas reinhardtii* Pc and *Phaeodactylum* Cyt were obtained by cloning and expression in *Escherichia coli* cells using synthetic genes from the available database protein sequences, but with the codon usage optimized for *E. coli*, and fused in the amino-terminal to the transit peptide of Cyt from *Nostoc* sp. PCC 7119 [23]. Protein purifications from the periplasmic fractions were carried out as previously described [17,21]. *Chlamydomonas* Pc mutants were generated by site-directed mutagenesis using oligonucleotide pairs containing the sequence change desired (Fig. S1, in Supplementary section). Single mutants were generated by mutagenic PCR using the synthetic Pc WT gene as template. However, in order to generate the double E85K/Q88R mutant, the E85K simple mutant was used as template. Mutant proteins purification was carried out with the same procedure than for the WT Pc with minor changes. In all cases protein fractions were concentrated and finally frozen at −80 °C until use. The correct expression of all the proteins was checked by MALDI-TOF mass spectrometry, to estimate molecular weights and to compare with the theoretical expected ones. The concentration of the Cyt and Pcs was calculated using the published extinction coefficients at 553 nm (reduced form, Cyt) or 597 nm (oxidized form, Pcs) [21]. The P700 content in PSI samples was calculated from the photoinduced absorbance changes at 820 nm using the absorption coefficient of 6.5 mM^{-1} cm^{-1} determined by Mathis and Sétif [24]. Chlorophyll concentration was determined according to Arnon [25].

2.2. Laser flash absorption spectroscopy

Kinetics of flash-induced absorbance changes associated to PSI reduction were followed at 830 nm as previously described [17,26]. Unless otherwise stated, the standard reaction mixture contained, in a final volume of 0.2 mL, 20 mM Tricine–KOH, pH 7.5, 10 mM $MgCl_2$, 0.03% β-DM, an amount of PSI particles equivalent to 0.5 mg Chl mL^{-1}, 0.1 mM methyl viologen, 2 mM sodium ascorbate and Cyt or Pc at the indicated concentration. To study the ionic strength effect, the NaCl concentration was progressively increased by adding small amounts of a concentrated salt solution to the reaction cell. All the experiments were performed at 22 °C in a 1 mm path-length cuvette. Kinetic data collection and analyses were as previously described [18,26]. Each kinetic trace was the average of 8–12 independent measurements. The estimated error in the observed rate constant (k_{OBS}) determination was ≤15%, based on reproducibility and signal-to-noise ratios. For the *Phaeodactylum* Cyt/PSI native system biphasic kinetic profiles were fitted according to a minimal three-step reaction mechanism involving intracomplex partners rearrangement [18,27]. Values for k_{ON} and k_{OFF}, the association and dissociation rate constants, respectively, for Cyt/PSI complex formation (and the equilibrium constant, $K_A = k_{ON} / k_{OFF}$), the ET first-order rate constant (k_{ET}), the first-order limiting rate constant at infinite protein concentration (k_{SAT}), and the amplitude of the fast phase for PSI reduction extrapolated to infinite Cyt concentration (R_{MAX}), were estimated as previously described [27] (Fig. S2, in Supplementary section). For the Pc/PSI cross-reactions, monophasic kinetic profiles were fitted according to a more simple two-step reaction mechanism [18,28]. Minimal values for k_{ON}, and k_{OFF} (and K_A), as well as the k_{SAT} values (equal to k_{ET}) were estimated applying the formalism previously described [28] (Fig. S2, in Supplementary section) by a nonlinear least-squares computer-fitting iterative procedure using the KaleidaGraph program fitting routine.

2.3. Redox titrations

The redox potential value for WT and each mutant *Chlamydomonas* Pc was determined as reported previously [15] by following the differential absorbance changes at 597 minus 500 nm. Errors in the experimental determinations were less than ±5 mV.

2.4. Structural modeling of proteins

The *Phaeodactylum* PSI complex was built as follows. PsaA and PsaB subunits were modeled based on the X-ray crystal structures of PsaA and PsaB from *Pisum sativum* (PDB 2WSC), with which they shared 79% and 76% sequence identity, respectively [29]. The *Phaeodactylum* PsaF subunit was modeled based on the theoretical model of PsaF [30] from *Phaseolus aureus* (PDB 1YO9), with which it shared 51% sequence identity (slightly better than PsaF from spinach in PDB 2WSC, which had 48% sequence identity with *Phaeodactylum* PsaF). In addition, in this theoretical model the well conserved positively-charged PsaF residues (K16, R17, K23, K24), expected to be involved in the binding to the donor metalloproteins, are more exposed (especially K23) and seem to be in a better orientation than in the spinach PsaF X-ray crystal structure of PDB code 2WSC. All the sequence alignments were performed using BLAST [31]. Then, a total of ten homology models were built using MODELLER version 9v10 with default settings [32], and the model with the best DOPE score [33] was finally selected.

Mutants of Pc from *Chlamydomonas* (E85K, Q88R, E85K/Q88R, E85V and V93K) were modeled with UCSF Chimera program [34], using the WT Pc crystal structure (PDB entry 2PLT) [35] as scaffold.

2.5. Protein–protein docking simulations

Docking simulations were performed by FTDock [36], with electrostatics and 0.7 Å grid resolution, and ZDOCK 2.1 [37], which generated 10,000 and 2000 rigid-body docking poses, respectively. All docking poses were evaluated by the energy-based pyDock 1.0 scoring scheme [38], based on desolvation and electrostatics, with limited van der Waals energy contribution. Cofactors and ions were both included during the sampling and the scoring calculations, using a recently revamped version (upcoming publication) of pyDock 3.0 [39]. After scoring, each docking pose was inserted in a bilayer lipid membrane using the Membrane Builder tool in the CHARMM-GUI website (http://www.charmm-gui.org) [40] with default parameters. Finally, all the docking solutions in which the soluble electron carrier was clashing with the membrane (i.e., any Pc or Cyt atom within a distance of less than 3 Å from any lipid molecule of the membrane) were removed. Atomic distances and clashes were computed with ICM-Browser program [41] (www.molsoft.com).

3. Results

3.1. Kinetic analysis

In *Phaeodactylum*, Cyt acts as the only electron donor to PSI since this organism, as most diatoms, lacks Pc [3,8]. Here, by using laser-flash absorption spectroscopy, we have analyzed cross-reactions between Pc donor proteins from cyanobacteria, green alga and plant species with *Phaeodactylum* PSI. The rationale of this study is to check if diatom PSI still recognizes the lost Pc donor, and thus to shed new light on the evolution of the ET mechanism to PSI in the different branches of the evolutionary tree of photosynthetic oxygenic organisms.

Control experiments analyzing the interaction of cloned WT diatom Cyt with PSI showed biphasic kinetics for the re-reduction of photo-oxidized P_{700}^{+} by Cyt (not shown), the Cyt concentration dependence of the k_{OBS} for both phases confirming the occurrence of the mechanism previously described for the *Phaeodactylum* Cyt/PSI couple and other eukaryotic donor/PSI systems [13,17,18,27,42–44]:

$$\mathrm{PSI_{OX}} + \mathrm{Cyt_{RED}} \underset{k_{\mathrm{OFF}}}{\overset{k_{\mathrm{ON}}}{\leftrightarrow}} [\mathrm{PSI_{OX}}\ldots\mathrm{Cyt_{RED}}] \overset{K_{\mathrm{R}}}{\leftrightarrow} [\mathrm{PSI_{OX}}\ldots\mathrm{Cyt_{RED}}]^{*} \overset{k_{\mathrm{ET}}}{\rightarrow} \mathrm{PSI_{RED}} + \mathrm{Cyt_{OX}}$$

where K_A corresponds to k_{ON}/k_{OFF}, K_R is the equilibrium constant for the rearrangement of the initial transient complex to achieve an optimized ($[\mathrm{PSI_{OX}}\ldots\mathrm{Cyt_{RED}}]^{*}$) ET configuration, and k_{ET} is the ET rate, which approximates to the k_{OBS} of the initial fast phase (ca. 20,000 s^{-1}; not shown), that in *Phaeodactylum* does not represent more than ≈30% of the kinetics amplitude (not shown) [17]. The K_A value, estimated by using the formalisms previously reported [27] is shown in Table 1 together with k_{SAT}, the first-order limiting rate constant of the predominant slower phase at infinite protein concentration (Fig. 1, left) that does not discriminate the rate for rearrangement from the pure intracomplex ET reaction. All the observed kinetics features are similar to those previously described for the Cyt purified from diatom cells [17].

When studying the interaction of Pcs with *Phaeodactylum* PSI, only monophasic kinetics were observed (Fig. 2A), even though the protein concentration dependence of k_{OBS} exhibits in some cases saturation profiles (some examples are shown in Fig. 1, left). It is important to mention that a possible fast kinetic component (if any) with a low amplitude (≤10% of total absorbance change) would not be detectable in our system, due to the signal to noise ratio of the kinetic traces (Fig. 2). Consequently the results have been here analyzed considering a more simple mechanism, consisting in transient complex formation without any detectable rearrangement within the complex previous to the ET step [13,19,28]:

$$\mathrm{PSI_{OX}} + \mathrm{Pc_{RED}} \overset{K_A}{\leftrightarrow} [\mathrm{PSI_{OX}}\ldots\mathrm{Pc_{RED}}] \overset{k_{ET}}{\rightarrow} \mathrm{PSI_{RED}} + \mathrm{Pc_{OX}}$$

where the estimated k_{SAT} limiting rate constant of the observed single kinetic phase at infinite Pc concentration now corresponds to the protein-independent k_{ET} rate [18,28]. Fig. 1 also includes, for comparative purposes, the concentration dependence of the slower (and predominant) phase of the *Phaeodactylum* Cyt/PSI system. The estimated K_A values (equal to k_{ON}/k_{OFF}) for acidic Pcs are comparable to that calculated for the *Phaeodactylum* native system (≈1×10^4 M^{-1}; Table 1). However, all these Pcs exhibited lower k_{SAT} (i.e., k_{ET}) values (ca. 300–400 s^{-1}) as compared both with the k_{SAT} of the slower phase (ca. 900 s^{-1}) (Table 1) and the k_{ET} (20,000 s^{-1}) of the diatom system, indicating the formation of less-optimized Pc/PSI ET complexes. On the other hand, in the case of Pcs from the cyanobacterium *Nostoc* and the fern *Dryopteris*, much slower kinetics (not shown) with k_{OBS} values depending linearly on Pc concentration were obtained (Fig. 1, left), indicating the occurrence of a simple oriented collisional mechanism, with no formation of transient complex [13,19]. Table 1 shows values for the second-order rate constants (k_2) inferred from such linear protein concentration dependence. Thus, whereas diatom PSI is able to effectively bind eukaryotic acidic Pcs, although with a minor ET efficiency as compared with *Phaeodactylum* Cyt (Fig. 1 and Table 1), both the positively-charged *Nostoc* Pc and the fern Pc, in which the relocation of the acidic region results in very distinct electrostatic properties [45,46], showed a drastic decrease in the affinity and the ET efficiency to PSI. Finally, the absence of a detectable fast phase when any Pc acts as electron donor to diatom PSI points to the subtle and precise interactions involved in the rearrangement process leading to the optimized configuration for ET in the native Cyt/PSI diatom couple [17].

Considering the different electrostatic features of *Phaeodactylum* Cyt and the different Pcs [8,17,20], an analysis of the effect of ionic strength on the process was also performed (Fig. 1, right; and Fig. 2B). As previously shown, the dependence of k_{OBS} for the predominant slower phase of PSI reduction in the native Cyt/PSI system on NaCl concentration showed a bell-shaped profile when increasing ionic strength, indicating the existence of some reorientation of redox partners inside the transient complex prior to the ET step [[17]; and see below]. Regarding the Pc/PSI interaction, cyanobacterial and eukaryotic Pcs behave in an opposite way when increasing NaCl concentration (Fig. 1, right). Thus, eukaryotic Pcs interact with diatom PSI by means of attractive forces, as inferred from the continuous decrease of the k_{OBS} at increasing salt concentrations. However, *Nostoc* Pc shows repulsive electrostatic interactions with the diatom PSI, as deduced from the increase of the k_{OBS}

Table 1
Kinetic parameters for *Phaeodactylum* PSI reduction by cytochrome c_6 and prokaryotic and eukaryotic plastocyanins.

Donor protein	k_2 (M^{-1} s^{-1})	K_A[a] (M^{-1})	k_{SAT}[a,b] (s^{-1})	k_2^{HI}[c] (M^{-1} s^{-1})	E_O (mV)
Phaeodactylum Cyt	–	$0.8 \pm 0.16 \times 10^4$	930 ± 19	$2.9 \pm 0.08 \times 10^6$	
Nostoc Pc	$2.4 \pm 0.02 \times 10^5$	–	–	$5.0 \pm 0.25 \times 10^5$	
Dryopteris Pc	$2.5 \pm 0.30 \times 10^4$	–	–	$6.0 \pm 0.50 \times 10^4$	
Arabidopsis Pc	–	$1.5 \pm 0.05 \times 10^4$	390 ± 13	$4.5 \pm 0.25 \times 10^5$	
Spinach Pc	–	$0.8 \pm 0.20 \times 10^4$	290 ± 70	$5.6 \pm 0.09 \times 10^5$	
Monorapidium Pc	–	$1.0 \pm 0.30 \times 10^4$	290 ± 77	$6.0 \pm 0.50 \times 10^5$	
Chlamydomonas WT Pc	–	$1.0 \pm 0.08 \times 10^4$	360 ± 29	$4.0 \pm 0.10 \times 10^5$	+370
Chlamydomonas E85K Pc	–	$0.5 \pm 0.07 \times 10^4$	690 ± 97	$1.1 \pm 0.20 \times 10^6$	+373
Chlamydomonas E85V Pc	–	$1.8 \pm 0.23 \times 10^4$	120 ± 15	$3.0 \pm 0.10 \times 10^5$	+364
Chlamydomonas Q88R Pc	–	$0.4 \pm 0.09 \times 10^4$	610 ± 120	$1.0 \pm 0.04 \times 10^6$	+368
Chlamydomonas E85K/Q88R Pc	–	$0.9 \pm 0.15 \times 10^4$	260 ± 43	$6.2 \pm 0.60 \times 10^5$	+389
Chlamydomonas V93K Pc	–	$1.0 \pm 0.08 \times 10^4$	170 ± 14	$4.0 \pm 0.25 \times 10^5$	+378

[a] Estimated according to the formalisms previously described [27,28].

[b] Value corresponding to the limiting rate constant at infinite protein concentration of the slower and major phase in the *Phaeodactylum* Cyt/PSI native system [17] and to the ET first-order rate constant, k_{ET}, in the Pc/PSI cross-reactions.

[c] Estimated at 200 mM NaCl. Error values are given by standard deviations. See the Experimental procedures section for more details.

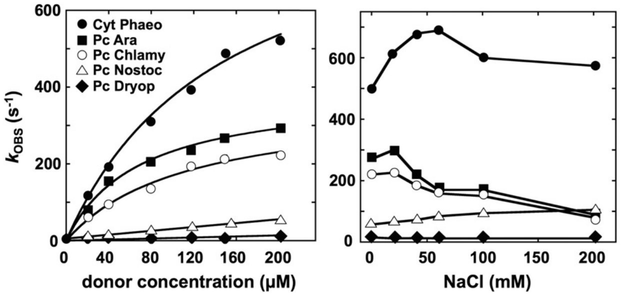


Fig. 1. *Phaeodactylum* PSI reduction by *Phaeodactylum* Cyt (slow predominant phase) or Pc from *Nostoc*, *Chlamydomonas*, *Dryopteris* and *Arabidopsis*, as indicated. (*Left*) Dependence of the observed rate constant (k_{OBS}) upon donor protein concentration. The standard reaction mixture contained, in a final volume of 0.2 mL, 20 mM Tricine–KOH, pH 7.5, 10 mM $MgCl_2$, 0.03% β-DM, an amount of PSI-enriched particles equivalent to 0.5 mg of chlorophyll mL^{-1}, 0.1 mM methyl viologen, 2 mM sodium ascorbate and the indicated concentrations of either Cyt or Pc. (*Right*) Plots of k_{OBS} versus NaCl concentration. The salt content of the samples was increased by adding small amounts of concentrated NaCl stock solutions. Cyt or Pc concentration was 200 μM. Other experimental conditions were as described in Experimental procedures section.

when increasing NaCl. Table 1 shows the second-order rate constants values (k_2^{HI}) estimated at high ionic strength (200 mM NaCl concentration) for the different donor/PSI systems. It is interesting to note that, with the exception of the divergent fern Pc, the k_2^{HI} values at high ionic strength are similar, but also sensibly lower than that obtained with the diatom Cyt, indicating that the intrinsic reactivity of the different Pcs with PSI in the absence of electrostatic interactions is similar, but lower than the native Cyt. This behavior could be explained by the different intrinsic efficiency of cofactors — exposed heme vs. a hindered Cu cofactor — or different surface steric properties.

From the solved crystal structure of *Phaeodactylum* Cyt, turns out that this protein has evolved towards a decrease in the acidic area thought to be involved in the interaction with PSI (revised in [47], and see [8,17]) (Fig. 3). Taking into account the different intensity in the electrostatic character of *green* Pcs and the diatom Cyt, a set of *Chlamydomonas* Pc mutants were constructed by replacing negative residues by neutral or positive groups (Fig. 3) in order to mimic the Cyt properties, and so trying to increase in this way the Pc ET efficiency with PSI. Although *Arabidopsis* Pc showed to be slightly more reactive than *Chlamydomonas* Pc, the green alga protein was selected because of its closer evolutive relation with the diatom. When analyzing the reduction of *Phaeodactylum* PSI by the mutated Pcs, no fast phase was observed in most cases, although the non-linear donor protein concentration dependence of k_{OBS} shown in Fig. 4 (left) again indicates the formation of transient bimolecular complexes. The estimated minimal values for K_A and k_{ET} for the different Pc mutants are shown in Table 1. The only exception to this behavior is the Q88R mutant, for which biphasic kinetics were observed at Pc concentration ≥150 μM (not shown). However, the very low amplitude of the initial fast phase (<15%) does not allow to obtain reliable k_{OBS} values (≈1000 s^{-1}, not shown), which are in any case of the same order of magnitude as the saturation value of the slower phase at high protein concentration (ca. 600 s^{-1}, Table 1). Thus the occurrence of the kinetic mechanism involving complex formation without intracomplex protein rearrangement was here assumed for the interaction of this mutant with PSI. In addition, this mutation brings back one arginine residue into the northern surface of *Chlamydomonas* Pc that has been previously described as important for the binding to PSI in prokaryotic Pcs [15].

The kinetic data shown in Table 1 indicate a moderate effect of the different Pc mutations at low ionic strength. Thus, the E85K and Q88R Pc showed a slightly increased efficiency in the ET to PSI, whereas mutants E85V, V93K and E85K/Q88R, displayed a diminished reactivity towards PSI. In the case of the E85K and Q88R mutants, the less saturated profiles at high donor concentration, as compared both with the WT Pc and the other mutants (Fig. 4, left), generates a higher inaccuracy in the determination of the kinetic constants (Table 1). However, this behavior already reflects a diminished K_A towards PSI, as also indicated by the lower estimated values. It is interesting to note that these two mutants

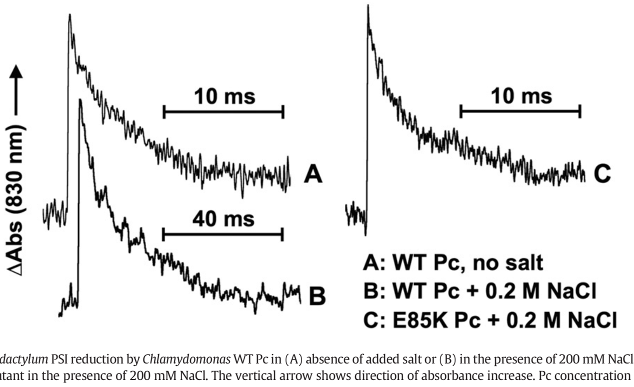


Fig. 2. Kinetic traces showing *Phaeodactylum* PSI reduction by *Chlamydomonas* WT Pc in (A) absence of added salt or (B) in the presence of 200 mM NaCl. (C) *Phaeodactylum* PSI reduction by the E85K *Chlamydomonas* Pc mutant in the presence of 200 mM NaCl. The vertical arrow shows direction of absorbance increase. Pc concentration was 200 μM. Other experimental conditions were as described in Fig. 1.

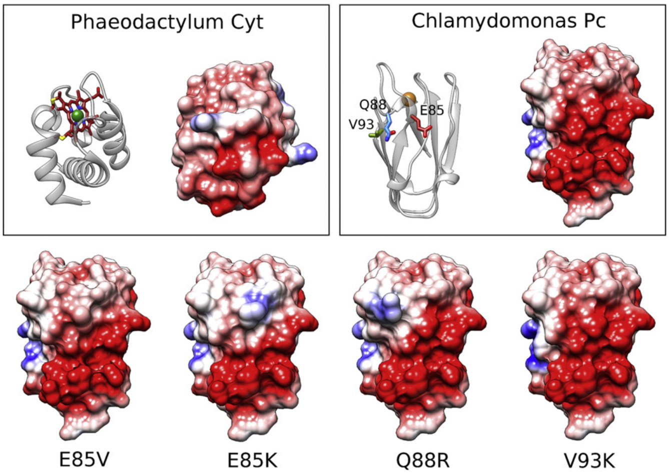


Fig. 3. (*Top*) Backbone and surface electrostatic potential distribution of WT Cyt from *Phaeodactylum* and WT Pc from *Chlamydomonas*. The replaced groups in Pc are depicted in red (E85), blue (Q88) and green (V93) on the structure. (*Bottom*) Surface electrostatic potential distribution of *Chlamydomonas* Pc mutants. The views display in front the protein surface proposed to be responsible for electrostatic interactions with PSI, as shown by the backbone draws. Calculations of surface electrostatic potential distribution were performed with UCSF Chimera program using default parameters and based on Coulomb's electrostatics with distance dependent dielectric constant (ε = 4*d*). Electrostatic potential values are shown in a scale from red to blue, corresponding to −10.0 and +10.0 kcal/(mol·*e*), respectively, at 298 K.

showing the lower K_A values (E85K and Q88R) also have the higher k_{ET} rates. The opposite effect is observed in the E85V Pc, in which a higher affinity in the binding to PSI is accompanied by the lower efficiency in the ET reaction (Table 1). The different effect of the mutations has to be explained mostly in terms of changes in the electrostatics of protein surfaces or steric modifications induced by the amino acid replacements, as the redox potential is not significantly altered in the Pc mutants (Table 1), with the only exception of the E85K/Q88R double mutant, for which a ca. 20 mV potential increase would only partially explain the diminished k_{ET}. In addition, the effect of the mutations can be highlighted when analyzing the effect of increasing ionic strength on PSI reduction by the different Pc mutants (Figs. 2C and 4, right). As indicated before, WT Pc showed a marked decrease of the k_{OBS} with increasing salt concentration, thus reflecting relatively strong attractive interactions with PSI. However, all the mutants present a dependence on ionic strength much smoother than the WT protein, which is again particularly relevant in the case of the E85K and Q88R Pcs, both mutants being significantly more reactive towards PSI than the WT Pc as the salt concentration increases (Figs. 2B and C and 4, right), and consequently showing in both cases higher k_2^{HI} values (Table 1).

3.2. Structural model of Cyt/PSI interaction by computational docking

To understand the structural and energetic determinants of the differences in efficiency observed in diatom PSI reduction with respect to the green systems [17], computational docking simulations were performed between the modeled structure of PSI from *Phaeodactylum* (see Experimental procedures section) and both its native Cyt (PDB entry 3DMI) [8] and the corresponding Cyt from the green alga *Monoraphidium* (PDB entry 1CTJ) [48], for which kinetic data for the ET to diatom PSI have been previously reported [17]. We note that while we were preparing this manuscript, a new structure for PSI from the plant *P. sativum* (PDB 4Y28) has been reported [49], with a slightly better sequence identity with *Phaeodactylum* PsaF (53%).

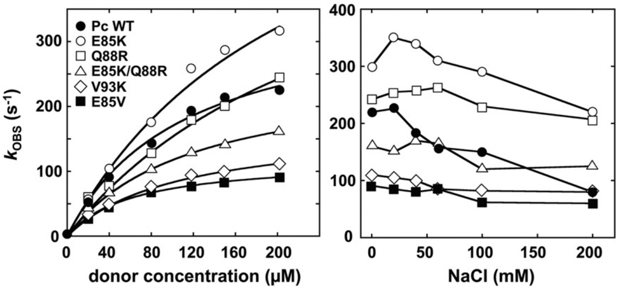


Fig. 4. *Phaeodactylum* PSI reduction by *Chlamydomonas* WT Pc and mutants, as indicated. (*Left*) Dependence of k_{OBS} upon donor protein concentration. (*Right*) Plots of k_{OBS} versus NaCl concentration at 200 μM Pc. Other experimental conditions were as described in Fig. 1.

Actually, the theoretical model that we used as template for PsaF has 1.6 Å Cα-RMSD from this new structure (4Y28), therefore much closer to it than the spinach PsaF structure (2WSC) that we used as template for PsaA and PsaB (2.8 Å Cα-RMSD). Even though the choice of one or another structure among the similar templates that are available for PsaF is not going to significantly affect the modeling, all the above considerations suggest that the choice of the theoretical model as template for PsaF was an appropriate decision.

The docking simulations between PSI and Cyt from *Phaeodactylum* showed a funnel-like binding energy landscape (Fig. S3, in Supplementary section), with the lowest-energy docking orientations in which the residues W652 of PsaA and W624 of PsaB (which form the expected hydrophobic recognition site and the ET pathway to P700 for Cyt and Pc [50,51]) and the propionate groups of the heme molecule of Cyt were located at a distance of around 10–20 Å (minimum distance among the top 20 docking poses is 8.1 Å) (Fig. S3, in Supplementary section and Table 2). There are also docking orientations with short-distance between W652 of PsaA, W624 of PsaB and the Cyt heme, but at higher docking energy. This is consistent with the reorganization process from the initial encounter complex observed for this interaction [[17], and see above].

Remarkably, the docking between *Phaeodactylum* PSI and *Monoraphidium* Cyt yielded a much larger population of low-energy docking orientations in which PsaA W652, PsaB W624 and the Cyt heme were at very short distance (minimum distance among the top 20 docking poses is 2.8 Å) as compared with the native *Phaeodactylum* complex (Fig. S3, in Supplementary section). This is consistent with the more efficient ET kinetics (and higher k_{ET}) to diatom PSI found when analyzing *Monoraphidium* Cyt as compared with the native protein, and with the smaller reorganization effect seen for this interaction [17]. In addition, the docking poses for *Monoraphidium* Cyt showed better scoring values (average scoring for the top 20 docking poses is −39.9 a.u.) than for *Phaeodactylum* Cyt (average scoring for the top 20 docking poses is −32.7 a.u.) (Table 2). Since the docking scoring values can be related to the binding affinity, these findings are also consistent with the better association constant (K_A) found in the cross-reaction with *Monoraphidium* Cyt [17].

We have analyzed in detail the most efficient docking models for ET (e.g., PsaA W652, PsaB W624 and Cyt heme groups at less than 3.0 Å distance) in the interaction of the diatom PSI with *Phaeodactylum* and *Monoraphidium* Cyt (Fig. 5). The best energy model from *Phaeodactylum* Cyt docking (PsaA W652, PsaB W624 and Cyt heme at 3.0 Å distance) does not show electrostatics interactions with the conserved positive patch in PsaF (Fig. 5B), neither favorable interactions at the binding interface (Fig. 5C). This can explain why the efficient docking orientations in the *Phaeodactylum* Cyt docking are energetically penalized. On the contrary, the best model from *Monoraphidium* Cyt docking shows strong electrostatic interactions between the conserved positive patch in PsaF and the Cyt residues D69, E70, D71, and E72 (Fig. 5D and E) in the Cyt acidic patch, previously proposed to be involved in the interaction with PSI in green systems [12,43,51]. Although this acidic patch is conserved in *Phaeodactylum* Cyt (residues S113, D114, E115 and E116) [8], as stated above these residues in *Phaeodactylum* Cyt are not interacting with PsaF in the ET efficient docking orientations. The reason for these differences in binding is that in *Monoraphidium* Cyt this docking orientation is further stabilized by electrostatic interactions between PsaA residues R747 and R648, and *Monoraphidium* Cyt residues D42 and H30 (Fig. 5E). However, in *Phaeodactylum* Cyt these residues are A86 and K74, respectively, which thus can explain why an equivalent docking orientation would be energetically penalized in the interaction with *Phaeodactylum* PSI, and, as a consequence, other less-efficient orientations for ET become more populated (Fig. S3, in Supplementary section).

3.3. Structural model of Pc/PSI interaction by computational docking

We also applied computational docking to investigate the interaction between *Phaeodactylum* PSI and Pc from *Chlamydomonas*, both for the WT protein and mutant variants previously designed to try to mimic *Phaeodactylum* Cyt electrostatic properties (Pc mutants E85K, Q88R, E85K/Q88R, E85V and V93K) (Fig. 3). In the WT Pc docking simulation, the 20 lowest-energy docking orientations had the PsaA W652, PsaB W624 residues and the Pc His87 residue [involved in the active-site catalytic triad [35,43,50]] at a distance above 4.2 Å (Fig. S4, in Supplementary section). The average docking scoring for these top 20 docking poses is −37.4 a.u. (Table 2). There are other docking poses with shorter distances between the redox groups (and therefore, expectedly more efficient for ET) but with clearly worse docking energies. In any case, the fact that the lowest-energy docking orientations have redox groups not very far from each other is consistent with the absence of reorganization effects for these interactions. According to the docking model, the minor ET efficiency of *Chamydomonas* Pc to *Phaeodactylum* PSI — as compared with the diatom Cyt — is not justified by a longer distance between redox groups in docking, but it should be again explained by an intrinsic different efficiency of cofactors and/or surface composition, given that they are essentially two different systems. A deep theoretical analysis of all these considerations would involve large quantum and molecular mechanics calculations, which are beyond the focus of the present work. On the other hand, when performing docking with Pc E85K and Q88R mutants, the 20 lowest-energy docking poses included models that had PsaA W652, PsaB W624 and Pc His87 located at very short distance (1.4 Å), with an average docking scoring of around −30 a.u. (Fig. S4, in Supplementary section; and Table 2). This is consistent with the above described observation that these mutants slightly increased ET but decreased binding affinity.

Table 2
Computational docking results for the interaction of *Phaeodactylum* PSI with cytochromes c_6 and plastocyanins.

Donor protein	*Low-energy docking models*[a]				*Best ET docking models*[b]		
	Best ET model[c]						
	Energy	Rank	Distance	Average energy	Lowest energy	Rank	Distance
	(a.u.)		(Å)	(a.u.)	(a.u.)		(Å)
Phaeodactylum Cyt	−29.8	20	8.1	−32.7	−17.6	209	3.0
Monoraphidium Cyt	−39.2	14	2.8	-39.9	−39.2	14	2.8
Chlamydomonas WT Pc	−40.7	2	4.2	−37.4	−29.0	79	1.9
Chlamydomonas E85K Pc	−29.8	14	1.4	−31.1	−29.8	14	1.4
Chlamydomonas E85V Pc	−33.5	15	4.2	−35.2	−28.5	60	1.4
Chlamydomonas Q88R Pc	−28.6	9	1.4	−29.9	−28.6	9	1.4
Chlamydomonas E85K/Q88R Pc	−33.0	8	3.2	−33.0	−28.4	22	1.9
Chlamydomonas V93K Pc	−40.3	3	4.1	−37.3	−28.7	74	1.9

[a] The 20 lowest-energy docking models.
[b] The most efficient docking models for ET, in which PsaA W652/PsaB W624 are located at less than 3.0 Å from Cyt heme, or less than 2.0 Å from Pc His87.
[c] The most efficient docking orientation for ET (i.e., shortest distance between PsaA W652/PsaB W624 and Cyt heme or Pc His87 groups), among the 20 lowest-energy docking models.

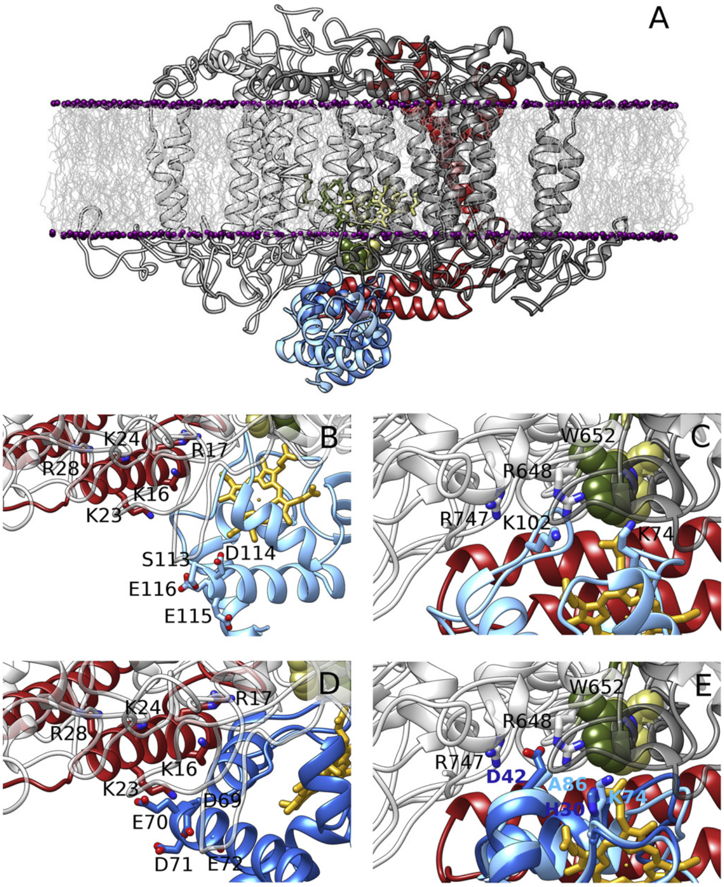


Fig. 5. Representative docking models between *Phaeodactylum* PSI and *Phaeodactylum* or *Monoraphidium* Cyts. (A) Best-energy docking models for efficient ET (PsaA W652/PsaB W624 and Cyt heme groups at less than 3.0 Å distance) are shown for *Phaeodactylum* (light blue; rank 209, docking energy −17.6 a.u., distance between Trp residues and cofactors 3.0 Å) and *Monoraphidium* (dark blue; rank 14, docking energy −39.2 a.u., distance between Trp residues and cofactors 2.8 Å) Cyts. (B–D) Details of atomic interactions for (B, C) *Phaeodactylum* Cyt and (C, D) *Monoraphidium* Cyt docking models. In (D) *Phaeodactylum* Cyt is shown as superimposed onto the *Monoraphidium* Cyt, for the sake of comparison. The PsaA, PsaB, and PsaF subunits of PSI are depicted in light gray, dark gray, and red, respectively.

It is interesting to analyze in atomic detail the results from the Pc/PSI docking. Fig. 6 shows two representative models from the docking between *Phaeodactylum* PSI and *Chamydomonas* WT Pc. One (Fig. 6B, C) is the best-energy model among the most efficient docking orientations for ET (e.g., PsaA W652, PsaB W624 and Pc His87 at less than 2.0 Å distance). More specifically, this docking model has PsaA W652, PsaB W624 and Pc His87 at a distance of 1.9 Å, and shows a strong electrostatic interaction between the conserved positive patch in PSI subunit PsaF and the Pc residues D42, E43 and D44 in the "east" negative patch [35], thus delineating an orientation equivalent to that above described for *Monoraphidium* Cyt (compare Figs. 5D and 6B). Interestingly, PsaA R747 is interacting with Pc residue N64 (located in same position as *Monoraphidium* Cyt D42 in the ET efficient docking model), and PsaA W652 is interacting with Pc H87 (Fig. 6C). This shows that green Pc is able to form similar contacts as green Cyt with *Phaeodactylum* PSI for efficient ET. This docking orientation would not be affected in the E85K and Q88R mutants, and thus the reason of the effect of these mutations must be found in other docking orientations that are not so efficient for ET (see below).

In this sense, Fig. 6 also shows the expectedly most efficient orientation for ET among the 20 lowest-energy docking models in which PsaA W652, PsaB W624 and Pc His87 are located at 4.2 Å distance (Fig. 6D, E), which in principle would be less efficient for ET than the previously described docking model. In this new model, while there is still some electrostatic interactions with the conserved positive patch of PsaF subunit in PSI (Fig. 6D), there are other interactions that could additionally contribute to its binding affinity, such as the one between Pc Q88 and PsaA K638, or more especially, between Pc E85 and PsaA R463 and R648 (Fig. 6E). Replacement of these two Pc residues by positively charged ones, as in E85K and Q88R mutations, will cause destabilization of this docking

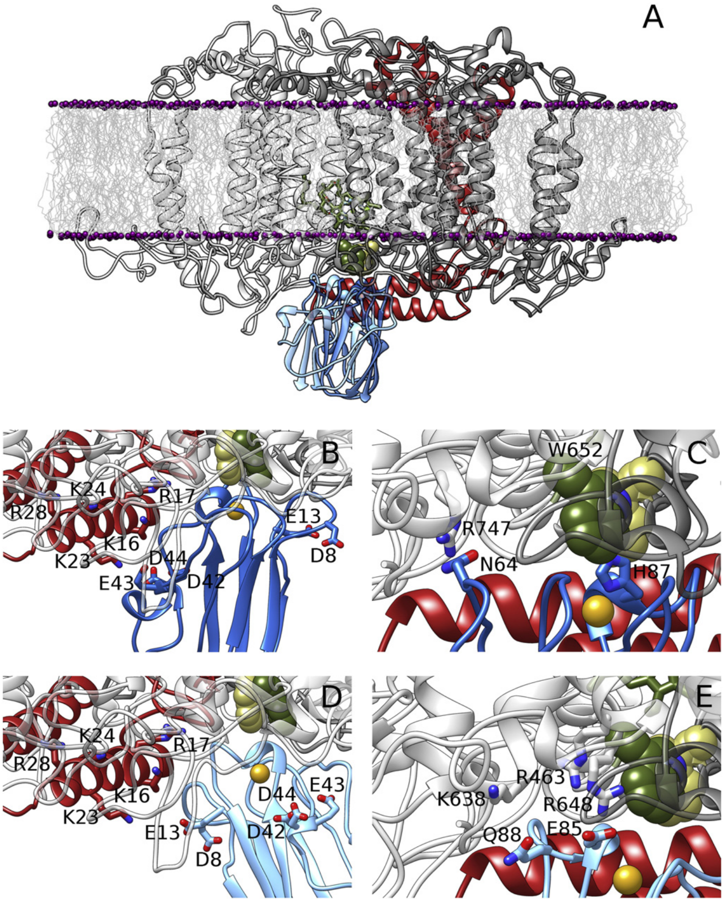


Fig. 6. Representative docking models between *Phaeodactylum* PSI and *Chlamydomonas* Pc. (A) Rank 79 orientation (dark blue; docking energy −29.0 a.u.), with PsaA W652/PsaB W624 and Pc His87 located at 1.9 Å; rank 2 orientation (light blue; docking energy −40.7 a.u.), with PsaA W652, PsaB W624 and Pc His87 located at 4.2 Å distance. (B–D) Details of atomic interactions for ET efficient (B, C) and low-energy (D, C) docking models. The PsaA, PsaB, and PsaF subunits of PSI are depicted in light gray, dark gray, and red, respectively.

orientation, which has redox centers at medium distance (and therefore not optimal for ET), and now the docking orientations with short distance between redox groups (expectedly more efficient for ET) would become more populated.

On the other side, in the docking simulations with the Pc mutants E85K/Q88R, E85V and V93K, no significant differences were found in the distribution of the low-energy docking pose orientations with respect to the WT (not shown). Indeed, the average docking energies for the top 20 docking poses were of −33.0, −35.2 and −37.3 a.u., for which the minimum distance between PsaA W652, PsaB W624 residues and Pc H87 were 3.2, 4.2 and 4.1 Å respectively (Table 2).

4. Discussion

PSI reduction has been extensively analyzed in vitro and in vivo in a wide variety of organisms, revealing that the kinetic mechanisms for the reaction of either Pc or Cyt with PSI from the same organism are similar, although they have increased in complexity and efficiency while evolving from prokaryotic cyanobacteria to green alga and plant eukaryotic organisms [13]. Thus, PSI reduction by the donor proteins, isolated from different sources, can follow either an oriented collisional mechanism (type I), a mechanism involving transient complex formation (type II), or complex formation with rearrangement of the interface (type III), the latter mechanism classically observed in green alga and plant eukaryotic systems [13].

It has been previously described that the ET reaction from Cyt to PSI in the diatom *Phaeodactylum* follows the type III three-steps mechanism found in eukaryotic green systems [17], in which an initial Cyt/PSI encounter complex reorganizes to a more productive final configuration. This is consistent with the docking models shown here, in which the most stable docking orientations are not expected to be efficient for ET due to the longer distance between redox centers. In addition, the diatom system shows lower efficiencies than the green systems both in the formation of the properly arranged [Cyt–PSI] complex and in the ET reaction itself [13,17,22,27,43,50]. This apparent decreased reactivity is the consequence of diminished basic patches on PsaF and acidic regions on Cyt, both resulting in a weaker electrostatic interaction between partners. This feature of diatoms has been proposed to denote a compromise between ET efficiency and optimal protein donor turnover [17], as in the green systems it has been suggested that the strong donor/PSI electrostatic interaction limits the donor exchange and so the overall ET turnover [12,44].

