UNIVERSITAT DE BARCELONA

Environment matters: the impact of urea and macromolecular crowding on proteins

Michela Candotti

UNIVERSITAT DE BARCELONA

Facultat de Biologia

Doctorat en Biomedicina

Environment matters: the impact of urea and macromolecular crowding on proteins

Michela Candotti

November 2015

UNIVERSITAT DE BARCELONA

Facultat de Biologia

Doctorat en Biomedicina

Environment matters: the impact of urea and macromolecular crowding on proteins

Memòria presentada per Michela Candotti per optar al grau de doctora per la Universitat de Barcelona

Tutor i Director:

Modesto Orozco Lopez

Doctoranda:

Michela Candotti

Acknowledgments

This work would not have been possible without the help of many people, starting from my supervisor, Prof. Modesto Orozco, who have always trusted me and gave me freedom. The entire lab, which has changed so much during these years, has been an incredible source of help and friendship. A special thanks goes to Adam and Jose that have always found the time to sort out my informatic-doubts. A mention of honor for La Fede & Palbo, that, beside all the good-laughs, are a great reference in the lab; Ivan for his Balkan humor and for all the shared links (some of them even useful) and Antonija for inspiring well-crafted work (in the lab and in the kitchen) and for all the philosophical conversation (with a vermut). To all the people that I have met in the lab: Pedro, Agusti, Hansel, Nacho, Oscar, Nadine, Rima, Floriane, Rosana, Guillem, Antonella, Annalisa and Marga... it has been an enriching experience to spend these years with all of you. IRB has been almost a second home for me thanks to all the wonderful people that I've met here and the good friends that I've found. A big thanks also to the entire administration department at IRB Barcelona, that I've visited often and where I always felt listened. At IRB I also had the chance to work with many inspiring people: prof. Salvatella for a fruitful collaboration; Santi for the good brainstormings; the PhD Symposium Committee and the Student Council for being such cool teams.

Un agraïment a tot l'equip de Tempesta per ser la font de tantes esperances i idees! Ed infine un grazie speciale anche a tutti quei fili che mi mantengono connessa a casa (Cinzia parlo sopratutto di te!) e, banale ma vero, alla mia famiglia per l'appoggio constante e l'infinita comprensione!

CONTENTS

List of Figures

List of Tables

Abbreviations

AIM Atoms in Molecules
BB Backbone
CC Contact Coefficient
CM Contact Map
CROW proteic crowding
DoF Degree of Freedom
EM Electron microscopy
FRET Fluorescence Resonance Energy Transfer
FSS First Solvation Shell
HB, H-bond Hydrogen Bond
HW Hot Water
IDP Intrinsically Disordered Protein
IOD Intrinsically Ordered Protein
MD Molecular Dynamics
MG Molten Globule
MM Molecular Mechanics
MSD Mean Square Displacement
NMR Nuclear Magnetic Resonance
NOE Nuclear Overhauser Effect
PBC Periodic Boundary Conditions
PC Protein Core
PDB Protein Data Bank
PEG Polyethyleneglycol
PME Particle Mesh Ewald
QM Quantum Mechanics
RDC Residual Dipolar Coupling
Rgyr Radius of Gyration
RMSD Root Mean Square Deviation
RMSF Root Mean Square Fluctuation
SASA Solvent Accesible Surface Area
SAXS Small-angle X-ray scattering
SC Sidechains
SCOP Structural Classification of Proteins
SI Structure Index
SS Secondary Structure
UBQ Ubiquitin
VdW Van der Waals Interactions

"One day I will find the right words, and they will be simple"

Jack Kerouac, The Dharma Bums

OVERVIEW

Life is all about adaption to the different environment. Proteins, similarly to the organisms they belong to, should function under ever-changing conditions. Their flexibility and structural multiplicity promote adjustment to the versatile context. This work aims to understand analytically the impact of two diametric opposite environments on protein structure and dynamics and compared them to the most common solvent on earth: water.

The first environment is a traditional denaturing solution (urea 8M), which has served for years in protein science laboratories to investigate protein stability. The second environment instead moves towards a more physiological representation of proteins: the crowded cell cytoplasm. Despite years of work, the nature of proteins in these two conditions is still unclear, which limits our knowledge of the solvent-dependent polymorphism of proteins.

The presented work aims to overcome this lack of knowledge and it

offers a comparative study of the most general systems: an assorted spectrum of proteins folds, several stages along the reaction path (early stages or end-point) and/or various protein force-fields. Our primary objective was to challenge the specific experimental settings and, when possible, determine the standard pattern and general rules valid at proteome level.

Here we focus on three major aspects of proteins: the structure, the dynamic and the interactions with the solvent molecules – studied at both global (molecular level) as well as local (atomic level). Molecular dynamics (MD) simulation is a suitable tool to study such properties thanks to its powerful capability to: *i)* analyze proteins at a broad range of resolutions (from single atom to single-molecule); *ii)* access the direct time-resolved dynamic of the system; and *iii)* dissect the specific interactions that arise in the new environmental settings.

1.1. Thesis organization

This thesis is a compilation of three published (or in the process of publication) works; the first two investigate proteins in urea-aqueous solution while the last one focuses on macromolecular crowding. They are presented following the chronological order of publication, and the trend follows an increased in the system resolution (single protein, many folds, many protein types). In all of them the analyses are anchored on three aspects of proteins: structure, dynamic and interactions with the solvent; however the specific relevance given to each of these features depends on the project. Given the essentiality of these three aspects, **Chapter 1** introduces the central concepts related to proteins that are relevant to understand this work. **Chapter 2** moves into the realm of the methodology employed here, MD simulations, presenting its theoretical framework within the field of computational biophysics and molecular modeling. **Chapter 3** is a handbook that aims to facilitate the understanding of the many analysis employed in this work, most common to all the three projects while some exclusive to one. The handbook is supposed to be read alongside with the results section (see later) and, therefore, it contains cross-references to the figures in the publications, where the analysis is applied. All together

this information should provide a solid ground to understand better the details and the relevance of the three publications, included in **Chapter 4** (urea-related) and **Chapter 5** (crowding-related). A brief synopsis contextualizes each work and specifies the objectives for that project; each article then has its introduction, methodology, results, discussion and supplementary information sections following the structure dictated by the journal. A summary of the major results is presented in **Chapter 6** which also contains a general discussion that connects and compares the three projects, leading, then, to the main conclusions of this work.

"He said science was going to discover the basic secret of life some day"
"Didn't I read in the paper the other day where they'd finally found out what it was?'
"I missed that", I murmured.
"I saw that," said Sandra.
"About two days ago."
"That's right", said the bartender.
"What is the secret of life?" I asked.
"I forget", said Sandra.
"Protein" the bartender declared.
"They found out something about proteins."

KURT VONNEGUT - THE CAT'S CRADDLE

CHAPTER 1

Proteins as flexible structures

This chapter introduces proteins and their three features that are the leitmotif through all my research: their structure, dynamics and interactions with the solvent. Since the primary aim here is to provide a background for the study at hand, only relevant topics are addressed. For a more comprehensive view, the reader is referred to specific books, such as [1], [2].

1.1. THE WORKING CLASS OF THE CELL

When I imagine a cell, I picture in my mind a busy city filled with strange inhabitants. I imagine the cell landscape dominated by a large dome-like building; a temple? A library? It stores all the available knowledge that is carefully read, transcribed and interpreted by dedicated lit-

erates. Outside, the noisy streets are filled with a chaotic crowd – a heterogeneous group of tireless hard-workers of all kinds: bulky globules, elongated elastic structures and even some amorphous ones that change shape definitely too often. All together they represent the working class at the basis of the cell hierarchy. Possibly the Swedish chemist Berzelius had a similar view when, in 1838, he was looking for an appropriated name for those industrious creatures that populate the cell. He called them proteins from Greek *protos*, meaning "the first" or "of primary importance".

To sustain the work of a complex system such a cell, proteins need to exert an enormous variety of function: transport, regulation, message delivery, product manufacture, waste cleanup, etc. But how can proteins achieve such a broad array of duties? As engineers or industrial designers would say, a look at the structure of a machine gives a first glimpse into what it is capable of. Similarly, in biology the many functions of proteins are intimately linked to their incredibly complicated and precisely detailed structure - some examples of which can be found in **Figure 1.1**. Despite the complexity, the 3D structure relies on few principles of its essential building unit: the amino-acids. Frederick Sanger gave chemical dignity to proteins when he identified the sequence of insulin [3]; however it was Kaj Ulrick Linderstrøm-Lang who in 1951 classified their structural features in a clear and practical scheme, ubiquitously used in every biology textbook [4]. In the next sections, I will follow his bottom-up classification (from primary to quarterly); first in the list: the primary structure.

1.2. The bricks of protein structure: the aminoacids

Proteins are linear chains (polymers) build by an ordered sequence of amino acids. Only 20 amino acids are available in nature but the number that hardly frustrate the protein variety; after all even our language doesn't suffer from scarcity with only twenty-one letters available. For example 20^{10} possible sequences of 100 amino acids exist, with each sequence uniquely defining a protein. In each of the 20 amino acids we can distinguish two parts **Figure 1.2**:

- The discriminatory **side-chain**, often identified as –R and responsible

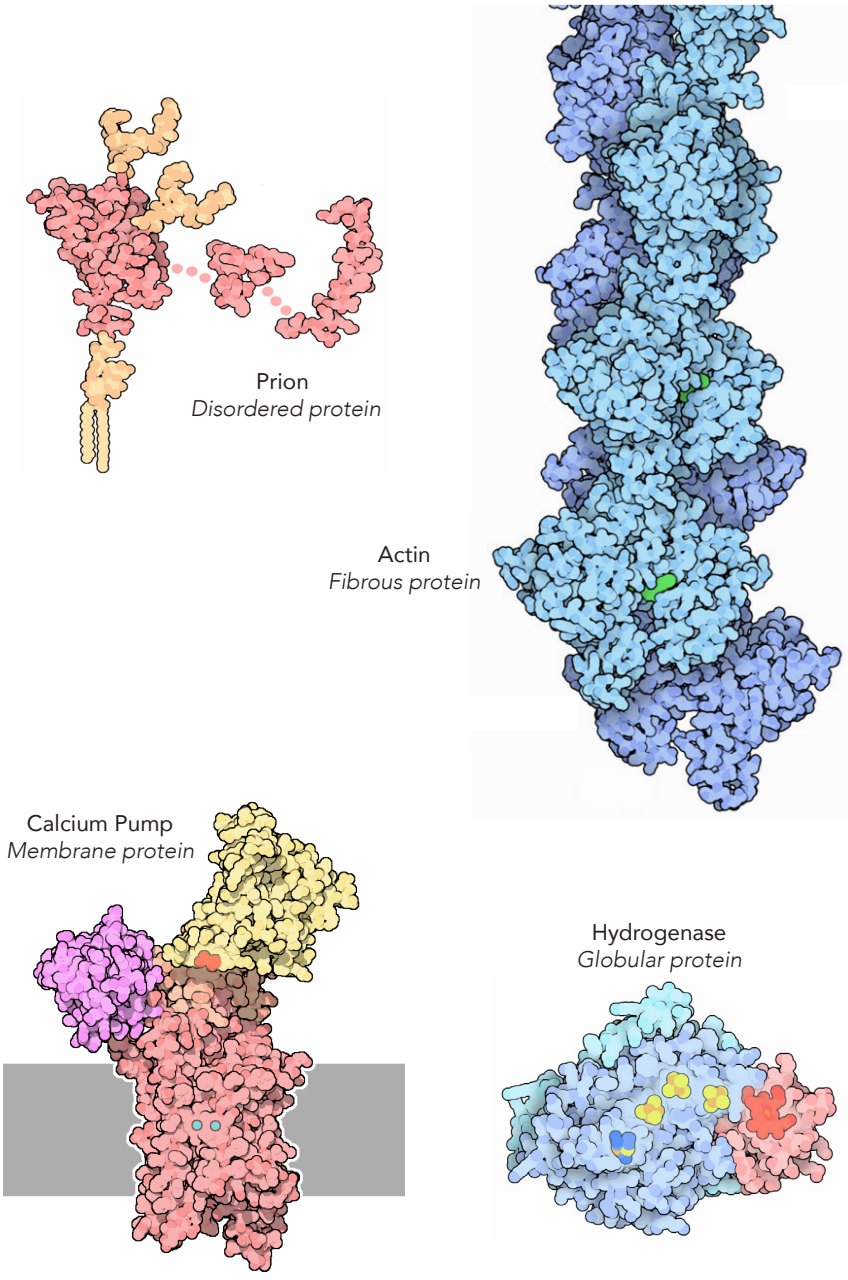


Figure 1.1. The variety of protein types. The structural representation of four proteins representing four different classes - from the RCSB PDB Molecule of the Month feature by David Goodsell.

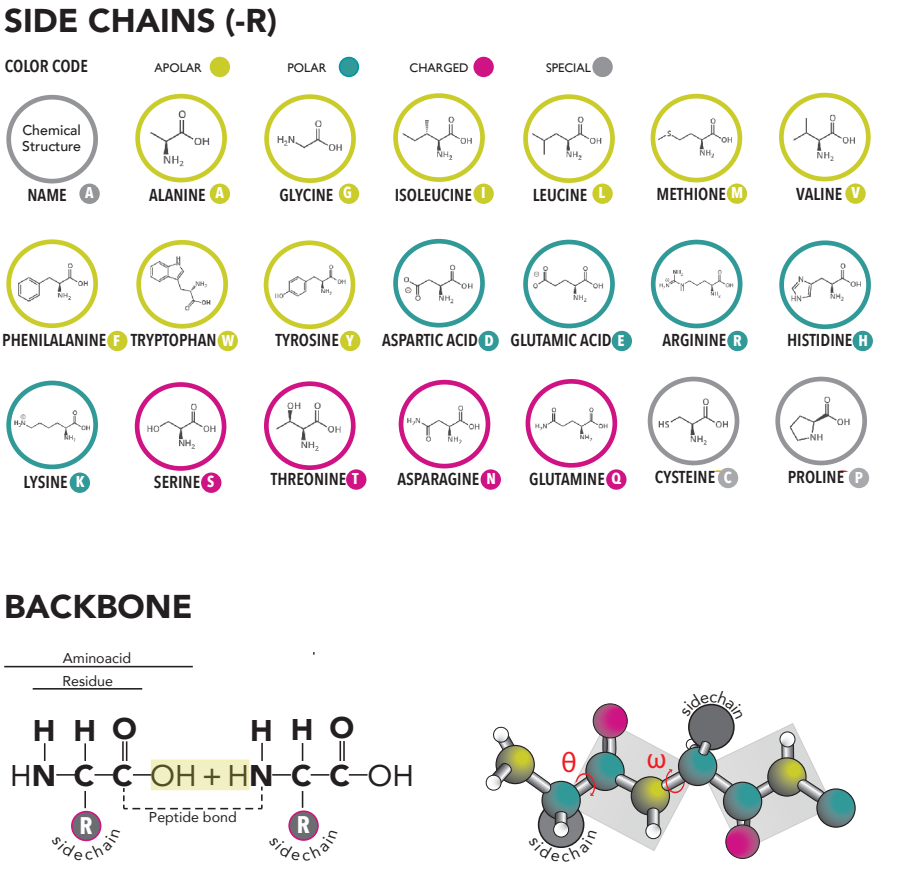


Figure 1.2. The protein aminoacids. On top: sidechains grouped according to the polarity; on the bottom the chemical structure of the backbone on the left, on the right the 3D rearrangements of the two dihedral planes.

for the peculiar physicochemical properties of each amino-acid. For the purpose of this thesis I will classify them according to their chemical polarity - the tendency to be attracted or repelled by electrical charges due to the asymmetrical charge distribution. Low polarity is typical of apolar residues, followed by polar ones and taken to the extreme by charged residues

- The **backbone**, common to all the aminoacids and responsible for linking adjacent units via a covalent peptide bond between the carbon

C and nitrogen N atoms, releasing a water molecule H_2O.

Two covalently bound residues –the portion of the free amino acid that remains after the polymerization- create a planar unit due to the partial double bond character of the peptide bond. This formation arises from the delocalization of the electron cloud covering the C-N bond atoms that prevent any rotations around the covalent bond. Such planarity severely constrains the backbone flexibility, allowing only rotations of the peptide planes around the adjacent bonds. The related degrees of freedom are quantitatively described by two dihedral (torsional angles) θ,ω that capture the position of atomic groups about the peptide bond. In the circular range [–180°, +180°] of θ,ω values, few combinations prevent steric clashes between atoms. The allowed ones are generally reported on a two-dimensional map, the Ramachandran plot (named after the scientists that introduced it), the shape of which can vary with specific features of each residues and, as we will see later, not all the allowed regions in the Ramachandran plot are equally likely [5].

1.3. The determination of the protein structure

We generally accept that a protein spontaneously self-assemble in a process called folding and reach a *unique* functional structure, known as native structure [6]. Unique refers to a structure in which each atom can be represented at high resolution (Å) in an average position over an ensemble of small fluctuations. These structures are often derived by physical methods such as nuclear magnetic resonance (NMR) spectroscopy or X-ray crystallography and collected in a large database, the protein data bank (PDB; available online at www.rcsb.org).

The first explosion of available protein structures was driven by a method invented by Max Perutz, the heavy atom replacement that allowed, in principle, solving structure for any protein that could be crystallized [7]. X-ray crystallography determines the mean position of atoms and their chemical bonds based on the diffraction of a beam of X-rays by a single crystal into many specific directions. Multiple two-dimensional images taken at different rotations allow reconstructing the three-dimensional

model of the electron density within the crystal, but its resolution depends on, among other factors, the quality of the crystal. Indeed over expressing a protein and obtaining a crystal is extremely challenging and it still poses considerable difficulties in the case of flexible systems or membrane proteins.

Later, in the 1980s, when enough powerful equipment and techniques were available, NMR started to be applied to protein structure determination, thanks to pioneers such as Kurt Wuthrich [8]. From the PDB statistics, the 5% of all the newly deposited protein structures are solved by NMR spectroscopy. NMR method for protein determination is based on the magnetic features of atomic nuclei possessing a nuclear spin. The chemical environment of those nuclei gives rise to a set of observables such as the Nuclear Overhauser Effect (NOE) and the chemical shift and its derivative spin-spin coupling (J-Coupling), both extremely useful in structure determination. NOEs quantitatively describe the intensity enhancement experienced by protons when another spatially close proton is saturated or inverted. The strength of the NOE signal then depends only on the spatial proximity of protons and it can be used to originate distance restraints between atoms within 1.8 Å and 6 Å. Angles restraints instead can be obtained from J-coupling between active spin nuclei (such as 13C and 15N usually employed in NMR experiments) but only for atoms linked by 2-3 covalent bonds. Therefore, they can estimate the two torsional angles θ, ω of the protein backbone. The intrinsically low sensitivity of NMR and the high complexity of obtained spectra also hamper its application to protein over 40-60 kDa.

Overall such information defines well the rigid or semi-rigid elements of the protein structure, while information on the dynamics up to the millisecond range became accessible only later with the development of special NMR techniques. Among them residual dipolar coupling (RDC) stand out as a resource for both structural information at the long distance and dynamical information on the time scales slower than a nanosecond, as we will see in Section 1.10 .

For tricky systems, other techniques can help to define structural details, despite at a lower resolution. Electron microscopy (EM), and particularly

the cryo-EM, can be combined with the single-particle reconstruction method to define large macromolecular complexes difficult to crystallize [9]; electron crystallography is primarily used for membrane proteins [10]; Small-angle X-ray and neutron scattering (SAXS and SANS) are applied to structures in solution [11], and lastly homology modeling exploits the structure details of a homologue proteins with an high sequence identity [12].

1.4. Protein folding

The capabilities of NMR and X-ray crystallography had biased the resolvable structures; the information collected in the past 50 years represent mainly a specific subset of proteins: globular proteins, soluble in water and with a well-ordered structure that facilitate its isolation and characterization. These proteins, to which I will refer as intrinsically ordered proteins (IOPs), have been so well characterized for years that they have become the stereotype of protein structure. However, as we will see in Section 1.9 , disorder in proteins exists at physiological conditions. Despite elusive to experiments for long, a new class of protein, the intrinsically disordered proteins IDPs, has become prominent in the last 10 years. That said, here I will concentrate on basic principles of protein structure and folding found by pioneering studies on IOPs. Some of those principles extend to all protein types while others, that I will punctually stressed, apply only to IOPs.

In the 60s Christian B. Anfinsen made a discovery that nowadays puzzle for its simplicity [13]. He proved that a protein, the IOP ribonuclease A, could refold back into its native functional structure, lost during a process known as denaturation. The reversibility of folding had one strong implication: it is the sequence of the protein (the primary structure) that encodes all the necessary information to adopt the native conformation and manifest the functional structural feature.

The spontaneous folding of an IOP into a single structure puzzled Cyrus Levinthal [14]. He described it as a paradox of nature: a polypeptide can assume a vast number of conformations, so large that it would be impossible to adjust randomly to the correct one on a micro/millisecond

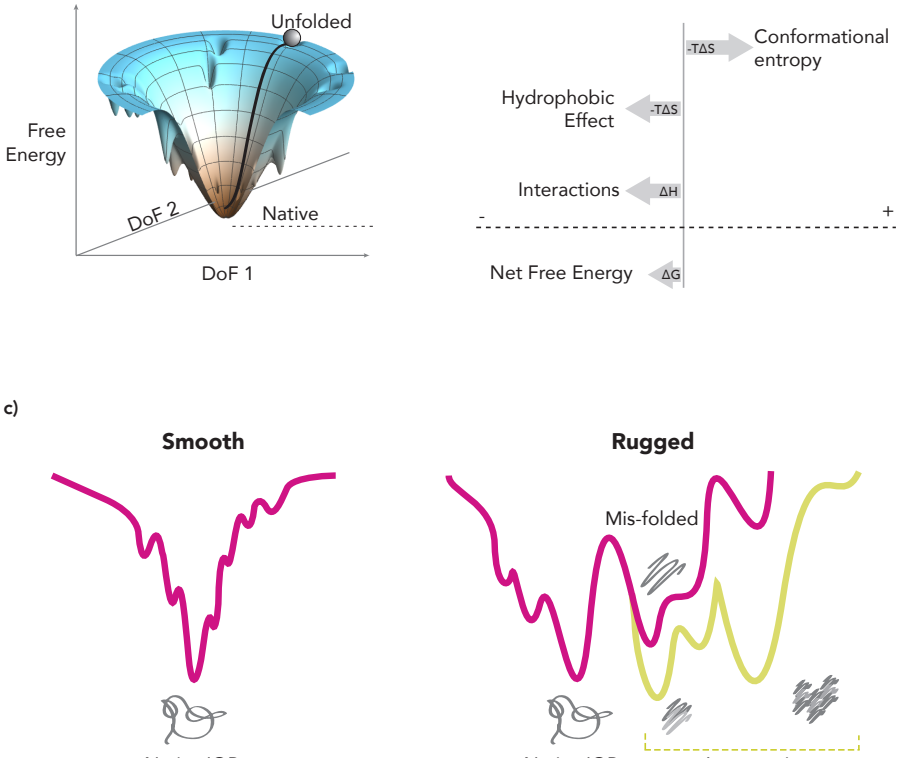


Figure 1.3. The protein energy landscape. a) Example of a smooth 3D energy landscape; b) The energetic balance in the protein folding; c) The cross-section of two schematic energy landscapes. In the rugged landscapes, the overlap between the yellow and magenta surface represent species that can lead to aggregation.

timescale –the typical folding time for an IOP. A quantitative perspective: if we assume that each θ,ω dihedral has three torsions options, then each peptide unit can access up to nine conformers. For an average protein of 100 residues, it would result in 9^{100} or 10^{95} possibilities to explore. Even considering an ultra fast sub-picosecond timescale for each torsional change, it would take an incommensurable time to go through each of these conformations and pick the native one. This simple paradox is a vivid demonstration that IOPs attain their native state by a guided research instead of a random chaotic exploration. Anfinsen proposed a thermodynamic hypothesis to solve the Levinthal paradox: proteins follow the

principle of minimum energy [15]. They inevitably reach the native structure that bears the minimum in energy through distinct intermediates states through a folding pathway, which underline the non-randomness of the process. A new theoretical view about folding emerged later with the application of statistical mechanics models. These models consider each macroscopic state of a protein (folded, unfolded, intermediated) as a distribution or an ensemble of conformations. Folding can follow multiple unpredictable routes passing many intermediates conformations. The new view then replaces the sequential folding pathway with broader concepts: the energy landscape and the folding funnel.

An energy landscape summarizes all the conformations accessible to a protein as a function of its free energy. Each point on the surface associates the energy, quantified in the vertical axis, with the conformation, defined by its degree of freedom (DoF - i.e. the backbone dihedrals) collected in the many lateral axes (for clarity restricted to two in the example in **Figure 1.3**).

The horizontal section of the energy surface at a given depth then identifies the conformational entropy at that energy and proportional to the accessible configuration. High-energy conformations stand on top of hills in the energy landscape; more favorable ones instead rest in valleys. The kinetic of folding then seems, as described by Ken A. Dill, like "a rolling ball on this energy surface, following a trajectory winding through the hills and the valleys": from high-energy starting points the protein can following many routes (changes in conformations) to reduce its energy [16]. For IOPs, the energy landscape is modeled as a funnel in which a protein, while moving energetically downhill, it narrows its accessible conformational space until it falls into the native state (the energy minimum or the spout) [17]

Energy landscapes vary the features to reflect different folding behavior **(Figure 1.3c)**. A minimally frustrated landscape with few unwanted traps or local minima (smooth landscape) leads a protein to quickly fold into its functional structure; instead a more rugged and heterogeneous one allows many competing folding pathways, which should anyway inevitably fall into the unique minimum [16]. In that scenario the folding process can

be difficult and potentially dangerous: an energy landscape riddled with several deep local minima enhance the chances to be trapped in misfolded structures, which in turn could even lead to aggregation [18]. Protein landscapes can deviate from the traditional funnel shape: when they have multiple wide minima, separated by low energy barrier, the protein can quickly fluctuate between several conformations. That's the case for IDPs and molten globules (MGs), as we will see in the following Section 1.9

The shape of the energy landscape ultimately depends on the primary sequence of a protein, while the quantification of the free energy of each conformation, at stable temperature and pressure, can be calculated as Gibbs free energy:

$$G = H - TS$$

where H is the enthalpy of the system, T the fix temperature and S the entropy; this relationship fundamentally links structure (H, enthalpy as interactions), dynamics (S as the arrangements that a system assumes) and function (G). As just discussed, during folding the protein conformational entropy of an IOP decreases unfavorably. To undergo to favorable changes in Gibbs free energy (ΔG), a protein then needs to counterbalance the entropy loss with contribution from the entropy (- TΔS) and enthalpy (ΔH) of the entire system, composed by the protein in the solvent at a specific temperature. In the following section I will review the major contribution to protein stability (**Figure 1.3c**): the formation of internal non-covalent interactions and the hydrophobic effect. However, solvation, the subject of Section 1.11 , is essential in the final energy account. The vertical extent of the energy landscape, in fact, can be modified with changes in the solvent.

1.5. Architects of ordered structures

Here I will briefly introduce the main non-covalent interactions that a protein can form (**Figure 1.4**), while Chapter 2 will describes how to express them quantitatively according to the classical mechanics.

- **Van der Waals** interactions (VdW) are universal forces that occur be-

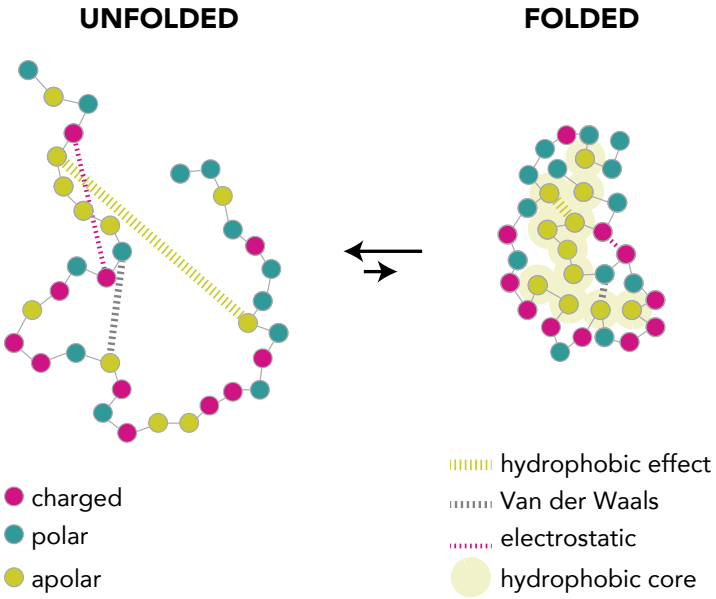


Figure 1.4. The forces in protein folding. Schematic overview of the forces based on the polarity of the residues involved. The protein is represented as a chain of balls.

tween any pair of atoms. They take their name after the Dutch scientists who first described them in 1873 in the attempt to explain the deviation of a real gas from the ideal one. Van der Waals interactions result from a balance between two terms: the attraction that occurs at large distances due to the instantaneous charge fluctuation induced by nearby particles (also known as London or dispersion forces); and the repulsion that occur at smaller distances due to the overlap of electron shells (due to Pauli exclusion principle). The related binding energy is one to three orders of magnitude smaller than the covalent ones.

- **Electrostatic** interactions occur at a longer distance compared to dispersion and are related to the repulsion or attraction between charges or permanent multi-poles. They arise when atoms in a covalent bond share electrons non-uniformly, creating a dipole, which in a complex molecule will arrange in multi-poles. The strength of such interactions

is then dependent on the electronegativity of the atoms involved. A particular type of attractive electrostatic interaction is the hydrogen bond. This relationship invariably occurs when a hydrogen atom H- covalently bound to an electronegative one (often oxygen or nitrogen) approach another electronegative atom. The atom are listed as hydrogen donor (D), the one to which the H- is covalently bound, and hydrogen acceptor (A). The strength of the hydrogen bond is sensitive to atoms orientations and distance. The ideal, strongest hydrogen bond –stronger than any other non-covalent interactions (2-10 kcal/mol)- needs the linearity between D-H:::A and it stands change of ~30 ° in the angle formed with the donor atom.

Besides the protein interactions, another force that favors the compaction of the protein arises from the surrounding environment - often a water solution. In solution, water molecules avoid mixing with oil-like or apolar substances, from there called hydrophobic. At the interface with apolar moieties i.e. hydrocarbons, the water molecules are forced into an ordered network that limits their rotations and translations. To avoid such a high price in entropy, the unfavorable interface between apolar residues and water needs to be minimized. Proteins in water then prefer a compact structure (less exposed surface) with non-polar side chains buried in the protein interior – forming a hydrophobic core. This phenomenon is referred as the **hydrophobic effect** or imprecisely as hydrophobic bonds - even if preferential interactions between hydrophobic residues don't exist as such. As we will see in the following sections, hydrophobic effect is affecting the protein disorder too: a protein rich in charged residues has no need to form a core, neither to fold.

1.6. Optimized patterns in nature

"I didn't have any molecular model with me in Oxford but I took a sheet of paper and sketched the atoms with the bonds between them and then folded the paper to bend one bond at the right angle - what I thought it should be relative to the other - and kept doing this, making a helix, until I could form hydrogen bonds between one turn of the helix and the next turn of the helix, and it only took a few hours of doing that to discover the alpha-helix."

Linus Pauling

As beautifully explained by Linus Pauling, few specific backbone motifs effectively satisfy the constraints imposed by the planarity of the peptide bond and simultaneously optimize the intra-molecular hydrogen bonds [19]. Such intuition guided him, together with Corey and Branson, to correctly model the two most common secondary structures in proteins shown in **Figure 1.5** : *i)* the alpha-helix in which the amino -NH group of a residue (i+4) forms a hydrogen bond with the carbonyl -C=O group four residues earlier (i+4→i); and *ii)* the beta-sheets (parallel or antiparallel depending on the orientation) with hydrogen bonds between facing polypeptide portions. Similarly the most common regions in the Ramachandran plot correspond to dihedral combination typical of these two structures: around (-60, 50) for alpha helices and around (-130, +120) for beta sheets **(Figure 1.6).**

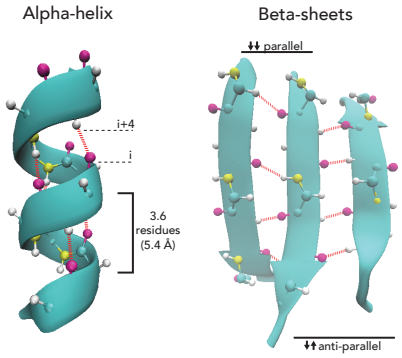


Figure 1.5. The two most common secondary structure elements. Intra molecular hydrogen bonds are marked in red.

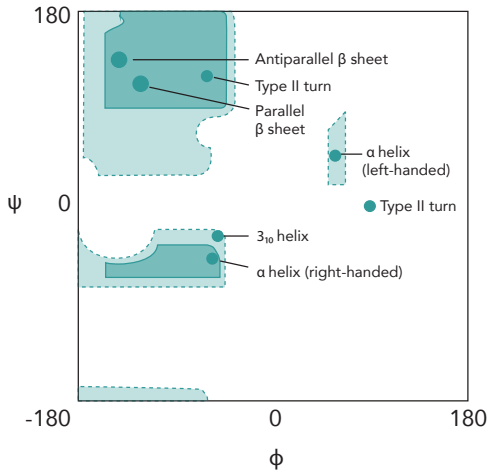


Figure 1.6. Ramachandran plot. The dihedral values found in proteins are shown as blue areas and often correspond to some of the most typical secondary structures of proteins (dots).

While the repeated i+4 →i hydrogen bonding pattern defines the alpha helix, other helical patterns are possible, even if less frequent. For example, they include a i+3 →i pattern typical for the 3^{10} helix - usually of shorter length and tighter than the alpha helix; and a i+5 →i pattern typical of pi-helix. Not every protein residues adopt secondary structure; unordered portions exist even for the most structure IOP (sometimes called turn, bridge, coil…). The unordered regions often lie at the protein surface where they interact with the solvent; indeed they contain mainly hydrophilic residues, in line with the hydrophobic effect described previously.

1.7. Super-motifs, tertiary and quaternary structure

According to the scheme by Linderstrøm-Lang, the secondary structure of the proteins is followed, unsurprisingly, by the tertiary structures. The logic of its name is, however, more than adequate since it refers to the exact topological 3D arrangement in space of the protein motifs. At this level, all the forces discussed previously come into play: the optimization of interactions, the compaction of the structure, the hydrophobic effect, and the apparent steric hindrance. Together they are responsible for the

uniqueness of the native tertiary structure that ultimately depends on the protein sequence and the environmental conditions.

Some folding patterns are redundant in proteins: super-secondary structures are relatively simple i.e. the β-hairpin in **Figure 1.7**; instead domains involve larger portions and behave independently from the rest of the structure. Proteins can also be composed of multiple, independently folded, chains bound together - often in a symmetric way. The network of interactions that keep them together have the same nature of the previously described ones, and it was logically named quaternary structures.

1.8. Protein classification

Beside globular, proteins were traditionally classified also in used to fibrous, and membrane proteins (see **Figure 1.1**). While the latter are the ones inserted into the apolar cellular membrane, the fibrous ones are elongated structures that provide structural support to cells and tissues. However, both differ from the globular ones for their water-insolubility, which complicated their structural characterization. Indeed one of the most used classification schemes, SCOP - Structural Classification of

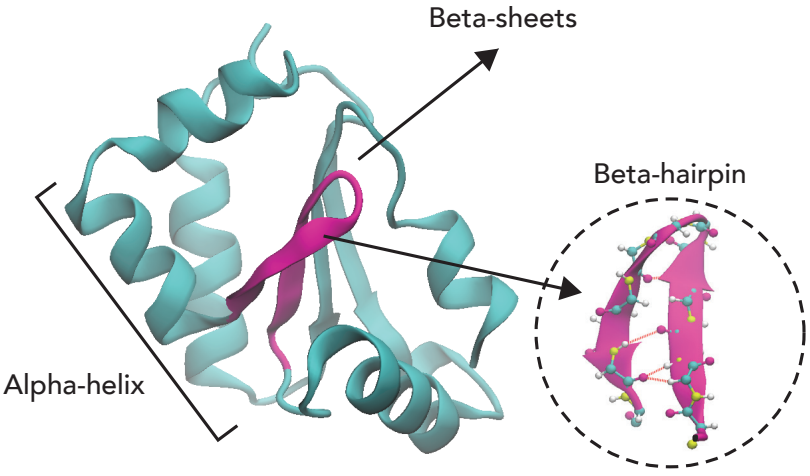


Figure 1.7. Example of tertiary structure for the protein ubiquitin (PDB: 1UBQ). Helices are represented as coils, beta-sheets as arrows.

Proteins (scop.mrc-lmb.cam.ac.uk/scop) - proposed by Murzin and colleagues in 1995, heavily relies on globular proteins [20]. SCOP organized proteins according to their structural and evolutionary relationships in a tree-like hierarchy. The highest level (most general) divide proteins into classes that cluster them into common global topologies of secondary structure. The classes all-α and all-β display mainly one type of secondary structure in their core (see 1OPC and 1CQY in **Figure 1.9**) while α/β and α+β mix them (for example 1KTE).

1.9. Embracing chaos: disorder in proteins

The need for a new classification method arose from the evidence that some proteins don't fit the one-fold description and lack a defined core. These exotic molecules, known as intrinsically disordered proteins (IDPs), are better represented by an ensemble of conformations that dynamically interchange [22]. Their rugged energy landscape allows several minima that represent the variety of conformations among which the protein can fluctuate. A simple description based on the average position of each atom will condense their complexity into a meaningless structure (**Figure 1.8**).

Shreds of evidence about IDPs emerged already at the beginning of the 50s but they have been neglected until recently [23]. Even today they re-

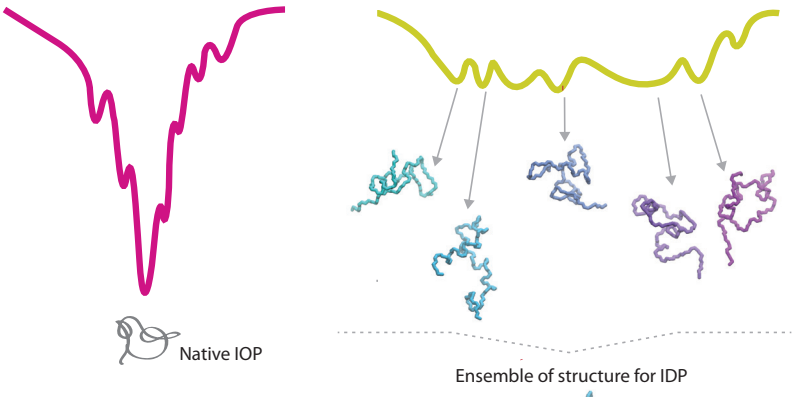
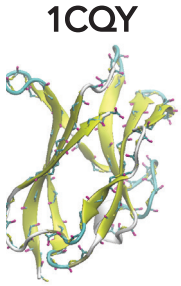
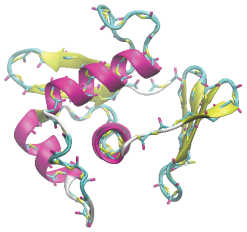


Figure 1.8. Energy landscape for an IOP vs IDP. One-dimensional cross section through two examples of energy landscapes which illustrate the difference between an IOP (one-fold) and a IDP (multiple conformations).



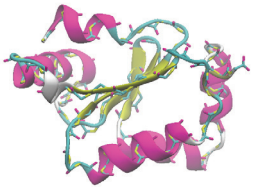
1CQY

Class: **All beta**
Fold: Prealbumin-like
core: sandwich; 7 strands in 2 sheets
Superfamily: Starch-binding domain-like
Family: Starch-binding domain
Protein: beta-amylase
Species: Bacillus cereus



1OPC

Class: **All alpha**
Fold: DNA/RNA-binding 3-helical bundle
core: 3-helices; bundle
Superfamily: C-terminal effector domain of the bipartite response regulators binds to DNA and RNA polymerase; the N-terminal, receiver domain belongs to the CheY family
Family: PhoB-like
Protein: OmpR
Species: Escherichia coli



1KTE

Class: **Alpha and beta (a/b)**
Fold: Thioredoxin fold
Core: 3 layers, a/b/a; mixed beta-sheet of 4 strands
Superfamily: Thioredoxin-like
Family: Thioltransferase
Protein: Glutaredoxin (Grx, thioltransferase)
Species: Pig (Sus scrofa)

Figure 1.9. The SCOP classification: the top-down hierarchy for the three proteins (PDB code on the left) that are selected to represent each SCOP class (see Section 4.2).

main a challenge in biophysics due to the need to collect high-resolution data for an ensemble of conformations and at relevant time scales [24]. However, an explosion of interest started in the 90s thanks to the advances in techniques suited to study dynamic systems. For example small angle X-ray scattering (SAXS) can quantify the collapse of a protein[25]; fluorescence resonance energy transfer (FRET) can single out each conformation unpacking the information in ensemble average [26]; NMR

can describe the relative positioning of atoms and MD simulations can follow their dynamic behavior over several time scales [27]. Nowadays in-cell NMR experiments represent one of the most unique environments to characterize disorder under near physiological conditions [28].

The distinction between IDPs and IOPs already starts at the primary structure: the IDPs sequence is enriched in polar, charged residues and prolines resembling regions at the surface of IOPs: the lack of hydrophobic residues limits the chances to form a compact hydrophobic core and keep the protein into a non-rigid structure. However each IDP shows a specific pattern of flexibility and compaction - from molten globule forms (collapsed and with secondary structure elements) to random coils (extended and purely disorder) [29]. This distinction is encoded, similarly to IOPs, in the protein sequence and depends on the charge distribution: extended conformations tend to have a net charge randomly distributed along the sequence while collapsed conformations possess clusters of opposite charges that allow a degree of compaction. The understanding of these features, along with others, has encouraged the development of algorithms to predict the degree of disorder based solely on the protein sequence i.e. PONDR, the predictor of natural disorder regions [30]. A collection of disordered proteins along with their features can be found in a dedicated database (www.disprot.org [31]). These predictions estimate that the human proteome contains between 35% and 50% of disordered regions – a data that will convince even the most reluctant to embrace them as functional entities inside the cell [32]. IDPs adapt several functions that would be hardly covered by their ordered counterparts: thanks to the conformational interchange they adapt to various situations, bond several partners and sense environmental changes. Indeed IDPs are generally involved in signaling, regulation and control - complementing the functions of IOPs that are dedicated for example to catalysis, binding of small ligands or transport across membrane [32].

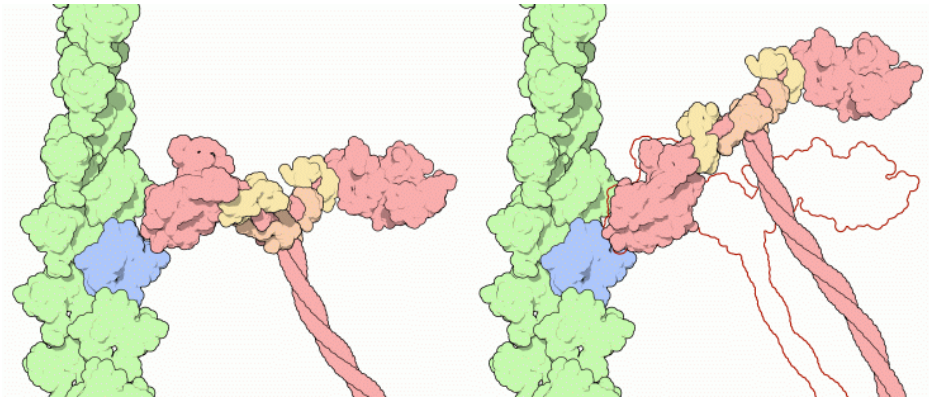


Figure 1.10. Example of a structure in motions: the two conformations of myosin. When ATP is cleaved, the straight "rigor" form, as shown on the right myosin (PDB entry 2mys) adopts a bent, flexed form, like in the structure on the left (PDB entry 1br1). Image credit to PDB.

1.10. Structures in motion

IDPs were the ultimate example to demonstrate the naiveté of the classic structure-function paradigm, bringing to an extreme freedom the conformational interchange. However many biological processes rely on changes in the structure of proteins, which amplitude in time and space are usually proportional [33], [34]. In this thesis, words like "dynamics" and "flexibility" refer to such time-dependent changes in the structure that ultimately are related to the concerted and thermally driven movements of the atoms.

Protein flexibility is related with the exploration of the protein energy landscape [34]. The transition between minima (at the actual condition) provides the protein with an asset of different conformations, selected and designed to exert a mechanism [23]. The transition can take place at local level i.e. a rearrangement of a secondary structure element (ns-microsecond timescale) or at global level i.e. a large domain motions (milli- or seconds) - see **Figure 1.11**. The motor protein myosin, responsible of the muscle contraction, is an example of the latter. Myosin uses ATP-cleavage to adopt a bend and flex conformation; the new form grabs the actin filament and allows the protein to climbs, like an arm along, an actin

filament **(Figure 1.10).**

Less evident changes but equally essential can occur during enzyme catalysis, protein-protein binding or ligand-recognition. To understand how a protein exerts its function then, a single structure provides a limited hint – in a similar way that a static sculpture of a nutcracker fails to explain its utility. A comprehensive view instead should complement structural details with information about the motion of the system. From small amplitude vibrations to large-scale domain rearrangements, modern protein biophysics should not neglect any of them.

Protein dynamics is so ubiquitous and persistent that even X-ray crystallography, which provides a frozen picture of an energetically minimized conformation under the crystallization conditions, contains information about it [35], [36]. Occasionally, large structural rearrangements are captured in two crystals of the same protein i.e. open/close or apo/holo conformations (alone or in complex with a ligand) and before/after ATP cleavage as in the case of myosin (see **Figure 1.10**). More often, the dynamic information is reduced to the amplitude of the atomic fluctuations that attenuate the X-ray scattering signal (B-factor - quantitative described in Section 3.2). Lastly, the mere absence of a region from the X-ray spectra classifies it as highly flexible and possibly disordered.

A more detailed view on protein flexibility at different scale, both temporal and spatial, was possible thanks to the development of Molecular Dynamics (MD) simulations in the 70s [37] and the application of NMR in the 80s [38]; and later confirmed by more recent techniques such as time-resolved crystallography y, FRET or neutral scattering.

Thanks to the large variety of experimental setting and observables, NMR covers a rich time scale of protein movements, from picoseconds (spin-relaxation) up to seconds (amide proton exchange saturation or real-time NMR) [39]. Among the several NMR-observables, residual dipolar coupling (RDC) provide both structural information at the long distance and information on the dynamics slower than a nanosecond. RDCs are obtained in special field-oriented NMR and provide spatially and temporally averaged information about an angle between the external magnetic field and a bond vector in a molecule. Rather than distance

restraints (as NOEs) they can provide orientational constraints about the relative orientation of parts of the molecule, even when they lie far apart. However, two issues complicate the data interpretation: first, the definition of the ever-changing tensor that describes the alignment of a flexible molecule with respect to the laboratory field; and secondly, the decoding of the information packed into ensemble averages, which often requires the support of theoretical models to be transformed into atomic positions [40], [41].

MD simulations instead provide a direct gateway to protein flexibility over several time scales (up to micro-millisecond, **Figure 1.11**) and most importantly without sacrificing the atomic resolution [42]. As crucial method utilized in this work, MD simulation deserved an entirely dedicated chapter (Chapter 2).

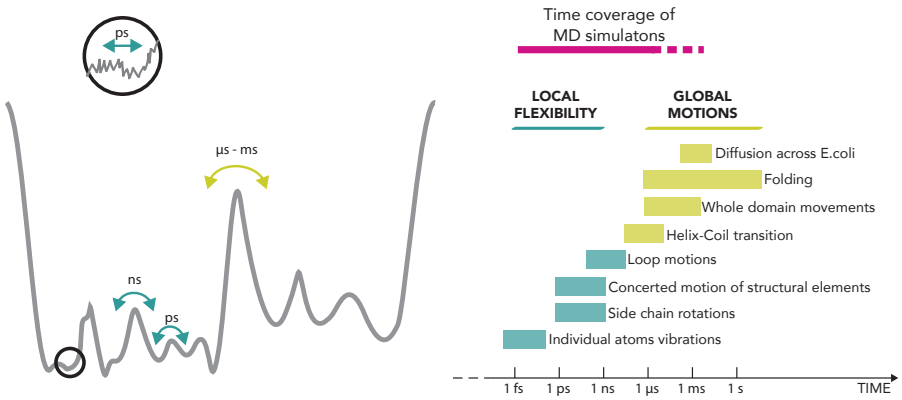


Figure 1.11. Timescale of protein motions. Local (smaller amplitude) and global (large amplitude) motions timescales are shown on the right; on the left a schematic representation of the associated barriers in the energy landscape.

1.11. Solvent and protein stability

As flexible entities, proteins modulate their architectures to accommodate the ever-changing environment. Beside the internal forces reviewed so far, the context in which protein operates equally influence their shape. The majority of proteins exert their function inside the cell: a crowded

aqueous phase with a temperature of ~300K, a pressure of ~1 bar, pH between 6-8 and ionic strength between 100 – 200 mM KCl or NaCl. Laboratory and simulated conditions both mimic the physiological environment with water -awkwardly called aqueous solution- and carefully stabilize all the other factors. Alteration in any of the involved variables creates a new milieu affecting the energetic balance and consequently the structural equilibrium of proteins. This work focuses on one of these: the solvent composition.

Opposite to the "aqueous solution" that contains only water and ions, non-aqueous solvents are still based on water, the portion of which can vary. Water is rarely totally excluded from the protein environment because of its fundamental role as "lubricant of life". Other molecules, named co-solvents, act then as companions for water. At microscopic level the protein solvation consists of *i)* the formation of interactions between the protein and the solvents; and *ii)* the change in the interactions between the solvent molecules due to the cavity created to accommodate the protein. The kind of interactions in the solvation shell, composed of the solvent molecules in the protein surrounding, have the same nature of the ones inside the protein and described in Section 1.5 . Ultimately the protein structure arises from a complex interplay of enthalpic and entropic contributions from protein-protein, protein-solvent, and solvent-solvent interactions. Clearly a small change in these forces can tip the energetic balance towards other structures. Small organic molecules, known as osmolytes, perfectly exemplify the concept. When added as co-solvent to water, they can push the folded/unfolded equilibrium of IOP towards opposite directions: protecting osmolytes favor the native state, whereas denaturing osmolytes the unfolded one. Protecting osmolytes such as trimethylamine N-oxide (TMAO), glycine, glycerol and others are ubiquitous in nature because they contribute to stabilizing proteins against adverse conditions. On the contrary, urea, a denaturing osmolyte found in mammalian kidney, is a crucial reagent employed in protein stability studies.

Indeed, urea and crowding agents, natural or synthetic, are the three co-solvents protagonists of this thesis. Their effect on protein stability

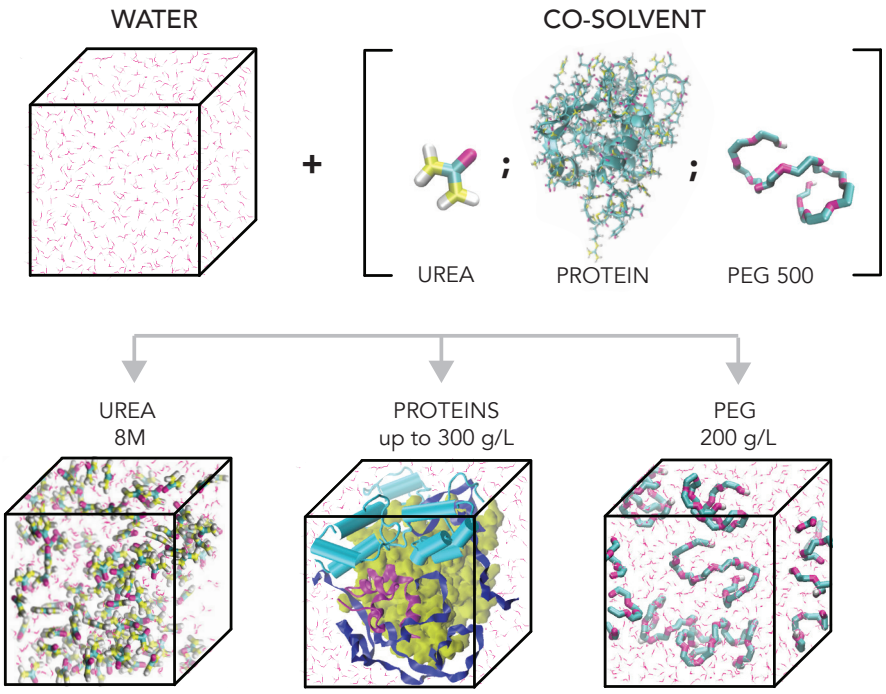


Figure 1.12. The three co-solvents. Each cosolvent is added to the normal water solution to reach the desired concentration. For each of them, the molecular structure and an example of a unit box is shown. The structure of PEG 500 (dodecaethylene glycol) is taken from the PDB: 2XP6.

is addressed in Chapter 4 and 5; while here I will only briefly introduce them (**Figure 1.12**).

Urea is highly soluble in water, generally used in concentration between 8 to 10 M; its high solubility in water reflects its ability to form an extensive network of hydrogen bonds with water. Overall urea molecules integrate well into water structure, showing only a minimal tendency for self-aggregation [43]–[45]. The structural perturbations on the solvent constitution are then excluded as driving force for protein unfolding; instead the direct interactions between urea and the protein appear as better candidates. The latter will be investigated in Chapter 4.

Crowding agents, natural or synthetic macromolecules, are used as co-solvents to fill the void and mimic the crowding effect of the cytoplasm. When macromolecules as co-solvents occupy at least 20-30% of

the volume, they cause several restrictions on the solution, for example in the volume accessible to the protein structure and in the mobility of water [46]–[48].

Synthetic crowding agents, such as poly(ethylene glycol) (PEG), poly-vinylpyrolidone (PVP), dextran and Ficoll, are generally employed thanks to their "inert "character, meaning their inability to interact specifically with proteins, which correctly mimics the excluded-volume. Among them, **PEG**, also known as PEO (polyethylene oxide), is the most controversial. Its structure, a water-soluble chain of ethylene oxide (–CH2–CH2–O–) which adopt an helical elongated shape PEG [49], is however less "inert" than expected and suffer of attractive interactions with proteins [50]. However PEG is still used as a reference agent for macromolecular crowding [51], [52]. **Proteins** instead represent the most common *natural* crowders and they might also exert their influence beyond the general volume-exclusion: a network of nonspecific intermolecular interactions, which alters protein stability with unexpected consequences on protein structure and dynamics. The latter are the main subjects of Chapter 5.

Once defined these environments, we can move on to adress the main objectives of the present thesis.

OBJECTIVES

The main aim of this thesis is to extract rules valid at proteome level about the repercussions of two co-solvents, namely the urea-aqueous solution and a crowded environment, on three major aspects of proteins: structure, dynamics and solvent interactions. To accomplish such aim we employed MD simulation on three system systems strategically selected to addressed a specific objective (**Figure 1.13**):

1. To understand the nature of the **urea-induced unfolded state**. Combining MD simulations and available NMR data we aimed to: *i)* define the unfolded state ensemble of the model protein ubiquitin, *(ii)* understand the energetics stabilizing unfolded structures in urea, and *(iii)* describe the differential nature of the interactions of the fully unfolded proteins with urea and water.

1. To understand the **early stages of urea-induced unfolding**. Performing a vast number of simulations on a representative dataset of folded proteins (representing the major folds of globular proteins) we aimed to: *i)* identify common patterns on the early staged of the unfolding reaction; *ii)* challenge at proteome level the observations related to the solvent/protein interactions derived from the previous project, and *iii)* investigate the kinetic role of urea in triggering protein unfolding.

1. To understand the effect of **effect of crowding** in protein structure, dynamics and interaction properties. Analyzing a variety of crowding conditions on different proteins we aimed to: *i)* understand the general effect of synthetic (polyethyleneglycol; PEG) and natural (proteins) crowded agents in the structure and dynamics of folded proteins, and *ii)* understand the differential effect of crowding in folded, versus unfolded and molten globule proteins.

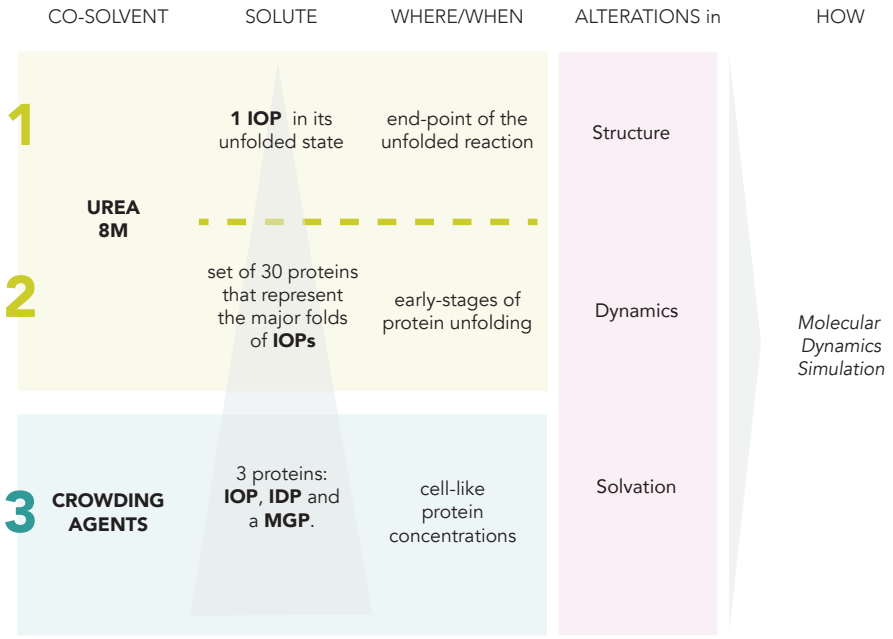


Figure 1.13. Schematic organization of the three projects. Each project is classified according to the co-solvent used, the proteins (solute) under study and the spatio-temporal definition of the system; and have a different degree of generalisation (single protein, many folds, many protein types). In all of them the analyses are anchored on three aspects of proteins (structure, dynamic and interactions with the solvent), adressed by means of MD simulations.

Bibliography Chapter 1

[1] A. V. Finkelstein and O. B. Ptitsyn, Protein physics: a course of lectures. Amsterdam ; Boston: Academic Press, 2002.

[2] A. Fersht, Structure and mechanism in protein science: a guide to enzyme catalysis and protein folding. New York: Freeman, 1999.

[3] A. O. W. Stretton, "The First Sequence: Fred Sanger and Insulin," Genetics, vol. 162, no. 2, pp. 527–532, Oct. 2002.

[4] K. U. Linderstrom-Lang, Proteins and enzymes: Lane medical lectures, 1951. Stanford University Press; Oxford University Press, 1952.

[5] G. N. Ramachandran, C. Ramakrishnan, and V. Sasisekharan, "Stereochemistry of polypeptide chain configurations," J. Mol. Biol., vol. 7, no. 1, pp. 95–99, Jul. 1963.

[6] G. D. Rose, P. J. Fleming, J. R. Banavar, and A. Maritan, "A backbone-based theory of protein folding," Proc. Natl. Acad. Sci. U. S. A., vol. 103, no. 45, pp. 16623–16633, Nov. 2006.

[7] A. R. Fersht, "Max Ferdinand Perutz OM FRS," Nat. Struct. Mol. Biol., vol. 9, no. 4, pp. 245–246, Apr. 2002.

[8] A. G. Palmer and D. J. Patel, "Kurt Wüthrich and NMR of Biological Macromolecules," Structure, vol. 10, no. 12, pp. 1603–1604, Dec. 2002.

[9] S. Jonic and C. Vénien-Bryan, "Protein structure determination by electron cryo-microscopy," Curr. Opin. Pharmacol., vol. 9, no. 5, pp. 636–642, Oct. 2009.

[10] H.-T. Chou, J. E. Evans, and H. Stahlberg, "Electron crystallography of membrane proteins," Methods Mol. Biol. Clifton NJ, vol. 369, pp. 331–343, 2007.

[11] H. D. T. Mertens and D. I. Svergun, "Structural characterization of proteins and complexes using small-angle X-ray solution scattering," J. Struct. Biol., vol. 172, no. 1, pp. 128–141, Oct. 2010.

[12] Z. Xiang, "Advances in Homology Protein Structure Modeling," Curr. Protein Pept. Sci., vol. 7, no. 3, pp. 217–227, Jun. 2006.

[13] C. B. Anfinsen, E. Haber, M. Sela, and F. H. White, "The Kinetics Of Formation Of Native Ribonuclease During Oxidation Of The Reduced Polypeptide Chain," Proc. Natl. Acad. Sci. U. S. A., vol. 47, no. 9, pp. 1309–1314, Sep. 1961.

[14] C. Levinthal, "How to Fold Graciously," presented at the Mossbauer Spectroscopy in Biological Systems: Proceedings of a meeting held at Allerton House, Monticello, Illinois, 1969, pp. 22–24.

[15] C. B. Anfinsen, "Principles that govern the folding of protein chains," Science, vol. 181, no. 4096, pp. 223–230, Jul. 1973.

[16] K. A. Dill and H. S. Chan, "From Levinthal to pathways to funnels," Nat. Struct. Mol. Biol., vol. 4, no. 1, pp. 10–19, Jan. 1997.

[17] J. D. Bryngelson, J. N. Onuchic, N. D. Socci, and P. G. Wolynes, "Funnels, pathways, and the energy landscape of protein folding: a synthesis," Proteins, vol. 21, no. 3, pp. 167–195, Mar. 1995.

[18] A. Gershenson, L. M. Gierasch, A. Pastore, and S. E. Radford, "Energy landscapes of functional proteins are inherently risky," Nat. Chem. Biol., vol. 10, no. 11, pp. 884–891, Nov. 2014.

[19] L. Pauling, R. B. Corey, and H. R. Branson, "The structure of proteins: Two hydrogen-bonded helical configurations of the polypeptide chain," Proc. Natl. Acad. Sci., vol. 37, no. 4, pp. 205–211, Apr. 1951.

[20] A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia, "SCOP: a structural classification of proteins database for the investigation of sequences and structures," J. Mol. Biol., vol. 247, no. 4, pp. 536–540, Apr. 1995.

[21] A. Andreeva, D. Howorth, C. Chothia, E. Kulesha, and A. G. Murzin, "SCOP2 prototype: a new approach to protein structure mining," Nucleic Acids Res., vol. 42, no. D1, pp. D310–D314, Jan. 2014.

[22] P. E. Wright and H. J. Dyson, "Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm," J. Mol. Biol., vol. 293, no. 2, pp. 321–331, Oct. 1999.

[23] K. Teilum, J. G. Olsen, and B. B. Kragelund, "Functional aspects of protein flexibility," Cell. Mol. Life Sci. CMLS, vol. 66, no. 14, pp. 2231–2247, Jul. 2009.

[24] V. N. Uversky, "A decade and a half of protein intrinsic disorder: biology still waits for physics.," Protein Sci. Publ. Protein Soc., vol. 22, pp. 693–724, 2013.

[25] P. Bernadó and D. I. Svergun, "Analysis of intrinsically disordered proteins by small-angle X-ray scattering," Methods Mol. Biol. Clifton NJ, vol. 896, pp. 107–122, 2012.

[26] B. Schuler, S. Müller-Späth, A. Soranno, and D. Nettels, "Application of Confocal Single-Molecule FRET to Intrinsically Disordered Proteins BT - Intrinsically Disordered Protein Analysis," in Intrinsically Disordered Protein Analysis, New York, NY: Springer New York, 2012, pp. 21–45.

[27] M. R. Jensen, R. W. Ruigrok, and M. Blackledge, "Describing intrinsically disordered proteins at atomic resolution by NMR," Curr. Opin. Struct. Biol., vol. 23, no. 3, pp. 426–435, Jun. 2013.

[28] F.-X. Theillet, A. Binolfi, T. Frembgen-Kesner, K. Hingorani, M. Sarkar, C. Kyne, C. Li, P. B. Crowley, L. Gierasch, G. J. Pielak, A. H. Elcock, A. Gershenson, and P. Selenko, "Physicochemical Properties of Cells and Their Effects on Intrinsically Disordered Proteins (IDPs)," Chem. Rev., vol. 114, no. 13, pp. 6661–6714, Jul. 2014.

[29] R. K. Das, K. M. Ruff, and R. V. Pappu, "Relating sequence encoded information to form and function of intrinsically disordered proteins," Curr. Opin. Struct. Biol., vol. 32, pp. 102–112, Jun. 2015.

[30] P. Romero, Z. Obradovic, X. Li, E. C. Garner, C. J. Brown, and A. K. Dunker,

"Sequence complexity of disordered protein," Proteins, vol. 42, no. 1, pp. 38–48, Jan. 2001.

[31] M. Sickmeier, J. A. Hamilton, T. LeGall, V. Vacic, M. S. Cortese, A. Tantos, B. Szabo, P. Tompa, J. Chen, V. N. Uversky, Z. Obradovic, and A. K. Dunker, "DisProt: the Database of Disordered Proteins," Nucleic Acids Res., vol. 35, no. Database issue, pp. D786–793, Jan. 2007.

[32] C. J. Oldfield and A. K. Dunker, "Intrinsically Disordered Proteins and Intrinsically Disordered Protein Regions," Annu. Rev. Biochem., vol. 83, no. 1, pp. 553–584, 2014.

[33] M. Orozco, "A theoretical view of protein dynamics," Chem. Soc. Rev., vol. 43, no. 14, pp. 5051–5066, Jun. 2014.

[34] K. Henzler-Wildman and D. Kern, "Dynamic personalities of proteins," Nature, vol. 450, no. 7172, pp. 964–972, Dec. 2007.

[35] A. Kuzmanic, N. S. Pannu, and B. Zagrovic, "X-ray refinement significantly underestimates the level of microscopic heterogeneity in biomolecular crystals," Nat. Commun., vol. 5, p. 3220, Feb. 2014.

[36] R. B. Fenwick, H. van den Bedem, J. S. Fraser, and P. E. Wright, "Integrated description of protein dynamics from room-temperature X-ray crystallography and NMR," Proc. Natl. Acad. Sci., vol. 111, no. 4, pp. E445–E454, Jan. 2014.

[37] J. A. McCammon, B. R. Gelin, and M. Karplus, "Dynamics of folded proteins," Nature, vol. 267, no. 5612, pp. 585–590, Jun. 1977.

[38] K. Wuthrich, "Protein structure determination in solution by nuclear magnetic resonance spectroscopy," Science, vol. 243, no. 4887, pp. 45–50, Jan. 1989.

[39] I. R. Kleckner and M. P. Foster, "An introduction to NMR-based approaches for measuring protein dynamics," Biochim. Biophys. Acta, vol. 1814, no. 8, pp. 942–968, Aug. 2011.

[40] C. Camilloni and M. Vendruscolo, "A Tensor-Free Method for the Structural and Dynamical Refinement of Proteins using Residual Dipolar Couplings," J. Phys. Chem. B, vol. 119, no. 3, pp. 653–661, Jan. 2015.

[41] S. Esteban-Martín, R. B. Fenwick, and X. Salvatella, "Refinement of Ensembles Describing Unstructured Proteins Using NMR Residual Dipolar Couplings," J. Am. Chem. Soc., vol. 132, no. 13, pp. 4626–4632, Apr. 2010.

[42] D. E. Shaw, P. Maragakis, K. Lindorff-Larsen, S. Piana, R. O. Dror, M. P. Eastwood, J. A. Bank, J. M. Jumper, J. K. Salmon, Y. Shan, and W. Wriggers, "Atomic-Level Characterization of the Structural Dynamics of Proteins," Science, vol. 330, no. 6002, pp. 341–346, Oct. 2010.

[43] M. C. Stumpe and H. Grubmüller, "Aqueous urea solutions: structure, energetics, and urea aggregation," J. Phys. Chem. B, vol. 111, no. 22, pp. 6220–6228, Jun. 2007.

[44] A. K. Soper, E. W. Castner, and A. Luzar, "Impact of urea on water structure: a clue to its properties as a denaturant?," Biophys. Chem., vol. 105, no. 2–3, pp. 649–666, Sep. 2003.

[45] F. Vanzi, B. Madan, and K. Sharp, "Effect of the Protein Denaturants Urea and Guanidinium on Water Structure: A Structural and Thermodynamic Study," J. Am. Chem. Soc., vol. 120, no. 41, pp. 10748–10753, Oct. 1998.

[46] R. Harada, Y. Sugita, and M. Feig, "Protein Crowding Affects Hydration Structure and Dynamics," J. Am. Chem. Soc., vol. 134, no. 10, pp. 4842–4849, Mar. 2012.

[47] K. Luby-Phelps, "Cytoarchitecture and physical properties of cytoplasm: volume, viscosity, diffusion, intracellular surface area," Int. Rev. Cytol., vol. 192, pp. 189–221, 2000.

[48] K. Luby-Phelps, "The physical chemistry of cytoplasm and its influence on cell function: an update," Mol. Biol. Cell, vol. 24, no. 17, pp. 2593–2596, 2013.

[49] S. A. Oelmeier, F. Dismer, and J. Hubbuch, "Molecular dynamics simulations on aqueous two-phase systems - Single PEG-molecules in solution," BMC Biophys., vol. 5, no. 1, p. 14, Aug. 2012.

[50] J. Tyrrell, K. M. Weeks, and G. J. Pielak, "Challenge of Mimicking the Influences of the Cellular Environment on RNA Structure by PEG-Induced Macromolecular Crowding," Biochemistry (Mosc.), Oct. 2015.

[51] A. J. Boersma, I. S. Zuhorn, and B. Poolman, "A sensor for quantification of macromolecular crowding in living cells," Nat. Methods, vol. 12, no. 3, pp. 227–229, Mar. 2015.

[52] L. A. Ferreira, P. P. Madeira, L. Breydo, C. Reichardt, V. N. Uversky, and B. Y. Zaslavsky, "Role of solvent properties of aqueous media in macromolecular crowding effects," J. Biomol. Struct. Dyn., vol. 0, no. 0, pp. 1–12, Jan. 2015.

"Art is the lie that helps tell the truth"
Pablo Picasso

CHAPTER 2

Theory and modeling: MD simulations

When people hazard a question about my research field I provide a large list of synonyms: computational structural biology, in silico biophysics, molecular modeling… usually the list goes on until I recognize a receptive (or scared) expression in my interlocutor's face. It's difficult to contain a discipline in few keywords but here I will try to briefly review them and correctly place the present work in its field. This chapter is, then, intended as a broad and conceptual introduction to the theoretical framework of the main method used here: molecular modeling and, more specifically, MD simulations. Further technical details regarding the specifications of each project can be found in the related method section of each publication.

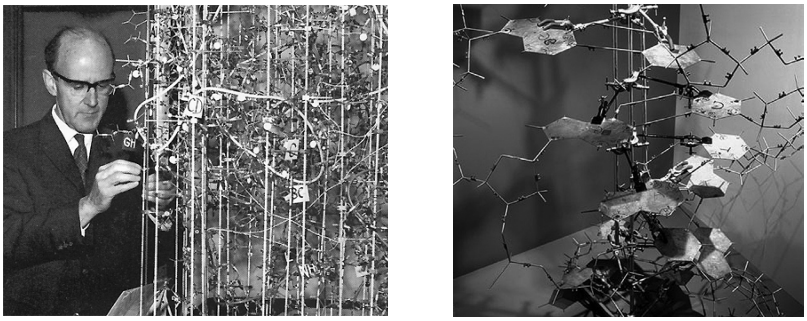


Figure 2.1. Wire-models for macromolecules. From the left: John Kendrew and its wire model, hand-built to fit the electron density for the protein myoglobine (153 aminoacids and over 2600 atoms); the double helices model for the DNA molecule built by James Watson and Francis Crick.

2.1. Molecular modeling

"Building a model of a small protein is like doing a three-dimensional jigsaw puzzle with thousand pieces. It is painful, slow work but at the end you really know the molecule. You also so want to computerize it!"

Michael Levitt, Nobel prize Lecture 2013

Molecular modeling is the study of the behavior of molecules through model building that rationalizes the properties of a biological system through the laws of physics (biophysics). Despite a model remains a simplified reproduction of the complexity of life, simplifications and approximations are needed to rationalize quantities, discern patterns and add insights otherwise difficult to observe. In other words it helps to grasp and make sense of the complexity. Nowadays we rarely use wire hand-built model - as James Watson and Francis Crick did for the double helices or John Kendrew for the myoglobine (**Figure 2.1**) – instead we handle larger set of variables exploiting the capabilities of computers (from here the expression in silico). In general computational models are appropriately developed for a variety of issues, from the motions of galaxies to economical modeling and many more. In structural biophysics the obvious focus is the structure and, as already stressed out in Chapter 1, the dynamics of macromolecules.

Molecular models differ mainly in resolution: quantum mechanics provide the most accurate results however it is limited to short time scale and to systems with few atoms. For macromolecules such as DNA or proteins methods based on classical mechanics are more suitable; they can handle millions of atoms and reach longer timescale (micro/millisecond) relevant in biology. Among those techniques, MD simulation is probably the most widely used; it provides the time-resolved dynamic of the system at such high resolution that once was defined the "computational microscope" for molecular biology [1]. In 2013 the Nobel Prize in Chemistry to A. Warshel, M. Karplus, and M. Levitt praised their pioneering work on "multiscale models for chemical systems", which started the rise of MD simulations.

2.2. MD simulation as a computational microscope

MD simulations apply first principle physics to model the motions of atoms in complex systems i.e. macromolecule immerse in a specific environment. The first MD simulation of a protein (BPTI), published in 1977, covered a simulation time of 8.8 picoseconds (ps) [2]. Nowadays the exponential progresses in computational power gave access to larger timescale - several orders of magnitude longer (micro-milli second)- but the basic principles of those first simulations remain fully effective. This chapter is a overview of such principles (**Figure 2.2**): I will review the assumptions that allows treating atoms as particles under the laws of classical physics and the forces acting on these particles that guide their evolution in time. The expression of these forces together with a set of parameters are the tool-kit to run MD simulations and they come nicely packed inside the so-called force fields. Finally I will very briefly describe the logic (aka the algorithm) used to sample among conformation.

2.3. From QM to MM: principles of molecular dynamics

According to quantum mechanics, the only accurate description of molecular behavior at sub-atomic level implies to resolution of the time-de-

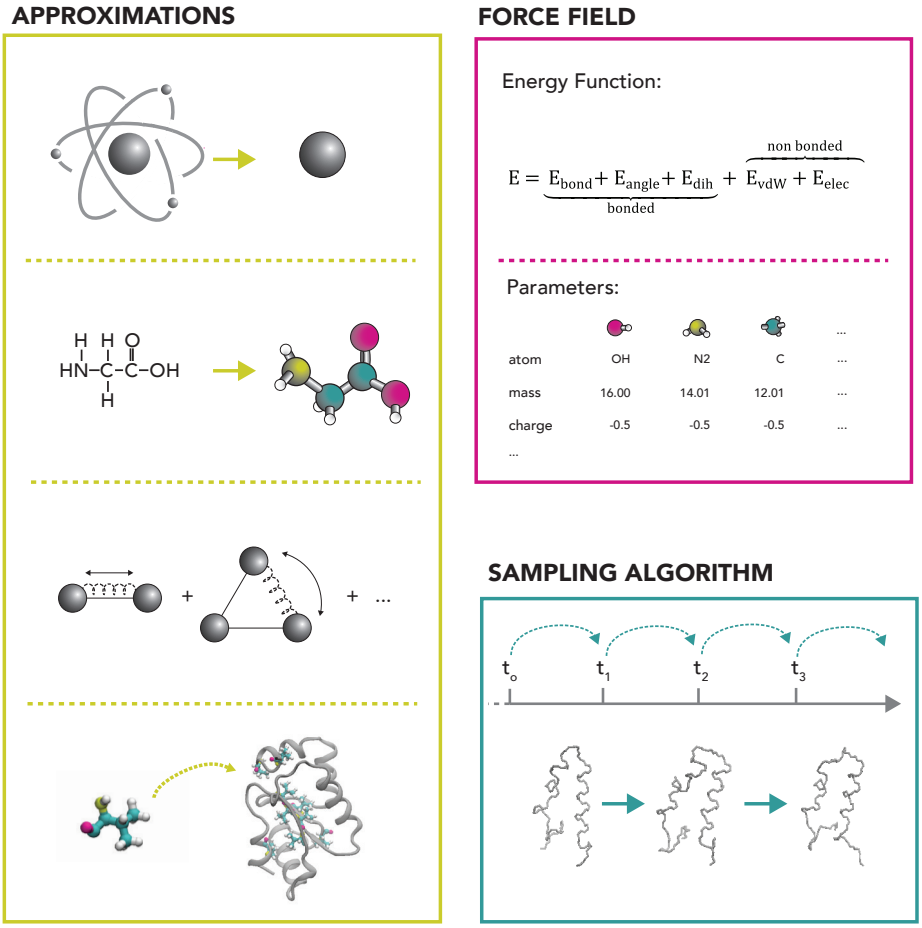


Figure 2.2. Principles of MD simulations. MD simulations rely on three pillars: the approximations used to described molecules (in the yellow box from top to bottom: Born-Oppenheimer, classical, pair-wise additivity and transferability approximation) - see Section 2.3; a force-field (box in magenta) - see Section 2.5; and an algorithm to calculate the movements - see Section 2.7.

pendent Schrödinger equation. With apparent simplicity, this equation relates the structures of electrons and nuclei to the 3D structure, energies and many other associated properties of molecules. In practice the direct solution of Schrödinger equation for macromolecules is so expensive that it is simply unfeasible. Luckily several approximations can drastically reduce the complexity of the energy models and the related computational effort:

1. The **Born-Oppenheimer approximation** separates the electronic and nuclear degrees of freedom and assumes that electrons follow instantaneously the nuclear motion. This is generally a good approximation because the nuclei—much heavier than the electrons—are typically fixed on the timescale of electronic vibration

1. Once atoms are simplified as nuclei, the molecule can be regarded as a classical mechanical body formed by masses centered at the nuclei (atoms) connected by springs (bonds). In the **classical approximation** then the time-dependent Schrodinger equation, which dominates quantum mechanics, is replaced by the Newton's equations of motion, typical of classical mechanics.

1. The molecule, as a mechanical body, stretches, bends, and rotates about those bonds in response to inter and intra molecular forces. According to the **pair-wise additivity approximation**, these forces calculated for each pair of atoms in the system can be summed up with an energy function that provides the energy of the system.

1. The **transferability approximatio**n implies that the energy function developed on a small set of molecules applies to a wider range of molecules with similar chemical groups. If the energy parameters are not dependent on the local environment, a relatively small number of atom types derived from small molecules should be enough to describe whatever macromolecule.

In summary, MD simulations provide a reasonable compromise between accuracy and computational efficiency and their correctness depends on

the validity of these assumptions.

2.4. The potential energy function

Once agreed that a molecule can be treated as a mechanical body, the forces acting on every particle of the system can be expressed through computationally manageable functions (**Figure 2.3**). Forces can be classified as:

1. Short-range between bonded atoms, in particular

- **vibrations** of the bond lenght r (two adjacent atoms) and the bond angle α (two adjacent bonds) that account for small-scale deviations; bonds and angles are considered as an elastic or spring body described by an harmonic potentials based on the Hooke's equation:

$$E_{bond} = \sum_{\text{all bonds}} k_r(r - r_0)^2$$

$$E_{angle} = \sum_{\text{all angles}} k_\alpha(\alpha - \alpha_0)^2$$

According to this equations angles and bonds, when displaced from the equilibrium values (r_0 and $\alpha_0$contained in the forcefield parametrization), experience a restoring force proportional to the displacement and to a constant factor k characteristic for the stiffness of each bond or angle, depending on the atoms involved.

- **torsion** of each dihedral ω, formed by two bonded atoms and the two atoms adjacent to them. Dihedral terms cannot be described by a harmonic term because of their periodicity; instead they are expressed by a trigonometric term; for example:

$$E_{dih} = \sum_{\text{all dihedrals}} \frac{V_{tor}}{2}(1 - \cos(n\omega - \gamma))$$

where the parameters depending in the atoms involved are the torsional barrier V_{tor}, the periodicity n and the phase angle γ.

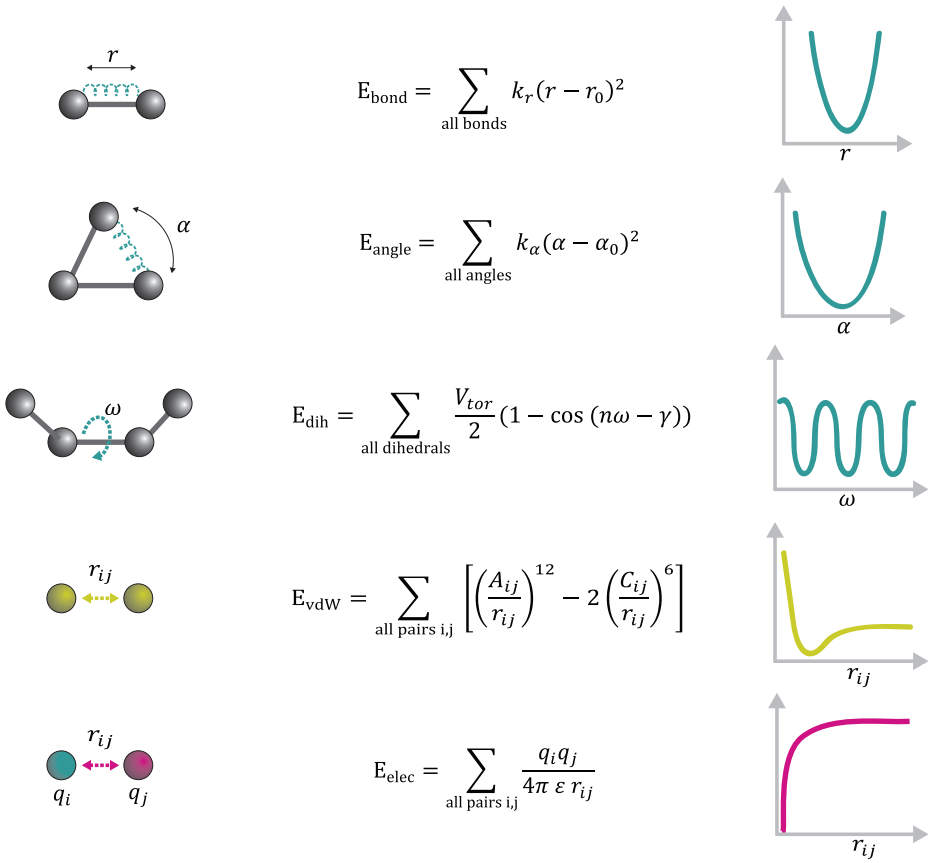


Figure 2.3. Interactions energies in MD simulation. The energy functions (equations in the middle, form on the right) used to describe the interactions between atoms, schematically illustrated on the left. From top to bottom: bond-stretching, bond-angle vibrations, dihedral torsions, Van der Waals and Coulomb interactions.

2. Long-range between non-bonded atoms:

- **Van der Waals interactions** between all pairs of atoms and expressed by a two-terms potential [12-6] to take into account: i) the attraction that arise at great distance when electronic clouds don't overlap; it is expressed by the London's attraction term that weakens with a factor of $(1/r)^6$ where r is the distance between the centers of the two atoms involved; ii) the repulsion between orbitals that bear

already a pair of electrons when their electron clouds overlap at lower distance. The repulsive term counterbalanced the attractive term and has been approximated by Lennard–Jones using a proportion to $(1/r)^m$ where m=12 mainly for computational convenience.

$$E_{vdW} = \sum_{\text{all pairs i,j}} \left[\left(\frac{A_{ij}}{r_{ij}} \right)^{12} - 2 \left(\frac{C_{ij}}{r_{ij}} \right)^{6} \right]$$

Since the overall interaction fall off quickly with the distance they are generally evaluated only for nearby pairs of atoms (within a certain cutoff).

- **Electrostatic interactions** between two particles i, j with partial charges q_i and q_j are modeled by Coulomb's law:

$$E_{elec} = \sum_{\text{all pairs i,j}} \frac{q_i q_j}{4\pi \varepsilon r_{ij}}$$

where ε is the dielectric constant of the medium and r_{ij} the interatomic distance. Columbic attraction also describe hydrogen bonds as attractions between the partial negative charge of the A atom and the partial positive charge of H.

Among all the interactions the electrostatic ones require the most time consuming calculation: unlike the van der Waals one, the Columbic term decay slower with the distance and therefore includes a large subset of atoms. To handle them efficiently a common method used is the **particle mesh Ewald (PME)** that separates the electrostatic energy into two terms: a short-range potential calculated in the real space and a long-ranged potential evaluated on a discrete interlaced Fourier space. Both terms converge quickly when evaluated in their respective spaces and cutoffs can be used safely without sacrificing accuracy. However the methods requires the periodic symmetry of the system. In MD simulations this can be deliberately attained by **periodic boundary conditions**: the entire system (biomolecules and solvent molecules) is placed inside a unit cell and infinitely replicated, creating a repeating pattern similar to the ones formed by azulejos tiles (**Figure 2.4**).

In each of the replicated units, the periodic image of each atom will move exactly the same way as the original one in the central box. When

an atom leaves the central box, its image will enter through the opposite face – thus the central unit still contain all the information to track the entire system. The system, then, can only adopt shape that full fill the surrounding space with translational copies of itself. This unit cells can be a cuboid, a dodecahedron, or a truncated octahedron and are usually large enough to avoid that the protein atom are too close to the box boundary (minimum distance 1.5 nm).

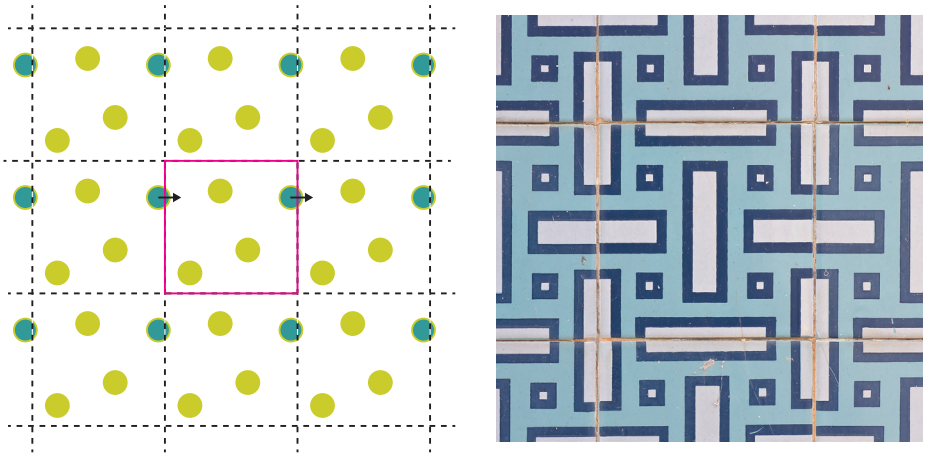


Figure 2.4. The periodic boundary conditions. Schematic illustration in two dimensions of the periodic unit (in magenta) which contains all the information about the system. On the right an example of a pattern from *azulejos* tiles.

2.5. Force-fields: the wikiHow of MD simulations

Force-fields provide the simulation software with two main tools:

- a *recipe* to evaluate the energies expressed by the **energy function** in which the terms described so far are summed together to give the potential energy of the system:

$$E = \underbrace{E_{bond} + E_{angle} + E_{dih}}_{\text{bonded}} + \overbrace{E_{vdW} + E_{elec}}^{\text{non bonded}}$$

- the *ingredients* that need to be mix to reproduce a specific system – expressed by a set of **parameters** that model each atom type in the

system. As we have seen, each term of the energy function contains several variables that depend on the atoms involved; for example hydrogen H and oxygen O surely have a different partial charge q or stiffness of bond stretching k . Current force-fields further distinguish subcategories (atom types) that depend on the molecular environment (e.g., aromatic carbon in a nitrogenous base) and/or hybridization state (e.g., sp2 or sp3).

No universal force-field for proteins exists and the used parameters are at the service of that energy function and shouldn't be regarded as an intrinsic property of proteins. However all the energy functions in the current forcefields are almost the same as the one conveniently devised in the seminal work of Lifson and coworkers [3] . The parameters instead are constantly improved to reproduce more precise properties of molecules, leading to several updated versions of the same forcefield. For example the first force fields were parameterized to reproduce mainly structural properties and vibrational spectra of sample molecules and they were tested primary in gas-phase simulations and with quantum mechanics calculations. Later parameterization instead included liquid phase properties (densities, heats of vaporization) and ab-initio quantum mechanical calculations. An extensive review on the topic can be found in [4], while here I will briefly go through the versions of the force fields used in the present work:

- **AMBER** (Assisted Model Building with Energy Refinement) - version parm99 [5] , parm99SB [6] with improved torsional angles for the backbone and parm99SB-ILDN with improved side-chain torsion potential [7].
- **OPLS-AA** (Optimized Potentials for Liquid Simulations – All Atom) [8]; the only force field born with the specific purpose of correctly fit experimental properties of liquids;
- **CHARMM** (Chemistry at HARvard Macromolecular Mechanics) - versions CHARMM22 [9] and its dihedral potential corrected variant CHARMM22/CMAP [10]. CHARMM force field still enforces neutral charge groups (of adjacent atoms must have zero net charge).

In the past, many efforts have been done by our group to accurately evaluate the performance of available force fields through systematic analyses of a large data set of simulations (MoDEL)[11] . The study concluded that current force fields yield to reasonable consensus when used to simulate native structures and they correctly reproduced experimental data available at that time [12]. Recent studies on long-timescale simulations further confirmed such conclusion [13]. Despite the encouraging results, improvements in current force fields are expected, especially concerning:

- The bias towards the native structure of IOPs – that **misrepresents the disorder** of IDPs [14], [15]. The reason for this bias is quite simple: historically parameters were tuned employing data from IOPs and consequently current force-fields tend to hyper-stabilize secondary structures forms (i.e helices) and collapsed structure. Improved solvent models has been recently employed to recover the correct compactness of IDPs [16], [17].
- The **lack of polarization** to correctly redistribute the electronic cloud around each atom in response of changes in environment, an essential feature to model bond breakings or metal binding. Current force-fields instead assume a fixed partial charge q assigned to each atom type [18].
- The **lack of other quantum effects** (charge transfer among others) that imply a change in the intrinsic properties of atoms, even of the topology of the system depending on the environment.

Bearing in mind these caveats is fundamental to discern the relevant conclusion that can be extracted from MD simulation and discard others. In general comparative studies that focus on qualitative trends; employ several force fields or complement experimental data can overcome these issues.

2.6. Protein solvation

Beside the obvious importance for protein parameters, other fundamental entities need to be treated explicitly – solvent and ions. Two possibilities exist to treat solvent and ions in MD simulations: *i)* implicit representations using models derived from continuum electrostatic theory and *ii)* explicit representations where solvent and ions are represented at the same level of detail than protein atoms. I will limit the explanation here to explicit solvent models, those that have been used in this thesis.

Efforts have been made to accurately parametrize **water**, the most common and important solvent. Two of the most common model have been introduced in the early 80s (SPC [19] and TIP3P [20]). Both models posses three interaction points corresponding to the three atoms of the water molecule, very similar fixes charges and the same OH equilibrium bond length but they differ in the Lennard-Jones parameters and equilibrium HOH angle. In my studies I employed the model TIP3P (transferable intermolecular potential 3P), compatible and included in all the used force fields.

Parameters for **non-standard solvents**, such as urea or polyethylene glycol (PEG), are unlikely to be part of the standard package of forcefields; however scientific literature offers several references to correctly parametrize those molecules, the selected parameters were then adapted to the force-field format (**Figure 2.5**).

Urea parameters: several models for urea molecule exist; in this work we selected three models, one for each forcefield family that we have

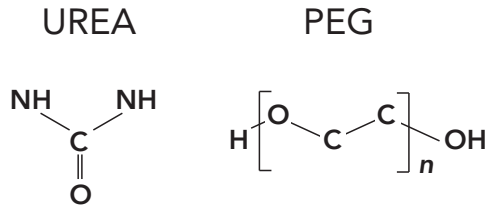


Figure 2.5. Urea and PEG chemical structures. In the structure,the atoms and bonds whose parameters are necessary to correctly described the two molecules in force-fields.

used. In OPLS we implemented the parameters developed by Smith and coworkers using the Kirkwood-Buff (KB) theory for solution [21]; in PARM99 the Caflish -Karplus parameters derived in analogy with the amide of asparagine and glutamine sidechains [22]; and in CHARMM the parameters developed by Caballero-Herrera & Nilsson, again in analogy of asparagine sidechain and fitted to the non empirical intermolecular potentials for urea-water systems [23], [24].

The three models differ mainly on the charged distribution and consequently on the dipole moment, as shown in **Table 2.1.**

PEG parameters: PEG is a water-soluble straight chain polymer composed of repeating ethylene oxide monomeric units (Ch2-Ch2-O). We

Partial Charge	PARM99	CHARMM	OPLS
O	-0.510	-0.502	-0.390
C	0.510	0.142	0.142
N	-0.620	-0.569	-0.542
H	0.310	0.333 / 0.416	0.333
Dipole Moment	4.85 D	5.27 D	4.65 D

Table 2.1 Urea models. Partial charges and related dipole moments for the three urea models used in this work.

chose to use the Fischer et al. TraPPE-UA parameters since they were already successfully tested in the PARM99 forcefield and GROMACS [25], [26]. We treated the beginning and end hydrogen atoms like methyl/methylene groups, with a +0.25 electric charge and the equivalent non-bonded parameters for a hydroxyl hydrogen from the TraPPE-UA forcefield (Transferable Potentials for Phase Equilibria)[27]. PEG parameters are often derived from those of dimethoxyethane (DME), since both molecules share a common backbone of C – O – C and C – C – O

bonds. The DME parameters were optimized to reproduce the experimentally observed backbone torsional conformations in aqueous solution.

As we have seen before, the periodic boundary condition imposed a unit cell of the system with a box shape, often a cuboid or a dodecahedron. The solvation consists on placing the protein structure inside the box and fills it with solvent molecules or with the solvate module. The solvate module is a smaller box filled only with solvent molecules at the desired conditions, i.e. concentration; for example the modules used here are the urea aqueous solution at 8M and the PEG solution at 200 g/l. Ions, usually in the form of sodium chloride, are also added to preserve the neutrality of the system.

2.7. Moving through the conformational space: the algorithm

Force fields allow calculating the forces acting on each atom given a structure with atomic position x, since the force F is the negative gradient of the potential energy E:

$$F = -\frac{u}{d}$$

However this static picture becomes alive only when velocities come into play; initially they can be randomly assigned from a Maxwell-Boltzmann distribution that depends on the temperature. The movement of the atoms can,then, be simulated by numerically solving the Newton's law of motions:

$$\frac{F}{m} = \frac{d\,v}{d\,t} = \frac{d^2x}{dt^2}$$

where v and x are the velocities and the positions at time t while m are the masses. An iterative, step by step numerical integration is used to obtain an approximate solution: the integration is broken down into many small stages, each separated by a fixed interval of time Δt, the simulation timestep, during which the forces are assumed to remain constant. A review on the major algorithm employed nowadays can be find here [28] and a more extensive explanation, together with further details, can be

found in textbooks such as [29], [30]. Here I simply the topic stressing only that the repetition of this step form a trajectory along time, similarly to a moving picture composed by several still images. When frames are visualized together, the outlook reminds of the high-speed photography developed in the 50s by Harold Eugene Edgerton using the stroboscope (**Figure 2.6**).

While MD simulation uses the integrations of velocities at each time step, the "strobe" uses brief repetitive flashes of light that freeze subse-

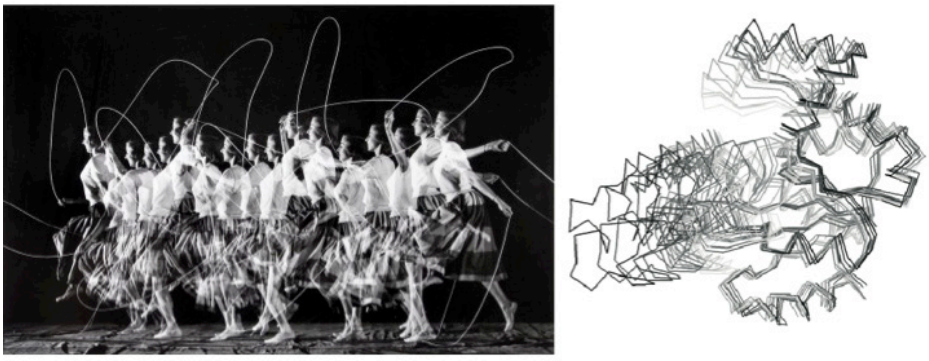


Figure 2.6. Frames of dynamic motions. High-speed photography, developed by Edgerton, creates still images of a motion (for example, the jump rope on the left). Similarly, MD alghoritm calculates the movements of a protein in a step-wise manner (on the right).

quent still images. In both cases the frequencies of recorded frames affects the final image: when a vibrating object is observed at its vibration frequency (or a multiple), it appears stationary. Thus the fast vibrations limit the resolution of both stroboscope and MD simulations. In the context of molecules the fastest motion (under the femtosecond) occur in bonds involving hydrogen. Those bonds vibration are irrelevant at biological level hence they can be frozen by means of specific **restrained algorithms** among which we can recall SHAKE [31], its modification RATTLE [32] and LINCS [33], all of them based on the use of Lagrange multipliers. Longer timesteps (2 femtosecond) then can be applied enhancing the time reachable by MD simulation.

By integrating the Newtonian equations of motion, energy is conserved producing a microcanonical ensemble, where, the number of particles N,

the volume V, and the total energy E are kept constant (**NVE** ensemble). Experiments on the other hand are often conducted at constant temperature and/or constant pressure. Therefore MD simulations are often performed in other ensembles such as the canonical ensemble (**NVT**), where the temperature is kept constant while the total energy can vary, or the isothermal–isobaric ensemble (**NPT**), where both temperature and pressure are kept constant and allowing the volume to change. Again for a deeper overview on the most common thermostat and barostat used in MD simulations the reader is referred to [28]. Most of the technical details depends on the **software** used to perform the MD simulation. In this work I employed several programs: the open sources codes NAMD (Nano scale Molecular Dynamics) [34] and GROMACS (GROningen Machine for Chemical Simulation) [35]; and the licensed package AceMD [36]- a counterpart of NAMD compatible with graphics processing units (GPUs). For the exact specification on each project we referred to the method section of each publication (Chapter 4 and 5).

The solvated, electroneutral system is not yet suited as starting structure for MD simulation. The structure extracted from PDB might suffer from steric clashes or bond-angle deformations and the addition of solvent molecules and ions might have introduced further van der Waals overlaps. The system is generally relaxed through an **energy minimization**. As we have seen, the energy landscape of an IOP has a funnel like shape with a global minimum and a very large number of local minima. Given a starting configuration, it is possible to find the nearest local minimum that can be reached by systematically moving down the steepest local gradient of the energy [37]. Following minimization, the system still needs some refinement: it has to reach the desire conditions (temperature and pressure) and it can still contain unphysical arrangements due to the incorrect solvent placement around the protein. An **equilibration** run, then, is usually performed prior to the actual production run for data collection. At this phase position restrains are applied to the protein structure allowing the solvent molecules to re-orient and move to find a more natural positions with respect to one another, and to the protein. Once the system is well-equilibrated at the desired conditions, the position restraints

are released and the **MD run** for data collection can begin, following the iteration steps described in the previous section. During an MD run, the protein is exploring a portion of its energy landscape depending on the starting position (where the ball is placed in the landscape) and the assigned velocities (the push). Indeed copies of the same structure that differ only in the starting velocities might lead to different trajectories. When many possibilities exist for the starting conformation, as is in the case of an ensemble of structure, calculations can be performed in parallel using a variety of starting structures (See Section 4.1). Despite there are many sophisticated methods to enhance the sampling, such as replica-exchange molecular dynamics or metadynamics [38], in the second project (Section 4.2) we simply enhance the sampling abilities of MD simulations by employing a mildly higher temperature, which has the effect to lower the barrier between minima and encourage conformational exchange (see **Figure 2.7**).

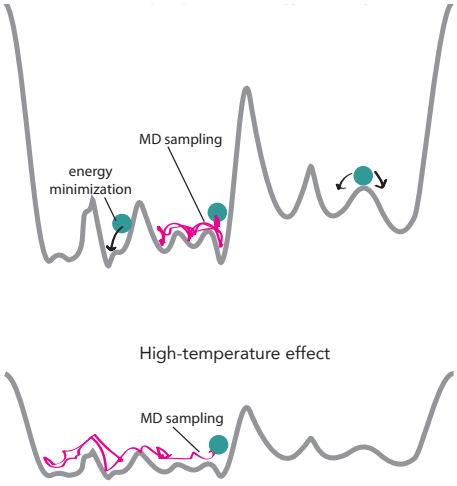


Figure 2.7. MD sampling of the energy landscape. The sampling of a single MD simulation on the energy surface (cross-section) is represented by a magenta line, and the starting conformation as a ball on the surface. While minimization will inevitably make it fall to the closest local minimina, MD can cross certain barriers. The stochastic nature of simulation can lead the sampling towards different sides of the energy landscape (on the right). On the bottom: high temperatures smooth the energey barriers and can be use to increased the sampling capabilities of simulations.

Bibliography Chapter 2

[1] R. O. Dror, R. M. Dirks, J. P. Grossman, H. Xu, and D. E. Shaw, "Biomolecular simulation: a computational microscope for molecular biology," Annu. Rev. Biophys., vol. 41, pp. 429–452, 2012.

[2] J. A. McCammon, B. R. Gelin, and M. Karplus, "Dynamics of folded proteins," Nature, vol. 267, no. 5612, pp. 585–590, Jun. 1977.

[3] S. Lifson and A. Warshel, "Consistent Force Field for Calculations of Conformations, Vibrational Spectra, and Enthalpies of Cycloalkane and n-Alkane Molecules," J. Chem. Phys., vol. 49, no. 11, pp. 5116–5129, Dec. 1968.

[4] J. W. Ponder and D. A. Case, "Force fields for protein simulations," Adv. Protein Chem., vol. 66, pp. 27–85, 2003.

[5] J. Wang, P. Cieplak, and P. A. Kollman, "How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules?," J. Comput. Chem., vol. 21, no. 12, pp. 1049–1074, Sep. 2000.

[6] V. Hornak, R. Abel, A. Okur, B. Strockbine, A. Roitberg, and C. Simmerling, "Comparison of multiple Amber force fields and development of improved protein backbone parameters," Proteins, vol. 65, no. 3, pp. 712–725, Nov. 2006.

[7] K. Lindorff-Larsen, S. Piana, K. Palmo, P. Maragakis, J. L. Klepeis, R. O. Dror, and D. E. Shaw, "Improved side-chain torsion potentials for the Amber ff99SB protein force field," Proteins, vol. 78, no. 8, pp. 1950–1958, Jun. 2010.

[8] W. L. Jorgensen, D. S. Maxwell, and J. Tirado-Rives, "Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids," J. Am. Chem. Soc., vol. 118, no. 45, pp. 11225–11236, Jan. 1996.

[9] A. D. MacKerell, D. Bashford, M. Bellott, R. L. Dunbrack, J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. Reiher, B. Roux, M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiórkiewicz-Kuczera, D. Yin, and M. Karplus, "All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins," J. Phys. Chem. B, vol. 102, no. 18, pp. 3586–3616, Apr. 1998.

[10] A. D. Mackerell, M. Feig, and C. L. Brooks, "Extending the treatment of backbone energetics in protein force fields: Limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations," J. Comput. Chem., vol. 25, no. 11, pp. 1400–1415, Aug. 2004.

[11] T. Meyer, M. D'Abramo, A. Hospital, M. Rueda, C. Ferrer-Costa, A. Pérez, O. Carrillo, J. Camps, C. Fenollosa, D. Repchevsky, J. L. Gelpí, and M. Orozco, "MoDEL (Molecular Dynamics Extended Library): A Database of Atomistic Molecular Dynamics Trajectories," Structure, vol. 18, no. 11, pp. 1399–1409, Oct. 2010.

[12] M. Rueda, C. Ferrer-Costa, T. Meyer, A. Pérez, J. Camps, A. Hospital, J. L. Gelpí, and M. Orozco, "A consensus view of protein dynamics," Proc. Natl. Acad. Sci., vol. 104,

no. 3, pp. 796–801, Jan. 2007.

[13] K. Lindorff-Larsen, P. Maragakis, S. Piana, M. P. Eastwood, R. O. Dror, and D. E. Shaw, "Systematic Validation of Protein Force Fields against Experimental Data," PLoS ONE, vol. 7, no. 2, p. e32131, Feb. 2012.

[14] M. Knott and R. B. Best, "A Preformed Binding Interface in the Unbound Ensemble of an Intrinsically Disordered Protein: Evidence from Molecular Simulations," PLoS Comput Biol, vol. 8, no. 7, p. e1002605, Jul. 2012.

[15] F. Palazzesi, M. K. Prakash, M. Bonomi, and A. Barducci, "Accuracy of Current All-Atom Force-Fields in Modeling Protein Disordered States," J. Chem. Theory Comput., vol. 11, no. 1, pp. 2–7, Jan. 2015.

[16] R. B. Best, W. Zheng, and J. Mittal, "Balanced Protein–Water Interactions Improve Properties of Disordered Proteins and Non-Specific Protein Association," J. Chem. Theory Comput., vol. 10, no. 11, pp. 5113–5124, Nov. 2014.

[17] S. Piana, A. G. Donchev, P. Robustelli, and D. E. Shaw, "Water Dispersion Interactions Strongly Influence Simulated Structural Properties of Disordered Protein States," J. Phys. Chem. B, vol. 119, no. 16, pp. 5113–5123, Apr. 2015.

[18] C. M. Baker, "Polarizable force fields for molecular dynamics simulations of biomolecules," Wiley Interdiscip. Rev. Comput. Mol. Sci., vol. 5, no. 2, pp. 241–254, Mar. 2015.

[19] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, and J. Hermans, "Interaction Models for Water in Relation to Protein Hydration," in Intermolecular Forces, B. Pullman, Ed. Springer Netherlands, 1981, pp. 331–342.

[20] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein, "Comparison of simple potential functions for simulating liquid water," J. Chem. Phys., vol. 79, no. 2, pp. 926–935, 1983.

[21] L. J. Smith, H. J. C. Berendsen, and W. F. van Gunsteren, "Computer Simulation of Urea–Water Mixtures: A Test of Force Field Parameters for Use in Biomolecular Simulation," J. Phys. Chem. B, vol. 108, no. 3, pp. 1065–1071, Jan. 2004.

[22] A. Caflisch and M. Karplus, "Structural details of urea binding to barnase: a molecular dynamics analysis," Structure, vol. 7, no. 5, pp. 477–S2, May 1999.

[23] A. Caballero-Herrera, K. Nordstrand, K. D. Berndt, and L. Nilsson, "Effect of Urea on Peptide Conformation in Water: Molecular Dynamics and Experimental Characterization," Biophys. J., vol. 89, no. 2, pp. 842–857, Aug. 2005.

[24] A. W. P. -O. Åstrand, "Nonempirical intermolecular potentials for urea-water systems," J. Chem. Phys., vol. 100, no. 2, pp. 1262–1273, 1994.

[25] J. Fischer, D. Paschek, A. Geiger, and G. Sadowski, "Modeling of aqueous poly(oxyethylene) solutions: 1. Atomistic simulations," J. Phys. Chem. B, vol. 112, no. 8, pp. 2388–2398, Feb. 2008.

[26] H. Lee, R. M. Venable, A. D. MacKerell, and R. W. Pastor, "Molecular Dynamics

Studies of Polyethylene Oxide and Polyethylene Glycol: Hydrodynamic Radius and Shape Anisotropy," Biophys. J., vol. 95, no. 4, pp. 1590–1599, Aug. 2008.

[27] J. M. Stubbs, J. J. Potoff, and J. I. Siepmann, "Transferable Potentials for Phase Equilibria. 6. United-Atom Description for Ethers, Glycols, Ketones, and Aldehydes," J. Phys. Chem. B, vol. 108, no. 45, pp. 17596–17605, Nov. 2004.

[28] S. Hug, "Classical Molecular Dynamics in a Nutshell," in Biomolecular Simulations, L. Monticelli and E. Salonen, Eds. Humana Press, 2013, pp. 127–152.

[29] D. Frenkel and B. Smith, Understanding molecular simulation: from algorithms to applications. Academic Press, 2002.

[30] D. Rapaport, The art of molecular dynamics simulation. Cambridge University Press, 2004.

[31] J.-P. Ryckaert, G. Ciccotti, and H. J. Berendsen, "Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes," J. Comput. Phys., vol. 23, no. 3, pp. 327–341, 1977.

[32] H. C. Andersen, "Rattle: A 'velocity' version of the shake algorithm for molecular dynamics calculations," J. Comput. Phys., vol. 52, no. 1, pp. 24–34, Oct. 1983.

[33] B. Hess, H. Bekker, H. J. C. Berendsen, and J. G. E. M. Fraaije, "LINCS: A linear constraint solver for molecular simulations," J. Comput. Chem., vol. 18, no. 12, pp. 1463–1472, Sep. 1997.

[34] J. C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R. D. Skeel, L. Kalé, and K. Schulten, "Scalable molecular dynamics with NAMD," J. Comput. Chem., vol. 26, no. 16, pp. 1781–1802, Dec. 2005.

[35] S. Pronk, S. Páll, R. Schulz, P. Larsson, P. Bjelkmar, R. Apostolov, M. R. Shirts, J. C. Smith, P. M. Kasson, D. van der Spoel, B. Hess, and E. Lindahl, "GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit.," Bioinforma. Oxf. Engl., vol. 29, pp. 845–54, 2013.

[36] M. J. Harvey, G. Giupponi, and G. D. Fabritiis, "ACEMD: Accelerating Biomolecular Dynamics in the Microsecond Time Scale," J. Chem. Theory Comput., vol. 5, no. 6, pp. 1632–1639, Jun. 2009.

[37] M. Christen and W. F. van Gunsteren, "On searching in, sampling of, and dynamically moving through conformational space of biomolecular systems: A review," J. Comput. Chem., vol. 29, no. 2, pp. 157–166, Jan. 2008.

[38] T. Schlick, "Molecular dynamics-based approaches for enhanced sampling of long-time, large-scale conformational changes in biomolecules," F1000 Biol. Rep., vol. 1, Jul. 2009.

"Definition belongs to the definers, not to the defined"

TONI MORRISON, BELOVED

CHAPTER 3

Little handbook for the analysis of MD simulations

The outcome of an MD simulation is a trajectory with the positions and velocities of every atom of the system available at each time step. This enormous set of data needs a careful data mining, which relies on relevant observables and suitable descriptors for the given objectives. In the case of this thesis, which focus on changes in the protein features that depend on the environment, several tools were used, first, to define a control state and, second, to approach and quantify the change in *i)* protein structure; *ii)* protein dynamics and *ii)* interactions with the solvent. In this section I will review the main analysis, which are often common to all the projects. When possible I will point to a specific figure in the publications (#P1 for Publication 1, Section 4.1; #P2 for Publication 2, Section 4.2; and #P3 for Publication 3, Section 5.1)

3.1. Observables of protein structure

RMSD – Root Mean Square Deviation

RMSD quantifies the structural difference of a structure X with respect to a reference structure X^0. To appropriately calculate the RMSD the two conformations X and X^0 need to undergo to a structural superposition. This can be done via a simple least-squares fitting algorithm that finds the optimal rotations R and translations T to minimizing the sum of the squared distances among all structures. Subsequently the RMSD can be calculate according to:

$$RMSD = \min_{\{R,T\}} \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\vec{x}_i - \vec{x}_i^0\right)^2}$$

where x_i and x_i^0 are the coordinates of each of the N selected atom respectively in conformation X and in the reference structure X^0. RMSD efficiently condense the behavior of a structure and it can be applied in several ways: i) RMSD values calculated over time and against a single reference structure (i.e. the starting structure or PDB) can identify thermal fluctuations (within 3 Å) or conformational changes (larger values). For the latter i.e. folding/unfolding transition, the RMSD calculated for backbone atoms is mostly relevant; (#P1 Fig 1A and 1B; #P2 Fig 1, 3, S1, S2, Tab S1, S2, #P3 Fig S4) ii) the average RMSD calculated in shorter time windows (time lag i.e. from 2 ns up to 200 ns) and using the first structure in that window as reference instead gives insight on protein mobility on shorter timescales (#P2 Fig S4B); iii) the all-against-all distribution of RMSD values (pairwise RMSD) calculated for an ensemble of structures can quantify the structural diversity and be employed in the clustering of structures (#P1 Fig S5).