It is interesting to compare the *Phaeodactylum* native Cyt/PSI docking complex (Fig. 5B, C) with those described previously in green systems [10,12,43,51]. It is widely accepted that the lumenal loops i/j of PsaA/B in PSI, including the PsaA W651 and PsaB W627 residues (*Chlamydomonas* numbering), form the hydrophobic recognition site for binding of Pc and Cyt, by means of complementary hydrophobic areas around the donors ET site [43,50]. Electrostatic interactions are also established between negatively charged residues of Pc and Cyt with the positively charged N-terminal domain of PsaF [12,51]. Particularly, *Chlamydomonas* Cyt seems to establish specific interactions involving residues K23/K27 of PsaF and the E69/E70 groups located at the "eastern" negatively charged area of Cyt. Additionally, the positive charge on the "northern" site of Cyt (R66) and the adjacent D65 can form a strong salt bridge with the R623/D624 pair of PsaB [51]. According to this model, the distance between the donor/acceptor redox cofactors is ≈14 Å [12,51]. The *Phaeodactylum* Cyt/PSI docking complex described here (Fig. 5B, C) has a different orientation than the *Chlamydomonas* Cyt/PSI complex. The reason is that the D65 group in *Chlamydomonas* Cyt is not conserved in *Phaeodactylum* Cyt (equivalent residue is Gly109), and thus it cannot stabilize this orientation. As a consequence, it also loses the electrostatic interactions with PsaB and the overall binding energy is less favorable.

Cyt is the only electron carrier between the b_6f and PSI complexes in *Phaeodactylum*, and thus a pertinent question is if diatom PSI is still able to recognize the lost Pc donor. It is first interesting to note that *Nostoc* cyanobacterial Pc reacts with low efficiency with diatom PSI by means of a simple oriented collisional mechanism. This is in agreement with the low reactivity previously described in cross-reactions of Cyt/PSI systems from cyanobacteria and *Phaeodactylum* [17]. On the other hand, most cross-reactions involving acidic green Pcs and diatom PSI show kinetic parameters comparable in the main limiting steps (intermolecular affinity and efficiency at saturating donor concentration) to the native diatom system (Table 1), even though they follow different kinetic mechanisms. Thus, the interaction of green alga and plant Pcs with diatom PSI proceeds via transient complex formation, but apparently with the absence of the final rearrangement step observed in the native *Phaeodactylum* system, and with sensibly lower k_{ET} rates, indicating the formation of less-optimal functional complexes. This emphasizes the fine adjustment involved in the formation of the final productive structure in the Cyt/PSI diatom couple, that is lost in the interactions with the non-native Pc donors [17].

The results obtained in the cross-reactions should be explained according to the different structural and electrostatic features of Pcs from different sources, as well as of the donor docking site in diatom PSI. Thus, whereas the low efficiency in the interaction of the positively-charged *Nostoc* Pc agrees with the occurrence of repulsive electrostatic interactions with the also positively-charged binding site on the diatom PSI [17], eukaryotic acidic Pcs seem to interact with diatom PSI by means of attractive forces, as previously described for the interaction of *Monoraphidium* Cyt with *Phaeodactylum* PSI [17]. However, it is interesting to note that, in spite of the different electrostatic character of Pcs from the different sources, the reactivity towards PSI at high ionic strength is very similar in most cases, indicating a comparable intrinsic reactivity of the different Pcs in the absence of electrostatic forces. This intrinsic reactivity is, nevertheless, about 4–5 times lower when compared with the diatom native system (Fig. 1, and Table 1), suggesting that other factors beyond the pure electrostatics, i.e., hydrophobic and/or solvent effects or structural steric factors, contribute to this difference in reactivity.

Fern Pc represents an interesting exception to the main features of the reactivity of green Pcs with diatom PSI, as *Dryopteris* Pc shows the lowest efficiency of the systems here analyzed (up to ten times less efficient than the cyanobacterial *Nostoc* Pc). It has been previously reported that fern Pc conserves both the same global structure and negative electrostatic character of eukaryotic Pcs, but its acidic region has moved from the canonical east position and is surrounding the hydrophobic ET north site, this change resulting in very distinct electrostatic and steric properties [45,46]. Thus, the unusual structure of *Dryopteris* Pc impedes an efficient interaction with diatom PSI, as previously described in its interaction with spinach PSI [46], confirming that fern Pc has followed a relatively independent evolutionary pathway since ferns diverged from other vascular plants [45,46].

Previous results obtained with cross-reactions of different Cyt/PSI eukaryotic systems suggested that the different electrostatic properties of Cyt, more than the PSI, mainly make the difference in behavior of diatoms with respect to other photosynthetic eukaryotes from the green lineage [17]. This has been confirmed here by computational modeling. The *Monoraphidium* Cyt/*Phaeodactylum* PSI docking complex shows virtually the same orientation as the native *Chlamydomonas* Cyt/PSI green complex, and is able to form similar interactions with the positive patch in PsaF (Fig. 5D) [51]. In addition, the salt-bridges formed by D65 and R66 of *Chlamydomonas* Cyt with PsaB R623 and D624 residues are conserved in the *Monoraphidium* Cyt/diatom PSI interaction (equivalent residues: D65 and R67; and R620 and D621, respectively). The key interface D42/R747 salt-bridge found in our *Monoraphidium* Cyt/*Phaeodactylum* PsaA model was not previously reported for the *Chlamydomonas* Cyt/PSI complex [51], but since these residues are conserved (equivalent ones are D41 and R746), we can expect that this salt-bridge is also formed in *Chlamydomonas* Cyt/PSI complex. Interestingly, the redox centers in *Monoraphidium* Cyt/*Phaeodactylum* PSI are found at a shorter distance (11.6 Å) than in *Chlamydomonas* Cyt/PSI (≈14 Å) [51]. In addition, the reduction of diatom PSI by the strongly acidic Cyt from green alga showed an increased affinity and k_{ET} but a lower efficiency in the formation of the properly arranged Cyt/PSI complex as compared with the native Cyt, because the too strong electrostatic interactions [17]. Thus, *Chlamydomonas* Pc mutants are here designed by replacing negative groups of the acidic patch — widely accepted to be responsible for electrostatic interactions with PSI [12,35,47] — by neutral or positive residues (Fig. 3). The rationale for these designs has been to mimic the Cyt electrostatic properties, trying to increase the efficiency of a green Pc in reducing diatom PSI by decreasing the negative character of its acidic patch (Fig. 3).

The effect of the different Pc mutations, although moderate, gives interesting information about the binding mechanism to PSI. Thus, higher k_{ET} rates, as compared with the WT Pc, are observed with the two mutants (E85K and Q88R) showing about half of the K_A value of the WT protein (Table 1). Just the opposite effect is however obtained in the E85V Pc, in which the lower k_{ET} rate goes together with the higher affinity towards PSI (Table 1). Actually, an inverse exponential relationship between the estimated K_A and k_{ET} values is observed for all the *Chlamydomonas* Pc variants (Table 1). Thus, WT *Chlamydomonas* Pc seems to be fixed, by means of strong electrostatic interactions, in a less productive complex configuration, that can be improved in mutants showing an increased flexibility in the binding to PSI. Our docking model shows that E85K and Q88R mutants are destabilizing this less productive complex configuration, which effectively increases the population of the productive orientations and therefore are more efficient for ET. In this sense it is interesting again to compare the docking model of *Chlamydomonas* Pc/diatom PSI with the previously proposed Pc/PSI interactions in green systems, in which electrostatic interactions involve D42/D44 and E43/E45 of Pc with residues K17/K23/K30 in PsaF [11,12]. The *Chlamydomonas* Pc/*Phaeodactylum* PSI most productive docking model (Fig. 6B) conserves such interactions and thus would be able to yield efficient orientations for ET. However, Pc E85 and Q88 residues are stabilizing alternative, but less productive, orientations in the *Chlamydomonas* Pc/diatom PSI complex (Fig. 6E). This is consistent with the smaller ET efficiency found for *Chlamydomonas* Pc, and the ET increase in E85K and Q88R mutants. Interestingly, *Chlamydomonas* Pc does not possess a positively charged amino acid at a position equivalent to the R66 found in *Chlamydomonas* and *Phaeodactylum* Cyts (corresponding to the R87 position of prokaryotic Pcs). In cyanobacteria, this positively charged amino acid is important for efficient ET to PSI [15]. Thus, by bringing back this arginine residue in the Q88R mutant of the green alga Pc, an improved reactivity has been observed.

On the other side, we should note that the effect of the two individual E85K and Q88R mutations is counteracted in the double mutant E85K/Q88R, which shows a similar K_A and a slightly diminished k_{ET} compared with the WT Pc. This would be at least partially explained in terms of the small increase of the double mutant redox potential. However, there must be some additional effect that cannot be described in our rigid-body docking simulations, like a conformational change of the two new positive residues that avoids the destabilization effect of the lesser productive configuration by the two individual mutations. Lastly, the results obtained with the V93K protein indicates that this hydrophobic residue is relevant in the ET process, as this mutant shows a significantly decreased k_{ET}, in spite of maintaining the same affinity for PSI than the WT Pc (Table 1).

To conclude, the kinetic and mutagenic analysis herein reported for *Phaeodactylum* PSI reduction by green Pcs contrasts with the results previously obtained with a eukaryotic Cyt [17]. Whereas the green alga Cyt overall reacts more efficiently with diatom PSI than the native Cyt — both in affinity and ET rate — because its stronger electrostatic character [17], green Pcs are together less efficient in the ET process, while maintaining a similar affinity than the *Phaeodactylum* Cyt towards diatom PSI. In addition, our analysis with mutated green alga Pcs shows that introducing positive groups, and thus weakening the interaction with PSI, can in some cases enhance the ET step. This is the result of an improved intracomplex flexibility to optimize ET, indicating that in the WT Pc too strong electrostatic interactions determine a non-optimal complex configuration of the redox partners. These differences in the Cyt/Pc interaction with the diatom PSI cannot be explained only in terms of dissimilarities in the electrostatics of the studied systems, but also in the existence of differential structural and steric factors in the two families of soluble electron carriers.

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.bbabio.2015.09.006.

Conflict of interest

The authors declare no conflict of interest.

Acknowledgements

Research work was supported by FEDER and the Spanish Ministry of Economy and Competitiveness (MINECO, Grants BIO2012-35271 to JAN, and BIO2013-48213-R to JFR) and the Andalusian Government (PAIDI BIO-022, to JAN). The authors thank Prof. Marcellus Ubbink (Leiden Institute of Chemistry, The Netherlands) for the supply of fern plastocyanin, and to Rocío Rodríguez (Proteomic Service, IBVF) for technical assistance.

References

[1] C. Bowler, A. Vardi, A.E. Allen, Oceanographic and biogeochemical insights from diatom genomes, Ann. Rev. Mar. Sci. 2 (2010) 333–365.

[2] D.M. Nelson, P. Tréguer, M.A. Brzezinski, A. Leynaert, B. Quéguiner, Production and dissolution of biogenic silica in the ocean: revised global estimates, comparison with regional data and relationship to biogenic sedimentation, Glob. Biogeochem. Cycles 9 (1995) 359–372.

[3] C. Bowler, A.E. Allen, J.H. Badger, et al., The *Phaeodactylum* genome reveals the evolutionary history of diatom genomes, Nature 456 (2008) 239–244.

[4] I. Grouneva, P.J. Gollan, S. Kangasjärvi, M. Suorsa, M. Tikkanen, E.-M. Aro, Phylogenetic viewpoints on regulation of light harvesting and electron transport in eukaryotic photosynthetic organisms, Planta 237 (2013) 399–412.

[5] R. Nagao, A. Moriguchi, T. Tomo, A. Niikura, S. Nakajima, T. Suzuki, A. Okumura, M. Iwai, R.-S. Shen, M. Ikeuchi, I. Enami, Binding and functional properties of five extrinsic proteins in oxygen-evolving photosystem II from a marine centric diatom, *Chaetoceros gracilis*, J. Biol. Chem. 285 (2010) 29191–29199.

[6] R. Nagao, A. Ishii, O. Tada, T. Suzuki, N. Dohmae, A. Okumura, M. Iwai, T. Takahashi, Y. Kashino, I. Enami, Isolation and characterization of oxygen-evolving thylakoid membranes and photosystem II particles from a marine diatom *Chaetoceros gracilis*, Biochim. Biophys. Acta 1767 (2007) 1353–1362.

[7] T. Veith, C. Büchel, The monomeric photosystem I-complex of the diatom *Phaeodactylum tricornutum* binds specific fucoxanthin chlorophyll proteins (FCPs) as light-harvesting complexes, Biochim. Biophys. Acta 1767 (2007) 1428–1435.

[8] H. Akazaki, F. Kawai, M. Hosokawa, T. Hama, H. Chida, T. Hirano, B.-K. Lim, N. Sakurai, W. Hakamata, S.-Y. Park, T. Nishio, T. Oku, Crystallization and structural analysis of cytochrome c_6 from the diatom *Phaeodactylum tricornutum* at 1.5 Å resolution, Biosci. Biotechnol. Biochem. 73 (2009) 189–191.

[9] G. Peers, N.M. Price, Copper-containing plastocyanin used for electron transport by an oceanic diatom, Nature 441 (2006) 341–344.

[10] A. Ben-Shem, F. Frolow, N. Nelson, Crystal structure of plant photosystem I, Nature 426 (2003) 630–635.

[11] P. Fromme, A. Melkozernov, P. Jordan, N. Krauss, Structure and function of photosystem I: interaction with its soluble electron carriers and external antenna systems, FEBS Lett. 555 (2003) 40–44.

[12] A. Busch, M. Hippler, The structure and function of eukaryotic photosystem I, Biochim. Biophys. Acta 1807 (2011) 864–877.

[13] M. Hervás, J.A. Navarro, M.A. De la Rosa, Electron transfer between soluble proteins and membrane complexes in photosynthesis, Acc. Chem. Res. 36 (2003) 798–805.

[14] M. Hervás, J.A. Navarro, A. Díaz, M.A. De la Rosa, A comparative thermodynamic analysis by laser-flash absorption spectroscopy of photosystem I reduction by plastocyanin and cytochrome c_6 in *Anabaena* PCC 7119, *Synechocystis* PCC 6803, and spinach, Biochemistry 35 (1996) 2693–2698.

[15] F.P. Molina-Heredia, M. Hervás, J.A. Navarro, M.A. De la Rosa, A single arginyl residue in plastocyanin and cytochrome c_6 from the cyanobacterium *Anabaena* sp. PCC 7119 is required for efficient reduction of photosystem I, J. Biol. Chem. 276 (2001) 601–605.

[16] M. Hervás, A. Díaz-Quintana, C. Kerfeld, D. Krogmann, M.A. De la Rosa, J.A. Navarro, Cyanobacterial photosystem I lacks specificity in its interaction with cytochrome c_6 electron donors, Photosynth. Res. 83 (2005) 329–333.

[17] P. Bernal-Bayard, F.P. Molina-Heredia, M. Hervás, J.A. Navarro, Photosystem I reduction in diatoms: as complex as the green lineage systems but less efficient, Biochemistry 52 (2013) 8687–8695.

[18] M. Hervás, J.A. Navarro, A. Díaz, H. Bottin, M.A. De la Rosa, Laser-flash kinetic analysis of the fast electron transfer from plastocyanin and cytochrome c_6 to photosystem I. Experimental evidence on the evolution of the reaction mechanism, Biochemistry 34 (1995) 11321–11326.

[19] A.B. Hope, Electron transfer amongst cytochrome *f*, plastocyanin and photosystem I: kinetics and mechanisms, Biochim. Biophys. Acta 1456 (2000) 5–26.

[20] M.A. De la Rosa, J.A. Navarro, A. Díaz-Quintana, B. De la Cerda, F.P. Molina-Heredia, A. Balme, P.S. Murdoch, I. Díaz-Moreno, R.V. Durán, M. Hervás, An evolutionary analysis of the reaction mechanisms of photosystem I reduction by cytochrome c_6 and plastocyanin, Bioelectrochemistry 55 (2002) 41–45.

[21] J.A. Navarro, M. Hervás, M.A. De la Rosa, Purification of plastocyanin and cytochrome c_6 from plants, green algae, and cyanobacteria, in: R. Carpentier (Ed.), Photosynthesis Research Protocols, 684, Humana Press Inc., Totowa, NJ 2011, pp. 79–94.

[22] F.P. Molina-Heredia, J. Wastl, J.A. Navarro, D. Bendall, M. Hervás, C. Howe, M.A. De la Rosa, A new function for an old cytochrome? Nature 424 (2003) 33–34.

[23] F.P. Molina-Heredia, A. Balme, M. Hervás, J.A. Navarro, M.A. De la Rosa, A comparative structural and functional analysis of cytochrome c_M, cytochrome c_6 and plastocyanin from the cyanobacterium *Synechocystis* sp. PCC 6803, FEBS Lett. 517 (2002) 50–54.

[24] P. Mathis, P. Sétif, Near infra-red absorption spectra of the chlorophyll *a* cations and triplet state *in vitro* and *in vivo*, Isr. J. Chem. 21 (1981) 316–320.

[25] D.I. Arnon, Copper enzymes in isolated chloroplasts, Plant Physiol. 24 (1949) 1–15.

[26] M. Hervás, J.A. Navarro, Effect of crowding on the electron transfer process from plastocyanin and cytochrome c_6 to photosystem-I: a comparative study from cyanobacteria to green algae, Photosynth. Res. 107 (2011) 279–286.

[27] K. Sigfridsson, S. He, S. Modi, D.S. Bendall, J. Gray, Ö. Hansson, A comparative flash-photolysis study of electron transfer from pea and spinach plastocyanins to spinach photosystem I. A reaction involving a rate-limiting conformational change, Photosynth. Res. 50 (1996) 11–21.

[28] G. Tollin, T.E. Meyer, M.A. Cusanovich, Elucidation of the factors which determine reaction-rate constants and biological specificity for electron-transfer proteins, Biochim. Biophys. Acta 853 (1986) 29–41.

[29] A. Amunts, H. Toporik, A. Borovikova, N. Nelson, Structure determination and improved model of plant photosystem I, J. Biol. Chem. 285 (2010) 3478–3486.

[30] C. Jolley, A. Ben-Shem, N. Nelson, P. Fromme, Structure of plant photosystem I revealed by theoretical modeling, J. Biol. Chem. 280 (2005) 33627–33636.

[31] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman, Basic local alignment search tool, J. Mol. Biol. 215 (1990) 403–410.

[32] N. Eswar, B. Webb, M.A. Marti-Renom, M.S. Madhusudhan, D. Eramian, M.Y. Shen, U. Pieper, A. Sali, Comparative protein structure modeling using MODELLER, Curr. Protoc. Protein Sci. 2 (9) (2007) 1–31 (Unit).

[33] M.Y. Shen, A. Sali, Statistical potential for assessment and prediction of protein structures, Protein Sci. 15 (2006) 2507–2524.

[34] E.F. Pettersen, T.D. Goddard, C.C. Huang, G.S. Couch, D.M. Greenblatt, E.C. Meng, T.E. Ferrin, UCSF chimera-a visualization system for exploratory research and analysis, J. Comput. Chem. 25 (2004) 1605–1612.

[35] M.R. Redinbo, D. Cascio, M.K. Choukair, D. Rice, S. Merchant, T.O. Yeates, The 1.5-Å crystal structure of plastocyanin from the green alga *Chlamydomonas reinhardtii*, Biochemistry 32 (1993) 10560–10567.

[36] H.A. Gabb, R.M. Jackson, M.J. Sternberg, Modelling protein docking using shape complementarity, electrostatics and biochemical information, J. Mol. Biol. 272 (1997) 106–120.

[37] R. Chen, Z. Weng, A novel shape complementarity scoring function for protein–protein docking, Proteins 51 (2003) 397–408.

[38] T.M. Cheng, T.L. Blundell, J. Fernandez-Recio, pyDock: electrostatics and desolvation for effective scoring of rigid-body protein-protein docking, Proteins 68 (2007) 503–515.

[39] B. Jiménez-García, C. Pons, J. Fernández-Recio, pyDockWEB: a web server for rigid-body protein-protein docking using electrostatics and desolvation scoring, Bioinformatics 29 (2013) 1698–1699.

[40] S. Jo, T. Kim, V.G. Iyer, W. Im, CHARMM-GUI: a web-based graphical user interface for CHARMM, J. Comput. Chem. 29 (2008) 1859–1865.

[41] R. Abagyan, W.H. Lee, E. Raush, L. Budagyan, M. Totrov, M. Sundstrom, B.D. Marsden, Disseminating structural genomics data to the public: from a data dump to an animated story, Trends Biochem. Sci. 31 (2006) 76–78.

[42] K. Sigfridsson, S. Young, Ö. Hansson, Electron transfer between spinach plastocyanin mutants and photosystem 1, Eur. J. Biochem. 245 (1997) 801–812.

[43] F. Sommer, F. Drepper, M. Hippler, The luminal helix l of PsaB is essential for recognition of plastocyanin or cytochrome c_6 and fast electron transfer to photosystem I in *Chlamydomonas reinhardtii*, J. Biol. Chem. 277 (2002) 6573–6581.

[44] F. Drepper, M. Hippler, W. Nitschke, W. Haehnel, Binding dynamics and electron transfer between plastocyanin and photosystem I, Biochemistry 35 (1996) 1282–1295.

[45] T. Kohzuma, T. Inoue, F. Yoshizaki, Y. Sasakawa, K. Onodera, S. Nagatomo, T. Kitagawa, S. Uzawa, Y. Isobe, Y. Sugimura, M. Gotowda, Y. Kai, The structure and unusual pH dependence of plastocyanin from the fern *Dryopteris crassirhizoma*. The protonation of an active site histidine is hindered by π–π interactions, J. Biol. Chem. 274 (1999) 11817–11823.

[46] J.A. Navarro, C.E. Lowe, R. Amons, T. Kohzuma, G.W. Canters, M.A. De la Rosa, M. Ubbink, M. Hervás, Functional characterization of the evolutionarily divergent fern plastocyanin, Eur. J. Biochem. 271 (2004) 3449–3456.

[47] A. Díaz-Quintana, M. Hervás, J.A. Navarro, M.A. De la Rosa, Plastocyanin and cytochrome c_6: the soluble electron carriers between the cytochrome b_6f complex and photosystem I, in: P. Fromme (Ed.), Photosynthetic Protein Complexes: A Structural Approach, Wiley-VCH Verlag Gmbh & Co. KGaA, Weinheim 2008, pp. 181–200.

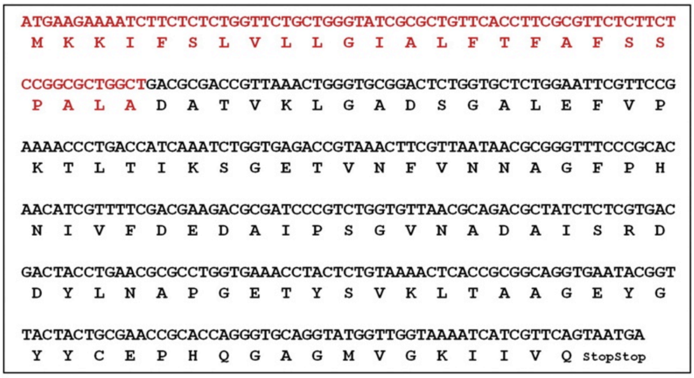
[48] C. Frazão, C.M. Soares, M.A. Carrondo, E. Pohl, Z. Dauter, K.S. Wilson, M. Hervás, J.A. Navarro, M.A. De la Rosa, G.M. Sheldrick, Ab initio determination of the crystal structure of cytochrome c_6 and comparison with plastocyanin, Structure 3 (1995) 1159–1169.

[49] Y. Mazor, A. Borovikova, N. Nelson, The structure of plant photosystem I supercomplex at 2.8 Å resolution, eLife 4 (2015), e07433http://dx.doi.org/10.7554/eLife.07.

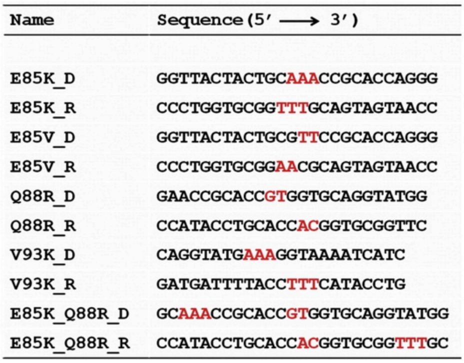
[50] F. Sommer, F. Drepper, W. Haehnel, M. Hippler, The hydrophobic recognition site formed by residues PsaA-Trp651 and PsaB-Trp627 of photosystem I in *Chlamydomonas reinhardtii* confers distinct selectivity for binding of plastocyanin and cytochrome c_6, J. Biol Chem. 279 (2004) 20009–20017.

[51] F. Sommer, F. Drepper, W. Haehnel, M. Hippler, Identification of precise electrostatic recognition sites between cytochrome c_6 and the photosystem I subunit PsaF using mass spectrometry, J. Biol. Chem. 281 (2006) 35097–35103.

Supporting information

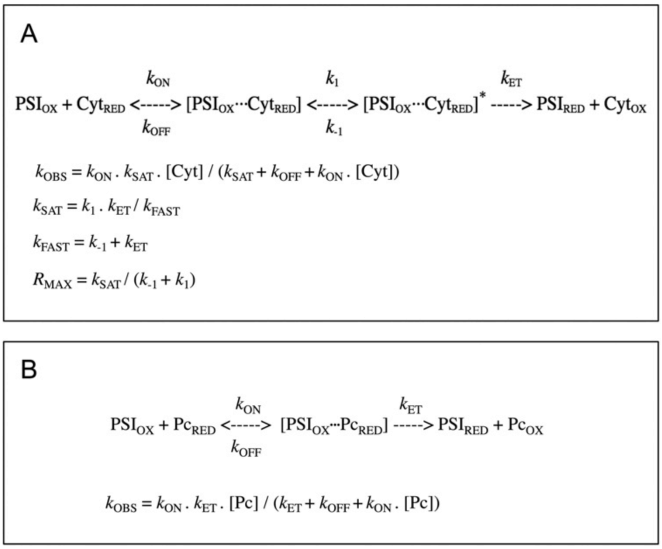


```
ATGAAGAAAATCTTCTCTCTGGTTCTGCTGGGTATCGCGCTGTTCACCTTCGCGTTCTCTTCT
 M  K  K  I  F  S  L  V  L  L  G  I  A  L  F  T  F  A  F  S  S

CCGGCGCTGGCTGACGCGACCGTTAAACTGGGTGCGGACTCTGGTGCTCTGGAATTCGTTCCG
 P  A  L  A  D  A  T  V  K  L  G  A  D  S  G  A  L  E  F  V  P

AAAACCCTGACCATCAAATCTGGTGAGACCGTAAACTTCGTTAATAACGCGGGTTTCCCGCAC
 K  T  L  T  I  K  S  G  E  T  V  N  F  V  N  N  A  G  F  P  H

AACATCGTTTTCGACGAAGACGCGATCCCGTCTGGTGTTAACGCAGACGCTATCTCTCGTGAC
 N  I  V  F  D  E  D  A  I  P  S  G  V  N  A  D  A  I  S  R  D

GACTACCTGAACGCGCCTGGTGAAACCTACTCTGTAAAACTCACCGCGGCAGGTGAATACGGT
 D  Y  L  N  A  P  G  E  T  Y  S  V  K  L  T  A  A  G  E  Y  G

TACTACTGCGAACCGCACCAGGGTGCAGGTATGGTTGGTAAAATCATCGTTCAGTAATGA
 Y  Y  C  E  P  H  Q  G  A  G  M  V  G  K  I  I  V  Q StopStop
```



Name	Sequence (5′ ⟶ 3′)
E85K_D	GGTTACTACTGCAAACCGCACCAGGG
E85K_R	CCCTGGTGCGGTTTGCAGTAGTAACC
E85V_D	GGTTACTACTGCGTTCCGCACCAGGG
E85V_R	CCCTGGTGCGGAACGCAGTAGTAACC
Q88R_D	GAACCGCACCGTGGTGCAGGTATGG
Q88R_R	CCATACCTGCACCACGGTGCGGTTC
V93K_D	CAGGTATGAAAGGTAAAATCATC
V93K_R	GATGATTTTACCTTTCATACCTG
E85K_Q88R_D	GCAAACCGCACCGTGGTGCAGGTATGG
E85K_Q88R_R	CCATACCTGCACCACGGTGCGGTTTGC

Fig. S1. (*Top*) Sequence of *Chlamydomonas reinhardtii* Pc synthetic gen. (*Bottom*) DNA primers used for site-directed mutagenesis of Chlamydomonas Pc.



$$PSI_{OX} + Cyt_{RED} \underset{k_{OFF}}{\overset{k_{ON}}{\longleftrightarrow}} [PSI_{OX} \cdots Cyt_{RED}] \underset{k_{-1}}{\overset{k_{1}}{\longleftrightarrow}} [PSI_{OX} \cdots Cyt_{RED}]^{*} \overset{k_{ET}}{\longrightarrow} PSI_{RED} + Cyt_{OX}$$

$$k_{OBS} = k_{ON} \cdot k_{SAT} \cdot [Cyt] / (k_{SAT} + k_{OFF} + k_{ON} \cdot [Cyt])$$

$$k_{SAT} = k_{1} \cdot k_{ET} / k_{FAST}$$

$$k_{FAST} = k_{-1} + k_{ET}$$

$$R_{MAX} = k_{SAT} / (k_{-1} + k_{1})$$

$$PSI_{OX} + Pc_{RED} \underset{k_{OFF}}{\overset{k_{ON}}{\longleftrightarrow}} [PSI_{OX} \cdots Pc_{RED}] \overset{k_{ET}}{\longrightarrow} PSI_{RED} + Pc_{OX}$$

$$k_{OBS} = k_{ON} \cdot k_{ET} \cdot [Pc] / (k_{ET} + k_{OFF} + k_{ON} \cdot [Pc])$$

Fig. S2. Kinetic models and equations used in the fitting of the experimental data for the determination of kinetic rate constants for Phaeodactylum PSI reduction by (A) its native Cyt (27) or (B) by plastocyanins (28). k_{ET}, first-order electron transfer rate constant; k_{FAST}, observed first-order rate constant for the first fast phase of PSI reduction by Cyt; k_{OBS}, observed pseudo first-order rate constant; k_{ON} and k_{OFF}, association and dissociation rate constants, respectively, for complex formation; k_{SAT}, observed pseudo first-order rate constant extrapolated to infinite donor protein concentration (equal to k_{ET} for PSI reduction by plastocyanins); k_1 and k_{-1}, forward and reverse rate constants, respectively, for complex rearrangement; R_{MAX}, amplitude of the fast phase for PSI reduction by Cyt extrapolated to infinite donor concentration.

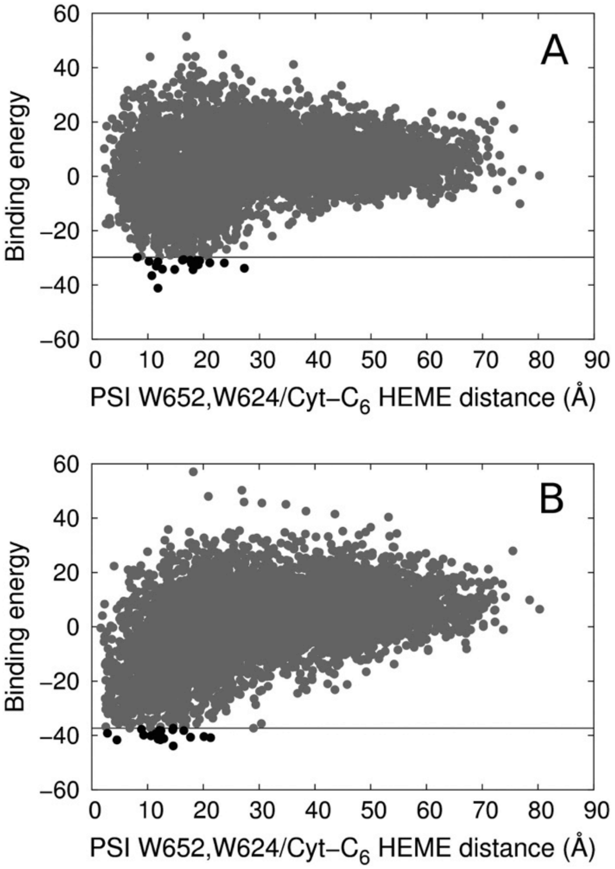


Fig. S3. Computational docking results for the interaction between *Phaeodactylum* PSI and Cyt from (A) *Phaeodactylum* or (B) *Monoraphidium*. The 20 lowest-energy docking orientations are highlighted.

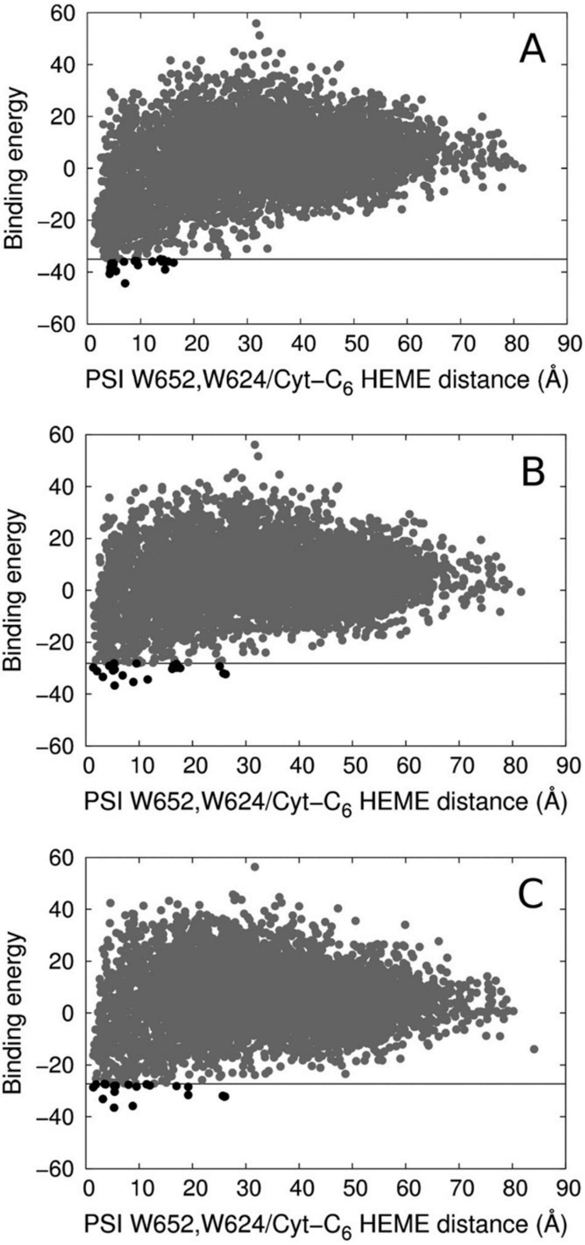


Fig. S4. Computational docking results for the interaction between *Phaeodactylum* PSI and *Chlamydomonas* WT Pc (A), or the E85K (B) and Q88R (C) mutants. The 20 lowest-energy docking orientations are highlighted.

3.4. Description of protein plasticity: an example of biomedical interest

Understanding the effects of pathologic mutations on protein function at molecular level requires the consideration of several factors, such as protein stability, molecular recognition or conformational flexibility. Protein kinases constitute a paradigmatic example of the close link between dynamics and function. These enzymes regularly switch between distinctive inactive and active states undergoing large conformational changes, whose complete computational description is still highly challenging.

This section will report the application of (i) large-scale conventional Molecular Dynamics and (ii) state-of-the-art enhanced sampling with metadynamics to elucidate the structural and dynamic basis of several protein kinase dysfunctional mutations involved in severe pathologies.

Manuscripts presented in this section:

1. Pallara C, Glaser F, Fernández-Recio J. **Structural and dynamic effects of MEK1 pathological mutations (I): unphosphorylated apo and phosphorylated ATP-bound.** (in preparation)

2. Pallara C, Sutto L, Gervasio FL, Fernández-Recio J. **Structural and dynamic effects of MEK1 pathological mutations (II): enhanced sampling Metadynamics simulations.** (in preparation)

3.4.1. Structural and dynamic effects of MEK1 pathological mutations (I): unphosphorylated apo and phosphorylated ATP-bound

Chiara Pallara, [1] Fabian Glaser,[2] Juan Fernández-Recio[1*]

[1]Joint BSC-CRG-IRB Research Program in Computational Biology, Barcelona Supercomputing Center, Barcelona, Spain.

[2]Bioinformatics Knowledge Unit, The Lorry I. Lokey Interdisciplinary Center for Life Sciences and Engineering Technion, Israel

* Corresponding author

Structural and dynamic effects of MEK1 pathological mutations (I): unphosphorylated apo and phosphorylated ATP-bound

Chiara Pallara, [1] Fabian Glaser,[2] Juan Fernández-Recio[1]*

[1]Joint BSC-CRG-IRB Research Program in Computational Biology, Life Sciences Department, Barcelona Supercomputing Center, Barcelona, Spain

[2]Bioinformatics Knowledge Unit, The Lorry I. Lokey Interdisciplinary Center for Life Sciences and Engineering Technion, Israel

*Corresponding author

Abstract

MAPK/ERK signaling pathway constitutes one of the principal and better known signal transduction processes in cells, which is involved in the tight regulation of many biological events such as cell proliferation, differentiation and apoptosis. Dysregulation of MEK1/2, a key component of this cascade, is known to be related to different pathologies including several cancer types (melanoma, lung and ovarian cancer) or many congenital RASopathies, such as the Cardio-Facio-Cutaneous (CFC) syndrome. Apart from the conventional hallmarks shared within essentially all known protein kinases, MEK1 N-lobe features a peculiar N-terminal α-helix (αA-helix), which has been reported to play a negative regulatory role on MEK1 catalytic activity. In agreement with these findings, some oncogenic mutations, as well as all the mutations associated to CFC-syndrome described so far, lie on or face to this regulatory helix and result in MEK1 overactivation. Despite the known biomedical consequences of such specific mutations, the mechanistic explanation underlying their functional impact remains elusive. Hence, we used Molecular Dynamics (MD) simulations to investigate the structural and dynamic effects in different biologically relevant states of MEK1 protein kinase of selected pathological mutations related to cancer and/or CFC syndrome. In light of the present results, all the mutations described here seem to favor the transition from the inactive to active state by either increasing αA-helix structural flexibility or promoting the close-

to-open transition of the activation loop. This study provides a better understanding of the effects of these mutations at molecular level.

Introduction

The MAPK/ERK signaling cascade is a central cellular pathway that controls several biological processes such as proliferation, differentiation, development, and, under some conditions, also apoptosis. During the last decades, different studies confirmed that up regulation of this pathway plays a key role in the pathogenesis and progression of various diseases, including many cancer types (e.g., pancreas, colon, lung, ovary) [1-12]. Moreover, it was recently observed that germline mutations on this pathway are associated with a class of developmental disorders, the so-called RASopathies [13-16], which include cardio-facio-cutaneous (CFC) syndrome [17-18]. This is a rare genetic condition that typically affects the heart (cardio), facial features (facio) and skin (cutaneous) and which is generally associated with a varying degree of learning difficulty and developmental delay [19-25].

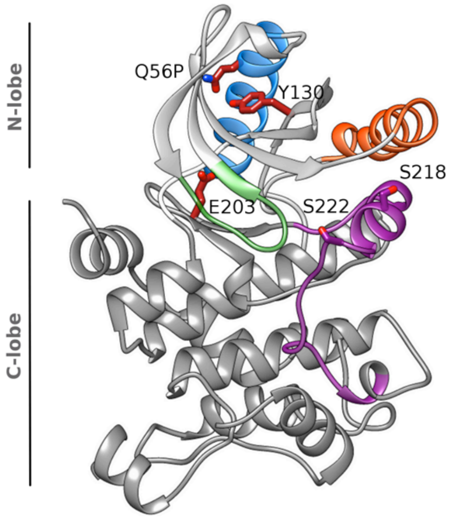


Fig 1. Structure of MEK1 protein kinase. The main structural hallmarks are shown in different colors: αA-helix in blue, αC-helix in orange, P-loop in green and A-loop in magenta. The phosphorylation sites and the mutated residues described herein (i.e., Q56, Y130 and E203) are highlighted.

Among the numerous and heterogeneous proteins involved in MAPK/ERK pathway, MEK1 or MAP2K1 (mitogen-activated protein kinase kinase 1), as well as its highly homologous MEK2, plays a crucial regulatory role since it generally functions as modulating funnel in the transmission of various up- to downstream signals [26]. Like other serine/threonine protein kinases, MEK1 protein structure contains functionally critical regions shared within essentially all known protein kinases (**Fig 1**): the Glycine-rich P-loop, consisting in a highly conserved sequence motif (GxGxGG), which locks the catalytic cleft and is involved in the ATP phosphate positioning; αC-helix, which contributes to forming or breaking the ATP binding site; and finally the activation segment (A-loop), involved in the modulation of MEK1 activity and hosting two putative phosphorylation sites (i.e., S218 and S22). Apart from these conventional hallmarks, MEK1 N-lobe features a peculiar N-terminal α-helix (αA-helix), which has been reported to play a negative regulatory role on MEK1 catalytic activity [27] related to the stabilization of αC-helix outward displacement [28].