TM Score – Template Modeling Score

The TM-score [1] captures similarity between proteins with different tertiary structure, measuring the global similarity and relying less on local

structural variations. Therefore it offers a more accurate measure for large conformational changes of the same protein (i.e. folding / unfolding transition) than RMSD.

$$\text{TM-score} = \frac{1}{L}\left[\sum_{i=1}^{L_{ali}} \frac{1}{1 + d_i^2/d_0^2}\right]_{max}$$

where L and L_{ali} are the lengths of the target protein and the aligned region respectively; d_i is the distance of the *i-th* pair of residues between the two structures; the 'max' implies the procedure that maximize the superposition matrix; and d_0 is defined as $\sqrt[3]{L-15} - 1.8$ to normalize the average TM-score to become independent on the size of the protein. In this way the TM-score assumes values between (0,1] where 1 indicates a perfect match between two structures. Generally scores below 0.20 corresponds to randomly unrelated structures whereas score higher than 0.5 assume structures with the same fold (#P2 Fig 1, 3, S1, Tab S1, S2) [1]

RGYR – Radius of Gyration

The radius of gyration gives a rough measure for the level of compaction in the structure. It can be calculated using the distances between each atom and the center of mass:

$$R_g = \sqrt{\frac{\sum_i r_i^2 \cdot m_i}{\sum_i m_i}}$$

where m_i is the mass of atom *i* and r_i the position of atom *i* with respect to the center of mass of the molecule. For examples see #P1 Fig 1C; #P2 Fig 1, 3, S1, Tab S1.

SASA - Solvent Accessible Surface Area

As clearly stated by its name, the SASA is the portion of the surface area of a biomolecule that is accessible to the solvent. SASA was first described by Lee & Richards in 1971 and is traditionally calculated using the 'rolling ball' algorithm [2]. Indeed one of the software used here (NACCESS [3]) uses a probe of given radius rolled around the surface of the molecule, the path traced out by its center is the accessible surface

(**Figure 3.1**). Typically, the probe has the radius of a water molecule (1.4 Angstroms) and hence it is referred as *solvent* ASA. Thin slices through the 3D molecular volume are used to calculate the accessible surface of individual atoms, the thinner the slices the higher the accuracy (#P2 Fig 1-3, S1, S3). A more computationally efficient procedure is implemented in the standard analysis package of GROMACS (g_sas) and it computes hydrophobic, hydrophilic and total solvent accessible surface area using the double cubic lattice method [4] (#P3 Fig 3).

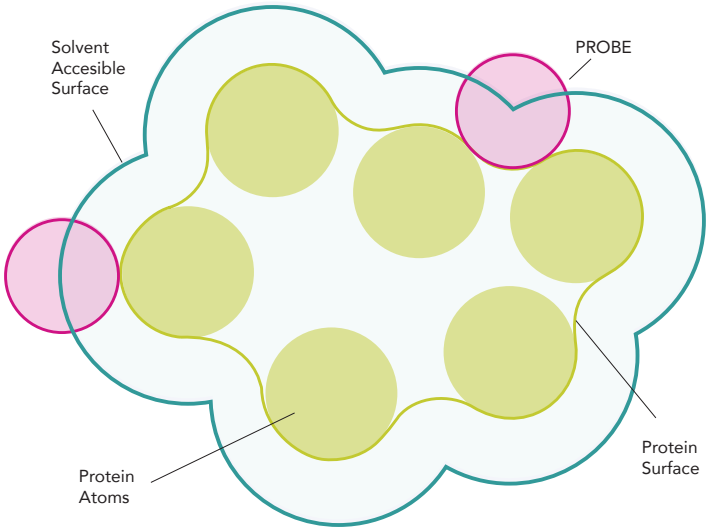


Figure 3.1. Calculation of SASA. The solvent accessible surface (blue line) defined by a probe (in magenta in two sample positions) rolling over the molecule atoms (yellow).

SS - Secondary Structure detection

Computational algorithms can apply the same logic used by Linus Pauling (Section 1.6) to predict the secondary structure of a protein based on its atomic coordinates. Two software packages are widely used to assign SS, DSSP and STRIDE, the latter has been selected here because its assignments are based on both the detection of hydrogen pattern and the values of the backbone dihedrals [5], instead DSSP rely solely on hydrogen bonding pattern [6]. The assignment function in STRIDE has an

hydrogen-bond term containing a Lennard-Jones like 8-6 distance-dependent potential and two angular dependence factors reflecting the planarity of the optimized hydrogen bond geometry and a more detailed explanation of the method is available in [5]. The output of STRIDE assign each residue to one of, the main SS categories are described in **Table 3.1**. In addition, if a helix or sheet is too short, the residues involved are designated as turns or bridges, respectively. Everything that doesn't fit in

	HB	Dihedrals {φ,ψ}	Minimum length
Alpha helix	i+4→i	−60, −50	4
3_{10} helix	i+3→i	−50, − 25	3
PI-helix	i+5→i	−60, +125	5
Beta-sheets	Between strands	−120, +115 or −140, +135	2

Table 3.1 Criteria for secondary structure assignments in STRIDE. Hydrogen bond (HB) connectivity for the i-th residue; its backbone dihedrals; and the minimum number of subsequent residues needed to form a structural element.

the categories above is designated as random coil.

The assignment of SS in a MD trajectory can be visualize in a plot time vs sequence in which each SS element has a specific color code, however for long trajectory the delivered information becomes cryptic and confusing. Therefore here I often summarized the SS data either in tables or in a plot sequence vs frequency, in which for each residue the frequency in which it adopts a SS element is reported. Values are generally grouped together for each type of helix and each type of beta-sheets to allow a direct comparison of the two main types of SS (#P1 Fig 3A; #P2 Tab S3; #P3 Fig S8). The low resolution of such representation however hides important details such as the effective length of helices. For example, it is possible to resolved such information calculating for each helical element its length and starting residue.

CM – Contact Maps

The contact map is a powerful tool to detect the local rearrangements of

a protein: it visualizes the pairwise distances between residues (intra-protein or inter-protein in the case of protein complexes) either through a binary code (1 contact, 0 no contact, see #P2 Fig 6, #P3 Fig S1) or through a color code proportional to the frequency in which two residues are found in contact -the latter is mostly used with MD trajectories or ensemble of structures. In this work two residues are defined in contact when the distance between their C-alpha is smaller than 10 Å (#P1 Fig 4, #P2 Fig 6, #P3 Fig 2, S1, S6, S7 and for inter-protein contacts #P3 Fig S9, S10). In the case of IOPs the specific network of contacts between proteins residues creates a unique and reduced representation of the native structure **(Figure 3.2).** For example from the contact map some of the secondary elements are already recognizable: helices are identified by strips directly adjacent to the diagonal while beta-sheets appear as parallel or perpendicular (anti-parallel) lines to the diagonal. More interestingly,

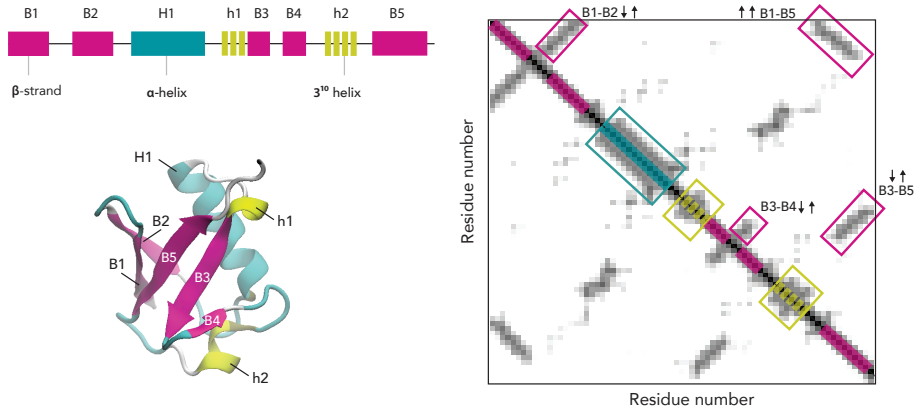


Figure 3.2. The contact map of the protein ubiquitin. Each structural element (B for beta strand, H for alpha helix, h for 3^{10} helix) is shown along the protein sequence, its position in the 3D structure and in the 2D contact map.

CM provides a direct, coarse-grained, and robust picture of the global fold of the protein, which is especially useful to characterize large conformational changes.

The definition of native features in IOPs

The concept of native structure applies exclusively to IOPs and it is tra-

ditionally defined as the structure (atomic coordinates, contact and secondary structure elements) determined experimentally and deposited in PDB. However the frozen picture of the native structure obtained from PDB can be improved adding information about the fluctuations around that structure, which are defined by MD simulations starting from the PDB reference structure. According to this dynamic definition of native structure:

- The native contacts and native secondary structure are defined as those present more than 80% of the time in the control simulation.
- The native protein core is defined by residues that exhibit a stable "buried" pattern in the control simulation: only residues with an average SASA and a standard deviation below the threshold of 10 Å2 are part of the protein core, regardless of their hydrophobic nature (#P2 Fig 4A).

The definition of such reference features is essential to make the right comparison. For example the protein native contacts can be lost during large protein transition i.e. the protein unfolding; CMs visualize where these rearrangements are located but for a more precise description a definition of lost contacts is needed. Here I defined a lost contact when it faces a reduction greater than 30% of the total simulated time - compared to the control simulation (#P2 Fig 6, S4A).

In studying large transition, i.e. protein unfolding, a suitable coordinate to follow the progress is the fraction of native structure that persists at each time frame. In this work the selected coordinate was the **Structure Index SI** defined as the sum of the existing fraction of native secondary structure S2 and the sum of existing native contacts S3 as in [7] **(Figure**

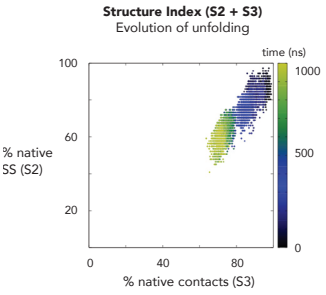


Figure 3.3. SI and unfolding. Both components of the Structural Index decrease along the unfolding simulation of a protein (PDB:1CQY).

3.3 and #P2 Fig 1, 3). This metric is extremely useful to compare the degree of unfolding experiences by different protein because it is independent on the size and the structural features of each system.

Bi-dimensional map to overview the protein sampling

When analyzing a large ensemble of structurally different conformations of the same protein, i.e. IDPs or conformational changes of IOPs, we need an appropriate visualization of the explored portion of the energy landscape (sampling). Here I often employed bi-dimensional maps that project two variables over time or according to their contingent frequency. In each case the observables should be carefully chosen to discern the relevant features of the dynamic ensemble under study. For example a plot with the Rgyr/SASA resolves the compactness of the explored structures (#P1 Fig 2, S6); the ΔSASA/SI plot allows to follow the opening of the structure during several unfolding transition (#P2 Fig 2); similarly to the plots with Rgyr/RMSD and SASA/SASA-Polar, which allowed to compare the outcome of simulations of the same protein but in different environments (#P3 Fig 3, Fig S5). Two variables related to the relative 3D positioning of helices were also used to follow the sampled structural arrangements in different environments (#P3 Fig S3)

Clustering

Clustering method can distill the salient structural feature while reducing the dimensionality of the trajectory and are very useful to condense structural information obtained in complex trajectories. Clustering detects easily the most common conformations and the rare ones that otherwise would be difficult to discern from average values. For example gromos [8], the algorithm used in this work, divide the conformations, regardless of when and where they occur, in sub-groups (clusters) based on their structural similarity - measured by the RMSD. The pair-wise RMSD of all the conformations is used to create a neighbors network: two structures are neighbors if their RMSD is below some user-defined cutoff. The conformation with the largest number of neighbors is the central member, or centroid, of the cluster whose properties are assumed to represent the entire group. All of its neighbors are then removed and the iteration is re-

peated for the remaining structures in the pool. The number of centroids that the clustering returns is sensitive to the RMSD cutoff: the larger the value, the fewer the identified clusters. For example in one of the projects presented here I used a two-steps clustering: first a cutoff of 1.5 Å was applied on separated trajectories and later a 3.5 Å cutoff on a joint ensemble of several trajectory with the aim to recognized commonalities (#P3 Fig 1, S2). The time-resolved output of a clustering can be used also to monitor conformational changes, as explained in the following section.

3.2. Observables of protein dynamics

Local Level

RMSF - Root Mean Square Fluctuation

The RMSF quantifies local changes along the protein chain: it gives the average fluctuations (deviations of the position) over time for each atom –RMSD instead gives the time resolved average over all the particles. For each atom *i:*

$$RMSF_i = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(\vec{x_i}(t) - \langle\vec{x_i}\rangle)^2}$$

where T is the overall trajectory time, t is the selected time frame, x_i is the position of atom i after superposition on the reference structure, and $\langle x_i \rangle$ is the average reference position over time T. Usually in a protein the tails (N- and C-terminal) fluctuate more than any other part while secondary structure elements tend to be more rigid than the unstructured parts. The RMSF of the protein can also be correlated with the experimental x-ray B-factor (or temperature or Debye-Waller factors) via the equation:

$$B = 8\pi^2 \frac{RMSF^2}{3}$$

Similarly to RMSF, B-factors are defined as a measure of spatial fluctuations of atoms around their average position. They can be obtained not

only from MD simulations, but also from crystal data. However, experimental values include effects of noise due to refinement errors, lattice defects, crystal contacts, and rigid-body motions and an exact correspondence between simulated and experimental values should not be expected (#P2 Fig 6, S4D). As for RMSD, the group of atoms used for the superimposition and for the RMSF calculation can differ. For example, during large transition that critically affect the backbone a protein superimposition will fail to grasp local motion – in this work I instead calculated the RMSF of each side-chain after aligning the structures based solely on the backbone atoms of the same residue (#P2, page 4, line 12; #P3 Fig S6, S11 and Tab S1). This metric allows quantifying the local motion experienced by each residue, independently from the global structure rearrangements.

Contact disruption and exploration

Contacts experience fluctuation in the form of brief opening events. Here the inter-atomic distance is used to distinguish two conformations: open when the minimum distance between two heavy atoms is larger than 5 Å and closed when it is smaller than 4 Å. In this work the focus was on the

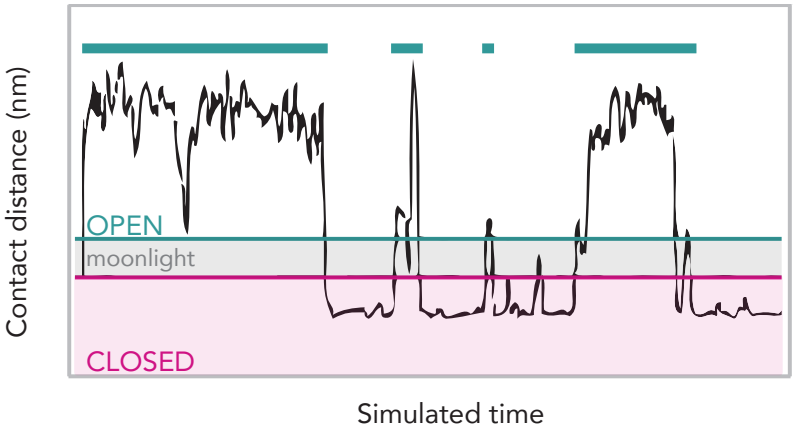


Figure 3.4. Contact dynamics. The distance between two residues is used to define when the contact is open (> 5 Å) or closed (<4 Å), marked by the blue and the magenta lines, respectively. The bold blue line on top defined the time interval where the contact is considered as open.

average opening time, the moonlight zone between 4 and 5 Å is considered a neutral area in which the contact assume the conformation of the previous frame to avoiding over detection of fast noise movements [9] (#P2 Tab 2, Fig S4C).
It is useful in the context of trajectories involving large conformational changes to evaluate the portion of all the possible intra-protein contacts that have been explored during the MD simulation. Here I simply defined a contact as explored, when it is present for more than a frame in the trajectory. The amount of the explored contacts among all the possible ones (n^2-n , with n the number of residues) can be used to quantify the conformational plasticity that a protein experience. It is also related to the fuzziness observed in the contact maps: the larger the amount of contacts explored, the fuzzier the contact-plot, which imply a prevalence of transient contacts and a larger structural plasticity (#P3 Fig 6, S11 and Table S1).

- **Global Level**

Conformational entropy
The conformational entropy is associated with the number of conformations explored by a biomolecule and depends on the complete energy landscape of a molecule and it is defined in terms of probabilities of to occupy an individual conformation of a molecule (microstate). The conformational entropy is:

$$S^C = -k_B \sum_i p_i \ln p_i$$

where k_B is the Boltzmann constant and p_i is the probability of occupancy of each microstate. Entropy cannot be calculated as a simple average from an MD simulation but it needs an extensive sampling of all the degrees of freedom. However a reasonably long MD trajectory can be used to estimate the upper limit for the configurational entropy S^C for conformation C. One of the most widely used methods is the quasi-harmonic approximation [10], which assumes that atomic fluctuations have a Gauss-

ian distribution. This is the best bet because it has the highest entropy among all the statistical models with the same variance. A mechanical model that reproduced Gaussian-distributed coordinate displacements is the harmonic oscillator .Within this approach potentials are fitted on the observed coordinated covariance σ, smoothing over any anharmonicity:

$$S^C = \frac{1}{2} nk_B + \frac{1}{2} k_B \ln[(2\pi)^n \sigma]$$

The input data is the covariance matrix σ of atomic fluctuations extracted from the Cartesian coordinates of an MD simulation trajectory . For each of the 3N Cartesian coordinates, where N is the number of considered atoms, the individual elements in the (3N,3N) covariance matrix are:

$$\sigma_{i,j} = \langle (x_i - \overline{x_i})(x_j - \overline{x_j}) \rangle = \frac{\sum_{f=1}^{Nf} (x_{if} - \overline{x_i})(x_{jf} - \overline{x_j})}{(N_f - 1)\sigma_i \sigma_j}$$

where the covariance $\sigma_{i,j}$ measures how much the two *i-th* and *j-th* coordinates x_i, x_j deviate together in reference of their pre-calculated average coordinate *x*; and <...> denotes the average across the considered frames or trajectory. GROMACS standard package easily implement entropy calculation using a two-step procedure: first it calculates the covariance matrix and then estimates the entropy based on the quasi harmonic approach. This approach was employed in to calculate values in #P3 Fig 6; S11 and Table S1.

Time lapses for reconfigurational events

The time-resolved output of a clustering (each frame has an associated cluster, see Section 3.1 for an explanation) provides valuable information to evaluate the time lapses between conformational changes. Here, we assumed that when two sequential frames belong to different clusters a reconfigurational event has occurred; with this information in hand it then easy to extract the average time between conformational changes. Such information quantifies the frequency of structural rearrangements, which then can be used to compare simulations in different environments (#P3 Fig 6; S11 and Table S1). The number of reconfigurational events is independent on the number of clusters since a change in cluster only need

a minimum of two options.

3.3. Protein and the solvent

FSS – First Solvation Shell and Bulk

The molecules surrounding the protein behave differently from those in pure solvent. Thus, two groups of solvent can be then defined in a simulation: the first solvation shell (FSS) is formed by solvent molecules within 5 Å from the protein; the bulk instead by molecules that lie at a distance from the protein larger than 6Å [11] (#P1 Tab 2, Fig S1A ; #P2 Tab 3).

Identification of protein/solvent contacts

To evaluate the contacts that solvent molecules can form with protein residues, I used a criterion based on atomic distance: a contact is formed when at least two heavy atoms belonging to each molecule are closer than 3.5 Å (#P2 Fig 5, 6, S6, #P3 Fig 5,6). Contacts can then be easily grouped according to *i)* the nature of the residues involved (polar or apolar); *ii)* the portion that forms the contact (sidechain or backbone) and *iii)* the position in the native structure (part of the native protein core or not). Once these contacts are defined other solvent features can be easily extracted, for example the ratio of apolar/polar contacts (#P1 Tab 2, Fig S1, #P2 Tab 3, Fig S5A) or the residence time of a solvent molecule around the protein (#P2 Fig 4B). Among all the contacts, the peculiar nature of the hydrogen bonds allows to identify them with a simple geometrical criterion: a cutoff of 3.5 Å for the distance between the two electronegative atoms (the donor and the acceptor) and 120° for the angle between the donor-hydrogen and the acceptor (#P1 Tab 3, #P2 Tab 4, S4).

When two or more co-solvent types are present, the contact preference with either one co-solvent or the other can be quantified with the Contact Coefficient (CC). Here I specifically employed this metrics to evaluate solvation of proteins by urea and water molecules (CC_{UW}) [12], defined as:

$$CC_{UW} = \frac{N_{XU}}{N_{XW}} \cdot \frac{M_W}{M_U}$$

where N_{XU} and N_{XW} are the numbers of atomic contacts of amino acid X with urea and water molecules, respectively. CC_{UW} is normalized using the total numbers of atoms belonging to urea molecules M_U or to water molecules M_W in the system. A contact coefficient of 1 means that the amino acid has no contact preference for one of the co-solvents; values above 1 indicate preferential interaction with urea while values below 1 with water (#P1 Tab 2, Fig 3B, S7, #P2 Fig 4A).

Energetics of hydrogen bonds (HB): Atoms in Molecule (AIM)

To evaluate the strength of different hydrogen bonds a deeper analysis of the HB structure is needed. The quantum theory of Atoms in Molecule (AIM), pioneered by Richard Bader [13], comes in hand by considering a bond as a 3D entity, the topology of which quantifies its physical and chemical properties. A powerful observable is the spatial topological decomposition of the electron density ρ(r). According to AIM theory the electron density ρ(r) is at maximum at the atomic nuclei, which allows to clearly identifies the atomic position; chemical bonds can then be easily traced to unite the atomic nuclei. The saddle point, the minimum in electron density ρ(r), along the bonding direction identifies the bond critical point. In non covalent interactions, such as the hydrogen bond, the properties of the density field at the critical point, i.e. the density itself or its Laplacian[1] , are proportional to the strength and the energy of the corresponding interaction [14]. In this work I compared the strength of several HB between the protein backbone and the solvent molecules, by evaluating the electronic

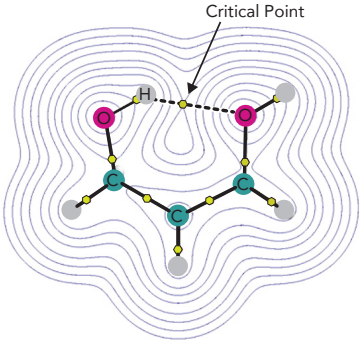


Figure 3.5. Bond critical points. Contour map of the electron density. The atomic symbols in denote the positions of the nuclei; yellow spheres the bond critical points and black lines the bond paths.

1 The scalar derivative of the gradient vector field of the electron density. It determines where electronic charge is locally concentrated (negative values) and depleted (positive values).

distribution at their critical points. The structure were first extracted by MD simulations runs and geometrically optimized by QM calculation from which electron densities are derived (#P1 Tab 4).

Energetics of solvent interactions

A post processing procedure allows calculating the interactions energies between two subgroups of molecule in a MD trajectory. Similarly to an MD run, the forces can be computed following the same pairwise equation given in Chapter 2. In my projects I calculated the electrostatic and Van der Waals energy contribution between each solvent molecule and the rest of the system, taking into account the relative position to the protein (FSS or bulk) at each frame. The comparison of the interaction energy distribution for all the molecules in the FSS, influenced by the presence of the protein, and in the bulk gives insights on the preferential and more favorable interactions with the protein [11](#P1 Fig S8, #P2 Fig 5; 6, S5B).

Solvent diffusion and MSD – Mean Square Displacement

The collective motion of all particles in a fluid is termed diffusion and its quantified by the diffusion coefficient. This macroscopic property relates to the microscopic thermal motion of individual molecules and relates with their average motility quantified in the mean square displacement MSD:

$$\mathrm{MSD} = \langle \Delta\vec{r}(t)^2 \rangle = \frac{1}{\mathrm{N}} \sum_{i=1}^{N} \left(\vec{r}_i(t) - \vec{r}_i(0) \right)^2$$

Where $r_i(t)$ and $r_i(0)$ are the position of particle i at time t and at the reference time 0, respectively. For molecules consisting of more than one atom (i.e. solvent molecules or even proteins), r_i can be taken as the center of mass positions of the molecules. Albert Einstein, in his PhD thesis, derived a relationship between the macroscopic D diffusion coefficient and the microscopic behavior collected in the MSD [15]:

$$\lim_{n \to \infty} \langle \Delta\vec{r}(t)^2 \rangle = 6Dt$$

After calculating the MSD in different time windows (n), the diffusion coefficient D can be derived from the slope of the fitting line. Generally only the last half of values is used for the fitting, since the Einstein relation is valid as time approaches infinity (#P2 Fig S5C; #P3 Table 1).

3.4. Comparison with experimental observables

NMR observables were used to validate the MD trajectories by back calculating values from the MD ensembles and compared to the experimental ones. In the first project we relied on two NMR high-resolution observables and one at low resolution (SAXS) (#P1 Tab 1, Fig S2-S4):

1. **J-couplings**, calculated from the dihedral angles through the Karplus equation [16]:

$$^{3}J_{HN-H\alpha} = 6.4\cos^2\theta - 1.4\cos\theta + 1.9$$

where θ is the dihedral angle between H-N-Cα-H atoms and the coefficients used, usually dependent on the atoms and substitutes involved, are those suggested by [17] for the protein backbone. The superscript 3 indicates that the amide proton (HN) is coupled to the proton of the Cα (Hα) three bonds away, via H-N-Cα-H bonds.

2. **RDC** depends on both the geometry of the nuclei and also the degree and direction of alignment of the molecule with respect to the laboratory frame, which depends on its structure and the mechanism by which alignment is induced. This information is contained in the alignment tensor A, with elements A_{ij}, which is a traceless and symmetric 3 × 3 matrix defined by five independent elements that are, in most cases, unknown and need to be determined empirically. RDCs are back calculated using the equation:

$$D^{\text{calc}} = -\frac{\mu_0 \gamma_X \gamma_Y h}{8\pi^3 r^3} \sum_{ij} A_{ij} \cos\phi_i \cos\phi_j$$

where i,j are the two nuclei, μ_0 is the magnetic susceptibility of vacuum,

γ_X is the gyromagnetic ratio of nucleus X, $\hbar$ is Planck's constant, r is the internuclear distance, and ϕ_i is the angle between the internuclear vector and axis i of the molecular reference frame where A_{ij} is defined.

3. **SAXS** (Small Angle X-ray Scattering) experiment employs an X-ray to provide information about the fluctuations of electronic densities in the matter. The output is a curve that registers the scattering intensity *I(q)* upon variation of the scattering angle and it contains information in the reciprocal space on the structure of the object in solution, carrying information about shape and size of macromolecules. Software like CRYSOL [18]calculates the scattering curve from the structure of macromolecules (i.e. pdb files) and it fits the theoretical scattering curve to the experimental one by minimizing the discrepancy (chi-square value).

Bibliography Chapter 3

[1] Y. Zhang and J. Skolnick, "TM-align: a protein structure alignment algorithm based on the TM-score," Nucleic Acids Res., vol. 33, no. 7, pp. 2302–2309, 2005.

[2] B. Lee and F. M. Richards, "The interpretation of protein structures: estimation of static accessibility," J. Mol. Biol., vol. 55, no. 3, pp. 379–400, Feb. 1971.

[3] S. Hubbard and J. Thornton, NACESS. 1992.

[4] D. Eisenberg, "The discovery of the α-helix and β-sheet, the principal structural features of proteins," Proc. Natl. Acad. Sci. U. S. A., vol. 100, no. 20, pp. 11207–11210, Sep. 2003.

[5] D. Frishman and P. Argos, "Knowledge-based protein secondary structure assignment," Proteins Struct. Funct. Bioinforma., vol. 23, no. 4, pp. 566–579, Dec. 1995.

[6] W. Kabsch and C. Sander, "Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features," Biopolymers, vol. 22, no. 12, pp. 2577–2637, Dec. 1983.

[7] M. C. Stumpe and H. Grubmüller, "Polar or Apolar—The Role of Polarity for Urea-Induced Protein Denaturation," PLoS Comput Biol, vol. 4, no. 11, p. e1000221, Nov. 2008.

[8] X. Daura, K. Gademann, B. Jaun, D. Seebach, W. F. van Gunsteren, and A. E. Mark, "Peptide Folding: When Simulation Meets Experiment," Angew. Chem. Int. Ed., vol. 38, no. 1–2, pp. 236–240, Jan. 1999.

[9] M. Lindgren and P.-O. Westlund, "The affect of urea on the kinetics of local unfolding processes in chymotrypsin inhibitor 2," Biophys. Chem., vol. 151, no. 1–2, pp. 46–53, Sep. 2010.

[10] M. Karplus and J. N. Kushick, "Method for estimating the configurational entropy of macromolecules," Macromolecules, vol. 14, no. 2, pp. 325–332, Mar. 1981.

[11] L. Hua, R. Zhou, D. Thirumalai, and B. J. Berne, "Urea denaturation by stronger dispersion interactions with proteins than water implies a 2-stage unfolding," Proc. Natl. Acad. Sci. U. S. A., vol. 105, no. 44, pp. 16928–16933, Nov. 2008.

[12] M. C. Stumpe and H. Grubmüller, "Interaction of Urea with Amino Acids: Implications for Urea-Induced Protein Denaturation," J. Am. Chem. Soc., vol. 129, no. 51, pp. 16126–16131, Dec. 2007.

[13] R. F. W. Bader, "A quantum theory of molecular structure and its applications," Chem. Rev., vol. 91, no. 5, pp. 893–928, Jul. 1991.

[14] null Espinosa, null Souhassou, null Lachekar, and null Lecomte, "Topological analysis of the electron density in hydrogen bonds," Acta Crystallogr. B, vol. 55, no. Pt 4, pp. 563–572, Aug. 1999.

[15] A. Einstein, "Über die von der molekularkinetischen Theorie der Wärme geforderte Bewegung von in ruhenden Flüssigkeiten suspendierten Teilchen," Ann. Phys.,

vol. 322, pp. 549–560, 1905.

[16] M. Karplus, "Contact Electron-Spin Coupling of Nuclear Magnetic Moments," J. Chem. Phys., vol. 30, no. 1, pp. 11–15, Jan. 1959.

[17] A. Pardi, M. Billeter, and K. Wüthrich, "Calibration of the angular dependence of the amide proton-C alpha proton coupling constants, 3JHN alpha, in a globular protein. Use of 3JHN alpha for identification of helical secondary structure," J. Mol. Biol., vol. 180, no. 3, pp. 741–751, Dec. 1984.

[18] D. Svergun, C. Barberato, and M. Koch, "CRYSOL a Program to Evaluate X-ray Solution Scattering of Biological Macromolecules from Atomic Coordinates}," J. Appl. Crystallogr., vol. 28, no. 6, pp. 768–773, Dec. 1995.

Barcelona, 20 Nov 2015

Comisión de doctorado.

La Sra. Michela Candotti ha estado realizando su tesis doctoral durante los últimos años en el IRB bajo mi dirección. El trabajo de su tesis doctoral se verá reflejado en una serie de publicaciones científicas algunas aún en fase de redacción, una que se ha sometido a revisión muy recientemente y dos ya publicadas.

1) M.Candotti; S.Esteban-Martín; X.Salvatella; M.Orozco "Towards an atomistic description of the urea-denatured state of proteins". *Proc. Natl. Acad. Sci. USA*. (2013) 110, 5933-5938

2) M.Candotti, A.Perez, C.Ferrer-Costa, M.Rueda, T.Meyer, J.L.Gelpí and M.Orozco. "Exploring the early stages of the chemical unfolding of proteins at the proteome scale". *PLOS Comput. Biol.*, (2013), 9, e1003393.

PNAS es una de las revistas multidisciplinares de más impacto con un IF cercano al 10. El artículo en concreto, publicado en 2013 tiene ya más de 20 citaciones, por encima de la media de la revista. PLOS Comput Biol., es la revista de referencia en biología computacional, con un IF cercano al 5. Los artículos enviados o en proceso de escritura apuntan también a revistas de alto impacto.

En todos los artículos la Sra. Candotti es la primera firmante y responsable de haber realizado y discutido la gran mayoría de simulaciones. También ha contribuido a la redacción de los artículos. Los trabajos de su tesis no forman parte de otra tesis doctoral

Prof. Modesto Orozco
Modesto.orozco@irbbarcelona.org

"Arriving at one goal is the starting point to another"

John Dewey

CHAPTER 4

Protein unfolding in urea-aqueous solution

"The compact and crystalline structure of the natural protein molecule, being formed by virtue of secondary valences, is easily destroyed by physical as well as chemical forces. Denaturation is disorganization of the natural protein molecule, the change from the regular arrangement of a rigid structure to the irregular, diffuse arrangement of the flexible open chain."

Hsein Wu - 1931

Hsien Wu, a Chinese biochemist, was the first to define the denatured state of protein at structural level [1]. Despite so, denaturation became a widely known concept only after Mirsky and Pauling presented very similar ideas in a milestone article in 1936 [2]. The definition of the unfolded state is a troublesome task due to its elusive nature; it can be simply regarded as major changes in the structure of a protein which leave

its primary sequence intact but prevent the protein from performing its function [3]. The unfolded/denatured state is not a single entity, rather is formed by many conformations with little native tertiary and secondary structure. The unfolded protein is then high in both conformational entropy and free energy. In fact, in the protein landscape, these conformations correspond to points that become sampled only after perturbations of the physiological conditions [4], [5].

As we have seen in Chapter 1, protein stability depends on the environment, and specifically on changes for example in the temperature, pH, pressure, or in the solvent composition i.e. by adding denaturants to a large concentration (for example the standard denaturing solution of 8M aqueous urea). At macroscopic level the folded and unfolded states of a protein coexist in a dynamic equilibrium, the portion of the population that is folded is experimentally measurable by some conformational probe. For example, the secondary structure elements in a protein absorb circularly polarize light; the amount of absorbed light then, measured by Circular Dichroism –CD- spectroscopy, becomes a marker of the degree of foldedness in the protein ensemble. At structural level, the definition of the denatured "state" is, however, not trivial and depends on the solution conditions. Generally, unfolded conformations, under strong denaturing conditions, are described as highly open and solvent-exposed, with little residual structure; however very little is known about the structural details of the unfolded state and on the differences/similarities with the folded one.

Many osmolytes become part of the standard tool-kit in biochemist laboratories as tools to address the stability of the native state of IOPs . For example, Anfinsen already employed urea in his seminal experiment about spontaneous folding; since then this denaturant has characterized the thermodynamic properties of many proteins, becoming one of the most popular ones [5]. Indeed the urea-induced unfolding transition is easy to follow over a population of the same protein that typically unfold in an all-or-none (cooperative) manner: aminoacid residues cannot be withdrawn from ordered region randomly and one at a time; instead the structure undergo to a "steep" transition which occurs within a narrow

range of concentration of denaturing agent.

In fact the fraction of folded proteins, measure for example by CD, decreases with the addition of urea in solution, following a typical sigmoidal curve **(Figure 4.1)**. At the midpoint half of the ensemble is folded while the rest is unfolded and once the plateau is reached the transition is complete. For most of the proteins a concentration of 8M of urea is enough to reach the completed transition to the denatured state.

The folding/unfolding process then is then an equilibrium for which the equilibrium constant K_{ex} can be extracted and consequently the difference in free energy ΔG between the folded and unfolded states follows the equation:

$$\Delta G = -RT \ln K_{ex}$$

where R is the gas constant and T the absolute temperature. Employing urea in such experiments characterized hundreds of proteins; their ΔG usually falls within a narrow range between -5 and -15 kcal/mol.

The thermodynamic description of folding leaves many open questions regarding the mechanism of action of urea, a subject that has been often debated in literature [7], [8]. Indeed experiments struggle in gaining atomistic insight on protein denaturation: the chaotic nature of the unfolding process makes it difficult to interpret the ensemble measurements and to distinguish the multiple unfolding pathways. The high resolution of MD simulations would easily overcome these issues but the typical millisecond timescale of the unfolding process is out of its reach.

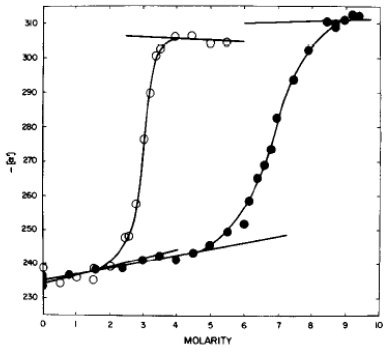


Figure 4.1. The unfolding sigmoidal curve. Optical rotation of ribonuclease (left) and lysozyme (right) as a function of urea (full dots) and Guanidium Cloride (circles). Taken from [6]. The two denaturant agents exert similar effects but at different concentrations.

As seen in Chapter 1, experimental and theoretical studies both agree that urea integrates well into the hydrogen-bonding network of water leaving almost unaffected the solvent structure; as a consequence the solvation term is reduced to the contribution of protein-solvent interactions. Indeed the "**direct mechanism**" theory states that protein-urea interactions are the major driving force of protein unfolding. This theory brought with it a debate around the **nature of such interactions** and the **involved protein moieties**, cause by the ambivalent nature of the urea molecule, which bears both polar and apolar features.

Urea is a roughly planar molecule with one central carbonyl group C=O attached to two NH2 groups (**Figure 4.2**). The H atoms are sufficiently polarized by N, which becomes a good hydrogen donor in H-bonds. Similarly the lone pair electrons on the carbonyl oxygen can accept one hydrogen bond as well. No surprise then that hydrogen bonds were the first to be addressed as driving forces for urea-solvation; however dispersion interactions (also known as van der Waals attractive part, London forces or soft interactions) have quickly stolen the limelight; leading to a vivid discussion: it was easy to spot contrasting conclusions between results, for example [9] and [10]. The introduction of the two publications presented here focus indeed in presenting this topic.

To address the issue with a consistent approach, we set up a first project with the main objective to extract information, valid at proteome level, on the role of those two interactions (Section 4.2, Urea-MoDEL project). While analyzing the data then other questions emerged and the scope was enlarged to include more specific analyses related to the endpoints of the folding/unfolding equilibrium: what is the exact nature of the urea-induced unfolded state of a protein? What is the role of urea in

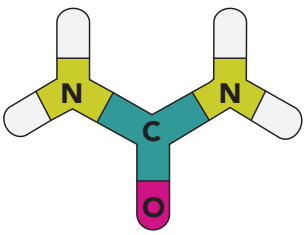


Figure 4.2. The chemical structure of urea.

triggering the unfolding? The two projects that I present in this chapter addressed those two questions independently, still keeping as leitmotif the protein-solvent interaction (**Figure 4.3**). The projects are ordered by year of publication; even in reality the Urea-MoDEL project (Project 2, Publication 2) was initiated before than the Urea-UBQ project (Project 1, Publication 1). Despite the two studied systems are at the opposite side of the unfolding path, they do complement each other offering together a holistic view of the mechanism of urea-induced denaturation.

The **first project (Section 4.1)**, in collaboration with the laboratory of Molecular Biophysics (Prof. Xavier Salvatella) at our institute aimed to characterize the urea-denatured state of ubiquitin, a small prototypical protein highly popular in nuclear magnetic resonance (NMR) studies.

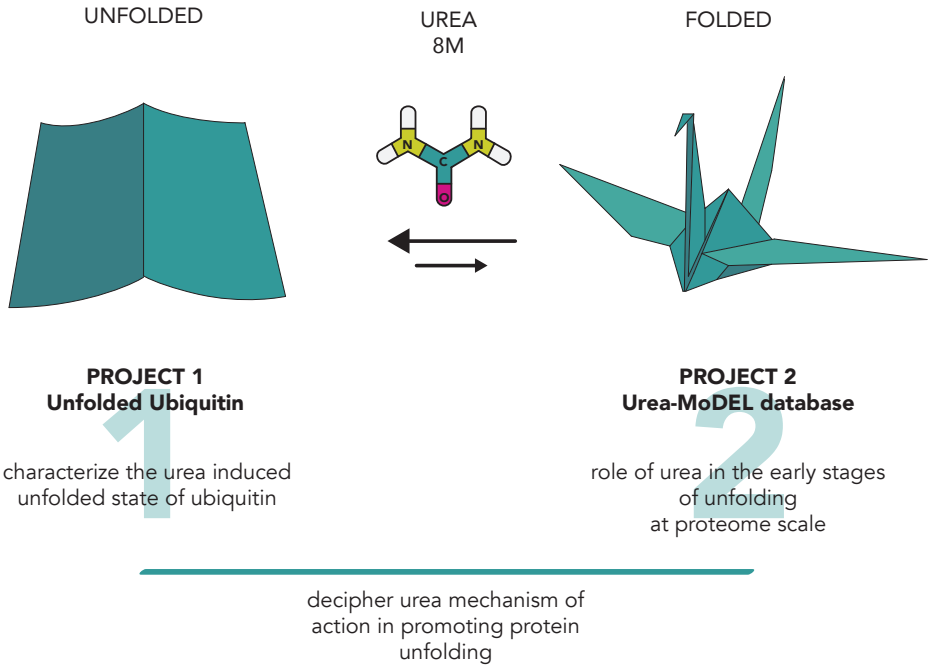


Figure 4.3. Overview of the two projects on urea-induced unfolding. While focusing on the opposite sides of the unfolding path, both projects keep as leitmotif the analysis of protein-solvent interactions and aim to understand the mechanism by which urea triggers the unfolding.

We profit from Prof. Salvatella's experience in representing unstructured proteins by refining their structure with NMR data: in analogy with IDPs, only an ensemble of conformations describes the heterogeneity of the chemically denatured ubiquitin. Prof. Salvatella provided us with the starting seeds to recreate by mean of MD simulation the correct urea / unfolded ubiquitin system. After the validation against the available experimental observables, the ensemble generated by our atomistic MD simulations became very useful to dissect the nature of the unfolded state of the protein and its solvent environment.

The **second project (Section 4.2)** is a high-throughput study that aimed to overcome the heterogeneity of the unfolding pathway by collecting simulations for all the most prevalent meta-folds in the Protein Data Bank (PDB). The analysis identifies common patterns that proteins experience during the early stages of the denaturation process. This system not only allowed us to scale the results derived from the first projects to reach a proteome level but also to investigate the kinetic role of urea in trigger the unfolding.

Bibliography Chapter 4

[1] H. Wu, "Studies on Denaturation of Proteins XIII. A Theory of Denaturation†," Adv. Protein Chem. - ADVAN PROT CHEM, vol. 46, pp. 6–26, 1995.

[2] A. E. Mirsky and L. Pauling, "On the Structure of Native, Denatured, and Coagulated Proteins," Proc. Natl. Acad. Sci. U. S. A., vol. 22, no. 7, pp. 439–447, Jul. 1936.

[3] C. Tanford, "Protein denaturation," Adv. Protein Chem., vol. 23, pp. 121–282, 1968.

[4] K. A. Dill and H. S. Chan, "From Levinthal to pathways to funnels," Nat. Struct. Mol. Biol., vol. 4, no. 1, pp. 10–19, Jan. 1997.

[5] K. A. Dill and D. Shortle, "Denatured States of Proteins," Annu. Rev. Biochem., vol. 60, no. 1, pp. 795–825, 1991.

[6] R. F. Greene and C. N. Pace, "Urea and guanidine hydrochloride denaturation of ribonuclease, lysozyme, alpha-chymotrypsin, and beta-lactoglobulin," J. Biol. Chem., vol. 249, no. 17, pp. 5388–5393, Sep. 1974.

[7] D. R. Canchi and A. E. García, "Cosolvent effects on protein stability," Annu. Rev. Phys. Chem., vol. 64, pp. 273–293, 2013.

[8] J. L. England and G. Haran, "Role of Solvation Effects in Protein Denaturation: From Thermodynamics to Single Molecules and Back," Annu. Rev. Phys. Chem., vol. 62, no. 1, pp. 257–277, 2011.

[9] L. Hua, R. Zhou, D. Thirumalai, and B. J. Berne, "Urea denaturation by stronger dispersion interactions with proteins than water implies a 2-stage unfolding," Proc. Natl. Acad. Sci. U. S. A., vol. 105, no. 44, pp. 16928–16933, Nov. 2008.

[10] F. Gabel, M. R. Jensen, G. Zaccaï, and M. Blackledge, "Quantitative modelfree analysis of urea binding to unfolded ubiquitin using a combination of small angle X-ray and neutron scattering," J. Am. Chem. Soc., vol. 131, no. 25, pp. 8769–8771, Jul. 2009.

4.1. The urea-unfolded state of ubiquitin (Publication 1)

Our colleagues at the laboratory of Molecular Biophysics in our institute had developed a method to structurally characterize disordered proteins using RDC data as refinement. Its first application was on the protein ubiquitin, denatured in urea 8M (reference [30] in the article). The method, called ERIDU (ensemble refinement of intrinsically disordered and unstructured molecules), finally overcomes the use of a single, average alignment tensor that, as discussed in Chapter 1, usually hampers the RDC refinement of flexible structures. The single tensor approach, in fact, neglects the rapid change in the shape, typical of unfolded proteins, which modifies the alignment with the laboratory framework. ERIDU instead efficiently computes the individual alignment tensor of each ensemble member, leading to more accurate results. In the case of the unfolded ubiquitin ERIDU allowed to select an ensemble of 100 structures that all together describe the denatured state of ubiquitin. However the presence of urea wasn't specifically addressed in none of the refinement stages, lacking the details that should emerge from accounting the protein-solvent interactions.

We decided then to explore the effect of the urea solution on the ERIDU-ensemble by using the 100 structures as starting seeds for as many independent MD simulations with explicit solvent **(Figure 4.4)**. We immersed the structures in either 8M urea solution or in water, the latter used as control, and as a way to learn on the first stages of refolding. Two new ensembles were then created by the structures collected during the MD runs, called respectively MD UREA and MD WATER. We first address the validity of these ensembles calculating observables, such as the J-coupling and SAXS curve, and comparing them with the ones available experimentally. Despite the fact that individual trajectories have deviated significantly from seed points, the MD UREA ensemble reproduces fairly well all the data without sacrificing the plasticity expected for a denatured state. On the other hand, MD WATER quickly deteriorates the agreement, moving towards a more compact "native-like" structure. These results indicated that the force-field was able to reproduce the expected

behavior of ubiquitin in the two solvents, and that the simulation time is enough as to reproduce a significant amount of movement in the protein.

At this point, we moved into the study of the interactions between solvent molecules and the unfolded protein in the MD UREA ensemble. We found that the unfolded state of a protein is more "ureaphilic" than the native globular state, attracting a large proportion of urea molecules in the protein surroundings, mostly due to Van der Waals interactions, especially between apolar side chains and urea. To investigate the role of the hydrogen bonds, we employed QM calculation (Bader's atom in molecules analysis) on the skeleton of hydrogen bonds taken from our MD UREA ensemble. We conclude that, although hydrogen bonds exist and contribute to the stabilization of the denatured protein, they unlikely represent the differential factor and the principal driving force of unfolding. Overall dispersion, rather than electrostatic, is the main energetic contribution that keeps the protein unfolded in the presence of urea.

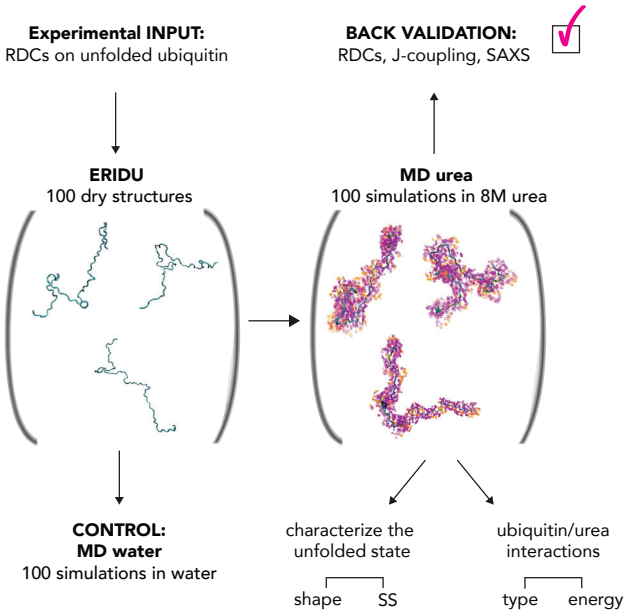


Figure 4.4. Schematic overview of the Urea-UBQ project. Experimental data were used as input for MD simulation and as post-validation. Control simulations were also perfomed in water. Once validated, the simulations were used to adress several aspect of urea-unfolding, as indicated.

Toward an atomistic description of the urea-denatured state of proteins

Michela Candotti[a,b], Santiago Esteban-Martín[a,b], Xavier Salvatella[a,b,c,1], and Modesto Orozco[a,b,d,1]

[a]Molecular Modelling and Bioinformatics Unit, Institute for Research in Biomedicine Barcelona, 08028 Barcelona, Spain; [b]Computational Biology Program, Barcelona Supercomputing Center, 08034 Barcelona, Spain; [c]Institució Catalana de Recerca i Estudis Avançats, Barcelona, Spain; and [d]Departament de Bioquímica, Facultat de Biologia, University of Barcelona, 08028 Barcelona, Spain

Edited by Arieh Warshel, University of Southern California, Los Angeles, CA, and approved February 26, 2013 (received for review September 24, 2012)

We present here the characterization of the structural, dynamics, and energetics of properties of the urea-denatured state of ubiquitin, a small prototypical soluble protein. By combining state-of-the-art molecular dynamics simulations with NMR and small-angle X-ray scattering data, we were able to: (*i*) define the unfolded state ensemble, (*ii*) understand the energetics stabilizing unfolded structures in urea, (*iii*) describe the dedifferential nature of the interactions of the fully unfolded proteins with urea and water, and (*iv*) characterize the early stages of protein refolding when chemically denatured proteins are transferred to native conditions. The results presented herein are unique in providing a complete picture of the chemically unfolded state of proteins and contribute to deciphering the mechanisms that stabilize the native state of proteins, as well as those that maintain them unfolded in the presence of urea.

denaturing mechanism | protein unfolding | random coil | ensemble simulation

It has been known for decades that protein structure is highly dependent on the solvent, and that certain chemical compounds induce protein destabilization and eventually unfolding (1). Biochemistry textbooks (2) state that under high denaturant concentration proteins adopt a "random coil" conformation, but very little is explained on the nature of such state. In fact, it has not yet been described how different the urea random coil is from the ensemble of conformations sampled by the unfolded state under native conditions and, even more important, no consensus exists on the physicochemical mechanisms explaining chemical denaturalization.

Urea is probably the most used chemical denaturant (3), but after decades of study there are still many unknowns in its mechanism of action. Within "direct" theory, urea denatures proteins by direct interaction with protein residues. These interactions are supposed to be stronger than those occurring with water, which would explain that groups not exposed to solvent in aqueous solution become exposed when urea is present. Several variants of the direct theory have been put forward. Thus, some authors have suggested that the major destabilizing effect of urea is related to its preferential interaction with the protein backbone (4, 5), side-chains (6), polar or charged residues (7), hydrophobic residues (8–10), or a mixture of hydrophobic and polar residues (11). The physical nature of the direct interactions between proteins and urea is also subject of debate, with some authors suggesting that it is mostly electrostatic and related to the formation of direct hydrogen bonds (5, 7, 12, 13), and others suggest that dispersion interactions are the main factor (14, 15). Some authors supported the idea the denaturing role of urea is not related to the formation of direct urea–protein interactions, but to its ability to "dry" the protein, weakening the hydrophobic effect responsible for stabilizing protein structures (16–19). However, recent consensus is that this indirect mechanism is not the main explanation of the effect of urea (18–20), pointing instead to the direct mechanism.

Massive experimental efforts have been directed to characterizing the nature of the urea-unfolded state of proteins (3, 21–23), because this is expected to be instrumental for the understanding of protein folding (24). However, structural experimental techniques face great difficulties to characterize the unfolded ensemble, mostly a result of its large conformational flexibility. Only recently low- and medium-resolution models of the unfolded state have been derived from spectroscopic techniques, such as small angle X-ray scattering data (SAXS) or NMR (25, 26).

The history of simulation techniques (mostly molecular dynamics, MD) in the field of chemical unfolding of proteins starts in the late 1990s (27, 28). Such pioneering works faced many problems, the most important one being the vast difference between the time scale of the unfolding process and that accessible to simulation. Even now, 15 y later, plain MD simulations are still limited to reproducing the early stages of unfolding for most proteins (10, 14), forcing the use of advanced sampling techniques that bias trajectories to populate unfolded states, a strategy that has provided spectacular results (13, 29) but that has obvious bias risks in terms of the reliability of the sampled unfolded state and of the unfolding pathway.

In this article we overcome the intrinsic time-dependent problems of MD to describe urea-unfolded proteins and the risks of biasing techniques by running multiple unrestrained simulations started from representative conformations of the unfolded state, as defined by NMR data collected under denaturing conditions (pH 2.0 and 8 M urea concentration): the ERIDU [Ensemble Refinement of Intrinsically Disordered and Unstructured proteins (30)]. MD simulations validated by NMR and SAXS data allowed us to describe with unprecedented accuracy the structural and physico-chemical properties of the unfolded state of a protein in urea and to advance the understanding of the mechanisms of protein folding and unfolding.

Methods

Starting Configurations. We used a finite set of structures collected in the ERIDU ensemble of ubiquitin (30) as starting configurations for our simulations. In short, this ensemble contains 100 different conformers of the protein, which were determined by refinement of a statistical coil model (31) using residual dipolar couplings (RDCs) as restraints. The 100-member ensemble was found sufficient to properly describe the RDCs and known residual native contacts.

System Set-Up. Proteins were titrated to pH 2.0 using MDWEB procedures (32), and immersed in a pre-equilibrated box of water/urea (28), adjusting the concentration of denaturant to 8 M (matching the experimental denaturing conditions). After test calculations with different urea force-field models, we selected the widely used refined parameters from the Optimized Potentials for Liquid Simulations (OPLS) forcefield, (5, 8–11, 13, 15, 20, 28). Note, however, that although incorrect models can yield to biased results (9, 33) most current urea models provide very similar results (34, 35). Chloride ions were added to keep electroneutrality. All systems were then energy-minimized and preequilibrated by MD for 8 ns, keeping the backbone restrained by intramolecular harmonic potentials. This unusually large

Author contributions: M.C., S.E.-M., X.S., and M.O. designed research; M.C. performed research; M.C. and S.E.-M. analyzed data; and M.C., S.E.-M., X.S., and M.O. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. E-mail: modesto.orozco@irbbarcelona.org or xavier.salvatella@irbbarcelona.org.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1216589110/-/DCSupplemental.

preequilibration period was required to avoid artifactual movements of the backbone related to an incorrect solvent arrangement around the protein (Fig. S1). Preequilibrated structures were then relaxed by removing backbone constraints during a 1-ns equilibration, which was then followed by 10-ns production runs for the 100 conformations in the ERIDU ensemble. Additionally, 100-ns simulations for the 10 most dissimilar structures in the ERIDU ensemble were performed to check the effect of extending simulation times.

In parallel to urea simulations we performed control MD simulations in water. Thus, proteins were solvated in a water box using the TIP3P model and chloride ions to keep neutrality of the system. Solvated systems were minimized, preequilibrated, and equilibrated using a protocol identical to that used for the urea/water simulations, except for the shorter preequilibration in water because 2 ns were enough to equilibrate the solvent. Production trajectories were collected for 10 ns under the same simulation conditions used for urea simulations.

Bader's Atoms in Molecule Analysis. To examine the electronic structure of the hydrogen bonds between protein backbone and solvent, a set of four representative structures (of protein–water and protein–urea complexes) were taken from the MD trajectories. The starting geometries were reduced to protein backbone atoms of the residue involved in hydrogen bonds with urea or water molecules. These geometries were optimized at the MP2(full)/6–311++G** (6 d, 10 f) level of theory using Gaussian09 (36, 37). The topological analysis of the electron distribution was performed with the AIMPAC package (38).

Trajectories Analysis. Analysis was performed using visual molecular dynamics (39) as well as in-house programs and the analysis tools, most of them available at the MDWEB application (32). Secondary structure was evaluated using STRIDE software (40). Further details are provided in *SI Methods*.

Trajectory Validation. The scalar couplings were calculated from the ensemble by using the Karplus equation as shown elsewhere (22) and compared with experimentally measured values. The RDCs were calculated as described in Esteban-Martín et al. (30). For this process the alignment tensor of each conformation was computed explicitly using the method developed by Almond and Axelsen (41) and the RDCs were averaged linearly. To account for the absolute degree of alignment the ensemble-averaged RDCs were globally scaled to the experimentally measured RDCs. The SAXS profile was computed using Crysol (42) software with default parameters.

Results

Validation of MD Simulations. Recent refinement of force-fields guarantees that MD simulations reproduce well the folded state of proteins (43, 44), but there is not such a guarantee for the unfolded state, especially in the presence of chemical denaturants. A first step should be then to check the quality of the MD ensembles by direct comparison with experimental observables. Table 1 presents a comparison between experimental and calculated NMR parameters for the ERIDU and MD simulations performed (MD simulations were performed under the same conditions used to measure the experimental data). The agreement between experimental and calculated NMR parameters, which include three-bond scalar couplings (^{3}J) and RDCs, was quantified using the Spearman correlation coefficient (ρ). As shown in Table 1, MD simulations in 8 M urea provide an accurate description of ^{3}J scalar couplings of ubiquitin (Fig. S2), with a ρ in fact slightly better than that obtained for the reference ERIDU ensemble (0.70 Hz; no restraints based on ^{3}J were introduced to derive the ERIDU structures). For the case of RDCs, we find a good correlation between calculated and measured RDCs ($\rho = 0.80$) (Table 1 and Fig. S3), in fact only slightly worse than that of the ERIDU ensemble ($\rho = 0.98$), where RDCS where explicitly restrained.

Table 1. Comparison between calculated and experimentally measured parameters for ubiquitin at pH 2.0 and 8 M urea

Ubiquitin and urea	^{3}J couplings ρ	RDCs ρ	SAXS χ^2
ERIDU	0.64	0.98	2.1
MD urea	0.66	0.72	1.4
MD urea (all trajectories)	0.64	0.80	1.4
MD water	0.51	0.61	9.8

To further check the quality of the MD ensembles in 8 M urea, we analyzed their ability to reproduce coarse grained experimental observables derived from SAXS experiments performed using identical conditions to those of our simulations (25). The agreement between MD-results (postprocessed using the Crysol protocol) and experiment is very good ($\chi^2 = 1.4$) (Fig. S4), even improving the good behavior of the original ERIDU ensemble ($\chi^2 = 2.1$; no SAXS restraints were included in ERIDU definition). Note that no experimental restraints or constraints were imposed in our MD simulations, which means that the agreement with these experimental observables should be interpreted as an independent validation of our trajectories.

In summary, unbiased MD ensembles obtained in strong denaturing conditions (8 M urea pH 2) satisfactorily reproduce fairly well all available experimental data on the chemical unfolded state of ubiquitin. It is however unclear whether or not this behavior is just reflecting the good quality of the original ERIDU ensemble, which was not much deteriorated in multiple 10-ns simulations as those considered here (see below). To check this point we computed all SAXS and NMR observables using an MD ensemble obtained under identical MD conditions and started from the same conformations, but in pure aqueous solvent. If the observed agreement between calculated and measured SAXS and NMR observables stems simply from memory artifacts, we should find similar behavior in urea and aqueous simulations. As noted in Table 1, it is however clear that MD-water simulations deteriorate the quality of the original ensembles (correlation with experimental RDCs and ^{3}J scalar couplings decreases to 0.61 and 0.51, respectively, and χ^2 for the fitting to SAXS curve increases up to 9.8) in terms of experimental observables of the urea unfolded state, suggesting that we are not facing a memory artifact. To further reject this possibility and to detect any other artifact related to the limited length of the simulations, we extended the simulations of the 10 most distinct structures in the ERIDU ensemble to 100 ns (*Methods*). Results in Table S1 demonstrate the lack of significant artifacts and the very close similarity, in terms of structural and experimental parameters of the ensembles obtained by 10- and 100-ns trajectories. In summary, we are confident that our MD simulations capture well the nature of the urea–protein system, allowing us to analyze the ensemble as a reliable representation of the conformational space of chemically denatured ubiquitin.

Characterization of the Unfolded Ensemble in Water and Water/Urea. The ERIDU ensemble is expected to capture reasonably well the nature of the urea-denatured state of ubiquitin, but the constituting structures have little sense individually, and might be in fact never populated. It is interesting then to analyze the evolution of the MD trajectories collected in 8 M urea using as a reference those obtained in pure aqueous solution. Fig. 1 demonstrates that structures sampled by MD simulations in urea are far from native structure, as anticipated by the ERIDU ensemble, showing backbone rmsds in the range 15–40 Å. MD-sampled structures in urea are extended, showing typical radii of gyration (Rg) in the range 20–40 Å (compared with 11.7 Å of the native protein), with an average value of 29.4 Å, matching the ERIDU distribution and average ($Rg_{ERIDU} = 29.9$ Å) and reproducing well the Rg derived from independent SAXS data (28–32 Å) (25). The macroscopic similarity between MD and ERIDU ensembles also becomes clear in Fig. 2*A*, which displays the average Rg and solvent accessible surface area (SASA) for the 100 ($\times$ 10 ns) MD trajectories and for the corresponding ERIDU structures. In summary, all macroscopic descriptors of the ERIDU and MD ensembles are very similar, but this does not imply that MD sampled structures match individual ERIDU conformations. Thus, as noted in Fig. 1, individual trajectories move around 10 Å in 10 ns from starting ERIDU, and more than 20 Å in 100-ns trajectories. These conformational movements, which as noted above do not modify experimental observables of the ensemble, often imply convergence

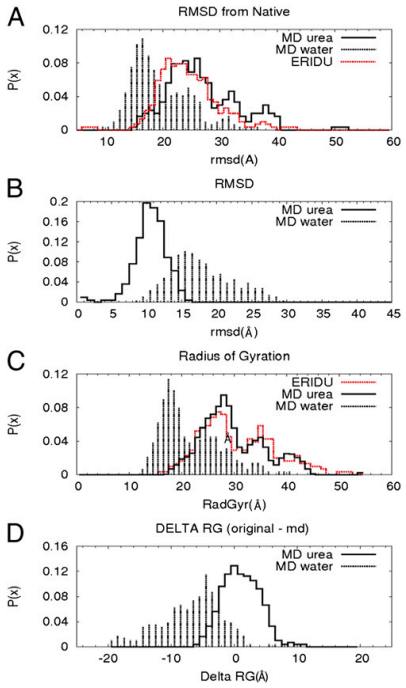


Fig. 1. Global conformational changes of the NMR conformational ensemble after MD simulations. (*A*) Backbone rmsd between native ubiquitin and final conformers of the ERIDU and MD ensembles. (*B*) Backbone rmsd between initial and final conformers of the ERIDU and MD ensembles. (*C*) Rg distributions for ERIDU and MD ensembles. (*D*) Difference in Rg between the initial and final conformers of the ERIDU and MD ensembles.

to other ERIDU structures (Fig. S5), indicating the dynamic nature of the unfolded ensemble.

Despite starting from the same seed conformations, the trajectories collected in pure water show a completely different behavior from those collected in the presence of urea. Thus, instead of navigating around the ERIDU representative structures, trajectories in water very quickly diverge (rmsd from 10 to 30 Å) (Fig. 1*B*), approaching to the native state (Fig. 1*A*). The major macroscopic effect of the transfer of the ERIDU ensemble to water is a dramatic

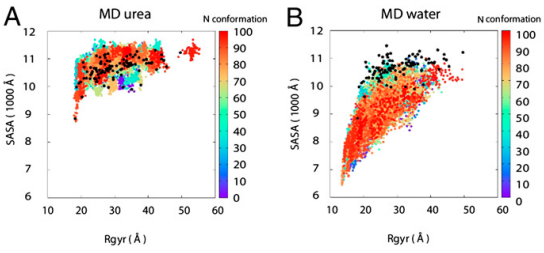


Fig. 2. Sampling of MD simulations. Rg and solvent accessible surface area SASA used as collective variables to visualize the sampling during the MD simulation of all of the ensemble (*A*) in urea and (*B*) in water. Black dots represent ERIDU starting conformations and frames belonging to the simulation of a particular conformation are plotted with the same color according to the color bar.

collapse of the structures, which becomes evident as a very large reduction of the Rg and of the SASA of the structures (Figs. 1*D* and 2*B*). A comparison of the bi-dimensional Rg vs. SASA plot in Fig. 2 for urea and water simulations clearly shows that although the former is typical of a stable unfolded ensemble, close to the seeding one (note that when 100-ns trajectories are used, no changes in the plots are observed) (Fig. S6), the latter is typical of a system in the first stages of folding, fast evolving toward a collapsed state akin to a molten globule. This finding is even clearer when looking at the conformation interchange plots in Fig. S5, which show that although in the urea simulations random movements and exchange along conformers occur, in the water simulations all trajectories tend to converge toward a similar collapsed state.

Residual Structural Elements. A long debate in studies dealing with unfolded proteins is whether or not residual structural elements of the native structure are preserved in the chemically denatured state, and whether or not, if they exist, these elements can act as nucleation points to guide refolding in the absence of denaturant. Very few amounts of secondary structure elements are detected experimentally (45), and MD-simulations confirm the lack of persistent elements of secondary structure in urea. In fact, we detected a very small (less than 2%) α-helical annotation spread along residues in both N- and C-terminal sections (Fig. 3), corresponding to short-transient motifs, which never propagate to well-defined stable elements of secondary structure. However, despite the lack of stable secondary elements, a few 3D native arrangements, are present in our simulations (Fig. 4), the most important is a native-like β-hairpin between (originally) β-strands 1 and 2 (residues MET1-GLU18). This contact was present in 12% of ERIDU ensemble (30), and it was detected around 9% of the time in our MD ensembles. Replication of a small number (13) of trajectories using randomizing velocities leads to no change in the population of β1-β2 contacts, and extension of 10 trajectories to 100 ns just increases slightly (to 12%) the population of this contact, matching ERIDU estimates and supporting claims by Meier et al. based on direct measures of transhydrogen-bond scalar couplings (45) (not introduced by any means as restraints in our simulations).

Trajectories in water lead to an increase in interresidue contacts, which start to signal (Fig. 4) some long-distance interactions resembling those occurring in the native state. The β1-β2 contact is reinforced because it is present in more than 35% of the aggregated simulation time in water (three times more than in urea). Interestingly, when the urea-unfolded structure is immersed in water a significant amount of secondary structure is generated. Thus, short turns are much better defined; the β-strand that was very marginal in urea is detected in 2–4% of the simulated time, and the much local α-helix conformation increases its population to 4%, with some segments sampling α-helix conformation nearly 10% of the simulation time (Fig. 3*A*). The locations of the segments forming secondary structures correlate (Fig. 3*A*), although not perfectly, with the regions of sequence where secondary structure elements exist in the native structure. It is clear that urea not only facilitates the expansion of the structure but also reduces dramatically the amount of secondary structure. Indeed, in the absence of urea, secondary structures form concomitantly with the hydrophobic collapse. Some contacts, like the β1-β2 contact (marginal in urea), become reinforced, acting as early-formed native contacts, which might help in guiding the refolding process toward a productive pathway once denaturant is removed.

Nature of Protein–Urea Interactions. As already anticipated by Fig. S1, chemically denatured ubiquitin captures very efficiently urea from the bulk solution (Table 2). For example, the ratio water/urea in the first solvation shell (FSS) is around 0.9, which compares with 5.4 in distant regions of the simulation box. The enrichment of urea in the FSS was previously reported in MD simulations of the early stages of the unfolding of urea (28), but the magnitude of such an enrichment is much larger than that found for folded forms of the protein (in fact, a control simulation performed for the folded protein in urea reveals a ratio water/urea in the FSS of only

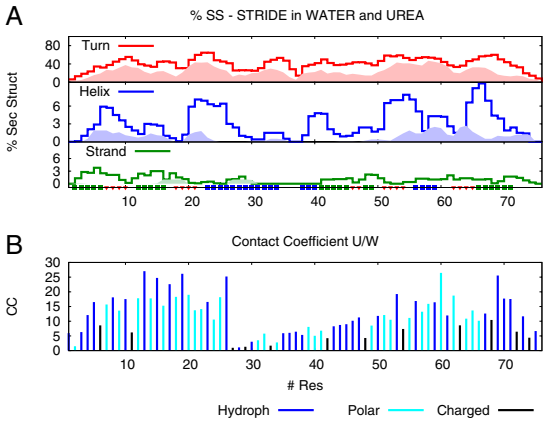


Fig. 3. Secondary structure and contacts with urea molecules. (*A*) Residual secondary structure calculated with STRIDE (40) (coil, turns, helixes, and strands) in MD simulations in urea (filled curve) and in water (line). (*B*) CC_{UW} along protein sequence. In both graph the secondary structure in the folded protein is shown on the *x* axis.

around 2.0) (Fig. S1), suggesting that those residues exposed when the protein is unfolded are those with higher preference to urea compared with water. Interestingly, in the unfolded ensemble apolar (A) and polar (P) residues were exposed to solvent to the same extent (Table S2), but urea molecules tend to localize closer to apolar residues (AP ratio 0.81) compared with water molecules (AP ratio 0.57) (Table 2). The differential solvation effect of urea is also evident in the contact coefficient (CC_{UW}) (9), which reaches a value of 14.0 for apolar and only 10.6 for polar residues. As anticipated by others (8–10, 14), hydrophobic residues are the main differential target for the preferential interaction of urea with unfolded conformations of the protein. Clearly, our simulations suggest that at least for this protein and in the context of a fully unfolded ensemble, the preferential solvation of apolar solutes by urea is the main stabilizing factor of the open state. As discussed by Mountain and Thirumalai (34), other balance of interactions can occur for other conformational states of the proteins.

Urea has been suggested to have a tendency to interact with specific structural motifs of proteins (14, 15). In the absence of clear cavity regions in the unfolded state of ubiquitin, we centered our analysis in the region showing residual α-helix content, which appear enriched ($P < 0.001$) in contacts with urea (Fig. 4*B* and Fig. S7), as well as in the vicinities of the contact β1-β2 findings; in this case there were no differences in urea interactions with respect to the background. It is difficult to determine how much of the improved urea binding found in regions with residual α-helixes is related to sequence bias, but it is clear that the presence of residual 3D motives is not always stabilized by a preferential binding of urea molecules.

Energetics of Protein–Urea Interactions. Previous studies have favored the direct mechanism to explain the denaturing effect of urea, but have not clarified what is the physical nature of protein–urea interactions: the electrostatic term (mainly hydrogen bonds) or the van der Waals contacts (mostly dispersive effects). It has been proposed by different authors that urea denaturation is driven by hydrogen bonding to the backbone or polar residues (7, 13), a hypothesis that seems to be supported by the structure of urea, which looks optimized to form hydrogen bonds with peptide backbones. To check whether this hypothesis is correct, we analyzed the occurrence of protein-solvent hydrogen bonds in our denatured ensemble (*Methods* and Table 3). Considering the structure of water and urea, we should expect two- (water) and four- (urea) times more hydrogen-bond donor than acceptor interactions with the protein, but in fact more events of water acting as hydrogen-bond acceptor than as a donor are found (ratio

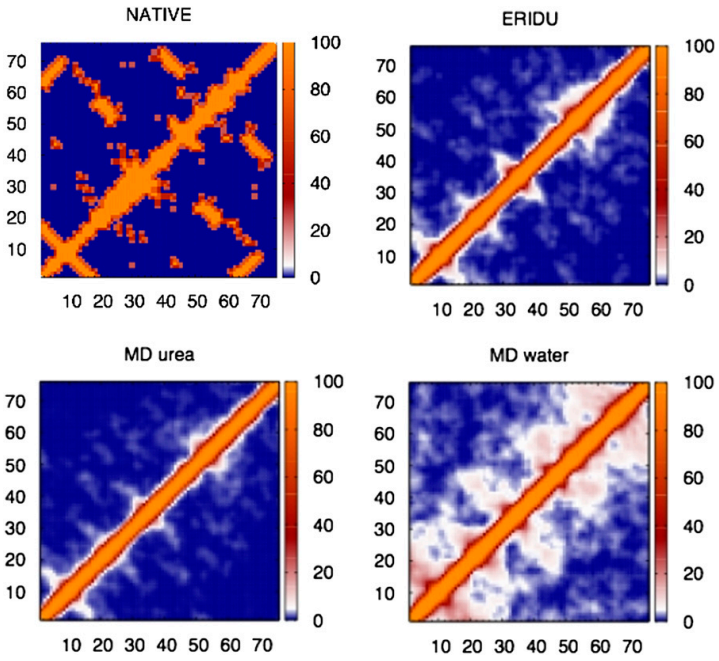


Fig. 4. Structural changes and stability of the hairpin. Contact maps for the folded ubiquitin, the initial (ERIDU) and final MD ensembles. A cut-off of 10 Å was used.

Table 2. Characterization of the urea/water-protein structural organization

Solvent descriptor	Average ± SD
R_{WU}* FSS	0.89 ± 0.18
R_{WU}* bulk	5.36 ± 0.31
AP ratio† - water	0.57 ± 0.08
AP ratio† - urea	0.81 ± 0.05
CC_{UW}‡ apolar	13.98 ± 7.03
CC_{UW}‡ polar	10.62 ± 5.44
CC_{UW}‡ backbone	7.73 ± 1.76
CC_{UW}‡ side-chains	5.39 ± 1.16

*R_{WU}, ratio of water/urea molecules in FSS (< 5 Å) and bulk (> 6 Å).
†AP ratio, fraction of solvent molecules close to apolar and to polar/charged residues in the FSS.
‡CC_{UW}, average contact coefficient for apolar, polar/charged residues and for backbone and side-chain atoms.

donor/acceptor for water equal to 0.63) (Table 3), and for the urea the ratio donor/acceptor is just 2.55 (instead of the expected 4) (Table 3), indicating that intrinsic acceptor capabilities of both urea and water are larger than their donor capabilities, even in the case of urea, the largest number of donors overcome this intrinsic preference. The fraction of hydrogen bonds formed by the solvent with the protein backbone is higher compared with side-chains, the difference being more pronounced in urea (~91% for backbone and ~9% for side-chain). In addition, the majority of hydrogen bonds formed by urea are made with apolar residues, but the main hydrogen-bond partners for water are the polar ones (Table 3). Our MD-derived results suggest then that urea hydrogen-bond capabilities contribute to stabilize exposed apolar residues directly by electrostatic interactions with the backbone, and indirectly by increasing the local density of urea around them, favoring the formation of short-range dispersion interactions. However, it is very unlikely that urea will direct protein unfolding through hydrogen bonding, as otherwise we should expect higher side-chain contacts and more abundant interactions with polar residues, where hydrogen-bond interactions are expected to be stronger.

Some authors (27, 46) have suggested that hydrogen bonds involving urea are intrinsically stronger than those involving water molecules and, therefore, that even a small number of hydrogen bonds can significantly stabilize the open state of the protein. To analyze this point with accuracy, we performed quantum mechanics Bader's analyses (47) of complexes of water and urea with a model of a peptide unit (*Methods*). As discussed elsewhere (48), the analysis of the hydrogen-bond critical points in quantum mechanical optimized geometries provides a direct unbiased measure of the intrinsic stability of hydrogen bonds. Results shown in Table 4 reveal that (when contacting an amide moiety) water intrinsically prefers to act as a hydrogen-bond donor than acceptor, contrary to what is found in MD simulations of water/urea mixtures, suggesting a certain frustration of hydrogen-bond capabilities of water in the FSS. For urea the situation is different, because the acceptor capabilities of the carbonyl group are much larger than the donor ones of the N-H bonds. Electron densities at the hydrogen-bond critical points and associated Laplacians clearly show that water is a much stronger hydrogen-bond donor than urea, but it is only slightly poorer than urea as an acceptor. Considering that there are 2.5-more hydrogen bonds with urea acting as donor than as acceptor (see above), we can rule out the hypothesis that urea-protein hydrogen bonds are intrinsically stronger than the water-protein ones.

We also tested the recent suggestion by Blackledge and colleagues (25) that urea hydrogen bonds to the protein backbone (with *x* number of urea molecules per residue), acting mostly (or exclusively) as acceptor. We found first a much smaller number of hydrogen-bond urea-protein contacts than that suggested by the authors, and in fact the hydrogen-bond donor role of urea is 2.5-times more prevalent than the role as an hydrogen-bond acceptor. These findings, which are very robust to simulation details, argue against a major effect of urea acceptor capabilities as a main reason for the stabilization of the unfolded form of the protein. However, to double-check this and to gain a deeper detail on the physical nature of urea–protein interactions we postprocessed our trajectories following the energy interaction scheme proposed by Hua et al. (14). For the case of water molecules (Fig. S8), the energy distributions obtained for the molecules in the FSS and in the bulk overlapped, suggesting that water is not really protein-philic (when the protein is unfolded in the presence of urea). The situation for urea is quite different; although the electrostatic term distributions in bulk and FSS overlap, indicating that hydrogen bonds and other polar contacts do not favor the migration of urea from bulk to the FSS, urea dispersive interactions are clearly better in the FSS than in the bulk. Clearly, as anticipated by others from unfolding trajectories of native proteins (14, 15, 29), direct dispersive interactions of urea, mostly with apolar residues of the protein, seem the major driving force for the stabilization of expanded conformations of ubiquitin. However, as noted above, hydrogen bonds and other polar contacts are not negligible, as they might be important to stabilize exposed polar moieties of the protein that will generate otherwise strong local fields, which would destabilize the unfolded state.

Table 3. Percentage of urea-protein and water-protein hydrogen bonds (HB) formed

Bond	Sites	No. HBs	Percent*	Partners (% of total)*			
					Protein	Apol	Pol
Urea H-donor	4	36,229	39	All		52	48
				SC	9	2	98
				BB	91	55	45
Urea H-acceptor	1	14,181	61	All		52	48
				SC	13	6	94
				BB	86	52	48
Ratio D/A	4	2.55					
Water H-donor	2	3,409	24	All		44	56
				SC	23	1	99
				BB	77	56	44
Water H-acceptor	1	5,364	76	All		36	64
				SC	25	5	95
				BB	75	44	56
Ratio D/A	2	0.63					

Apol, apolar; BB, backbone; Pol, polar/charged; Ratio D/A, Ratio donor/acceptor; SC, side-chains.
*Percent of total number, normalized according to the number of donor or acceptors sites.

Discussion and Conclusion

Urea plays two major roles in unfolding: (*i*) a kinetic effect reducing the free-energy barrier associated to protein unfolding, by favoring the disruption of the key elements of the native structure, and (*ii*) a thermodynamic effect, stabilizing extended forms of the protein, that will collapse toward the molten globule when

Table 4. Bader's atom in a molecule analyses of hydrogen bonds

Water or urea	$\rho \times 10^{2}$*	$\nabla^2\rho \times 10^{2}$†
Water		
H-Donor	1.88	8.51
H-Acceptor	1.46	6.47
Urea		
H-Donor	1.15	5.04
H-Acceptor	1.69	6.79

*Electron density at the bond critical points.
†Associated Laplacian.

transferred to water. Most MD simulations published to date have investigated the effect of urea in the unfolding mechanism, giving very valuable information on the effect of urea in the early stages of protein unfolding but providing little information on the nature of urea–protein interactions in the fully unfolded state. The present study, where the seed for MD simulation is not the native structure but a NMR-derived ensemble of the unfolded state of the protein provides a complementary picture of the effect of urea in protein unfolding, giving direct information on the role of urea in stabilizing the unfolded state of the protein.

The protein in 8 M urea is fully extended and flexible, making no attempts to recover native-like 3D structure. We found structural and flexibility patterns for the protein in urea that fit with the concept of "random coil" outlined in biochemistry textbooks (49). When the chemically denatured ensemble is suddenly introduced in water, a fast hydrophobic collapse occurs, outlining sections of secondary structure (mostly inexistent in urea) and defining some near-native 3D contacts, which might act as the seed for regeneration of the native structure. Clearly, the chemically unfolded state seems quite different to the "unfolded" state in aqueous solution.

The unfolded protein is more urea-philic than hydrophilic, and captures very efficiently denaturant molecules to the FSS. Urea forms abundant hydrogen bonds with the protein backbone but detailed analysis of urea population reveals that apolar residues are the main targets for urea-specific solvation and that dispersion, rather than electrostatic interactions, is the main energetic contribution to explain the stabilization of the unfolded state of the protein and the irreversibility of the unfolding process in the presence of urea. Whether or not urea uses the same physicochemical mechanism to accelerate the kinetics of unfolding is an intriguing issue that will require further investigation.

ACKNOWLEDGMENTS. This work has been supported by Grants CTQ2009-08850 (to X.S.) and BIO2012-32868 (to M.O.) from the Ministerio de Economía y Competitividad-Spain; Grant 102030 from Marató de TV3 (to X.S.); the European Research Council-Advanced grant (to M.O.); the Instituto Nacional de Bioinformatica (M.O.); the Consolider E-Science Project (M.O.); the Framework VII Scalalife Project (M.O.); and the Fundación Marcelino Botin (M.O.). M.C. is a fellow of the Spanish Ministries of Education; S.E.-M. is a Juan de la Cierva researcher; and M.O. is an Institució Catalana de Recerca i Estudis Avançats Academia Researcher.

1. Kauzmann W (1959) Some factors in the interpretation of protein denaturation. *Adv Protein Chem* 14:1–63.
2. Lehninger A-L, Cox M-M, Nelson D-L (2008) *Lehninger Principles of Biochemistry* (W.H. Freeman, New York), 5th Ed, pp 140–142.
3. England JL, Haran G (2011) Role of solvation effects in protein denaturation: From thermodynamics to single molecules and back. *Annu Rev Phys Chem* 62:257–277.
4. Auton M, Holthauzen LM, Bolen DW (2007) Anatomy of energetic changes accompanying urea-induced protein denaturation. *Proc Natl Acad Sci USA* 104(39): 15317–15322.
5. Klimov DK, Straub JE, Thirumalai D (2004) Aqueous urea solution destabilizes Abeta (16-22) oligomers. *Proc Natl Acad Sci USA* 101(41):14760–14765.
6. Canchi DR, García AE (2011) Backbone and side-chain contributions in protein denaturation by urea. *Biophys J* 100(6):1526–1533.
7. O'Brien EP, Dima RI, Brooks B, Thirumalai D (2007) Interactions between hydrophobic and ionic solutes in aqueous guanidinium chloride and urea solutions: Lessons for protein denaturation mechanism. *J Am Chem Soc* 129(23):7346–7353.
8. Stumpe MC, Grubmüller H (2007) Interaction of urea with amino acids: Implications for urea-induced protein denaturation. *J Am Chem Soc* 129(51):16126–16131.
9. Stumpe MC, Grubmüller H (2008) Polar or apolar—The role of polarity for urea-induced protein denaturation. *PLOS Comput Biol* 4(11):e1000221.
10. Stumpe MC, Grubmüller H (2009) Urea impedes the hydrophobic collapse of partially unfolded proteins. *Biophys J* 96(9):3744–3752.
11. Mountain RD, Thirumalai D (2003) Molecular dynamics simulations of end-to-end contact formation in hydrocarbon chains in water and aqueous urea solution. *J Am Chem Soc* 125(7):1950–1957.
12. Lim WK, Rösgen J, Englander SW (2009) Urea, but not guanidinium, destabilizes proteins by forming hydrogen bonds to the peptide group. *Proc Natl Acad Sci USA* 106(8):2595–2600.
13. Berteotti A, Barducci A, Parrinello M (2011) Effect of urea on the β-hairpin conformational ensemble and protein denaturation mechanism. *J Am Chem Soc* 133(43): 17200–17206.
14. Hua L, Zhou R, Thirumalai D, Berne BJ (2008) Urea denaturation by stronger dispersion interactions with proteins than water implies a 2-stage unfolding. *Proc Natl Acad Sci USA* 105(44):16928–16933.
15. Lindgren M, Westlund PO (2010) On the stability of chymotrypsin inhibitor 2 in a 10 M urea solution. The role of interaction energies for urea-induced protein denaturation. *Phys Chem Chem Phys* 12(32):9358–9366.
16. Rezus YL, Bakker HJ (2006) Effect of urea on the structural dynamics of water. *Proc Natl Acad Sci USA* 103(49):18417–18420.
17. Finer E, Franks F, Tait M (1972) Nuclear magnetic resonance studies of aqueous urea solutions. *J Am Chem Soc* 94(13):4424–4429.
18. Stumpe MC, Grubmüller H (2007) Aqueous urea solutions: Structure, energetics, and urea aggregation. *J Phys Chem B* 111(22):6220–6228.
19. Vanzi F, Madan B, Sharp K (1998) Effect of the protein denaturants urea and guanidinium on water structure: A structural and thermodynamic study. *J Am Chem Soc* 120(41):10748–10753.
20. Soper AK, Castner EW, Luzar A (2003) Impact of urea on water structure: A clue to its properties as a denaturant? *Biophys Chem* 105(2-3):649–666.
21. Pardi A, Billeter M, Wüthrich K (1984) Calibration of the angular dependence of the amide proton-C alpha proton coupling constants, 3JHN alpha, in a globular protein. Use of 3JHN alpha for identification of helical secondary structure. *J Mol Biol* 180(3): 741–751.
22. Guinn EJ, Pegram LM, Capp MW, Pollock MN, Record MT, Jr. (2011) Quantifying why urea is a protein denaturant, whereas glycine betaine is a protein stabilizer. *Proc Natl Acad Sci USA* 108(41):16932–16937.
23. Schellman JA (2002) Fifty years of solvent denaturation. *Biophys Chem* 96(2–3): 91–101.
24. Rösgen J, Pettitt BM, Bolen DW (2005) Protein folding, stability, and solvation structure in osmolyte solutions. *Biophys J* 89(5):2988–2997.
25. Huang JR, Gabel F, Jensen MR, Grzesiek S, Blackledge M (2012) Sequence-specific mapping of the interaction between urea and unfolded ubiquitin from ensemble analysis of NMR and small angle scattering data. *J Am Chem Soc* 134(9):4429–4436.
26. Bernadó P, Blackledge M (2009) A self-consistent description of the conformational behavior of chemically denatured proteins from NMR and small angle scattering. *Biophys J* 97(10):2839–2845.
27. Bennion BJ, Daggett V (2003) The molecular basis for the chemical denaturation of proteins by urea. *Proc Natl Acad Sci USA* 100(9):5142–5147.
28. Tirado-Rives J, Orozco M, Jorgensen WL (1997) Molecular dynamics simulations of the unfolding of barnase in water and 8 M aqueous urea. *Biochemistry* 36(24):7313–7329.
29. Canchi DR, Paschek D, García AE (2010) Equilibrium study of protein denaturation by urea. *J Am Chem Soc* 132(7):2338–2344.
30. Esteban-Martín S, Fenwick RB, Salvatella X (2010) Refinement of ensembles describing unstructured proteins using NMR residual dipolar couplings. *J Am Chem Soc* 132(13):4626–4632.
31. Jha AK, Colubri A, Freed KF, Sosnick TR (2005) Statistical coil model of the unfolded state: Resolving the reconciliation problem. *Proc Natl Acad Sci USA* 102(37): 13099–13104.
32. Hospital A, et al. (2012) MDWeb and MDMoby: An integrated web-based platform for molecular dynamics simulations. *Bioinformatics* 28(9):1278–1279.
33. Tsai J, Gerstein M, Levitt M (1996) Keeping the shape but changing the charge: A simulation study of urea and its iso-steric analogues. *J Chem Phys* 104(23):9417–9430.
34. Mountain RD, Thirumalai D (2004) Importance of excluded volume on the solvation of urea in water. *J Phys Chem B* 108(21):6826–6831.
35. Xiu P, et al. (2011) Urea-induced drying of hydrophobic nanotubes: Comparison of different urea models. *J Phys Chem B* 115(12):2988–2994.
36. Møller C, Plesset MS (1934) Note in an approximation treatment for many-electron system. *Phys Rev* 46(7):618–622.
37. Frisch MJ, et al. (2009) Gaussian 09, (Revision A.1, Inc., Wallingford CT). Available at http://www.gaussian.com/g_tech/g_ur/m_citation.htm.
38. Biegler-König FW, Bader RFW, Tang TH (1982) Calculation of the average properties of atoms in molecules. II. *J Comput Chem* 13(2):317–328.
39. Humphrey W, Dalke A, Schulten K (1996) VMD: Visual molecular dynamics. *J Mol Graph* 14(1):33–38, 27–28.
40. Frishman D, Argos P (1995) Knowledge-based protein secondary structure assignment. *Proteins* 23(4):566–579.
41. Almond A, Axelsen JB (2002) Physical interpretation of residual dipolar couplings in neutral aligned media. *J Am Chem Soc* 124(34):9986–9987.
42. Svergun DI, 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.
43. Rueda M, et al. (2007) A consensus view of protein dynamics. *Proc Natl Acad Sci USA* 104(3):796–801.
44. Lindorff-Larsen K, et al. (2012) Systematic validation of protein force fields against experimental data. *PLoS ONE* 7(2):e32131.
45. Meier S, Strohmeier M, Blackledge M, Grzesiek S (2007) Direct observation of dipolar couplings and hydrogen bonds across a beta-hairpin in 8 M urea. *J Am Chem Soc* 129(4):754–755.
46. Holthauzen LM, Rösgen J, Bolen DW (2010) Hydrogen bonding progressively strengthens upon transfer of the protein urea-denatured state to water and protecting osmolytes. *Biochemistry* 49(6):1310–1318.
47. Bader R (1991) A quantum theory of molecular structure and its applications. *Chem Rev* 91(5):893–928.
48. Tang TH, Deretey E, Knak Jensen SJ, Csizmadia IG (2006) Hydrogen bonds: Relation between lengths and electron densities at bond critical points. *Eur Phys J D* 37:217–222.
49. Smith LJ, Fiebig KM, Schwalbe H, Dobson CM (1996) The concept of a random coil. Residual structure in peptides and denatured proteins. *Fold Des* 1(5):R95–R106.

Supporting Information

Candotti et al. 10.1073/pnas.1216589110

SI Methods

System Set-Up. Final system sizes ranges from 40,000 and 150,000 atoms with simulation boxes around one million $Å^3$ (minimum 550.000, maximum 2.100.000 $Å^3$), large enough as to guarantee the lack of close contacts with images (shortest protein-protein image contacts above 15 Å).

Simulation Details. Protein force-field parameters were taken from latest modification of Amber parm99 (P99 SBILDN) (1) and, for consistency with this force-field, urea parameters were taken from Smith et al. (2), which refined previous an Optimized Potentials for Liquid Simulations (OPLS) urea model, and the water model was TIP3P (3). All production runs were carried out in NVT ensemble (T = 300 K) using periodic boundary conditions and the Particle Mesh Ewald procedure to introduce long-range electrostatic effects (4). All bonds involving hydrogens were constrained using SHAKE (5), which allowed us to integrate Newton's equations of motion every 2 fs. Trajectories were obtained using NAMD (6) for equilibration, and the AceMD program (7) on graphical processing units at the Barcelona Supercomputing Center (Minotauro Supercomputer) and Institute for Research in Biomedicine, Barcelona, for production.

Trajectory Analysis. Contact maps were calculated using distances between α-carbon atoms and a cut-off of 10 Å. Structures containing a hairpin were identified using the crystal structure of ubiquitin as reference (PDB ID code 1UBQ) and calculating the rmsd for residues MET1 to GLU18, the cut-off used for the rmsd was 6 Å. The water/urea ratio (R_{WU}) was calculated in the first solvation shell (FSS), defined as solvent molecules within 5 Å of the protein, and in the bulk, defined as solvent molecules with a distance larger than 6 Å from any atom of the protein. The apolar/polar (AP) ratio was determined as the fraction of solvent molecules close to apolar and polar/charged residues for each solvent species in the FSS. The contact coefficient (CC_{UW}) for amino acid x was calculated as the ratio of the number of atomic contacts of amino acid x with urea to the number with water molecules (8). Hydrogen bonds were defined according to a cut-off of 3.5 Å for the distance between donor and acceptor atoms and 120° for the angle between donor-hydrogen and acceptor. To compute the energy of each urea or water molecule with the rest of the system in the FSS and in the bulk we used the same procedure presented in Hua et al. (9). The solvent accessible surface area has been calculated with NACESS (10).

1. Lindorff-Larsen K, et al. (2010) Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins* 78(8):1950–1958.
2. Smith LJ, Berendsen HJC, van Gunsteren WF (2004) Computer simulation of urea-water mixtures: A test of force field parameters for use in biomolecular simulation. *J Phys Chem B* 108(3):1065–1071.
3. Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML (1983) Comparison of simple potential functions for simulating liquid water. *J Chem Phys* 79(2):926–935.
4. Essmann U, et al. (1995) A smooth particle mesh Ewald method. *J Chem Phys* 103(19): 8577.
5. Kräutler V, van Gunsteren WF, Hünenberger PH (2001) A fast SHAKE algorithm to solve distance constraint equations for small molecules in molecular dynamics simulations. *J Comput Chem* 22(5):501–508.
6. Phillips JC, et al. (2005) Scalable molecular dynamics with NAMD. *J Comput Chem* 26(16):1781–1802.
7. Harvey C, Giupponi G, De Fabritiis G (2009) ACEMD: Accelerated molecular dynamics simulations in the microseconds timescale. *J Chem Theory Comput* 5:1632.
8. Stumpe MC, Grubmüller H (2008) Polar or apolar—The role of polarity for urea-induced protein denaturation. *PLOS Comput Biol* 4(11):e1000221.
9. Hua L, Zhou R, Thirumalai D, Berne BJ (2008) Urea denaturation by stronger dispersion interactions with proteins than water implies a 2-stage unfolding. *Proc Natl Acad Sci USA* 105(44):16928–16933.
9. Hua L, Zhou R, Thirumalai D, Berne BJ (2008) Urea denaturation by stronger dispersion interactions with proteins than water implies a 2-stage unfolding. *Proc Natl Acad Sci USA* 105(44):16928–16933.
10. Hubbard SJ, Thornton JM (1993) 'NACCESS', Computer Program, Department of Biochemistry and Molecular Biology, University College London.

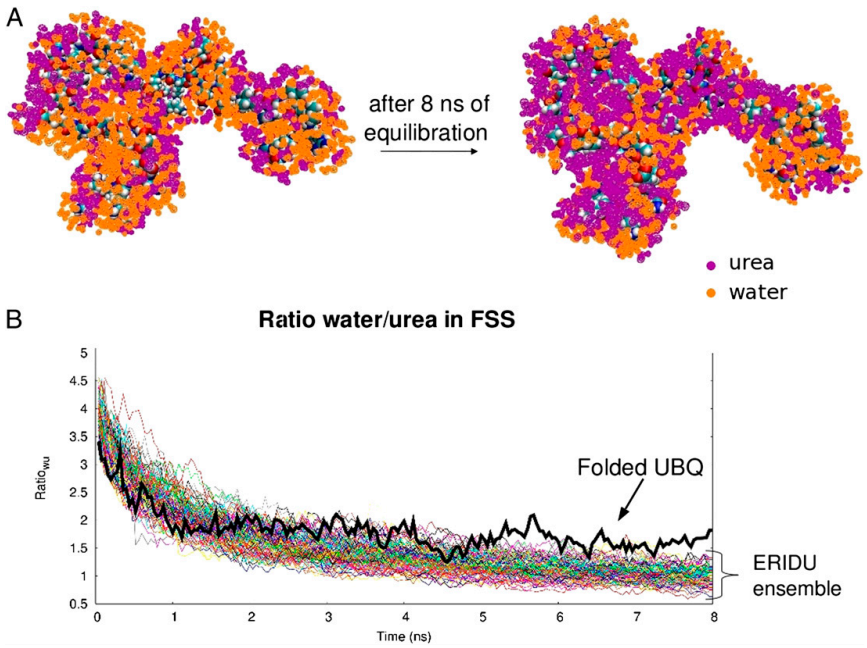


Fig. S1. Equilibration of the FSS. (*A*) The figure shows two snapshots corresponding to the starting (*Left*) and end (*Right*) points of the equilibration run (8 ns). Urea and water molecules in the FSS are shown, respectively, in purple and orange. (*B*) Time evolution of the water to urea ratio (R_{WU}) in the FSS during the equilibration period (8 ns). Native ubiquitin is shown in black line.

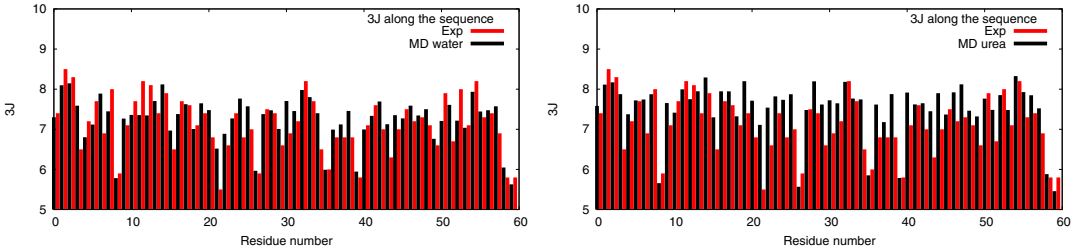


Fig. S2. Comparison with experimental ^3J scalar couplings. Experimental ^3J scalar couplings along the protein sequence (in red) are compared with those calculated from molecular dynamic (MD) simulations in water (*Left*) and urea (*Right*).

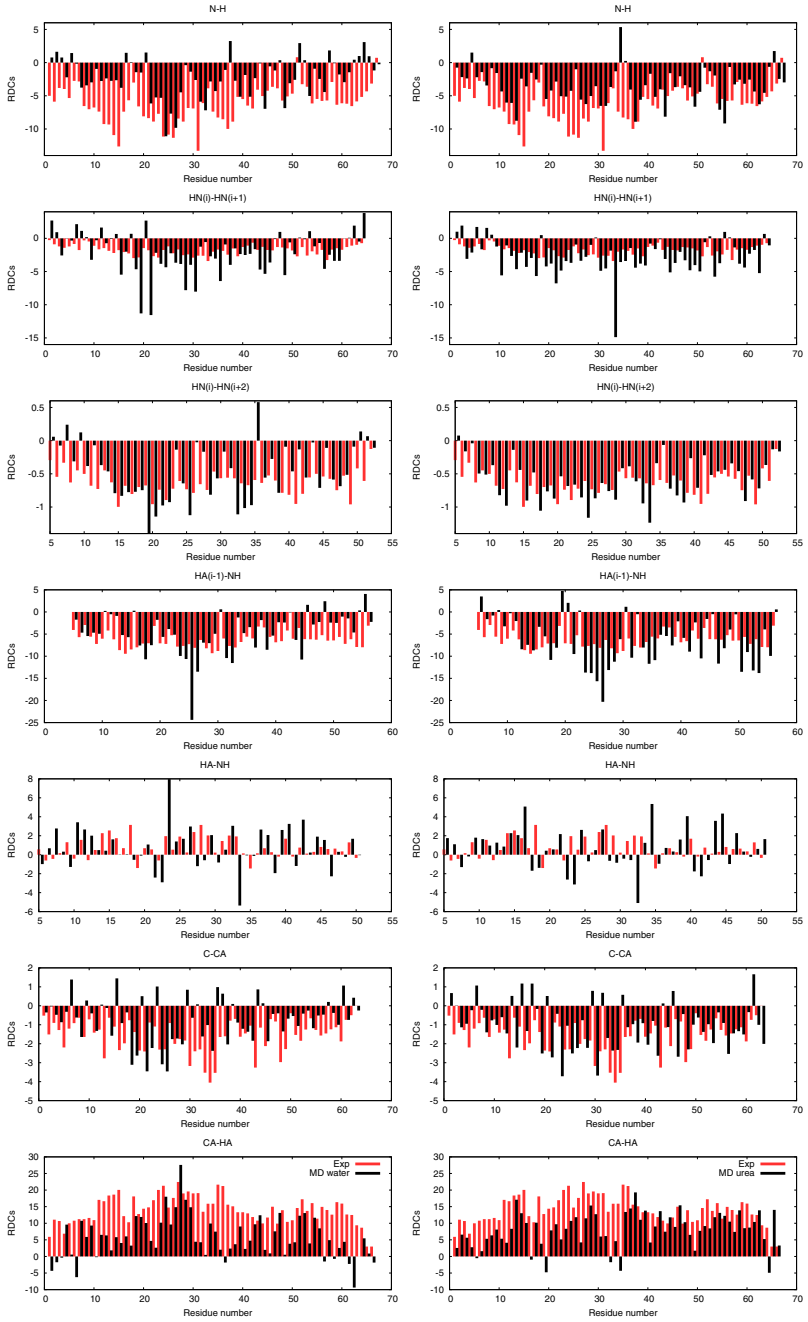


Fig. S3. Comparison with experimental residual dipolar couplings (RDCs). The different types of experimental RDCs along the protein sequence (in red) are compared with those calculated from MD simulations in water (*Left*) and urea (*Right*).

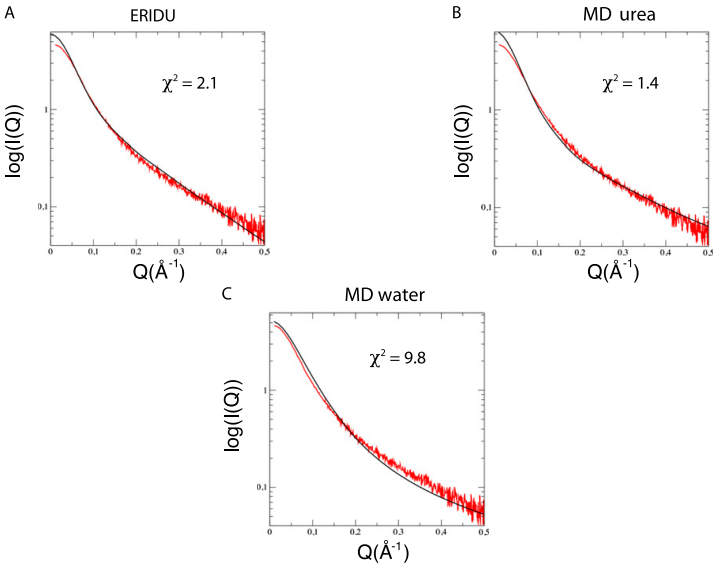


Fig. S4. Comparison of experimental and calculated small-angle X-ray scattering data (SAXS) curves. Experimental data (red line) are compared with curve calculated from (*A*) the ERIDU (Refinement of Intrinsically Disordered and Unstructured molecules ensemble) and (*B*) snapshots taken from MD simulation in urea and (*C*) in water.

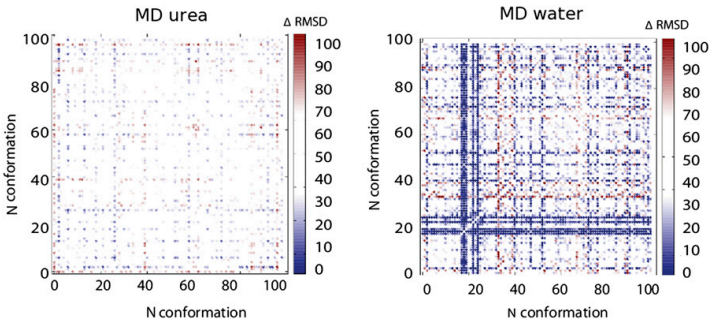


Fig. S5. Rmsd matrix between all of the conformations. Each dot represents the difference in rmsd before and after 10 ns of MD simulation between two conformations. A negative value (blue dots) show two conformations that are more similar after MD simulations, and positive values (red dots) imply that the conformation become less alike. Notice that although in the urea simulations random movements and interchange along conformers occur [noted as a mixture of cases of conformations becoming more similar (blue) and more different (red) dots], in the water simulations all trajectories tend to converge toward a similar collapsed conformation (blue dots massively dominating).

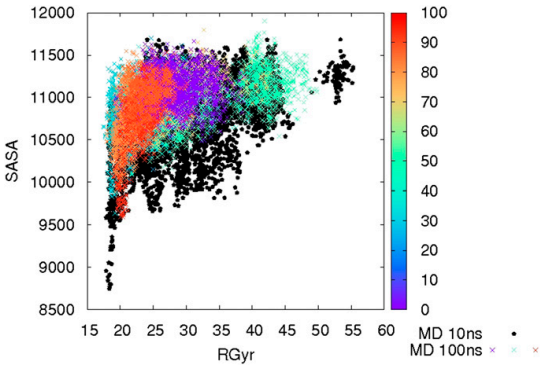


Fig. S6. MD sampling. Radius of gyration (RG) and solvent accessible surface area (SASA) used as collective variables to visualize the sampling during 100 ns of MD simulation for 10 conformations. Black dots represent the sampling of 100 × 10-ns trajectories; the 10 simulations of 100 ns are plotted with the different colors according to the conformation number.

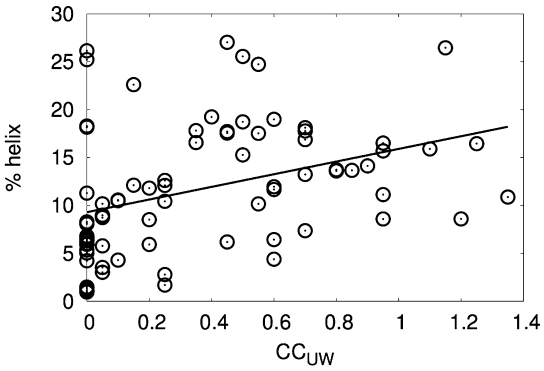


Fig. S7. Helical propensity and urea preference. Correlation between the percentage of helical propensity and the CC_{UW} at the residue level in MD urea simulations.

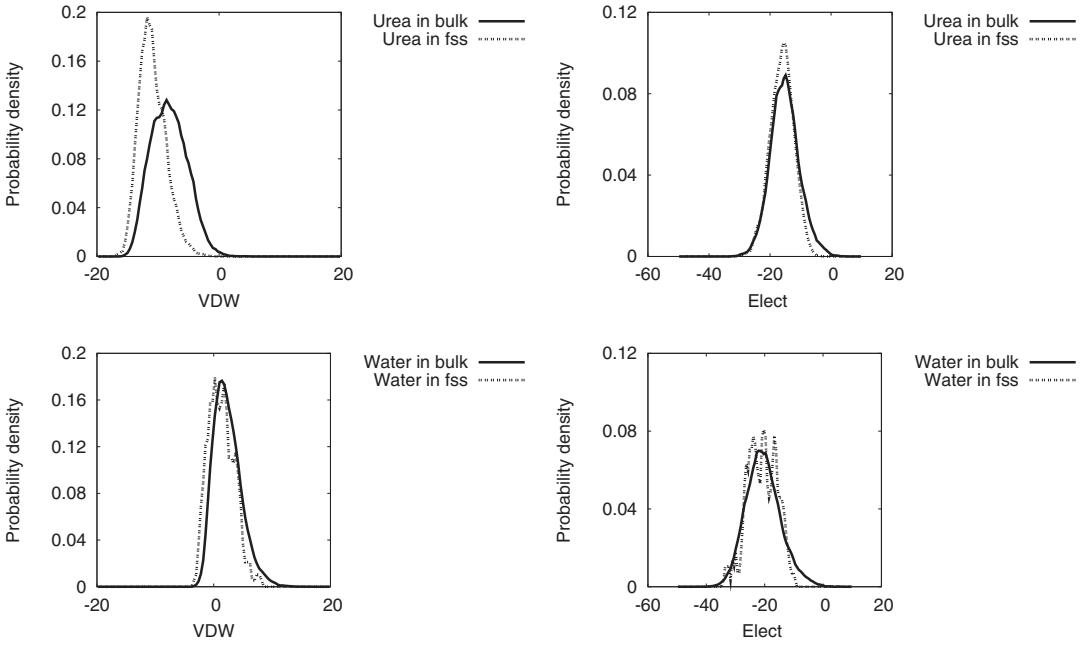


Fig. S8. Electrostatic and dispersion energies of the solvent. Probability distribution function of Van der Waals and electrostatic energy of urea and water in the FSS and in the bulk with the rest of the system.

Table S1. Conformation descriptors for the two ensemble of simulations (10 × 100 ns and 100 × 10 ns)

Conformation descriptor	10 × 100 ns	100 × 10 ns
SASA (Å^2)	10,810 ± 364	10,964 ± 297
Radius of gyration (Å)	29.4 ± 6.42	27.8 ± 5.96
^{3}J agreement (ρ)	0.64	0.60
RDCs agreement (ρ)	0.80	0.78

Table S2. Average CC_{UW} calculated for polar and hydophobic residues and their SASA calculated; in addition the specific contribution of side-chains and backbone are reported

Residue	CC_{UW}	SASA total (Å^2)	SASA side-chain (Å^2)	SASA backbone (Å^2)
Polar and charged (40 residues)	10.322	148.918	126.886	22.0321
Hydrophobic (28 residues)	13.755	147.898	127.842	20.0565

4.2. The early stages of the chemical unfolding at proteome scale (Publication 2)

At the beginning of my Ph.D, the Prof. Orozco's group had just released the largest European database of MD trajectories (MoDEL) (reference [21] in the article). Among this huge amount of data, MoDEL hosted two interesting comparative sets that have been already used to test the performance of several force-fields in water (reference [20] in the article). The first set was formed by 30 proteins, which represent the most populated SCOP folds. The second one was a subset of the first, formed by three ultra-representative protein for the main SCOP folds (all-α, /β and all-β, see **Figure 1.9**). For the latter, simulations at the microsecond timescale were collected, quite a rarity at that time. The next steps consisted in setting up a similar comparative database for the denatured state of these proteins, taking advantage of the available data collected in water. The aim was to understand the molecular driving of the urea-induced unfolding at the proteomic level, avoiding any bias due to the force-field or the protein used. Such project was possible thanks the efforts by Manuel Rueda, Alberto Perez, and Carles Ferrer-Costa, which started the system setups.

First, the three ultra-representative proteins were used as a probe test where we validate our procedure at the microsecond timescale **(Figure 4.5)**. Since denaturation is a slow process, we speed up some observables simulating proteins in urea 8M at a mildly high temperature (368K). As controls, then, these three proteins were simulated in water at both 300K, to check the stability of the used force-field; and at 368K to separate the effect of thermal and chemical unfoldings. In the latter case we could address more specifically the changes in protein dynamics, profiting from the same system temperature.

We found that proteins clearly begin to unfold, but none of them reaches a fully unfolded state: they all preserve a certain degree of secondary structure and a compacted shape. Both temperature-alone and urea-alone produce a similar degree of unfolding and located in the same or very close by positions of the protein.

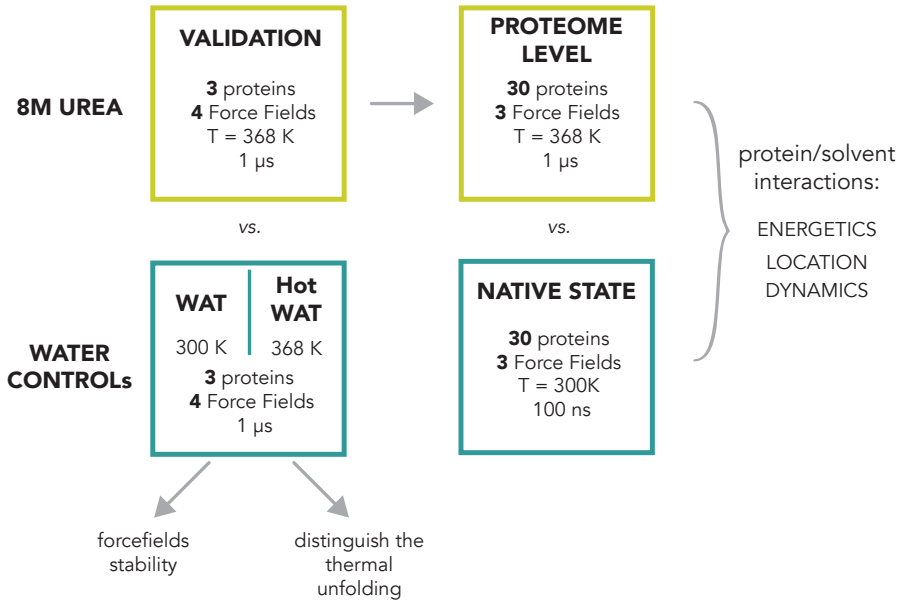


Figure 4.5. Schematic overview of Urea-MoDEL. After a first validation on a smaller subset of proteins (against equally-long simulations in water), the analyses were extended on a larger subset, using the dynamic native state for comparison.

However, a deeper look revealed several differences in the urea-induced unfolded structure: the apolar side chains are preferentially exposed to the solvent, and the atomic motions are slowed down prolonging the average time of local unfolding events.

The spotted differences pointed out towards a complex role for urea, far from a passive stabilizer of the thermal unfolding. We challenged then this hypothesis at the proteomic scale simulating the other 27 proteins in denaturing condition (30 in total counting the three ultra-representatives too). The analysis, guided by the earlier findings, focused this time on the location and energetics of the interactions between protein and solvent **(Figure 4.5).** The central question this time was the relation between the interactions of urea molecule with the protein and the observed local unfolding: are they related? Does urea trigger the unfolding, specifically where it locates?

Overall, our conclusions are in good agreement with those from our first project: the van der Waals interactions are responsible for the attrac-

tion of a lot of urea molecules in the protein surroundings. Specific and location-dependent interactions (both H-bonds and VdW) instead stabilize few urea molecules in cavities within the hydrophobic core. Those interactions often happen consistently in all the simulations of the same protein, and regardless of the forcefield used. We also found a common mechanism that brings urea molecules within the protein core, regardless of the protein under study (**Figure 4.6**). The intrusion of urea molecules is possible thanks to local unfolding events that happen at access points of the protein core (hinge points), usually formed by loops or turns at the protein surface. The longer average time of local unfolding events possibly facilitates the entering of urea molecules facilitating irreversible unfolding events. These are the "weak points" in each protein that are responsible for initiating the unfolding.