As the majority of protein kinases, MEK1 exists in at least two conformational states generally referred to as inactive and active state, whose interconversion is driven by phosphorylation/dephosphorylation cycles. Indeed, although conserving a weak basal efficiency, MEK1 catalytic activity dramatically increases upon A-loop phosphorylation [29]. The recently solved X-ray crystal structure of MEK1 inactive conformation, as well as its similarity to the other STE group kinases have provided some hints about the structural basis for MEK1 regulatory mechanism. Large conformational changes, mainly involving A-loop and αC-helix, could be reasonably expected: (i) the highly packed and helical conformation of the activation loop in the inactive state should assume a full unpacked and extended conformation upon its phosphorylation; (ii) αC-helix, initially locked in a *αC-out* conformation should tilt toward the N-lobe, completing the catalytic cleft and thus assuming the so-called *αC-in* conformation. Such alternative orientations of αC-helix are defined by specific contacts of E114 residue, a highly conserved glutamate located on the αC-helix N-terminal region, which appears involved in a hydrogen bond with the A-loop Q214 residue or engaged in a salt bridge with K97 (the catalytic lysine located on the β3-strand) in the *αC-out* or *αC-in* conformation, respectively.

In agreement with the above mentioned regulatory role of MEK1 αA-helix, some oncogenic mutations as well as all the mutations associated to CFC-syndrome described so far lie on or face to the regulatory αA-helix [28] and lead to an increase of the MEK1 kinase activity with respect to the wild type (WT) [17]. Nevertheless clear differences have arisen between cancer and CFC syndrome related mutations. Indeed, the former are usually related to a constitutive activation of MEK1 protein kinase while the latter are associated with a milder increase in MEK1 catalytic activity [17, 30-32].

Despite all the above considerations, the mechanism underlying the MEK1 inactive-to-active transition is not yet understood at atomic level and only a few studies have been reported so far concerning the functional and mechanistic impact of some MEK1 mutations [33-34]. Here, we have studied the structural and dynamics effects in different biologically relevant states of MEK1 of selected pathological mutations related to cancer and/or CFC syndrome. For this, long Molecular Dynamics (MD) simulations were performed on the WT MEK1 and three pathological mutants associated to an increase of the MEK1 kinase activity: one lying on the regulatory helix (Q56P) and other two falling within the MEK1 kinase domain boundaries and facing αA-helix (Y130C and E203K). These mutations show different biochemical and clinical effects: Y130C is known to cause a slight increase of the MEK1 basal activity [17], occurs in one of the most documented mutation site found in CFC syndrome [35] but has never been described in any type of cancer; E203K is related to a constitutive activation of MEK1 [31, 33], occurs in 8% of melanoma [31, 36] but is not associated with CFC syndrome; finally Q56P causes a significant increase of the kinase activity *in vitro* [37], and has been reported to occur in patients with CFC syndrome [38] as well as in lung adenocarcinoma cell lines [37, 39-40].

Materials and Methods

Structural models of MEK1 variants

The crystal structure of MEK1 in an inactive conformation (PDB ID: 3EQD; resolution 2.1 Å) was used as a scaffold to model all the MEK1 variants studied here (i.e., WT, Y130C, Q56P and E203K), either in the unphosphorylated apo or in the phosphorylated ATP-bound state. All crystallographic water molecules and non-catalytic ions were removed. The unphosphorylated apo state of MEK1 was modeled by removing the AGS molecule and the Mg+ ion contained in the crystallographic structure. The phosphorylated ATP-bound state of MEK1 was modeled by substituting the co-crystallized cofactor by ATP and adding phosphate groups on S218 and S222 residues from the A-loop. Mutants, phosphorylated residues and other structural manipulations were done with UCSF Chimera program [41].

Molecular Dynamics simulation protocol

Classical MD simulations were performed on each MEK1 structure using GROMACS 4.6.7 program [42]. The CHARMM22* force field [43] was used to parameterize the proteins (together with cofactor, ions and water molecules). Phosphorylated serine residues were described by Gromos43a1p force field.

Each system was solvated in TIP3P water molecules and enclosed in a dodecahedron box with periodic boundary conditions and a minimum distance of 12 Å between the protein and the boundaries. The molecular charges were neutralized by adding the proper number of positive ions (i.e., Na+). Van der Waals interaction cutoffs were set to 10 Å and the long-range electrostatic interactions were calculated by the particle-mesh Ewald algorithm [44], with mesh spaced 1.2 Å. Each solvated system underwent a short energy minimization and a three-step equilibration (as previously described [45]) and then was used as input for three 1-μs-long unbiased simulations in isothermal-isobaric ensemble (setting the pressure to 1 atm and temperature to 300K). Finally, a total of 24 1-μs-long unbiased trajectories were collected.

GROMACS g_rms and g_mindist tools were executed to compute RMSD and residues contacts respectively. The most representative structure along the multiple simulations was selected as follows: for each system the three trajectories were merged and then clusterized according to the c-alpha atoms positions using the single-linkage clustering algorithm of *g_cluster* program from the GROMACS package and setting an RMSD cutoff value of 2 Å with each cluster, finally the most populated cluster was selected and the structure with smallest distance to all the other members was chosen.

Results

Dynamic effect of mutations on αA-helix

To compare and understand the structural, dynamics and energetics impact of selected pathological mutants that are reported to induce over-activating effects on MEK1 catalytic activity, the unphosphorylated apo and phosphorylated ATP-bound MEK1 protein kinase structural models were constructed both for the WT as well as for the three pathological mutants described here (i.e., Y130C, Q56P and E203K). Each of the eight models was then subjected to three 1-μs-long MD simulations, and the trajectories analyzed as follows.

In addition to the conventional hallmarks shared among virtually all the kinases, MEK1 N-lobe features αA-helix, which has been reported to play a negative regulatory role on MEK1 basal activity [27]. Moreover, since all the mutants studied here lie on or are located close to this helix, our first goal was to investigate the impact of the mutations on this region. To this aim, for each MEK1 variant (i.e., WT, Y130C, Q56P and E203K mutants) we calculated the RMSD of the αA-helix in the unphosphorylated apo model derived from MD simulations with respect to that in the X-ray crystal structure of MEK1 inactive state (PDB 3EQD). As shown in **Fig 2** and **Table 1**, all the mutated MEK1 structures had larger deviation compared with the WT.

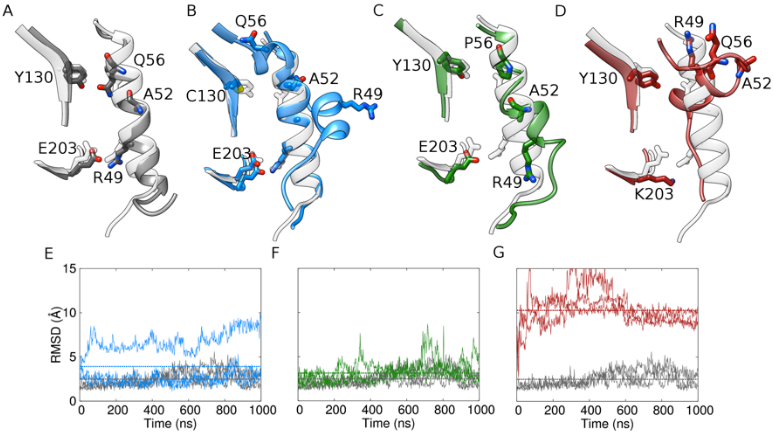


Fig 2. Dynamic effects of mutations on MEK1 αA-helix. (A-D) αA-helix most representative structures for unphosphorylated apo WT, Y130C, Q56P and E203K MEK1 simulations (in gray, blue, green and red respectively). For Y130C mutant two representative structures (corresponding to the WT-like and odd simulations) are depicted. (E-F) RMSD of the αA-helix with respect to that in 3EQD MEK1 X-ray crystal structure: WT, Y130C, Q56P and E203K (same color code as above). Dotted lines indicate the RMSD average value among the three simulations.

Table 1. Average RMSD and contact frequency values along the MD simulations.

	RMSD*$_{avg}$ *(Å)				***Pair contact frequency (%)***			
	αA-helix[a]	*P-loop*[a]	*αC-helix*[b]	*A-loop*[b]	*A52/Q56*[a]	*Q56/Y130C*[a]	*R49/E203*[a]	*E114/Q214*[b]
WT	2.5	2.6	3.8	4.7	84	76	80	83
Y130C	3.9	3.3	4.3	5.5	63	5.7	59	29
Q56P	3.2	3.6	4.3	5.7	0	18	62	54
E203K	10.3	3.0	4.4	5.9	37	1	0	45
WT	2.5	2.6	3.8	4.7	84	76	80	83

A pair contact is defined as residue-residue minimal interatomic distance < 4 Å; [a]unphosphorylated apo MEK1 structure; [b]phosphorylated ATP-bound MEK1 structure. RMSD values are referred to backbone atoms

Interestingly, not only these findings appeared in agreement with the increase of MEK1 activity caused by the mutations (as previously reported) but also confirmed the differences in the extent of the mutants impact. Indeed, the conformational flexibility led by E203K mutation (the one with the highest MEK1 constitutively activation) appeared much higher with respect to Y130C and Q56P. However, it is important to notice that while all the Q56P and E203K trajectories showed quite comparable behaviors, two out of three Y130C simulations results similar to the WT system while the third exhibited much higher fluctuations. These data suggested that the effects induced by this mutation might be milder than suggested by its average RMSD value.

In order to obtain more in-depth understanding on the effects caused by the mutations, we analyzed more in detail the structural changes involving αA-helix as well as some crucial contacts between such helix and the kinase core (**Table 1** and **Fig S1**). As shown in **Fig 2**, αA-helix remained folded and tightly packed against the kinase core region along the WT trajectories. On the contrary, all the mutations induced weakening of αA-helix/core contacts typically associated with its partial unfolding. Regarding Y130C case, the tyrosine residue in the WT structure makes extensive hydrophobic contacts with F53 and K57, and a long-range hydrogen bond with Q56 side chain, all located on αA-helix. All these interactions are deeply affected by the mutation of the tyrosine to a cysteine, a much smaller and polar residue, leading to the weakening of the contacts between αA-helix and the core region. Moreover, the loose of the contact with Q56 residue is combined with the break of the salt bridge between K203 and R49 and perturbation of Q56/A52 backbone hydrogen bond. Unlike Y130, Q56 is located on αA-helix and is involved in fewer interactions with the kinase domain. However the introduction of a proline residue on this position causes a structural disruption of the helix by introducing a strong kink, which completely displaces the N-terminal half of the αA-helix and strongly impact αA-helix/core interaction. Finally, the dramatic consequences associated to the mutation of E203 residue to a lysine might be explained not only by the disruption of the salt bridge between E203K and R49 but also by the electrostatic repulsion between the positively charged K203 and R49 residues.

Increase of P-loop flexibility upon mutations

Phosphate-binding loop (P-loop) plays a crucial role in the cofactor binding in virtually all protein kinases, acting as a lid for the ATP binding pocket. Moreover its opening may be useful to facilitate ATP intake in the apo state. For this reason, we investigated the effects of the mutations on P-loop flexibility along the MD trajectories. Indeed, the MD simulations showed that all mutants had much broader deviations for the residues of the P-loop as compared to the WT (**Fig 3E-G**) (**Table 1**). Interestingly, since the slowest step in the phosphorylation cycle corresponds in most kinases to product release [46], such increase of the P-loop intrinsic flexibility might be related to a higher cofactor turnover a thus explain the over-activation caused by the mutations. Moreover, since steric blockage of the binding site is a very common mechanism used by kinases to maintain their inactive state, in the apo state P-loop is likely to collapses onto the C-lobe, adopting a conformation that disfavor cofactor binding and restricting the ATP binding site. Actually we found that in WT, as well as in Y130C and E203K simulations, P-loop tended to flop into the active site, adopting a conformation that would have a significant number of steric clashes with ATP. On the contrary, in Q56P mutation P-loop conformation would have fewer clashes with ATP (**Fig 3 A, D**).

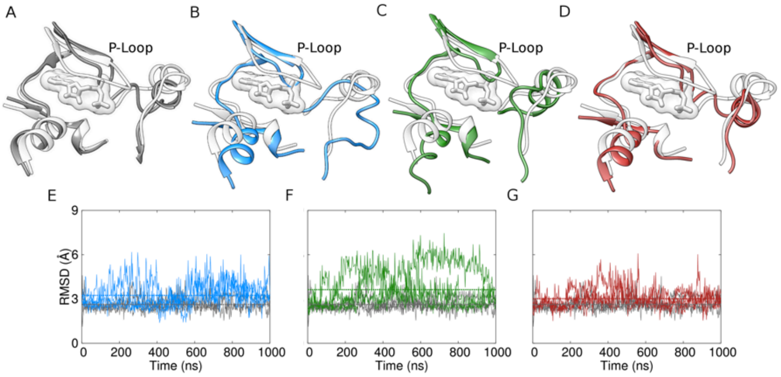


Fig 3. Effects of mutations on P-loop flexibility. (A-D) P-loop most representative structures for unphosphorylated apo WT, Y130C, Q56P and E203K MEK1 simulations (in gray, blue, green and red, respectively). (E-F) Residue RMSD of the P-loop with respect to 3EQD MEK1 X-ray crystal structure: WT, Y130C, Q56P and E203K (same color code as above). Dotted lines indicate the RMSD average value among the three simulations.

Conformational changes in the activation loop (A-loop) upon phosphorylation

It is well known that phosphorylation is required for the activation of many kinases [47] and the activation loop (A-loop) undergoes conformational changes upon switching between inactive and active state of the kinase. The number and the location of phosphorylation sites within A-loop vary among kinases [48]. S218 and S222 are known phosphorylation sites in the activation loop of MEK1 protein kinases [29].

First of all, we were interested in understanding the structural changes in A-loop (residues 207-233) in MEK1 protein kinases upon phosphorylation in physiological conditions. Thus, we compared the structural flexibility of A-loop along the MD trajectories between the wild type of unphosphorylated apo MEK1 and phosphorylated ATP-bound MEK1. Interestingly, we found that MEK1 phosphorylation induced a significant increase of the conformational flexibility not only in the activation segment but also in the αC-helix, as indicated by their average RMSD values along the MD simulation, switching from 3.8 to 4.7 Å and from 2.4 to 3.8 Å, respectively (**Fig 4 A, B**).

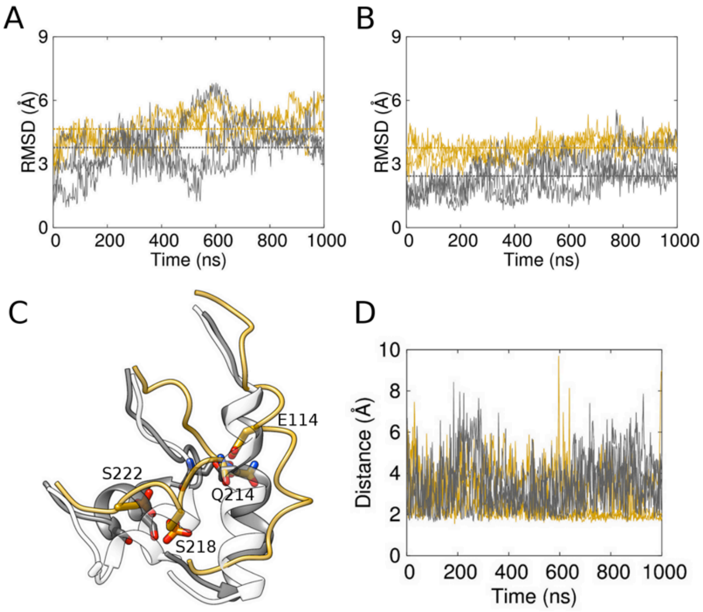


Fig 4. A-loop and αC-helix conformational flexibility upon phosphorylation. (A) A-loop and (B) αC-helix RMSD in unphosphorylated apo (grey) and phosphorylated ATP-bound (yellow) MEK1 states with respect to the X-ray crystal structure (3EQD) of MEK1 inactive state; (C) Most representative structures along the MD simulation for unphosphorylated apo (gray) and phosphorylated ATP-bound MEK1 (yellow) states as compared to the X-ray crystal structure (white); (D) Minimal distance between Q214 backbone and E114 side chain residues (same color code as above). Dotted lines indicate the RMSD average value among the three simulations.

As shown in **Fig 4C**, considerable structural variations occurred both in A-loop and in αC-helix upon phosphorylation. Indeed, A-loop underwent a partial unfolding of the highly packed and helical conformation adopted in the X-ray crystal structure, whereas αC-helix suffered a significant distortion and outward shift. Nevertheless, we could not observe a complete inactive-to-active transition of A-loop even upon phosphorylation. Indeed, is still assumed a rather packed conformation strongly interacting with αC-helix, which, in turn, appeared locked in a αC-out orientation by a highly stable hydrogen bond between E114 and Q214 residues (**Fig 4D**). Finally, it is interesting to notice that also in absence of any phosphorylation, the A-loop maintained a rather considerable instability, undergoing a slight tilt towards C-terminal lobe. These finding might explain the existence of a significant, although low basal activity of MEK1 protein kinase [29].

Mutations favor MEK1 inactive-to-active transition

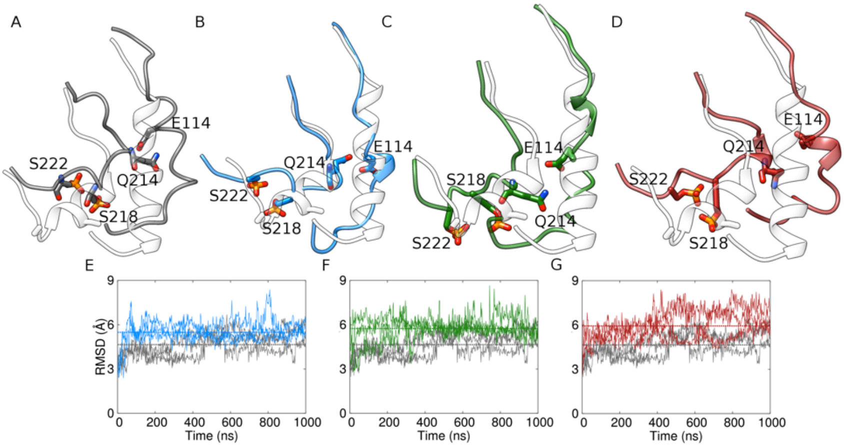


Fig 5. Effects of mutations on A-loop flexibility. (A-D) A-loop most representative structures for unphosphorylated apo WT, Y130C, Q56P and E203K MEK1 simulations (in gray, blue, green and red respectively). (E-F) RMSD of the A-loop with respect to that in 3EQD MEK1 X-ray crystal structure for WT, Y130C, Q56P and E203K (same color code as above). Dotted lines indicate the RMSD average value among the three simulations.

The above results clearly suggest that A-loop phosphorylation triggers structural variation in both the activation segment and αC-helix, which are supposed to facilitate transition to the active state. Thus, in order to understand the functional impact of the mutations on such process, we compared the structural variations in the activation segment and the αC-helix upon phosphorylation for the mutated system with the ones observed in the WT. The RMSD plots showed that any mutated system induced larger deviations compared with the WT in both the A-loop (**Fig 5E-G**) as well as the αC-helix (**Fig S2A**). Moreover, some differences in the extent of the effects led by the mutations could be disclosed, as indicated by the average RMSD values of A-loop along the simulations: more significant effects were led by E203K mutant, followed by Q56P and finally Y130C (**Table 1**).

To examine in more detail the structural consequences induced by the mutations, we compared the most populated A-loop conformations explored along the simulation by the WT and mutated systems. Actually, none of the mutated structures underwent a complete inactive-to-active transition during any of the three simulations and the extent of the induced A-loop unfolding resulted comparable with that exhibited by the WT. (**Fig 5A-D**). Nevertheless, they did induce a dramatic destabilization of the hydrogen bond between αC-helix E114 and A-loop Q214 residues, whose interacting frequency dropped from 85% in the WT to a range between 29% and 54% for the mutants (**Table 1** and **Fig S2**). These data suggested that not only the mutations cause an increase of the A-loop flexibility but even more interestingly they could result in an acceleration of the inactive-to-active transition process, which would prompt the formation of the salt bridge between E114 and the catalytic lysine K97.

Discussion

In order to investigate the intrinsic propensity for inactive-to-active transition upon either oncogenic or CFC syndrome-related mutations, 1-μsec MD simulations were performed on two biologically relevant forms (i.e., the unphosphorylated apo and the phosphorylated ATP-bound state) of MEK1 protein kinase for the WT and three pathological mutants: Y130C, exclusively associated with CFC-syndrome, E203K exclusively related to cancer and Q56P observed in both the diseases.

In light of the present results, all the mutations described here seemed to favor the transition from inactive to active state in MEK1 protein kinase. All the mutations resulted in increasing αA-helix structural flexibility and promoting the close-to-open transition of the activation loop. However the effects produced by E203K resulted more dramatic as compared to Q56P and even more with respect to Y130C. Indeed these findings are in agreement not only with the constitutive activation induced by E203K, but also with the pathological degrees of the different mutations. Finally a distinctive over-activating effect was found in Q56P with respect to Y103C and E203K mutants involving the increase of P-loop flexibility that could be related to the promotion of the cofactor turnover.

Conclusions

The findings described herein can help to rationalize and quantify the activating effects induced by Y130C, Q56P and E203K mutations offering a mechanistic explanation to the different extent of MEK1 over-activation observed for each specific mutation.

References

1. Dhillon AS, Hagan S, Rath O, Kolch W. MAP kinase signalling pathways in cancer. Oncogene. 2007 May 14;26(22):3279-90.
2. Shields JM, Pruitt K, McFall A, Shaub A, Der CJ. Understanding Ras: 'it ain't over 'til it's over'. Trends Cell Biol. 2000 Apr;10(4):147-54.
3. Davies H, Bignell GR, Cox C, Stephens P, Edkins S, Clegg S, et al. Mutations of the BRAF gene in human cancer. Nature. 2002 Jun 27;417(6892):949-54.
4. Rajagopalan H, Bardelli A, Lengauer C, Kinzler KW, Vogelstein B, Velculescu VE. Tumorigenesis: RAF/RAS oncogenes and mismatch-repair status. Nature. 2002 Aug 29;418(6901):934.
5. Mercer KE, Pritchard CA. Raf proteins and cancer: B-Raf is identified as a mutational target. Biochim Biophys Acta. 2003 Jun 5;1653(1):25-40.
6. Singer G, Oldt R, 3rd, Cohen Y, Wang BG, Sidransky D, Kurman RJ, et al. Mutations in BRAF and KRAS characterize the development of low-grade ovarian serous carcinoma. J Natl Cancer Inst. 2003 Mar 19;95(6):484-6.
7. Vos MD, Martinez A, Ellis CA, Vallecorsa T, Clark GJ. The pro-apoptotic Ras effector Nore1 may serve as a Ras-regulated tumor suppressor in the lung. J Biol Chem. 2003 Jun 13;278(24):21938-43.
8. Sieben NL, Macropoulos P, Roemen GM, Kolkman-Uljee SM, Jan Fleuren G, Houmadi R, et al. In ovarian neoplasms, BRAF, but not KRAS, mutations are restricted to low-grade serous tumours. J Pathol. 2004 Mar;202(3):336-40.
9. Hingorani SR, Jacobetz MA, Robertson GP, Herlyn M, Tuveson DA. Suppression of BRAF(V599E) in human melanoma abrogates transformation. Cancer Res. 2003 Sep 1;63(17):5198-202.
10. Sharma A, Trivedi NR, Zimmerman MA, Tuveson DA, Smith CD, Robertson GP. Mutant V599EB-Raf regulates growth and vascular development of malignant melanoma tumors. Cancer Res. 2005 Mar 15;65(6):2412-21.
11. Hoeflich KP, Gray DC, Eby MT, Tien JY, Wong L, Bower J, et al. Oncogenic BRAF is required for tumor growth and maintenance in melanoma models. Cancer Res. 2006 Jan 15;66(2):999-1006.
12. Sumimoto H, Hirata K, Yamagata S, Miyoshi H, Miyagishi M, Taira K, et al. Effective inhibition of cell growth and invasion of melanoma by combined suppression of BRAF (V599E) and Skp2 with lentiviral RNAi. Int J Cancer. 2006 Jan 15;118(2):472-6.

13. Tidyman WE, Rauen KA. The RASopathies: developmental syndromes of Ras/MAPK pathway dysregulation. Curr Opin Genet Dev. 2009 Jun;19(3):230-6.

14. Sarkozy A, Carta C, Moretti S, Zampino G, Digilio MC, Pantaleoni F, et al. Germline BRAF mutations in Noonan, LEOPARD, and cardiofaciocutaneous syndromes: molecular diversity and associated phenotypic spectrum. Hum Mutat. 2009 Apr;30(4):695-702.

15. Aoki Y, Niihori T, Kawame H, Kurosawa K, Ohashi H, Tanaka Y, et al. Germline mutations in HRAS proto-oncogene cause Costello syndrome. Nat Genet. 2005 Oct;37(10):1038-40.

16. Schubbert S, Zenker M, Rowe SL, Boll S, Klein C, Bollag G, et al. Germline KRAS mutations cause Noonan syndrome. Nat Genet. 2006 Mar;38(3):331-6.

17. Rodriguez-Viciana P, Tetsu O, Tidyman WE, Estep AL, Conger BA, Cruz MS, et al. Germline mutations in genes within the MAPK pathway cause cardio-facio-cutaneous syndrome. Science. 2006 Mar 3;311(5765):1287-90.

18. Niihori T, Aoki Y, Narumi Y, Neri G, Cave H, Verloes A, et al. Germline KRAS and BRAF mutations in cardio-facio-cutaneous syndrome. Nat Genet. 2006 Mar;38(3):294-6.

19. Roberts A, Allanson J, Jadico SK, Kavamura MI, Noonan J, Opitz JM, et al. The cardiofaciocutaneous syndrome. J Med Genet. 2006 Nov;43(11):833-42.

20. Narumi Y, Aoki Y, Niihori T, Neri G, Cave H, Verloes A, et al. Molecular and clinical characterization of cardio-facio-cutaneous (CFC) syndrome: overlapping clinical manifestations with Costello syndrome. Am J Med Genet A. 2007 Apr 15;143A(8):799-807.

21. Aoki Y, Niihori T, Narumi Y, Kure S, Matsubara Y. The RAS/MAPK syndromes: novel roles of the RAS pathway in human genetic disorders. Hum Mutat. 2008 Aug;29(8):992-1006.

22. Rodriguez-Viciana P, Rauen KA. Biochemical characterization of novel germline BRAF and MEK mutations in cardio-facio-cutaneous syndrome. Methods Enzymol. 2008;438:277-89.

23. Kratz CP, Rapisuwon S, Reed H, Hasle H, Rosenberg PS. Cancer in Noonan, Costello, cardiofaciocutaneous and LEOPARD syndromes. Am J Med Genet C Semin Med Genet. 2011 May 15;157C(2):83-9.

24. Rauen KA, Banerjee A, Bishop WR, Lauchle JO, McCormick F, McMahon M, et al. Costello and cardio-facio-cutaneous syndromes: Moving toward clinical trials in RASopathies. Am J Med Genet C Semin Med Genet. 2011 May 15;157C(2):136-46.

25. Anastasaki C, Rauen KA, Patton EE. Continual low-level MEK inhibition ameliorates cardio-facio-cutaneous phenotypes in zebrafish. Dis Model Mech. 2012 Jul;5(4):546-52.

26. Herskowitz I. MAP kinase pathways in yeast: for mating and more. Cell. 1995 Jan 27;80(2):187-97.

27. Mansour SJ, Candia JM, Matsuura JE, Manning MC, Ahn NG. Interdependent domains controlling the enzymatic activity of mitogen-activated protein kinase kinase 1. Biochemistry. 1996 Dec 3;35(48):15529-36.

28. Fischmann TO, Smith CK, Mayhood TW, Myers JE, Reichert P, Mannarino A, et al. Crystal structures of MEK1 binary and ternary complexes with nucleotides and inhibitors. Biochemistry. 2009 Mar 31;48(12):2661-74.

29. Yan M, Templeton DJ. Identification of 2 serine residues of MEK-1 that are differentially phosphorylated during activation by raf and MEK kinase. J Biol Chem. 1994 Jul 22;269(29):19067-73.

30. Brunet A, Pages G, Pouyssegur J. Constitutively active mutants of MAP kinase kinase (MEK1) induce growth factor-relaxation and oncogenicity when expressed in fibroblasts. Oncogene. 1994 Nov;9(11):3379-87.

31. Delaney AM, Printen JA, Chen H, Fauman EB, Dudley DT. Identification of a novel mitogen-activated protein kinase kinase activation domain recognized by the inhibitor PD 184352. Mol Cell Biol. 2002 Nov;22(21):7593-602.

32. Bentivegna S, Zheng J, Namsaraev E, Carlton VE, Pavlicek A, Moorhead M, et al. Rapid identification of somatic mutations in colorectal and breast cancer tissues using mismatch repair detection (MRD). Hum Mutat. 2008 Mar;29(3):441-50.

33. Nikolaev SI, Rimoldi D, Iseli C, Valsesia A, Robyr D, Gehrig C, et al. Exome sequencing identifies recurrent somatic MAP2K1 and MAP2K2 mutations in melanoma. Nat Genet. 2012 Feb;44(2):133-9.

34. Kiel C, Serrano L. Structure-energy-based predictions and network modelling of RASopathy and cancer missense mutations. Mol Syst Biol. 2014;10:727.

35. Dentici ML, Sarkozy A, Pantaleoni F, Carta C, Lepri F, Ferese R, et al. Spectrum of MEK1 and MEK2 gene mutations in cardio-facio-cutaneous syndrome and genotype-phenotype correlations. Eur J Hum Genet. 2009 Jun;17(6):733-40.

36. Bromberg-White JL, Andersen NJ, Duesbery NS. MEK genomics in development and disease. Brief Funct Genomics. 2012 Jul;11(4):300-10.

37. Emery CM, Vijayendran KG, Zipser MC, Sawyer AM, Niu L, Kim JJ, et al. MEK1 mutations confer resistance to MEK and B-RAF inhibition. Proc Natl Acad Sci U S A. 2009 Dec 1;106(48):20411-6.

38. Procaccia S, Seger R. Mek. In: Choi S, editor. Encyclopedia of Signaling Molecules: Springer New York; 2012. p. 1051-8.

39. Marks JL, Gong Y, Chitale D, Golas B, McLellan MD, Kasai Y, et al. Novel MEK1 mutation identified by mutational analysis of epidermal growth factor receptor signaling pathway genes in lung adenocarcinoma. Cancer Res. 2008 Jul 15;68(14):5524-8.

40. Arcila ME, Drilon A, Sylvester BE, Lovly CM, Borsu L, Reva B, et al. MAP2K1 (MEK1) Mutations Define a Distinct Subset of Lung Adenocarcinoma Associated with Smoking. Clin Cancer Res. 2015 Apr 15;21(8):1935-43.

41. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, et al. UCSF Chimera--a visualization system for exploratory research and analysis. J Comput Chem. 2004 Oct;25(13):1605-12.

42. Hess B, Kutzner C, van der Spoel D, Lindahl E. GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. Journal of Chemical Theory and Computation. [doi: 10.1021/ct700301q]. 2008 2008/03/01;4(3):435-47.

43. Piana S, Lindorff-Larsen K, Shaw DE. How robust are protein folding simulations with respect to force field parameterization? Biophys J. 2011 May 4;100(9):L47-9.

44. Essmann U, Perera L, Berkowitz ML, Darden T, Lee H, Pedersen LG. A smooth particle mesh Ewald method. The Journal of Chemical Physics. 1995;103(19):8577-93.
45. Sutto L, Gervasio FL. Effects of oncogenic mutations on the conformational free-energy landscape of EGFR kinase. Proc Natl Acad Sci U S A. 2013 Jun 25;110(26):10616-21.
46. Adams JA. Kinetic and catalytic mechanisms of protein kinases. Chem Rev. 2001 Aug;101(8):2271-90.
47. Steichen JM, Iyer GH, Li S, Saldanha SA, Deal MS, Woods VL, Jr., et al. Global consequences of activation loop phosphorylation on protein kinase A. J Biol Chem. 2010 Feb 5;285(6):3825-32.
48. Johnson LN, Noble ME, Owen DJ. Active and inactive protein kinases: structural basis for regulation. Cell. 1996 Apr 19;85(2):149-58.

Supporting information

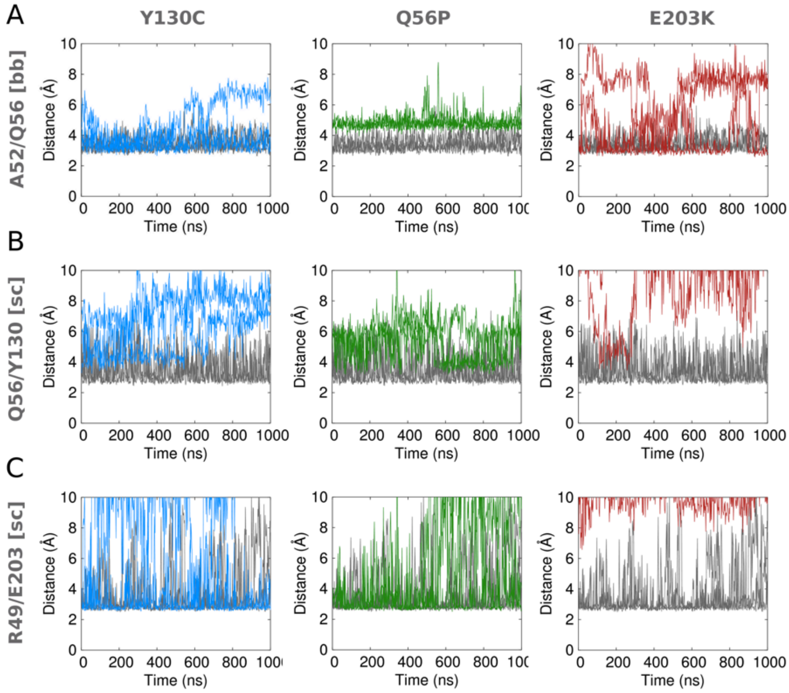


Fig S1. Molecular interaction between αA-helix and kinase core. Residue-residue minimal distance of (A) A52/Q56, (B) Q56/Y130C and (C) R49/E203 contacts. WT, Y130C, Q56P and E203K unphosphorylated apo MEK1 data are represented in gray, blue, green and red respectively.

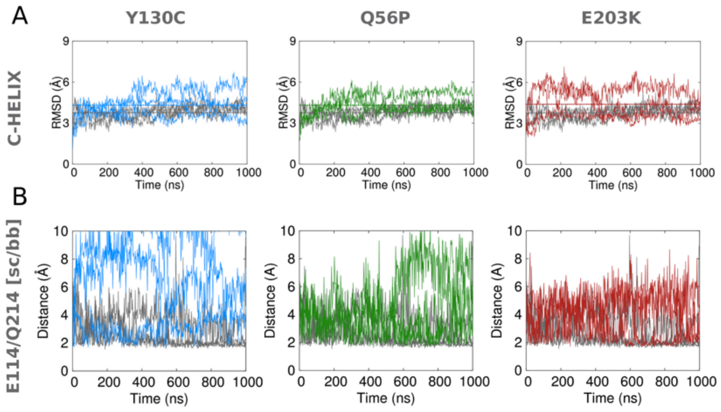


Fig S2. Effects of mutations on αC-helix flexibility. (A) αC-helix RMSD with respect to the X-ray crystal structure (3EQD) of MEK1 inactive state; (B) Minimal distance between Q214 backbone and E114 side chain residues. WT, Y130C, Q56P and E203K phosphorylated ATP-bound MEK1 data are in gray, blue, green and red respectively.

3.4.2. Structural and dynamic effects of MEK1 pathological mutations (II): enhanced sampling Metadynamics simulations

Chiara Pallara,[1] Ludovico Sutto,[2] Francesco Luigi Gervasio,[2*] Juan Fernández-Recio,[1*]

[1]*Joint BSC-CRG-IRB Research Program in Computational Biology, Barcelona Supercomputing Center, Barcelona, Spain.*

[2]*Department of Chemistry, University College London, London, U.K.*

* Corresponding authors

Structural and dynamic effects of MEK1 pathological mutations (II): enhanced sampling Metadynamics simulations

Chiara Pallara,[1] Ludovico Sutto,[2] Francesco Luigi Gervasio,[2*] Juan Fernández-Recio,[1*]

[1]Joint BSC-CRG-IRB Research Program in Computational Biology, Life Sciences Department, Barcelona Supercomputing Center, Barcelona, Spain

[2]Department of Chemistry, University College London, London, U.K.

*Corresponding authors

Abstract

Protein kinases are key regulators of eukaryotic living cells, since they are involved in crucial biochemical functions and signaling networks. These enzymes share a common fold and, in response to specific cellular signals, switch between distinctive inactive and active states undergoing large conformational changes. Describing these large conformational changes exhaustively is highly challenging. One relevant example within the protein kinase family is MEK1 (Mitogen-activated protein kinase kinase 1), involved in the MAPK/ERK pathway. Given its regulatory role in many important cell processes (such as gene expression, cell differentiation and apoptosis), over-activating mutations on MEK1 are known to cause different serious pathologies, such as several cancer types (melanoma, lung and ovarian cancer) or different congenital anomaly disorders, like the Cardio-Facio-Cutaneous (CFC) syndrome. Thus, a comprehensive elucidation of the large-scale conformational transitions that rule MEK1 functional mechanism is of great value to identify druggable spots, which play a key role during such motions and thus guide the rational design of selective inhibitors. In order to investigating the intrinsic propensity for inactive-to-active transition of MEK1 of pathological mutations, we recently employed conventional Molecular Dynamics (MD) on MEK1 in different biologically relevant states. Indeed, the analysis of such simulations helped to disclose interesting dynamic events regarding MEK1 activation process and the specific destabilization effects caused by each mutation. Nevertheless, given the long time scales involved in the protein kinases activation process

and the intrinsic limitations of the MD simulations in the exhaustive sampling of biomolecule energy landscape, some important aspects of this topic remained hidden. Therefore the main purpose of the present study consisted on enhancing the conformational exploration by using a state-of-the-art enhanced sampling method, such as PT-MetaD.

The present findings, combined with those previously reported, could not only help to rationalize and quantify the activating effects induced by pathological mutations but also offer a mechanistic explanation to the different extent of MEK1 over-activation observed for oncogenic or CFC-related mutations and eventually open the path for the development of disease specific therapeutic approaches.

Introduction

One of the largest and most functionally diverse protein families, kinases represent key regulators of eukaryotic living cells, since they are involved in crucial biochemical functions and signaling networks. Among the over 500 already characterized members, the large majority of human protein kinases, has been found to share a common fold and switch between distinctive inactive and active states in response to specific cellular signals [1].

One relevant example within this multifunctional protein family is MEK1 (Mitogen-activated protein kinase kinase 1), involved in the MAPK/ERK pathway. As a result of its regulatory role in many important cell processes, like gene expression, cell differentiation and apoptosis, deregulation of MEK1/2 is known to cause several important pathologies. At least 15 activating mutations of MEK1/2 are associated with the CFC syndrome [2-3], while at least other 10 have been identified in several cancer types (melanoma, lung and ovarian cancer) [4-15]. As the majority of the protein kinases, MEK1 is known to exist in at least two states, associated to either high or low catalytic activity and fine-tuned by protein phosphorylation. These two states can adopt different forms (usually referred as active and inactive) structurally featured by specific hallmarks described as follow. By analogy with other STE group kinases, MEK1 activation is thought to be mainly controlled by extensive conformational arrangements in three conserved structural motifs close to the active site: the activation loop (A-loop), the Asp-Phe-Gly (DFG) motif and αC-helix (**Fig 1**).

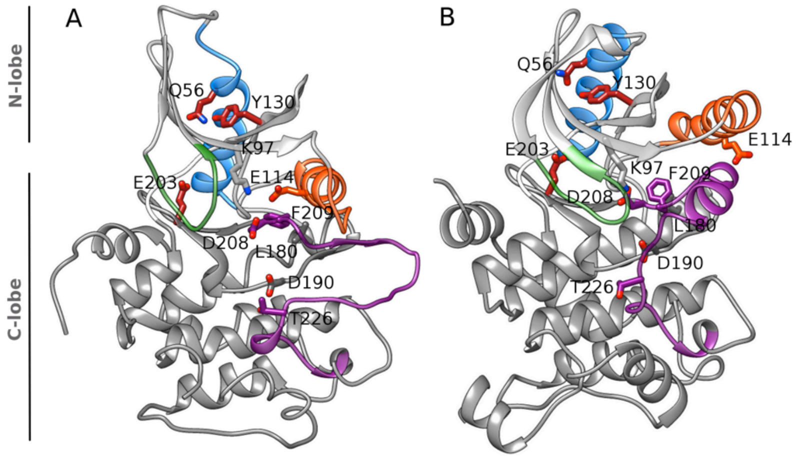


Fig 1. Comparison of (A) the active homology model and (B) inactive crystal structure (PDB 3EQD) of MEK1 protein kinase. Key structural elements are colored in blue (αA-helix), orange (αC-helix), green (P-Loop) and magenta (A-loop). Residues either high-conserved or used to define the collective variables (K97, E114, L180, D190, D208, F209, T226) are shown in stick. The mutated residues described herein (i.e., Q56, Y130 and E203) are in red. It is noteworthy that, although the inactive state is generally marked by the *DFG-out* conformation, 3EQD crystal structure shows D208 pointing inward the ATP binding site.