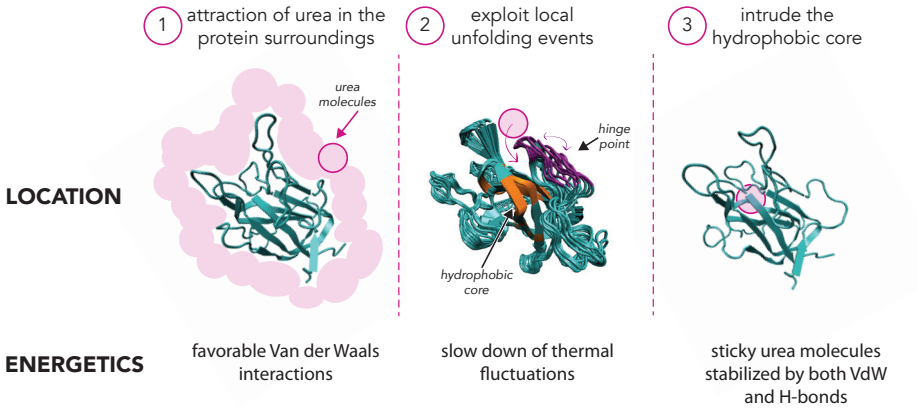


Figure 4.6. The route of urea to enter the protein core. Schematic exemplification of the steps that an urea molecule follows to intrude inside the core of a protein (1CZT). See also Figure 4 in the article's main text.

OPEN ACCESS Freely available online

PLOS COMPUTATIONAL BIOLOGY

Exploring Early Stages of the Chemical Unfolding of Proteins at the Proteome Scale

Michela Candotti[1,2], Alberto Pérez[3], Carles Ferrer-Costa[2], Manuel Rueda[4], Tim Meyer[5], Josep Lluís Gelpí[1,6], Modesto Orozco[1,2,6]*

1 Institute for Research in Biomedicine (IRB Barcelona), Barcelona, Spain, **2** Joint Research Program in Computational Biology, Institute for Research in Biomedicine and Barcelona Supercomputing Center, Barcelona, Spain, **3** Laufer Center, Stony Brook University, Stony Brook, New York, United States of America, **4** Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California San Diego, La Jolla, California, United States of America, **5** Theoretische und Computergestützte Biophysik, Max-Planck-Institut für Biophysikalische Chemie, Göttingen, Germany, **6** Department of Biochemistry and Molecular Biology, University of Barcelona, Barcelona, Spain

Abstract

After decades of using urea as denaturant, the kinetic role of this molecule in the unfolding process is still undefined: does urea actively induce protein unfolding or passively stabilize the unfolded state? By analyzing a set of 30 proteins (representative of all native folds) through extensive molecular dynamics simulations in denaturant (using a range of force-fields), we derived robust rules for urea unfolding that are valid at the proteome level. Irrespective of the protein fold, presence or absence of disulphide bridges, and secondary structure composition, urea concentrates in the first solvation shell of quasi-native proteins, but with a density lower than that of the fully unfolded state. The presence of urea does not alter the spontaneous vibration pattern of proteins. In fact, it reduces the magnitude of such vibrations, leading to a counterintuitive slow down of the atomic-motions that opposes unfolding. Urea stickiness and slow diffusion is, however, crucial for unfolding. Long residence urea molecules placed around the hydrophobic core are crucial to stabilize partially open structures generated by thermal fluctuations. Our simulations indicate that although urea does not favor the formation of partially open microstates, it is not a mere spectator of unfolding that simply displaces to the right of the folded$\longleftrightarrow$unfolded equilibrium. On the contrary, urea actively favors unfolding: it selects and stabilizes partially unfolded microstates, slowly driving the protein conformational ensemble far from the native one and also from the conformations sampled during thermal unfolding.

Citation: Candotti M, Pérez A, Ferrer-Costa C, Rueda M, Meyer T, et al. (2013) Exploring Early Stages of the Chemical Unfolding of Proteins at the Proteome Scale. PLoS Comput Biol 9(12): e1003393. doi:10.1371/journal.pcbi.1003393

Editor: David van der Spoel, University of Uppsala, Sweden

Received July 16, 2013; **Accepted** October 29, 2013; **Published** December 12, 2013

Funding: This work has been supported by grants CTQ2009-08850 (XS) and BIO2012-32868 (MO) from MINECO-Spain, the European Research Council (ERC-Advanced Grant, MO), the Instituto Nacional de Bioinformatica (INB; MO), the Consolider E-Science Project (MO), Framework VII Scalalife Project (MO), and the Fundación Marcelino Botin (MO). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: modesto.orozco@irbbarcelona.org

Introduction

Urea is a protein denaturant that has been used for decades in the study of protein folding/unfolding; however, after many years of research the ultimate reasons of the denaturing properties of urea remain elusive [1,2]. The dominant paradigm for unfolding (the "direct" mechanism) claims that the denaturant properties of urea are related to its capacity to interact with exposed protein residues more strongly than water [3–15]. However, the nature of such a preferential interaction is not so clear. Thus, while some authors suggest that it is mostly electrostatic and related to the formation of direct hydrogen bonds [7–9,16–17], others claim that preferential dispersion is the leading term [13–15]. It is also unclear whether the major destabilizing effect of urea is related to interaction with the backbone [6–7] or with side chains [8–12]. In the latter case, there is also discussion regarding the preferential side chains: polar and charged [9] or apolar [4,10–12].

We recently combined multi-replica molecular dynamics (MD) simulations and direct NMR measures of ubiquitin to characterize the "urea unfolded ensemble" of this model protein [15]. Our results suggest that urea stabilizes flexible over-extended conformations of the protein, which are unlikely to be sampled in the "unfolded" state of aqueous proteins. Extended conformations of the protein with exposed hydrophobic surfaces are more ureaphilic than the native globular state, due mostly to extensive London dispersion interactions (the attractive contribution in Van der Waals interactions between instantaneous dipoles) between apolar side chains and urea molecules in the first solvation shell of unfolded conformations. We believe that our results in reference 15 clarify the molecular basis of the effect of urea on the thermodynamics of the folded$\longleftrightarrow$unfolded equilibrium, but unfortunately, they do not provide information on the kinetic role of urea in the unfolding process. In other words: does urea actively induce protein unfolding? Or, on the contrary, does it passively stabilize the unfolded state by selectively binding to unfolded conformations? To analyze this point, we should characterize the effect of urea in the first stages of thermochemical unfolding, when the protein structure is still close to the native conformation and internal residues are not fully exposed. Clearly, a study of this nature presents many difficulties, the most important being that the effect of urea on early stages of unfolding might be dependent on the native structure. Therefore, to obtain conclusions of general

Author Summary

The delicate equilibrium between the folded and functional structure of a protein and its unfolded state is highly dependent on environmental variables such as the solvent. For example the co-solvent urea is a well-known protein denaturant that displaces the equilibrium towards unstructured and non-functional conformations of proteins. However the molecular mechanism behind its ability remains an enigma and the interpretation of the experimental data is still ambiguous. By analyzing a set of representative proteins through extensive molecular dynamics simulations in urea, we provide a robust and consensus picture of the first stages of urea-driven protein unfolding and elucidate the role of urea in accelerating protein unfolding. Our results suggest that urea, thanks to its stickiness and slow diffusion, benefits from the intrinsic flexibility of proteins and stabilizes partially open-states, slowly driving the protein toward unfolding.

validity, all representative protein folds should be addressed. Also, results can be force-field-dependent, so if we aim to obtain robust conclusions, we should perform simulations with a variety of force-fields.

Given the typical kinetics of the folding/unfolding transitions of small globular proteins [18], microsecond (μsec) long simulations should trace the first stages of these processes. In the current work, we investigate the first stages of urea-driven protein unfolding using μsec-long atomistic simulation; to gain universality, we used 30 proteins representative of all protein folds, while to protect our conclusions from force-field-related uncertainties, we used several of the most popular force-fields. The results derived from this study provide a robust and complete picture of the role of urea in destabilizing folded states of proteins, and more importantly, on the molecular mechanisms by means of which urea contributes to accelerating protein unfolding.

Results

Protocol validation using three ultra-representative proteins

We first validated our protocol using three ultra-representative proteins (in bold in Table 1), one for each of the main classes in the Structural Classification of Proteins (SCOP, [19]). We monitored the protein stability in three environments: i) in chemical unfolding conditions, in 8M urea and with a mildly high temperature (T = 368K) to speed up the observable effects; ii) in thermal unfolding condition, in water with the same high temperature; this control allowed us to distinguish the effect of urea and temperature on protein unfolding; iii) in water at room temperature as final control. Four force-fields were used (OPLSAA - ON2; CHARMM - C22; AMBER99 - P99 and P99SBILDN) for each system (see Methods for the description of the force-fields used), collecting in total 36 simulations of 1-μsec length each.

Control simulations at room temperature. Analysis of the trajectories in water at room temperature for the 3 ultra-representative proteins confirmed that current force-fields can accurately represent the native conformation of soluble proteins in the μsec range [20,21,22], reproducing the global and local structure of proteins well. The structures in the last segment of the trajectory (and the corresponding ones collected just after equilibration) showed, in general, little structural drift from the experimental conformation (see Figure 1, Suppl. Figures S1, S2 and Suppl. Table S1). This was noted in the small values of root mean squared deviation (RMSD) from native structure at the end of the simulation (typically around 1.5 Å), and the good preservation of the fold structure (average TMscore around 0.8), the shape descriptors (radius of gyration, RadGyr and solvent accessible surface area, SASA) and the secondary structure (SS) composition. We found only one significant discrepancy: simulation of 1CQY using the C22 force-field showed a non-negligible transition in the 100-ns time scale, leading to the sampling of conformations that were 3 Å away from the experimental structure; see Suppl. Figures S1, S2.

Control simulation of thermal unfolding (hot water). The mild high temperature applied in the simulations in hot water (below water boiling point: T = 368K) significantly enhanced the global fluctuations of the protein (see Suppl. Figures S2), while advances in the unfolding were still moderate. Thus, after 1 μsec of MD in hot water, the RMSD from experimental

Table 1. Structures representative of the 30 most populated protein meta-folds.

Symb Fig. 3	PDB code	Molecule name
a	1AGI	Angiogenin-1
b	1CHN	Chemotaxis protein CheY
c	1FVQ	Copper-transporting ATPase
d	1GND	Rab GDP dissociation inhibitor alpha
e	**1KTE**	**Glutaredoxin-1 (Thioltransferase)**
f	1LIT	Lithostathine-1-alpha
g	1PDO	Mannose Permease – IIA domain
h	1SDF	Stromal cell-derived factor-1
i	1SUR	PAPS Reductase
j	2HVM	Hevamine
k	1BFG	Basic fibroblast growth factor
l	1BJ7	Allergen Bos D2
m	**1CQY**	**β-amylase, Starch-binding domain**
n	1CSP	Cold shock protein B
o	1CZT	Coagulation factor V, C2 domain
p	1J5D	Plastocyanin
q	1KXA	Sindbis virus capsid protein
r	1NSO	Protease
s	1PHT	P13-kinase, SH3 domain
t	1BSN	F1-ATPase, ε subunit
u	1EMR	Leukemia Inhibitory Factor
v	1IL6	Interleukin-6
w	1JLI	Interleukin-3
x	1K40	Focal adhesion kinase, FAT domain
y	1LKI	Leukemia Inhibitory Factor
z	1OOI	Odorant binding protein LUSH
α	**1OPC**	**OMPR, Dna-binding domain**
β	1FAS	Fasciculin-1
γ	1I6F	Alpha-like toxin CsEv5
δ	1SP2	SP1F2, zinc-finger dna binding domain

The list is divided according to the SCOP fold group (in order *all-α, all-β and α/β*). The three ultra representative proteins used in the protocol validation are in bold.

doi:10.1371/journal.pcbi.1003393.t001

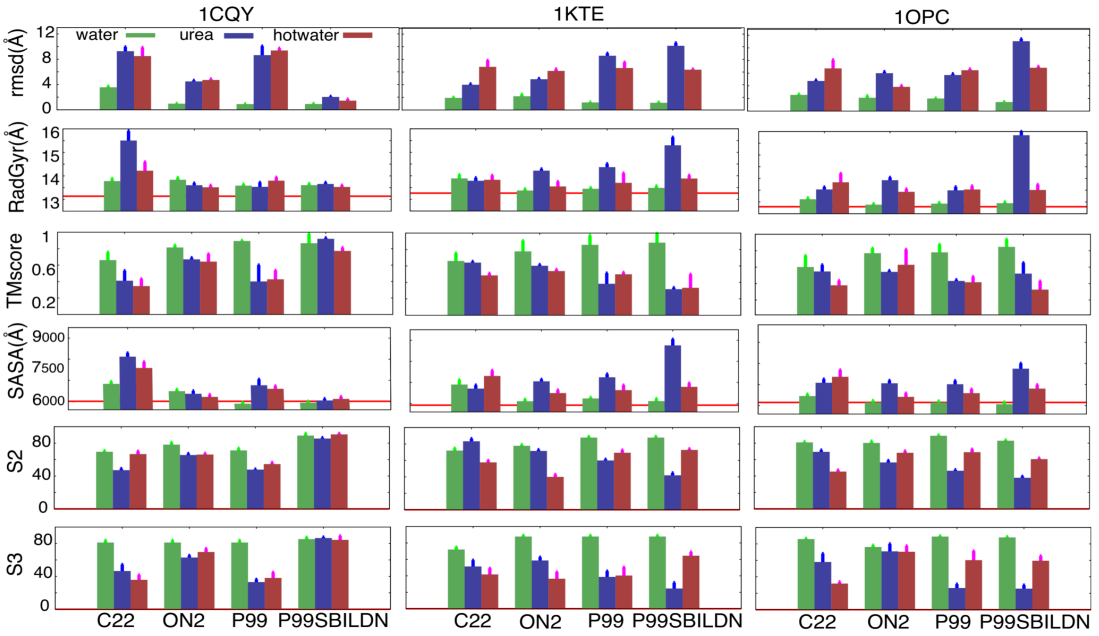


Figure 1. Shape and unfolding descriptors for the three ultra-representative proteins. Root-mean-squared-deviation (RMSD) from the starting conformation, radius of gyration (RadGyr), TMscore, solvent accessible surface area (SASA), native secondary structure index (S2) and native contacts index (S3) were calculated in water, urea and hot water in the four force-fields (OPLSAA - ON2; CHARMM - C22; AMBER99 - P99 and last-modified P99SBILDN). Average values and relative standard deviations are calculated in the last 10 ns of the simulation. For radius of gyration and SASA, the value found for the starting conformation is reported as a red line. See Methods and Suppl. Text S1 for a description of the metrics. Error bars mark the standard deviation.
doi:10.1371/journal.pcbi.1003393.g001

structures reached the range 5–7 Å, and shape descriptors (RadGyr and SASA) indicated a moderate increase in the size of the protein (see Figure 1 and Suppl. Table S2). The fold-architecture started to be corrupted (TMscore values around 0.5), with a moderate loss of native contacts (S3) and native secondary structure (S2 - see Figure 1 and Suppl. Table S3).

Detailed analysis of the 12 simulations (three proteins and four force-fields) provides interesting information on the behavior of the force-fields. In general, the overall picture at the beginning of thermal denaturation of proteins was quite robust to force-field changes. However, we found two clear discrepancies. First, C22 appeared to facilitate unfolding in hot water (see Suppl. Figure S2), yielding more flexible structures than those obtained with the other force-fields. Second, in P99SBILDN the 1CQY protein remained fully preserved at the end of the high temperature trajectory. Five independent replicas of the same system with different starting geometries and velocities failed to detect significant unfolding for 1CQY with P99SBILDN. This observation points to a potential problem of over-stabilization of the folded structure for this all-β protein.

The differential effect of urea in the early stages of chemical unfolding. We first analyzed the impact of high concentrations of urea on the three ultra-representative proteins. Overall, and contrary to previous suggestions [13], in the microsecond scale the unfolding efficiency of urea did not change dramatically from that in hot water simulation. In the same simulation period, proteins in urea display RMSD values that were marginally larger than in hot water (see Figure 1 and Supp. Table S2), and TMscore, S2 (secondary structure) and S3 (native contacts) values at the end of the simulations in urea were not much different to those obtained in hot water (see Figure 1 and Supp. Table S2), except for a certain enlargement in the disruption of β-sheets when urea was added (see Suppl. Table S3).

We found a significant correlation ($r = 0.701$; p-value$<2.2\ 10^{-16}$) between the time that each native contact remained lost in water and in urea at the same temperature (Suppl. Figure S3-A). This observation suggests that urea does not attack specific parts of the protein, but rather benefits from the intrinsic breathing movements of the protein at high temperature. However, the role of urea in guiding unfolding is reflected by the different nature of the structural deformations that occurred in hot water and urea simulations. Thus, the latter sampled conformations that were slightly more extended (higher RadGyr) and clearly more exposed (higher SASA) than those sampled in hot water (Figure 2). It is worth noting (see Figure 2 and Suppl. Figure S3) that in urea-driven unfolding the solvent-accessible surface (SASA) corresponding to apolar residues increased dramatically, a behavior reminiscent of the surfactant action, while this increase was moderate in hot water simulations. The urea-induced increase of the apolar-exposed area was not accompanied by a dramatic enlargement of RadGyr or to a large decrease in the structural indexes, thereby suggesting that the exposure of the hydrophobic core occurs through the creation of small cavities (filled with urea) and the exposure of apolar side

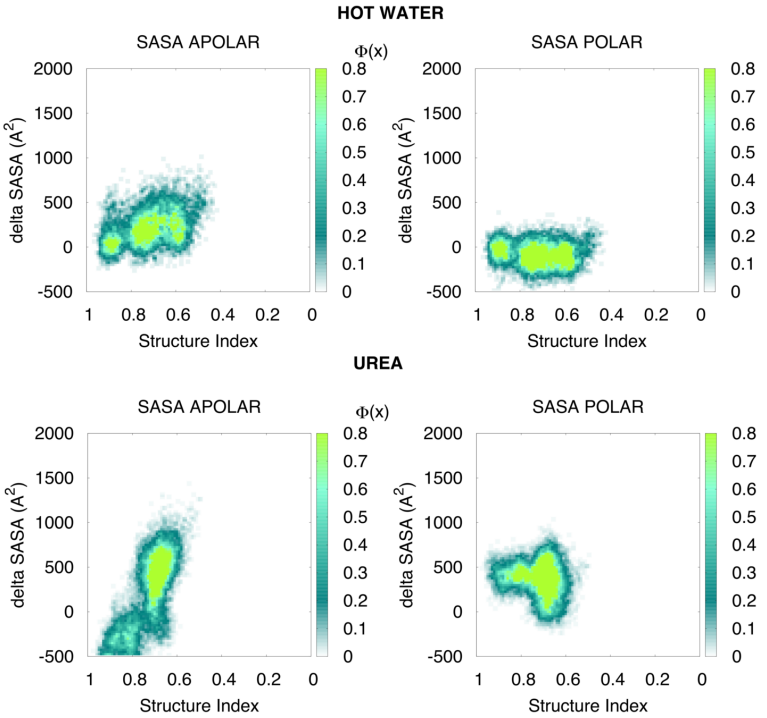


Figure 2. Variation of the solvent accessible surface area (SASA) during the unfolding. Values are reported for apolar and polar residues in hot water and urea, for the three ultra-representative proteins in all force-fields. SASA is normalized for each protein using the average value calculated in the water simulations (to take into account the structure rearrangements and the mobility in water), while the structure index is used to follow the unfolding process (from 1 - fully native folded protein - towards 0). The color marks the density. For more detailed pictures for each protein see Suppl. Figure S3. Note that in urea the increase in SASA is larger than that in water, mostly due to the exposure of apolar moieties, similarly to the action of surfactant.
doi:10.1371/journal.pcbi.1003393.g002

chains, without a dramatic extension of the protein or an explosion of the hydrophobic core.

A second major difference between the unfolding yielded by hot water and by urea was revealed by the analysis of the dynamics of the protein. Intuition suggests that proteins will show greater fluctuation (at the same temperature) in the presence of a denaturant like urea. This is certainly true for a fully unfolded protein [15], but not during early stages of unfolding, when the protein is still close to its native state, as noted in the values of RMSD calculated in various time-windows (see Suppl. Figure S3-B). The explanation of this apparently counterintuitive finding is that urea solutions are more viscous, thus reducing the fastest movements of the proteins, including the oscillations of side chains (66% of side chains were stiffer in urea than in water). This reduction causes a slow down of the atomic-motions, which in fact opposes unfolding. However, the slower mobility of urea and its sticky nature may explain the longer life time of lost contacts (see Suppl. Text S1) in urea (see Table 2 and Suppl. Figure S4-C), a feature that clearly favors unfolding (see below).

Regarding differences related to force-fields, we detected the same discrepancies as in hot water. C22 simulations showed more mobility and distortions (Suppl. Figure S2), but conformations were still similar to those obtained with other force-fields. With the P99SBILDN force-field, the full-β protein 1CQY remained stable when simulated at high temperature in the presence (but also absence) of urea. Five 1-μsec replicas of this trajectory failed again to detect any significant unfolding of this protein, a finding that suggests caution in the use of P99SBILDN (a force-field refined to reproduce folded structures) in unfolding studies of full-β proteins. Given our observation, the P99SBILDN force-field was not considered in the rest of the study.

Table 2. Change in flexibility of contacts (opening time) maintained in hot water and urea.

Force-field	U* (%)	W* (%)	Tot (%)	Δ(U-W) Normalized (%)
ON2	10.54	7.4	17.94	+17.5%
P99	12.19	4.64	16.83	+44.8%
C22	2.76	5.85	8.61	−35.8%
P99SBILDN	7.02	3.51	10.53	+33.3%

*Percentage of native contacts that present a longer opening time in urea (U) or water (W) (difference between opening time larger than | 0.1 | ns). The average total number of contacts is 1110 in C22, 1148 in ON2, 1140 in P99 and 1143 in P99. Each protein has ~380 native contacts - defined as those occurring for more than 80% of the time in the 0.1 microsecond simulation in water at 300K.
doi:10.1371/journal.pcbi.1003393.t002

Proteome-level study of urea unfolding

After the validation of our protocol, we extended the chemical unfolding simulations to a larger set of proteins, to avoid any bias in the conclusion due to the native structure. We performed 1 μsec of simulation in urea at high temperature (T = 398K) for 30 proteins covering all the major protein folds (Table 1 and Suppl. Dataset S1). Each system was simulated in three force-fields (C22, ON and P99), excluding P99SBILDN as reported above, and collecting in total 90 simulations. To have a more realistic picture of the native state, instead of using the crystal structure, we used as control 0.1 μsec-long simulations in water at room temperature for all the 90 systems. The analysis described here reveals some common robust trends that illustrate the effect of urea during the early stages of protein unfolding.

Global denaturation. As anticipated from simulations in the small set of ultra-representative proteins, urea led to an enlargement of the protein and to a deviation from its native structure (Figure 3), without reaching, however, full unfolding in any of the 90 simulations in urea at high temperature. On average, our simulations produced RMSD values (from experimental native conformation) around 4 Å larger than those found at the end of the control simulation in water, while for these proteins a fully unfolded structure should yield ΔRMSD values above the range 20 Å [15] and a random structure above 10 Å [23]. Only a few proteins lost their fold integrity and native contacts after 1-μsec simulations in urea at high temperature, as noted in the reductions beyond 0.5 in the TMscore and in native contact (S3 structural index) around 0.2–0.3. However, in general, the urea-induced disruption of core structural elements was moderate (reduction of TMscore around 0.3–0.4 and S3 indexes around 0.4–0.6 at the end of the trajectory; see Figure 3).

The selected force-fields showed a consistent representation of unfolding and in general the urea- labile or resistant proteins defined among the entire set matched in all of them. For example, all force-fields detected minuscule advances in unfolding (as determined by the set of metrics in Figure 3) for 1GND(d), 2HVM(j), 1CSP(n), 1OPC(α) and 1KTE(e), whereas the same force-fields detected significant progresses in others (for example 1CQY-m, 1BSN-t, 1OOI-z, 1K40-x or 1SP2-δ). Only in a few cases was there apparent large discrepancy between force-fields (example 1FVQ) and these corresponded to simulations where structural alterations were already seen in the reference simulations (example C22 for 1FVQ). In summary, despite the stochastic nature of unfolding and the uncertainties implicit to the force-field, the general picture of urea unfolding detected here is robust.

Urea-induced unfolding and loss of secondary structure. We did not find any correlation between the presence/absence of disulphide bridges and the extent of urea-induced unfolding (note that to exclusively analyze the effects of urea, disulphide bridges were not reduced in our calculations). All changes in urea sensitivity related to fold type, secondary structure composition or the presence or absence of disulphide bridges were small during the first stages of unfolding.

The distribution of urea around the protein. As anticipated in previous studies [4,8,10,13,15], proteins are urea-philic. All the proteins studied here (for all force-fields) quickly recruited urea into the first solvation shell (in agreement with osmometric experiments [24]), where the water/urea ratio reached values in the range 3–3.5 water/urea molecules, while the background ratio was around 6 (see Table 3). However, this enrichment was smaller than that found for a fully unfolded protein (0.9 for unfolded ubiquitin; [15]), thereby suggesting that the most urea-philic groups remained buried in the interior of the protein. Urea did not preferentially solvate any residue (see Figure 4A) and showed preferential binding to the backbone rather than to the side chains during the first stages of unfolding. Although larger in size, urea

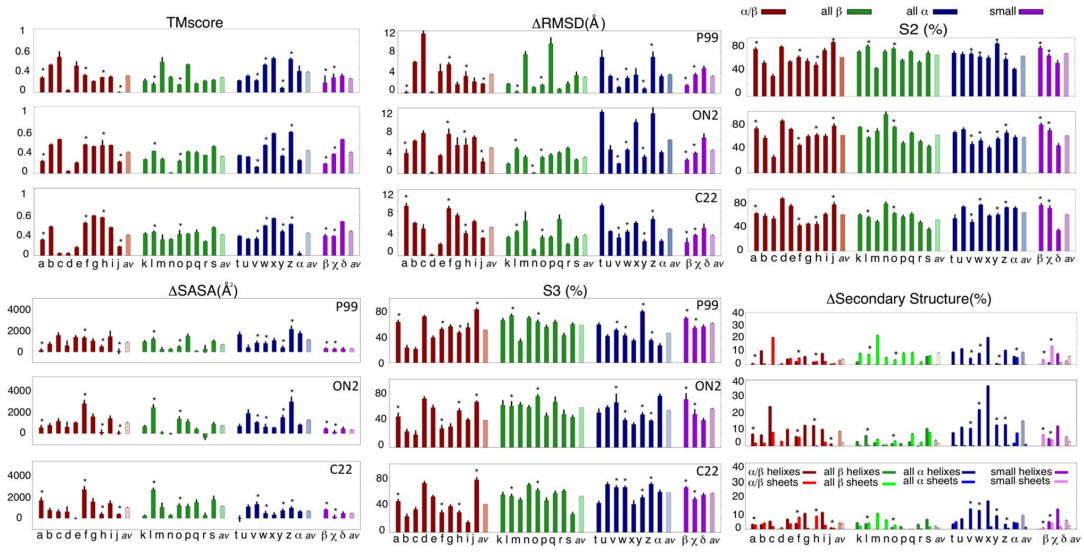


Figure 3. Shape and unfolding descriptors for the 30 representative proteins. The difference of TMscore, RMSD and SASA between values in urea and in water (to allow comparison between proteins of different size) calculated in urea in the three force-fields. The native contacts index (S3), native secondary structure index (S2) and the difference in Secondary Structure content (Δ% Sec. Structure) are also shown. To facilitate discussion, proteins are grouped following the SCOP classification. The correspondent pdb code is reported in Table 1; the group average is reported as "*av*" while the symbol * marks proteins with disulfide bonds. Error bars mark the standard deviation.
doi:10.1371/journal.pcbi.1003393.g003

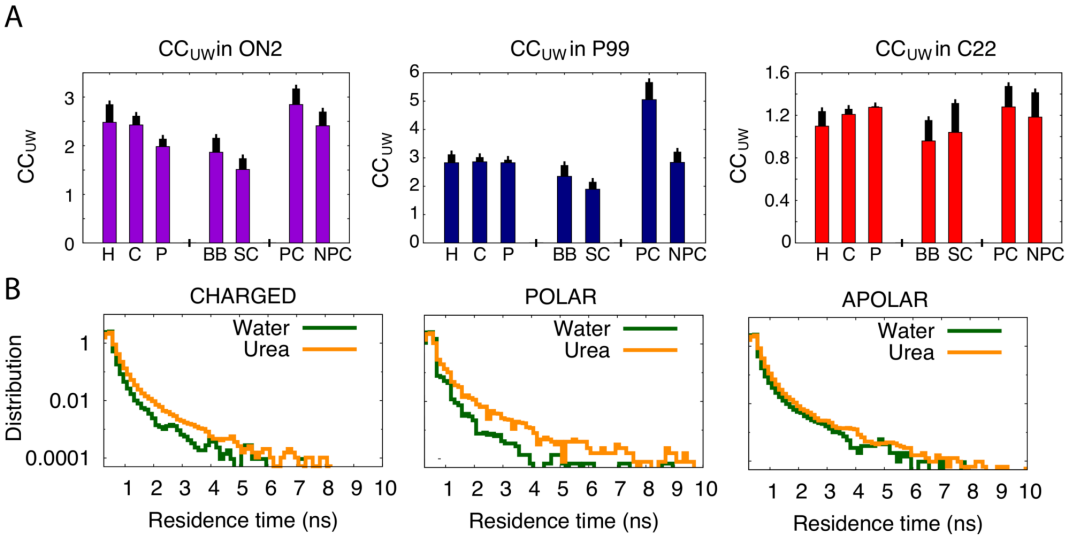


Figure 4. Location of solvent molecules in urea. A) Preference for urea solvation measured by CC_{UW} (the ratio for each amino acid between atomic contacts with urea and with water molecules - see Suppl. Text S1) for different parts of the protein: hydrophobic (H), polar (P), charged (C) residues, side chains (SC), backbone (BB), protein core (PC) and non-protein core (NPC). Error bars mark the standard deviation. Note that in all the force-fields the PC shows the largest values, meaning a larger preference for urea. **B**) Distribution of the residence time for urea and water molecules during a 1-μsec trajectory.
doi:10.1371/journal.pcbi.1003393.g004

has a higher affinity than water to interact with residues placed in narrow cavities near the hydrophobic core (see below). Long residence urea molecules placed in these cavities led to a partial exposure of the hydrophobic core of the protein (see Figures 1–3).

At this point we wish to comment on the urea distribution for C22 simulations, since we detected significantly less urea in proximity of the protein as compared to the other force-fields. This unusual behavior of C22 urea simulations is evident in Table 3 and Suppl. Figure S5-AB, where some trends found in P99, ON2 (or P99SBILDN) differ from those in the C22 simulations. Urea densities around the proteins in the C22 simulations may have been too low, possibly reflecting the excessive polarity of the urea model used in the C22 trajectories (dipole moment 5.3 D, compared with the dipoles around 4.7 D of the other models) [11].

Table 3. Comparison of ratio water-urea in the first solvation shell (FSS) and in the bulk (BULK).

$Ratio_{WU}$*	ON2 FSS[1] - BULK[2]	P99 FSS - BULK	C22 FSS - BULK
All α	3.39–6.26	3.01–6.45	5.03–5.87
All β	3.56–6.43	3.13–6.52	5.12–5.89
α/β	3.22–6.24	3.09–6.21	4.93–5.94
Small	3.46–6.17	2.94–6.33	4.95–6.19

Values are the average along the simulation, SD is always lower than 0.2.
*values are the average along the simulation, standard deviation is always lower than 0.2,
[1]FSS defined by a maximum 5 Å cutoff to protein,
[2]BULK defined by a minimum cutoff of 6 Å.
doi:10.1371/journal.pcbi.1003393.t003

The energetics of protein-urea interaction. The nature of the interaction between urea and proteins has been the subject of intense discussion (see *Introduction*). Our previous results [15] suggest that in the fully unfolded state there are many urea-protein hydrogen bonds, mostly with the backbone, but that the main factor responsible for the urea-philicity shown by proteins is the differential dispersion interaction of bulk and protein-bound urea. However, these conclusions for the unfolded state might not be valid when the protein is still compact during the early stages of unfolding. Analysis of current data (see Table 4 and Suppl. Table S4) shows that already in these early stages of unfolding 30% of the protein-solvent hydrogen bonds are with urea, and the ratio is even higher (36%) when considering only stable contacts. In 2/3 of the cases, urea acts as a hydrogen donor when H-bonding to the backbone, and, in general, urea-protein H-bonds display longer life times than water-protein ones, a feature that appears to be crucial to stabilize partially exposed residues (see below). Nevertheless, the formation of these H-bond interactions (mostly electrostatic in nature) is not the driving force that explains the urea-philicity of the nearly-native conformation, since the migration of urea from background to the first solvation shell (FSS) of the protein does not alter global electrostatics (see Suppl. Figure S5-B), but improves the van der Waals interactions [13–15]. This effect and the gain in water entropy related to the replacement of several water molecules by a single urea molecule [10,11] may drive denaturation in the early stages of urea unfolding.

Urea and protein dynamics. Urea diffuses quite slowly (Suppl. Figure S5-C) and limits protein fluctuations, which leads to an apparent paradox: a denaturant that slows down the dynamics of proteins compared to the equivalent simulations in water (see also Suppl. Figure S4-B). However, analysis of trajectories show that such a paradox does not exist. Urea migration to the protein surface was slower than that of water, but once it reached the

Table 4. Hydrogen bond interactions of urea/water with proteins during the last 10 ns of trajectories.

H- bonds (% of total)*:	Urea/water as H-donor		Urea/water as H-acceptor	
	66/65		34/35	
H-bonds with protein:	**Backbone**	**Side chains**	**Backbone**	**Side chains**
% of total	58/48	42/52	67/69	33/31
Av. Lifetime(ps)[$]	74/64	107/74	327/68	451/249

*Distance cutoff is 3.50 Å, angle cutoff is 120.00 degrees. Hydrogen bonded solvent molecules are defined for occupancies (total time) larger than 0.5 ns,
[$]Life-time refers to the percentage of analyzed trajectory (10 ns).
doi:10.1371/journal.pcbi.1003393.t004

surface, urea remained for longer periods (see Figure 4B), especially when located in cavities near the hydrophobic core of the protein (see Figure 5 and Suppl. Figure S6-A for examples). Interestingly, the positions of long-lasting urea interactions are consistent among all four force-fields and seems associated with a sizeable improvement in van der Waals interactions and electrostatic energies and with the formation of strong long-living H-bonds (see examples in Figure 5, Suppl. Figure S6-A and Suppl. Table S4). These findings demonstrate that even if H-bonding is not the driving force behind the urea-philicity of proteins, it is important to stabilize urea molecules at specific positions at the protein interior.

As noted above, residues that are very mobile in urea are also highly mobile in water at high temperature (see Suppl. Figure S4-A). Furthermore (see Suppl. Figure S4-C), with the exception of C22 simulations, there was a slight but significant ($r>0.2$; p-value:$<2.2\ 10^{-16}$) correlation between oscillating residues in urea simulations and in native simulations (water at 300K). Interestingly, long-residence urea molecules were typically bound to rigid regions of the protein adjacent to mobile residues, i.e. they are located at putative hinge-points at the interface between the more rigid core of the protein and flexible loops or tails (see examples in Figure 6C and Suppl. Figure S6-C). The presence of sticky urea in these regions is expected to have a major role in guiding unfolding (see below).

Discussion

MD simulations with additive potentials and explicit solvent have become very popular to explore chemical unfolding of protein. There is little doubt that the use of the technique has produced sizeable advances in the field, but we cannot ignore some potential caveats in the beginning of this discussion. First, for computational reasons we (and most authors in the field) are using classical non-polarizable force-fields, which might not be accurate enough to deal with a complex process such as unfolding. Previous studies [4–5,7–15,25] have however demonstrated that urea/water/protein effective parameters are able to reproduce a variety of experimental observables, such as mass densities and radial distribution functions of urea/water solutions derived from neutron scattering experiments [25], the experimental water/urea transfer free energies of tripeptides [10], and the urea density

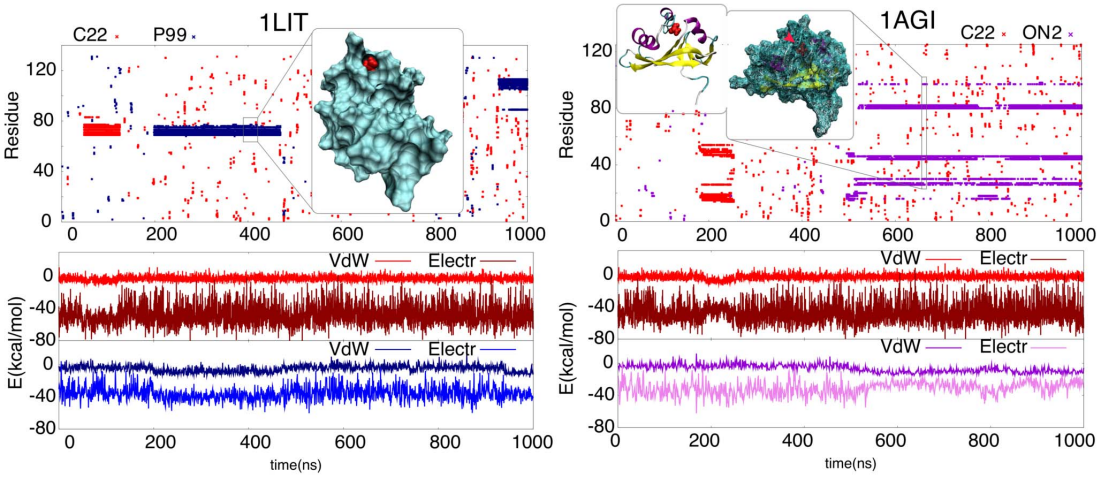


Figure 5. Long residence urea molecules. Examples of urea contacts with the protein residues (y-axis) along 1 μsec of simulation (x-axis). Each dot defines a contact between a urea molecule and protein residues. A contact is defined when at least one pair of heavy atoms comes closer than 3.5 Å (see Suppl. Text S1). Examples of urea molecules trapped in the protein core are shown in the top panels. Note that in the same protein but simulated in different force-fields, a long residence urea is trapped in a very similar area of the protein core. The panels below show the evolution of electrostatic and dispersion energies for the urea molecules (calculation details as in Suppl. Figure S5-B; see also Suppl. Text S1). Note the reduction mainly in dispersion energies upon the binding of urea.
doi:10.1371/journal.pcbi.1003393.g005

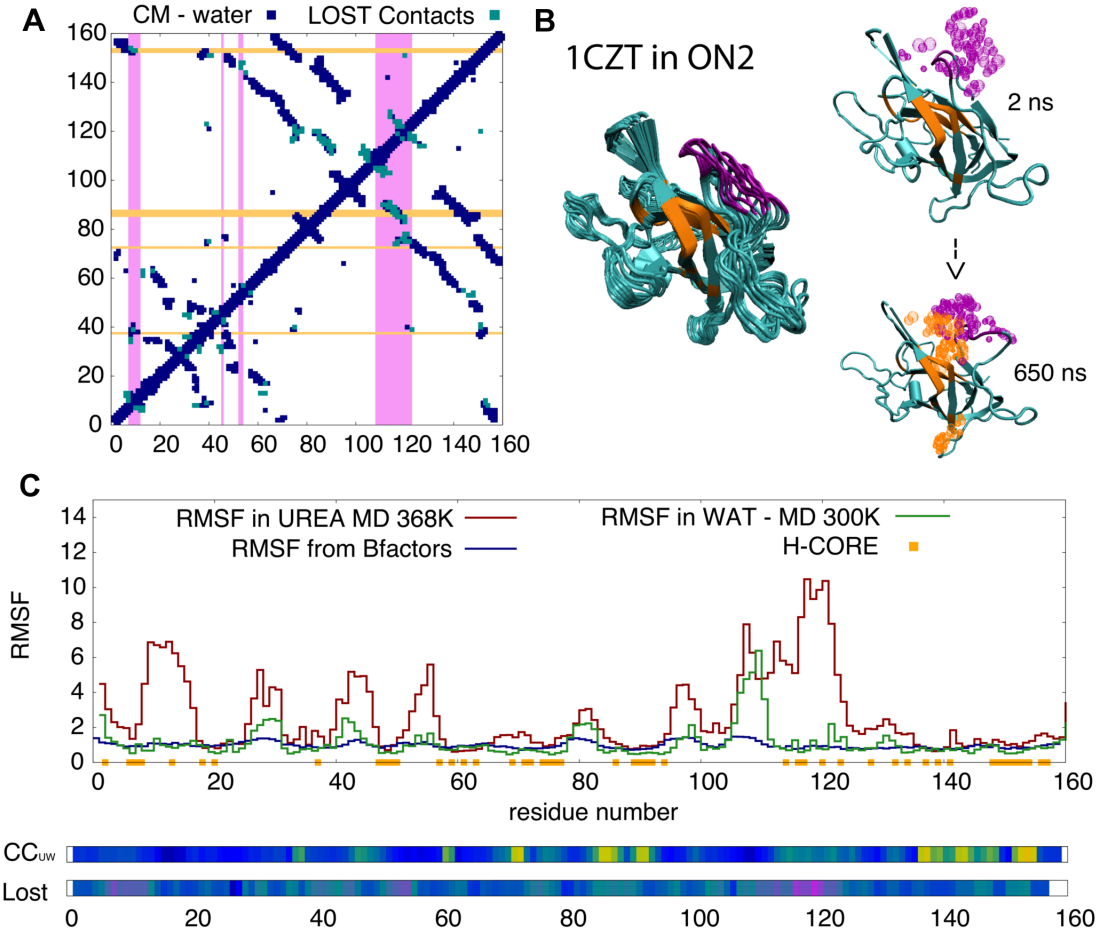


Figure 6. Urea intrusion into the core of 1CZT. A) Contact map from the crystal structure of 1CZT in blue (each dot represent a contact), the contacts lost during 1 μsec of simulation in urea are shown in light blue. Areas in magenta mark residues with a large flexibility while those in orange mark residues with a high preference to contact urea. **B**) Snapshots showing the temporal evolution of the protein structure; the areas in magenta and orange follow the same color code as in panel A. Urea molecules within 4 Å of these areas are shown in the same color. Note that flexible areas (in magenta) on the surface of the protein - mainly loops - undergo opening events, and the loss of contacts (panel A) connecting these areas to the protein core (in orange) triggers urea intrusion. **C**) The residue root mean square fluctuation (RMSF; a measure of flexibility), the contact coefficient CC_{UW} (measure of the binding preference of protein to contact urea rather than water) and the % of lost time (measure of local unfolding) along the protein sequence. These metrics allow us to locate areas with large % of lost contact time and high flexibility in urea (magenta), while orange and yellow regions illustrate large values of CC_{UW}, meaning a remarkable preference to contact urea. For more examples see Suppl. Figure S6-B.
doi:10.1371/journal.pcbi.1003393.g006

around unfolded proteins found by vapor pressure osmometry measures [4,8,10,13,15,24]. Furthermore, our recent work [15] has demonstrated that unbiased MD simulations in 8M urea reproduce very accurately the unfolded ensemble as determined from a variety of spectroscopic techniques (including SAXS and NMR) under the same conditions. Thus, despite their simplicity current force-fields reproduce reasonably well urea/water/protein mixtures. We should remember that since we are exploring a microsecond-long process, no direct experimental data is available for comparison and accordingly caution is required. This move us to use a consensus approach, running the simulations with different force-fields to extract those results that seem robust to force-field changes.

A second reason of concern is related to the stochastic nature of unfolding, where individual trajectories can show different degree of unfolding [13]. Again, by comparing different trajectories we tried to define robust findings, but we cannot ignore that the experimental result is the averaging a near-Avogadro number of trajectories. A third reason of concern, common to many experimental studies, is the generality of the results, i.e. how general are the results obtained with a few model proteins. To convince ourselves on the general validity of our results we repeated the unfolding studies for a large number of proteins representative of all prevalent folds. Despite the obvious caveat of any theoretical study, this approach provided a picture of unprecedented, to

our knowledge, completeness and robustness of the early stages of urea unfolding.

Under our simulation conditions (8M urea at T = 368 K), we detected clear signals of unfolding in the microsecond range, but progress in denaturation was smaller than that reported for model proteins of reduced stability [13]. Overall, the advance in unfolding of proteins at T = 368 after 1 μsec of MD is not dependent on the fold, nor secondary structure composition, and was similar for proteins with and without disulphide bridges in the native form. No dramatic differences in the advance of unfolding were found between water and urea simulations performed at the same temperature; however, unfolding paths in the presence or absence of urea differed, since the partially unfolded structures sampled in hot water maintained the hydrophobic residues hidden in the core of the protein, while such residues were more accessible in presence of urea.

Urea solutions are more viscous than pure water, which in our simulation reduced high-frequency movements in the protein, generating an unexpected slow down of the atomic-motions. Urea residence times around protein residues were large, especially when urea molecules diffuse close to the hydrophobic core or to the interface between rigid and thermally mobile regions (hinge points). We consider that the sticky nature of urea and its preferential placement at hinge points is crucial for unfolding, since it favors the rapid trapping of residues that become exposed as a consequence of stochastic thermal motions. The stabilizing effect of urea on exposed residues slowly biases the trajectory towards the unfolded state, by decreasing the chances of microscopic refolding [12,26]. The effect of stabilization of exposed residues is especially productive in terms of unfolding when residues are apolar, since in this case urea (but not water) traps very efficiently the residue, increasing the accessibility of other apolar residues in the vicinity. The ensuing greater recruitment of urea in the region leads to a cooperative effect resulting in the acceleration of protein unfolding. Our data shows that, similar to the unfolded state [13,15], it is the van der Waals interactions that drive the accumulation of urea on the surface of the folded protein. However, the role of H-bonding cannot be dismissed, as these bonds are crucial for the stabilization of long-living urea interactions near hinge points, which in turn are required to bias intrinsic protein dynamics towards unfolding. Clearly, ''direct'' effects not only are the main factors responsible for the urea-mediated stabilization of the unfolded state [15], but are also relevant in guiding the first steps of urea unfolding.

Microscopic unfolding events are related to stochastic thermal motions, which are in principle similar to those that occur spontaneously in water at room temperature. However, urea is not a mere passive spectator that simply stabilizes the small percentage of unfolded protein coexisting within the native ensemble and leading to a displacement in the folded$\longleftrightarrow$unfolded equilibrium towards the denatured state. On the contrary, urea has a dual function: i) it takes advantage of microscopic unfolding events, decreasing their chances of refolding, and favoring further unfolding [12,26]; and ii) among these microscopic unfolding events it selects and stabilizes microstates with exposed hydrophobic regions [4] (see Suppl. Figure S7). These effects lead to a slow divergence in the temperature-unfolding pathways in water and urea, and, as shown for ubiquitin [15], to distinct unfolded states. Consequently, concepts such as folded and unfolded states or folding and unfolding pathways need to be revisited and reformulated considering the nature of the denaturant used.

Methods

Selected proteins

As model structures for the main protein-folds we used the same structures selected in our previous work in reference 20. We first explored the early stages of urea unfolding using three ultra-representative proteins for the most populated fold in the three main classes in the SCOP database (*all-α* 1OPC, *all-β* 1CQY and *α/β* 1KTE; [19,20]). Once the simulation protocols had been validated with these proteins, the study was extended to a larger set, consisting of 30 structures (110 residues on average) representative of the most populated protein folds ([20,27] and Suppl. Dataset S1)

Simulation set-up

All starting structures were taken from the Protein Data Bank (PDB; [28]) and processed using our standard procedure implemented in the MDWeb server [29]: experimental structures were titrated to define the major ionic state at neutral pH, neutralized by ions (sodium and chloride), minimized for 1000 steps, heated up to the final temperature, and solvated using a 8M urea/water octahedron box with a spacing distance of 15 Å around the system. The box was previously equilibrated in a Monte Carlo simulation using the BOSS program [30]. The water model was taken from Jorgensen's TIP3P [31], while ion and urea force-field parameters were those considered as the default of each force-field. Urea parameters from Smith et al. [32] were used for OPLS and P99SBILDN simulations, the same charges but scaled according to the amber force-field were used in PARM 99, while Nilsson's parameters were used in the CHARMM 22 force field [33]. Systems were then pre-equilibrated for 0.5 ns with parm99-AMBER force field in keeping the backbone restrained by intra-molecular harmonic potentials and then equilibrated (0.5 ns) in each force field parameters removing backbone constraints.

Simulation details

For the small set of ultra-representative proteins, three sets of simulations corresponding to water at room temperature (T = 300 K), hot water (T = 368 K), and urea at high temperature (T = 368 K) were carried out. For each condition, we performed 1 μsec simulations using four force-fields: three general purpose ones (OPLSAA -ON2- [34]; CHARMM -C22- [35]; AMBER99 -P99- [36]), and a last-generation force-field able to accurately reproduce folded proteins (P99SBILDN, [37]). For the extended set of 30 proteins, control simulations in water were limited to 0.1 μsec at room temperature, while the 8M urea simulations were performed, as above, for 1 μsec at T = 368 K. Simulations for the extended set of proteins were carried out using ON2, C22 and P99. All simulations were performed using periodic boundary conditions and particle Mesh Ewald [38] corrections for the representation of long-range electrostatic effects using a 1.0 Å grid spacing and a 9 Å cutoff. All trajectories were collected with the NAMD2 [39] program. Integration of equations of motions was performed every 2 fs after removing vibrations of bonds involving hydrogen atoms using SHAKE/RATTLE algorithm [40,41]. All simulations were carried out in the isothermal (T = 300 or 368 K)/isobaric ensemble (P = 1 atm) using the Langevin thermostat and barostats [42,43]. The trajectories were analyzed using VMD [44] and the MdWeb server [29], as well as Flexserver which can be accessed at: http://mmb.pcb.ub.es/FlexServ/ (see also Suppl. Text S1 for a detailed explanation of the metrics).

Supporting Information

Dataset S1 List of the structures selected to represent the 30 most populated folds according to SCOP, CATH, Dali and Dagget's databases [20]. When available, we included the denaturation midpoint, as measurement of protein intrinsic stability.
(XLS)

Figure S1 Structural descriptors for the ultra representative proteins. Structural descriptors (and associated standard deviations) for the 3 ultra representative proteins along the first and last 10 ns of the simulated time (1 microsecond) in water at 300K. The red line reports values for the starting conformation. Error bars mark the standard deviation.
(TIF)

Figure S2 Root mean square deviations for the ultra representative proteins. A) RMSd evolution and **B**) distribution among 1 microsecond for the 3 ultra representative proteins in the three environments: water at 300K, urea 8M at 368K and water at 368K. Each color identifies a force-field: red for C22, violet for ON2, blue for P99 and green for P99SBILDN.
(TIFF)

Figure S3 Evolution of solvent accessible areas for the ultra representative proteins. Correlation between solvent accessible surface area (ΔSASA) of polar (left side) and apolar residues (right) and the global structure index. For each force-field, values for the three ultra-representative proteins: 1KTE (green), 1CQY (red) and 1OPC (blue) are reported. **Δ**SASA is defined as the difference to the average values of the corresponding control simulations, the global structure index is used to follow the progress in the unfolding process (from 1 - fully native folded protein - towards 0).
(TIFF)

Figure S4 Comparison between unfolding in hot water and urea for the ultra representative proteins. A) Correlation between the percentages of lost contact time for each residue in urea and in hot water ($r = 0.701$; p-value$<2.2\ 10^{-16}$). The percentage of lost contact time is calculated as contact time lost during 1 microsecond (using water simulation at 300 K as a reference) **B**) Average RMSd measured in different time windows (time lag), from 2 ns up to 200 ns, in hot water (blue) and urea (green). Reference structure for RMSd calculations is always the first frame in the window, which means tha this metrics gives an estimate of the short time scale oscillations of the protein **C**) Force-field dependent distribution of average opening times(temporal unfold – see Suppl. Text S1) in urea (green) and hot water (orange) during the first 100 ns of simulations for the three ultra-representative proteins. **D**) Correlation between the root mean square fluctuation (RMSF) of the residues between simulations in urea (368K) and water (300K). P-value is always smaller than $2.2\ 10^{-16}$.
(TIFF)

Figure S5 Solvent features in urea unfolding simulations. A) Average ratio water/urea molecules in the first solvation shell of the 30 representative proteins in urea (values for every force-field are presented using normal color code). Average values and relative standard deviations are calculated in the last 10 ns of the simulation. To facilitate discussion proteins are grouped according to the SCOP classification, the group average is reported as AV while the symbol * marks proteins with disulfide bonds. Error bars mark the standard deviation. **B**) Distribution of Van der Waals and electrostatic energies for urea and water in the first solvation shell and in the bulk. **C**) Urea and water mean square displacement in different time windows (tau) among the last 10 ns of the trajectories. The diffusion coefficient is calculated using the Einstein equation, more details in Suppl. Text S1.
(PDF)

Figure S6 Examples of urea contacts during protein unfolding. A) Examples of urea-protein contacts along simulation time (μsec). Each dot in the plot defines a contact between that particular urea molecule and a residue in the protein. Examples of urea molecules trapped in the protein core are shown. **B**) Variation along the sequence of the residue RMSF (measure for the flexibility), the contact coefficient CC_{UW} (measure for the preference of protein to contact urea vs. water) and the % of lost contact time (Lost; a measure for the unfolding). The three examples are randomly chosen among the 30 simulations; results are shown for all the three force-field. The RMSF for each residue is calculated in water (green) and in urea (red) while the B-factors (appropriately scaled to maintain same units as RMSF) are from the PDB structure (blue). Residues that are part of protein core are marked in yellow along the x-axes. The color scale for CC_{UW} along the protein sequence ranges from blue (low preference for urea) to orange (large preference for urea). Areas of high urea preference are mostly located in rigid regions flanking highly flexible segments. The % of lost time is calculated as the average percentage of lost time from all the native contacts that each residue forms. The color scale ranges from blue (low unfolding) to magenta (large unfolding).
(TIFF)

Figure S7 A scheme to illustrate the action of urea on micro-folding events. Two residues exposed due to local unfolding oscillation - that quickly re-collapse in water - can remain exposed for longer time in presence of urea. Urea, that has a greater ability than water to form dispersion interactions, can stabilize parts of the protein that are usually hidden from the solvent, such as hydrophobic residues, and that can become exposed during these unfolding oscillation. The summation of many of these events moves the equilibrium towards the unfolding state of a protein.
(TIF)

Table S1 Comparison of structural descriptors for 3 ultra-representative proteins in the periods (10–100 ns) and (910–1000 ns).
(DOCX)

Table S2 Comparison of structural descriptors for 3 ultra-representative proteins in the period (990–1000 ns) calculated in hotwater (HW) and urea (U) and their difference with water (W) among the same period. Values are displayed as mean(standard deviation).
(DOCX)

Table S3 Comparison of % secondary structure for 3 ultra-representative proteins in the period (990–1000 ns) calculated in hotwater (HW), urea (U) and water (W).
(DOCX)

Table S4 Hydrogen bond interactions of urea/water with proteins during the last 10 ns of trajectories for different force-fields. Life-time refers always to the 10 ns window analyzed.
(DOCX)

Text S1 Methods. Description of the analysis performed.
(DOCX)

Acknowledgments

We acknowledge Dr. Anton Feenstra for critical reading of the manuscript and whose suggestions helped improve and clarify it.

Author Contributions

Conceived and designed the experiments: MC MO AP JLG TM. Performed the experiments: MC AP CFC MR. Analyzed the data: MC. Contributed reagents/materials/analysis tools: TM. Wrote the paper: MO MC TM AP CFC MR JLG.