More in detail, as virtually all the protein kinases, MEK1 active state conformation is characterized by a highly unpacked and extended conformation of A-loop, exhibiting a hairpin in its N-terminal region. Such peculiar orientation is fixed by a conserved hydrogen bond between the carboxyl of the D190 and the hydroxyl group of T226 (i.e., the aspartic acid of the HRD motif and the threonine of the conserved A-loop GT motif, respectively) which in turn is responsible for the correct orientation of the P-site hydroxyl acceptor group of the substrate during the MEK1/ERK phosphoryl transfer reaction [16]. Moreover αC-helix tilts toward the N-lobe, completing the active site and assuming the so-called *αC-in* conformation, which is sealed by a salt bridge between E114, the conserved glutamate of αC-helix and K97, the catalytic lysine located on the β3-strand. Finally, the aspartate of the DFG motif assumes a conformation generally referred as *DFG-in*, facing its side chain into the ATP-binding pocket in order to coordinate the Mg2+ ion and properly orientate the ATP substrate.

On the contrary, in the inactive conformation, this latter interaction is often disrupted by the turning of the DFG phenilanine toward the ATP site (defined as *DFG-out* conformation) and is usually coupled with marked changes in A-loop, which adopts a highly packed and helical conformation, causing an extensive displacement of T226 hydroxyl group, that results placed about 9 Å away from D190 carboxyl. Finally, the αC-helix adopts the so-called *αC-out* conformation, rotating out of the catalytic cleft and tilts away from the N-terminal lobe. As a result, E144 side chain is unable to form the critical ion pair with the catalytic lysine, resulting involved in a hydrogen bond with the A-loop Q214 residue.

With conventional MD and the currently available computer hardware, events on the microsecond time-scale can be sampled without the need of any *a priori* knowledge of the relevant reaction coordinates, which can be useful to observe unexpected events. We recently used MD simulation to investigate the intrinsic propensity for inactive-to-active transition upon either oncogenic or CFC syndrome-related mutations in different biologically relevant states of MEK1 protein kinase. The analysis of such simulations helped to disclose interesting dynamic events regarding MEK1 activation process and the specific destabilization effects caused by each mutation. Nevertheless, given the large conformation changes involved in protein kinases activation, the complete process cannot be fully described by conventional MD and thus some important aspects may remain hidden.

Reported solutions to overcome this limitation consist in the application of novel algorithms for enhanced Molecular Dynamics (eMD), such as metadynamics [17-18], where large energy barriers are artificially reduced, allowing proteins to shift between conformations that would not be accessible within the time scales of MD. Indeed, the efficiency of such technique has been recently boosted by its combination with multiple-replica approaches, as in the case of the parallel-tempering metadynamics simulations approach (PT-MetaD), which has been successfully applied to the exploration of very complex conformational free-energy surfaces of kinases [19-21]. Thus, the application of PT-MetaD appeared a promising strategy to fully address the exhaustive exploration of MEK1 protein kinase energy landscape under both physiological and pathological conditions.

Extensive PT-MetaD simulations were performed on the wild type (WT-) MEK1 and the previously described pathological mutations. As already mentioned, they mainly differ in the pathogenic effects as well as in the extent of their impact on MEK1 catalytic activity. More in detail, Q56P mutant has been reported to be involved in CFC syndrome [3] as well as in lung adenocarcinoma cell lines [22-23], while Y130C and E203K are associated with CFC-syndrome [2] and cancer [24-25], respectively. Moreover, E203K mutation generally results in MEK1 constitutive activation [26], whereas Q56P and Y130C have been found to be related to lower overactivating effects [27-28].

Materials and Methods

Protein structural models

Inactive MEK1 structure was taken from the Protein Data Bank (www.rcsb.org) [29], with PDB code 3EQD [30] and it was used as scaffold to model the mutants studied here. Mutants and missing residues were modeled using UCSF Chimera program [31]. Although a crystal structure of a presumably activated MEK1 (3W8Q) is available in the PDB, we decided not to use it. The main reasons of this choice are explained below. Firstly the unavailability of the corresponding manuscript prevents us from knowing the protocol used for the crystallization; secondly, the DFG-motif adopts a *DFG-out* orientation that should not be found in a canonical active state. Thus, MEK1 active conformation was built by homology modeling as follows: the kinase domain (65-382 residues) was built based on the human MST3 kinase X-ray structure (PDB 3A7I) [32], while αA-helix (39-54 residues), missing in the template, was derived from the MEK1 inactive structure (PDB 3EQD). The sequence alignment between MEK1 and MST3 protein kinase was performed using BLAST [33], and then a total of 20 homology models were built using MODELLER [34]. Finally, the model with the best DOPE score was selected and subjected to 1000 steps of steepest descent energy minimization. The human MST3 kinase was used as template because it has the highest sequence similarity to MEK1 (34% according to BLAST) among the STE kinases for which a canonical active structure is available in the PDB.

Simulation setup

The MD simulations were performed using GROMACS 4.6.7 [35] with the PLUMED plug-in [36] for Metadynamics calculations. Each system was described by the CHARMM22* force field with the correction of Piana et al. [37] and enclosed in a dodecahedron box with periodic boundary conditions containing a 12 Å buffer of TIP3P water molecules. Long-range electrostatics interactions were calculated by the particle-mesh Ewald algorithm [38], with mesh spaced 1.2 Å. Both electrostatics and van der Waals interactions cutoffs were set to 10 Å. All the simulations were performed in a canonical ensemble, coupled with a velocity-rescale thermostat [39] and a time step of 2 fs.

Each solvated system was prepared performing a short energy minimization and a three-step equilibration protocol as described before [21]. Then, a preliminary parallel tempering metadynamics (PT-MetaD) was performed with 10 replicas at increasing temperatures (from 300 to 450K), biasing only the potential energy. Each Gaussian was added every 500 timesteps, with an initial hill height of 10 kJ/mol and a width of 1000 kJ/mol. The bias factor was set to 200. This protocol allowed to enlarge the energy fluctuations at different temperatures, improve the overlap of the potential energy distributions and thus sample the Well-Tempered Ensemble (WTE).

At this point, PT-WTE was combined with a metadynamics simulation (PTMetaD-WTE [40]). Two collective variables (CV1 and CV2) were used to study the inactive-to-active transition of the A-loop, namely the distance in contact map space to the closed and open A-loop conformation (as defined before [21]). Thus, when their values are close to 0, A-loop assumes a conformation similar to that found in the reference structure while higher values correspond to higher similarity to the alternative conformation. Moreover, as the DFG motif was found in an in-conformation in both the inactive and active reference structures used herein and given its relevance as hallmark to define kinases activity states, an additional collective variable (CV3) was specified describing its in-to-out transition as the combination of the distances between the center of mass of K97 and D208 as well as F209 and L180 side chain residues, setting the weighing factors to -0.73 and -0.68, respectively. Thus, values ranging around -1.0 stand for the DFG motif pointing in toward the ATP active site (*DFG-in* conformation) while values centered around -1.5 correspond to the DFG motif pointing outward (DFG-out conformation), finally values higher than -2.0 indicate intermediate DFG conformations (e.g., Asp- up or down conformation). During the PTMetaD-WTE simulations, each Gaussian was added every 1000 time steps, with an initial hill height of 10 kJ/mol and a width of 0.5 units for CV1 and CV2 and 0.1 for CV3. The bias factor was set to 10.

PTMetaD-WTE simulations analysis

The free energy surfaces (FES) of the WT and the three mutants were obtained for the replica at 300K by integrating the deposited bias during the PTMetaD-WTE simulations, as required by the metadynamics algorithm. For convenience, they are shown as function of two CVs at a time (CV1, CV2 and CV1, CV3).

To obtain the representative structure of each basin observed in all the 300K FESs, a set of structures within a small patch surrounding each basin was selected setting the range cutoff values to ±1 for CV1 and CV2 and ±0.1 for CV3 around the center. The structures were then clusterized according to the Cα atoms positions using the single-linkage clustering algorithm of *g_cluster* program from the GROMACS package [35] and setting an RMSD cutoff value of 2 Å within each cluster. The residues located on unstructured modeled loop (residues 278-306) were excluded during the RMSD calculations. Finally the most populated cluster for each basin was selected and the structure with smallest distance to all the other members was picked as the most representative structure of the basin.

Results

PTMetaD-WTE simulations on MEK1 WT and pathological mutations

To elucidate the structural and energetic consequences of Y130C, Q56P and E203K mutations on the catalytic domain of MEK1 protein kinase, four extended PTMetaD-WTE simulations were performed. The total sampling time for each system was at least 10 μs, leading to well-converged free-energy surfaces. A total of three collective variables (CVs) were selected in order to characterize both A-loop transition from closed to open state (CV1 and CV2) and the flip of DFG-motif (CV3) (see Materials and Methods). The simulations were run until the free energy projected along each collective variable at a time did not change noticeably in the last 50 ns (**Fig S1**). This convergence criterion led to 1.3 μs long simulations for WT MEK1 and to 1.0, 1.3 and 1.0 μs for each Y130C, Q56P and E203K replica, respectively. To further check the convergence, the free energy profiles in the mono-dimensional projection obtained by PLUMED module *sumhills* were compared with the ones obtained by the time-independent re-weighting technique of Tiwary and Parrinello [41]. The profiles extracted by the two methods did not change significantly and thus confirmed the convergence of the simulations (**Fig S2**).

For each system, either the WT MEK1 protein kinase and the mutants, the FES was projected along two collective variables at a time, CV1, CV2 and CV1, CV3 respectively, where CV1 and CV2 account for a broad characterization of A-loop conformation (i.e., the distance in contact map space to A-loop closed and open conformation, respectively) while CV3 describes the DFG-motif orientation. On the whole, the projections of the FES as function of CV1 and CV2 revealed that all mutations led to a considerable destabilization of A-loop closed conformation, a clear stabilization of the open state and a remarkable lowering of the free energy barrier for the closed-to-open transition. In addition, significant differences were found in the FESs projected on the CV1 and CV3 dimensions: all the mutants lead to a flattening of the free energy barrier for DFG in-to-out transition although to different extents (**Fig S3**).

Effect of the mutations on the A-loop closed-to-open transition

Wild type MEK1

In the FES projected along the first two collective variables, the deepest free-energy minimum of WT-MEK1 protein kinase corresponds to the auto-inhibited state, in which the αA-helix is folded and tightly packed against the kinase core region (**Fig 2**). This state appears strongly stabilized by a crucial salt bridges network, which acts as a clamp between the αA-helix and the N-lobe: R49, located on the αA-helix interacts with E144 (lying on the β5-strand), which in turn is linked to E203 and K205 (of the β7/β8 loop). These interactions aid in keeping the αC-helix shifted away from its active state orientation through a cascade of intramolecular contacts, which spread from the αA-helix across the entire N-terminal region and involve the hydrogen bonds between the side chain of Q56 and K57 (on the αA-helix) with Y130 and H119 (on the β4-strand and the αC-helix), respectively. Moreover, similarly to the inactive MEK1 crystal structure, the αC-out orientation is stabilized by a network of hydrophobic interactions between L215 and V211 A-loop residues with I99, L101, I111 and L115 located on αC-helix and β1-strand. As a result, the highly conserved glutamate 114 of the αC-helix is pointing out of the active site and is thus unable to form the critical ion pair with the catalytic K97 (kept at a distance of roughly 17 Å). On the contrary, the E114 outward orientation is stabilized by the formation of a hydrogen bond between the side chains of E114 and Q214.

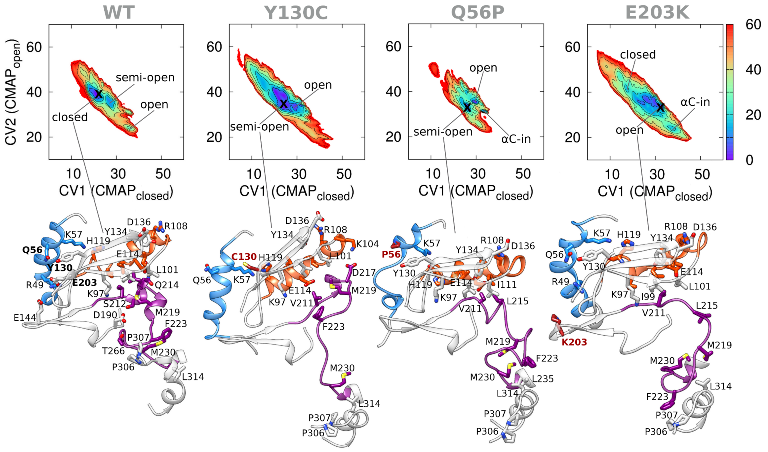


Fig 2. Free-energy surface of wild-type MEK1 and the three mutants as a function of CV1 (x-axis) and CV2 (y-axis). A cross indicates the global free energy minimum for which a representative structure is shown below. Each minimum is also named according to the corresponding main feature. The contour lines are drawn every 10kJ/mol. In the structures below, the αA-helix is show in light blue, the αC-helix in orange, the A-loop in magenta. The mutated residues are shown in red.

On the other hand, the activation loop appears blocked in a rather folded conformation, which consists in a short helix partially broken in the middle but stabilized by the crystallographic hydrogen bond between the A220 and D217 backbone atoms. This A-loop close orientation appears fixed by the formation of an additional hydrogen bond between the hydroxyl group of S212 (highly conserved among all the MEK1 family members) and the carboxyl of D190, involved in the HRD motif. D190 is thus unable to interact with the threonine of the GT motif (i.e., T226), which in turn is kept at roughly 10 Å far from the aspartate. Indeed, the D190/S212 contact, despite it has never been observed, might reasonably prevent the production of the proper A-loop rearrangement in a fully active conformation and appears in agreement with the previously observed increase of MEK1 basal activity led by the mutation of S212 to alanine [42].

Apart from the closed state minimum, two additional low-energy basins were identified, located at 10 kJ/mol and 20 kJ/mol higher energy and which correspond to semi- of fully-open A-loop structures, respectively. In the semi-open state, the A-loop assumes an intermediate conformation between canonical open and closed ones. Thus, it adopts a reasonably closed but rather unfolded rearrangement in which only one turn of the original two-turn helix is kept folded. On the contrary, the fully open state recalls the A-loop observed in several protein kinases active structures although the αC-helix appears in the *out-conformation*, rotated out of the catalytic site (**Fig S4**).

It is worth to note that none of the states found hereby corresponds to the fully active structure, which confirms that the active conformation of WT, in absent of phosphorylation, is energetically unfavorable and infrequently sampled [43]. This is further supported by the comparison between the free energy barriers extracted by the closed to semi-open, semi-open to closed and semi-open to open transition, corresponding to around 30kJ/mol, 20kJ/mol and 40kJ/mol, respectively.

Y130C-MEK1 mutant

The Y130C mutant dramatically changes the conformational free-energy landscape projected along CV1 and CV2, shifting the equilibrium toward the active conformation. Unlike the WT, Y130C FES shows only one significantly broad minimum, corresponding to a rather heterogeneous bunch of structures that mainly differ in the conformation of A-loop. The most populated cluster is marked by a partially unfolded A-loop that forms a short one-turn helix at its N-terminal (still stabilized by L215/S212 backbone hydrogen bond) and an extended disordered region at its C-terminal. As shown in **Fig 2**, the single point mutation on the 130 residue causes the break of the hydrogen bond with the αA-helix Q56 and leads to a slight outward slipping of β4 and β5-strand, locked by the formation of a new salt bridge between D136 and R108 and coupled with Y134 side chain outward flipping. This leads to the distortion and inward shift of central region of the αC-helix and thus the weakening of some hydrophobic contacts with the A-loop (e.g., V211/I111 and V211/L115) which in turn lead to a partial unfolding of the A-loop, stabilized by a new K104/D217 salt bridge. Although the extent of the αC-helix turn results to be insufficient to prompt the formation of the salt bridge between E114 and the catalytic lysine K97, it does reduce the distance between the two catalytic residues which are placed at a distance of roughly 12 Å. On the other hand, the second cluster, much less populated than the previous one, shows a full distension of the A-loop and a partial deformation of the αC-helix in which the first three turns are maintained while the other two appear totally disordered (**Fig S4**).

Interestingly, not only both the semi- and fully- open states are energetically very similar, suggesting that in the Y130C mutant, in contrast to the WT, are equally likely to occur but also they share the same basin representing alternative structures that can rapidly convert from one to the other.

Q56P-MEK1 mutant

Similarly to Y130C mutant, Q56P clearly shifts the equilibrium toward the active conformation, indicating that the mutant, unlike the WT, spends most of its time in an A-loop semi-open or open state. However, the changes observed in the conformational free-energy landscape appear even more dramatic than Y130C resulting in the complete loss of the basin corresponding to the closed state. Indeed, Q56P mutant FES shows three narrow minima located at roughly CV1 25-35, which are energetically equivalent and divided by rather similar energy barriers.

The most populated basin contains structures marked by (i) an almost-open A-loop that results in a full unfolded but not broadly extended conformation and (ii) the αC-helix rotated out of the catalytic site (*αC-out* conformation). Interestingly enough, the opening extent of A-loop caused by Q56P appears larger than the one observed in Y130C lower-energy basin. As shown in **Fig 2**, the mutation of glutamine 56 to a proline obviously prevents the formation of the native Q56/Y130 hydrogen bond and lead to an outward shift of the αA-helix, which in turn end with the switch of the H119 side chain with its backbone in the αA-helix/αC-helix hydrogen bond with K57 side chain. As a result, a slight drop on the C-terminal region of the αC-helix (stabilized by D136/R108 salt bridge, as found in Y130C) promotes the extensive unfolding of A-loop helical region. This new A-loop conformation is featured by the loss of several hydrophobic interactions with the αC-helix (such as V211/L115, V211/V117, M219/I99, L215/I111 and L215/L101) and sealed by the formation of new contacts, specific of the canonical open conformation (namely F223/L314 F223/L235 and M230/L314).

The second basin is characterized by a fully open conformation of A-loop and a partial unfolding of the αC-helix coupled with its rearrangement in an atypical orientation, which corresponds to an intermediate state rather equidistant between the canonical *αC-in* and *αC-out* conformation.

On the other hand, the third minimum corresponds to an ensemble of conformations in which the total extension of A-loop is combined by the stabilization of the αC-helix rotation toward the catalytic site similarly to many active kinases structures (**Fig S4**). Although the salt bridge between E114 and K97 is never formed, as in the fully active conformation, the side chains of the two catalytic residues are found at a minimum distance of 8.5 Å, which result remarkably lower than the one observed in the WT closed structure, equivalent to 17 Å. As shown in **Fig 3B**, this αC-helix orientation is triggered by the strong hydrophobic interaction between the αA-helix mutated residue P56 and the αC-helix L118 side chain and thus stabilized by the formation of a hydrogen bond between the H119 imidazole ring (released by the interaction with K57) and G213 backbone.

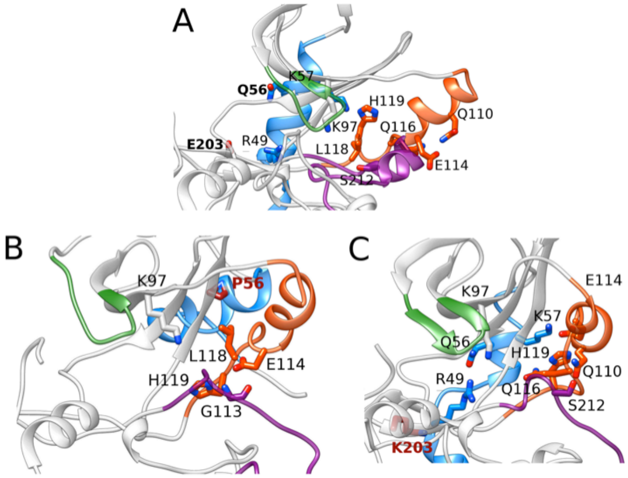


Fig 3. Comparison of (A) WT A-loop closed conformation with (B) Q56P and (C) E203K mutants structures of the free energy minima corresponding to the fully open state showing a partial αC-helix rotation toward the catalytic site (αC-in): the αA-helix is show in light blue, the P-loop in green, the αC-helix in orange, A-loop in magenta. The mutated residues are in red.

On the whole, the structural and energetic impact of Q56P mutant not only appears stronger than that exerted by Y130C mutant, but also induces additional effects on the αC-helix displacement, that were not observed in Y130C simulations. All these findings, in agreement with the effects previously observed with the MD, could be related to its deeper response on the kinase activity observed *in vitro* biochemical studies [27] and the oncogenic nature of this mutation.

E203K-MEK1 mutant

Similarly to Y130C and Q56P mutants, E203K also induces a shift toward the active conformation. However, the changes observed in the CV1/CV2 FES appear even more striking than the mutations described above. Similarly to what observed for Q56P mutant, the fully closed state appears rather destabilized. On the contrary, the lower-energy basin contains an ensemble of structures, which are marked by an almost-open A-loop (resulting fully unfolded and partially extended) combined with the αC-helix rotated out of the catalytic site (**Fig 2**). Indeed the electrostatic repulsion between the positively charged K203 and R49 residues leads to the weakening of the contacts between the αA-helix and the core region which produces very similar effects as the ones observed in Q56P mutant. Thus, the formation of a new hydrogen bond between K57 side chain amine group and H119 backbone leads to an extensive distortion of the αC-helix (stabilized by D136/R108 salt bridge, as found in the other mutants), which in turn weakens the packing contacts with the A-loop (such as V211/L115, V211/V117, L215/I111 and L215/L101) and thus boosts its unfolded and opened conformation. Moreover, similarly to Q56P, the new A-loop state appears stabilized by new hydrophobic interaction involving M230 and F223 residues (e.g., M230/L314 and F223/P306 and F223/P307 contacts)

On the other hand, the FES reveals an additional minimum at CV1=38 and CV2=22 corresponding to the rather active conformation involving the complete unfolding of A-loop that appears in the fully extended conformation (as previously observed in Y130C and Q56P simulation), as well as a significant rearrangement of the αC-helix (as found only in Q56P simulation but missing in Y130C). Interestingly, although a complete αC-helix rotation was never observed and the salt bridge between E114 and K97 is never formed, the ensemble of states found in this basin corresponds to a quite stable intermediate state, which shows a partial unfolding of the αC-helix N-terminal region that might precedes the canonical αC-helix in-conformation. As shown in **Fig 3C**, the new αC-helix orientation is triggered by the formation of the hydrogen bond between the guanidine of arginine 49 and the amide group of the glutamine 56 residues causing an extensive rearrangement of the αA-helix and thus the break of the K57/H119 native contact. As a result, the above mentioned residues appear involved in two new hydrogen bonds, engaging E114 and Q116 respectively. This causes a partial unfolding and broad inward switch of the αC-helix, which in turn is further stabilized by an additional hydrogen bond between Q110 and S212 residue. It also worth noting that, although E114 always remains at an average distance of roughly 15 Å from K97, its salt bridge with K57 observed herein might boost the inward rotation of E114 and thus prelude the E114/K97 contact formation.

In summary, the effect of E203K mutant appears deeper than that exerted by both Y130C and Q56P mutant. Moreover, as well as Q56P (but not Y130C) a significant αC-helix displacement was observed. All these findings, in agreement with the effects previously observed in the MD, could be related to the constitutive activation led by this mutation and its oncogenic nature.

Effect of the mutations on the DFG-motif in-to-out transition

Wild type MEK1

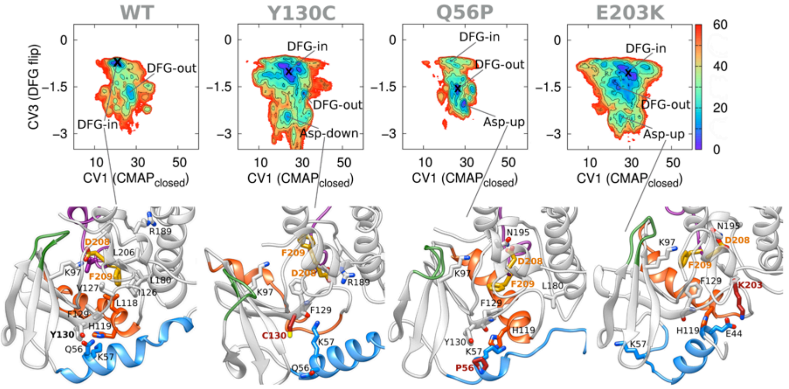


Fig 4. Free-energy surface of wild-type MEK1 and the three mutants as a function of CV1 (x-axis) and CV3 (y-axis). Each minimum is named according to the corresponding main feature. The contour lines are drawn every 10 kJ/mol. In the structures below, the αA-helix is show in light blue, the P-loop in green, the αC-helix in orange, the A-loop in magenta. DFG-motif D208 and F209 residues are shown in yellow, and the mutated residues in red.

The wild type MEK1 free energy surface computed as function of CV1 and CV3 shows two deepest free-energy minima (**Fig 4**). These two minima correspond to ensembles of inactive structures where the αA-helix is folded and tightly packed against the core region, the αC-helix is rotated out of the active site and the A-loop is rather folded. However, the global minimum corresponds to structures in which the DFG motif results in an active conformation with D208 forming a salt bridge with K95 and F209 tightly interacting with L180 (*DFG-in* conformation) and perfectly fitted in the hydrophobic pocket delimited by L118, I126, V127, L180, I186, L206 and V211.

In agreement with previous proposals that claimed the existence of certain degree of basal activity of MEK1 protein kinase, the *DFG-in* conformation was found to be energetically favorable with respect to the *DFG-out* conformation [44].

Y130C MEK1 mutant

In addition to the shift between the A-loop closed to open conformation, the free energy surface projected as a function of CV1 and CV3 also reveals a number of crucial differences between the WT and Y130C mutant (**Fig 4**). First of all, the basin corresponding to the *DFG-in* conformation is much broader than the corresponding basin for the WT, suggesting that this state can adopt many more conformations in the mutant than in the WT. However, the most remarkable change consists in a significant flattening of the free energy barrier for the *DFG-in* to *DFG-out* transition (reduced from roughly 40kJ/mol to 30kJ/mol). Indeed, this could promote the ADP release [45], which is known to be the rate-limiting step in kinase catalysis [46-47] and thus explain the catalytic activity increase observed in the mutant with respect to the WT. A further difference involves the appearing of a third local minimum, virtually absent in WT, that lies 10 kJ/mol above the other minima and corresponds to an intermediate conformation of the DFG flip transition in which the aspartate is pointing downward (Asp-down state). Interestingly, as shown in **Fig 4**, this state is triggered by the formation of a hydrogen bond between the K57 positively charged ε-amino group (released from the interaction with H119 side chain) and the backbone of F129 residue located on the underlying β4-strand. This contact stabilizes a new orientation of the bulky phenylalanine side chain, which in turn fills the position usually occupied by F209 residue when DFG-motif adopts an *in-conformation*.

Q56P MEK1 mutant

As in Y130C mutant, the DFG-motif flip also results deeply affected by the Q56P mutation. The free energy landscape computed as a function of CV1 and CV3 shows a total of three minima: the canonical DFG-out and DFG-in conformations, plus an additional local minimum, located at CV1=30 and CV3=-2.1. Interestingly, unlike

Y130C mutant, virtually all the structures corresponding to the third basin are characterized by an intermediate conformation of the DFG flip transition in which the aspartate is pointing upward (Asp-up state). As shown in **Fig 4**, the release of the αA-helix from the packing interaction with the kinase core region leads to the formation of a new hydrogen bond involving the backbone of F129 residue located on the underlying β4-strand. However, unlike the previous mutant, the contact with the H119 imidazole ring (released from the interaction with K57 side chain) stabilizes a different, less bulky orientation of the F129 side chain, that does not reject the DFG-motif aromatic residue but instead complete the hydrophobic pocket formed by I141, I139, L101, L115, V127, M143 and Y134 thus stabilizing this intermediate state.

Finally, it is worth noting that, although the free energy of the transition state between DFG-in and DFG-out conformation (equivalent to around 40 kJ/mol) does not significantly change with respect to the WT, the destabilization of the in-conformation, which lies 20kJ/mol above the out-conformation, considerably reduce the free energy barrier of the in-to-out transition (shifted from roughly 40kJ/mol to 20kJ/mol) and thus increase the ATP turnover rate. These findings could properly explain the increase in kinase activity observed with respect to both the WT and Y130C mutant. Nevertheless, DFG-out appears as a lower-energy state with respect to DFG-in conformation, which in principle would be in disagreement with what one would expect. Thus, deeper analysis will be needed in the future in order to clarify this issue.

E203K MEK1 mutant

As previously observed in A-loop close-to-open transition, the changes promoted by E203K on the CV1/CV3 free energy landscape appear more profound than the mutations described above. Similar to Y130C, E230K FES shows two energetically equivalent low-energy basins that correspond to the DFG-in and DFG-out conformations. However, as found in Q56P but not in Y130C, a third local minimum corresponding to an intermediate Asp-up state was observed. Interestingly, the molecular mechanism underlying its stabilization is virtually the same as the one previously described for Q56P, although it is triggered by different interactions. Indeed, the ion pair between E144 and K203 side chain residues favors the hydrogen bond formation between the H119 imidazole ring (released from the interaction with H119 side chain) and the backbone of F129. This in turn stabilizes a beneficial orientation of its side chain that can thus extensively interact with the DFG-motif phenylalanine and stabilize this intermediate state. Moreover, as seen for Q56P, the new orientation of D208 results further fastened by a new hydrogen bond with N195 side chain (**Fig 4**).

Nevertheless, the most significant effect promoted by E203K mutation, consists in an additional flattening of the energy barrier for the DFG in-to-out transition (reduced to roughly to 10 kJ/mol) with respect to both the mutations described herein and which could be related to the MEK1 constitutive activation experimentally observed.

Discussion

Both oncogenic and CFC related mutations are bound to lead an increase of the MEK1 kinase activity although to different extents. We recently employed large-scale Molecular Dynamics (MD) in order to investigate the intrinsic propensity for inactive-to-active transition upon such mutations in different biologically relevant states. Indeed, MD simulations helped to disclose interesting dynamic events regarding MEK1 activation process and the specific destabilization effects caused by each mutation. Nevertheless, given the large conformation changes involved in protein kinases activation process and the limits of the conventional MD simulation in the exhaustive sampling of biomolecule energy landscape, some important aspects of this topic remained hidden. Thus, in order to enhance the exploration of the energy landscape of our system, a state-of-the-art enhanced sampling method, such as PTMetaD-WTE, was applied on the same case study.

As observed through MD simulations and in agreement with the experimental data reported, all the mutations described herein result in easing the shift of the equilibrium from the inactive to active state to a different extent. As a rule, the effects produced by E203K results generally more dramatic if compered with Q56P and even more with respect to Y130C. Nevertheless, in light of the present study, additional hints about specific molecular mechanisms accounting for oncogenic and non-oncogenic mutations were disclosed.

First of all, E203K and Q56P show remarkable effects on the closed-to-open transition of A-loop, stabilizing a virtually equivalent intermediate state in which the A-loop result completely unfolded while the consequences of Y130C mutation (although clearly appreciable) results less striking, promoting the stabilization of an intermediate state in which the A-loop is still partially folded. Another significant difference arises from the αC-helix flexibility, consisting on a higher propensity for out-to-in transition observed only in oncogenic mutations (Q56P and E203K). Finally, it is worth noting the crucial effects on the DFG-motif plasticity caused by the mutations. Indeed, although by different mechanisms, all the mutants significantly flatten the energy barrier for the DFG in-to-out transition thus promoting the ADP release and increasing the ATP turnover rate. Nevertheless, the effects produced by E203K results generally more dramatic if compered with Q56P and even more with respect to Y130C.

Conclusions

In light of the present results, all the mutations described here reveal a double effect regarding the energetics of the A-loop as well as the DFG-motif: they might facilitate the protein kinase phosphorylation by stabilizing A-loop open conformations and increase the ATP turnover rate by lowering the free energy barrier of the DFG in-to-out transition. Overall, the present findings, combined to those previously reported, could help to rationalize and quantify the activating effects induced by Y130C, Q56P and E203K mutations, offer a mechanistic explanation to the different extent of MEK1 over-activation observed for oncogenic or CFC-related mutations and eventually open the path for the development of disease specific therapeutic approaches.

References

1. Manning G, Whyte DB, Martinez R, Hunter T, Sudarsanam S. The protein kinase complement of the human genome. Science. 2002 Dec 6;298(5600):1912-34.

2. Dentici ML, Sarkozy A, Pantaleoni F, Carta C, Lepri F, Ferese R, et al. Spectrum of MEK1 and MEK2 gene mutations in cardio-facio-cutaneous syndrome and genotype-phenotype correlations. Eur J Hum Genet. 2009 Jun;17(6):733-40.

3. Procaccia S, Seger R. Mek. In: Choi S, editor. Encyclopedia of Signaling Molecules: Springer New York; 2012. p. 1051-8.

4. Dhillon AS, Hagan S, Rath O, Kolch W. MAP kinase signalling pathways in cancer. Oncogene. 2007 May 14;26(22):3279-90.

5. Shields JM, Pruitt K, McFall A, Shaub A, Der CJ. Understanding Ras: 'it ain't over 'til it's over'. Trends Cell Biol. 2000 Apr;10(4):147-54.

6. Davies H, Bignell GR, Cox C, Stephens P, Edkins S, Clegg S, et al. Mutations of the BRAF gene in human cancer. Nature. 2002 Jun 27;417(6892):949-54.

7. Rajagopalan H, Bardelli A, Lengauer C, Kinzler KW, Vogelstein B, Velculescu VE. Tumorigenesis: RAF/RAS oncogenes and mismatch-repair status. Nature. 2002 Aug 29;418(6901):934.

8. Mercer KE, Pritchard CA. Raf proteins and cancer: B-Raf is identified as a mutational target. Biochim Biophys Acta. 2003 Jun 5;1653(1):25-40.
9. Singer G, Oldt R, 3rd, Cohen Y, Wang BG, Sidransky D, Kurman RJ, et al. Mutations in BRAF and KRAS characterize the development of low-grade ovarian serous carcinoma. J Natl Cancer Inst. 2003 Mar 19;95(6):484-6.
10. Vos MD, Martinez A, Ellis CA, Vallecorsa T, Clark GJ. The pro-apoptotic Ras effector Nore1 may serve as a Ras-regulated tumor suppressor in the lung. J Biol Chem. 2003 Jun 13;278(24):21938-43.
11. Sieben NL, Macropoulos P, Roemen GM, Kolkman-Uljee SM, Jan Fleuren G, Houmadi R, et al. In ovarian neoplasms, BRAF, but not KRAS, mutations are restricted to low-grade serous tumours. J Pathol. 2004 Mar;202(3):336-40.
12. Hingorani SR, Jacobetz MA, Robertson GP, Herlyn M, Tuveson DA. Suppression of BRAF(V599E) in human melanoma abrogates transformation. Cancer Res. 2003 Sep 1;63(17):5198-202.
13. Sharma A, Trivedi NR, Zimmerman MA, Tuveson DA, Smith CD, Robertson GP. Mutant V599EB-Raf regulates growth and vascular development of malignant melanoma tumors. Cancer Res. 2005 Mar 15;65(6):2412-21.
14. Hoeflich KP, Gray DC, Eby MT, Tien JY, Wong L, Bower J, et al. Oncogenic BRAF is required for tumor growth and maintenance in melanoma models. Cancer Res. 2006 Jan 15;66(2):999-1006.
15. Sumimoto H, Hirata K, Yamagata S, Miyoshi H, Miyagishi M, Taira K, et al. Effective inhibition of cell growth and invasion of melanoma by combined suppression of BRAF (V599E) and Skp2 with lentiviral RNAi. Int J Cancer. 2006 Jan 15;118(2):472-6.
16. Adams JA. Activation loop phosphorylation and catalysis in protein kinases: is there functional evidence for the autoinhibitor model? Biochemistry. 2003 Jan 28;42(3):601-7.
17. Sutto L, Marsili S, Gervasio FL. New advances in metadynamics. Wiley Interdisciplinary Reviews: Computational Molecular Science. 2012;2(5):771-9.
18. Laio A, Parrinello M. Escaping free-energy minima. Proc Natl Acad Sci U S A. 2002 Oct 1;99(20):12562-6.
19. Berteotti A, Cavalli A, Branduardi D, Gervasio FL, Recanatini M, Parrinello M. Protein conformational transitions: the closure mechanism of a kinase explored by atomistic simulations. J Am Chem Soc. 2009 Jan 14;131(1):244-50.
20. Lovera S, Sutto L, Boubeva R, Scapozza L, Dolker N, Gervasio FL. The different flexibility of c-Src and c-Abl kinases regulates the accessibility of a druggable inactive conformation. J Am Chem Soc. 2012 Feb 8;134(5):2496-9.
21. Sutto L, Gervasio FL. Effects of oncogenic mutations on the conformational free-energy landscape of EGFR kinase. Proc Natl Acad Sci U S A. 2013 Jun 25;110(26):10616-21.
22. Marks JL, Gong Y, Chitale D, Golas B, McLellan MD, Kasai Y, et al. Novel MEK1 mutation identified by mutational analysis of epidermal growth

factor receptor signaling pathway genes in lung adenocarcinoma. Cancer Res. 2008 Jul 15;68(14):5524-8.

23. Arcila ME, Drilon A, Sylvester BE, Lovly CM, Borsu L, Reva B, et al. MAP2K1 (MEK1) Mutations Define a Distinct Subset of Lung Adenocarcinoma Associated with Smoking. Clin Cancer Res. 2015 Apr 15;21(8):1935-43.

24. Bromberg-White JL, Andersen NJ, Duesbery NS. MEK genomics in development and disease. Brief Funct Genomics. 2012 Jul;11(4):300-10.

25. Delaney AM, Printen JA, Chen H, Fauman EB, Dudley DT. Identification of a novel mitogen-activated protein kinase kinase activation domain recognized by the inhibitor PD 184352. Mol Cell Biol. 2002 Nov;22(21):7593-602.

26. Nikolaev SI, Rimoldi D, Iseli C, Valsesia A, Robyr D, Gehrig C, et al. Exome sequencing identifies recurrent somatic MAP2K1 and MAP2K2 mutations in melanoma. Nat Genet. 2012 Feb;44(2):133-9.

27. Emery CM, Vijayendran KG, Zipser MC, Sawyer AM, Niu L, Kim JJ, et al. MEK1 mutations confer resistance to MEK and B-RAF inhibition. Proc Natl Acad Sci U S A. 2009 Dec 1;106(48):20411-6.

28. Rodriguez-Viciana P, Tetsu O, Tidyman WE, Estep AL, Conger BA, Cruz MS, et al. Germline mutations in genes within the MAPK pathway cause cardio-facio-cutaneous syndrome. Science. 2006 Mar 3;311(5765):1287-90.

29. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. Nucleic Acids Res. 2000 Jan 1;28(1):235-42.

30. Fischmann TO, Smith CK, Mayhood TW, Myers JE, Reichert P, Mannarino A, et al. Crystal structures of MEK1 binary and ternary complexes with nucleotides and inhibitors. Biochemistry. 2009 Mar 31;48(12):2661-74.

31. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, et al. UCSF Chimera--a visualization system for exploratory research and analysis. J Comput Chem. 2004 Oct;25(13):1605-12.

32. Ko TP, Jeng WY, Liu CI, Lai MD, Wu CL, Chang WJ, et al. Structures of human MST3 kinase in complex with adenine, ADP and Mn2+. Acta Crystallogr D Biol Crystallogr. 2010 Feb;66(Pt 2):145-54.

33. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990 Oct 5;215(3):403-10.

34. Eswar N, Webb B, Marti-Renom MA, Madhusudhan MS, Eramian D, Shen MY, et al. Comparative protein structure modeling using MODELLER. Curr Protoc Protein Sci. 2007 Nov;Chapter 2:Unit 2 9.

35. Hess B, Kutzner C, van der Spoel D, Lindahl E. GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. Journal of Chemical Theory and Computation. [doi: 10.1021/ct700301q]. 2008 2008/03/01;4(3):435-47.

36. Tribello GA, Bonomi M, Branduardi D, Camilloni C, Bussi G. PLUMED 2: New feathers for an old bird. Comput Phys Commun 2014;185(2):604-13.

37. Piana S, Lindorff-Larsen K, Shaw DE. How robust are protein folding simulations with respect to force field parameterization? Biophys J. 2011 May 4;100(9):L47-9.

38. Petersen HG. Accuracy and efficiency of the particle mesh Ewald method. The Journal of Chemical Physics. 1995;103(9):3668-79.

39. Bussi G, Donadio D, Parrinello M. Canonical sampling through velocity rescaling. J Chem Phys. 2007 Jan 7;126(1):014101.
40. Deighan M, Bonomi M, Pfaendtner J. Efficient Simulation of Explicitly Solvated Proteins in the Well-Tempered Ensemble. Journal of Chemical Theory and Computation. [doi: 10.1021/ct300297t]. 2012 2012/07/10;8(7):2189-92.
41. Tiwary P, Parrinello M. A time-independent free energy estimator for metadynamics. J Phys Chem B. 2015 Jan 22;119(3):736-42.
42. Gopalbhai K, Jansen G, Beauregard G, Whiteway M, Dumas F, Wu C, et al. Negative regulation of MAPKK by phosphorylation of a conserved serine residue equivalent to Ser212 of MEK1. J Biol Chem. 2003 Mar 7;278(10):8118-25.
43. Zheng CF, Guan KL. Activation of MEK family kinases requires phosphorylation of two conserved Ser/Thr residues. EMBO J. 1994 Mar 1;13(5):1123-31.
44. Yan M, Templeton DJ. Identification of 2 serine residues of MEK-1 that are differentially phosphorylated during activation by raf and MEK kinase. J Biol Chem. 1994 Jul 22;269(29):19067-73.
45. Kannan N, Neuwald AF. Did protein kinase regulatory mechanisms evolve through elaboration of a simple structural component? J Mol Biol. 2005 Sep 2;351(5):956-72.
46. Lew J, Taylor SS, Adams JA. Identification of a partially rate-determining step in the catalytic mechanism of cAMP-dependent protein kinase: a transient kinetic study using stopped-flow fluorescence spectroscopy. Biochemistry. 1997 Jun 3;36(22):6717-24.
47. Shaffer J, Sun G, Adams JA. Nucleotide release and associated conformational changes regulate function in the COOH-terminal Src kinase, Csk. Biochemistry. 2001 Sep 18;40(37):11149-55.

Supporting Informations

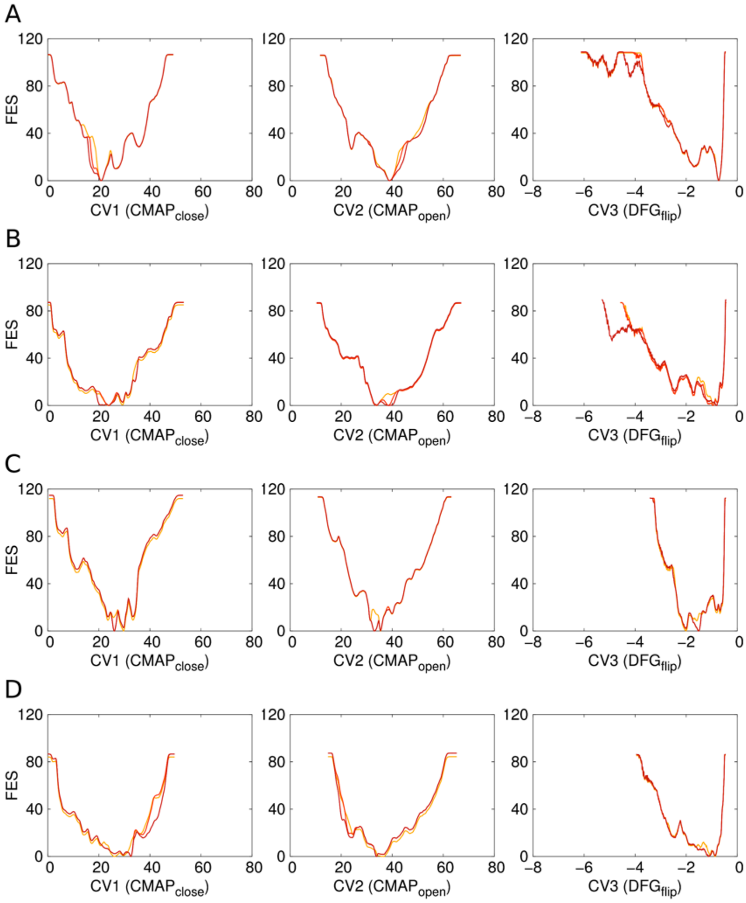


Fig S1. Free-energy mono-dimensional projection as a function of CV1, CV2 (the distance in contact map space to A-loop closed and open state, respectively) and CV3 (DFG-motif orientation): (A) wild-type MEK1 protein kinase and the three mutants: (B) Y130C, (C) Q56P and (D) E203K**.** In each plot three curves are shown, separated by 50 ns of simulation time and corresponding to the last 150 ns of each simulation. The final converged FES is shown in red.

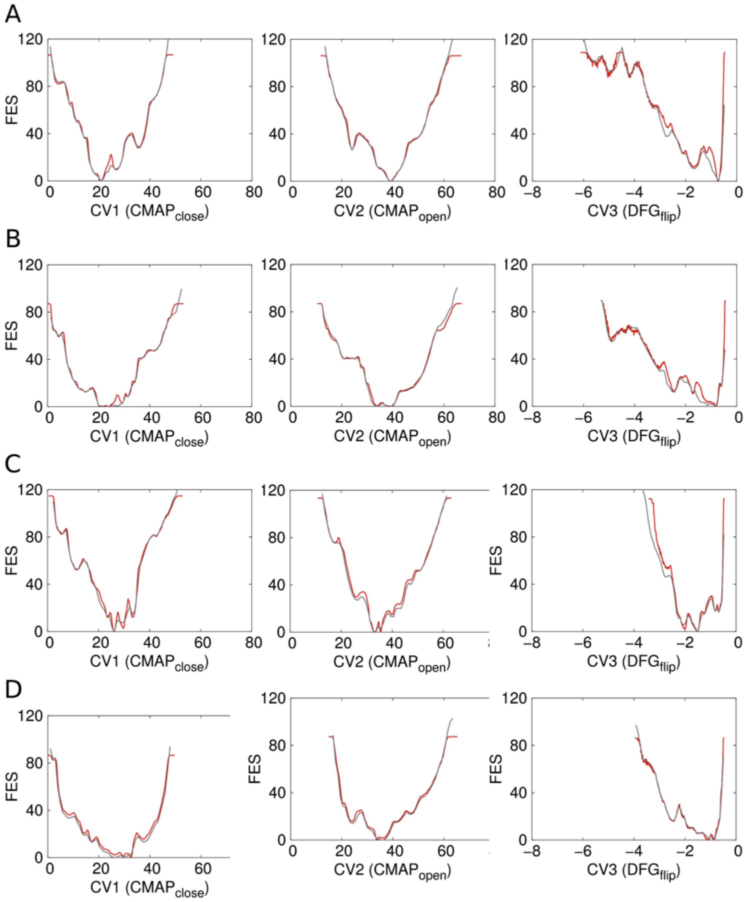


Fig S2. Comparison between the free-energy profile in the mono-dimensional projection obtained by PLUMED module sumhills (in red) and the ones obtained by the time-independent re-weighting technique of Tiwary and Parrinello (in gray): CV1, CV2 (the distance in contact map space to the A-loop closed and open state, respectively) and CV3 (DFG-motif orientation) for (A) wild-type MEK1 protein kinase and the three mutants: (B) Y130C, (C) Q56P and (D) E203K.

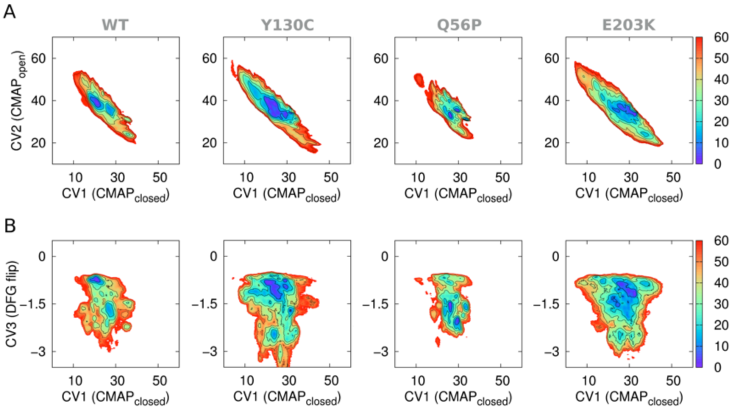


Fig S3. Comparison between the free-energy profile in the bi-dimensional projection: (A) CV1, CV2 and (B) CV1, CV3 for wild-type MEK1 protein kinase and the three mutants. The contour lines are drawn every 10kJ/mol.

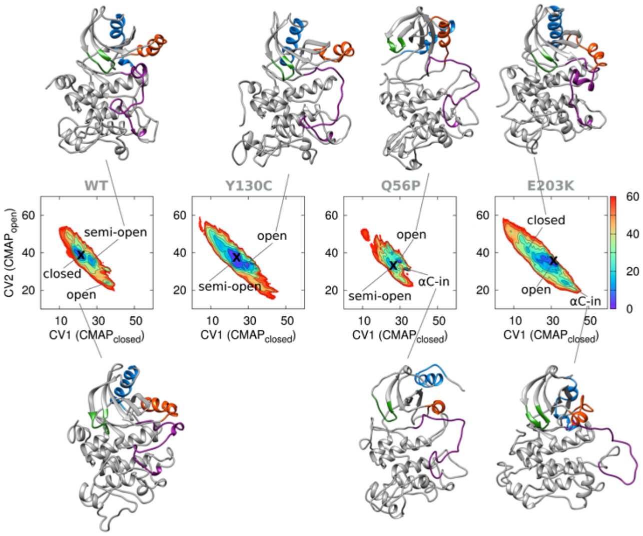


Fig S4. Free-energy surface of wild-type MEK1 and the three mutants as a function of CV1 (x-axis) and CV2 (y-axis). A cross indicates the global free energy minimum. The contour lines are drawn every 10kJ/mol. The representative structures of the secondary free energy minima are shown with the αA-helix in light blue, the P-loop in green, the αC-helix in orange, the A-loop in magenta.

4. Results summary

“if I don't know something,
I look into it.”

Luis Pasteur

4.1. Advances and new challenges in modeling of protein interactions

Given the growing interest in protein-protein interactions and the technical advances in computational field, an increasing number of *in silico* tools have been developed with the aim of (i) identifying residues that significantly contribute to binding, and (ii) modeling protein complexes starting from the isolated component structures (*docking problem*). Testing and comparing these computational methodologies is fundamental in order to assess their performance, identify their limitations, and finally guide new developments in the field. In this context, CAPRI experiment provides a common ground for testing the predictive capability of currently available docking methods.

Firstly, this section will be focused on the analysis of several existing computational protocols for the characterization of protein-protein interfaces. Secondly, the performance of our pyDock protocol (Cheng et al., 2007) on the last CAPRI round (Lensink and Wodak, 2013) will be evaluated and discussed.

4.1.1. Energetic characterization of protein-protein interfaces

The performance of four different computational methods for the characterization of protein-protein interfaces was assessed using the recently solved complex between MEK1 and BRAF protein kinases (PDB 4MNE) (Haling et al., 2014) as a test case. The analysis was focused on (i) ConSurf (Glaser et al., 2003), based on the degree of conservation of each amino acid site among their close sequence homologues (see **section 1.2.6** and **3.1.1**); (ii) pyDock module pyDockNIP (Grosdidier and Fernandez-Recio, 2008), which computes the tendency of a given residue to be located at the interface based on rigid-body docking poses (see section 1.2.6); (iii) the residue contribution to complex binding energy computed with pyDock (upcoming publication); and finally, (iv) *in silico* Alanine (Ala) - scanning based on free-energy calculations involving Molecular Mechanics/Generalized Born Surface Area method (MM-GBSA) using MMPBSA.py script (Miller et al., 2012) from AMBER14 tools (Salomon-Ferrer et al., 2013).

Indeed, rather similar predictions were obtained using the different methods (**Fig 1** from **section 3.1.1**) although some peculiarities should be highlighted. Firstly, MM-GBSA Ala-scanning resulted much more computationally expensive and time consuming than the other methods. In addition, NIP and ConSurf have the considerable advantage that *a priori* knowledge of the complex is not required. On the other side, pyDock binding energy residue contribution and MM-GBSA Ala-scanning have the convenience of providing a quantitative analysis that can be easily compared with the experimental data. Finally, it is noteworthy that pyDock scheme revealed a rather similar performance to MM-GBSA Ala-scanning in the prediction of binding energy residue contribution but at much lower computational cost.

4.1.2. New challenges in protein docking: improving energetics and description of flexibility

pyDock docking program (Cheng et al., 2007) is regularly tested by participating in the CAPRI experiment (http://www.ebi.ac.uk/msd-srv/capri/), a community-wide blind prediction of macromolecular interactions based on the three-dimensional structures of the interacting proteins (see **section 1.2.6**).

The fifth CAPRI edition (2010-2012) (Lensink and Wodak, 2013) consisted of a total of 15 targets that, in addition to the standard docking predictions, included binding affinity predictions and free energy changes upon mutation (Moretti et al., 2013), as well as prediction of sugar binding and interface water molecules (Lensink et al., 2014). We have participated in all the proposed targets with high success, since our models were globally placed among the top 5 ranked groups out of more than 60 participants (Lensink and Wodak, 2013) (**Fig 1** from **section 3.1.2** and **Table S11A** from Lensink and Wodak, 2013).

Major determinants for the success were the generation of docking poses with FTDock (Gabb et al., 1997) at a grid resolution of 0.7 Å (instead of 1.2 Å as in the past), as well as with SwarmDock program (Moal and Bates, 2010) for part of the targets. In selected targets (T47, T48 and T58), distance restraints were used, although this hardly made a difference. In target T58, SAXS data were used for complementary scoring with pyDockSAXS (Pons et al., 2010), which slightly improved the predictive performance. Overall, pyDock submitted consistently good models for all non-difficult cases, although they were far from being trivial, since their subunits were in the unbound conformation or needed to be modeled based on homology templates. In all cases but one, pyDock successful models were ranked within the first five submitted solutions, being ranked as first in two out of six successful cases.

As mentioned above, T47, T55, T56 and T57 were non-standard targets for which new *ad hoc* protocols had to be developed. Target T47 was a rather trivial modeling of a protein-protein complex structure but the real challenge consisted in the prediction of the location of water molecules within the complex interface. Our approach was based on DOWSER *ab initio* optimization procedure (Zhang and Hermans, 1996). Indeed, this choice was reasonably successful, although *a posteriori* analysis showed that the most successful approaches were based on deriving initial water positions from interfaces of related complexes followed by energy minimization (as in Nakamura's group protocol). More details on the protocols used by all the participants and the results obtained are reported in Lensink et al., 2014. T55 and T56 targets aimed to predict the binding affinity changes upon mutations on two designed influenza hemagglutinin protein binders. By applying a machine learning protocol based on 85 energy descriptors, our predictions were placed within the top 3 out of 22 groups. Overall, the CAPRI community faced a big challenge with these targets (Moretti et al., 2013). Finally, for T57, which involved the prediction of a protein-sugar interaction, we developed a new protocol based on a combination of different scoring functions, such as PScore and AMBER (Case et al., 2005). Although no correct models were submitted, at least one almost acceptable model was found (11.2 and 4.3 Å ligand- and interface- RMSD, respectively).

In the on-going (not yet evaluated) CAPRI edition (2013-to date), one of the new challenges consisted in the prediction of eight protein/peptide complexes (T60-67), in which pyDock obtained highly successful results as predictor and rather satisfying as server with 7 and 5 successfully predicted targets, respectively. However a much more demanding test was recently proposed for CAPRI Round 30, the first joint CASP-CAPRI experiment, consisting in 25 targets of homo-oligomers from the CASP11 2014 Round. pyDock submitted at least one acceptable model in 11 out of 12 easy homodimer targets, either as predictor or scorers, while, as scorers, it successfully predicted two out of six difficult homodimer targets and one out of two hetero-complex targets. On the contrary, pyDock failed in the prediction or evaluation of any tetramer target, where the inaccuracy of the homology-built subunit models and the smaller pair-wise interfaces severely limited the ability to derive the correct assembly mode. Globally, pyDock predictions were placed among the top 10 ranked groups out of about 25 predictors, and among the top 5 ranked groups out of about 12 scorers participating in Round 30 (**Table 4** from Lensink M et al., submitted).

Overall, the recent results of the CAPRI experiment demonstrated a general robustness of current docking algorithms, since their performance remained relatively consistent when facing targets of different difficulty. Two of the most challenging targets were T46, where subunits needed to be modelled based on distantly related templates, and T51, where the isolated structures of the docking partners were largely different from those of the native complex. Regarding the estimation of binding affinity changes upon mutation, although this is still far from being solved (as shown in T55 and T56), it is interesting to note that the incorporation of backbone flexibility and more extensive sampling of side-chain conformations were in general positive for the predictions (Lensink and Wodak, 2013).

All these observations show that the intrinsic molecular flexibility plays a crucial role in the protein-protein association mechanism and thus should deserve special consideration for the structural and energetic modeling of protein-protein complexes.

4.2. Protein plasticity improves protein-protein docking

Given the key role of conformational flexibility in protein-protein association, accurate description of protein plasticity in docking protocols appears crucial for the successful prediction of complex structure (see **sections 3.1.2** and **4.1.2**). However, in addition to the higher computational cost, the development of efficient flexible docking algorithms is hampered by the limited theoretical knowledge about the protein-protein association mechanism.

Thus, an important aim of this PhD thesis has been to study the role of protein conformational heterogeneity in protein-protein recognition and the exploration of ways to integrate unbound conformational ensembles within protein-protein docking.

4.2.1. Protein plasticity enhances protein-protein recognition

This study has explored whether a minimal description of the conformational heterogeneity of the interacting proteins could significantly improve their binding capabilities.

To this aim, first of all, small conformational ensembles for all interacting proteins in the protein-protein docking benchmark 3.0 (Hwang et al., 2008) were automatically generated, starting from the unbound docking partners, by using two distinct conformational sampling procedures: (i) a fast energy optimization as implemented in MODELLER (Eswar et al., 2007), and (ii) Molecular Dynamics simulation, as implemented in AMBER package (Salomon-Ferrer et al., 2013). As shown in **Fig 1** and **S2** from **section 3.2.1**, the deviation of the interface atoms from the initial unbound structure were rather modest in both approaches, reaching a maximum of about 1.2 and 2.7 Å RMSD values for MODELLER and MD ensembles, respectively. Moreover, only in about 20% of the benchmark cases, the conformational ensembles (either from MODELLER or MD simulation) contained conformers that were significantly more similar to the native conformation than the unbound X-ray structure. Interestingly, in the majority of cases the conformational ensembles contained conformers that showed better binding energy capabilities than the unbound X-ray structure.

The following step was to evaluate whether the conformational ensembles contained conformers that could be beneficial for docking. To this aim, we selected these conformers that seemed more promising for binding, either because of higher similarity to the bound state or higher complementarity to the docking partner (in terms of binding energy or number of clashes). Finally, a total of five conformers were collected for each benchmark case and consequently used as input for rigid-body docking simulations, as implemented in pyDock (Cheng et al., 2007). Surprisingly, the conformers that were structurally more similar to the reference did not yield better docking results than the unbound structures. On the contrary, when using the conformers selected according to the binding energy or the smallest number of clashes when in the native orientation, the docking performance significantly improved (**Fig 4A** from **section 3.2.1**). Moreover, a clear dependence of the docking improvement on the conformational rearrangement of the interacting proteins upon binding was observed: the ensemble success rates were particularly good in the low-flexible cases, for which they reached values similar to those when using the bound X-ray structures; on the contrary, in rigid and highly-flexible cases the selected conformers yielded similar results to the unbound structures (**Fig 4B** and **4C** from **section 3.2.1**).

Globally, the ensemble size and method of sampling did not have a large effect on the docking performance. The selected conformers from the 1000-member ensembles generated by MODELLER or MD yielded similar results to those selected from the 100-member ensembles (**S4 Fig** from **section 3.2.1**), although a small improvement was observed in the flexible cases when using the larger ensembles (**S5 Fig** from **section 3.2.1**). Moreover, no better docking performance was observed by extending the MD simulation time to 100 ns (**S1 Table** from **section 3.2.1**). However, new sampling approaches based on NMA did appear useful to produce a significant improvement of the success rate in the most flexible cases (**Table 2** from **section 3.2.1**).

In summary, this study showed that considering conformational heterogeneity in the unbound state of the interacting proteins can improve their binding capabilities in cases of moderate unbound-to-bound mobility. In these cases, the existence of conformers with better binding energy (not necessarily related to bound geometries) in the native orientation is associated to a significantly improvement in the docking predictions. These findings set useful guidelines for the development of novel protocols in practical docking prediction.

4.2.2. Development of a new protocol for ensemble docking

The previous study showed that a simple molecular mechanics approach, as the one implemented in MODELLER program (Eswar et al., 2007), can generate conformers with better binding properties and thus improve docking predictions (see **section 3.2.1**). Based on this study, we have devised a novel protocol to integrate unbound conformational ensemble within protein-protein docking (using pyDock) and tested it on a data set of 124 protein-protein docking cases (Chen et al., 2003). In order to reduce the computational costs associated with a cross-docking approach, a total of only 100 docking runs were preformed randomly pairing every conformer from the receptor with another one from the ligand. All the docking poses generated were thus merged together and finally ranked according to pyDock scoring function (filtering out similar conformations with ligand RMSD cutoff 10 Å).

As shown in **Fig 1A** and **1B** from **section 3.2.2**, ensemble docking clearly showed better success rates (32.3%) than the unbound subunits (19.4%). Moreover, in order to provide a statistical significance for these results, 100 different docking runs were performed using random initial rotations of the unbound receptor and ligand molecule. The resulting docking performance obtained (26.6% success rate) stood halfway between the ensemble docking and the unbound results. These results suggest that conformational heterogeneity in the interacting subunits improves the binding capabilities of the unbound X-ray structures. Interestingly, lowering the number of random initial rotations (i.e., five rotations) did not substantially change the docking performance (27.4%), showing the importance of performing a small sampling of the position of the starting molecules instead of using only the unbound X-ray structures. Moreover, it was observed that the docking improvement when using conformational ensembles depended on the flexibility of the interacting proteins upon binding. Indeed, as shown in **Fig 1C** and **1D** from **section 3.2.2**, the largest improvement occurs in medium-flexible cases, those with I-RMSD$_{C\alpha}$ between 1.0 and 2.0 Å.

Finally, we also explored the question of which is the minimal number of conformers that would be needed in order to observe a significant improvement in the docking results. As shown in **Fig. 2** from **section 3.2.2**, it was found that the docking performance increases linearly with the size of the ensemble. For instance, when only five conformers were used for the ensemble docking protocol, the prediction success rates (24.2%) were similar to random. Interestingly, when only the high-affinity, low-flexible cases are considered, the docking success improves dramatically with just a few conformer pairs, so that 30 conformers provide similar success rates as 100 conformers. All these data suggest that the results might be further improved by using a higher number of conformer pairs, increasing the number of receptor-ligand conformer combinations, although for this, new algorithms would need to be developed and optimized for high-performing computing in order to deal with the humongous computational cost.

In conclusion, this study shows that a minimal conformational heterogeneity can be used in a practical docking protocol to improve the results of unbound docking. Moreover, this proved to have the potential of further improving the predictive results by extending conformational sampling and/or considering larger ensembles, although this would involve an enormous computational cost.

4.3. Modeling protein interactions: application to cases of interest

The expertise acquired during the first part of this PhD thesis on the theoretical basis of protein interactions has favored my participation in several multidisciplinary projects, basically through the application of computational methods to the modeling of protein interactions of biological interest. Two of the most interesting collaborative projects will be shown here in detail.

The first study consisted on the energetic characterization of host-pathogen protein interactions, which are key steps of virtually every infection process (i.e, pathogen replication and survival within the host system). Such processes are typically achieved by precise mechanisms developed by the pathogen to subvert and exploit normal host cell processes, very often by mimicking specific host cell interactions. The second project included the *ab initio* modeling of redox complexes, which usually involve transient interactions characterized by the existence of *encounter complexes* formed by “microcollisions” that properly align the reactive groups.

4.3.1. Energetics of host-pathogen protein interactions

L*egionella pneumophila* effector Vacuolar Inhibitor protein D (VipD) localizes to the host early endosomal membrane where succeeds in binding GTPase Rab5 host protein, competing with its endogenous ligands. This key interaction triggers the activation of the phospholipase A1 (PLA1) activity of VipD, which in turn leads to the alteration of the endosomial membranes composition and thereby protect the pathogen from endosomial partitioning (Gaspar and Machner, 2014). Thus, the energetic description of such interaction as well as the disclosure of the molecular basis of this host-pathogen competing process is of fundamental importance to open the path for the development of novel antimicrobial therapeutics.

For this purpose, systematic *in silico* Alanine scanning calculations were performed on all the interface residues of GTPase Rab5 in complex with L*egionella pneumophila* VipD as well as in complex with the endogenous human ligand proteins EEA1, RAbaptin-5 and Rabenosyn-5. Indeed, the analysis of the VipD-Rab5 interface pinpointed a core of hydrophobic contacts complemented by several polar interactions in the surrounding rim area (**Fig 3B** from **section 3.3.1***).* In particular, the VipD binding epitope in Rab5 included non-polar residues in the switch I (I54, G55, A56 and F58), the interswitch (W75) and the switch II element (Y83, L86, M89 and

Y90) as well as polar/charged residues in the interswitch (T60, K71 and E73) and switch II element (R82 and R92). On the other hand, the corresponding epitope in VipD included several hydrophobic residues in helixα17 (F442, A446, A450 and L454) and in helixα18 (Y473, I480, V483) that wrapped around an elongated hydrophobic path in Rab5 formed by the conserved triad (F58, W75 and Y90) and L86.

Moreover, a much more interesting observation arose from the comparison with the endogenous Rab5 complexes, which revealed a shared hydrophobic core (F58, W75 and Y90) complemented by a unique polar interaction in VipD-Rab5 interface surrounding this hydrophobic triad (R92 from Rab5 with D484 and D479 from VipD) and which provide a rather stronger energetic contribution to the interaction in comparison with the other cellular ligands (**Fig 7A** from **section 3.3.1**). The role of R92 in VipD-Rab5 complex (not existing in those between Rab5 and the endogenous ligands) was confirmed by experimental mutagenesis. These findings offer a useful help for the development of novel treatments aimed at selectively blocking the VipD activation process rather than the enzyme's active site.

4.3.2. Modeling redox complexes by docking

The interaction of photosystem I (PSI) with electron transfer proteins (such as plastocyanin or cytochrome) catalyzes the first step of the photosynthesis process. Unlike most cyanobacteria and unicellular green algae, *Phaeodactylum tricornutum* alga, as most diatoms, lacks plastocyanin and thus leaves cytochrome c_6 (Cyt) as the only electron donor to photosystem I (Akazaki et al., 2009). Nevertheless, diatom PSI is still able to recognize and functionally interact with different eukaryotic acidic plastocyanins, although with lower efficiency than the native electron carrier (Bernal-Bayard et al., 2013). Thus, the principal goal of this study was to understand the structural and energetic determinants of the differences in the efficiency observed in diatoms PSI reduction with respect to the green systems.

To this aim, docking simulations were initially performed between the homology-built PSI model from *Phaeodactylum* and both the native and the green alga *Monoraphidium braunii* Cyt structures. Additional docking analyses were performed to investigate the interaction of PSI from *Phaeodactylum* with the Pc of green alga *Chlamydomonas reinhardtii* and particular efforts were addressed to identify energetic differences between the wild type and some mutants designed to mimic *Phaeodactylum* Cyt electrostatic properties (i.e., E85K, Q88R, E85K/Q88R, E85V and V93K).

The docking simulations between PSI and Cyt from *Monoraphidium* yielded a much larger population of low-energy productive orientations and showed better binding energy values as compared with the native *Phaeodactylum* (**Table 2** and **Fig S3** from **section 3.3.2**). Further analyses revealed that the *Monoraphidium Cyt* best-energy productive docking model was stabilized by electrostatic interactions between PsaA residues R747 and R648 and Cyt residues D42 and H30, which corresponded to an alanine and a lysine (i.e., A86 and K74) in *Phaeodactylum* Cyt. Indeed, these differences can explain why an equivalent docking orientation would be energetically penalized in the interaction with *Phaeodactylum* PSI (**Fig 5** from **section 3.3.2**).

On the other hand, in agreement with its lower similarity to Cyt electrostatic patches, the wild type Pc yielded a much larger population of low-energy less productive docking orientations in comparison with some of the mutants, such as E85K and Q88R. Indeed, these orientations revealed to be stabilized by electrostatic interactions between PSI positively charged residues (namely R747, R648 and K638) with *Chlamydomonas* Pc E85 and Q88 residues (**Fig 6** from **section 3.3.2**). On the contrary, the loss of these contacts by the substitution of E85 and Q88 with positively charged residues facilitated a higher population of the more productive binding modes, as observed in the docking landscapes of the E85K and Q88R Pc mutants.

All these results not only improve the understanding of the mechanism and energetics of PSI reduction but also shed new light on the evolution of the electron transfer mechanism to PSI in the different branches of the evolutionary tree of photosynthetic oxygenic organisms.

4.4. Description of protein plasticity: an example of biomedical interest

In order to understand the mechanistic aspects of several biological processes, such as the functional effects of a single mutation, the consideration of protein plasticity is of paramount importance. Protein kinases represent a paradigmatic example of the importance of a close link between dynamics and function. These enzymes regularly switch between characteristic inactive and active states undergoing large conformational changes, whose complete computational description is still challenging. Standard all-atom Molecular Dynamics (MD) simulations are useful computational tools to explore small-to-medium conformational rearrangement, but large conformational transitions in proteins can only be described by enhanced sampling methods.

Thus, this PhD thesis concluded with the application of conventional Molecular Dynamics (MD) and state-of-the-art enhanced sampling using metadynamics to elucidate the effects at molecular level of pathological mutations in MEK1 protein kinase (Y130C, exclusively associated with Cardiofaciocutaneous- (CFC) syndrome; E203K, exclusively related to cancer; and Q56P, observed in both diseases) (see **sections 3.4.1** and **3.4.2**).

Structural and dynamic effects of pathological mutations by extended Molecular Dynamics

In order to understand the effect of the above mentioned pathological mutations at molecular level as well as their impact on the intrinsic propensity for MEK1 inactive-to-active transition, 1-μs-long MD simulations were initially performed on the MEK1 unphosphorylated *apo* as well as on the phosphorylated ATP-bound state. Then, four additional 1-μs-long simulations were run using PTmetaD-WTE protocol (Deighan et al., 2012), an enhanced sampling approach which combines parallel tempering with well-tempered metadynamics.

According to the MD simulations, all the mutations described herein resulted in easing the *inactive-to-active* transition in both the unphosphorylated *apo* and the phosphorylated ATP-bound MEK1 state. More in detail, in the basal state they increased αA-helix structural flexibility, promoting both its partial unfolding and the loose of αA-helix/core native contacts (e.g., Y130/Q56, R49/E203) (**Fig 2, S1** and **Table 1** from **section 3.4.1**). However, a distinctive effect was found in Q56P with respect to Y103C and E203K mutants, involving the increase of P-loop flexibility, which could be causing a higher cofactor turnover (**Fig 3** and **Table 1** from **section 3.4.1**). Over-activating effects were also found in the phosphorylated A-loop, where all of the mutants studied herein induced a significant increase of the A-loop flexibility as compared with that in the WT (**Fig 5** and **Table 1** from **section 3.4.1**) and favoured a dramatic destabilization of the hydrogen bond between αC-helix E114 and A-loop Q214 residues, whose interacting frequency dropped from 85% in the WT to a range between 29 and 54% for the mutants (**Fig S2** and **Table 1** from **section 3.4.1**). All these data suggest that the mutants could result in the acceleration of active-to-inactive process with significant impact on both the biological state analyzed. Nevertheless, the effects produced by E203K resulted more dramatic if compared with Q56P and even more with respect to Y130C.

In agreement with MD, the free-energy landscapes obtained using PTMetaD-WTE confirmed that all the mutations seemed to favour the inactive to active state transition, although to a different extent. In addition, new hints about specific molecular mechanisms accounting for oncogenic and non-oncogenic mutations were discovered. First of all, E203K and Q56P showed significant effects on the *closed-to-open* transition of the A-loop, stabilizing an intermediate state in which the A-loop resulted completely unfolded. On the other hand, Y130C mutation showed milder effects (although clearly appreciable) on the stabilization of an intermediate state in which the A-loop is still partially folded (**Fig 2** from **section 3.4.2**). Regarding αC-helix flexibility, a higher propensity for out-to-in transition was observed only in oncogenic mutations (Q56P and E203K) (**Fig 3** from **section 3.4.2**). Finally, it is worth noting the crucial effects of the mutations on the DFG-motif plasticity. Indeed, all the mutants, although through different mechanisms, significantly flattened the energy barrier for the DFG in-to-out transition, which could promote the ADP release and increase the ATP turnover rate.

In conclusion, the combination of MD and PTMetaD-WTE simulations could help to rationalize and quantify the activating effects induced by Y130C, Q56P and E203K pathological mutations in MEK1, as well as to propose a mechanistic explanation to the different extent of MEK1 over-activation observed for oncogenic or CFC-related mutations and eventually open the path for the development of disease specific therapeutic approaches.

5. Discussion

“The important thing
is not to stop questioning.”

Albert Einstein

After sequencing the complete genomes of several organisms we are starting to unravel their intricate protein-protein interaction networks, which is essential to understand biological processes at molecular level, with the ultimate goal of contributing to improve therapeutic intervention. In this context, one of the current biological challenges is to provide structural details at atomic level for such interactomes. However, given the intrinsic limitations of available experimental methods to determine 3D structures (e.g., X-ray crystallography or NMR), large-scale structural determination of complete interactomes seems beyond current capabilities.

Fortunately, computational methods can complement experimental efforts by providing structural and energetic large-scale modeling of protein interactions. In spite of the most recent advances in computational modeling, many important challenges remain. Among them, efficient consideration of conformational flexibility is arguably the most important problem to solve in order to improve the structural and energetics modeling of proteins and their interactions. The consideration of protein dynamics appears essential in several *in silico* studies, from the prediction of their interactions to the study of the mechanistic aspects of their function. Indeed, this issue is not well addressed in most of the currently available docking programs, which this is actually one of the main limitations in their predictive performance.

The work of this thesis has focused on analyzing the importance of structural, energetics and dynamic aspects of protein interactions, the development of new computational tools to help solving current problems, and the application of computational modeling to cases of biological and biomedical interest.

5.1. Energetic characterization of protein-protein interfaces: advances and new challenges

Residues in protein-protein interfaces do not equally contribute to the overall binding energy; in fact a few *hot-spot* residues typically contribute the most to the binding energy (Clackson and Wells, 1995; Bogan and Thorn, 1998). The identification of such residues and thus the functional characterization of protein interactions at molecular level is mandatory, not only to understand biological and pathological phenomena, but also to design improved, or even new interfaces, or to develop new therapeutic approaches (Wells and McClendon, 2007; Kar et al., 2012; Thangudu et al., 2012).

Considering that the experimental characterization of protein interfaces remains expensive, time-consuming and labor-intensive, computational approaches represent a significant breakthrough in proteomics, assisting or even replacing experimental efforts. Thanks to the technological advances in computing and data processing, these techniques now cover a vast range of protocols, from the estimation of the evolutionary conservation of amino acid positions in a protein, to the energetic contribution of each residue to the binding affinity. In this thesis, several existing computational protocols to model the phylogenetic, structural and energetic properties of residues within protein-protein interfaces have been reviewed and their performance compared, using MEK1-BRAF complex (PDB ID 4MNE) as example (see **section 3.1.1**). Although the different protocols differed in time and computational costs, type of the analysis provided (e.g., qualitative or quantitative) and the degree of structural information required, the predictions obtained were quite consistent between themselves (**Fig 1** from **section 3.1.1**).

Interestingly, a new protocol for pyDock per-residue free energy decomposition had a similar performance to the widely used MM-GBSA Ala-scanning although at much lower computational cost. Based on this observation, future systematic comparisons between these two methods would be interesting in order to produce new developments in the field.

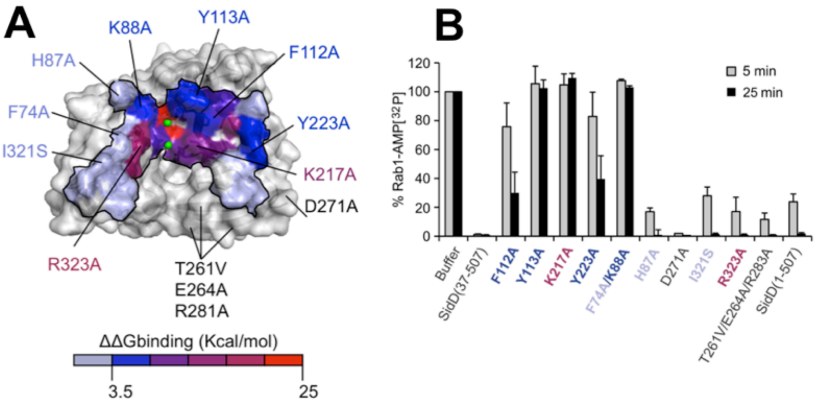


Figure 11. (A) SidD interface surface contacting AMPylated Rab1, colored according to the predicted ΔΔG; (B) *In vitro* de-AMPylation assay on selected mutants located within and outside the interface. From Chen et al., 2013.

Apart from the above cited study, MM-GBSA Ala-scanning protocol has been successfully applied in an additional multidisciplinary project during this thesis for the characterization of Rab1/SidD complex in collaboration with Aitor Hierro research group (CIC bioGUNE) (Chen et al., 2013). As shown in **Figure 11** *in silico* predicted *hot-spots* are in substantial agreement with the experimental data on *in vitro* SidD-mediated de-AMPylation of Rab1.

Overall, all these observations confirm that *in silico* protocols provide a reliable aid to the energetic characterization of protein-

protein interfaces.

5.2. Structural prediction of protein interactions: from energetics to molecular flexibility and other challenges

The last CAPRI edition, in which I have actively participated, sparked a large number of ideas and new strategies for structural modeling of protein interactions. In addition, our overall experience has been highly rewarding, since pyDock docking scheme confirmed its high performance in protein complexes prediction being placed among the top5 ranked groups out of more than 60 participants (see **section 3.1.2** and **Table 3**).

Table 3. Overall pyDock performance among the top 10 ranked groups.

Rank	Group	#Targets / *** + ** + *
1	Bonvin	9 / 1 *** + 3 ** + 5 *
2	Bates	8 / 2 ** + 6 *
3	Vakser	7 / 1 *** + 6 *
4	Vajda	6 / 2 *** + 3 ** + 1 *
5	**Fernandez-Recio**	**6 / 1 *** + 3 ** + 2 ***
5	Shen	6 / 1 *** + 3 ** + 2 *
7	Zou	6 / 1 *** + 2 ** + 3 *
8	Zacharias	6 / 1 *** + 5 *
9	ClusPro	6 / 4 ** + 2 *
10	Eisenstein	5 / 1 *** + 2 ** + 2 *

Predictions are classified as acceptable (*), medium (**), and high (***).

Besides this, many targets were proposed that go beyond the traditional protein-protein complex structure prediction, but that constitute important current challenges in modeling protein interactions. For instance, regarding the prediction of binding energy changes upon mutation, the application of multiparametric regression models, as implemented in our protocol, showed excellent results.

The overall evaluation showed that current protocols are more successful in predicting deleterious mutations than beneficial ones. Furthermore, the best procedures were those that took into account the effect of the mutation on the stability of the monomer, and that considered both sterical and energetic parameters (i.e, packing metrics, Lennard–Jones type potentials, electrostatic and solvation terms). Finally, an additional advantage was conferred by extensive sampling of side chain conformations and the incorporation of some backbone flexibility (Moretti et al., 2013).

The general analysis of the protocols used by the CAPRI community for predicting the location of water molecules within the complex interface showed interesting results. First of all, it seems that high- to medium-quality models for the protein complex are required for successful interface water predictions. Moreover, the combination of established molecular mechanics force fields with some conformational sampling step and a final energy minimization, as well as the consideration of initial water positions derived from interfaces of related complexes, proved to be more successful than much simpler water placement methods (Lensink et al., 2014).

Regarding last CAPRI round in combination with CASP (i.e., Round 30) for the modeling of homo-oligomeric proteins, the majority of participants showed a general poor success in docking of homology-built subunits (especially if modeled on distantly related templates) (Lensink et al, submitted), thus confirming what was indeed previously reported (Lensink and Wodak, 2010) and showing again that more work is needed in this direction.

As for the most traditional targets involving the prediction of a protein-protein complex structure, which is still the main focus of

CAPRI, the analysis of the results showed that dealing with protein flexibility in docking simulations still remains a major challenge. Indeed, docking performance dramatically drops in cases that undergo medium-to-large conformational changes upon binding or in which the interacting subunits need to be modeled based on low-homology templates (Lensink and Wodak, 2013; Pallara et al., 2013). This, in principle, could be explained by poor geometrical (and thus, energetic) complementarity of the docking partners in the unbound form.

Therefore, in order to develop new strategies to include flexibility in docking, a much better understanding of molecular recognition process at structural and energetic level is required, which has been a major motivation in this PhD thesis.

5.3. The role of conformational heterogeneity in protein-protein association process

Conformers providing better binding energy in the native orientation are more likely to produce more effective docking encounters.

The systematic study on the role of conformational heterogeneity in protein-protein association process, as described in **section 3.2.1** and summarized in **Figure 12A,** showed that sets of discrete conformers representing the conformational heterogeneity of the unbound structure yielded better docking results than the unbound structures themselves.