References

1. Kauzmann W (1959) Some factors in the interpretation of protein denaturation. Adv Protein Chem; 14 :1–63
2. England JL, Haran G (2011) Role of solvation effects in protein denaturation: from thermodynamics to single molecules and back. Annu Rev Phys Chem; 62 :257–77
3. Soper AK, Castner EW, Luzar A (2003) Impact of urea on water structure: a clue to its properties as a denaturant? Biophys Chem; 105 (2–3): 649–66.
4. Tirado-Rives J, Orozco M, Jorgensen WL. (1997) Molecular dynamics simulations of the unfolding of barnase in water and 8 M aqueous urea. Biochemistry; 36 (24): 7313–29.
5. Bennion BJ, Daggett V (2003) The molecular basis for the chemical denaturation of proteins by urea. Proc Natl Acad Sci U S A; 100 (9): 5142–7.
6. Auton M, Holthauzen LM, Bolen DW (2007) Anatomy of energetic changes accompanying urea-induced protein denaturation. Proc Natl Acad Sci U S A; 104 (39): 15317–22.
7. Klimov DK, Straub JE, Thirumalai D (2004) Aqueous urea solution destabilizes Abeta(16–22) oligomers. Proc Natl Acad Sci U S A; 101 (41): 14760–5.
8. Canchi DR, García AE (2011) Backbone and side-chain contributions in protein denaturation by urea. Biophys J; 100 (6):1526–33.
9. O'Brien EP, Dima RI, Brooks B, Thirumalai D (2007) Interactions between hydrophobic and ionic solutes in aqueous guanidinium chloride and urea solutions: lessons for protein denaturation mechanism. J Am Chem Soc; 129 (23): 7346–53.
10. Stumpe MC, Grubmüller H. (2007) Interaction of urea with amino acids: implications for urea-induced protein denaturation. J Am Chem Soc; 129 (51): 16126–31.
11. Stumpe MC, Grubmüller H (2008) Polar or apolar–the role of polarity for urea-induced protein denaturation. PLoS Comput Biol; 4 (11): e1000221.
12. Stumpe MC, Grubmüller H (2009) Urea impedes the hydrophobic collapse of partially unfolded proteins. Biophys J; 96 (9): 3744–52.
13. Hua L, Zhou R, Thirumalai D, Berne BJ (2008) Urea denaturation by stronger dispersion interactions with proteins than water implies a 2-stage unfolding. Proc Natl Acad Sci U S A; 105 (44): 16928–33.
14. Lindgren M, Westlund PO (2010) On the stability of chymotrypsin inhibitor 2 in a 10 M urea solution. The role of interaction energies for urea-induced protein denaturation. Phys Chem Chem Phys; 12 :9358–9366
15. Candotti M, Esteban-Martin S, Salvatella X, Orozco M. (2013) Towards an atomistic description of the urea-denatured state of proteins. Proc Natl Acad Sci U S A 2013 ; 110(15):5933–8.
16. Lim WK, Rösgen J, Englander SW (2009) Urea, but not guanidinium, destabilizes proteins by forming hydrogen bonds to the peptide group. Proc Natl Acad Sci U S A; 106 (8): 2595–600.
17. Berteotti A, Barducci A, Parrinello M (2011) Effect of urea on the -hairpin conformational ensemble and protein denaturation mechanism. J Am Chem Soc 133(43) :17200–6
18. Mayor U, Johnson CM, Daggett V, Fersht AR (2000) Protein folding and unfolding in microseconds to nanoseconds by experiment and simulation. Proc Natl Acad Sci U S A; 97 (25): 13518–22
19. 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.
20. Rueda M, Ferrer-Costa C, Meyer T, Pérez A, Camps J et al. (2007) A consensus view of protein dynamics. Proc Natl Acad Sci U S A; 104 (3): 796–801.
21. Meyer T, D'Abramo M, Hospital A, Rueda M, Ferrer-Costa C et al. (2010) MoDEL (Molecular Dynamics Extended Library): a database of atomistic molecular dynamics trajectories. Structure; 18:1399–409
22. Shaw DE, Maragakis P, Lindorff-Larsen K, Piana S, Dror RO et al (2010) Atomic-Level Characterization of the Structural Dynamics of Proteins. Science; 330(6002):341–346.
23. Maiorov VN, Crippen GM (1994) Significance of root-mean-square deviation in comparing three-dimensional structures of globular proteins. J Mol Biol; 235(2):625–34.
24. Courtenay ES, Capp MW, Record MT Jr (2001) Thermodynamics of interactions of urea and guanidinium salts with protein surface: relationship between solute effects on protein processes and changes in water-accessible surface area. Protein Sci; 10(12):2485–97.
25. Stumpe MC and Grubmüller H. (2007) Aqueous Urea Solutions: Structure, Energetics, and Urea Aggregation. J Phys Chem B; 111(22), 6220–6228
26. Lindgren M, Westlund PO (2010) The affect of urea on the kinetics of local unfolding processes in chymotrypsin inhibitor 2. Biophys Chem; 151 (1–2): 46–53.
27. Day R, Beck D, Armen R, Daggett V (2003) A Consensus View of Fold Space: Combining SCOP, CATH, and the Dali Domain Dictionary. Protein Sci; 12: 2150–2160.
28. Bernstein FC, Koetzle TF, Williams GJ, Meyer EE Jr, Brice MD et al. (1977) The Protein Data Bank: A Computer-based Archival File For Macromolecular Structures. J Mol Biol; 112: 535.
29. Hospital A, Andrio P, Fenollosa C, Cicin-Sain D, Orozco M et al. (2012) MDWeb and MDMoby: an integrated web-based platform for molecular dynamics simulations. Bioinformatics; 28 (9): 1278–9.
30. Jorgensen WL, Tirado-Rives J (2005) Molecular modeling of organic and biomolecular systems using BOSS and MCPRO. J Comput Chem; 26: 1689–1700.
31. Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML (1983) Comparison of simple potential functions for simulating liquid water. J Chem Phys; 79: 926–935.
32. Weerasinghe S, Smith PE (2003) Kirkwood-Buff derived force field for mixtures of urea and water. J Phys Chem B; 107: 3891–8
33. Caballero-Herrera A and Nilsson L (2006) Urea parametrization for molecular dynamics simulations. Journal of Molecular Structure: THEOCHEM; 758: 139–148
34. Jorgensen WL, Tirado-Rives J (1988) The OPLS Force Field for Proteins. Energy Minimizations for Crystals of Cyclic Peptides and Crambin. J Am Chem Soc; 110 (6): 1657–1666
35. MacKerell AD, Bashford D, Dunbrack MRL, Evanseck JD, Field MJ et al. (1998) All-atom empirical potential for molecular modeling and dynamics studies of proteins. J Phys Chem B; 102: 3586–3616
36. Wang J, Cieplak P, Kollman PA (2000) How well does a restrained electrostatic potential (resp) model perform in calculating conformational energies of organic and biological molecules. J Comp Chem; 21: 1049–1074.
37. Piana S, Lindorff-Larsen K, Shaw DE (2011) How robust are protein folding simulations with respect to force field parameterization? Biophys J 100(9):L47–L49
38. Essmann U, Perera L, Berkowitz ML, Darden T, Lee H et al. (1995) A Smooth Particle Mesh Ewald Method, J Chem Phys; 103: 8577
39. Phillips J, Braun R, Wang W, Gumbart J, Tajkhorshid E et al. (2005) Scalable molecular dynamics with NAMD. J Comput Chem; 26:1781–1802
40. Andersen HC (1983) RATTLE: A "Velocity" Version of the SHAKE Algorithm for Molecular Dynamics Calculations. Journal of Computational Physics 52: 24–34.
41. Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML (1983) Comparison of simple potential functions for simulating liquid water. J Chem Phys 79(2): 926–935.
42. Martyna GJ, Tobias DJ and Klein ML (1994) Constant pressure molecular dynamics algorithms. J Chem Phys 101(5): 4177
43. Feller SE, Zhang Y, Pastor RW, Brooks BR (1995) Constant pressure molecular dynamics simulation: The Langevin piston method. J Chem Phys; 103(11): 4613
44. Humphrey, Dalke A, Schulten K (1996) VMD - Visual Molecular Dynamics. J Molec Graphics; 14(1): 33–38

Supplementary Information

PDB CODE	Fig 3 symbol	Denaturation Midpoint **	SCOP Class	SEGMENT LENGTH
1AGI	a	Tm = 63° C * human (33%)	alpha-beta	125
1BFG	k	Tm = 64° C	mainly-beta	126
1BJ7	l		mainly-beta	150
1BSN	t		mainly-beta	88
1CHN	b	Tm = 56° C	alpha-beta	126
1CQY	m	Tm = 60° C * B.circulans (33%)	mainly-beta	99
1CSP	n	Tm = 60° C	mainly-beta	67
1CZT	o	Tm= 65.5° C	mainly-beta	160
1EMR	u		mainly-alpha	159
1FAS	b		mainly-beta	61
1FVQ	c		alpha-beta	72
1GND	d		mainly-alpha	105
1I6F	c		alpha-beta	60
1IL6	v	Cm = 5.5 M urea	mainly-alpha	166
1J5D	p	Tm = 69° C	mainly-beta	98
1JLI	w		mainly-alpha	112
1K40	x	Tm = 72° C	mainly-alpha	126
1KTE	e	Tm = 55° C * E.Coli (33%)	alpha-beta	105
1KXA	q		mainly-beta	65
1LIT	f		alpha-beta	131
1LKI	y		mainly-alpha	172
1NSO	r	Cm = 3.4 M urea	mainly-beta	107
1OOI	z	Tm= 50° C	mainly-alpha	124
1OPC	a		mainly-alpha	99
1PDO	g	Tm = 94° C	alpha-beta	129
1PHT	s		mainly-beta	83
1SDF	h		mainly-beta	67
1SP2	d		small	31
1SUR	i		alpha-beta	215
2HVM	j		alpha-beta	273

Extract from the **Dataset 1.** The 30 most populated folds according to SCOP,. When available, the denaturation midpoint is included as measurement of protein intrinsic stability. **Denaturation midpoint defined as the temperature (Tm in °C) or denaturant concentration (Cm in M); * values reported for an homologous, the similarity to the one in the original species is reported in brackets.

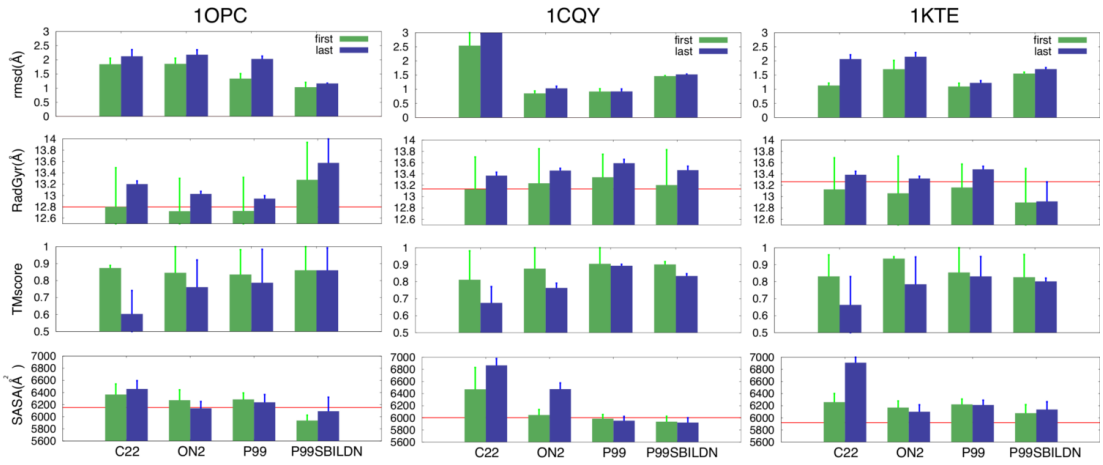


Figure S1. Structural descriptors for the ultra representative proteins. Structural descriptors (and associated standard deviations) for the 3 ultra representative proteins along the first and last 10 ns of the simulated time (1 microsecond) in water at 300K. The red line reports values for the starting conformation. Error bars mark the standard deviation.

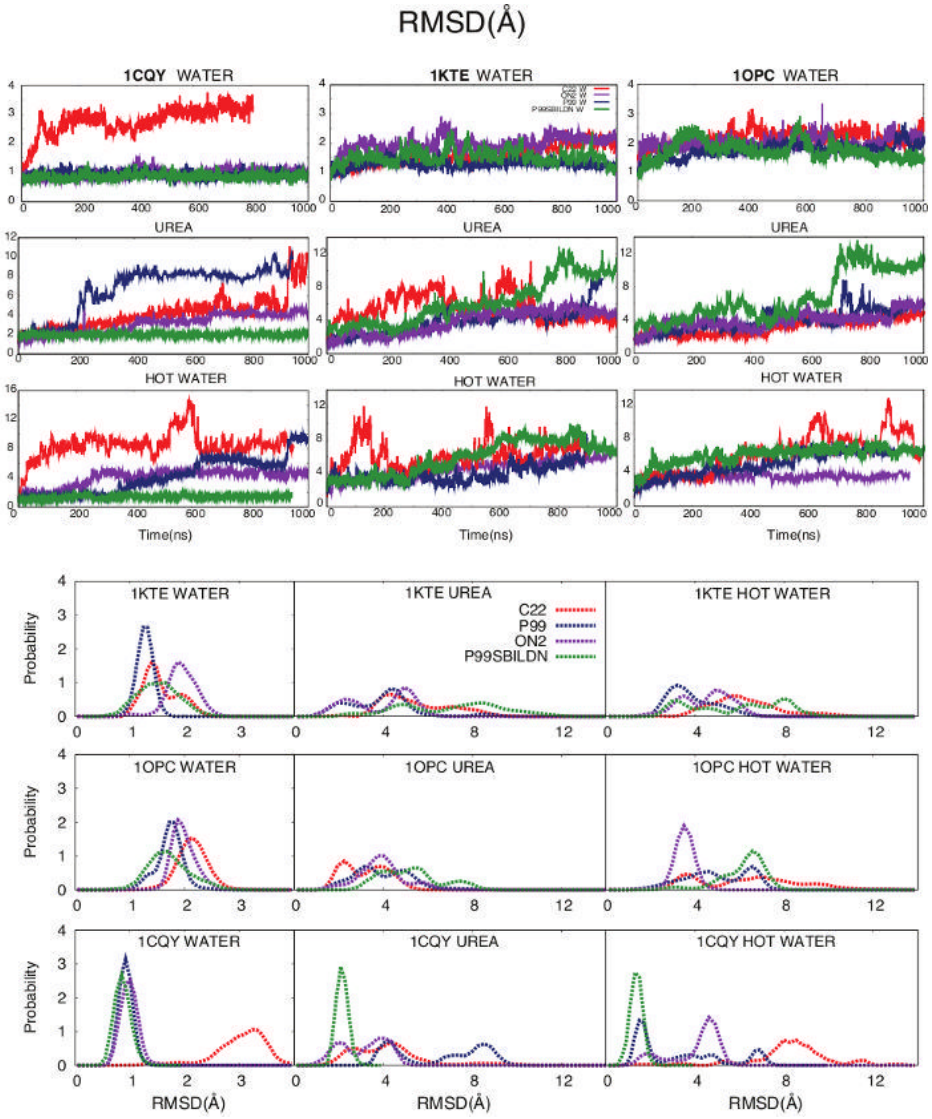


Figure S2. Root mean square deviations for the ultra representative proteins. A) RMSd evolution and B) distribution among 1 microsecond for the 3 ultra representative proteins in the three environments: water at 300K, urea 8M at 368K and water at 368K. Each color identifies a force-field: red for C22, violet for ON2, blue for P99 and green for P99SBILDN.

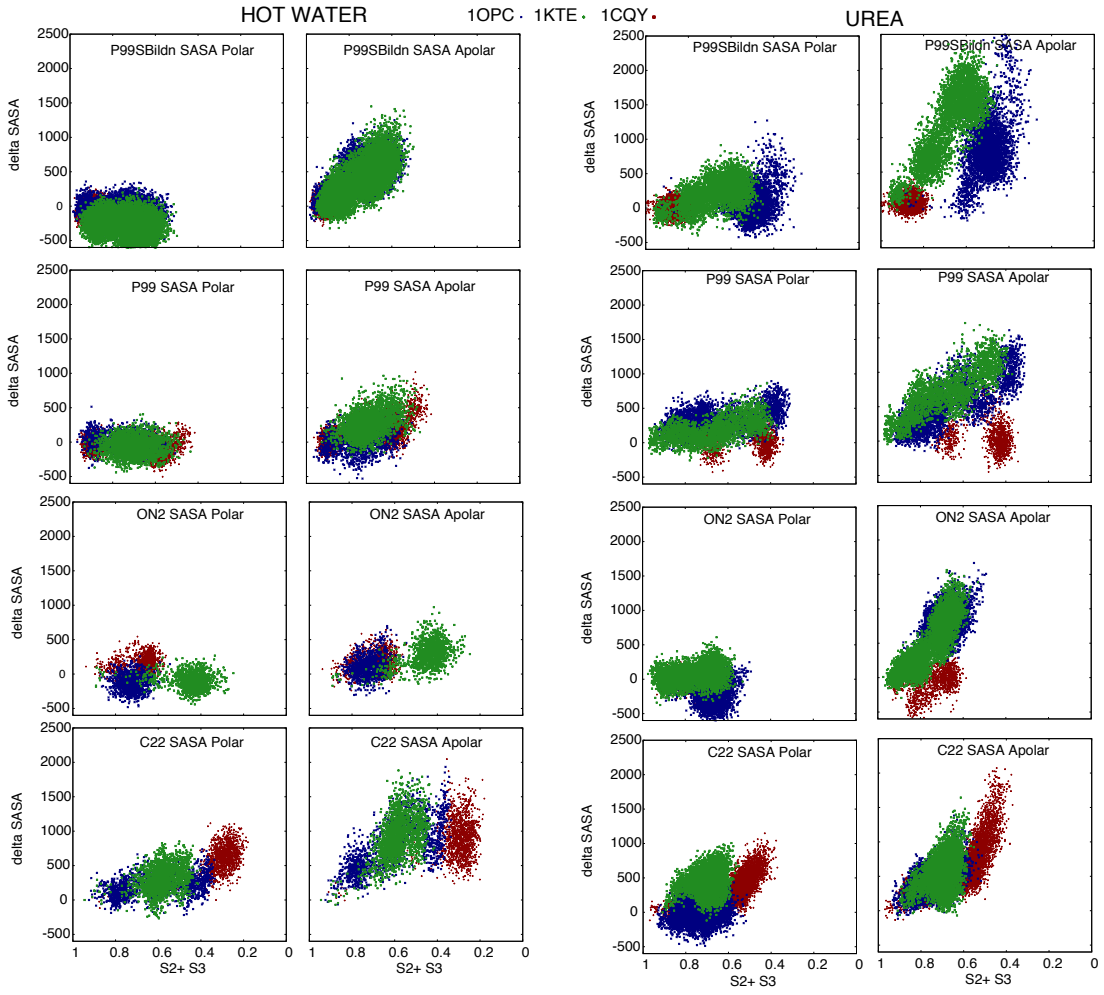


Figure S3. Evolution of solvent accessible areas for the ultra representative proteins. Correlation between solvent accessible surface area (ΔSASA) of polar (left side) and apolar residues (right) and the global structure index. For each force-field, values for the three ultra-representative proteins: 1KTE (green), 1CQY (red) and 1OPC (blue) are reported. ΔSASA is defined as the difference to the average values of the corresponding control simulations, the global structure index is used to follow the progress in the unfolding process (from 1 - fully native folded protein - towards 0).

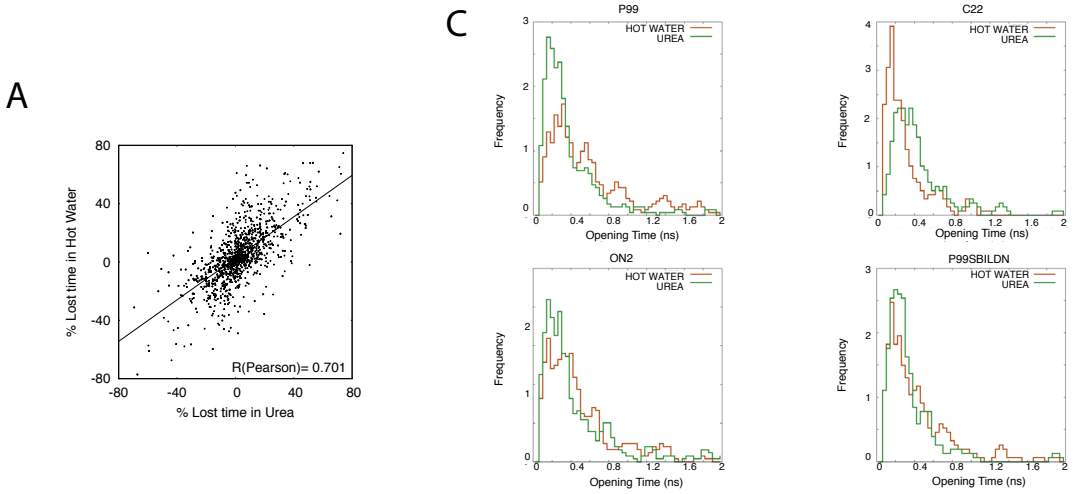
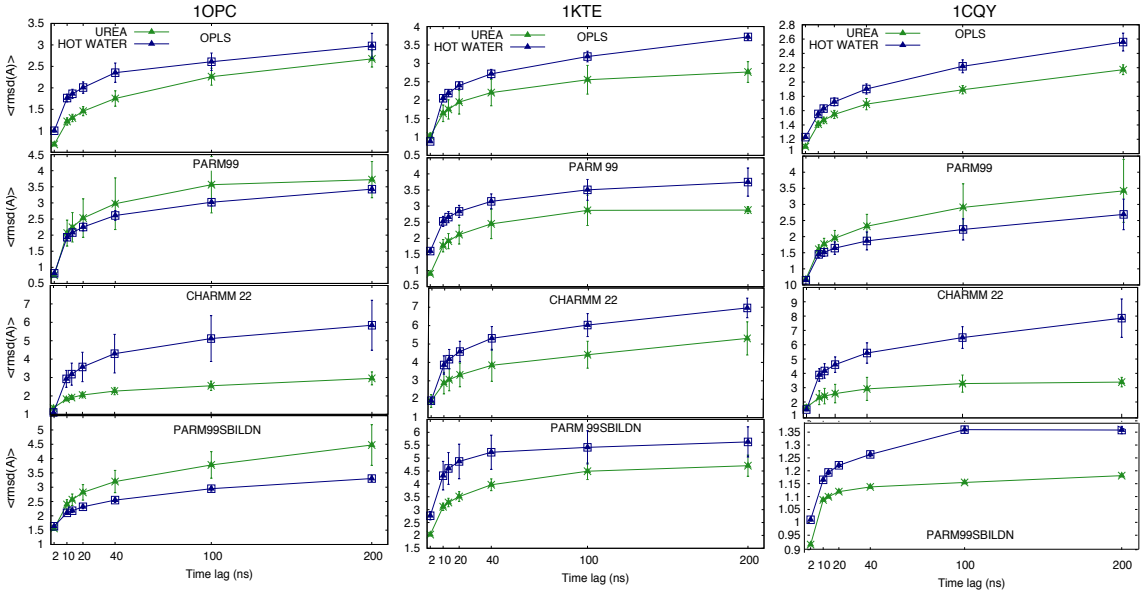
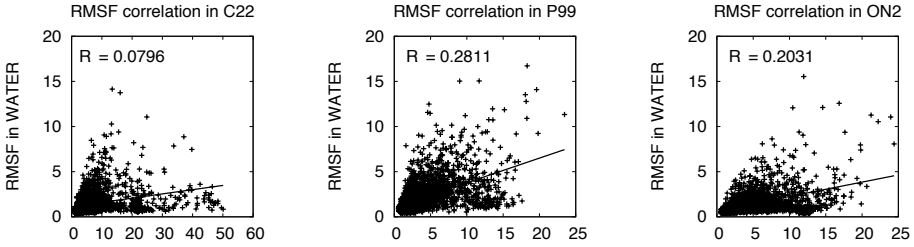


Figure S4. Comparison between unfolding in hot water and urea for the ultra representative proteins. A) Correlation between the percentages of lost contact time for each residue in urea and in hot water (r=0.701; p-value<2.2 10−16). The percentage of lost contact time is calculated as contact time lost during 1 microsecond (using water simulation at 300 K as a reference) B) Average RMSd measured in different time windows (time lag), from 2 ns up to 200 ns, in hot water (blue) and urea (green). Reference structure for RMSd calculations is always the first frame in the window, which means that this metrics gives an estimate of the short time scale oscillations of the protein C) Force-field dependent distribution of average opening times(temporal unfold – see Suppl. Text S1) in urea (green) and hot water (orange) during the first 100 ns of simulations for the three ultra-representative proteins. D) Correlation between the root mean square fluctuation (RMSF) of the residues between simulations in urea (368K) and water (300K). P-value is always smaller than 2.2 10−16.

B


1OPC
1KTE
1CQY
UREA
HOT WATER
OPLS
PARM99
CHARMM 22
PARM99SBILDN
<rmsd(A)>
Time lag (ns)

D


RMSF correlation in C22
R = 0.0796
RMSF correlation in P99
R = 0.2811
RMSF correlation in ON2
R = 0.2031
RMSF in WATER

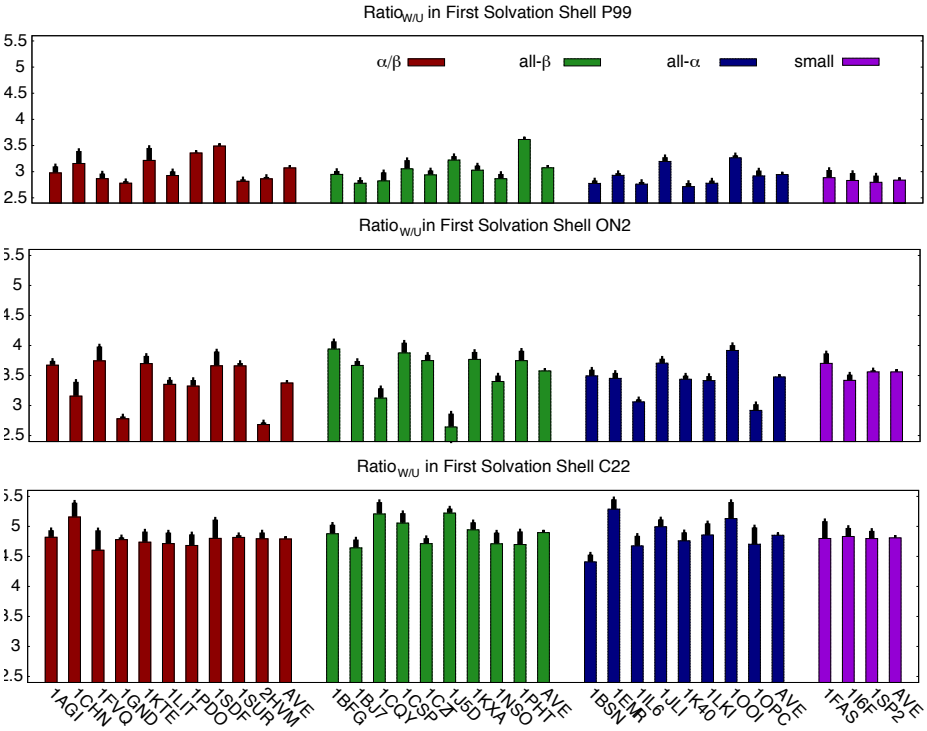


Figure S5. Solvent features in urea unfolding simulations. A) Average ratio water/urea molecules in the first solvation shell of the 30 representative proteins in urea (values for every force-field are presented using normal color code). Average values and relative standard deviations are calculated in the last 10 ns of the simulation. To facilitate discussion proteins are grouped according to the SCOP classification, the group average is reported as AV while the symbol * marks proteins with disulfide bonds. Error bars mark the standard deviation. B) Distribution of Van der Waals and electrostatic energies for urea and water in the first solvation shell and in the bulk. C) Urea and water mean square displacement in different time windows (tau) among the last 10 ns of the trajectories. The diffusion coefficient is calculated using the Einstein equation, more details in Suppl. Text S1.

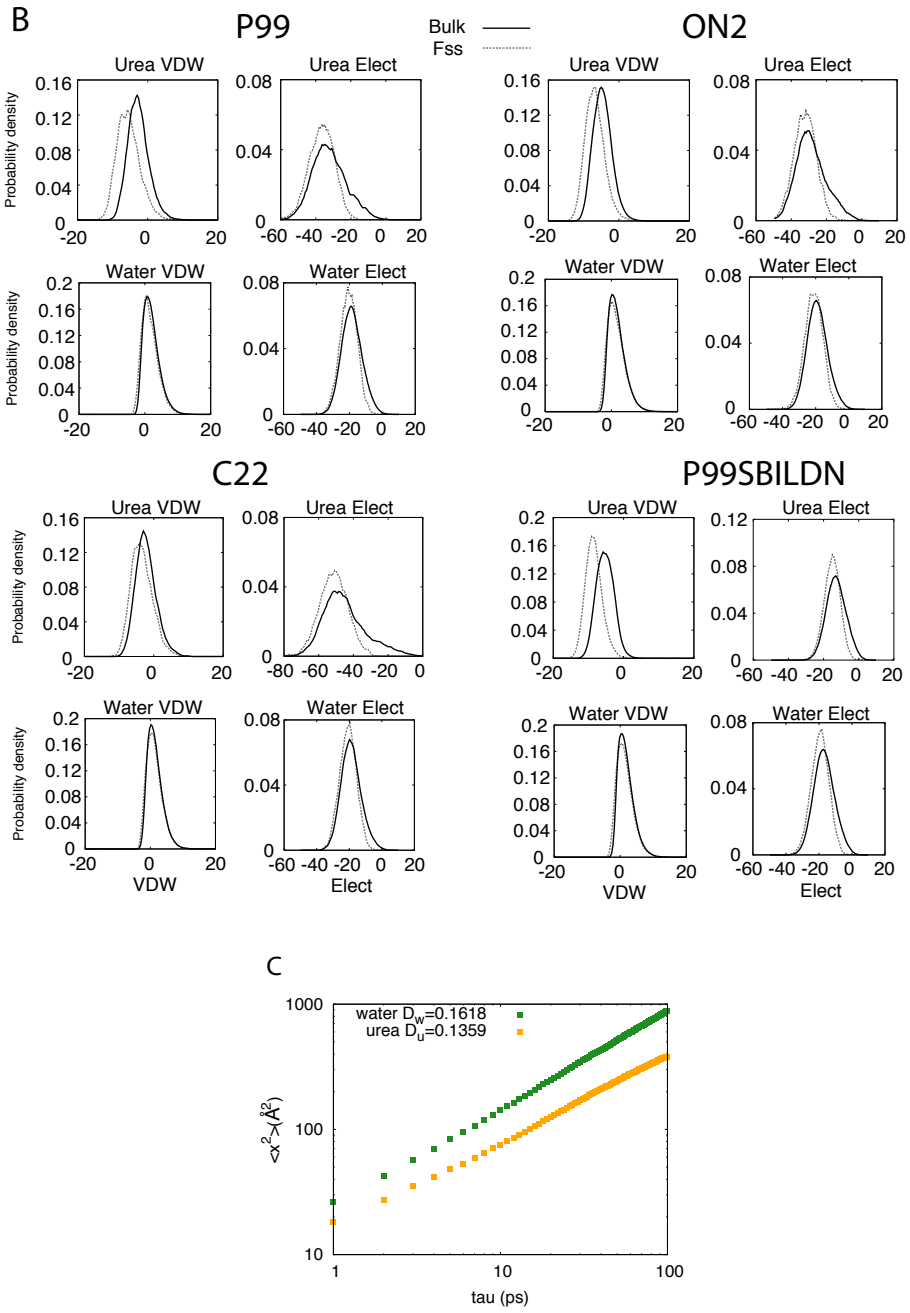
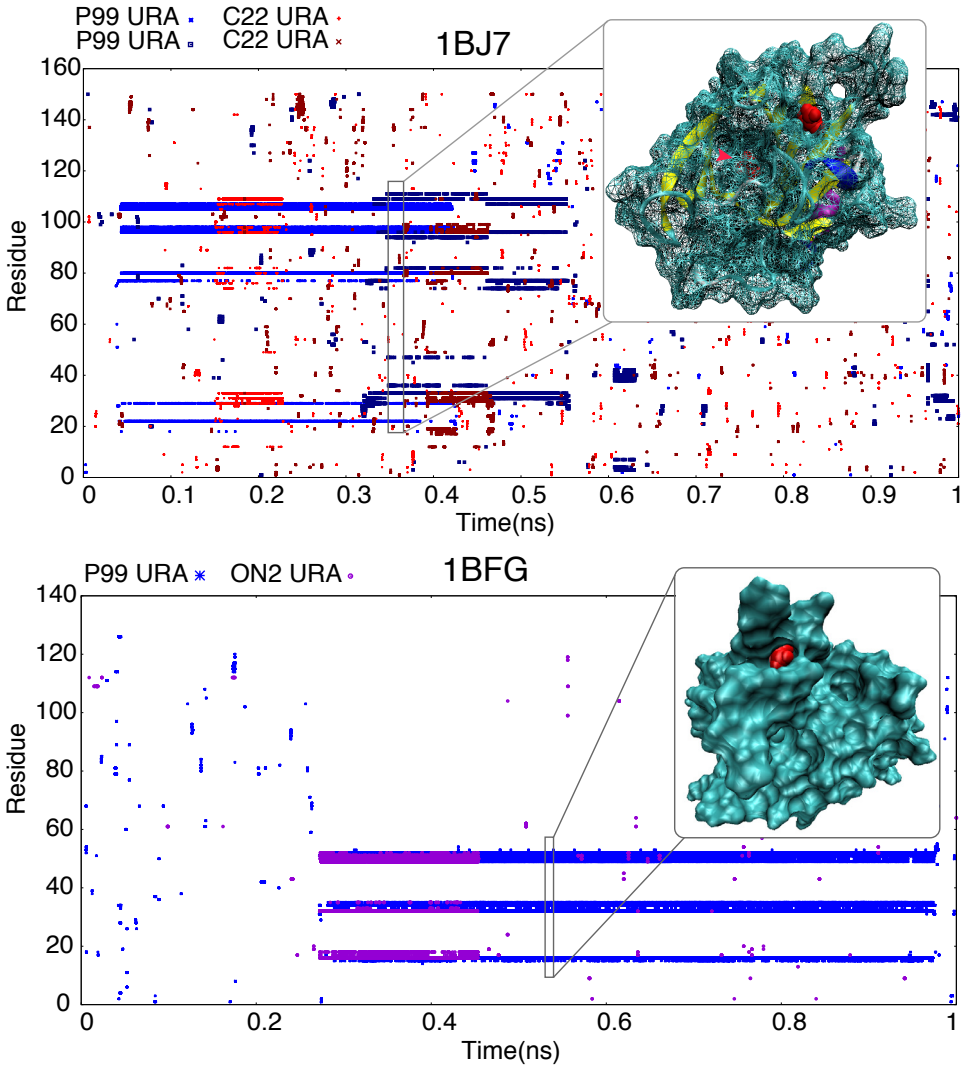
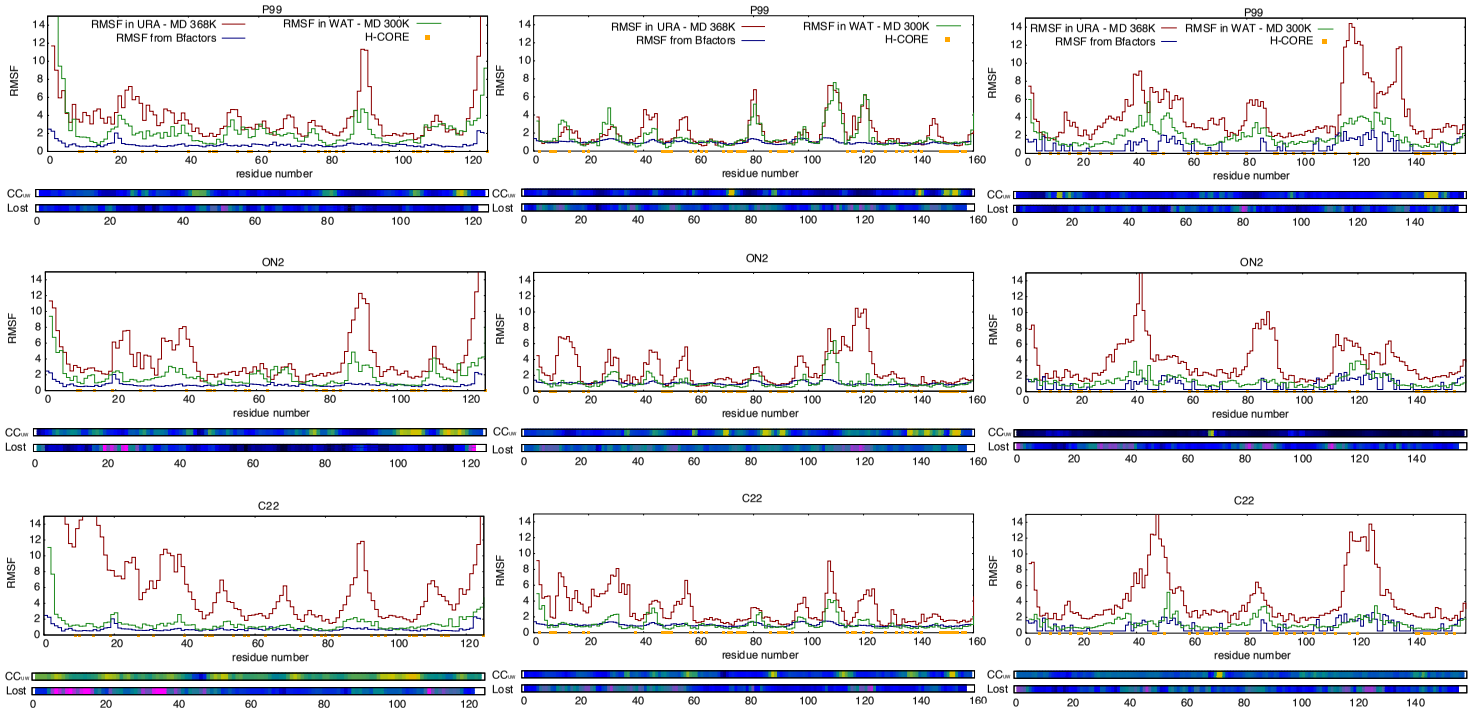

B
P99
Bulk
Fss
ON2
Urea VDW
Urea Elect
Water VDW
Water Elect
Probability density
C22
P99SBILDN
VDW
Elect
C
water D_w=0.1618
urea D_u=0.1359
<x^2> (Å^2)
tau (ps)

Figure S6. Examples of urea contacts during protein unfolding. A) Examples of urea-protein contacts along simulation time (μsec). Each dot in the plot defines a contact between that particular urea molecule and a residue in the protein. Examples of urea molecules trapped in the protein core are shown.





B) Variation along the sequence of the residue RMSF(protein flexibility), the contact coefficient CC_{UW} (protein contacts with urea), and the % of lost contact time (protein local unfolding). The RMSF for each residue is calculated in water (green) and in urea (red) while the B-factors are from the PDB structure (blue). Residues that are part of protein core are marked in yellow along the x-axes. The color scale for CC_{UW} along the protein sequence ranges from blue (low preference for urea) to orange (large preference for urea). Areas of high urea preference are mostly located in rigid regions flanking highly flexible segments. The % of lost time is calculated as the average percentage of lost time from all the native contacts that each residue forms. The color scale ranges from blue (low unfolding) to magenta (large unfolding).

Structural Descriptor		FIRST (10-100ns)	LAST (910-1000ns)	Δ(Last-First)	Z-score
Rmsd(Å)	C22	1.84(0.29)	2.54(0.19)	0.70(0.09)	1.34
	ON2	1.47(0.20)	1.78(0.13)	0.31(0.06)	0.77
	P99	1.12(0.13)	1.39(0.09)	0.27(0.04)	0.63
	P99*	1.35(0.08)	1.46(0.02)	0.11(0.05)	0.34
Tmscore	C22	0.84(0.10)	0.65(0.13)	-0.19(0.03)	-1.31
	ON2	0.89(0.12)	0.77(0.12)	-0.12(0.00)	0.78
	P99	0.87(0.16)	0.84(0.11)	-0.03(-0.05)	0.56
	P99*	0.86(0.11)	0.83(0.06)	-0.03(-0.05)	0.46
R_g (Å)	C22	13.02(0.61)	13.32(0.06)	0.30(-0.54)	0.59
	ON2	13.00(0.62)	13.27(0.04)	0.26(-0.57)	0.68
	P99	13.08(0.47)	13.34(0.06)	0.26(-0.41)	0.69
	P99*	13.12(0.63)	13.32(0.28)	0.19(-0.35)	0.43
SASA (Å^2)	C22	6365(225)	6744(147)	378(-80)	1.23
	ON2	6164(123)	6237(110)	74(-13)	0.89
	P99	6165(88)	6134(93)	-30(5)	0.23
	P99*	5984(107)	6050(147)	66(40)	0.65
% Helix	C22	25.3	25.3	0.0	
	ON2	24.6	25.9	1.3	
	P99	25.0	24.6	-0.4	
	P99*	25.3	25.6	0.3	
% Sheets	C22	27.9	24.3	-3.6	
	ON2	29.7	28.9	-0.8	
	P99	29.9	27.8	-2.1	
	P99*	30.1	30.0	-0.1	

Table 1. Comparison of structural descriptors for 3 ultra-representative proteins in the periods (10–100 ns) and (910–1000 ns) and their difference (Δ(Last-First)). When possible values are displayed as mean(standard deviation) and the Z-score has been calculated of the Δ(Last-First) value related to the differences between non consecutive windows of 10 ns.

		Hot Water(HW) (990-1000ns)	Urea (U) (990-1000ns)	ΔHW-W)	Δ(U-W)
Rmsd (Å)	C22	7.35(1.25)	6.01(0.33)	4.68(1.10)	3.33(0.18)
	ON2	4.91(0.24)	5.11(0.19)	3.15(0.00)	3.36(-0.04)
	P99	7.49(0.51)	7.64(0.68)	6.13(0.44)	6.28(0.61)
	P99*	4.89(0.20)	7.74(0.32)	3.72(0.10)	6.58(0.22)
Tmscore	C22	0.40(0.06)	0.53(0.07)	-0.24(-0.05)	-0.11(-0.03)
	ON2	0.60(0.11)	0.60(0.02)	-0.18(0.027)	-0.18(-0.05)
	P99	0.45(0.06)	0.41(0.11)	-0.39(-0.01)	-0.44(0.04)
	P99*	0.48(0.10)	0.58(0.05)	-0.39(-0.00)	-0.28(-0.05)
R_g(Å)	C22	13.96(0.31)	14.27(0.21)	0.37(0.20)	0.68(0.10)
	ON2	13.50(0.14)	13.91(0.09)	0.13(0.06)	0.55(0.02)
	P99	13.67(0.24)	13.79(0.16)	0.35(0.18)	0.47(0.10)
	P99*	13.64(0.15)	14.93(0.28)	0.29(0.08)	1.58(0.20)
SASA(Å^2)	C22	7418(295)	7302(176)	691(152)	575(33)
	ON2	6372(149)	6822(125)	126(55)	576(31)
	P99	6607(190)	7007(204)	516(124)	915(138)
	P99*	6570(160)	7517(210)	533(59)	1479(108)
S_2 (%)	C22	56.52(2.76)	66.57(2.53)	-17.56(0.95)	-7.52(0.72)
	ON2	58.10(2.28)	64.48(2.02)	-20.6(0.10)	-14.22(-0.15)
	P99	64.24(3.06)	51.34(1.24)	-18.35(1.3)	-31.26(-0.52)
	P99*	74.59(1.41)	55.10(2.01)	-12.0(-0.34)	-31.50(0.24)
S_3(%)	C22	36.39(5.28)	51.95(8.6)	-43.27(3.04)	-27.71(6.37)
	ON2	58.83(6.57)	64.13(5.29)	-22.87(4.39)	-17.57(3.11)
	P99	46.20(9.44)	32.66(5.14)	-39.72(7.68)	-53.26(3.38)
	P99*	69.43(5.16)	45.56(4.27)	-17.53(3.55)	-41.41(2.66)

Table S2. Comparison of structural descriptors for 3 ultra-representative proteins in the period (990-1000 ns) calculated in hotwater(HW) and urea (U) and their difference with water (W) among the same period. Values are displayed as mean(standard deviation).

		Hot Water(HW)	Urea (U)	Water(W)
		% alpha / beta	% alpha / beta	% alpha / beta
1KTE	C22	25 / 10	37 / 12	41 / 17
	ON2	14 / 22	26 / 17	41 / 16
	P99	34 / 12	24 / 9	39 / 17
	P99*	28 / 20	11 / 9	39 / 20
1OPC	C22	26 / 4.0	32 / 11	34 / 18
	ON2	28 / 18	27 / 18	31 / 23
	P99	32 / 16	26 / 5	35 / 19
	P99*	28 / 14	25 / 10	33 /23
1CQY	C22	1 / 22	0 / 30	0 / 52
	ON2	0 / 46	0 / 39	2 /56
	P99	9 / 31	7 / 34	1 /48
	P99*	2 / 46	2 / 46	3 / 48

Table S3. Comparison of % secondary structure for 3 ultra-representative proteins in the period (990-1000 ns) calculated in hotwater(HW), urea (U) and water (W).

a) OPLS

H- bonds*:	Urea / Water as H-donor		Urea /Water as H-acceptor	
% of total	64 / 62		36 / 38	
H-bonds with protein:	**BackBone**	**SideChains**	**BackBone**	**SideChains**
% of total	64 / 46	36 / 54	70 / 72	30 / 28
Lifetime %	0.69 / 0.58	1.08 / 0.80	4.25 / 1.01	4.18 / 3.79

b) CHARMM

H- bonds*:	Urea / Water as H-donor		Urea / Water as H-acceptor	
% of total	66 / 60		34 / 30	
H-bonds with protein:	**BackBone**	**SideChains**	**BackBone**	**SideChains**
% of total	53 / 55	47 / 45	64 / 69	36 / 31
Lifetime %	0.84 / 0.63	1.28 / 0.72	2.75 / 0.56	4.71 / 2.29

c) PARM 99

H- bonds*:	Urea / Water as H-donor		Urea / Water as H-acceptor	
% of total	69 / 65		31 / 35	
H-bonds with protein:	**BackBone**	**SideChains**	**BackBone**	**SideChains**
% of total	55 / 44	45 / 56	67 / 69	33 / 31
Lifetime %	0.73 / 0.66	0.94/ 0.68	2.67 / 0.54	3.94 / 2.09

Table S4. Hydrogen bond interactions of urea / water with proteins during the last 10 ns of trajectories for different force-fields. Life-time refers always to the 10 ns window analyzed.

Supplementary Text S1 - Methods

Analysis. Trajectories were analyzed using a variety of metrics. Protein structural descriptor include Root Mean Square Deviation (RMSD), TMscore, Radius of Gyration (RadGyr), Secondary Structure (SS - evaluated using STRIDE [1]), Solvent Accessible Surface Area (SASA - evaluated using NACESS [2]). The average RMSD was measured in different time windows, with a time lag from 2 ns up to 200 ns, in water and urea at 368K and always using as reference structure the first frame in the window.

To describe the unfolding , we calculated the change of protein features taking as the reference the native state described by the control simulation at 300K in water. Therefore the native contacts (tertiary structure) and native secondary structure were calculated as those occurring for more than 80% of the time in the control simulation, while the protein core was considered formed by residues with an average SASA and standard deviation lower then 10 $Å^2$ in the control simulation.

For the trajectories in urea and water at 368K we calculated the secondary structure index "S2" as the existing fraction of native secondary structure (see above) in each frame, and the tertiary structure index "S3" as the existing fraction of native contacts in each frame. Residues were considered to be in contact when their interesidue distance was shorter than 3.5 Å [3]. The global structure index [4] was defined as the sum of S2 and S3.

Regarding the stability of intra-protein contacts, we considered as lost contacts those with a reduced contact time in urea or water at 368K compared to water simulation at 300 K (reduction for more than 30% of the simulated time). The % of lost time for a residue was calculated as the average percentage of lost contact time for all the native contacts involving that residue, during 1 microsecond in urea or water at 368K and using water simulation at 300K as a reference. The flexibility of the contacts and the average opening time was calculated for each native contact at each snapshot in the first 100ns of hot water and urea simulations. This first part of the simulation contains the largest number of comparable contacts (see below), in later stages most of the contacts are generally unstable at least in one of the environments and therefore a comparison would be uninformative. A contact was considered "open" if the minimum distance between heavy atoms was larger than 5 Å and "closed" if the distance was smaller than 4 Å. In the moonlight zone (between 4 and 5 Å) the contact assumed the state of the previous frame, avoiding ambiguous classifications. We focused the analysis on comparable contacts that are still preserved in both urea and water at 368K (difference in contact time is less than 20% compared to water at 300 K, in both simulations). Contacts that are completely lost or fully maintained in at least one of the two environments were removed because they are uninformative regarding to changes in flexibility. We calculated the $rmsf_{SC}$ as the rmsf for a single sidechain after an alignment based only on the backbone of the same residue- thus the metric is only dependent on the local motion. The difference of $rmsf_{SC}$ between water and urea at 368K was used to evaluate the change in sidechain dynamics. We excluded differences smaller than 0.5 to avoid the comparison of residues with similar flexibility. Therefore the analysis was performed on values for Δ $rmsf_{SC}$ ($rmsf_{SC}$ in water − $rmsf_{SC}$ in urea) larger than 0.5 or smaller than -0.5.

Solvent features evaluated here include water/urea ratio in first solvation shell

(FSS; solvent molecules within 5 Å of the protein) and in the bulk (solvent molecules with a distance to the protein larger than 6Å). More detailed analysis were performed using the contact coefficient CC_{UW} metric. CC_{UW} is the ratio for each aminoacid between contacts with urea and with water molecules normalized with the total numbers of urea and water atoms [3]; a contact is formed when at least two heavy atoms are closer than 3.5 Å. The residence time for urea and water molecules during 1 microsecond trajectory was calculated as the time each solvent molecule is in contact (see previous definition of contact) with the same residues without any interruptions. Urea and water mean square displacements were calculated in different time windows (tau) among the last 10 ns of the trajectories. We used the Einstein equation [5] to calculate the diffusion coefficient (D) from the slope of the fitting line. Since the Einstein relation is valid as time approaches infinity, we used only the last half of values for the fitting. Solvent-protein hydrogen bonds were annotated with a heavy atom cutoff distance of 3.5 Å and a donor-hydrogen-acceptor angle greater than 120 degree. Stable H-bonds were defined as those detected for more than 5% of the analyzed time. Interaction energies for urea and water in the FSS and bulk were computed following Hua et al. [6] using a 13.0Å spherical cutoff . All the analyses were perform with MDWEB [7], VMD [8] , Ptraj [9] and in house software, while statistical analysis were performed with R [10].

Supplementary Bibliography

1. Heinig, M., Frishman, D. (2004). STRIDE: a Web server for secondary structure assignment from known atomic coordinates of proteins. Nucl. Acids Res. , 32, W500-2.

2. Hubbard,S.J.& Thornton, J.M. (1993), 'NACCESS', Computer Program, Department of Biochemistry and Molecular Biology, University College London."

3. Stumpe M. C. and Grubmüller H. (2007) Interaction of Urea with Amino Acids - Implications for Urea-Induced Protein Denaturation. *J.Am.Chem.Soc.*129(51):16126-31

4. Simms A.M., Toofanny R.D., Kehl C., Benson N.C., and Daggett V. (2008) Dynameomics: design of a computational lab workflow and scientific data repository for protein simulations. Protein Engineering Design & Selection 21: 369-377.

5. Allen, M. P., Tildesley, D. J. (1987) Computer Simulations of Liquids. Oxford: Oxford Science Publications.

6. Hua L,; Zhou R,; Thirumalai D,; Berne BJ. (2008) Urea denaturation by stronger dispersion interactions with proteins than water implies a 2-stage unfolding. Proc Natl Acad Sci U S A. 2008 Nov 4;105(44):16928-33

7. Hospital A, Andrio P, Fenollosa C, Cicin-Sain D, Orozco M, Gelpí JL. (2012) MDWeb and MDMoby: an integrated web-based platform for molecular dynamics simulations. Bioinformatics 28(9):1278-9.

8. Humphrey, W., Dalke, A. and Schulten, K., ``VMD - Visual Molecular Dynamics" *J. Molec. Graphics* **1996**, *14.1*, 33-38.

9. D.A. Case, et al (2012), *AMBER 12*, University of California, San Francisco.

10. R Development Core Team (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

"The hardest thing to see is what is in front of our eyes"

Johan Wolfgang von Goethe

CHAPTER 5

Macromolecular crowding and the physiological environment of proteins.

The diluted solutions generally employed by conventional biochemical experiments, and by most simulation studies are far from the physiological environment. The inside of a cell, the cytoplasm, is instead a very dense and inhomogeneous environment. For example, the bacterial cytoplasm is composed by protein and nucleic acids at a voume fraction of typically 20–30% (200-320 g/L of proteins[1] , 75-120 g/L of RNA and 11-18 g/L of DNA), various metabolites and inorganic ions at several concentrations, and water in the remaining space (~70%) (**Figure 5.1**). The case of eukaryotic cells is further complicated by the presence of organelles that

1 Of which ~10% are cytoskeletal filaments and ~90% are soluble globular proteins

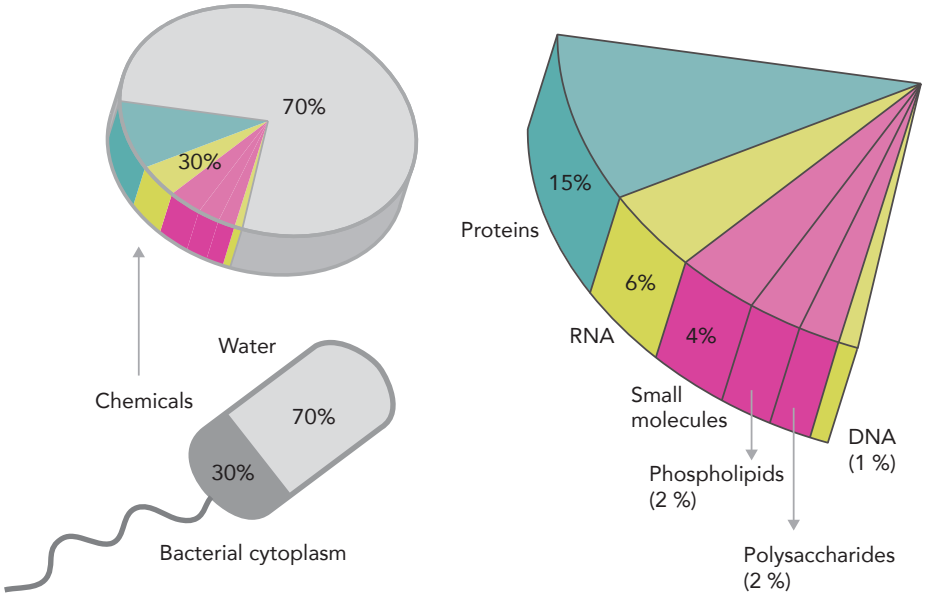


Figure 5.1. The composition of a bacterial cell. Most of the cell is water while the remaining 30% contains varying proportions of molecules; among those proteins prevail. Color code: blue for proteins, yellow for nucleic acids and magenta for the rest.

create separated compartments, and by a more extensive cytoskeleton. In their cytosol, the part of the cytoplasm outside the organelles, macromolecules occupy between 10–40% of the volume and the concentration varies with cell type (50-250 g/L of proteins and 20–50 g/L of nucleic acid). In both cases, citing Katherine Luby-Phelps, "the cell cytoplasm is more like a crowded party in a house full of furniture than a game in an empty field"[1]. Intuitively, many physical properties of the cytoplasm differ form a dilute solution: it has a higher viscosity and a reduced dielectric constant[2] compared to the infinite dilution limit. However at microscopic level most of the water molecules (~85%) still behaves similarly to pure solution (bulk) and only 15% of the water molecules has altered mobility (2-fold slower diffusion) [1]–[3].

Beside the changes in the solution properties, the most evident restric-

2 Due to the deplete the number of polarizable water molecules surrounding the proteins

tions of the cytoplasm are related to the mere presence of other macromolecules (crowding). One of the most academic definition of 'macromolecular crowding effects' was provided by Zhou, Rivas and Minton in a seminal review [4], which defines it as the alterations caused by "macromolecular cosolutes that are nominally inert with respect to the reaction of interest", where the term 'inert' implies the only interaction between the crowder and the other macromolecular components of the system is an excluded-volume (i.e. steric) interaction. This assumption has guided the first tentative to rationalize crowding which kept the volume exclusion as main player. The available volume modulates for example the effective concentration of a protein, with repercussions on its thermodynamic properties. Simple statistical thermodynamic models can, then, be used to predict the consequences on several processes such as protein folding, protein-protein association, conformational isomerization, enzyme activity and stability with respect to denaturation. In these models the structures are largely simplified and often described as spherical objects. For example, protein folding can be model as simple random walk in the presence of sphere obstacles [5] while the unfolded state is comparable to a compressable sphere [6]. By fine-tuning the size and the shape of both the crowders and the protein states, one can understand the impact of such variables in the model. The results are usually confirmed by experiments employing artificial crowding agents, which mimic the pure volume excluded effect [7]. These chemicals, such as dextrans, Ficoll and polyethylene(glycol) (PEG), are non-charged polymers that are expected to occupy space without interacting with proteins. Recently, many doubts have arose about the real inert character of crowding agents, questioning whether or not they reveal physiologically relevant information [8],[9].