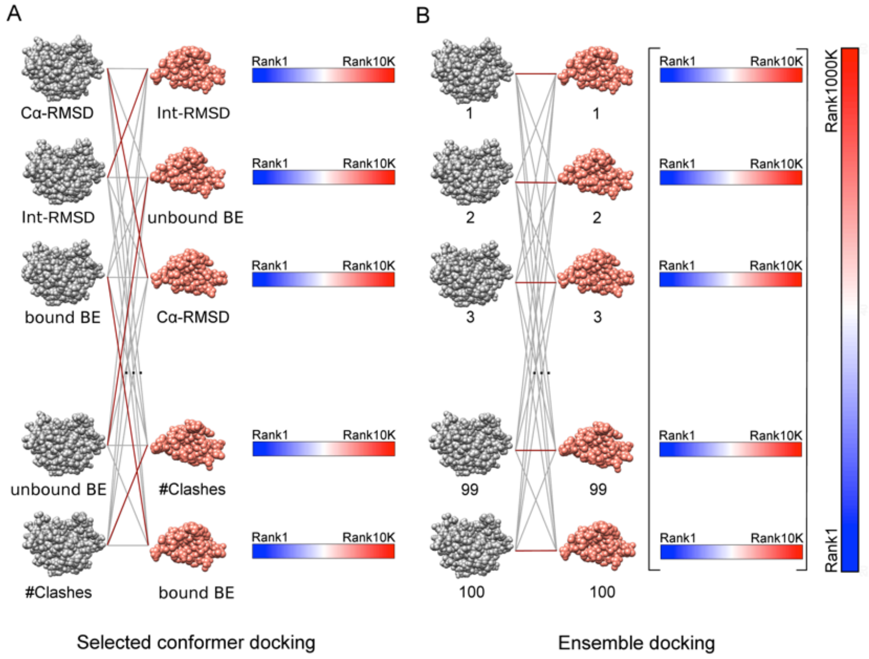


Figure 12. Schematic representation of (A) selected conformer docking (selection of the best ensemble conformers according to quality criteria based on the complex native orientation: Cα-RMSD, Int-RMSD, binding energy with the bound partner, binding energy with the unbound partner, number of clashes with respect to the bound partner); (B) ensemble docking (combination of every conformer from the receptor ensemble with another one randomly selected from the ligand ensemble*)*

In this context, it is interesting to analyze the reasons for the success of such conformers. Surprisingly, the conformers that were structurally more similar to the reference did not yield better docking results than the unbound structures. On the other side, selected conformers with the best binding energy in the native orientation performed better than the unbound structures. Thus, the capacity to provide favorable binding energy in the native orientation seems to be a major determinant for the success of docking, as opposed to the criterion of structural similarity to the native conformation. This might be due to the fact that in the majority of cases, ensembles were not exploring the conformational space close to the bound state, because sampling was limited to a region in the vicinity of the unbound.

Indeed, it was found that 90% of the successful cases (i.e., near-native solution ranked within top 10) had average conformer binding energy < -20.0 a.u. in the native orientation. Actually, 71% of the docking cases with conformers with binding energy in the native orientation < -40.0 a.u. were successful. This confirms that the existence of conformers with good optimal energy in the native orientation is determining the success of docking. In many of the cases that significantly improved (i.e., those which had a near-native ranked ≤ 10 when using the energy-based selected conformers but not when using the unbound structures), the unbound structures in the native orientation had binding energy < -20.0 a.u. but were not successful in unbound docking. In these cases, a little bit of conformational sampling seems to be sufficient to generate conformers that significantly improve docking results (**Fig 5** from **section 3.1.2**).

Binding mechanism: What can we learn from docking?

Docking simulations can provide interesting insights into the protein binding mechanism. Indeed, the different possible mechanisms that have been proposed for protein-protein association could be described by existing computational approaches (**Fig 6** from **section 3.1.2**). In this context, several possible scenarios can be considered. For protein complexes following rigid association (similar to *lock-and-key* mechanism), the use of rigid-body docking with the unbound subunits could be a suitable approach to describe the binding process and obtain good predictive models. Indeed, this seems to be the case for complexes with small conformational changes between unbound and bound states (I-RMSD$_{C\alpha}$ < 0.5 Å), in which unbound docking already gives as similar success rates as bound docking. In these cases, the use of energy-based selected conformers from unbound ensembles gave also similarly good docking rates as unbound and bound docking. Indeed, for these cases, the unbound proteins in the native orientation generally showed good average binding energy towards the bound partner, not far from that of the bound structures. Consistently, the average binding energy of the best conformers was typically similar to that of the unbound or bound pairs. However, when conformers were selected by criteria based on the structural similarity to bound state, docking success rates were much worse than unbound or bound docking, because in these cases conformational heterogeneity was more likely to produce conformers that are further from the bound state than the unbound one (given that the unbound was already close to the bound state). Indeed, in none of these cases there were a single conformer that was significantly closer to the bound state than the unbound structure.

On the other side, it is known that in complexes involving flexible association, rigid-body docking with the unbound structures is not going to produce correct models. For such cases, different binding mechanisms have been proposed, such as *conformational-selection* or *induced-fit*. For cases following the *conformational-selection* mechanism, the hypothesis is that the unbound proteins naturally sample a variety of conformational states, a subset of which are suitable to bind the other protein. Therefore, for these cases the use of precomputed unbound ensembles describing conformational variability of free proteins in solution should generate conformers that would improve rigid-body docking predictions with respect to those with the unbound structures. Indeed, this is the case for the complexes undergoing unbound to bound transitions between 0.5 and 1.0 Å I-RMSD$_{C\alpha}$. In these cases, selected conformers from the unbound ensembles yielded much better docking predictions than the unbound structures, virtually achieving the success of bound docking.

For cases undergoing unbound-to-bound transition between 1.0 and 2.0 Å I-RMSD$_{C\alpha}$, the use of unbound ensembles also improved the predictions with respect to the unbound docking results, although to a lesser extent. In these cases, the binding energy of the selected pair of conformers in the native orientation was typically better than the unbound structures and similar to the bound structures. Some residues in the best pair of conformers show better energy contribution than in the unbound state, which explains why this specific pair of conformers improves docking results. In these cases, the existence of a sub-population of "active" conformers, with good binding capabilities towards the bound partner, would be consistent with a conformational-selection mechanism. The fact that these conformers with improved binding capabilities are not geometrically closer to the bound state seems counterintuitive. However, recent views of binding mechanism show that active conformers that are selected by partner (initial encounters) do not necessarily need to be in the bound state, as they can adjust their conformations during the association process (Csermely et al., 2010). Indeed, these docking poses are likely to represent these initial encounters between the most populated conformational states of the interacting proteins and would be compatible with this extended conformational selection view (Csermely et al., 2010). However, in other cases the limited conformational sampling used might not be sufficient to explore all conformational states available in solution and therefore the specific binding mechanism cannot be easily identified.

As for the other extreme, in cases following an *induced-fit* mechanism the bound complexes would only be obtained after rearrangement of the interfaces when interacting proteins are approaching to each other, in which case the use of precomputed conformational ensembles in docking (even if generated by exhaustive sampling) would not produce favorable encounters around the native complex structure. This seems the case for complexes undergoing unbound to bound transitions above 3.0 Å I-RMSD$_{C\alpha}$. In all these cases, rigid-body docking, either with unbound structures or with selected conformers, failed to reproduce the experimental complex structure. In this context, an emblematic case is 1IRA, in which binding energy of the selected pair of conformers is similar to the unbound structures and much worse than the bound conformation. For these complexes, the use of precomputed unbound ensembles does not seem to see advantageous, and they would probably need to include flexibility during docking search, mimicking the *induced-fit* mechanism. However, in the flexible category (i.e., unbound to bound transitions between 2.0 and 3.0 Å I-RMSD$_{C\alpha}$,), there are also other cases (i.e., 1ACB), which seem to follow the (extended) *conformational-selection* mechanism, since the use of conformers helped to improve the docking results, and the conformers showed better energy than the unbound structures. Again, there might be other complexes under this category that could still follow the *conformational-selection* mechanism, but our conformational search was not sufficient to sample conformations that may exist in solution and could be productive for docking. This seems to be the case for 1I2M, in which ensembles based on MODELLER did not produce pairs of conformers with sufficiently good binding energy in the native orientation, but the docking rates improved when using extended sampling based on NMA.

Of course the use of docking calculations to learn about the binding mechanism has some limitations, in addition to the ones already mentioned. The timescale of transitions between inactive and active conformers can play an important role in controlling the binding mechanism (Zhou, 2010). In the work presented in this thesis, we can only assume that our ensembles are formed by conformers that are the most accessible in solution, so the existence of active conformers that can be preferentially selected by the bound partner would be compatible (but not exclusively) with a mainly *conformational-selection* mechanism. However, in a situation in which the active conformers are not easily accessible, as those that can only be generated with extended sampling, we could not identify the type of mechanism unless transition rates between conformers were considered.

5.4. Integration of unbound conformational ensembles in protein-protein docking: development and benchmarking of a novel protocol

Ensemble docking provides more near-native poses and as a consequence better predictive rates

In this thesis, we described a protocol to efficiently use in docking the conformational ensembles generated by MODELLER minimization (MM) from the unbound subunits (see **section 3.2.2** and **Figure 12B**). The results showed that a minimal structural heterogeneity provided by such ensembles can improve docking results with respect to the unbound X-ray structures.

Thus, in order to further study the reasons of such improvement, it was first explored for each receptor-ligand pair of conformers whether the docking energy of the best near-native solution (determinant for the docking success) depended on the number of near-native solutions obtained for such conformer pair (**Fig S1** from **section 3.2.2**). For the majority of cases it can be observed that the higher the number of near-native solutions generated by a given conformer pair, the higher the probability of obtaining good docking energies by such near-native solutions. In this line, the conformers that generated more near-native solutions than the unbound structure provided in general better near-native docking energies than those generated by the unbound structure. It was also observed that the bound X-ray structure typically yields more near-native solutions and with better docking energy than the unbound. Interestingly, for many cases, there were a few conformers that generated even more near-native solutions than the bound structure (e.g., 1NSN, 1DFJ, 1I2M). Therefore, increasing the ratio between near-native solutions and false positives is the main reason for the beneficial effect of some of the conformers found in the precomputed unbound ensembles. Indeed, this is consistent with the previously observed correlation between the number of near-native solutions generated by docking and the predictive success rates (Pons et al., 2010b).

For each case, the percentage of conformers that produced more near-native solutions than the unbound structure was also a determinant of the ensemble docking success. Cases with more than

70% of the conformers producing more near-native solutions than the unbound structure showed much higher success rate (72.7%) than the unbound docking (36.4%), almost reaching the optimal bound docking results. On the contrary, in those benchmark cases in which there were less than 70% of conformers that produced more near-native solutions, ensemble docking had similar success rate (41.2%) to when using unbound X-ray structures (35.3%) and far from the bound docking results.

Successful conformers are not necessarily more similar to bound state

As above mentioned, the consideration of conformational heterogeneity in docking can increase the number of near-native solutions generated by FTDock, as well as their docking energy, which is a key determinant for the docking success of each conformer pair. However, neither the number of near-native solutions found for each conformer pair or their best binding energy (and as a consequence the success rate) depended on the similarity of such conformer pair to the bound state (**Figs S2** and **S3** from **section 3.2.1**). This is in agreement with previous findings (see **sections 3.2.2**) and is consistent with an extended *conformational-selection* mechanism (Csermely et al., 2010). Nevertheless, in a few cases (e.g., 1AY7, 1MAH) we did observe that the most successful pair of conformers were the most similar to the bound state in terms of the RMSD of the predicted anchor residues (Rajamani et al., 2004; Meireles et al., 2010) (**Fig 3** from **section 3.2.2**). In these cases, unbound ensembles can explore bound-like orientations of specific interface key residues that can improve the interacting capability of such conformers upon the docking and thus yield better docking results with respect to the unbound.

Conformer pairs providing better binding energy in the native orientation are more likely to improve docking results

We reported that the structural similarity of the docking partners to the native conformation is not determinant for the docking success in general. However, what we found is that the better the binding energy of a conformer pair in the native orientation (i.e., after optimal superimposition on complex structure), the better the docking energy of the produced near-native solutions (and therefore the success rates) (**Fig S5** from **section 3.2.2**), which is also in agreement with our previous findings (see **section 3.2.1**). Thus, capacity to provide favorable binding energy in the native orientation seems to be a major determinant for the success of a given conformer pair.

In this regard, we found that for a given case the predictive success (i.e., best ranked near-native solution) of unbound docking strongly depended on the expected optimal binding energy of the unbound subunits as calculated in the native orientation (**Fig 4** from **section 3.2.2**). All successful docking cases (i.e., best near-native rank ≤ 10) had optimal binding energy of the unbound subunits in the native orientation < 0.0 a.u. A number of unsuccessful cases had also optimal unbound binding energy < 0.0 a.u., but the majority of them (70%) significantly improved in the ensemble docking. Only two cases out of 7 having a pair binding energy < -20.0 a.u. were unsuccessful in docking (1DFJ and 1MAH). Interestingly, the cases with pair binding energy between -20.0 and 0 a.u. seem the ones more benefited by the ensemble docking, since 62% of the successful cases had an optimal binding energy within such range. On the contrary, in cases with worse unbound optimal binding energy (> 0.0 a.u.), the docking results for the ensemble were as poor as those when using the unbound X-ray structures. After ensemble docking, 87% of the successful cases had optimal conformer pair binding energy < -20.0 a.u.. The majority of cases with optimal conformer pair binding energy > -20.0 a.u. were unsuccessful after ensemble docking. All this confirms that the existence of conformers capable of providing favorable binding energy in the native orientation is a major determinant for the success of the ensemble docking.

5.5. Characterization of protein interactions of biological interest: insights from computational models

This PhD thesis included the application of different computational methods to the study of specific cases of biomedical interest. These provided significant insights into the molecular and functional basis of such interactions and could potentially open the path for new challenges and opportunities in biology and biomedicine.

5.5.1. Energetics of host-pathogen protein interactions

Host-pathogen protein interactions control virtually all the key steps of every infection process (e.g., pathogen replication and survival within the host system). Such interactions are devised by the pathogen to subvert and exploit normal host cell processes and typically involve mimicking specific host cell interactions. Analysis of the X-ray crystallographic structure of a protein-protein complex between host and pathogen proteins is essential to provide details at atomic resolution, but it does not always help to identify the residues that are actually responsible for the interaction. In this context, either *in silico* or experimental mutational analysis are useful techniques for the characterization of protein complex interfaces.

One of the studies performed during this thesis consisted on elucidating the energetic basis of the interaction between VipD from *L. pneumophila* and human Rab5 GTPases using *in silico* alanine scanning (see **section 3.3.1**). Human Rab5 interacted with VipD through a helix-turn-helix element that was rather similar to that used by other endogenous ligands (i.e., EEA1, RAbaptin-5 and Rabenosyn-5). Despite the observed overlapping contacts, the energy for VipD binding was not distributed uniformly across the interface but instead concentrated into a combination of hot-spots that provide superior binding affinity and specificity (**Fig 7** from **section 3.3.1**). Notably, these data were confirmed by both experimental mutagenesis and evolutionary conservation analysis. Moreover the identification of a specific polar interaction on VipD-Rab5 interface, responsible for a rather stronger energetic contribution to the interaction in comparison with the other cellular ligands*,* provides the basis for future development of novel therapeutic approaches that, rather than directly targeting the enzyme's active site, could specifically disturb the host factor-mediated activation process of VipD and related microbial phospholipases.

5.5.2. Modeling redox complexes by docking

Despite impressive progress in automating procedures, experimental structure determination remains highly challenging in cases of multi-monomer assemblies, protein membrane and transient complexes. In this context, *ab initio* modeling is a promising alternative technique for the structural prediction of such complexes starting from the isolated component structures. In addition, the analysis of the docking energy landscapes can also provide useful insights into the energetic basis of the association mechanism. In this context, we have analyzed the structural and energetic determinants of *Phaeodactylum* photosystem I (PSI) reduction in different organisms by computational docking, revealing interesting aspects for the kinetics of the reaction of either cytochrome (Cyt) c_6 or plastocyanin (Pc) (see **sections 3.3.2**).

Phaeodactylum Cyt/Phaeodactylum PSI docking model

The analysis of the binding energy landscape of Cyt/PSI for the diatom *Phaeodactylum* showed that the most stable docking orientations are not expected to be efficient for ET due to the longer distance between redox centers. This is consistent with the type III three-steps mechanism found in eukaryotic green systems and previously described for the diatom *Phaeodactylum* (Bernal-Bayard et al., 2013), in which an initial Cyt/PSI encounter complex reorganizes to a more productive final configuration.

In addition, the diatom system shows lower efficiencies than the green systems both in the formation of the properly arranged [Cyt–PSI] complex and in the ET reaction itself (Sigfridsson et al., 1996; Sommer et al., 2002; Hervas et al., 2003; Molina-Heredia et al., 2003; Sommer et al., 2004; Bernal-Bayard et al., 2013)**.** This apparent decreased reactivity is the consequence of diminished basic patches on PsaF and acidic regions on Cyt, both resulting in a weaker electrostatic interaction between partners. This feature of diatoms has been proposed to denote a compromise between ET efficiency and optimal protein donor turnover (Bernal-Bayard et al., 2013), as in the green systems it has been suggested that the strong donor/PSI electrostatic interaction limits the donor exchange and so the overall ET turnover (Drepper et al., 1996; Busch and Hippler, 2011).

In this context, it is interesting to compare the *Phaeodactylum* native Cyt/PSI docking complex (**Fig 5** from **section 3.3.2**) with those described previously in green systems (Sommer et al., 2002; Ben-Shem et al., 2003; Sommer et al., 2006; Busch and Hippler, 2011). It is widely accepted that the lumenal loops i/j of PsaA/B in PSI, including the PsaA W651 and PsaB W627 residues (*Chlamydomonas* numbering), form the hydrophobic recognition site for binding of Pc and Cyt, by means of complementary hydrophobic areas around the donors ET site (Sommer et al., 2002; Sommer et al., 2004). Electrostatic interactions are also established between negatively charged residues of Pc and Cyt with the positively charged N-terminal domain of PsaF (Sommer et al., 2006; Busch and Hippler, 2011). Particularly, *Chlamydomonas* Cyt seems to establish specific interactions involving residues K23/K27 of PsaF and the E69/E70 groups located at the “eastern” negatively charged area of Cyt. Additionally, the positive charge on the “northern” site of Cyt (R66) and the adjacent D65 can form a strong salt bridge with the R623/D624 pair of PsaB (Sommer et al., 2006). According to this model, the distance between the donor/acceptor redox cofactors is ≈14 Å (Sommer et al., 2006; Busch and Hippler, 2011).

On the contrary, the *Phaeodactylum* Cyt/PSI complex built by docking has a different orientation than the *Chlamydomonas* Cyt/PSI complex. The reason is that the D65 group in Chlamydomonas Cyt is not conserved in *Phaeodactylum* Cyt (equivalent residue is Gly109), and thus it cannot stabilize this orientation. As a consequence, it also loses the electrostatic interactions with PsaB and the overall binding energy is less favorable.

Monoraphidium Cyt/Phaeodactylum PSI docking model

Previous results obtained with cross-reactions of different Cyt/PSI eukaryotic systems suggested that the different electrostatic properties of Cyt, more than the PSI, mainly make the difference in behavior of diatoms with respect to other photosynthetic eukaryotes from the green lineage (Bernal-Bayard et al., 2013). Indeed, this has been confirmed by the docking models. The *Monoraphidium* Cyt/*Phaeodactylum* PSI docking complex shows virtually the same orientation as the native *Chlamydomonas* Cyt/PSI green complex, and is able to form similar interactions with the positive patch in PsaF (**Figure 13**) (Sommer et al., 2006). In addition, the salt-bridges formed by D65 and R66 of *Chlamydomonas* Cyt with PsaB R623 and D624 residues are conserved in the *Monoraphidium* Cyt/diatom PSI interaction (equivalent residues: D65 and R67; and R620 and D621, respectively).

The key interface D42/R747 salt-bridge found in our *Monoraphidium* Cyt/*Phaeodactylum* PsaA model was not previously reported for the *Chlamydomonas* Cyt/PSI complex (Sommer et al., 2006), but since these residues are conserved (equivalent ones are D41 and R746), it can be expected that this salt-bridge is also formed in Chlamydomonas Cyt/PSI complex. Interestingly, the redox centers in Monoraphidium Cyt/Phaeodactylum PSI are found at a shorter distance (11.6 Å) than in Chlamydomonas Cyt/PSI (≈ 14 Å) (Sommer et al., 2006).

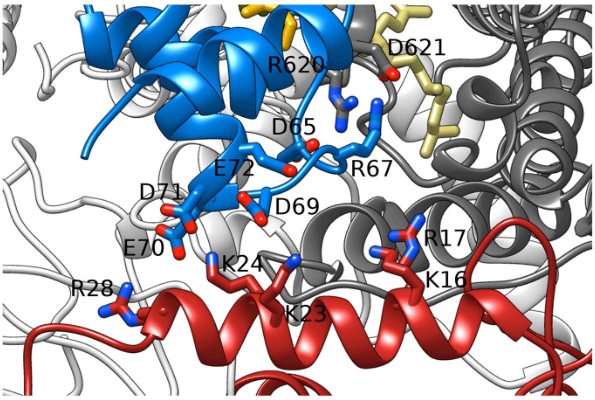


Figure 13. Representative docking model between Phaeodactylum PSI and *Monoraphidium* cytochrome c_6. Details of atomic interactions for best-energy docking models for efficient ET (PsaA W652 / PsaB W624 and Cyt heme groups at less than 3.0 Å distance), rank 14, docking energy -39.2 a.u., distance between Trp residues and cofactors 2.8 Å. *Monoraphidium* cytochrome c6 is depicted in dark blue; the PsaA, PsaB, and PsaF subunits of PSI are depicted in light grey, dark grey, and red, respectively.

Chlamydomonas Pc/Phaeodactylum docking models

The reduction of diatom PSI by the strongly acidic Cyt from green alga showed an increased affinity and k_{ET} but a lower efficiency in the formation of the properly arranged Cyt/PSI complex as compared with the native Cyt, because the too strong electrostatic interactions (Bernal-Bayard et al., 2013). Thus, *Chlamydomonas* Pc mutants were designed by replacing negative groups of the acidic patch –widely accepted to be responsible for electrostatic interactions with PSI (Redinbo et al., 1993; Díaz-Quintana et al., 2008; Busch and Hippler, 2011) -by neutral or positive residues. The rationale for these designs has been to mimic the Cyt electrostatic properties, trying to increase the efficiency of a green Pc in reducing diatom PSI by decreasing the negative character of its acidic patch (**Fig 3** from **section 3.3.2**).

The effect of the different Pc mutations, although moderate, gives interesting information about the binding mechanism to PSI. Thus, from the docking model, WT *Chlamydomonas* Pc seems to be fixed, by means of strong electrostatic interactions, in a less productive complex configuration, that can be improved in mutants showing an increased flexibility in the binding to PSI. Indeed, our docking model shows that E85K and Q88R mutants are destabilizing this less productive complex configuration, which effectively increases the population of the productive orientations and therefore are more efficient for ET.

In this sense it is interesting again to compare the docking model of *Chlamydomonas* Pc/diatom PSI with the previously proposed Pc/PSI interactions in green systems, in which electrostatic interactions involve D42/D44 and E43/E45 of Pc with residues K17/K23/K30 in PsaF (Redinbo et al., 1993; Busch and Hippler, 2011). The *Chlamydomonas* Pc/*Phaeodactylum* PSI most productive docking model conserves such interactions and thus would be able to yield efficient orientations for ET. However, Pc E85 and Q88 residues are stabilizing alternative, but less productive, orientations in the *Chlamydomonas* Pc/diatom PSI complex (**Fig 6** from **section 3.3.2**). This is consistent with the smaller ET efficiency found for *Chlamydomonas* Pc, and the ET increase in E85K and Q88R mutants. Interestingly, *Chlamydomonas* Pc does not possess a positively charged amino acid at a position equivalent to the R66 found in *Chlamydomonas* and *Phaeodactylum* Cyts (corresponding to the R87 position of prokaryotic Pcs). In cyanobacteria, this positively charged amino acid is important for efficient ET to PSI

(Molina-Heredia et al., 2001). Thus, by bringing back this arginine residue (or positively charged guanidinium group) in the Q88R mutant of the green alga Pc, an improved reactivity has been observed.

On the other side, it should be noted that the effect of the two individual E85K and Q88R mutations is counteracted in the double mutant E85K/Q88R, which shows a similar K_A and a slightly diminished k_{ET} compared with the WT Pc. This would be at least partially explained in terms of the small increase of the double mutant redox potential. However, there must be some additional effect that cannot be described in our rigid-body docking simulations, like a conformational change of the two new positive residues that avoids the destabilization effect of the lesser productive configuration by the two individual mutations. Lastly, the results obtained with the V93K protein indicates that this hydrophobic residue is relevant in the ET process, as this mutant shows a significantly decreased k_{ET}, in spite of maintaining the same affinity for PSI than the WT Pc (**Table 2** from **section 3.3.2**).

5.6. Dynamic basis of protein dysfunction: understanding the effect of pathological mutations

In order to understand the mechanistic aspects of several biological processes, such as the functional effects of a single mutation, the consideration of protein plasticity is of paramount importance. Protein kinases constitute a paradigmatic example of the importance of a close link between dynamics and function. These enzymes regularly switch between characteristic inactive and active states undergoing large conformational changes, whose exhaustive computational description is still challenging. Indeed, a comprehensive elucidation of these large-scale transitions is of great value to identify druggable spots, which play a key role during such motions and thus guide the rational design of selective therapeutics.

In this context, we elucidated the molecular basis of oncogenic and CFC-related mutations in MEK1 protein kinase by the application of both large-scale conventional Molecular Dynamics and a state-of-the-art enhanced sampling approach (i.e., PTMetaD-WTE protocol) (see **sections 3.4.1** and **3.4.2**).

According to the conventional MD simulations, all the mutations analyzed seemed to favor the transition from inactive to active state, resulted in destabilizing αA-helix (**Fig 2** from **section 3.4.1**) and promoting the close-to-open transition of the activation loop (**Fig 5** from **section 3.4.1**). Indeed the first event could be related to an inhibitory effect on the negative regulation of MEK1 basal activity modulated by this helix (Mansour et al., 1996; Fischmann et al., 2009), whereas the second could favor the substrate (i.e., ERK) recognition and turnover (Masterson et al., 2010; Masterson et al., 2011). It is also interesting to notice that the effects produced by E203K resulted more dramatic as compared to Q56P and even more with respect to Y130C. Indeed these findings are in agreement not only with the constitutive activation induced by E203K (Nikolaev et al., 2012), but also with the pathological degrees of the different mutations, as previously reported (Rodriguez-Viciana et al., 2006; Emery et al., 2009). Moreover, a distinctive over-activating effect was found in Q56P with respect to Y103C and E203K mutants involving the increase of P-loop flexibility (**Fig 3** from **section 3.4.1**) that could be related to a favouring effect on the cofactor turnover. In general, conventional MD simulations helped to disclose interesting dynamic events regarding MEK1 activation process and the specific destabilization effects caused by each mutation. Nevertheless, given the large conformation changes involved in protein kinases activation process and the intrinsic limits of this methodology in the exhaustive sampling of biomolecule energy landscape (Bowman, 2015), some important aspects of this topic remained hidden.

Thus, in order to improve the exploration of the energy landscape of our system, a state-of-the-art enhanced sampling method, such as PTMetaD-WTE, was applied. As observed through conventional MD simulations and in agreement with the experimental data reported, all the mutations described here facilitated, in more or less degree, the shift of the equilibrium from the inactive to active state. Interesting differences in the specific molecular mechanisms accounting for oncogenic and non-oncogenic mutations were observed. First of all, E203K and Q56P showed significant effects on the closed-to-open transition of the A-loop, stabilizing a virtually equivalent intermediate state in which the A-loop result completely unfolded while the consequences of Y130C mutation (although clearly appreciable) were milder, promoting the stabilization of an intermediate state in which the A-loop is still partially folded (**Fig 2** from **section 3.4.2**). Another significant difference arises from the αC-helix flexibility, with a higher propensity for out-to-in transition observed only in oncogenic mutations (Q56P and E203K) (**Fig 3** from **section 3.4.2**). Finally, it is worth noting the crucial effects on the DFG-motif plasticity caused by the mutations. Indeed, although by using different mechanisms, all the mutants significantly flattened the energy barrier for the DFG in-to-out transition thus promoting the ADP release and increasing the ATP turnover rate. Nevertheless, as previously observed, the effects produced by E203K are generally more dramatic if compared with Q56P and even more with respect to Y130C (**Fig 4** from **section 3.4.2**).

Indeed, the combination of conventional MD and PTMetaD-WTE succeeded in rationalizing and quantifying the activating effects induced by the mutations but also in offering a mechanistic explanation to the different extent of MEK1 over-activation observed for oncogenic or CFC-related mutations, which eventually opens the path for the development of disease specific therapeutic approaches.

6. Conclusions

“Science is forever a search not a discovery,
a journey never an arrival.”

Karl Popper

1. Current computational protocols aimed to model the phylogenetic, structural and energetic properties of residues within protein-protein interfaces show reasonably good predicting performance and consistency;

2. pyDock showed excellent performance in the blind CAPRI experiment, and was classified 5th out of more than 60 participants. The most difficult targets for pyDock included highly flexible proteins or the use of homology-built subunits;

3. The analysis of conformational heterogeneity in precomputed unbound ensembles revealed that docking encounters are favoured by improving the energetic complementarity of the docking partners rather than the geometrical similarity to the bound state;

4. The unbiased use of precomputed conformational ensembles is a successful strategy to incorporate flexibility into a docking approach for low-medium flexible cases, which might follow a conformational selection mechanism. On the contrary, new strategies to integrate flexibility during docking search are needed to improve the performance prediction in high-flexible cases;

5. Computational modeling can complement experimental data to improve our understanding of biological processes involving protein interactions, such as in host-pathogen interactions or electron transfer complexes;

6. Computational modeling of protein plasticity is essential to rationalize and quantify the effects induced by pathological mutations in MEK1.

7. Bibliography

“Wisdom is not a product of schooling
but of the lifelong attempt to acquire it.”

Albert Einstein

Abriata LA, Dal Peraro M. (2015) Assessing the potential of atomistic molecular dynamics simulations to probe reversible protein-protein recognition and binding. Sci Rep 5:10549.

Adamovic I, Mijailovich SM, Karplus M. (2008) The elastic properties of the structurally characterized myosin II S2 subdomain: a molecular dynamics and normal mode analysis. Biophys J 94:3779-3789.

Agrawal NJ, Helk B, Trout BL. (2014) A computational tool to predict the evolutionarily conserved protein-protein interaction hot-spot residues from the structure of the unbound protein. FEBS Lett 588:326-333.

Ahn YO, Mahinthichaichan P, Lee HJ, Ouyang H, Kaluka D, Yeh SR, Arjona D, Rousseau DL, Tajkhorshid E, Adelroth P, Gennis RB. (2014) Conformational coupling between the active site and residues within the K(C)-channel of the Vibrio cholerae cbb3-type (C-family) oxygen reductase. Proc Natl Acad Sci U S A 111:E4419-4428.

Akazaki H, Kawai F, Hosokawa M, Hama T, Chida H, Hirano T, Lim BK, Sakurai N, Hakamata W, Park SY, Nishio T, Oku T. (2009) Crystallization and structural analysis of cytochrome c(6) from the diatom Phaeodactylum tricornutum at 1.5 A resolution. Biosci Biotechnol Biochem 73:189-191.

Alberts B. (1998) The cell as a collection of protein machines: preparing the next generation of molecular biologists. Cell 92:291-294.

Alfalah M, Keiser M, Leeb T, Zimmer KP, Naim HY. (2009) Compound heterozygous mutations affect protein folding and function in patients with congenital sucrase-isomaltase deficiency. Gastroenterology 136:883-892.

Aloy P, Russell RB. (2002a) Interrogating protein interaction networks through structural biology. Proc Natl Acad Sci U S A 99:5896-5901.

Aloy P, Russell RB. (2002b) The third dimension for protein interactions and complexes. Trends Biochem Sci 27:633-638.

Alushin GM, Lander GC, Kellogg EH, Zhang R, Baker D, Nogales E. (2014) High-resolution microtubule structures reveal the structural transitions in alphabeta-tubulin upon GTP hydrolysis. Cell 157:1117-1129.

Anishchenko I, Kundrotas PJ, Tuzikov AV, Vakser IA. (2014) Protein models: the Grand Challenge of protein docking. Proteins 82:278-287.

Anishchenko I, Kundrotas PJ, Tuzikov AV, Vakser IA. (2015) Protein models docking benchmark 2. Proteins 83:891-897.

Antal MA, Bode C, Csermely P. (2009) Perturbation waves in proteins and protein networks: applications of percolation and game theories in signaling and drug design. Curr Protein Pept Sci 10:161-172.

Aoki Y, Niihori T, Banjo T, Okamoto N, Mizuno S, Kurosawa K, Ogata T, Takada F, Yano M, Ando T, Hoshika T, Barnett C, Ohashi H, Kawame H, Hasegawa T, Okutani T, Nagashima T, Hasegawa S, Funayama R, Nakayama K, Inoue S, Watanabe Y, Ogura T, Matsubara Y. (2013) Gain-of-function mutations in RIT1 cause Noonan syndrome, a RAS/MAPK pathway syndrome. Am J Hum Genet 93:173-180.

Aoki Y, Niihori T, Narumi Y, Kure S, Matsubara Y. (2008) The RAS/MAPK syndromes: novel roles of the RAS pathway in human genetic disorders. Hum Mutat 29:992-1006.

Arnold K, Kiefer F, Kopp J, Battey JN, Podvinec M, Westbrook JD, Berman HM, Bordoli L, Schwede T. (2009) The Protein Model Portal. J Struct Funct Genomics 10:1-8.

Assi SA, Tanaka T, Rabbitts TH, Fernandez-Fuentes N. (2010) PCRPi: Presaging Critical Residues in Protein interfaces, a new computational tool to chart hot spots in protein interfaces. Nucleic Acids Res 38:e86.

Atilgan AR, Durell SR, Jernigan RL, Demirel MC, Keskin O, Bahar I. (2001) Anisotropy of fluctuation dynamics of proteins with an elastic network model. Biophys J 80:505-515.

Atkinson DE. (1977) 5 - Enzymes as Control Elements. In: Atkinson DE, editor. Cellular Energy Metabolism and its Regulation. San Diego: Academic Press. p 109-174.

Azad AA, Zoubeidi A, Gleave ME, Chi KN. (2015) Targeting heat shock proteins in metastatic castration-resistant prostate cancer. Nat Rev Urol 12:26-36.

Bader GD, Betel D, Hogue CW. (2003) BIND: the Biomolecular Interaction Network Database. Nucleic Acids Res 31:248-250.

Bahadur RP, Zacharias M. (2008) The interface of protein-protein complexes: analysis of contacts and prediction of interactions. Cell Mol Life Sci 65:1059-1072.

Baral C, Gonzalez G, Gitter A, Teegarden C, Zeigler A, Joshi-Tope G. (2007) CBioC: beyond a prototype for collaborative annotation of molecular interactions from the literature. Comput Syst Bioinformatics Conf 6:381-384.

Barducci A, Bussi G, Parrinello M. (2008) Well-tempered metadynamics: a smoothly converging and tunable free-energy method. Phys Rev Lett 100:020603.

Bashir Q, Volkov AN, Ullmann GM, Ubbink M. (2010) Visualization of the encounter ensemble of the transient electron transfer complex of cytochrome c and cytochrome c peroxidase. J Am Chem Soc 132:241-247.

Beauchamp KA, Lin YS, Das R, Pande VS. (2012) Are Protein Force Fields Getting Better? A Systematic Benchmark on 524 Diverse NMR Measurements. J Chem Theory Comput 8:1409-1414.

Ben-Shem A, Frolow F, Nelson N. (2003) Crystal structure of plant photosystem I. Nature 426:630-635.

Bentires-Alj M, Kontaridis MI, Neel BG. (2006) Stops along the RAS pathway in human genetic disease. Nat Med 12:283-285.

Berman H, Henrick K, Nakamura H, Markley JL. (2007) The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. Nucleic Acids Res 35:D301-303.

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. (2000) The Protein Data Bank. Nucleic Acids Res 28:235-242.

Bernadó P. (2011) Low-resolution structural approaches to study biomolecular assemblies. Wiley Interdisciplinary Reviews: Computational Molecular Science 1:283-297.

Bernal-Bayard P, Molina-Heredia FP, Hervás M, Navarro JA. (2013) Photosystem I Reduction in Diatoms: As Complex as the Green Lineage Systems but Less Efficient. Biochemistry 52:8687-8695.

Biasini M, Bienert S, Waterhouse A, Arnold K, Studer G, Schmidt T, Kiefer F, Cassarino TG, Bertoni M, Bordoli L, Schwede T. (2014) SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. Nucleic Acids Res 42:W252-258.

Bizzarri AR, Cannistraro S. (2002) Molecular Dynamics of Water at the Protein−Solvent Interface. The Journal of Physical Chemistry B 106:6617-6633.

Bockmann RA, Grubmuller H. (2002) Nanoseconds molecular dynamics simulation of primary mechanical energy transfer steps in F1-ATP synthase. Nat Struct Biol 9:198-202.

Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, Pilbout S, Schneider M. (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res 31:365-370.

Boehr DD, Nussinov R, Wright PE. (2009) The role of dynamic conformational ensembles in biomolecular recognition. Nat Chem Biol 5:789-796.

Bogan AA, Thorn KS. (1998) Anatomy of hot spots in protein interfaces. J Mol Biol 280:1-9.

Bolhuis PG, Chandler D, Dellago C, Geissler PL. (2002) Transition path sampling: throwing ropes over rough mountain passes, in the dark. Annu Rev Phys Chem 53:291-318.

Bonomi M, Branduardi D, Bussi G, Camilloni C, Provasi D, Raiteri P, Donadio D, Marinelli F, Pietrucci F, Broglia RA, M P. (2009) PLUMED: a portable plugin for free-energy calculations with molecular dynamics. Comp. Phys. Comm. 180:1961–1972.

Bonomi M, Parrinello M. (2010) Enhanced sampling in the well-tempered ensemble. Phys Rev Lett 104:190601.

Bonvin AM. (2006) Flexible protein-protein docking. Curr Opin Struct Biol 16:194-200.

Bowers PM, Pellegrini M, Thompson MJ, Fierro J, Yeates TO, Eisenberg D. (2004) Prolinks: a database of protein functional linkages derived from coevolution. Genome Biol 5:R35.

Bowie JU, Luthy R, Eisenberg D. (1991) A method to identify protein sequences that fold into a known three-dimensional structure. Science 253:164-170.

Bowman GR. (2015) Accurately modeling nanosecond protein dynamics requires at least microseconds of simulation. J Comput Chem.

Bradford JR, Westhead DR. (2005) Improved prediction of protein-protein binding sites using a support vector machines approach. Bioinformatics 21:1487-1494.

Braun P, Gingras AC. (2012) History of protein-protein interactions: from egg-white to complex networks. Proteomics 12:1478-1498.

Breitkreutz BJ, Stark C, Reguly T, Boucher L, Breitkreutz A, Livstone M, Oughtred R, Lackner DH, Bahler J, Wood V, Dolinski K, Tyers M. (2008) The BioGRID Interaction Database: 2008 update. Nucleic Acids Res 36:D637-640.

Brocchieri L, Karlin S. (2005) Protein length in eukaryotic and prokaryotic proteomes. Nucleic Acids Res 33:3390-3400.

Brooks B, Karplus M. (1985) Normal modes for specific motions of macromolecules: application to the hinge-bending mode of lysozyme. Proc Natl Acad Sci U S A 82:4995-4999.

Brown KR, Jurisica I. (2007) Unequal evolutionary conservation of human protein interactions in interologous networks. Genome Biol 8:R95.

Busch A, Hippler M. (2011) The structure and function of eukaryotic photosystem I. Biochim Biophys Acta 1807:864-877.

Butland G, Peregrin-Alvarez JM, Li J, Yang W, Yang X, Canadien V, Starostine A, Richards D, Beattie B, Krogan N, Davey M, Parkinson J, Greenblatt J, Emili A. (2005) Interaction network containing conserved and essential protein complexes in Escherichia coli. Nature 433:531-537.

Caffarri S, Tibiletti T, Jennings RC, Santabarbara S. (2014) A comparison between plant photosystem I and photosystem II architecture and functioning. Curr Protein Pept Sci 15:296-331.

Caldwell CR. (1989) Temperature-Induced Protein Conformational Changes in Barley Root Plasma Membrane-Enriched Microsomes: III. Effect of Temperature and Cations on Protein Sulfhydryl Reactivity. Plant Physiol 91:1339-1344.

Cao R, Wang Z, Cheng J. (2014) Designing and evaluating the MULTICOM protein local and global model quality prediction methods in the CASP10 experiment. BMC Struct Biol 14:13.

Cao ZW, Xue Y, Han LY, Xie B, Zhou H, Zheng CJ, Lin HH, Chen YZ. (2004) MoViES: molecular vibrations evaluation server for analysis of fluctuational dynamics of proteins and nucleic acids. Nucleic Acids Res 32:W679-685.

Case DA, Cheatham TE, 3rd, Darden T, Gohlke H, Luo R, Merz KM, Jr., Onufriev A, Simmerling C, Wang B, Woods RJ. (2005) The Amber biomolecular simulation programs. J Comput Chem 26:1668-1688.

Castrignano T, De Meo PD, Cozzetto D, Talamo IG, Tramontano A. (2006) The PMDB Protein Model Database. Nucleic Acids Res 34:D306-309.

Cavalli A, Spitaleri A, Saladino G, Gervasio FL. (2015) Investigating drug-target association and dissociation mechanisms using metadynamics-based algorithms. Acc Chem Res 48:277-285.

Ceol A, Chatr-Aryamontri A, Licata L, Cesareni G. (2008) Linking entries in protein interaction database to structured text: the FEBS Letters experiment. FEBS Lett 582:1171-1177.

Clackson T, Wells JA. (1995) A hot spot of binding energy in a hormone-receptor interface. Science 267:383-386.

Collins SR, Kemmeren P, Zhao XC, Greenblatt JF, Spencer F, Holstege FC, Weissman JS, Krogan NJ. (2007) Toward a comprehensive atlas of the physical interactome of Saccharomyces cerevisiae. Mol Cell Proteomics 6:439-450.

Csermely P, Palotai R, Nussinov R. (2010) Induced fit, conformational selection and independent dynamic segments: an extended view of binding events. Trends Biochem Sci 35:539-546.

Chaudhury S, Gray JJ. (2008) Conformer selection and induced fit in flexible backbone protein-protein docking using computational and NMR ensembles. J Mol Biol 381:1068-1087.

Chen H, Sharp BM. (2004) Content-rich biological network constructed by mining PubMed abstracts. BMC Bioinformatics 5:147.

Chen H, Skolnick J. (2008) M-TASSER: an algorithm for protein quaternary structure prediction. Biophys J 94:918-928.

Chen R, Li L, Weng Z. (2003a) ZDOCK: an initial-stage protein-docking algorithm. Proteins 52:80-87.

Chen R, Mintseris J, Janin J, Weng Z. (2003b) A protein-protein docking benchmark. Proteins 52:88-91.

Chen Y, Tascon I, Neunuebel MR, Pallara C, Brady J, Kinch LN, Fernandez-Recio J, Rojas AL, Machner MP, Hierro A. (2013) Structural basis for Rab1 de-AMPylation by the Legionella pneumophila effector SidD. PLoS Pathog 9:e1003382.

Cheng TM, Blundell TL, Fernandez-Recio J. (2007) pyDock: electrostatics and desolvation for effective scoring of rigid-body protein-protein docking. Proteins 68:503-515.

Cheung AY, de Vries SC. (2008) Membrane trafficking: intracellular highways and country roads. Plant Physiol 147:1451-1453.

Cho KI, Kim D, Lee D. (2009) A feature-based approach to modeling protein-protein interaction hot spots. Nucleic Acids Res 37:2672-2687.

Chothia C, Lesk AM. (1986) The relation between the divergence of sequence and structure in proteins. EMBO J 5:823-826.

Davies H, Bignell GR, Cox C, Stephens P, Edkins S, Clegg S, Teague J, Woffendin H, Garnett MJ, Bottomley W, Davis N, Dicks E, Ewing R, Floyd Y, Gray K, Hall S, Hawes R, Hughes J, Kosmidou V, Menzies A, Mould C, Parker A, Stevens C, Watt S, Hooper S, Wilson R, Jayatilake H, Gusterson BA, Cooper C, Shipley J, Hargrave D, Pritchard-Jones K, Maitland N, Chenevix-Trench G, Riggins GJ, Bigner DD, Palmieri G, Cossu A, Flanagan A, Nicholson A, Ho JW, Leung SY, Yuen ST, Weber BL, Seigler HF, Darrow TL, Paterson H, Marais R, Marshall CJ, Wooster R, Stratton MR, Futreal PA. (2002) Mutations of the BRAF gene in human cancer. Nature 417:949-954.

Davis FP, Sali A. (2005) PIBASE: a comprehensive database of structurally defined protein interfaces. Bioinformatics 21:1901-1907.

Day R, Daggett V. (2003) All-atom simulations of protein folding and unfolding. Adv Protein Chem 66:373-403.

De Las Rivas J, Fontanillo C. (2010) Protein-protein interactions essentials: key concepts to building and analyzing interactome networks. PLoS Comput Biol 6:e1000807.

de Vries SJ, van Dijk AD, Bonvin AM. (2006) WHISCY: what information does surface conservation yield? Application to data-driven docking. Proteins 63:479-489.

Deighan M, Bonomi M, Pfaendtner J. (2012) Efficient Simulation of Explicitly Solvated Proteins in the Well-Tempered Ensemble. Journal of Chemical Theory and Computation 8:2189-2192.

Dhillon AS, Hagan S, Rath O, Kolch W. (2007) MAP kinase signalling pathways in cancer. Oncogene 26:3279-3290.

Di Russo NV, Estrin DA, Marti MA, Roitberg AE. (2012) pH-Dependent conformational changes in proteins and their effect on experimental pK(a)s: the case of Nitrophorin 4. PLoS Comput Biol 8:e1002761.

Díaz-Quintana A, Hervás M, Navarro JA, De la Rosa MA. (2008) Plastocyanin and Cytochrome c6: the Soluble Electron Carriers between the Cytochrome b6f Complex and Photosystem I. In: Photosynthetic Protein Complexes: Wiley-VCH Verlag GmbH & Co. KGaA. p 181-200.

Dominguez C, Boelens R, Bonvin AM. (2003) HADDOCK: a protein-protein docking approach based on biochemical or biophysical information. J Am Chem Soc 125:1731-1737.

Doms A, Schroeder M. (2005) GoPubMed: exploring PubMed with the Gene Ontology. Nucleic Acids Res 33:W783-786.

Douguet D, Chen HC, Tovchigrechko A, Vakser IA. (2006) DOCKGROUND resource for studying protein-protein interfaces. Bioinformatics 22:2612-2618.

Drepper F, Hippler M, Nitschke W, Haehnel W. (1996) Binding dynamics and electron transfer between plastocyanin and photosystem I. Biochemistry 35:1282-1295.

Echols N, Milburn D, Gerstein M. (2003) MolMovDB: analysis and visualization of conformational change and structural flexibility. Nucleic Acids Res 31:478-482.

Emekli U, Schneidman-Duhovny D, Wolfson HJ, Nussinov R, Haliloglu T. (2008) HingeProt: automated prediction of hinges in protein structures. Proteins 70:1219-1227.

Emery CM, Vijayendran KG, Zipser MC, Sawyer AM, Niu L, Kim JJ, Hatton C, Chopra R, Oberholzer PA, Karpova MB, MacConaill LE, Zhang J, Gray NS, Sellers WR, Dummer R, Garraway LA. (2009) MEK1 mutations confer resistance to MEK and B-RAF inhibition. Proc Natl Acad Sci U S A 106:20411-20416.

Esteban-Martín S, Bryn Fenwick R, Salvatella X. (2012) Synergistic use of NMR and MD simulations to study the structural heterogeneity of proteins. Wiley Interdisciplinary Reviews: Computational Molecular Science 2:466-478.

Eswar N, Webb B, Marti-Renom MA, Madhusudhan MS, Eramian D, Shen MY, Pieper U, Sali A. (2007) Comparative protein structure modeling using MODELLER. Curr Protoc Protein Sci Chapter 2:Unit 2 9.

Ewing RM, Chu P, Elisma F, Li H, Taylor P, Climie S, McBroom-Cerajewski L, Robinson MD, O'Connor L, Li M, Taylor R, Dharsee M, Ho Y, Heilbut A, Moore L, Zhang S, Ornatsky O, Bukhman YV, Ethier M, Sheng Y, Vasilescu J, Abu-Farha M, Lambert JP, Duewel HS, Stewart, II, Kuehl B, Hogue K, Colwill

K, Gladwish K, Muskat B, Kinach R, Adams SL, Moran MF, Morin GB, Topaloglou T, Figeys D. (2007) Large-scale mapping of human protein-protein interactions by mass spectrometry. Mol Syst Biol 3:89.

Faradjian AK, Elber R. (2004) Computing time scales from reaction coordinates by milestoning. J Chem Phys 120:10880-10889.

Fernandez-Recio J, Totrov M, Abagyan R. (2002) Soft protein-protein docking in internal coordinates. Protein Sci 11:280-291.

Fernandez-Recio J, Totrov M, Skorodumov C, Abagyan R. (2005) Optimal docking area: a new method for predicting protein-protein interaction sites. Proteins 58:134-143.

Fields S, Song O. (1989) A novel genetic system to detect protein-protein interactions. Nature 340:245-246.

Finn RD, Marshall M, Bateman A. (2005) iPfam: visualization of protein-protein interactions in PDB at domain and amino acid resolutions. Bioinformatics 21:410-412.

Fischer E. (1894) Einfluss der Configuration auf die Wirkung der Enzyme. Ber. Dtsch. Chem. Ges. 27:2984–2993.

Fischer TB, Arunachalam KV, Bailey D, Mangual V, Bakhru S, Russo R, Huang D, Paczkowski M, Lalchandani V, Ramachandra C, Ellison B, Galer S, Shapley J, Fuentes E, Tsai J. (2003) The binding interface database (BID): a compilation of amino acid hot spots in protein interfaces. Bioinformatics 19:1453-1454.

Fischmann TO, Smith CK, Mayhood TW, Myers JE, Reichert P, Mannarino A, Carr D, Zhu H, Wong J, Yang RS, Le HV, Madison VS. (2009) Crystal structures of MEK1 binary and ternary complexes with nucleotides and inhibitors. Biochemistry 48:2661-2674.

Flores S, Echols N, Milburn D, Hespenheide B, Keating K, Lu J, Wells S, Yu EZ, Thorpe M, Gerstein M. (2006) The Database of Macromolecular Motions: new features added at the decade mark. Nucleic Acids Res 34:D296-301.

Franklin J, Koehl P, Doniach S, Delarue M. (2007) MinActionPath: maximum likelihood trajectory for large-scale structural transitions in a coarse-grained locally harmonic energy landscape. Nucleic Acids Res 35:W477-482.

Frauenfelder H, Sligar SG, Wolynes PG. (1991) The energy landscapes and motions of proteins. Science 254:1598-1603.

Friedrichs MS, Eastman P, Vaidyanathan V, Houston M, Legrand S, Beberg AL, Ensign DL, Bruns CM, Pande VS. (2009) Accelerating molecular dynamic simulation on graphics processing units. J Comput Chem 30:864-872.

Fu W, Sanders-Beer BE, Katz KS, Maglott DR, Pruitt KD, Ptak RG. (2009) Human immunodeficiency virus type 1, human protein interaction database at NCBI. Nucleic Acids Res 37:D417-422.

Furge LL. (2008) Biochemistry of signal transduction and regulation, 4th Ed. by G. Krauss. Biochemistry and Molecular Biology Education 36:385-385.

Gabb HA, Jackson RM, Sternberg MJ. (1997) Modelling protein docking using shape complementarity, electrostatics and biochemical information. J Mol Biol 272:106-120.

Gaillard T, Martin E, San Sebastian E, Cossio FP, Lopez X, Dejaegere A, Stote RH. (2007) Comparative normal mode analysis of LFA-1 integrin I-domains. J Mol Biol 374:231-249.

Gao M, Skolnick J. (2010) iAlign: a method for the structural comparison of protein-protein interfaces. Bioinformatics 26:2259-2265.

Gao W, Bohl CE, Dalton JT. (2005) Chemistry and structural biology of androgen receptor. Chem Rev 105:3352-3370.

Gao Y, Douguet D, Tovchigrechko A, Vakser IA. (2007) DOCKGROUND system of databases for protein recognition studies: unbound structures for docking. Proteins 69:845-851.

Garzon JI, Lopez-Blanco JR, Pons C, Kovacs J, Abagyan R, Fernandez-Recio J, Chacon P. (2009) FRODOCK: a new approach for fast rotational protein-protein docking. Bioinformatics 25:2544-2551.

Gaspar AH, Machner MP. (2014) VipD is a Rab5-activated phospholipase A1 that protects Legionella pneumophila from endosomal fusion. Proc Natl Acad Sci U S A 111:4560-4565.

Gavin AC, Bosche M, Krause R, Grandi P, Marzioch M, Bauer A, Schultz J, Rick JM, Michon AM, Cruciat CM, Remor M, Hofert C, Schelder M, Brajenovic M, Ruffner H, Merino A, Klein K, Hudak M, Dickson D, Rudi T, Gnau V, Bauch A, Bastuck S, Huhse B, Leutwein C, Heurtier MA, Copley RR, Edelmann A, Querfurth E, Rybin V, Drewes G, Raida M, Bouwmeester T, Bork P, Seraphin B, Kuster B, Neubauer G, Superti-Furga G. (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature 415:141-147.

Geppert T, Hoy B, Wessler S, Schneider G. (2011) Context-based identification of protein-protein interfaces and "hot-spot" residues. Chem Biol 18:344-353.

Ghoorah AW, Devignes MD, Smail-Tabbone M, Ritchie DW. (2011) Spatial clustering of protein binding sites for template based protein docking. Bioinformatics 27:2820-2827.

Ginalski K. (2006) Comparative modeling for protein structure prediction. Curr Opin Struct Biol 16:172-177.

Ginalski K, Elofsson A, Fischer D, Rychlewski L. (2003) 3D-Jury: a simple approach to improve protein structure predictions. Bioinformatics 19:1015-1018.

Giot L, Bader JS, Brouwer C, Chaudhuri A, Kuang B, Li Y, Hao YL, Ooi CE, Godwin B, Vitols E, Vijayadamodar G, Pochart P, Machineni H, Welsh M, Kong Y, Zerhusen B, Malcolm R, Varrone Z, Collis A, Minto M, Burgess S, McDaniel L, Stimpson E, Spriggs F, Williams J, Neurath K, Ioime N, Agee M, Voss E, Furtak K, Renzulli R, Aanensen N, Carrolla S, Bickelhaupt E, Lazovatsky Y, DaSilva A, Zhong J, Stanyon CA, Finley RL, Jr., White KP, Braverman M, Jarvie T, Gold S, Leach M, Knight J, Shimkets RA, McKenna MP, Chant J, Rothberg JM. (2003) A protein interaction map of Drosophila melanogaster. Science 302:1727-1736.

Glaser F, Pupko T, Paz I, Bell RE, Bechor-Shental D, Martz E, Ben-Tal N. (2003) ConSurf: identification of functional regions in proteins by surface-mapping of phylogenetic information. Bioinformatics 19:163-164.

Golbeck JH. (1987) Structure, function and organization of the Photosystem I reaction center complex. Biochim Biophys Acta 895:167-204.

Gong S, Park C, Choi H, Ko J, Jang I, Lee J, Bolser DM, Oh D, Kim DS, Bhak J. (2005a) A protein domain interaction interface database: InterPare. BMC Bioinformatics 6:207.

Gong S, Yoon G, Jang I, Bolser D, Dafas P, Schroeder M, Choi H, Cho Y, Han K, Lee S, Lappe M, Holm L, Kim S, Oh D, Bhak J. (2005b) PSIbase: a database of Protein Structural Interactome map (PSIMAP). Bioinformatics 21:2541-2543.

Grosdidier S, Fernandez-Recio J. (2008) Identification of hot-spot residues in protein-protein interactions by computational docking. BMC Bioinformatics 9:447.

Grunberg R, Leckner J, Nilges M. (2004) Complementarity of structure ensembles in protein-protein binding. Structure 12:2125-2136.

Guerler A, Govindarajoo B, Zhang Y. (2013) Mapping monomeric threading to protein-protein structure prediction. J Chem Inf Model 53:717-725.

Guney E, Tuncbag N, Keskin O, Gursoy A. (2008) HotSprint: database of computational hot spots in protein interfaces. Nucleic Acids Res 36:D662-666.

Gunther S, May P, Hoppe A, Frommel C, Preissner R. (2007) Docking without docking: ISEARCH--prediction of interactions using known interfaces. Proteins 69:839-844.

Haling JR, Sudhamsu J, Yen I, Sideris S, Sandoval W, Phung W, Bravo BJ, Giannetti AM, Peck A, Masselot A, Morales T, Smith D, Brandhuber BJ, Hymowitz SG, Malek S. (2014) Structure of the BRAF-MEK complex reveals a kinase activity independent role for BRAF in MAPK signaling. Cancer Cell 26:402-413.

Hall DA, Vander Kooi CW, Stasik CN, Stevens SY, Zuiderweg ER, Matthews RG. (2001) Mapping the interactions between flavodoxin and its physiological partners flavodoxin reductase and cobalamin-dependent methionine synthase. Proc Natl Acad Sci U S A 98:9521-9526.

Hamelberg D, Mongan J, McCammon JA. (2004) Accelerated molecular dynamics: a promising and efficient simulation method for biomolecules. J Chem Phys 120:11919-11929.

Hansmann UHE. (1997) Parallel tempering algorithm for conformational studies of biological molecules. Chemical Physics Letters 281:140-150.

Heifetz A, Eisenstein M. (2003) Effect of local shape modifications of molecular surfaces on rigid-body protein-protein docking. Protein Eng 16:179-185.

Hervas M, Navarro JA, De La Rosa MA. (2003) Electron transfer between membrane complexes and soluble proteins in photosynthesis. Acc Chem Res 36:798-805.

Hingorani SR, Jacobetz MA, Robertson GP, Herlyn M, Tuveson DA. (2003) Suppression of BRAF(V599E) in human melanoma abrogates transformation. Cancer Res 63:5198-5202.

Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, Adams SL, Millar A, Taylor P, Bennett K, Boutilier K, Yang L, Wolting C, Donaldson I, Schandorff S, Shewnarane J, Vo M, Taggart J, Goudreault M, Muskat B, Alfarano C, Dewar D, Lin Z, Michalickova K, Willems AR, Sassi H, Nielsen PA, Rasmussen KJ, Andersen JR, Johansen LE, Hansen LH, Jespersen H, Podtelejnikov A, Nielsen E, Crawford J, Poulsen V, Sorensen BD, Matthiesen J, Hendrickson RC, Gleeson F, Pawson T, Moran MF, Durocher D, Mann M, Hogue CW, Figeys D, Tyers M. (2002) Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. Nature 415:180-183.

Hoeflich KP, Gray DC, Eby MT, Tien JY, Wong L, Bower J, Gogineni A, Zha J, Cole MJ, Stern HM, Murray LJ, Davis DP, Seshagiri S. (2006) Oncogenic BRAF is required for tumor growth and maintenance in melanoma models. Cancer Res 66:999-1006.

Hoffmann R, Valencia A. (2004) A gene network for navigating the literature. Nat Genet 36:664.

Holding AN. (2015) XL-MS: Protein cross-linking coupled with mass spectrometry. Methods.

Hosur R, Xu J, Bienkowska J, Berger B. (2011) iWRAP: An interface threading approach with application to prediction of cancer-related protein-protein interactions. J Mol Biol 405:1295-1310.

Hu ZZ, Mani I, Hermoso V, Liu H, Wu CH. (2004) iProLINK: an integrated protein resource for literature mining. Comput Biol Chem 28:409-416.

Huang YJ, Hang D, Lu LJ, Tong L, Gerstein MB, Montelione GT. (2008) Targeting the human cancer pathway protein interaction network by structural genomics. Mol Cell Proteomics 7:2048-2060.

Huber LA. (2003) Is proteomics heading in the wrong direction? Nat Rev Mol Cell Biol 4:74-80.

Hutagalung AH, Novick PJ. (2011) Role of Rab GTPases in membrane traffic and cell physiology. Physiol Rev 91:119-149.

Hwang H, Pierce B, Mintseris J, Janin J, Weng Z. (2008) Protein-protein docking benchmark version 3.0. Proteins 73:705-709.

Inbar Y, Benyamini H, Nussinov R, Wolfson HJ. (2003) Protein structure prediction via combinatorial assembly of sub-structural units. Bioinformatics 19 Suppl 1:i158-168.

Isralewitz B, Gao M, Schulten K. (2001) Steered molecular dynamics and mechanical functions of proteins. Curr Opin Struct Biol 11:224-230.

Isserlin R, El-Badrawi RA, Bader GD. (2011) The Biomolecular Interaction Network Database in PSI-MI 2.5. Database (Oxford) 2011:baq037.

Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y. (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proc Natl Acad Sci U S A 98:4569-4574.

Janin J, Bahadur RP, Chakrabarti P. (2008) Protein-protein interaction and quaternary structure. Q Rev Biophys 41:133-180.

Janin J, Henrick K, Moult J, Eyck LT, Sternberg MJ, Vajda S, Vakser I, Wodak SJ. (2003) CAPRI: a Critical Assessment of PRedicted Interactions. Proteins 52:2-9.

Jefferson ER, Walsh TP, Roberts TJ, Barton GJ. (2007) SNAPPI-DB: a database and API of Structures, iNterfaces and Alignments for Protein-Protein Interactions. Nucleic Acids Res 35:D580-589.

Jenssen TK, Laegreid A, Komorowski J, Hovig E. (2001) A literature network of human genes for high-throughput analysis of gene expression. Nat Genet 28:21-28.

Jones DT, Miller RT, Thornton JM. (1995) Successful protein fold recognition by optimal sequence threading validated by rigorous blind testing. Proteins 23:387-397.

Jones DT, Taylor WR, Thornton JM. (1992) A new approach to protein fold recognition. Nature 358:86-89.

Jones R, Ruas M, Gregory F, Moulin S, Delia D, Manoukian S, Rowe J, Brookes S, Peters G. (2007) A CDKN2A mutation in familial melanoma that abrogates binding of p16INK4a to CDK4 but not CDK6. Cancer Res 67:9134-9141.

Jones S, Thornton JM. (1996) Principles of protein-protein interactions. Proc Natl Acad Sci U S A 93:13-20.

Jones S, Thornton JM. (1997) Analysis of protein-protein interaction sites using surface patches. J Mol Biol 272:121-132.

Joo K, Lee J, Sim S, Lee SY, Lee K, Heo S, Lee IH, Lee SJ. (2014) Protein structure modeling for CASP10 by multiple layers of global optimization. Proteins 82 Suppl 2:188-195.

Jordan RA, El-Manzalawy Y, Dobbs D, Honavar V. (2012) Predicting protein-protein interface residues using local surface structural similarity. BMC Bioinformatics 13:41.

Kallberg M, Wang H, Wang S, Peng J, Wang Z, Lu H, Xu J. (2012) Template-based protein structure modeling using the RaptorX web server. Nat Protoc 7:1511-1522.

Kar G, Kuzu G, Keskin O, Gursoy A. (2012) Protein-protein interfaces integrated into interaction networks: implications on drug design. Curr Pharm Des 18:4697-4705.

Karplus K, Barrett C, Hughey R. (1998) Hidden Markov models for detecting remote protein homologies. Bioinformatics 14:846-856.

Karunatilaka KS, Rueda D. (2014) Post-transcriptional modifications modulate conformational dynamics in human U2-U6 snRNA complex. RNA 20:16-23.

Kastritis PL, Moal IH, Hwang H, Weng Z, Bates PA, Bonvin AM, Janin J. (2011) A structure-based benchmark for protein-protein binding affinity. Protein Sci 20:482-491.

Katchalski-Katzir E, Shariv I, Eisenstein M, Friesem AA, Aflalo C, Vakser IA. (1992) Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques. Proc Natl Acad Sci U S A 89:2195-2199.

Kerrien S, Aranda B, Breuza L, Bridge A, Broackes-Carter F, Chen C, Duesbury M, Dumousseau M, Feuermann M, Hinz U, Jandrasits C, Jimenez RC, Khadake J, Mahadevan U, Masson P, Pedruzzi I, Pfeiffenberger E, Porras P, Raghunath A, Roechert B, Orchard S, Hermjakob H. (2012) The IntAct molecular interaction database in 2012. Nucleic Acids Res 40:D841-846.

Kerrigan JE. (2013) Molecular dynamics simulations in drug design. Methods Mol Biol 993:95-113.

Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, Telikicherla D, Raju R, Shafreen B, Venugopal A, Balakrishnan L, Marimuthu A, Banerjee S, Somanathan DS, Sebastian A, Rani S, Ray S, Harrys Kishore CJ, Kanth S, Ahmed M, Kashyap MK, Mohmood R, Ramachandra YL, Krishna V, Rahiman BA, Mohan S, Ranganathan P, Ramabadran S, Chaerkady R, Pandey A. (2009) Human Protein Reference Database--2009 update. Nucleic Acids Res 37:D767-772.

Kiefer F, Arnold K, Kunzli M, Bordoli L, Schwede T. (2009) The SWISS-MODEL Repository and associated resources. Nucleic Acids Res 37:D387-392.

Kiel C, Serrano L. (2014) Structure-energy-based predictions and network modelling of RASopathy and cancer missense mutations. Mol Syst Biol 10:727.

Kim HJ, Bar-Sagi D. (2004) Modulation of signalling by Sprouty: a developing story. Nat Rev Mol Cell Biol 5:441-450.

Kim HW, Shen TJ, Sun DP, Ho NT, Madrid M, Tam MF, Zou M, Cottam PF, Ho C. (1994) Restoring allosterism with compensatory mutations in hemoglobin. Proc Natl Acad Sci U S A 91:11547-11551.

Kiss G, Pande VS, Houk KN. (2013) Molecular dynamics simulations for the ranking, evaluation, and refinement of computationally designed proteins. Methods Enzymol 523:145-170.

Koes DR, Camacho CJ. (2012) PocketQuery: protein-protein interaction inhibitor starting points from protein-protein interaction structure. Nucleic Acids Res 40:W387-392.

Koike R, Ota M. (2012) SCPC: a method to structurally compare protein complexes. Bioinformatics 28:324-330.

Konc J, Depolli M, Trobec R, Rozman K, Janezic D. (2012) Parallel-ProBiS: fast parallel algorithm for local structural comparison of protein structures and binding sites. J Comput Chem 33:2199-2203.

Kortemme T, Baker D. (2002) A simple physical model for binding energy hot spots in protein-protein complexes. Proc Natl Acad Sci U S A 99:14116-14121.

Koshland DE. (1958) Application of a Theory of Enzyme Specificity to Protein Synthesis. Proc Natl Acad Sci U S A 44:98-104.

Kovacs IA, Szalay MS, Csermely P. (2005) Water and molecular chaperones act as weak links of protein folding networks: energy landscape and punctuated equilibrium changes point towards a game theory of proteins. FEBS Lett 579:2254-2260.

Kozakov D, Brenke R, Comeau SR, Vajda S. (2006) PIPER: an FFT-based protein docking program with pairwise potentials. Proteins 65:392-406.

Kraszewski S, Drabik D, Langner M, Ramseyer C, Kembubpha S, Yasothornsrikul S. (2015) A molecular dynamics study of catestatin docked on nicotinic acetylcholine receptors to identify amino acids potentially involved in the binding of chromogranin A fragments. Phys Chem Chem Phys 17:17454-17460.

Krishnan VV, Rupp B. (2001) Macromolecular Structure Determination: Comparison of X-ray Crystallography and NMR Spectroscopy. In: eLS: John Wiley & Sons, Ltd.

Krissinel E, Henrick K. (2007) Inference of macromolecular assemblies from crystalline state. J Mol Biol 372:774-797.

Krogan NJ, Cagney G, Yu H, Zhong G, Guo X, Ignatchenko A, Li J, Pu S, Datta N, Tikuisis AP, Punna T, Peregrin-Alvarez JM, Shales M, Zhang X, Davey M, Robinson MD, Paccanaro A, Bray JE, Sheung A, Beattie B, Richards DP, Canadien V, Lalev A, Mena F, Wong P, Starostine A, Canete MM, Vlasblom J, Wu S, Orsi C, Collins SR, Chandran S, Haw R, Rilstone JJ, Gandi K, Thompson NJ, Musso G, St Onge P, Ghanny S, Lam MH, Butland G, Altaf-Ul AM, Kanaya S, Shilatifard A, O'Shea E, Weissman JS, Ingles CJ, Hughes TR, Parkinson J, Gerstein M, Wodak SJ, Emili A, Greenblatt JF. (2006) Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. Nature 440:637-643.

Kruger DM, Gohlke H. (2010) DrugScorePPI webserver: fast and accurate in silico alanine scanning for scoring protein-protein interactions. Nucleic Acids Res 38:W480-486.

Krumbholz M, Koehler K, Huebner A. (2006) Cellular localization of 17 natural mutant variants of ALADIN protein in triple A syndrome - shedding light on an unexpected splice mutation. Biochem Cell Biol 84:243-249.

Kumar S, Ma B, Tsai CJ, Sinha N, Nussinov R. (2000) Folding and binding cascades: dynamic landscapes and population shifts. Protein Sci 9:10-19.

Kundrotas PJ, Lensink MF, Alexov E. (2008) Homology-based modeling of 3D structures of protein-protein complexes using alignments of modified sequence profiles. Int J Biol Macromol 43:198-208.

Kundrotas PJ, Zhu Z, Janin J, Vakser IA. (2012) Templates are available to model nearly all complexes of structurally characterized proteins. Proc Natl Acad Sci U S A 109:9438-9441.

Laio A, Parrinello M. (2002) Escaping free-energy minima. Proc Natl Acad Sci U S A 99:12562-12566.

Larsson P, Skwark MJ, Wallner B, Elofsson A. (2011) Improved predictions by Pcons.net using multiple templates. Bioinformatics 27:426-427.

Lee SA, Chan CH, Tsai CH, Lai JM, Wang FS, Kao CY, Huang CY. (2008) Ortholog-based protein-protein interaction prediction and its application to inter-species interactions. BMC Bioinformatics 9 Suppl 12:S11.

Lee TI, Young RA. (2013) Transcriptional regulation and its misregulation in disease. Cell 152:1237-1251.

Lensink MF, Mendez R, Wodak SJ. (2007) Docking and scoring protein complexes: CAPRI 3rd Edition. Proteins 69:704-718.

Lensink MF, Moal IH, Bates PA, Kastritis PL, Melquiond AS, Karaca E, Schmitz C, van Dijk M, Bonvin AM, Eisenstein M, Jimenez-Garcia B, Grosdidier S, Solernou A, Perez-Cano L, Pallara C, Fernandez-Recio J, Xu J, Muthu P, Praneeth Kilambi K, Gray JJ, Grudinin S, Derevyanko G, Mitchell JC, Wieting J, Kanamori E, Tsuchiya Y, Murakami Y, Sarmiento J, Standley DM, Shirota M, Kinoshita K, Nakamura H, Chavent M, Ritchie DW, Park H, Ko J, Lee H, Seok C, Shen Y, Kozakov D, Vajda S, Kundrotas PJ, Vakser IA, Pierce BG, Hwang H, Vreven T, Weng Z, Buch I, Farkash E, Wolfson HJ, Zacharias M, Qin S, Zhou HX, Huang SY, Zou X, Wojdyla JA, Kleanthous C, Wodak SJ. (2014) Blind prediction of interfacial water positions in CAPRI. Proteins 82:620-632.

Lensink MF, Wodak SJ. (2010) Docking and scoring protein interactions: CAPRI 2009. Proteins 78:3073-3084.

Lensink MF, Wodak SJ. (2013) Docking, scoring, and affinity prediction in CAPRI. Proteins 81:2082-2095.

Levy ED, Pereira-Leal JB, Chothia C, Teichmann SA. (2006) 3D complex: a structural classification of protein complexes. PLoS Comput Biol 2:e155.

Li S, Armstrong CM, Bertin N, Ge H, Milstein S, Boxem M, Vidalain PO, Han JD, Chesneau A, Hao T, Goldberg DS, Li N, Martinez M, Rual JF, Lamesch P, Xu L, Tewari M, Wong SL, Zhang LV, Berriz GF, Jacotot L, Vaglio P, Reboul J, Hirozane-Kishikawa T, Li Q, Gabel HW, Elewa A, Baumgartner B, Rose DJ, Yu H, Bosak S, Sequerra R, Fraser A, Mango SE, Saxton WM, Strome S, Van Den Heuvel S, Piano F, Vandenhaute J, Sardet C, Gerstein M, Doucette-Stamm L, Gunsalus KC, Harper JW, Cusick ME, Roth FP, Hill DE, Vidal M. (2004) A map of the interactome network of the metazoan C. elegans. Science 303:540-543.

Liang S, Zhang C, Liu S, Zhou Y. (2006) Protein binding site prediction using an empirical scoring function. Nucleic Acids Res 34:3698-3707.

Licata L, Briganti L, Peluso D, Perfetto L, Iannuccelli M, Galeota E, Sacco F, Palma A, Nardozza AP, Santonico E, Castagnoli L, Cesareni G. (2012) MINT, the molecular interaction database: 2012 update. Nucleic Acids Res 40:D857-861.

Lijnzaad P, Argos P. (1997) Hydrophobic patches on protein subunit interfaces: characteristics and prediction. Proteins 28:333-343.

Lindahl ER. (2008) Molecular dynamics simulations. Methods Mol Biol 443:3-23.

Lise S, Buchan D, Pontil M, Jones DT. (2011) Predictions of hot spot residues at protein-protein interfaces using support vector machines. PLoS One 6:e16774.

Liu S, Gao Y, Vakser IA. (2008) DOCKGROUND protein-protein docking decoy set. Bioinformatics 24:2634-2635.

Liu Z, Sun C, Olejniczak ET, Meadows RP, Betz SF, Oost T, Herrmann J, Wu JC, Fesik SW. (2000) Structural basis for binding of Smac/DIABLO to the XIAP BIR3 domain. Nature 408:1004-1008.

Lo Conte L, Chothia C, Janin J. (1999) The atomic structure of protein-protein recognition sites. J Mol Biol 285:2177-2198.

Lo YS, Huang SH, Luo YC, Lin CY, Yang JM. (2015) Reconstructing genome-wide protein-protein interaction networks using multiple strategies with homologous mapping. PLoS One 10:e0116347.

Lu L, Lu H, Skolnick J. (2002) MULTIPROSPECTOR: an algorithm for the prediction of protein-protein interactions by multimeric threading. Proteins 49:350-364.

Lyskov S, Gray JJ. (2008) The RosettaDock server for local protein-protein docking. Nucleic Acids Res 36:W233-238.

Ma B, Kumar S, Tsai CJ, Nussinov R. (1999) Folding funnels and binding mechanisms. Protein Eng 12:713-720.

Ma B, Shatsky M, Wolfson HJ, Nussinov R. (2002) Multiple diverse ligands binding at a single protein site: a matter of pre-existing populations. Protein Sci 11:184-197.

Mansour SJ, Candia JM, Matsuura JE, Manning MC, Ahn NG. (1996) Interdependent domains controlling the enzymatic activity of mitogen-activated protein kinase kinase 1. Biochemistry 35:15529-15536.

Masterson LR, Cheng C, Yu T, Tonelli M, Kornev A, Taylor SS, Veglia G. (2010) Dynamics connect substrate recognition to catalysis in protein kinase A. Nat Chem Biol 6:821-828.

Masterson LR, Shi L, Metcalfe E, Gao J, Taylor SS, Veglia G. (2011) Dynamically committed, uncommitted, and quenched states encoded in protein kinase A revealed by NMR spectroscopy. Proc Natl Acad Sci U S A 108:6969-6974.

McCammon JA, Gelin BR, Karplus M. (1977) Dynamics of folded proteins. Nature 267:585-590.

Meireles LM, Domling AS, Camacho CJ. (2010) ANCHOR: a web server and database for analysis of protein-protein interaction binding pockets for drug discovery. Nucleic Acids Res 38:W407-411.

Mercer KE, Pritchard CA. (2003) Raf proteins and cancer: B-Raf is identified as a mutational target. Biochim Biophys Acta 1653:25-40.

Miller BR, McGee TD, Swails JM, Homeyer N, Gohlke H, Roitberg AE. (2012) MMPBSA.py: An Efficient Program for End-State Free Energy Calculations. Journal of Chemical Theory and Computation 8:3314-3321.

Moal IH, Bates PA. (2010) SwarmDock and the use of normal modes in protein-protein docking. Int J Mol Sci 11:3623-3648.

Moal IH, Fernandez-Recio J. (2012) SKEMPI: a Structural Kinetic and Energetic database of Mutant Protein Interactions and its use in empirical models. Bioinformatics 28:2600-2607.

Molina-Heredia FP, Hervas M, Navarro JA, De la Rosa MA. (2001) A single arginyl residue in plastocyanin and in cytochrome c(6) from the cyanobacterium Anabaena sp. PCC 7119 is required for efficient reduction of photosystem I. J Biol Chem 276:601-605.

Molina-Heredia FP, Wastl J, Navarro JA, Bendall DS, Hervas M, Howe CJ, De La Rosa MA. (2003) Photosynthesis: a new function for an old cytochrome? Nature 424:33-34.

Montelione GT. (2012) The Protein Structure Initiative: achievements and visions for the future. F1000 Biol Rep 4:7.

Moreira IS, Fernandes PA, Ramos MJ. (2007) Computational alanine scanning mutagenesis--an improved methodological approach. J Comput Chem 28:644-654.

Moretti R, Fleishman SJ, Agius R, Torchala M, Bates PA, Kastritis PL, Rodrigues JP, Trellet M, Bonvin AM, Cui M, Rooman M, Gillis D, Dehouck Y, Moal I, Romero-Durana M, Perez-Cano L, Pallara C, Jimenez B, Fernandez-Recio J, Flores S, Pacella M, Praneeth Kilambi K, Gray JJ, Popov P, Grudinin S, Esquivel-Rodriguez J, Kihara D, Zhao N, Korkin D, Zhu X, Demerdash ON, Mitchell JC, Kanamori E, Tsuchiya Y, Nakamura H, Lee H, Park H, Seok C, Sarmiento J, Liang S, Teraguchi S, Standley DM, Shimoyama H, Terashi G, Takeda-Shitaka M, Iwadate M, Umeyama H, Beglov D, Hall DR, Kozakov D, Vajda S, Pierce BG, Hwang H, Vreven T, Weng Z, Huang Y, Li H, Yang X, Ji X, Liu S, Xiao Y, Zacharias M, Qin S, Zhou HX, Huang SY, Zou X, Velankar S, Janin J, Wodak SJ, Baker D. (2013) Community-wide evaluation of methods for predicting the effect of mutations on protein-protein interactions. Proteins 81:1980-1987.

Morrison KL, Weiss GA. (2001) Combinatorial alanine-scanning. Curr Opin Chem Biol 5:302-307.

Mosca R, Ceol A, Aloy P. (2013) Interactome3D: adding structural details to protein networks. Nat Methods 10:47-53.

Moult J, Pedersen JT, Judson R, Fidelis K. (1995) A large-scale experiment to assess protein structure prediction methods. Proteins 23:ii-v.

Muegge I. (2006) PMF scoring revisited. J Med Chem 49:5895-5902.

Mukherjee S, Zhang Y. (2011) Protein-protein complex structure predictions by multimeric threading and template recombination. Structure 19:955-966.

Muller MP, Peters H, Blumer J, Blankenfeldt W, Goody RS, Itzen A. (2010) The Legionella effector protein DrrA AMPylates the membrane traffic regulator Rab1b. Science 329:946-949.

Murzin AG, Brenner SE, Hubbard T, Chothia C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 247:536-540.

Nakayama M, Kikuno R, Ohara O. (2002) Protein-protein interactions between large proteins: two-hybrid screening using a functionally classified library composed of long cDNAs. Genome Res 12:1773-1784.

Negi S. (2014) Effect of Calcium Ion Removal, Ionic Strength, and Temperature on the Conformation Change in Calmodulin Protein at Physiological pH. J Biophys 2014:329703.

Negi SS, Schein CH, Oezguen N, Power TD, Braun W. (2007) InterProSurf: a web server for predicting interacting sites on protein surfaces. Bioinformatics 23:3397-3399.

Negroni J, Mosca R, Aloy P. (2014) Assessing the applicability of template-based protein docking in the twilight zone. Structure 22:1356-1362.

Neuvirth H, Raz R, Schreiber G. (2004) ProMate: a structure based prediction program to identify the location of protein-protein binding sites. J Mol Biol 338:181-199.

Nikolaev SI, Rimoldi D, Iseli C, Valsesia A, Robyr D, Gehrig C, Harshman K, Guipponi M, Bukach O, Zoete V, Michielin O, Muehlethaler K, Speiser D, Beckmann JS, Xenarios I, Halazonetis TD, Jongeneel CV, Stevenson BJ, Antonarakis SE. (2012) Exome sequencing identifies recurrent somatic MAP2K1 and MAP2K2 mutations in melanoma. Nat Genet 44:133-139.

O'Connell MR, Gamsjaeger R, Mackay JP. (2009) The structural analysis of protein-protein interactions by NMR spectroscopy. Proteomics 9:5224-5232.

Ode H, Matsuyama S, Hata M, Neya S, Kakizawa J, Sugiura W, Hoshino T. (2007) Computational characterization of structural role of the non-active site mutation M36I of human immunodeficiency virus type 1 protease. J Mol Biol 370:598-607.

Ofran Y, Rost B. (2007a) ISIS: interaction sites identified from sequence. Bioinformatics 23:e13-16.

Ofran Y, Rost B. (2007b) Protein-protein interaction hotspots carved into sequences. PLoS Comput Biol 3:e119.

Oostenbrink C, Villa A, Mark AE, van Gunsteren WF. (2004) A biomolecular force field based on the free enthalpy of hydration and solvation: the GROMOS force-field parameter sets 53A5 and 53A6. J Comput Chem 25:1656-1676.

Orchard S, Ammari M, Aranda B, Breuza L, Briganti L, Broackes-Carter F, Campbell NH, Chavali G, Chen C, del-Toro N, Duesbury M, Dumousseau M, Galeota E, Hinz U, Iannuccelli M, Jagannathan S, Jimenez R, Khadake J, Lagreid A, Licata L, Lovering RC, Meldal B, Melidoni AN, Milagros M, Peluso D, Perfetto L, Porras P, Raghunath A, Ricard-Blum S, Roechert B, Stutz A, Tognolli M, van Roey K, Cesareni G, Hermjakob H. (2014) The MIntAct project--IntAct as a common curation platform for 11 molecular interaction databases. Nucleic Acids Res 42:D358-363.

Orchard S, Kerrien S, Abbani S, Aranda B, Bhate J, Bidwell S, Bridge A, Briganti L, Brinkman FS, Cesareni G, Chatr-aryamontri A, Chautard E, Chen C, Dumousseau M, Goll J, Hancock RE, Hannick LI, Jurisica I, Khadake J, Lynn DJ, Mahadevan U, Perfetto L, Raghunath A, Ricard-Blum S, Roechert B, Salwinski L, Stumpflen V, Tyers M, Uetz P, Xenarios I, Hermjakob H. (2012) Protein interaction data curation: the International Molecular Exchange (IMEx) consortium. Nat Methods 9:345-350.