However inertness doesn't appear to be a feature of physiological crowding: already in the 80s McConkey suggested that in vivo the transient interactions that a protein form with the surrounding macromolecules could impact the protein structure. As important source of constraints in the spatial rearrangements of atoms, McConkey referred to them as "**quinary structure**", following the scheme proposed by Linderstrøm-Lang [10]. To clearly discern the impact of such interactions on protein structure is

not trivial and needs accurate tools to mimic or preserve the physiological environments. Thanks to recent advances in NMR spectroscopy, we can now directly observe the behavior of proteins inside a cell or in cell-like environments (reconstituted cytosol or in solution with protein crowders) [11]. The introduction of isotopes (15N 13C 19F) on the test protein distinguishes its signal from the un-enriched (and therefore silent) proteic crowders. Unfortunately, the collected spectra are of difficult interpretation [11]: the quinary interactions and the high viscosity hinders the fast tumbling (i.e. rotation motion) of proteins resulting in a broader signal, especially evident for globular proteins (IOPs) [12,13].

It is clear then that macromolecular crowding can have severe repercussion on the structure and consequently on the stability of proteins. In the case of synthetic crowding agents, a simple estimation of the protein stability can be extracted from the melting temperature T_m or its free energy of unfolding ΔG **(Figure 4.1)** upon the addition of crowding agents. In cell-like environments instead it can be estimated by the opening free energy, measured by the NMR-detected amide proton exchange. However the measured effects on protein stability often disagree between synthetic crowder and cell-like environments, and depend dramatically on the protein under study (**Table 5.1**) [3], [13]–[15]. Together these contradictions hinder the extractions of general rules about the effect of crowding on protein structure, dynamics and stability.

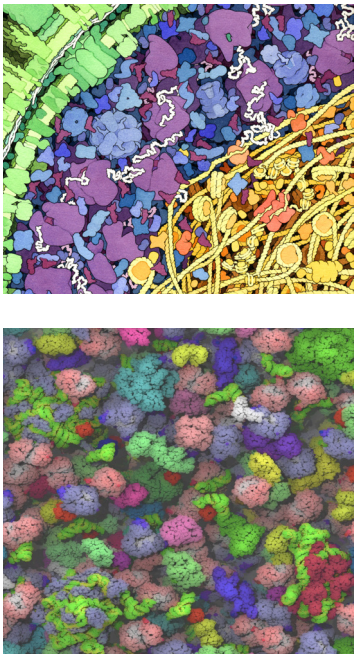


Figure 5.2. Representations of the crowded cytoplasm. From the top: Goodsell's rendering of the cellular environment [28] and a snapshot from a Brownian dynamics simulation of the cytoplasma of E.coli [27].

In silico simulations can be useful to resolve these apparent contradictions. In simulations, similarly to experiments, the degree of realism can vary too. At the simplest level the crowded environment can be included via coarse-grained and spheroidal models of

Type	Protein	Crowding	Effect on stability	Reference
IOPs	DNase I	PEG500	+ stability	Sasaki et al 2007 [16]
	C12	Ficoll	+ stability	Benton et al 2012, [17]
		Protein;	Destabilized	Sarkar et al. 2013 [18]
		E.coli lysate	Destabilized	
	GB1	In-cells	+ stability	Monteith et al. 2014, [19]
IDP	RTX (Ca2+)	Ficoll70	Stabilization of both apo and holo	Sotomayor-Perez et al. 2013 [20]
	Alpha synuclein	In cell	Remain disordered	Waudby et al, 2013 [21]
		vv crowders	Remain disordered	Munishkina et al 2004 [22]
	N-protein of bacteriophage λ	Protein BPTI	+ stability to compact conformations	Johansen et al, 2011 [23]
MGPs	Human α-lactalbumin (HLA)	Ficoll 70	+ stability	Zhang et al, 2012 [24]
		Dextran 70	+ stability	
		PEG 2000	destabilized (apo)	

Table 5.1 Crowding effects on protein stability. Collection of some of the results reported in literature for several proteins types, classified according to the crowding agent used and the obserbed effect (+: increased).

biomolecules [25] or be implicit represented by a low-dielectric continuum models [26]. A more realistic model consists of introducing in the simulations various biomolecules that mimic the biological diversity. Pioneer was the work of Elcock and coworker that recreated a model of the cytoplasm of E.Coli including the 50 most populated proteins (275 g/L) [27] **(Figure 5.2)**. Their Brownian dynamics simulation in implicit solvent allowed computing the cytoplasm's effects on the thermodynamics of protein folding, association and aggregation events. However it was thanks to

some recent MD simulation in explicit solvent that proteic crowders were linked to minor sub-populations of non-native states and structural perturbations that range from subtle changes to partial denaturation [29]–[31]. The third project, part of this thesis, follow this latter approach and aims to study by means of MD simulations in explicit solvent a crowded system which included proteins with different conformational landscapes. By studying the crowding effects with models that include more detailed information on proteins might help the design of better approximations.

5.1. Crowding and protein landscapes (Publication 3).

After the experience gained through the urea-unfolding projects, we decided to apply a similar consistent approach to address the issue of macromolecular crowding. We set up a system ad-hoc to extract information on the effect of proteic crowder on proteins with different energy landscape (i.e. MG, IDP, IOP), see **Figure 5.3.** To distinguish between universal effects related to crowding and specific ones related to proteic

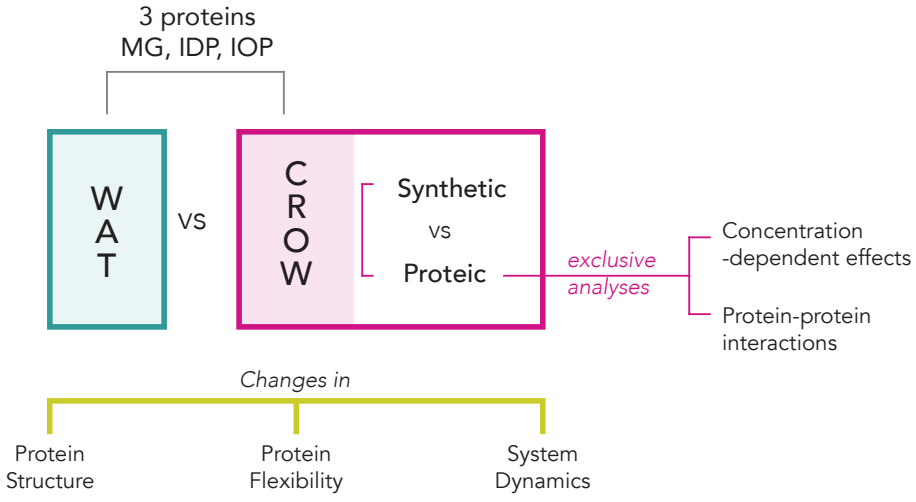


Figure 5.3. Schematic overview of the project. Three proteins with different energy landscape features were simulated in two crowding environments (proteins and PEG500) and compared to control simulations in water. The main aspects of the analysis are indicated in yellow (for all the systems) and in magenta (exclusive to proteic-crowding).

crowders, we included simulations of the same proteins in the presence of an artificial crowding agent (PEG500). As controls each protein was simulated in pure aqueous environment. The analysis was focused on changes, compared to what observed in water, in protein structure, flexibility and in the overall system dynamics (diffusion rates). For proteic crowders, we included five systems, each with a different protein concentration. In this way we could analyze concentration-dependent issues and enrich information on protein-protein interactions in dense environments.

In the proteic crowded system the selected proteins could play at the same time the role of crowding agents and the "subject" that experiences crowding. We specifically looked for proteins that could represent the variety of conformational landscapes (IDPs, IOPs and MGs) and form a biologically relevant network. It tuned out that our group was already familiar with a system that fitted in both categories [32]. The central player of the system is the protein NCBD (nuclear co-activator binding domain), a small all-helical 51-residue molten globule (MG). Despite possessing a structured core of ~40 residues, this protein is remarkably promiscuous, binding to seven different partners with multiple structures reported in the literature (PDB codes: 1KBH, 1ZOQ, 2KKJ, 2L14). Two of its partners are the small intrinsically disordered protein ACTR (IDP) and the large and structured protein IRF-3 (IOP). Thanks to it structural plasticity and depending on the ligand involved (ACTR or IRF), NCBD adopts diverse conformations upon binding, which are generated by different arrangements of well-defined helices [33].

Our aim was to place multiple copies of NCBD alongside with its two partners in the same simulation box, as illustrated in Figure 1 in the article (> 250000 atoms per system)[3].

We found that at structural level, crowding, both synthetic and proteic, favors open and moderately extended conformations with a higher content of secondary structure. Intriguingly, the malleable proteins (IDP and MG) in presence of proteic crowding gain in structures that could facil-

3 To efficiently simulate those huge systems we needed access to HPC resources, which we obtained during the 5th PRACE call (Grant #2013092029; 38,000,000 computational hours).

itate the binding. Regarding protein flexibility, the two crowders trigger opposite effects: PEG enhances the number of accessible conformations while proteic crowders limit them. The latter conformational restriction is not related to a reduction in the frequency of the conformational rearrangements events, which are in some cases even enhanced in presence of protein crowders. The dynamic of the entire system undergoes to a general reduction in presence of both crowders, with reduced diffusion rates for both water molecules and proteins. The volume exclusion effect depends on the concentration of the proteic crowders: the higher the concentration, the higher the degree of compactness of the protein. Proteins respond to crowding depending on their intrinsic disorder: the higher the disorder of a protein the higher the amount of aspecific intermolecular contacts that it forms. Intriguingly the changes observed in conformational entropy and protein flexibility follow the same trend, suggesting a major role for protein-protein contacts in driving the observed changes.

Bibliography Chapter 5

[1] K. Luby-Phelps, "The physical chemistry of cytoplasm and its influence on cell function: an update," Mol. Biol. Cell, vol. 24, no. 17, pp. 2593–2596, 2013.

[2] R. Harada, Y. Sugita, and M. Feig, "Protein Crowding Affects Hydration Structure and Dynamics," J. Am. Chem. Soc., vol. 134, no. 10, pp. 4842–4849, Mar. 2012.

[3] F.-X. Theillet, A. Binolfi, T. Frembgen-Kesner, K. Hingorani, M. Sarkar, C. Kyne, C. Li, P. B. Crowley, L. Gierasch, G. J. Pielak, A. H. Elcock, A. Gershenson, and P. Selenko, "Physicochemical Properties of Cells and Their Effects on Intrinsically Disordered Proteins (IDPs)," Chem. Rev., vol. 114, no. 13, pp. 6661–6714, Jul. 2014.

[4] H.-X. Zhou, G. Rivas, and A. P. Minton, "Macromolecular crowding and confinement: biochemical, biophysical, and potential physiological consequences.," Annu. Rev. Biophys., vol. 37, no. 1, pp. 375–397, 2008.

[5] H.-X. Zhou, "Protein folding and binding in confined spaces and in crowded solutions," J. Mol. Recognit. JMR, vol. 17, no. 5, pp. 368–375, Oct. 2004.

[6] A. P. Minton, "Models for excluded volume interaction between an unfolded protein and rigid macromolecular cosolutes: macromolecular crowding and protein stability revisited," Biophys. J., vol. 88, no. 2, pp. 971–985, Feb. 2005.

[7] A. Christiansen, Q. Wang, M. S. Cheung, and P. Wittung-Stafshede, "Effects of macromolecular crowding agents on protein folding in vitro and in silico," Biophys. Rev., vol. 5, no. 2, pp. 137–145, Feb. 2013.

[8] A. H. Elcock, "Models of Macromolecular Crowding Effects & the Need for Quantitative Comparisons with Experiment," Curr. Opin. Struct. Biol., vol. 20, no. 2, pp. 196–206, Apr. 2010.

[9] S. K. Mukherjee, S. Gautam, S. Biswas, J. Kundu, and P. K. Chowdhury, "Do Macromolecular Crowding Agents Exert Only an Excluded Volume Effect? A Protein Solvation Study," J. Phys. Chem. B, Oct. 2015.

[10] E. H. McConkey, "Molecular evolution, intracellular organization, and the quinary structure of proteins," Proc. Natl. Acad. Sci. U. S. A., vol. 79, no. 10, pp. 3236–3240, May 1982.

[11] D. I. Freedberg and P. Selenko, "Live Cell NMR," Annu. Rev. Biophys., vol. 43, no. 1, pp. 171–192, 2014.

[12] P. B. Crowley, E. Chow, and T. Papkovskaia, "Protein interactions in the Escherichia coli cytosol: an impediment to in-cell NMR spectroscopy," Chembiochem Eur. J. Chem. Biol., vol. 12, no. 7, pp. 1043–1048, May 2011.

[13] Q. Ma, J.-B. Fan, Z. Zhou, B.-R. Zhou, S.-R. Meng, J.-Y. Hu, J. Chen, and Y. Liang, "The Contrasting Effect of Macromolecular Crowding on Amyloid Fibril Formation," PLoS ONE, vol. 7, no. 4, p. e36288, Apr. 2012.

[14] I. M. Kuznetsova, K. K. Turoverov, and V. N. Uversky, "What macromolecular crowding can do to a protein," Int. J. Mol. Sci., vol. 15, no. 12, pp. 23090–23140, 2014.

[15] W. B. Monteith, R. D. Cohen, A. E. Smith, E. Guzman-Cisneros, and G. J. Pielak, "Quinary structure modulates protein stability in cells," Proc. Natl. Acad. Sci., vol. 112, no. 6, pp. 1739–1742, Feb. 2015.

[16] Y. Sasaki, D. Miyoshi, and N. Sugimoto, "Regulation of DNA nucleases by molecular crowding," Nucleic Acids Res., vol. 35, no. 12, pp. 4086–4093, 2007.

[17] L. A. Benton, A. E. Smith, G. B. Young, and G. J. Pielak, "Unexpected Effects of Macromolecular Crowding on Protein Stability," Biochemistry (Mosc.), vol. 51, no. 49, pp. 9773–9775, Dec. 2012.

[18] M. Sarkar, A. E. Smith, and G. J. Pielak, "Impact of reconstituted cytosol on protein stability," Proc. Natl. Acad. Sci. U. S. A., vol. 110, no. 48, pp. 19342–19347, Nov. 2013.

[19] W. B. Monteith and G. J. Pielak, "Residue level quantification of protein stability in living cells.," Proc. Natl. Acad. Sci. U. S. A., vol. 111, no. 31, pp. 11335–40, Aug. 2014.

[20] A.-C. Sotomayor-Pérez, O. Subrini, A. Hessel, D. Ladant, and A. Chenal, "Molecular Crowding Stabilizes Both the Intrinsically Disordered Calcium-Free State and the Folded Calcium-Bound State of a Repeat in Toxin (RTX) Protein.," J. Am. Chem. Soc., vol. 135, pp. 11929–34, 2013.

[21] C. A. Waudby, C. Camilloni, A. W. P. Fitzpatrick, L. D. Cabrita, C. M. Dobson, M. Vendruscolo, and J. Christodoulou, "In-cell NMR characterization of the secondary structure populations of a disordered conformation of α-synuclein within E. coli cells," PloS One, vol. 8, no. 8, p. e72286, 2013.

[22] L. A. Munishkina, E. M. Cooper, V. N. Uversky, and A. L. Fink, "The effect of macromolecular crowding on protein aggregation and amyloid fibril formation," J. Mol. Recognit. JMR, vol. 17, no. 5, pp. 456–464, Oct. 2004.

[23] D. Johansen, C. M. J. Jeffries, B. Hammouda, J. Trewhella, and D. P. Goldenberg, "Effects of macromolecular crowding on an intrinsically disordered protein characterized by small-angle neutron scattering with contrast matching.," Biophys. J., vol. 100, pp. 1120–1128, 2011.

[24] D.-L. Zhang, L.-J. Wu, J. Chen, and Y. Liang, "Effects of macromolecular crowding on the structural stability of human α-lactalbumin," Acta Biochim. Biophys. Sin., vol. 44, no. 8, pp. 703–711, Aug. 2012.

[25] M. S. Cheung, "Where soft matter meets living matter--protein structure, stability, and folding in the cell," Curr. Opin. Struct. Biol., vol. 23, no. 2, pp. 212–217, Apr. 2013.

[26] S. Tanizaki, J. Clifford, B. D. Connelly, and M. Feig, "Conformational Sampling of Peptides in Cellular Environments," Biophys. J., vol. 94, no. 3, pp. 747–759, Feb. 2008.

[27] S. R. McGuffee and A. H. Elcock, "Diffusion, Crowding & Protein Stability in a Dynamic Molecular Model of the Bacterial Cytoplasm," PLoS Comput Biol, vol. 6, no. 3, p. e1000694, Mar. 2010.

[28] D. S. Goodsell, The Machinery of Life. New York, NY: Springer New York, 2009.

[29] R. Harada, N. Tochio, T. Kigawa, Y. Sugita, and M. Feig, "Reduced native state stability in crowded cellular environment due to protein-protein interactions.," J. Am. Chem. Soc., vol. 135, pp. 3696–701, 2013.

[30] M. Feig and Y. Sugita, "Reaching new levels of realism in modeling biological macromolecules in cellular environments," J. Mol. Graph. Model., vol. 45, pp. 144–156, Sep. 2013.

[31] M. Feig and Y. Sugita, "Variable Interactions between Protein Crowders and Biomolecular Solutes are Important in Understanding Cellular Crowding," J. Phys. Chem. B, vol. 116, no. 1, pp. 599–605, Jan. 2012.

[32] A. N. Naganathan and M. Orozco, "The native ensemble and folding of a protein molten-globule: functional consequence of downhill folding.," J. Am. Chem. Soc., vol. 133, pp. 12154–12161, 2011.

[33] M. Kjaergaard, L. Andersen, L. D. Nielsen, and K. Teilum, "A folded excited state of ligand-free nuclear coactivator binding domain (NCBD) underlies plasticity in ligand recognition," Biochemistry (Mosc.), vol. 52, no. 10, pp. 1686–1693, Mar. 2013.

Is there a general effect of crowding on proteins?

Michela Candotti [1,2] and Modesto Orozco[1,2,3] (under preparation)*

[1] Institute for Research in Biomedicine (IRB Barcelona), Baldiri Reixac 10-12, 08028 Barcelona, Spain.

[2] Joint BSC-IRB Research Program in Computational Biology, Baldiri Reixac 10-12, 08028 Barcelona, Spain.

[3] Department of Biochemistry and Molecular Biology, University of Barcelona, 08028 Barcelona, Spain.

The habitat in which proteins exert their function is far from a dilute solution: it contains up to 300-400 g/L of several other macromolecules, primarily other proteins. The repercussions of this dense environment on protein behavior is often addressed employing synthetic crowding agents such as PEG. Such studies present a picture of crowding as an unspecific phenomenon that, by means of a volume exclusion effect, tends to favor folded states of any kind of protein. Here we use atomistic molecular dynamics simulations to analyze the effect of real (proteic) crowders in the structure and dynamics of three protein types: an intrinsically disordered (ACTR), a molten globule (NCBD) and a one-structure fold (IRF-3). We found that crowding doesn't stabilize a native compact structure and it prevents structural collapse. Physiological crowding generated by a dense protein environment leads to important changes in the structure and dynamics of proteins, often misrepresented by PEG questioning its utility as crowder model.

Keywords: quinary interaction, protein disorder, crowders

Introduction

Most in vitro and in silico experiments treat proteins as highly purified entities that act in isolation - neglecting that they perform their duties inside the cell. Their "habitat" – the cell cytoplasm - contains between 80 to 300 g/L of several other macromolecules, corresponding altogether to 5%-30% of volume occupancy [1]. Among the several effects that crowded environment can exert on protein behavior, the volume exclusion had originally believed to be the most relevant one [2-5]. According to the excluded volume paradigm, the presence of crowders limits the accessible space reducing the conformational entropy, favoring compact folded forms and restricting the prevalence of extended states [3].

Following this traditional view of crowding, most experimental studies on proteins in dense environments have been performed adding large polymers, such as poly(ethylene glycol) (PEG), Dextran or Ficoll. These polymers, often referred as "inert" crowders, should exclusively mimic the volume-exclusion effect [4], without adding any other more specific effects. However, in reality, experiments show a complex variety

of effects of "inert crowders" on protein stability, depending on the type and size of the crowder involved [3], [5], [6], warning on their effective "inert" nature [7]. Overall these findings arise many doubts on the ability of these polymers to represent the crowded cytosol, and on the capability of the volume exclusion model to explain the impact of crowding in protein structure and dynamics. Recent studies in cell-like environments have further challenged the volume exclusion model, showing that compacted conformations of proteins may not be always preferred in physiological crowded environments [9]–[15]. For example, NMR studies have shown that the native structure of a globular protein can be destabilized inside the cell [11], [12], in reconstituted cytosol [13] and in solution with proteic crowders [14] contrary to the situation found with large synthetic polymers [15].

Such results suggest that proteic crowders might have a dual nature: on one hand they display the classical volume-exclusion effect, and on the other they have form weak and transient (quinary) "soft" interactions with solute protein [9], [16], [17]. This generates a competition between destabilizing and stabilizing forces whose final result is difficult to predict [16], [18], [19]. To further complicate the scenario, we cannot ignore that crowding might affect not only the thermodynamics of folding, but also folding landscape, leading to the formation of alternative states not present in dilute solutions [20]. This can have dramatic impact on very dynamic proteins, such as intrinsically disordered (IDPs) or molten globules proteins (MGPs) [21]. Unfortunately, most crowding studies on these proteins employ synthetic polymers, often reporting only the expected increase in compactness of the structure [22]–[26]. Studies of IDPs or MGPs in cell-like crowder environments are more rare and have provided less clear conclusions [10], [18], [28–33].

Some of the problems in reaching a consensus theory on the nature of crowding from experimental data rely on the intrinsic limitations of experiments on highly dynamic systems, where single molecule information is lost within the experimentally detected structural ensemble [31]. Theoretical calculations, particularly molecular dynamics (MD), give direct access to atomic information on single-molecules in carefully controlled environments, and are then the perfect complement to experimentally ensemble-based techniques in the study of crowding effects [21], [35]–[38]. We take advantage here of the power of MD simulations to explore in detail the impact of synthetic (PEG) and physiological (proteins) crowders on the structure, dynamics and interactions of proteins showing: *i)* an intrinsically ordered protein (IOP): the 191-residues interferon regulatory transcription factor (IRF-3), *ii)* a molten-globule conformation (MGP): the 51-residues nuclear coactivator-binding domain of CREB (NCBD), and *iii)* an intrinsically disordered protein (IDP): the 47-residues activator for thyroid hormone and retinoid receptors (ACTR). These three proteins not only model the three major types of protein conformational landscapes, but also define a specific biological network, with NCBD as the central partner (the hub) able to transiently interact with IRF-3 and ACTR, thanks to its structural promiscuity [39]-[42]. Calculations present then the first systematic study of crowding on proteins showing different levels of structure and that define a biologically relevant crowded microenvironment.

Methods

Overview of the crowding models. We mixed NCBD, ACTR and IRF-3 to obtain five dense proteic solutions (175, 192,

239, 273 and 296 g of protein/ml; protein volume fraction 20-30%). A stoichiometry of 6:1:1 (NCBD, ACTR and IRF-3) was used to better reproduce the central protein of the system: NCBD, for which we considered 6 starting conformation (one per copy), three of them were taken from a NMR ensemble (PDB: 2KJJ), and corresponded to "folded" states (F1-3 in the remaining), while the other three were taken from snapshots collected from a 50 ns MD simulation at T=500K, corresponding a fully "unfolded" protein (U1-U3 in the remaining). Starting conformation for ACTR and IRF-3 were taken from PDB entries 1KBH and 1ZOQ respectively. The starting positions and orientations of the proteins in the simulation boxes were random (see below) to remove bias in the simulations. See *Figure 1* for a map of the simulations performed.

Control simulations. Control simulations at comparable timescale were performed in two environments: eight simulations (1 for ACTR, 6 for NCBD, and 1 for IRF-3) in pure water boxes; and eight additional simulations in water:PEG500 mixture (200 g/L concentration). In order to check for potential biases in the results originated from the finite size of the simulation box, and the use of a given set of relative orientation of the proteins we performed one additional simulation, but now considering an approximately an ~4 times larger box containing 24 NCBD, 4 IRF-3 and 4 ACTR proteins. This huge system (~850.000 atoms at 182 g/L of concentration) was simulated for 100ns and allowed us to have information of each protein copy in different protein surroundings.

To address the frustration of the contacts between NCBD and its partners in the crowded environment we extracted protein pairs formed by either a folded or an unfolded conformations of NCBD (F1 and U2 with ACTR, F3 and U3 with IRF-3) from the crowding simulation at 273 g/L and used them as starting seeds for multiple simulations in pure water and crowding conditions (273 g/L). For each of the four systems 10 simulations of 10 ns were performed (reaching a total of 400 ns in water and in proteic crowding respectively).

Simulation set-up. All starting structures were titrated, neutralized with monovalent ions, minimized, thermalized and pre-equilibrated using our standard procedure implemented in the MD-Web server [37]. In the case of PEG500 systems, proteins were immersed in a pre-equilibrated box of water/PEG molecules of 200g/L (starting PEG500 conformation from PDB- 4APO); the resulting systems were then pre-equilibrated by relaxing solvent for 10 ns prior to the general MD-Web equilibration procedure [37]. For the case of proteic crowding the starting positions and orientations of the different proteins were selected randomly by a Monte Carlo code that just avoids steric clashes between proteins. These systems were then hydrated to the desired concentration. The resulting systems were also pre-equilibrated for 10 ns prior to the general MD-Web equilibration procedure.

All the trajectories were collected with Gromacs 4.5 [38] using a time step of 2 fs in the isothermal (300 K) and isobaric (1atm) ensemble with Nose–Hoover thermostat and Berendsen barostat [39]–[41]. We applied periodic boundary conditions and particle Mesh Ewald corrections [42] for the representation of long-range electrostatic effects with a grid spacing of 1.0 nm and a cut-off of 1.0 nm for Lennard-Jones interactions. Constrains on chemical bonds were solved by SHAKE algorithm [43] with a relative tolerance of 0.0001. Parm99-SB-ILDN force field was employed for proteins [44], TIP3P for water molecules [45] and modified TraPPE-UA parameters from

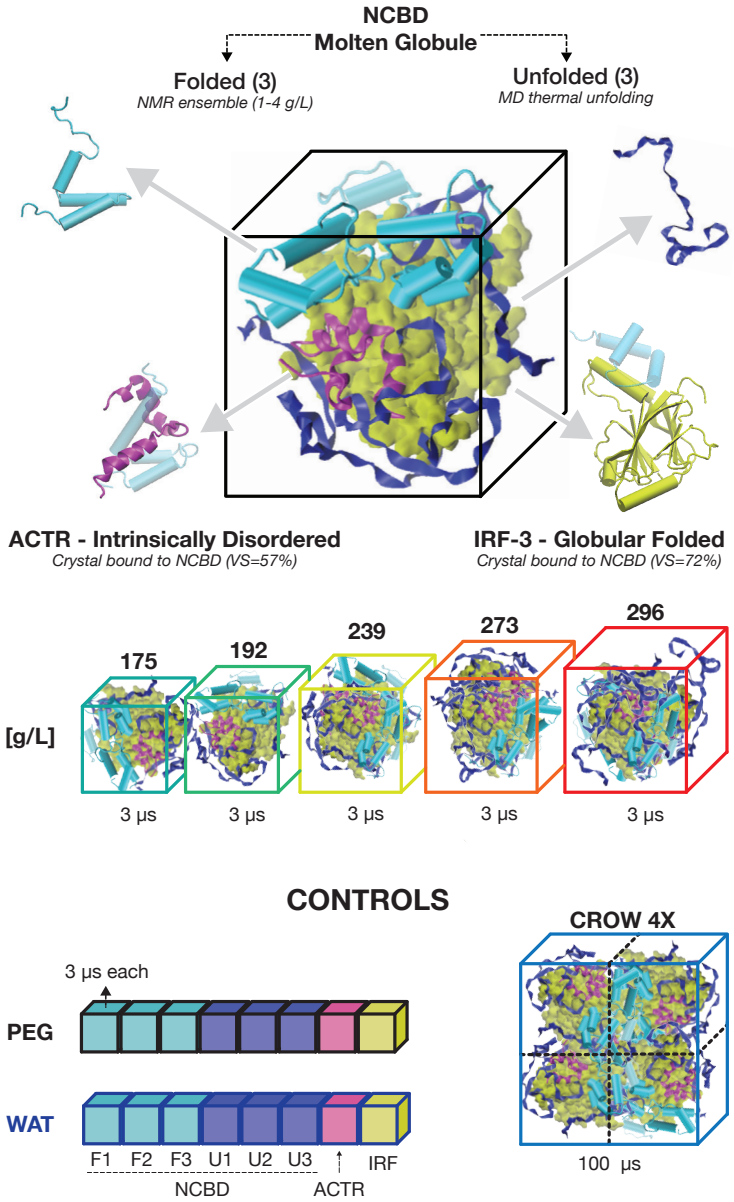


Figure 1. The simulated crowding system. From the top: example of one of the simulated boxes (192 g/L) composed by eight structures: three conformations of NCBD from the folded NMR ensemble (PDB 2KKJ); three unfolded conformations of NCBD from a simulation at 500K, one conformation of ACTR, bounded to NCBD (PDB: 1KBH) and one conformation of IRF-3, bounded to NCBD (PDB: 1ZOQ); the five concentrations used as proteic-crowders; and the control simulations. Below each box the simulated time for that system is indicated.

Fischer and colleagues for PEG molecules [46].

Analysis. Gromacs standard routines and in-house tools were used to mine the trajectories, with a minimum resolution of 20 picoseconds. We evaluate the overall protein compactness with the radius of gyration (Rgyr), the deviation from a reference structure with the root mean square deviation (RMSD), the exposed surface to the outside with the solvent accessible surface area (SASA) and the movements of each residues with the root means square fluctuations (RMSF). The secondary structure was evaluated by STRIDE [47]; VMD was used to visualize molecules and to analyze contacts [48]. Inter- and intra-protein contacts were defined by a cutoff of 0.8 nm between alpha Carbons. Intra-protein contacts were defined as "explored" if they were found in more than five frames. Conformations recurrently sampled were detected by using a two-steps clustering of backbone atoms using the GROMOS algorithm [49]. First we reduced the total number of conformations in each trajectory with a cutoff of 0.15 nm, and then, for each protein, the reduced ensembles in WAT, PEG and CROW were collected together and underwent to a second clustering with a cutoff of 0.35 nm. Following Knott-Best [50] the relative orientation of the helices of NCBD was used a coarse-grained descriptor of NCBD conformational space. The translational mean square displacements (MSD) of the center of mass of molecules were calculated to gain information on intermolecular movements (time windows of 10 and 25 ns were used for water and proteins respectively). Self-diffusion coefficients were determined using Einstein relationship as described elsewhere [51]. Conformational entropies were approximated at the quasi harmonic level using the last 1 μs of the simulations [52]. Finally, to detect reconfigurational events we clustered the all-atom trajectory employing the GROMOS algorithm [49] with a cutoff of 0.15 nm (0.1 nm for IRF-3) labeling as reconfigurational event any change in cluster.

Results and Discussion

Control simulations in water. Trajectories in water *(suppl. Figure S1* and *Figure 2*) show the expected behavior for the proteins under study. Thus, the folded protein (IOP: IRF3) is stable during the 4 μs of trajectory, maintaining the pattern of secondary structure, fold and shape. Native contacts are well preserved, with sizeable movements localized only at the C-tal helix, in a region with interface contacts in the crystal. A small, but detectable, tightening of the hydrophobic core of the protein is also present.

The Intrinsic disordered protein (IDP: ACTR) appears extremely mobile, sampling a wide repertoire of conformations: clustering analysis detected more than 250 different conformers, none of them populated more than 5.5% of the time, most of them compact. The contact map is very fuzzy, suggesting that no remote long-living contacts exist, which hinders the formation of stable folds. Some segments of ACTR tend to form secondary structure, especially evident in the α-helix at the N-terminal, in perfect agreement with NMR experiments ([53], [54]). However these helical elements are unstable and fuzzy, with local populations rarely above 50%, and undefined boundaries, therefore being unable to nucleate the global structure of the protein. Finally, the molten globule protein (MGP: NCBD) shows as expected a slow diffusion along the conformational space, with samplings showing strong memory effects ([35], [50], [55-56]) . When NCBD trajectory starts from the "folded" NMR structure significant plasticity is obtained (around

100 structural clusters), due to the different orientation of the three helical motives (h1, h2 and h3, see below and *Figure S3*), which generate a fuzzy contact map and confirming that helical arrangements of the ACTR-binding form prevail, while those required for IRF-3 recognition are rare ([50], [55]). When NCBD starts from the "unfolded" state, a fast collapses into an amorphous globule happens. The protein forms many remote and unstable contacts (282 structural clusters) and only small nascent elements of secondary structure (particularly in h1 and h2) are formed in the simulation time. All together these findings agree with previous claims on the slow dynamics of NCBD (>100 μs [55]), and illustrate the complexity of a folding landscape of a protein that was not evolutionary designed to collapse in a single well-defined minima. Overall, we can conclude that control simulations in water provide a reasonable picture of the conformational landscape of the three proteins considered here as models of well-ordered proteins, IDPs and MGPs in aqueous solutions. We can confidently use the same force-field and simulation protocol to explore crowded environments**.**

Crowding: synthetic vs proteic crowders. As described above, most theoretical and experimental studies on crowding have been performed using polymers (as PEG500) as co-solvents acting as "inert" crowders mimicking cellular crowding. However: *i)* are polymers such as PEG500 really "inert" crowders?, and *ii)* do they correctly mimic the proteic environment in the cell? In order to answer these two questions we compared trajectories in water, PEG500-crowding and proteic-crowding (using similar crowder concentrations in both cases) of the three model proteins considered here (*Figure 2*).

For IOP(IRF3) the effect of crowding is quite modest and neither proteins or PEG500 induce large changes in the local or global structure. Secondary structure is stabilized by crowding, including the C-tal helix that was fragile in water. Both type of crowders (specially the proteic ones) produce an increase in the protein surface, and an enlargement in the structure (RGYR/SASA), which is not consistent with the "exclude volume" theory. Only proteic crowders decrease the relative ratio of polar solvent accessible surface, suggesting that they attenuate the hydrophobic effect compared to water, where a collapse of the core is more visible (cartoons in *Figure 2* and *suppl. Figure S2*). Very interestingly, the crystal conformation of the protein is closer to those in a crowded environment (specially in the proteic media) than to those in dilute aqueous conditions, supporting the idea that crystals can in some cases mimic physiological conditions better than pure water (*suppl. Figure S4*).

For IDP(ACTR) crowding agents have a tremendous impact in the conformational landscape, but we cannot find a pattern of general "crowding" effects, since the changes induced by PEG in the conformational landscape of the protein are completely different to those produced by the proteic environment. Thus, PEG induces a dramatic enlargement of the sampled conformational space, which becomes dominated by extended conformers showing just a moderate amount of secondary structure. On the contrary, proteic crowders induce a reduction in the size of the sampled conformational space, which is now dominated by relatively compact structures. The conformational space is reduced and there is a dramatic increase in the level of secondary structure, which keeps the folding of the three helices required for NCBD binding [57], [58]. These results demonstrate the inability of PEG to reproduce cellular crowding around IDPs and strongly suggest that proteic (but not synthetic) crowd-

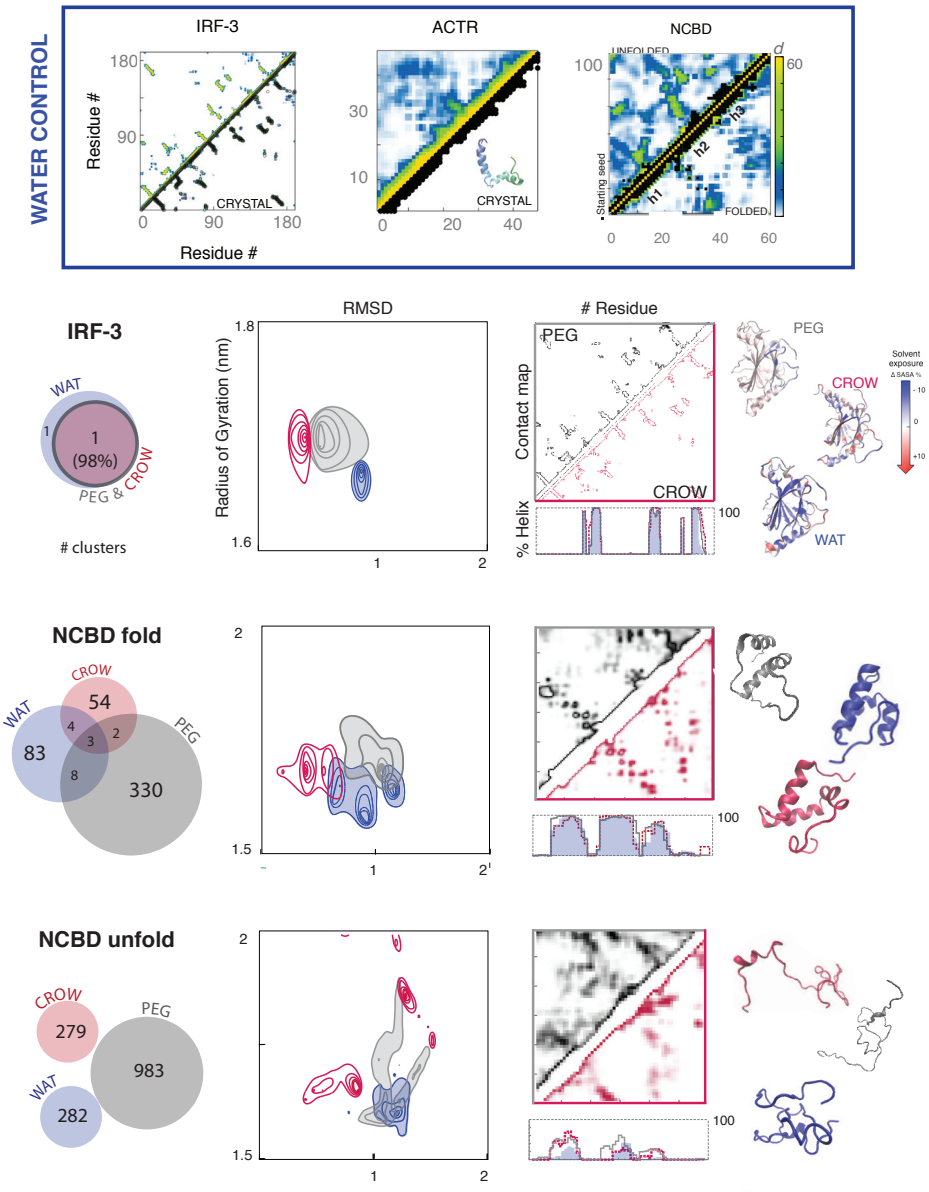


Figure 2. Structural changes in the three simulated environments. Top panel: contact maps in water, against the one of the PDB structure (in black). Bottom panel from left to right: conformational overlap between the clusters of each simulated environment; sampling maps based on the RMSD values from the starting conformation (x-axis) and the Radius of Gyration (y-axes); contact maps for the two crowding environments (see Fig S1 for water) and helical content along the sequence (calculated with STRIDE); and structures representing the most populated cluster in each environment. Color code: blue for water, red for crowding at the concentration of 192 g/L and grey for PEG500. For NCBD, the values from all the three conformations (three folded and three unfolded) are grouped together.

ing might help IDP to fold in the bioactive conformation.

For MGP(NCBD) the behavior of crowders largely depend on the starting conformation, mirroring the "memory effects" detected in dilute aqueous solution, and reinforcing the idea that NCBD (and probably other MGP) moves across a complex conformational landscape. In the trajectories collected starting from folded NCBD, crowders favor more extended conformations, with a fuzzy pattern of long-range contacts (*Figure 2*). The helical fragments are often arranged in bioactive conformations, often closer to the IRF-3-bound state, that is not sampled in water (*suppl. Figure S3*). For trajectories starting from the unfolded conformation of NCBD the effect of crowders is much larger. Both PEG and proteic crowders hinder the collapse observed in water and favor extended conformations. Interestingly, the native helices, which were hardly distinguishable in water, show significant populations and well-defined boundaries, especially for helix 1, an effect that as happens for an IDP can help in partner recognition. Compared to proteic-crowding, PEG again leads to a much larger flexible ensemble with a much more diffuse pattern of interactions, that differ from the one observed in proteic-crowding.

In summary, proteic crowders exert a complex effect in modulating protein conformation, which largely depends on the structural level of the native protein, not

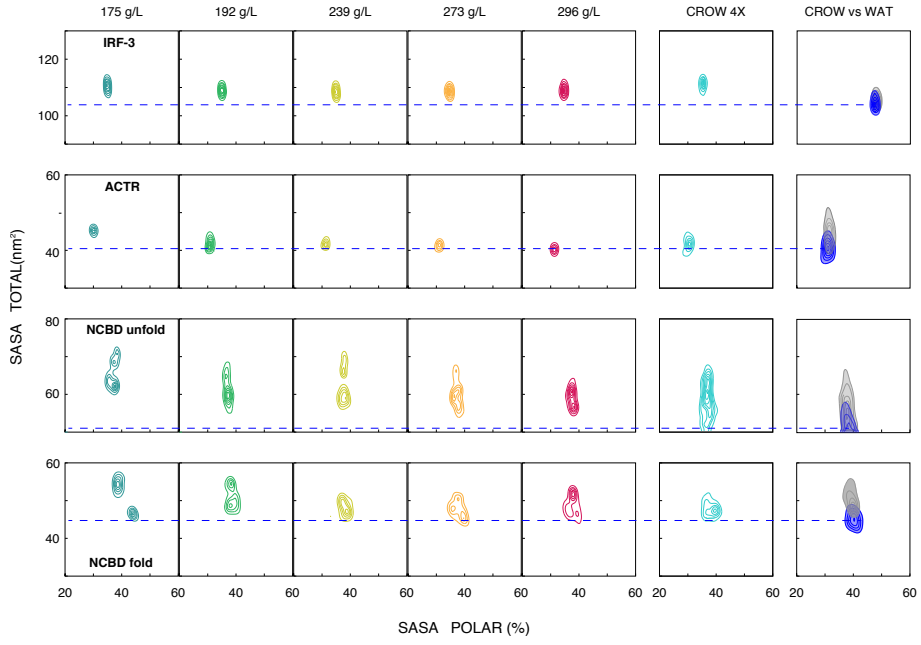


Figure 3. Changes in the solvent accessible surface area (SASA) of the protein in crowded systems. a) Sampling maps of the percentage of polar SASA (x-axis) and its total (y-axes) in nm2 calculated for the five concentrations of crowded systems and the other controls (CROW 4X = 100 ns at 182 g/L of a 4X larger system, PEG, water). For NCBD, the values from all the three conformations (three folded and three unfolded) are grouped together.

well reproduced by PEG.

Concentration effects in crowding. The analysis of 5 independent trajectories obtained at concentrations of protein from 175 to 296 g/l shows that the conformational landscape of proteins is quite robust to moderate changes on concentration of the proteic environment (see *Figure 3*, and *suppl. Figures S4-6*). However, detailed analysis shows some subtle, but systematic concentration-dependent changes in the crowding effect. For example, low concentration of proteic crowder favors extended conformations, while as the concentration increases more collapsed structures are preferred (*Figure 3*). This strongly suggest that proteic crowding is defined by the combination of two opposite effects: *i)* soft protein-protein interactions that favor the exposure of protein moieties and the prevalence of extended conformations, and *ii)* the "excluded volume" effect that favor collapsed structures. At low proteic concentration the first effect dominates, but, as the number of possible protein-protein contacts is satisfied, the "excluded volume" effect gains importance leading to more collapsed structures. The navigation of proteins above their energy landscape might be fine-tuned by playing with the proteic concentration.

Micro-enviornments in protein crowding. To evaluate the impact of the specific protein location in the box, we compared the 3 copies of folded and 3 copies of unfolded NCBD, which have different proteic neighbors and consequently different protein-protein contacts. If the specific protein surroundings played a major role, different behaviors would be expected for the 3 replicas. Instead their variability is consistent with that found in water or PEG (*Suppl. Table S1, suppl. Figure S4* and *S7*). To further confirm that specific interactions are not determinant in our systems, we simulate a larger variety of protein locations in a 4x larger box (4X CROW; 182 g/L proteic concentration). No remarkable differences are found between the sampling obtained here and the one in smaller simulation boxes (*Figure 3, suppl Figure S5* and *S8)*. The only remarkable exception is one of the copies of ACTR that has the N-tail exposed to a region of small proteic density (in magenta in *Figure 4*). There the lack of intra-protein contacts provoke an immediate response

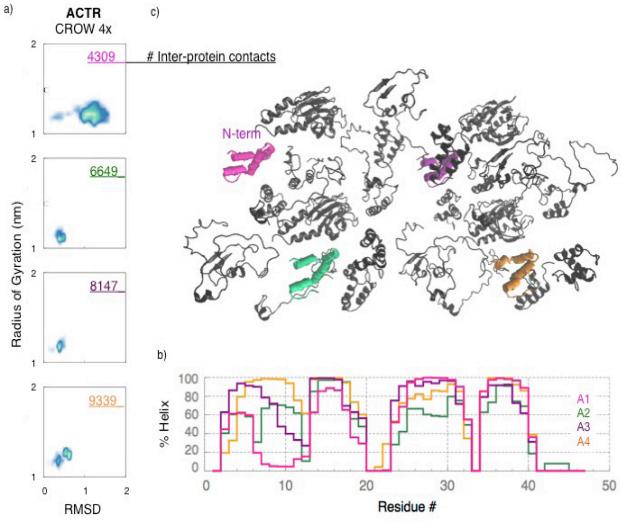


Figure 4. Structural descriptors for the four conformation of ACTR in the 4X box. a) Frequency maps of the RMSD values from the starting conformation (x-axis) and the Radius of Gyration (y-axes) in nm. The total number of inter-protein contacts is reported. b) Helical content along the sequence. c) Visualization of the 4X box with the four conformations of ACTR highlighted in colors, same color code as in the other graphs.

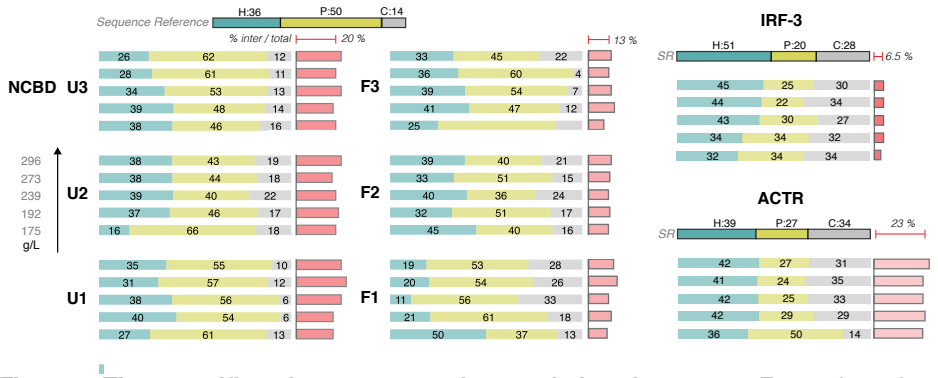


Figure 5. The aspecific quinary contacts in crowded environments. For each conformation, the distribution of the inter-protein contact according to the nature of the residues involved is report at increasing proteic crowding concentration (bottom - up). The darker boxes on top display the reference values in the protein sequence (H: hydrophobic in blue, P: polar in yellow and C: charged in gray). The percentage of the inter-protein contact among the total (inter and intra) is also reported in red; the average for each protein is reported at the top (see also Figure 6).

(within 100 ns) in ACTR, which undergoes to structural rearrangements (loss of helicity) never achievable when surrounded by proteins. The presence of aspecific contacts appears as a major determinant for the effect of proteic crowding.

Quinary contacts and crowding. Results above suggest that protein-protein interactions might be the responsible for the effect generated by a dense protein environment. A key issue is whether or not the explored inter-protein contacts correspond to unspecific transient contacts (quinary contacts) or specific interactions, which could not be assigned to a bona fide crowding effect.

We first analyzed the specific interactions that might occur in a biologically relevant cluster (IRF-3 and ACTR as partners of NCBD) but none of the contacts formed during the simulation recalls those of the bounded states *(suppl. Figure S9)*. This suggest that in the crowding environment such moieties might be busy interacting aspecifically with other proteins, preventing them from optimizing their binding interface and leading to a contact frustration. Indeed, the same complexes, when placed in water, rapidly adjust to form very specific pattern of contacts, that are not attainable under crowding conditions *(suppl. Figure S10)*. We are, then, reproducing a bona-fide "crowding effects", not contaminated by specific interactions that might occur in a biologically relevant cluster

Further analysis, based on the residues involved in protein-protein contacts, failed again to detect any prevalence of a specific type (*Figure 5)*. In this case, the protein crowder concentration leaves unaffected both the total number of protein-protein interactions and the type of the interacting residues *(Figure 5)*. We found only a clear trend: the higher the disorder of the protein, the larger the proportion of residues involved in protein-protein contacts (see *Figure 6*, left panel). This network of interaction could then trigger the changes in protein flexibility that show the same dependence on the degree of disorder (see next section). Overall we are, then, repro-

Diffusion Coefficient [μm² / s]

	Water		Proteins	
175 g/L	4900	(120)	11	(1)
192 g/L	4847	(199)	10	(3)
239 g/L	4291	(91)	7	(2)
273 g/L	4219	(162)	8	(1)
296 g/L	4041	(123)	11	(3)
WATER	5108	(226)	86	(26)
PEG	2432	(104)	33	(9)

Table 1. Diffusion coefficient of water molecules and the proteins in crowded environment, water and PEG. The table reports the average diffusion coefficient and its standard deviation calculated for all the water molecules and for all the protein present in the simulated box.

ducing a bona-fide "crowding effects", not contaminated by specific interactions that might occur in a biologically relevant cluster

The protein flexibility. We already observed that PEG and crowding environments alter, compared to water, the protein internal entropy in opposite ways: PEG enhances the number of conformations accessible to proteins; crowding instead limits the visited configurations (*Figure 2*, and *suppl. Table S1*) - both at global and local level as shown by the backbone configurational entropy and the amount of visited contacts, respectively. As observed for the interprotein contacts, the entropic change in proteic crowding (and not in PEG - *suppl Figure S11*) depends on the degree of disorder of each protein (*Figure 6*): the higher the expected disorder, the higher the reduction in conformational exploration. The protein dynamics at local level (RMSF) follows the same trend while the global dynamics is even enhanced (with the exception of the hyper-stabilized ACTR). In proteic crowding then conformational changes are still possible and even encouraged, but the accessible conformational space is limited and altered compared to both water and PEG. The crowding-induced entropic frustration is possibly balanced by the enthalpic contribution due to aspecific protein-protein contacts, the amount of which follow the same

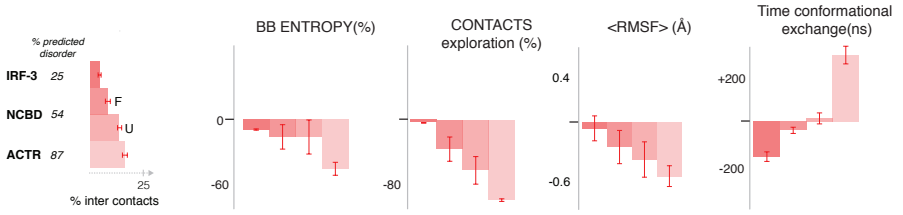


Figure 6. Effects of proteic crowding and protein disorder. From the left: for each protein the % of intrinsic disorder (calculated with PONDR-FIT); the percentage of inter protein contacts of the total (inter and intra) and the difference from the simulation in water in: the backbone conformational entropy; the % of explored intra-protein contacts; the average local root mean square fluctuation RMSF (Å) and the time between reconfigurational events (ns). Values are averages among all the crowding concentration.

trend *(Figure 6)*.

Diffusion. The additional interactions between macromolecules in crowding create a high viscosity that has repercussions on other aspect of the system mobility, i.e. diffusion rates and related binding kinetics [8], [59]. Considering translational micro-diffusion, i.e. involving small moieties, water molecules are slowed compared to dilute solutions *(Table 1)* [32]; however the slowest diffusion rate is found in PEG solution where the size of the PEG polymers, smaller than a protein, disturb the displacement of water molecules more effectively [60]. This behavior is not mirrored by macro-diffusion rates: protein displacement is mostly affected in crowding and within 10x as in [59], [61]–[63]. At the analyzed timescale (nanosecond), no remarkable differences exists between diffusion rates of folded or unfolded proteins. With due care [35],[66], our results provide a qualitative insight about the general reduction in mobility that affects all the proteins.

Conclusions

We detect several effects common to all the analyzed conformations and that provide an insight about the universal forces that proteins, at different extent, feel under crowding condition. Compared to dilute solution, both synthetic and proteic crowders favor open and moderately extended conformation with higher secondary structure content. However the contribution of the volume exclusion, that favors more compact structures, increases with the crowding concentration, confirming that macromolecular crowding is a battlefield between two opposite forces, the soft interactions and the hard-core repulsion, the balance of which depends on the concentration and type of the crowder [9].

The common outcomes between synthetic and proteic crowders are limited to this aspect; we detect, in fact, a divergent behavior in all the other observables. Proteins in PEG experience an increased conformational entropy, confirming recent observations from calorimetric analysis , which employ PEG molecules of similar dimension [7]. PEG also doesn't differentiate proteins depending on the degree of disorder. Special care then needs to be taken when studying flexible protein in presence of PEG.

The protein-crowded box appears as a stagnant system at the microsecond timescale, with slow diffusion rates and a general lower local dynamic. However at macro-level, changes involving the entire protein conformation can happen. Overall crowding favors conformations not explored in diluted solution and, in the case of molten globules, it can encourage the adaptability to multiple partners through structural rearrangements. The extent of these repercussions in fact depend on the protein type, with disordered proteins experiencing the most severe alterations. The unexpected stability and rigidity of an IDP suggests that this particular structure can exist in crowding condition independently from its partner, NCBD.

The work presented here is an attempt to rationalize the effect of homogeneously distributed proteic crowders – modeling an environment closer to the packed interior of a prokaryotic cell, with diffusion rates similar to E.coli [61]. However we should notice that eukaryotic cells are complex systems in which macromolecular crowding is just one of the players. Their cellular interior, far from a "bag full of molecules", might be filled with compartments where proteins could be stabilized or destabilized according to the specific surrounding [66], as observed in our case for protein portions exposed to diluted solutions.

Overall the picture that emerges from our

analysis clearly overcomes the classic view: crowding doesn't stabilize the native compact structure but instead prevents structural collapse. Despite the system stagnation, proteins upon proteic crowding retain a certain degree of malleability that could help to exert their function: the flexibility of the system is converged into a smaller ensemble of conformations (reduced conformational entropy) possibly leading to an efficient sampling among functional conformations and, as in the case of protein ACTR, can extending the lifetime of certain.

Both synthetic and proteic crowders behave like non-inert crowder; they both favor more open structures and with more helical content; however the protein structural details in the two environments diverge, and despite PEG mechanism of action remains outside the scope of this work, we join the concerns regarding its employment to mimic the cell interior.

Bibliography

1. S. B. Zimmerman and A. P. Minton, "Macromolecular crowding: biochemical, biophysical, and physiological consequences," Annu. Rev. Biophys. Biomol. Struct., vol. 22, pp. 27–65, 1993.

2. A. Christiansen, Q. Wang, M. S. Cheung, and P. Wittung-Stafshede, "Effects of macromolecular crowding agents on protein folding in vitro and in silico," Biophys. Rev., vol. 5, no. 2, pp. 137–145, 2013.

3. A. H. Elcock, "Models of macromolecular crowding effects and the need for quantitative comparisons with experiment.," Curr. Opin. Struct. Biol., vol. 20, no. 2, pp. 196–206, 2010.

4. L. R. Singh and S. Mittal, "Denatured State Structural Property Determines Protein Stabilization by Macromolecular Crowding: A Thermodynamic and Structural Approach," PLoS ONE, vol. 8, no. 11, p. e78936, Nov. 2013.

5. H.-X. Zhou, G. Rivas, and A. P. Minton, "Macromolecular crowding and confinement: biochemical, biophysical, and potential physiological consequences.," Annu. Rev. Biophys., vol. 37, no. 1, pp. 375–397, 2008.

6. J. Batra, K. Xu, and H.-X. Zhou, "Nonadditive effects of mixed crowding on protein stability," Proteins, vol. 77, no. 1, pp. 133–138, Oct. 2009.

7. M. Senske, L. Törk, B. Born, M. Havenith, C. Herrmann, and S. Ebbinghaus, "Protein Stabilization by Macromolecular Crowding through Enthalpy Rather Than Entropy," J. Am. Chem. Soc., vol. 136, no. 25, pp. 9036–9041, Jun. 2014.

8. I. M. Kuznetsova, B. Y. Zaslavsky, L. Breydo, K. K. Turoverov, and V. N. Uversky, "Beyond the excluded volume effects: mechanistic complexity of the crowded milieu.," Mol. Basel Switz., vol. 20, no. 1, pp. 1377–409, Jan. 2015.

9. A. E. Smith, Z. Zhang, G. J. Pielak, and C. Li, "NMR studies of protein folding and binding in cells and cell-like environments," Curr. Opin. Struct. Biol., vol. 30, pp. 7–16, Feb. 2015.

10. A. Politou and P. A. Temussi, "Revisiting a dogma: the effect of volume exclusion in molecular crowding," Curr. Opin. Struct. Biol., vol. 30, pp. 1–6, Feb. 2015.

11. K. Inomata, A. Ohno, H. Tochio, S. Isogai, T. Tenno, I. Nakase, T. Takeuchi, S. Futaki, Y. Ito, H. Hiroaki, and M. Shirakawa, "High-resolution multi-dimensional NMR spectroscopy of proteins in human cells.," Nature, vol. 458, no. 7234, pp. 106–109, 2009.

12. A. P. Schlesinger, Y. Wang, X. Tadeo, O. Millet, and G. J. Pielak, "Macromolecular crowding fails to fold a globular protein in cells.," J. Am. Chem. Soc., vol. 133, no. 21, pp. 8082–8085, 2011.

13. M. Sarkar, A. E. Smith, and G. J. Pielak, "Impact of reconstituted cytosol on protein stability.," Proc. Natl. Acad. Sci. U. S. A., vol. 110, no. 48, pp. 19342–19347, 2013.

14. A. C. Miklos, M. Sarkar, Y. Wang, and G. J. Pielak, "Protein crowding tunes protein stability.," J. Am. Chem. Soc., vol. 133, pp. 7116–7120, 2011.

15. D. Gnutt, M. Gao, O. Brylski, M. Heyden, and S. Ebbinghaus, "Excluded-Volume Effects in Living Cells," Angew. Chem. Int. Ed., vol. 54, no. 8, pp. 2548–2551, Feb. 2015.

16. W. B. Monteith, R. D. Cohen, A. E. Smith, E. Guzman-Cisneros, and G. J. Pielak, "Qui-

nary structure modulates protein stability in cells," Proc. Natl. Acad. Sci., vol. 112, no. 6, pp. 1739–1742, Feb. 2015.

17. Y. Wang, M. Sarkar, A. E. Smith, A. S. Krois, and G. J. Pielak, "Macromolecular crowding and protein stability.," J. Am. Chem. Soc., vol. 134, pp. 16614–8, 2012.

18. W. B. Monteith and G. J. Pielak, "Residue level quantification of protein stability in living cells.," Proc. Natl. Acad. Sci. U. S. A., vol. 111, no. 31, pp. 11335–40, Aug. 2014.

19. M. Sarkar, A. E. Smith, and G. J. Pielak, "Impact of reconstituted cytosol on protein stability," Proc. Natl. Acad. Sci. U. S. A., vol. 110, no. 48, pp. 19342–19347, Nov. 2013.

20. R. Harada, N. Tochio, T. Kigawa, Y. Sugita, and M. Feig, "Reduced native state stability in crowded cellular environment due to protein-protein interactions.," J. Am. Chem. Soc., vol. 135, pp. 3696–701, 2013.

21. V. N. Uversky, "A decade and a half of protein intrinsic disorder: biology still waits for physics.," Protein Sci. Publ. Protein Soc., vol. 22, pp. 693–724, 2013.

22. A. Soranno, I. Koenig, M. B. Borgia, H. Hofmann, F. Zosel, D. Nettels, and B. Schuler, "Single-molecule spectroscopy reveals polymer effects of disordered proteins in crowded environments," Proc. Natl. Acad. Sci. U. S. A., vol. 111, no. 13, pp. 4874–4879, Apr. 2014.

23. A. Roque, I. Ponte, and P. Suau, "Macromolecular crowding induces a molten globule state in the C-terminal domain of histone H1," Biophys. J., vol. 93, no. 6, pp. 2170–2177, Sep. 2007.

24. J. Hong and L. M. Gierasch, "Macromolecular crowding remodels the energy landscape of a protein by favoring a more compact unfolded state," J. Am. Chem. Soc., vol. 132, no. 30, pp. 10445–10452, Aug. 2010.

25. D. Johansen, C. M. J. Jeffries, B. Hammouda, J. Trewhella, and D. P. Goldenberg, "Effects of macromolecular crowding on an intrinsically disordered protein characterized by small-angle neutron scattering with contrast matching.," Biophys. J., vol. 100, pp. 1120–1128, 2011.

26. S. Qin and H.-X. Zhou, "Effects of Macromolecular Crowding on the Conformational Ensembles of Disordered Proteins," J. Phys. Chem. Lett., vol. 4, no. 20, Oct. 2013.

27. D. P. Goldenberg and B. Argyle, "Minimal effects of macromolecular crowding on an intrinsically disordered protein: a small-angle neutron scattering study," Biophys. J., vol. 106, no. 4, pp. 905–914, Feb. 2014.

28. C. S. Szasz, A. Alexa, K. Toth, M. Rakacs, J. Langowski, and P. Tompa, "Protein disorder prevails under crowded conditions," Biochemistry (Mosc.), vol. 50, no. 26, pp. 5834–5844, Jul. 2011.

29. A.-C. Sotomayor-Pérez, O. Subrini, A. Hessel, D. Ladant, and A. Chenal, "Molecular Crowding Stabilizes Both the Intrinsically Disordered Calcium-Free State and the Folded Calcium-Bound State of a Repeat in Toxin (RTX) Protein.," J. Am. Chem. Soc., vol. 135, pp. 11929–34, 2013.

30. C. A. Waudby, C. Camilloni, A. W. P. Fitzpatrick, L. D. Cabrita, C. M. Dobson, M. Vendruscolo, and J. Christodoulou, "In-cell NMR characterization of the secondary structure populations of a disordered conformation of α-synuclein within E. coli cells," PloS One, vol. 8, no. 8, p. e72286, 2013.

31. B. Schuler and H. Hofmann, "Single-molecule spectroscopy of protein folding dynamics—expanding scope and timescales," Curr. Opin. Struct. Biol., vol. 23, no. 1, pp. 36–47, 2013.

32. R. Harada, Y. Sugita, and M. Feig, "Protein Crowding Affects Hydration Structure and Dynamics," J. Am. Chem. Soc., vol. 134, no. 10, pp. 4842–4849, Mar. 2012.

33. A. V. Predeus, S. Gul, S. M. Gopal, and M. Feig, "Conformational Sampling of Peptides in the Presence of Protein Crowders from AA/CG-Multiscale Simulations," J. Phys. Chem. B, vol. 116, no. 29, pp. 8610–8620, Jul. 2012.

34. M. E. McCully, D. A. C. Beck, and V. Daggett, "Multimolecule test-tube simulations of protein unfolding and aggregation," Proc. Natl. Acad. Sci. U. S. A., vol. 109, no. 44, pp. 17851–17856, Oct. 2012.

35. A. N. Naganathan and M. Orozco, "The native ensemble and folding of a protein molten-globule: functional consequence of downhill folding.," J. Am. Chem. Soc., vol. 133, pp. 12154–12161, 2011.

36. M. Kjaergaard, K. Teilum, and F. M.

Poulsen, "Conformational selection in the molten globule state of the nuclear coactivator binding domain of CBP," Proc. Natl. Acad. Sci., vol. 107, no. 28, pp. 12535–12540, Jul. 2010.

37. A. Hospital, P. Andrio, C. Fenollosa, D. Cicin-Sain, M. Orozco, and J. L. Gelpí, "MDWeb and MDMoby: an integrated web-based platform for molecular dynamics simulations," Bioinformatics, vol. 28, no. 9, pp. 1278–1279, May 2012.

38. S. Pronk, S. Páll, R. Schulz, P. Larsson, P. Bjelkmar, R. Apostolov, M. R. Shirts, J. C. Smith, P. M. Kasson, D. van der Spoel, B. Hess, and E. Lindahl, "GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit.," Bioinforma. Oxf. Engl., vol. 29, pp. 845–54, 2013.

39. H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. DiNola, and J. R. Haak, "Molecular dynamics with coupling to an external bath," J. Chem. Phys., vol. 81, pp. 3684–3690, Oct. 1984.

40. S. Nosé, "Nosé, S.: A molecular-dynamics method for simulations in the canonical ensemble. Mol. Phys. 52, 255-268," Mol. Phys., vol. 52, no. 2, pp. 255–268, 1984.

41. W. G. Hoover, "Canonical dynamics: Equilibrium phase-space distributions," Phys. Rev. A, vol. 31, no. 3, pp. 1695–1697, Mar. 1985.

42. T. Darden, D. York, and L. Pedersen, "Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems," J. Chem. Phys., vol. 98, no. 12, pp. 10089–10092, Jun. 1993.

43. J.-P. Ryckaert, G. Ciccotti, and H. J. Berendsen, "Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes," J. Comput. Phys., vol. 23, no. 3, pp. 327–341, 1977.

44. K. Lindorff-Larsen, S. Piana, K. Palmo, P. Maragakis, J. L. Klepeis, R. O. Dror, and D. E. Shaw, "Improved side-chain torsion potentials for the Amber ff99SB protein force field," Proteins, vol. 78, no. 8, pp. 1950–1958, Jun. 2010.

45. W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein, "Comparison of simple potential functions for simulating liquid water," J. Chem. Phys., vol. 79, no. 2, pp. 926–935, 1983.

46. J. Fischer, D. Paschek, A. Geiger, and G. Sadowski, "Modeling of aqueous poly(oxyethylene) solutions: 1. Atomistic simulations," J. Phys. Chem. B, vol. 112, no. 8, pp. 2388–2398, Feb. 2008.

47. D. Frishman and P. Argos, "Knowledge-based protein secondary structure assignment," Proteins Struct. Funct. Bioinforma., vol. 23, no. 4, pp. 566–579, Dec. 1995.

48. W. Humphrey, A. Dalke, and K. Schulten, "VMD: visual molecular dynamics," J. Mol. Graph., vol. 14, no. 1, pp. 33–38, 27–28, Feb. 1996.

49. X. Daura, K. Gademann, B. Jaun, D. Seebach, W. F. van Gunsteren, and A. E. Mark, "Peptide Folding: When Simulation Meets Experiment," Angew. Chem. Int. Ed., vol. 38, no. 1–2, pp. 236–240, Jan. 1999.

50. M. Knott and R. B. Best, "A Preformed Binding Interface in the Unbound Ensemble of an Intrinsically Disordered Protein: Evidence from Molecular Simulations," PLoS Comput Biol, vol. 8, no. 7, p. e1002605, Jul. 2012.

51. M. P. Allen and D. J. Tildesley, Computer Simulations of Liquids. Oxford: Oxford Science Publications., 1987.

52. I. Andricioaei and M. Karplus, "On the calculation of entropy from covariance matrices of the atomic fluctuations," J. Chem. Phys., vol. 115, no. 14, pp. 6289–6292, Oct. 2001.

53. M. Kjaergaard, A.-B. Nø rholm, R. Hendus-Altenburger, S. F. Pedersen, F. M. Poulsen, and B. B. Kragelund, "Temperature-dependent structural changes in intrinsically disordered proteins: formation of alpha-helices or loss of polyproline II?," Protein Sci. Publ. Protein Soc., vol. 19, pp. 1555–1564, 2010.

54. V. Iešmantavičius, M. R. bing Jensen, V. Ozenne, M. Blackledge, F. M. Poulsen, and M. Kjaergaard, "Modulation of the Intrinsic Helix Propensity of an Intrinsically Disordered Protein Reveals Long-Range Helix–Helix Interactions," J. Am. Chem. Soc., vol. 135, no. 27, pp. 10155–10163, 2013.

55. M. Kjaergaard, L. Andersen, L. D. Nielsen, and K. Teilum, "A folded excited state of ligand-free nuclear coactivator binding domain (NCBD) underlies plasticity in ligand recognition," Biochemistry (Mosc.), vol. 52, no. 10, pp. 1686–1693, Mar. 2013.

56. M. Kjaergaard, F. M. Poulsen, and K. Teilum, "Is a Malleable Protein Necessarily Highly Dynamic? The Hydrophobic Core of the Nuclear Coactivator Binding Domain Is Well Ordered," Biophys. J., vol. 102, no. 7, pp. 1627–1635, Apr. 2012.

57. V. Iešmantavičius, J. Dogan, P. Jemth, K. Teilum, and M. Kjaergaard, "Helical Propensity in an Intrinsically Disordered Protein Accelerates Ligand Binding.," Angew. Chem. Int. Ed Engl., pp. 1–5, 2014.

58. J. Dogan, X. Mu, Å. Engström, and P. Jemth, "The transition state structure for coupled binding and folding of disordered protein domains," Sci. Rep., vol. 3, Jun. 2013.

59. J. A. Dix and A. S. Verkman, "Crowding effects on diffusion in solutions and cells," Annu. Rev. Biophys., vol. 37, pp. 247–263, 2008.

60. N. Kozer and G. Schreiber, "Effect of Crowding on Protein–Protein Association Rates: Fundamental Differences between Low and High Mass Crowding Agents," J. Mol. Biol., vol. 336, no. 3, pp. 763–774, Feb. 2004.

61. Y. Wang, C. Li, and G. J. Pielak, "Effects of proteins on protein diffusion," J. Am. Chem. Soc., vol. 132, no. 27, pp. 9392–9397, Jul. 2010.

62. S. R. McGuffee and A. H. Elcock, "Diffusion, Crowding & Protein Stability in a Dynamic Molecular Model of the Bacterial Cytoplasm," PLoS Comput Biol, vol. 6, no. 3, p. e1000694, Mar. 2010.

63. M. Feig and Y. Sugita, "Variable Interactions between Protein Crowders and Biomolecular Solutes are Important in Understanding Cellular Crowding," J. Phys. Chem. B, vol. 116, no. 1, pp. 599–605, Jan. 2012.

64. C. T. Andrews and A. H. Elcock, "Molecular Dynamics Simulations of Highly Crowded Amino Acid Solutions: Comparisons of Eight Different Force Field Combinations with Experiment and with Each Other," J. Chem. Theory Comput., vol. 9, no. 10, pp. 4585–4602, Oct. 2013.

65. D. Petrov and B. Zagrovic, "Are current atomistic force fields accurate enough to study proteins in crowded environments?," PLoS Comput. Biol., vol. 10, no. 5, p. e1003638, May 2014.

66. L. M. Gierasch and A. Gershenson, "Post-reductionist protein science, or putting Humpty Dumpty back together again," Nat. Chem. Biol., vol. 5, no. 11, pp. 774–777, Nov. 2009.

Supplementary Information

a)

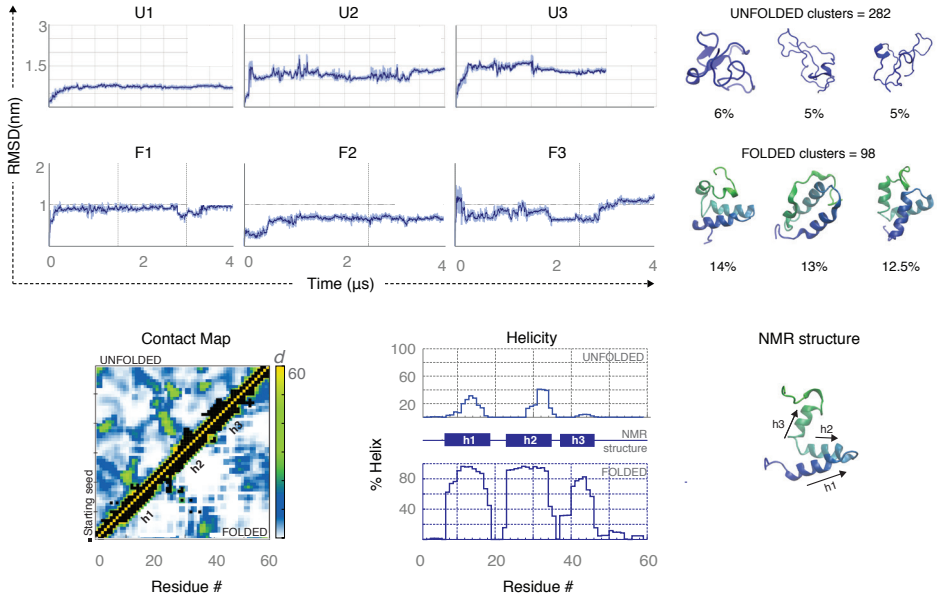
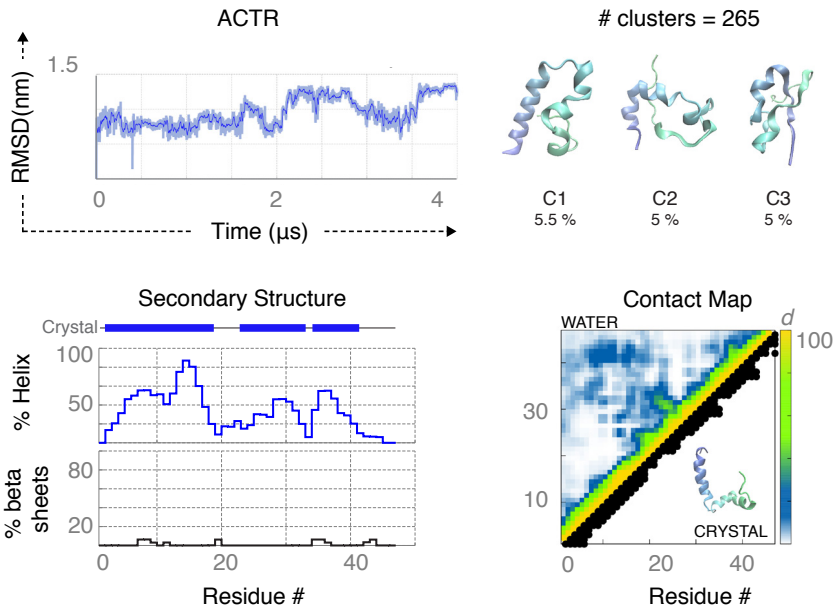
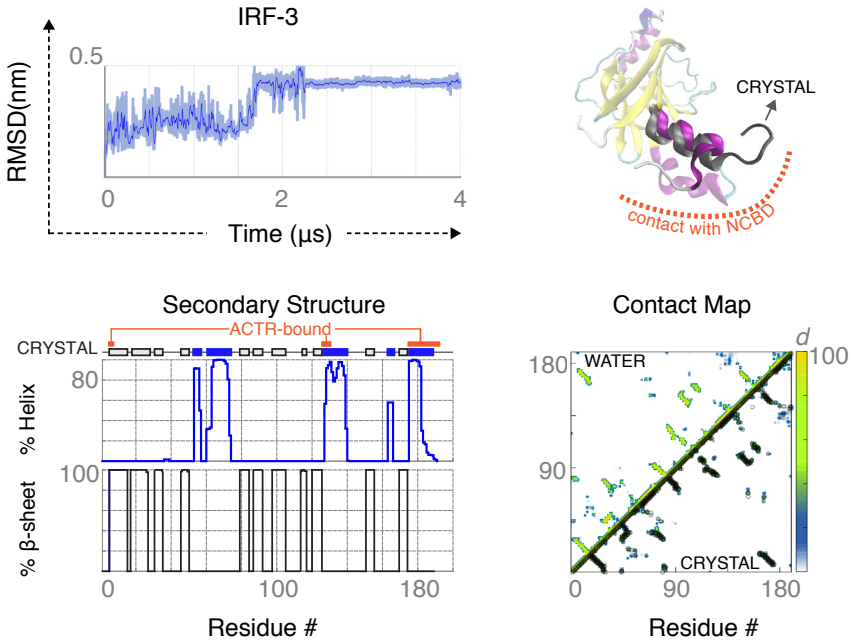


Figure S1. Control simulations in water. a) The RMSD evolution in time of each conformation of NCBD; the contact map and the helical content along the sequence (blue boxes are the helices in the crystal), calculated for the folded and unfolded conformations, grouped together. For each group the representative structures of the three most populated clusters are shown (the relative population is reported below) together with the reference structure from the PDB (NMR structure); **b)** same for ACTR; **c)** same for IRF-3 for which only the cartoon-like structure show the comparison between the crystal structure (in grey) and the second cluster found in water (population of 2%). Contact map of the PDB structure is shown in black while its secondary structure is shown with boxes (blue helices, black)

b)


ACTR
RMSD(nm)
1.5
0
2
4
Time (μs)
clusters = 265
C1
5.5 %
C2
5 %
C3
5 %
Secondary Structure
Crystal
% Helix
100
50
% beta sheets
80
20
0
20
40
Residue #
Contact Map
WATER
d
100
30
10
CRYSTAL
0
20
40
Residue #

c)


IRF-3
RMSD(nm)
0.5
0
2
4
Time (μs)
CRYSTAL
contact with NCBD
Secondary Structure
ACTR-bound
CRYSTAL
% Helix
80
100
% β-sheet
0
100
180
Residue #
Contact Map
d
100
180
WATER
90
CRYSTAL
0
90
180
Residue #

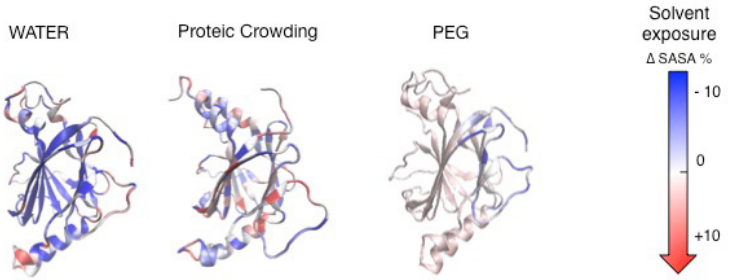


Figure S2. Changes of SASA in IRF-3, measured as the difference in SASA from the last to the fist frame for each protein residues. Residues with positive values, in red, open to the solvent during the simulated time, while residues with negative values, in blue, loose solvent exposure. Values are shown for water, PEG and crowding at 192 g/L.

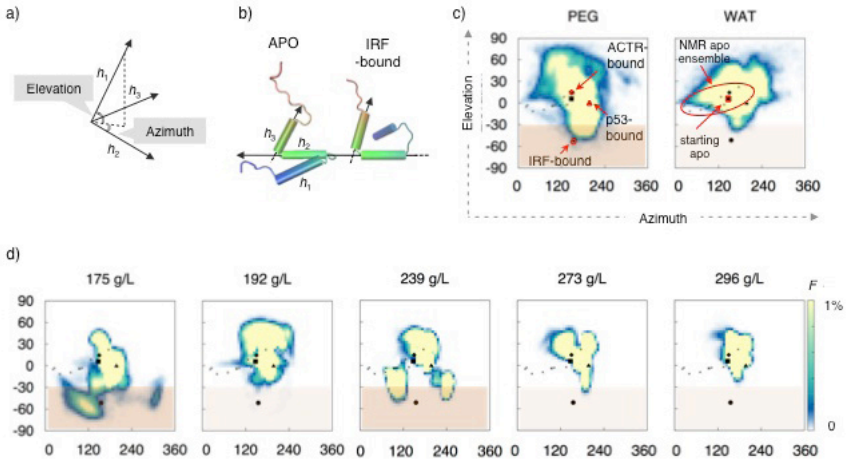


Figure S3. 3D rearrangements of NCBD helices. a) Cartoon explaining how elevation and azimuth are derived from the helix vectors h1-3 as seen in [Knott and Best, Plos Comp Biology, 2012]. b) Cartoon illustrating how helix vectors relate to the NCBD structure and its opposite positioning in two protein conformations. Each vector follows the principal axes of the atoms in the original helical each region, whether or not helices are formed in that moment. Panels in c) for control systems and d) for crowding system show the frequency of each 3D helical conformation defined by Azimuth (x-axes) and Elevation (y-axes). Results are collected for the three folded conformations together. The black signs mark values calculated from NCBD structure available in the PDB: ACTR-bound (PDB: 1KBH), p53 bound (2L14), IRF-bound (PDB: 1ZOQ), the NMR ensemble of unbound NCBD (2KJJ) and the structure used as starting point.

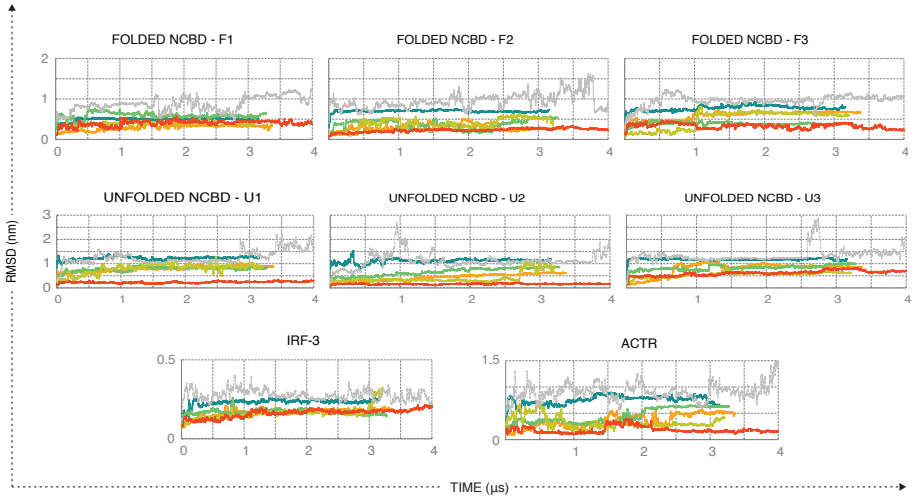


Figure S4. RMSD in crowded environments. RMSD values the starting conformation in crowded environment are displayed along time. Color code: gray for PEG500; blue to red from 175 g/L to 296 g/L, as in Figure 1.

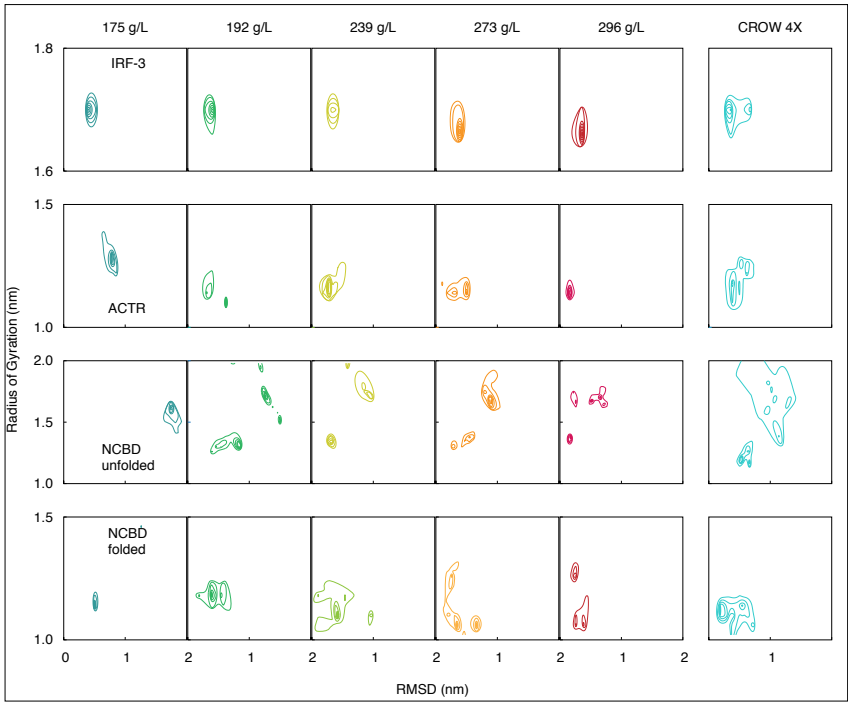


Figure S5. RMSD and Radius of Gyration sampling in crowding. The sampling map of the RMSD values from the starting conformation (x-axis) and the Radius of Gyration (y-axes) in nm calculated for the five concentrations of crowded systems and for CROW4x (182 g/L). For NCBD and for CROW4X results from the same protein are grouped together.

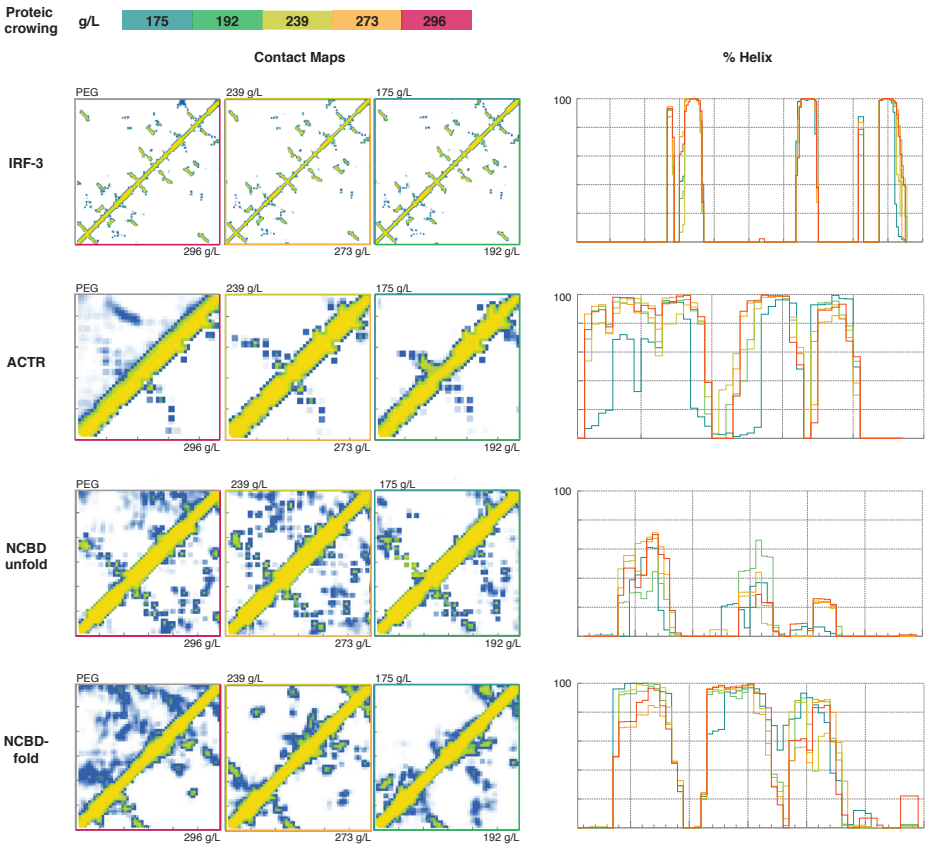


Figure S6. Structural descriptors with the crowding concentration. From the left: contact map for the crowded environments; and the helical content along the sequence calculated with STRIDE. In the case of NCBD are calculated for the three conformations together. Same color code applied as other figures.

	% Helical		BB conf. entropy (%)		Intra-Contacts exploration (%)		Time Conform. changes		Average RMSF	
	CROW	**PEG**	CROW	PEG	CROW	PEG	CROW	PEG	CROW	PEG
Folded NCBD	**2.64 (5.11)**	**1.33 3.01**	**-16.0 (11.1)**	**5.8 (16.8)**	**-32.6 (14.0)**	**-3.3 (15.6)**	**-37 (13)**	**-4 (10)**	**-0.24 (0.16)**	**0.13 (0.16)**
Conf F1	1.13 (0.66)	-2.79	-10.1 (13.2)	22.1	-20.5 (4.6)	14.7	-61 (27)	-10	-0.17 (0.15)	0.30 (0.16)
Conf F2	10.34 (3.05)	2.50	-26.3 (6.7)	6.7	-50.2 (2.6)	-1.3	-23 (12)	+1	-0.39 (0.15)	0.18 (0.16)
Conf F3	-3.40 (3.14)	4.29	-11.6 (4.1)	-11.5	-27.0 (8.5)	-23.3	-36 (12)	-3	-0.15 (0.16)	-0.10 (0.16)
Unfolded NCBD	**5.15 (3.25)**	**12.72 5.34**	**-16.1 (15.5)**	**12.6 (19.1)**	**-57.1 (15.9)**	**2.5 (27.8)**	**+13 (23)**	**-10 (11)**	**-0.36 (0.17)**	**0.25 (0.18)**
Conf U1	3.68 (3.03)	9.75	-1.4	34.5	-36.4 (4.0)	41.8	+30	-10	0.04 (0.15)	0.68 (0.15)
Conf U2	5.41 (3.51)	20.22	-26.0 (12.9)	0.9	-70.6 (6.8)	-19.0	-12	-16	-0.51 (0.19)	0.06 (0.20)
Conf U3	6.36 (2.54)	0.18	-9.3 (0.6)	2.3	-27.0 (8.5)	-23.3	+1	-5	-0.62 (0.17)	0.00 (0.19)

Table S1. Descriptors for the conformations of NCBD. For each conformations the average difference in several descriptors is reported for both proteic crowding (192 g/L) and PEG simulations. Values in bold are the average for each group while the standard deviation is reported in brackets.

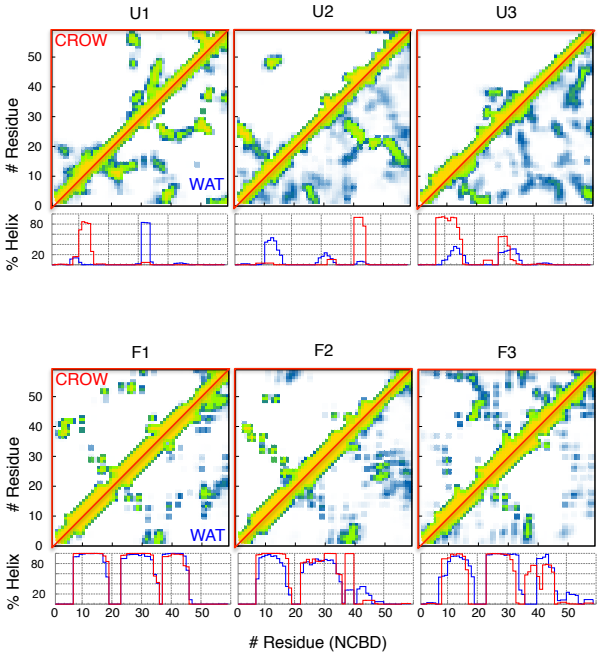


Figure S7. Structural details for the conformations of NCBD. For each conformations the contact map and the helicity along the sequence of the proteic crowder at 192 g/L, against the one in water (blue).

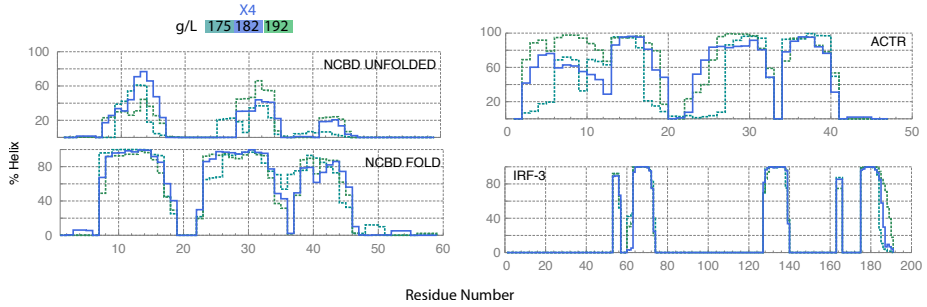


Figure S8. Helical content in the 4X control. For each protein we compared the helical content calculated in all the conformations in the 4X box(182 g/L) with the values taken from the crowding systems with comparable crowding concentration

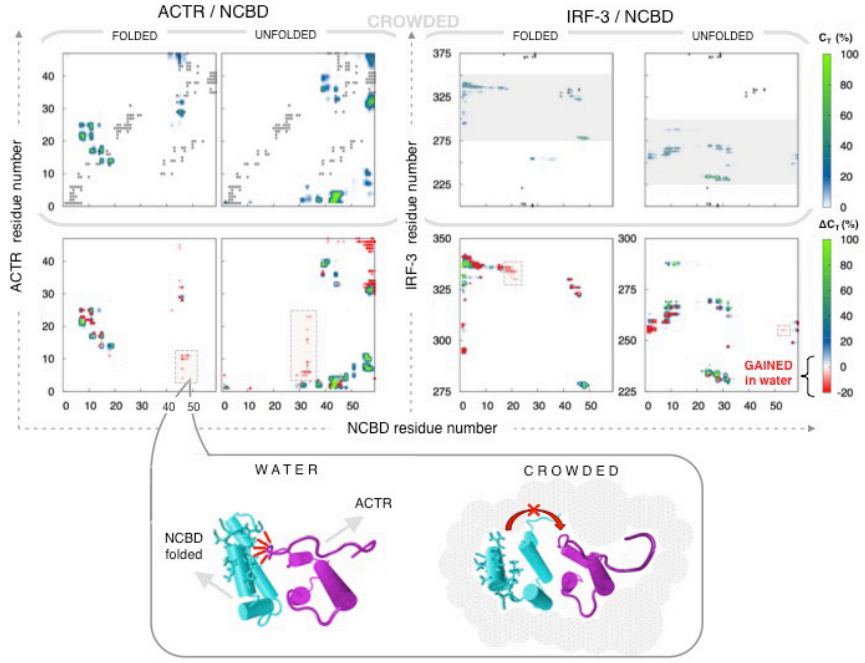


Figure S9. NCBD and its partners: complex formation and contact frustration. Contacts maps between NCBD residues (x-axes) and its partners (y-axes) ACTR on the left and IRF-3 on the right side. The plots in the first row display the contact time (% of the total simulated time) in the simulation at 273 g/L as an example of crowded system. The black dots mark contacts in the bounded structure available at the PDB. The second row displays the difference (Water - Crowded) in contact time calculated in the 10 copies of 10 ns in crowded conditions and in water. The contact map calculated from the 10 copies at crowded conditions is plotted in the background to identify contacts gained from scratch in water. The latter are marked with grey boxes. The cartoons at the bottom illustrate contacts newly formed in water (left side) between ACTR (in magenta) and NCBD (in cyan) while the crowded environment (right-side) prevented their formation.

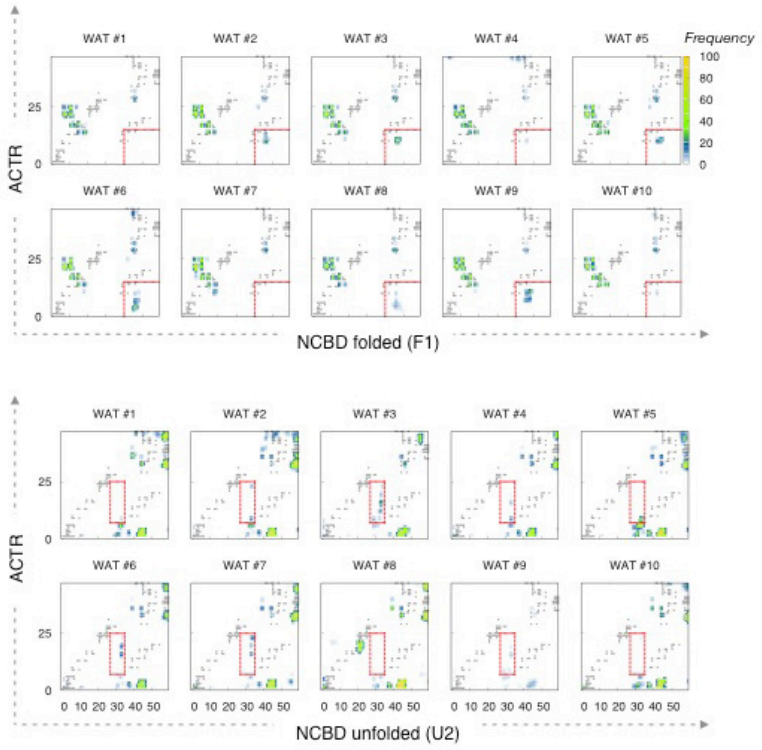


Fig S10. Contact maps of NCBD/ACTR complex in water. For each complex (ACTR with F1 or U2) the contacts maps for each of the ten copies of in water are shown. The red box highlights areas where often new contact (not present in crowding environment) are formed in water.

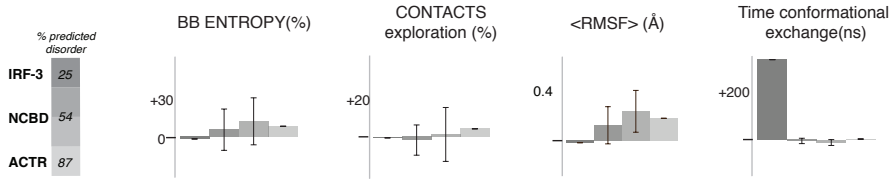


Figure S11. Effects of PEG depending on protein disorder. From the left: for each protein the % of intrinsic disorder (calculated with PONDR-FIT); the state (S) of the starting structure (F folded or U unfolded); and several differences using the simulation in water as reference: the backbone conformational entropy; the % of explored intra-protein contacts; the average local root mean square fluctuation RMSF (Å) and the time between reconfigurational events (ns).

"Real science is a revision in progress, always. It proceeds in fits and starts of ignorance."

Stuart Firestein, Ignorance: How It Drives Science

CHAPTER 6

Summary of the results and general discussion

6.1. Summary of the results

1) The urea-induced unfolded state of ubiquitin

We found that the simulations of the unfolded state of ubiquitin:

- In urea 8M reproduce fairly well the available experimental data, without sacrificing the plasticity expected from an unfolded state.
- In water the ensemble moves very quickly towards compact and "native-like" structures, starting a folding pathway.

The unfolded state of ubiquitin in presence of urea aqueous solution is

more ureaphilic than the native globular state, attracting a large proportion of urea molecules in the protein surroundings. The responsible for this attraction are mostly Van der Waals interactions, especially between apolar side-chains and urea. Although hydrogen bonds are present and contribute to the stabilization of the unfolded state, they don't represent the differential factor and the main driving force.

2) The early stages of protein unfolding in urea

In simulations of the early stages of unfolding for three ultra-representative proteins and in presence of either urea-aqueous solution or higher temperature, we found that:

- Proteins preserve a significant degree of secondary structure and a compact shape in both unfolding conditions.
- A similar degree of local unfolding in the three proteins is induced by both denaturing conditions, often even located in similar positions in the protein sequence.
- Only in the urea-induced unfolded structures the apolar sidechains are preferentially exposed to the solvent, and the atomic motions are slowed down prolonging the average time of local unfolding events.

In simulations of the urea-induced unfolding at proteome scale, we found that:

- Van der Waals interactions are responsible for a general attraction of urea around the protein
- Punctual interactions (hydrogen-bonds or Van der Waals) stabilize urea molecules in cavities within the hydrophobic core. These interactions are protein-specific and consistent among forcefields.
- Many examples suggest that the intrusion of urea is possible thanks to unfolding events that happen at access-points (hinge points) of the protein core.

3) Proteins in macromolecular crowding

- Compared to dilute solution, both synthetic and proteic crowders fa-

vor open and moderately extended conformation with higher secondary structure content. The contribution of the volume exclusion, that favors more compact structures, increases with the crowding concentration.

- Opposite to PEG, proteic crowders decrease the conformational entropy; this entropic frustration is balanced by the enthalpic contribution due to soft and non-specific inter-protein contacts with the surrounding macromolecule.
- Proteins respond to crowding depending on their intrinsic disorder: the higher the disorder of a protein the higher the amount of non-specific intermolecular contacts that it forms. Intriguingly the changes observed in conformational entropy and protein flexibility follow the same trend, suggesting a major role for protein-protein contacts in driving the observed changes.
- Crowders, especially the proteic ones seems to favor the population of bioactive conformation of disordered proteins (both intrinsic disordered and molten-globule).

6.2. General discussion

- **Building a consensus view.**

In this comparative study we created ad-hoc systems to derived general rules about the repercussions of two co-solvents (the urea-aqueous solution and a crowded environment) on three major features of proteins: their structure, dynamics and interactions with the surrounding solvent. Both phenomena are generally biased by the specifications of the system under study: type of protein involved, size of the crowder, parameters used, etc. While general trends had already emerged from previous studies, still special care needs to be taken when making comparison between results, which indeed even leaded to opposite conclusions [1-2]. In our case we also benefited from the joint efforts of previous studies and we tried to rationalize them in a consistent fashion; our approach consisted

in studying multiple versions of a system by varying, when possible, only one variable at time i.e. the protein class or fold, the forcefield, the type of crowder, the stage of the process etc. In this way we could discern the impact of such variables and extract results that are robust to their changes. For the urea-induced unfolding we studied independently the two end-points of folded/unfolded reaction, limiting our study to globular proteins that have a clear separation between folded and unfolded state (see the following section for a comprehensive discussion on the two projects). Macromolecular crowding instead is a phenomenon that could in principle affect all the proteins. In this case we included proteins with very different conformational landscapes and flexibility patterns in order to detect possible differential effects. Beside the general rules that could be derived from results that are consistent in all the systems, the spotted differences resulted equally informative in unveiling the mechanism beyond both urea-induced unfolding and macromolecular crowding.

- **The two sides of urea-induced unfolding: a comprehensive view**

When taken together, the two projects on urea-induced protein unfolding give a broader overview on the mechanism by which urea unfold the proteins. For example we detected a universal urea-philicity of the protein: at both the beginning and the end point of the unfolding process, urea molecules are attracted in the protein surroundings via dispersion forces. This enrichment is more noticeable for the unfolded state compared to its folded counterpart. And while in both cases the hydrogen bonds and other polar contact are not negligible, they never represent the driving force for the unfolding, as their magnitudes are comparable to those of water.

A difference aspect in the two stages of the unfolding is related to the protein areas that interact with the urea molecules. While in the fully-unfolded protein the apolar residues are the preferential partners, in the partially-unfolded state these moieties are still buried in the protein interior, preventing them to become the main partner. During the early stages of unfolding, instead, urea molecules are preferentially found in areas characterized by a specific topology: cavities of the protein core that become accessible thanks to local unfolding events or mobile loops in

the vicinities. The fundamental role of these events, part of the natural breathing pattern of each protein, suggest that urea doesn't attack areas on the protein surface depending of their nature, as it could have instead suggested the pattern found in the fully-unfolded protein. Instead two concomitant events initiate the unfolding: i) the general high presence of urea molecules around the protein and ii) the transient local mobility that gives a temporary access to the protein core. Once unfolded, the protein (now represented by an diverse ensemble of structures) becomes even more urea-philic compared to its corresponding folded state; indeed urea molecules readily surround and stabilize the newly exposed hydrophobic areas, preventing any attempt to recover the native state. The forces that drive unfolding in urea are, then, very different from the ones behind other chemical denaturants; for example guanidinium chloride (Gdm+ Cl-), induce denaturation by disrupting salt bridges that stabilize the folded conformation [3]. It should not be surprising then that some proteins have different responses and resistance to the two denaturing agents[4], confirming that the unfolded state of a protein, together with its structural features, cannot be separated from its unfolding conditions.

- **Proteins dynamics in non-aqueous environments**

Compared to pure water solutions, both urea 8M and crowded solutions have a higher viscosity, which slows down the overall dynamic of the system. Small co-solvents, such as urea and PEG, usually affect micro diffusion more than larger ones, like proteic crowders [5]. Indeed we found self-diffusion rates of water strongly reduced in PEG or urea 8M (0.5X PEG or 0.3X Urea), and only marginally in proteic crowders (0.8 X).

However, while in the case of PEG it doesn't affect the local dynamic of the protein, for urea and proteic crowders the viscosity is translated into a reduction of its fast local movements (i.e. rmsf of the local side chains). In presence of urea, the slower local mobility has important repercussion in the protein-solvent interactions: the longer opening time enhances the chances for urea molecules to intrude in the protein core and guide towards the unfolding of the protein. In the case of proteic crowders the reduction in the protein mobility depend on the amount of non-specific

contacts that a protein can form with the surrounding: the higher the number of interactions, the higher the reduction in local flexibility.

Interestingly in all the non-aqueous environments the global protein plasticity is preserved, if not enhanced compared to water: in the early stages of the unfolding process, proteins are accumulating changes that increase their degree of unfolding; the unfolded conformations interchange with other members of the unfolded ensemble; and in the crowded environment a degree of conformational exchanges is observed too. Such plasticity is fundamental to drive the sampling towards new structures that are not usually observed in water. This is not valid for our IDP in proteic crowding: the extraordinary plasticity observed in water is mitigated when placed in the new solvent. This single-case observation prevents us from deriving a general rule valid for IDPs but it encourages further studies on the reaction of several IDPs and MGs to cell-like environments.

- **Proteins outside water: common behaviors**

Proteins in either urea 8M or macromolecular crowding solutions present a more exposed surface to the new solvent and, when present, a destabilized hydrophobic core. Both phenomena come along with a larger exposure of the apolar surface of the protein, and new protein conformation are stabilized by the concerted presence of universal soft/dispersive interaction. Overall then, both denaturant and crowders appear more accommodating for the entire protein structure, which include a heterogeneous degree of polarity. Clearly, the ability of both types of solvent to remove water molecules from the vicinities of the protein surface explains this common effect, which is obviously more pronounced in the case of urea.

- **Sequence-dependent behaviors**

The reaction of each protein to the cosolvent (urea or crowder) is very specific; however from the background noise it is possible to discern patterns that might respond on specific protein features. For example, while all the proteins appear to be equally urea-philic regardless of their ami-

no-acid composition, each protein has specific weak points where the unfolding nucleates and that remain consistent among the several simulating conditions. As we already pointed out, the unfolding process exploits the protein local flexibility that gives temporary access to the protein core. The intrinsic flexibility pattern is sequence-dependent and explains the presence of protein-specific nucleation points, suggesting a possible mechanism by which some proteins resist to urea-denaturation. Since the urea-philicity is a universal feature for all the proteins, it is lack of these weak areas (aka unfolding nucleation points) that instead might prevent the urea-denaturation. For example proteins that have a native state with a tight, rigid and non-accessible protein core, despite being surrounded by urea molecules, don't create the right opportunity for solvent molecules to intrude inside the protein core. Such behavior might resemble the one of thermophilic proteins [4].

In proteic crowding, the differential response depends on a similar feature: the lack of a stable hydrophobic core in water, which is typical of intrinsically disordered or molten globule proteins. In water this feature is translated into a high degree of disorder and a restless structure, often collapsed to minimize the few hydrophobic residues from the exposure to water. Proteic crowders instead mitigate the disorder and the mobility and partially-structured conformations can have a longer lifetime.

- **Complexity and simplification of cell-like environments.**

Strictly speaking, the physiological environment of a protein is the one that occurs in nature, which can only be approximated in laboratory conditions and (probably) never be exactly reproduced. It follows that both experiments and simulations are models that only approximate reality based on a certain degree of simplification. Diluted solutions, despite being an over-simplification of the cell-like environment, have been extremely useful (and will continue to) in macromolecular biochemistry. However the study of cell-like environments will bring more awareness about the consequences of such a simplification. Most biophysical studies aiming to approach physiological-like crowding condition use idealized crowding agents such as PEG, or even the more inert Ficoll and Dextran.

We found that in some cases PEG and proteic crowders have similar, but in others (specially for disordered proteins) have quite different effects. In summary, our results suggest that while they remain interesting non-aqueous environments to study, they possibly do not represent the complexity of the cellular environment

- **Strength and limitations of simulations in biology**

To accomplish our aim and derive general rules, here we employed MD simulation on strategically selected systems. A major strength of in silico simulations is the full control of the system set-up, which allows the tuning of specific variables with a precision rather difficult to achieve in experimental settings. On the other hand, a major downside of simulations is their dependence on force-fields parameters. Special care then needs to be taken in selecting the adequate parameters, especially in absence of a direct comparison with experimental results. Comparative studies with multiple force-fields help to overcome such bias and lead to more robust results.

Nowadays we often exploit computational simulations to reproduce and predict the reality, and only seldom to perform absurd experiments, which would be impossible with any other method. However pushing a system to one extreme can reveal a lot about its nature: for example in the case of urea, Stumpe and Grübmuller performed a "Gedankenexperiment", in which urea polarity was scaled to create a hyper- and hypo polar urea molecule [6]. These cutting-edge works, in order to be realistic - although impossible, need a solid background and a detailed knowledge of the system under study. The large amount of available results on the urea aqueous solution, from both theoretical and experimental side, gave the right confidence to play with the system. On the other hand, crowded systems have become accessible to theoretical and experimental studies only recently; it is not surprising then that they miss a more solid realism. We are now at the right stage to use all the available tools to dig into the crowding issue: these concerted efforts will define better the challenges and guide the development of even more suitable methods to tackle the issue.

In summary, taking all the lessons from the simplified models used so far, we could soon move towards the study of proteins in their biological habitats. Although it seems science fiction to correctly simulate or precisely observe the behavior of proteins inside the cell, I believe that the joint efforts of experiments and simulations will soon meet at half way. In that sense I join the "call to arms" proposed by Elcock, who suggested to "stop comparing experimental apples with simulated oranges (or bananas)" and instead incite to directly compare numbers from theory and experiment, as precisely as possible [7]. The challenge is to shorten the gap between experiments and theory, and to combine both to approach the long-term dream of understanding living organisms from the basic rules of physics and chemistry.

Bibliography Chapter 6

[1] H.-X. Zhou, G. Rivas, and A. P. Minton, "Macromolecular crowding and confinement: biochemical, biophysical, and potential physiological consequences.," Annu. Rev. Biophys., vol. 37, no. 1, pp. 375–397, 2008.

[2] D. R. Canchi and A. E. García, "Cosolvent effects on protein stability," Annu. Rev. Phys. Chem., vol. 64, pp. 273–293, 2013.

[3] H. Meuzelaar, M. R. Panman, and S. Woutersen, "Guanidinium-Induced Denaturation by Breaking of Salt Bridges," Angew. Chem. Int. Ed Engl., Oct. 2015.

[4] P. Del Vecchio, G. Graziano, V. Granata, G. Barone, L. Mandrich, M. Rossi, and G. Manco, "Denaturing action of urea and guanidine hydrochloride towards two thermophilic esterases.," Biochem. J., vol. 367, no. Pt 3, pp. 857–863, Nov. 2002.

[5] J. A. Dix and A. S. Verkman, "Crowding effects on diffusion in solutions and cells," Annu. Rev. Biophys., vol. 37, pp. 247–263, 2008.

[6] M. C. Stumpe and H. Grubmüller, "Polar or Apolar—The Role of Polarity for Urea-Induced Protein Denaturation," PLoS Comput Biol, vol. 4, no. 11, p. e1000221, Nov. 2008.

[7] A. H. Elcock, "Models of macromolecular crowding effects and the need for quantitative comparisons with experiment.," Curr. Opin. Struct. Biol., vol. 20, no. 2, pp. 196–206, 2010.

CONCLUSIONS

In the urea-induced unfolded state of ubiquitin:

1. Simulations were able to reproduce the behavior of the unfolded ubiquitin in 8M urea solution.

2. Overall dispersion, rather than electrostatic interactions, is the main energetic contribution to explain the stabilization of the unfolded state of the protein and the irreversibility of the unfolding process in the presence of urea.

In the early stages of the urea induced unfolding:

1. The partially unfolded states expose to the solvent the apolar residues buried in the protein interior, mainly via cavitation.

2. Similar to the unfolded state, it is the dispersion interactions that drive urea accumulation in the solvation shell. H-bonds instead are crucial to stabilize long-living interactions strategically placed at hinge points;

3. Urea molecules take advantage of microscopic unfolding events to penetrate the protein interior, suggest a more sophisticated role for urea, far from the passive stabilizer of the thermal unfolding.

Regarding the impact of macromolecular crowding:

1. The universal effect of crowding is exerted via the soft interactions and favors open and moderately extended conformation with higher secondary structure. This phenomenon counterbalances the volume-exclusion, which prevails at higher crowding concentrations;

2. The impact of proteic crowding is proportional to the degree of disorder of the protein;

3. The artificial crowder PEG fails to reproduce correctly the effects of proteic crowders, arising concerns about its general use as a surrogate of cell-like environments.