Orchard S, Salwinski L, Kerrien S, Montecchi-Palazzi L, Oesterheld M, Stumpflen V, Ceol A, Chatr-aryamontri A, Armstrong J, Woollard P, Salama JJ, Moore S, Wojcik J, Bader GD, Vidal M, Cusick ME, Gerstein M, Gavin AC, Superti-Furga G, Greenblatt J, Bader J, Uetz P, Tyers M, Legrain P, Fields S, Mulder N, Gilson M, Niepmann M, Burgoon L, De Las Rivas J, Prieto C, Perreau VM, Hogue C, Mewes HW, Apweiler R, Xenarios I, Eisenberg D, Cesareni G, Hermjakob H. (2007) The minimum information required for reporting a molecular interaction experiment (MIMIx). Nat Biotechnol 25:894-898.

Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM. (1997) CATH--a hierarchic classification of protein domain structures. Structure 5:1093-1108.

Pagel P, Kovac S, Oesterheld M, Brauner B, Dunger-Kaltenbach I, Frishman G, Montrone C, Mark P, Stumpflen V, Mewes HW, Ruepp A, Frishman D. (2005) The MIPS mammalian protein-protein interaction database. Bioinformatics 21:832-834.

Palma PN, Krippahl L, Wampler JE, Moura JJ. (2000) BiGGER: a new (soft) docking algorithm for predicting protein interactions. Proteins 39:372-384.

Patey GN, Valleau JP. (1975) A Monte Carlo method for obtaining the interionic potential of mean force in ionic solution. The Journal of Chemical Physics 63:2334-2339.

Pavelka A, Chovancova E, Damborsky J. (2009) HotSpot Wizard: a web server for identification of hot spots in protein engineering. Nucleic Acids Res 37:W376-383.

Pereira-Leal JB, Levy ED, Teichmann SA. (2006) The origins and evolution of functional modules: lessons from protein complexes. Philos Trans R Soc Lond B Biol Sci 361:507-517.

Petoukhov MV, Svergun DI. (2005) Global rigid body modeling of macromolecular complexes against small-angle scattering data. Biophys J 89:1237-1250.

Phillips JC, Braun R, Wang W, Gumbart J, Tajkhorshid E, Villa E, Chipot C, Skeel RD, Kale L, Schulten K. (2005) Scalable molecular dynamics with NAMD. J Comput Chem 26:1781-1802.

Pieper U, Webb BM, Barkan DT, Schneidman-Duhovny D, Schlessinger A, Braberg H, Yang Z, Meng EC, Pettersen EF, Huang CC, Datta RS, Sampathkumar P, Madhusudhan MS, Sjolander K, Ferrin TE, Burley SK, Sali A. (2011) ModBase, a database of annotated comparative protein structure models, and associated resources. Nucleic Acids Res 39:D465-474.

Pierce BG, Hourai Y, Weng Z. (2011) Accelerating protein docking in ZDOCK using an advanced 3D convolution library. PLoS One 6:e24657.

Ponder JW, Case DA. (2003) Force fields for protein simulations. Adv Protein Chem 66:27-85.

Pons C, D'Abramo M, Svergun DI, Orozco M, Bernado P, Fernandez-Recio J. (2010a) Structural characterization of protein-protein complexes by integrating computational docking with small-angle scattering data. J Mol Biol 403:217-230.

Pons C, Fenwick RB, Esteban-Martín S, Salvatella X, Fernandez-Recio J. (2013) Validated Conformational Ensembles Are Key for the Successful Prediction of Protein Complexes. Journal of Chemical Theory and Computation 9:1830-1837.

Pons C, Grosdidier S, Solernou A, Perez-Cano L, Fernandez-Recio J. (2010b) Present and future challenges and limitations in protein-protein docking. Proteins 78:95-108.

Pons C, Talavera D, de la Cruz X, Orozco M, Fernandez-Recio J. (2011) Scoring by intermolecular pairwise propensities of exposed residues (SIPPER): a new efficient potential for protein-protein docking. J Chem Inf Model 51:370-377.

Pronk S, Pall S, Schulz R, Larsson P, Bjelkmar P, Apostolov R, Shirts MR, Smith JC, Kasson PM, van der Spoel D, Hess B, Lindahl E. (2013) GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. Bioinformatics 29:845-854.

Rajagopalan H, Bardelli A, Lengauer C, Kinzler KW, Vogelstein B, Velculescu VE. (2002) Tumorigenesis: RAF/RAS oncogenes and mismatch-repair status. Nature 418:934.

Rajamani D, Thiel S, Vajda S, Camacho CJ. (2004) Anchor residues in protein-protein interactions. Proc Natl Acad Sci U S A 101:11287-11292.

Raman S, Vernon R, Thompson J, Tyka M, Sadreyev R, Pei J, Kim D, Kellogg E, DiMaio F, Lange O, Kinch L, Sheffler W, Kim BH, Das R, Grishin NV, Baker D. (2009) Structure prediction for CASP8 with all-atom refinement using Rosetta. Proteins 77 Suppl 9:89-99.

Rauen KA. (2013) The RASopathies. Annu Rev Genomics Hum Genet 14:355-369.

Ravikant DV, Elber R. (2010) PIE-efficient filters and coarse grained potentials for unbound protein-protein docking. Proteins 78:400-419.

Rebholz-Schuhmann D, Arregui M, Gaudan S, Kirsch H, Jimeno A. (2008) Text processing through Web services: calling Whatizit. Bioinformatics 24:296-298.

Redinbo MR, Cascio D, Choukair MK, Rice D, Merchant S, Yeates TO. (1993) The 1.5-.ANG. crystal structure of plastocyanin from the green alga Chlamydomonas reinhardtii. Biochemistry 32:10560-10567.

Rigaut G, Shevchenko A, Rutz B, Wilm M, Mann M, Seraphin B. (1999) A generic protein purification method for protein complex characterization and proteome exploration. Nat Biotechnol 17:1030-1032.

Ritchie DW, Kemp GJ. (2000) Protein docking using spherical polar Fourier correlations. Proteins 39:178-194.

Rodier F, Bahadur RP, Chakrabarti P, Janin J. (2005) Hydration of protein-protein interfaces. Proteins 60:36-45.

Rodriguez-Viciana P, Tetsu O, Tidyman WE, Estep AL, Conger BA, Cruz MS, McCormick F, Rauen KA. (2006) Germline mutations in genes within the MAPK pathway cause cardio-facio-cutaneous syndrome. Science 311:1287-1290.

Rolland T, Tasan M, Charloteaux B, Pevzner SJ, Zhong Q, Sahni N, Yi S, Lemmens I, Fontanillo C, Mosca R, Kamburov A, Ghiassian SD, Yang X, Ghamsari L, Balcha D, Begg BE, Braun P, Brehme M, Broly MP, Carvunis AR, Convery-Zupan D, Corominas R, Coulombe-Huntington J, Dann E, Dreze M, Dricot A, Fan C, Franzosa E, Gebreab F, Gutierrez BJ, Hardy MF, Jin M, Kang S, Kiros R, Lin GN, Luck K, MacWilliams A, Menche J, Murray RR, Palagi A, Poulin MM, Rambout X, Rasla J, Reichert P, Romero V, Ruyssinck E, Sahalie JM, Scholz A, Shah AA, Sharma A, Shen Y, Spirohn K, Tam S, Tejeda AO, Trigg SA, Twizere JC, Vega K, Walsh J, Cusick ME, Xia Y, Barabasi AL, Iakoucheva LM, Aloy P, De Las Rivas J, Tavernier J, Calderwood MA, Hill DE, Hao T, Roth FP, Vidal M. (2014) A proteome-scale map of the human interactome network. Cell 159:1212-1226.

Roskoski R, Jr. (2012) MEK1/2 dual-specificity protein kinases: structure and regulation. Biochem Biophys Res Commun 417:5-10.

Roux B. (2002) Computational studies of the gramicidin channel. Acc Chem Res 35:366-375.

Rual JF, Venkatesan K, Hao T, Hirozane-Kishikawa T, Dricot A, Li N, Berriz GF, Gibbons FD, Dreze M, Ayivi-Guedehoussou N, Klitgord N, Simon C, Boxem M, Milstein S, Rosenberg J, Goldberg DS, Zhang LV, Wong SL, Franklin G, Li S, Albala JS, Lim J, Fraughton C, Llamosas E, Cevik S, Bex C, Lamesch P, Sikorski RS, Vandenhaute J, Zoghbi HY, Smolyar A, Bosak S, Sequerra R, Doucette-Stamm L, Cusick ME, Hill DE, Roth FP, Vidal M. (2005) Towards a proteome-scale map of the human protein-protein interaction network. Nature 437:1173-1178.

Salomon-Ferrer R, Case DA, Walker RC. (2013) An overview of the Amber biomolecular simulation package. Wiley Interdisciplinary Reviews: Computational Molecular Science 3:198-210.

Salsbury FR, Jr. (2010) Molecular dynamics simulations of protein dynamics and their relevance to drug discovery. Curr Opin Pharmacol 10:738-744.

Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, Eisenberg D. (2004) The Database of Interacting Proteins: 2004 update. Nucleic Acids Res 32:D449-451.

Sato Y, Kameya M, Arai H, Ishii M, Igarashi Y. (2011) Detecting weak protein-protein interactions by modified far-western blotting. J Biosci Bioeng 112:304-307.

Schneidman-Duhovny D, Inbar Y, Nussinov R, Wolfson HJ. (2005) PatchDock and SymmDock: servers for rigid and symmetric docking. Nucleic Acids Res 33:W363-367.

Schubbert S, Shannon K, Bollag G. (2007) Hyperactive Ras in developmental disorders and cancer. Nat Rev Cancer 7:295-308.

Schwede T. (2013) Protein modeling: what happened to the "protein structure gap"? Structure 21:1531-1540.

Schymkowitz J, Borg J, Stricher F, Nys R, Rousseau F, Serrano L. (2005) The FoldX web server: an online force field. Nucleic Acids Res 33:W382-388.

Segura J, Fernandez-Fuentes N. (2011) PCRPi-DB: a database of computationally annotated hot spots in protein interfaces. Nucleic Acids Res 39:D755-760.

Shakhnovich BE, Dokholyan NV, DeLisi C, Shakhnovich EI. (2003) Functional fingerprints of folds: evidence for correlated structure-function evolution. J Mol Biol 326:1-9.

Sharma A, Trivedi NR, Zimmerman MA, Tuveson DA, Smith CD, Robertson GP. (2005) Mutant V599EB-Raf regulates growth and vascular development of malignant melanoma tumors. Cancer Res 65:2412-2421.

Sheinerman FB, Norel R, Honig B. (2000) Electrostatic aspects of protein-protein interactions. Curr Opin Struct Biol 10:153-159.

Shields JM, Pruitt K, McFall A, Shaub A, Der CJ. (2000) Understanding Ras: 'it ain't over 'til it's over'. Trends Cell Biol 10:147-154.

Shingate P, Manoharan M, Sukhwal A, Sowdhamini R. (2014) ECMIS: computational approach for the identification of hotspots at protein-protein interfaces. BMC Bioinformatics 15:303.

Shoemaker BA, Panchenko AR, Bryant SH. (2006) Finding biologically relevant protein domain interactions: conserved binding mode analysis. Protein Sci 15:352-361.

Shulman-Peleg A, Shatsky M, Nussinov R, Wolfson HJ. (2007) Spatial chemical conservation of hot spot interactions in protein-protein complexes. BMC Biol 5:43.

Sieben NL, Macropoulos P, Roemen GM, Kolkman-Uljee SM, Jan Fleuren G, Houmadi R, Diss T, Warren B, Al Adnani M, De Goeij AP, Krausz T, Flanagan AM. (2004) In ovarian neoplasms, BRAF, but not KRAS, mutations are restricted to low-grade serous tumours. J Pathol 202:336-340.

Sigfridsson K, He S, Modi S, Bendall DS, Gray J, Hansson O. (1996) A comparative flash-photolysis study of electron transfer from pea and spinach plastocyanins to spinach Photosystem 1. A reaction involving a rate-limiting conformational change. Photosynth Res 50:11-21.

Singer G, Oldt R, 3rd, Cohen Y, Wang BG, Sidransky D, Kurman RJ, Shih Ie M. (2003) Mutations in BRAF and KRAS characterize the development of low-grade ovarian serous carcinoma. J Natl Cancer Inst 95:484-486.

Singh R, Park D, Xu J, Hosur R, Berger B. (2010) Struct2Net: a web service to predict protein-protein interactions using a structure-based approach. Nucleic Acids Res 38:W508-515.

Skjaerven L, Hollup SM, Reuter N. (2009) Normal mode analysis for proteins. Journal of Molecular Structure: THEOCHEM 898:42-48.

Smith GR, Sternberg MJ, Bates PA. (2005) The relationship between the flexibility of proteins and their conformational states on forming protein-protein complexes with an application to protein-protein docking. J Mol Biol 347:1077-1101.

Smyth MS, Martin JH. (2000) x ray crystallography. Mol Pathol 53:8-14.

Soding J, Biegert A, Lupas AN. (2005) The HHpred interactive server for protein homology detection and structure prediction. Nucleic Acids Res 33:W244-248.

Sommer F, Drepper F, Haehnel W, Hippler M. (2004) The hydrophobic recognition site formed by residues PsaA-Trp651 and PsaB-Trp627 of photosystem I in Chlamydomonas reinhardtii confers distinct selectivity for binding of plastocyanin and cytochrome c6. J Biol Chem 279:20009-20017.

Sommer F, Drepper F, Haehnel W, Hippler M. (2006) Identification of precise electrostatic recognition sites between cytochrome c6 and the photosystem I subunit PsaF using mass spectrometry. J Biol Chem 281:35097-35103.

Sommer F, Drepper F, Hippler M. (2002) The luminal helix l of PsaB is essential for recognition of plastocyanin or cytochrome c6 and fast electron transfer to photosystem I in Chlamydomonas reinhardtii. J Biol Chem 277:6573-6581.

Standley DM, Kinjo AR, Kinoshita K, Nakamura H. (2008) Protein structure databases with new web services for structural biology and biomedical research. Brief Bioinform 9:276-285.

Stein A, Russell RB, Aloy P. (2005) 3did: interacting protein domains of known three-dimensional structure. Nucleic Acids Res 33:D413-417.

Stelzl U, Worm U, Lalowski M, Haenig C, Brembeck FH, Goehler H, Stroedicke M, Zenkner M, Schoenherr A, Koeppen S, Timm J, Mintzlaff S, Abraham C, Bock N, Kietzmann S, Goedde A, Toksoz E, Droege A, Krobitsch S, Korn B, Birchmeier W, Lehrach H, Wanker EE. (2005) A human protein-protein interaction network: a resource for annotating the proteome. Cell 122:957-968.

Stites WE. (1997) Protein−Protein Interactions: Interface Structure, Binding Thermodynamics, and Mutational Analysis. Chemical Reviews 97:1233-1250.

Straub FB, Szabolcsi G. (1964) O dinamicseszkij aszpektah sztukturü fermentov (On the dynamic aspects of protein structure). In: Braunstein AE, editor. Molecular Biology, Problems and Perspectives. Moscow: Izdat. Nauka. p 182-187.

Stumpf MP, Thorne T, de Silva E, Stewart R, An HJ, Lappe M, Wiuf C. (2008) Estimating the size of the human interactome. Proc Natl Acad Sci U S A 105:6959-6964.

Sugita Y, Okamoto Y. (1999) Replica-exchange molecular dynamics method for protein folding. Chemical Physics Letters 314:141-151.

Suhre K, Sanejouand YH. (2004) ElNemo: a normal mode web server for protein movement analysis and the generation of templates for molecular replacement. Nucleic Acids Res 32:W610-614.

Sumimoto H, Hirata K, Yamagata S, Miyoshi H, Miyagishi M, Taira K, Kawakami Y. (2006) Effective inhibition of cell growth and invasion of melanoma by combined suppression of BRAF (V599E) and Skp2 with lentiviral RNAi. Int J Cancer 118:472-476.

Sutto L, Marsili S, Gervasio FL. (2012) New advances in metadynamics. Wiley Interdisciplinary Reviews: Computational Molecular Science 2:771-779.

Szilagyi A, Zhang Y. (2014) Template-based structure modeling of protein-protein interactions. Curr Opin Struct Biol 24:10-23.

Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, Simonovic M, Roth A, Santos A, Tsafou KP, Kuhn M, Bork P, Jensen LJ, von Mering C. (2015) STRING v10: protein-protein interaction networks, integrated over the tree of life. Nucleic Acids Res 43:D447-452.

Tang C, Iwahara J, Clore GM. (2006) Visualization of transient encounter complexes in protein-protein association. Nature 444:383-386.

Tartaglia M, Gelb BD. (2005) Noonan syndrome and related disorders: genetics and pathogenesis. Annu Rev Genomics Hum Genet 6:45-68.

Tartaglia M, Zampino G, Gelb BD. (2010) Noonan syndrome: clinical aspects and molecular pathogenesis. Mol Syndromol 1:2-26.

Teilum K, Olsen JG, Kragelund BB. (2009) Functional aspects of protein flexibility. Cell Mol Life Sci 66:2231-2247.

Teyra J, Doms A, Schroeder M, Pisabarro MT. (2006) SCOWLP: a web-based database for detailed characterization and visualization of protein interfaces. BMC Bioinformatics 7:104.

Thangudu RR, Bryant SH, Panchenko AR, Madej T. (2012) Modulating protein-protein interactions with small molecules: the importance of binding hotspots. J Mol Biol 415:443-453.

Thorn KS, Bogan AA. (2001) ASEdb: a database of alanine mutations and their effects on the free energy of binding in protein interactions. Bioinformatics 17:284-285.

Tidow H, Nissen P. (2013) Structural diversity of calmodulin binding to its target sites. FEBS J 280:5551-5565.

Tovchigrechko A, Vakser IA. (2006) GRAMM-X public web server for protein-protein docking. Nucleic Acids Res 34:W310-314.

Tsai CJ, del Sol A, Nussinov R. (2008) Allostery: absence of a change in shape does not imply that allostery is not at play. J Mol Biol 378:1-11.

Tsai CJ, Kumar S, Ma B, Nussinov R. (1999) Folding funnels, binding funnels, and protein function. Protein Sci 8:1181-1190.

Tsai CJ, Lin SL, Wolfson HJ, Nussinov R. (1996) Protein-protein interfaces: architectures and interactions in protein-protein interfaces and in protein cores. Their similarities and differences. Crit Rev Biochem Mol Biol 31:127-152.

Tsai CJ, Lin SL, Wolfson HJ, Nussinov R. (1997) Studies of protein-protein interfaces: a statistical analysis of the hydrophobic effect. Protein Sci 6:53-64.

Tuncbag N, Gursoy A, Guney E, Nussinov R, Keskin O. (2008) Architectures and functional coverage of protein-protein interfaces. J Mol Biol 381:785-802.

Tuncbag N, Keskin O, Gursoy A. (2010) HotPoint: hot spot prediction server for protein interfaces. Nucleic Acids Res 38:W402-406.

Tuncbag N, Keskin O, Nussinov R, Gursoy A. (2012) Fast and accurate modeling of protein-protein interactions by combining template-interface-based docking with flexible refinement. Proteins 80:1239-1249.

Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, Lockshon D, Narayan V, Srinivasan M, Pochart P, Qureshi-Emili

A, Li Y, Godwin B, Conover D, Kalbfleisch T, Vijayadamodar G, Yang M, Johnston M, Fields S, Rothberg JM. (2000) A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. Nature 403:623-627.

Ulrich EL, Akutsu H, Doreleijers JF, Harano Y, Ioannidis YE, Lin J, Livny M, Mading S, Maziuk D, Miller Z, Nakatani E, Schulte CF, Tolmie DE, Kent Wenger R, Yao H, Markley JL. (2008) BioMagResBank. Nucleic Acids Res 36:D402-408.

UniProt C. (2013) Update on activities at the Universal Protein Resource (UniProt) in 2013. Nucleic Acids Res 41:D43-47.

Vajda S, Vakser IA, Sternberg MJ, Janin J. (2002) Modeling of protein interactions in genomes. Proteins 47:444-446.

Vakser IA. (2013) Low-resolution structural modeling of protein interactome. Curr Opin Struct Biol 23:198-205.

Valdar WS. (2002) Scoring residue conservation. Proteins 48:227-241.

Valencia A, Pazos F. (2002) Computational methods for the prediction of protein interactions. Curr Opin Struct Biol 12:368-373.

Velankar S, Kleywegt GJ. (2011) The Protein Data Bank in Europe (PDBe): bringing structure to biology. Acta Crystallogr D Biol Crystallogr 67:324-330.

Venkatesan K, Rual JF, Vazquez A, Stelzl U, Lemmens I, Hirozane-Kishikawa T, Hao T, Zenkner M, Xin X, Goh KI, Yildirim MA, Simonis N, Heinzmann K, Gebreab F, Sahalie JM, Cevik S, Simon C, de Smet AS, Dann E, Smolyar A, Vinayagam A, Yu H, Szeto D, Borick H, Dricot A, Klitgord N, Murray RR, Lin C, Lalowski M, Timm J, Rau K, Boone C, Braun P, Cusick ME, Roth FP, Hill DE, Tavernier J, Wanker EE, Barabasi AL, Vidal M. (2009) An empirical framework for binary interactome mapping. Nat Methods 6:83-90.

von Behren MM, Volkamer A, Henzler AM, Schomburg KT, Urbaczek S, Rarey M. (2013) Fast protein binding site comparison via an index-based screening technology. J Chem Inf Model 53:411-422.

von Mering C, Krause R, Snel B, Cornell M, Oliver SG, Fields S, Bork P. (2002) Comparative assessment of large-scale data sets of protein-protein interactions. Nature 417:399-403.

Vos MD, Martinez A, Ellis CA, Vallecorsa T, Clark GJ. (2003) The pro-apoptotic Ras effector Nore1 may serve as a Ras-regulated tumor suppressor in the lung. J Biol Chem 278:21938-21943.

Vreven T, Moal IH, Vangone A, Pierce BG, Kastritis PL, Torchala M, Chaleil R, Jimenez-Garcia B, Bates PA, Fernandez-Recio J, Bonvin AM, Weng Z. (2015) Updates to the Integrated Protein-Protein Interaction Benchmarks: Docking Benchmark Version 5 and Affinity Benchmark Version 2. J Mol Biol 427:3031-3041.

Wang W, Donini O, Reyes CM, Kollman PA. (2001) Biomolecular simulations: recent developments in force fields, simulations of enzyme catalysis, protein-ligand, protein-protein, and protein-nucleic acid noncovalent interactions. Annu Rev Biophys Biomol Struct 30:211-243.

Warshel A. (2003) Computer simulations of enzyme catalysis: methods, progress, and insights. Annu Rev Biophys Biomol Struct 32:425-443.

Wells JA, McClendon CL. (2007) Reaching for high-hanging fruit in drug discovery at protein-protein interfaces. Nature 450:1001-1009.

Whitmarsh J, Govindjee. (1999) The Photosynthetic Process. In: Singhal GS, Renger G, Sopory SK, Irrgang KD, Govindjee, editors. Concepts in Photobiology: Springer Netherlands. p 11-51.

Winter C, Henschel A, Kim WK, Schroeder M. (2006) SCOPPI: a structural classification of protein-protein interfaces. Nucleic Acids Res 34:D310-314.

Wlodarski T, Zagrovic B. (2009) Conformational selection and induced fit mechanism underlie specificity in noncovalent interactions with ubiquitin. Proc Natl Acad Sci U S A 106:19346-19351.

Wodak SJ, Janin J. (1978) Computer analysis of protein-protein interaction. J Mol Biol 124:323-342.

Wortzel I, Seger R. (2011) The ERK Cascade: Distinct Functions within Various Subcellular Organelles. Genes Cancer 2:195-209.

Wuthrich K. (1990) Protein structure determination in solution by NMR spectroscopy. J Biol Chem 265:22059-22062.

Xia B, Tsui V, Case DA, Dyson HJ, Wright PE. (2002) Comparison of protein solution structures refined by molecular dynamics simulation in vacuum, with a generalized Born model, and with explicit water. J Biomol NMR 22:317-331.

Xia JF, Zhao XM, Song J, Huang DS. (2010) APIS: accurate prediction of hot spots in protein interfaces by combining protrusion index with solvent accessibility. BMC Bioinformatics 11:174.

Xu D, Lin SL, Nussinov R. (1997) Protein binding versus protein folding: the role of hydrophilic bridges in protein associations. J Mol Biol 265:68-84.

Yamada Y, Banno Y, Yoshida H, Kikuchi R, Akao Y, Murate T, Nozawa Y. (2006) Catalytic inactivation of human phospholipase D2 by a naturally occurring Gly901Asp mutation. Arch Med Res 37:696-699.

Yang J, Yan R, Roy A, Xu D, Poisson J, Zhang Y. (2015) The I-TASSER Suite: protein structure and function prediction. Nat Methods 12:7-8.

Yang LW, Liu X, Jursa CJ, Holliman M, Rader AJ, Karimi HA, Bahar I. (2005) iGNM: a database of protein functional motions based on Gaussian Network Model. Bioinformatics 21:2978-2987.

Zacharias M. (2003) Protein-protein docking with a reduced protein model accounting for side-chain flexibility. Protein Sci 12:1271-1282.

Zacharias M. (2004) Rapid protein-ligand docking using soft modes from molecular dynamics simulations to account for protein deformability: binding of FK506 to FKBP. Proteins 54:759-767.

Zacharias M. (2010) Accounting for conformational changes during protein-protein docking. Curr Opin Struct Biol 20:180-186.

Zahiri J, Bozorgmehr JH, Masoudi-Nejad A. (2013) Computational Prediction of Protein–Protein Interaction Networks: Algo-rithms and Resources. Current Genomics 14:397-414.

Závodszky P, Abaturov LV, Varshavsky YM. (1966) Structure of glyceraldehyde-3-phosphate dehydrogenase and its alteration by coenzyme binding. . Acta Biochim. Biophys. Acad. Sci. Hung. 1:389–403.

Zhang L, Hermans J. (1996) Hydrophilicity of cavities in proteins. Proteins 24:433-438.

Zhang Y. (2008) Progress and challenges in protein structure prediction. Curr Opin Struct Biol 18:342-348.

Zhou H, Pandit SB, Skolnick J. (2009) Performance of the Pro-sp3-TASSER server in CASP8. Proteins 77 Suppl 9:123-127.

Zhou HX. (2010) From induced fit to conformational selection: a continuum of binding mechanism controlled by the timescale of conformational transitions. Biophys J 98:L15-17.

Zhu X, Mitchell JC. (2011) KFC2: a knowledge-based hot spot prediction method based on interface solvation, atomic density, and plasticity features. Proteins 79:2671-2683.

Zoete V, Michielin O, Karplus M. (2002) Relation between sequence and structure of HIV-1 protease inhibitor complexes: a model system for the analysis of protein flexibility. J Mol Biol 315:21-52.

List of publications and thesis advisor report

The PhD thesis of Chiara Pallara is based on eight scientific articles, six of them as first author. Four of these articles have been already published in international peer-reviewed journals with impact factor between 1.290 and 9.674 (as indexed in ISI). One more article is currently under consideration by a journal of similar impact factor, and another three articles are soon to be submitted. Here is a list of all the articles to which Chiara Pallara has contributed during the PhD thesis. Only the articles in bold are part of the thesis. Authors marked with # contributed equally to the work.

1. **Pallara C#, Jiménez-García B#, Pérez-Cano L, Romero-Durana M, Solernou A, Grosdidier S, Pons C, Moal IH, Fernandez-Recio J. (2013) Expanding the frontiers of protein-protein modeling: From docking and scoring to binding affinity predictions and other challenges. *Proteins* 81(12), 2192-200.** (Impact factor: 2.921; 6 citations)

   CP actively participated in the generation, analysis and submission of the docking models for all the proposed targets within the fifth CAPRI edition (2010-2012), collected and analyzed pyDock results, and partially wrote the draft.

2. Chen Y, Tascón I, Neunuebel MR, Pallara C, Brady J, Kinch LN, Fernández-Recio J, Rojas AL, Machner MP, Hierro A. (2013) Structural basis for Rab1 de-AMPylation by the

Legionella pneumophila effector SidD. *PLoS Pathog* 9:e1003382. (Impact factor: 8.057; 6 citations)

3. Moretti R, Fleishman SJ, Agius R, Torchala M, Bates PA, Kastritis PL, Rodrigues JP, Trellet M, Bonvin AM, Cui M, Rooman M, Gillis D, Dehouck Y, Moal I, Romero-Durana M, Perez-Cano L, Pallara C, Jimenez B, Fernandez-Recio J, Flores S, Pacella M, Praneeth Kilambi K, Gray JJ, Popov P, Grudinin S, Esquivel-Rodriguez J, Kihara D, Zhao N, Korkin D, Zhu X, Demerdash ON, Mitchell JC, Kanamori E, Tsuchiya Y, Nakamura H, Lee H, Park H, Seok C, Sarmiento J, Liang S, Teraguchi S, Standley DM, Shimoyama H, Terashi G, Takeda-Shitaka M, Iwadate M, Umeyama H, Beglov D, Hall DR, Kozakov D, Vajda S, Pierce BG, Hwang H, Vreven T, Weng Z, Huang Y, Li H, Yang X, Ji X, Liu S, Xiao Y, Zacharias M, Qin S, Zhou HX, Huang SY, Zou X, Velankar S, Janin J, Wodak SJ, Baker D. (2013) Community-wide evaluation of methods for predicting the effect of mutations on protein-protein interactions. *Proteins* 81(11), 1980-7 (Impact factor: 2.921; 33 citations)

4. Lensink MF, Moal IH, Bates PA, Kastritis PL, Melquiond AS, Karaca E, Schmitz C, van Dijk M, Bonvin AM, Eisenstein M, Jimenez-Garcia B, Grosdidier S, Solernou A, Perez-Cano L, Pallara C, Fernandez-Recio J, Xu J, Muthu P, Praneeth Kilambi K, Gray JJ, Grudinin S, Derevyanko G, Mitchell JC, Wieting J, Kanamori E, Tsuchiya Y, Murakami Y, Sarmiento J, Standley DM, Shirota M, Kinoshita K, Nakamura H, Chavent M, Ritchie DW, Park H, Ko J, Lee H, Seok C, Shen Y, Kozakov D, Vajda S, Kundrotas PJ, Vakser IA, Pierce BG, Hwang H, Vreven T, Weng Z, Buch I, Farkash E, Wolfson HJ, Zacharias M, Qin S, Zhou HX, Huang SY, Zou X, Wojdyla JA, Kleanthous C, Wodak SJ. (2014) Blind prediction of interfacial water positions in CAPRI. *Proteins* 82:620-632. (Impact factor: 2.627; 2 citations)

5. **Lucas M, Gaspar A, Pallara C, Rojas A, Fernández-Recio J, Machner M, Hierro A. (2014) Structural basis for the**

recruitment and activation of the Legionella phospholipase VipD by the host GTPase Rab5. *Proc Natl Acad Sci USA* 111(34): E3514-23. (Impact factor: 9.674; 7 citations)

CP performed the interface characterization of the human Rab5 GTPase in complex with several endogenous and pathogen ligands through in silico Alanine (Ala) scanning calculations, participated in the analysis and interpretation of results, and partially wrote the draft.

6. Lang V, Pallara C, Zabala A, Lobato-Gil S, Lopitz-Otsoa F, Farrás R, Hjerpe R, Torres-Ramos M, Zabaleta L, Blattner C, Hay RT, Barrio R, Carracedo A, Fernandez-Recio J, Rodríguez MS, Aillet F. (2014) Tetramerization-defects of p53 result in aberrant ubiquitylation and transcriptional activity. *Mol Oncol*. 8(5):1026-42. (Impact factor: 5.331; 2 citations)

7. **Romero-Durana M#, Pallara C#, Glaser F, Fernández-Recio J. Modeling Binding Affinity of Pathological Mutations for Computational Protein Design. *Methods Mol Biol.* (accepted)** (Impact factor: 1.290)

CP performed and described in silico Alanine (Ala) scanning protocol, actively participate in the assessment and comparison of the computational tools discussed, and partially wrote the draft.

8. **Bernal-Bayard P, Pallara C, Castell MC, Molina-Heredia FP, Fernández-Recio J, Hervás, Navarro JA. Interaction of photosystem I from Phaeodactylum tricornutum with plastocyanins as compared with its native cytochrome c6: reunion with a lost donor. *Biochim Biophys Acta.* 2015 Dec; 1847(12):1549-59** (Impact factor: 5.353)

CP generated the 3D structure Phaeodactylum tricornutum photosystem I (PSI) by homology-based modeling, performed computational docking simulations between PSI and different native and non-native electron transfer proteins, participated in the analysis and interpretation of results, and partially wrote the draft.

9. Lensink MF, Velankar S, Kryshtafovych A, Huang SY, Schneidman D, Sali A, Segura J, Fernandez-Fuentes N, Viswanath S, Elber R, Grudinin S, Popov P, Neveu E, Lee H, Baek M, Park S, Heo L, Lee GR, Seok C, Qin S, Zhou HX, Ritchie DW, Maigret B, Devignes MD, Ghoorah A, Torchala M, Chaleil RAG, Bates PA, Ben-Zeev E, Eisenstein M, Negi SS, Weng Z, Vreven T, Pierce BG, Borrman TM, Yu J, Ochsenbein F, Guerois R, Vangone A, Rodrigues JPGLM, van Zundert G, Nellen M, Xue L, Karaca E, Melquiond ASJ, Visscher K, Kastritis PL, Bonvin AMJJ, Xu X, Qiu L, Chengfei Yan C, Li J, Ma Z, Cheng J, Zou X, Shen Y, Peterson LX, Kim HR, Roy A, Han X, Esquivel-Rodriguez J, Kihara D, Yu X, Bruce NJ, Fuller JC, Wade RC, Anishchenko I, Kundrotas PJ, Vakser IA, Imai K, Yamada K, Oda T, Nakamura T, Tomii K, Pallara C, Romero-Durana M, Jiménez-García B, Moal IH, Fernández-Recio J, Joung JY, Kim JY, Joo K, Lee J, Kozakov D, Vajda S, Mottarella S, Hall DR, Beglov D, Mamonov A, Xia B, Bohnuud T, Del Carpio CA, Ichiishi E, Gray J, Chermak E, Cavallo L, Oliva R and Wodak SJ. Prediction of homo- and hetero-protein complexes by ab-initio and template-based docking: a CASP-CAPRI experiment. *Proteins* (submitted) (Impact factor: 2.627)

10. **Pallara C, Rueda M, Abagyan R, Fernández-Recio J. Conformational Heterogeneity in Unbound State Enhances Recognition in Protein-Protein Encounters *PLoS Comput Biol.* (submitted)** (Impact factor: 4.620)

CP generated the MM and MD-based conformational ensembles, performed the docking simulations, participated in the analysis and interpretation of the results, and partially wrote the draft.

11. **Pallara C, Fernández-Recio J. Protein-protein ensemble docking at low-cost: improving predictive performance for medium-flexible cases (in preparation)**

    CP generated the MM-based conformational ensembles, performed the docking simulations, implemented the ensemble docking procedure, participated in the analysis and interpretation of the results, and partially wrote the draft.

12. **Pallara C, Glaser F, Fernández-Recio J. Structural and dynamic effects of MEK1 pathological mutations: unphosphorylated apo and phosphorylated ATP-bound (in preparation)**

    CP performed the MD simulations on MEK1 protein kinase structure, participated in the analysis and interpretation of the results, and wrote the draft.

13. **Pallara C, Sutto L, Gervasio FL, Fernández-Recio J. Structural and dynamic effects of MEK1 pathological mutations: enhanced sampling Metadynamics simulations (in preparation)**

    CP performed the PTMetaD-WTE simulations on MEK1 protein kinase structure, participated in the analysis and interpretation of the results, and wrote the draft.

14. Pallara C#, Gallastegui N#, Carbó LR, Zieliñska K, MacKinnon JAG, Chen D, Fernández-Recio J, Estébanez-Perpiñá E. Suppression of Androgen Receptor Activities by the NEDD8 Activating Enzyme UBA3 (in preparation)

15. Pallara C, Gallastegui N, Carbó LR, Osguthorpe DJ, Hagler AT, Estébanez-Perpiñá E, Fernández-Recio J. Profiling the pharmacological outcomes and coregulator binding repertoire of ligand-bound androgen receptor (in preparation)

Congress contributions

"You do not really understand something until you can explain it to your grandmother."

Albert Einstein

Posters

2012

1. Pallara C, Fernández-Recio J. (2012) **Precalculated conformational ensembles to improve protein-protein docking results.** Stanford-Sweden multiresolution Molecular simulation workshop. Uppsala (Sweden).

2. Pallara C, Fernández-Recio J. (2012) **Unbound conformational ensembles can improve protein-protein docking predictions.** XI Jornadas de Bioinformatica (JBI2012). Barcelona (Spain).

3. Pallara C, Fernández-Recio J. (2012) **Protein-protein docking with precomputed conformational sampling.** BioNMR 2012 Meeting. Barcelona (Spain).

4. Pallara C, Fernández-Recio J. (2012) **Exploring conformational selection mechanism in protein-protein association by docking.** IUBMB-FEBS 2012. Seville (Spain).

5. Pallara C, Glaser F, Fernández-Recio J. (2012) **Understanding MEK1 mutation causing the CFC (Cardio-Facio-Cutaneous) syndrome by Molecular Dynamics.** Biochemistry, biology and pathology of MAP kinases conference. Jerusalem (Israel)

2013

6. Jiménez-García B, Pallara C, Romero M, Triki D, Fernández-Recio J. (2013) **pyDock version 3: Improvement for high-performence docking and general applicability to non peptidic molecules.** 5th CAPRI Evaluation Meeting. Utrecht (Netherlands)

7. Pallara C, Glaser F, Fernández-Recio J. (2013) **Insight into the effect of CFC syndrome related MEK1 mutations from MD simulations.** Technical Meeting on High-Throughput Molecular Dynamics 2013. Barcelona (Spain).

8. Pallara C, Glaser F, Fernández-Recio J. (2013) **Structural basis for the CFC syndrome-related mutations in MEK1 protein kinase from MD simulations.** I Jornada en Bioinformàtica i Biologia Computacional. Barcelona (Spain).

9. Pallara C, Rueda M, Fernández-Recio J. (2013) **Conformational selection mechanism in protein-protein association: insights from docking.** XIV Congress of the Spanish Biophysical Society. Alcalá de Henares (Spain).

2014

10. Pallara C, Gallastegui N, Carbó LR, Osguthorpe DJ, Hagler AT, Estébanez-Perpiñá E, Fernández-Recio J. (2014) **Profiling the pharmacological outcomes and coregulator binding repertoire of ligand-bound androgen receptor.** Androgens 2014. London (UK).

11. Pallara C, Gallastegui N, Carbó LR, Zieliñska K, MacKinnon JAG, Don Chen J, Fernández-Recio J, Estébanez-Perpiñá E. (2014) **Suppression of AR Activities by the NEDD8 Activating Enzyme UBA3.** Androgens 2014. London (UK).

12. Pallara C, Rueda M, Fernández-Recio J. (2014) **Conformational selection mechanism in protein-protein association: precomputed conformational ensembles improve docking predictions.** Societat Catalana de Biologia and Bioinformatics Barcelona (BIB). Barcelona (Spain)

2015

13. Pallara C, Fernández-Recio J. (2015) **Protein plasticity improves protein-protein description.** 29th Annual Symposium of The Protein Society. Barcelona (Spain)

14. Pallara C, Fernández-Recio J. (2015) **Conformational heterogeneity enhances protein-protein recognition.** SEBBM-Sociedad Española de Bioquímica y Biología Molecular. Valencia (Spain)

Oral communications

2013

1. Pallara C, Fernández-Recio J. (2013) **Unbound conformational ensembles for docking: a complete benchmark study.** Life Sciences Seminars (BSC). Barcelona (Spain)

2014

2. Pallara C, Rueda M, Fernández-Recio J. (2014) **Conformational selection mechanism in protein-protein association: insights from docking.** XIV Congress of the Spanish Biophysical Society. Alcalá de Henares (Spain)

3. Pallara C, Rueda M, Fernández-Recio J. (2014) **Conformational selection mechanism in protein-protein association: precomputed conformational ensembles improve docking predictions.** II Jornada de Bioinformàtica i Biologia Computacional. Barcelona (Spain)

4. Pallara C, Fernández-Recio J. (2014) **Understanding MEK1 pathological mutations by enhanced Molecular Dynamics.** Severo Ochoa Research Seminar Lecture Series (SORS). Barcelona (Spain)

2015

5. Pallara C, Jiménez-García B, Romero M, Fernández-Recio J. (2015) **pyDock performance in 5th CAPRI edition: from docking and scoring to binding affinity predictions and other challenges.** BSC 2nd International Doctoral Symposium 2015. Barcelona (Spain)

6. Pallara C, Fernández-Recio J. (2015) **Identification of the structural and energetic basis of MEK1 pathological mutations: an MD and PTMetaD study.** XXII Jornades de Biologia Molecular. Barcelona (Spain)

7. Pallara C, Fernández-Recio J. (2015) **Conformational heterogeneity enhances protein-protein recognition.** XXXVIII Congreso SEBBM 2015. Valencia (Spain)

8. Pallara C, Bernal-Bayard P, Castell MC, Molina-Heredia FP, Hervás M, Navarro JA, Fernández-Recio J. (2015) **The binding of photosystem I from the diatom Phaeodactylum tricornutum with alternative donors: mechanistic and evolutionary insights from computational docking.** IX Reunión Temática de la Red de Estructura y Función de Proteínas. Seville (Spain)