UNIVERSITAT POMPEU FABRA
*Departament de Ciències Experimentals i de la Salut*

TESI DOCTORAL UPF 2012

# Mechanisms of evolutionary innovation in mammalian genes

**Macarena Toll i Riera**

Department of Experimental and Health Science

Evolutionary Genomics Group, Research Programme on Biomedical Informatics (GRIB), IMIM-UPF

Dra. Maria del Mar Albà
*Director*
Institut Municipal d'Investigació Mèdica
(IMIM), Universtitat Pompeu Fabra
Institució Catalana de Recerca i Estudis
Avançats (ICREA)

Dr. Eduardo Eyras
*Co-Director*
Universitat Pompeu Fabra
Institució Catalana de Recerca i Estudis
Avançats (ICREA)

Barcelona, 2012

*A tots els que han fet possible que hagi arribat fins aquí,*
*però en especial, als meus pares.*

The investigation of the truth is in one way hard, in another easy. An indication of this is found in the fact that no one is able to attain the truth adequately, while, on the other hand, no one fails entirely, but every one says something true about the nature of things, and while individually they contribute little or nothing to the truth, by the union of all a considerable amount is amassed

Aristotle
*Metaphysics, 993a, 30-993b, 9.*

# Acknowledgments

Quan comences la tesi et sembla que et queda un llarg camí per recórrer i sense adonar-te et trobes escrivint els agraïments de la tesi que és la part que tant has esperat perquè precisament l'has guardada per l'últim moment. Varies persones m'han acompanyat en aquest camí ja sigui centíficament o no, i a continuació les mencionaré intentant no deixar-me a ningú, però per si de cas, gràcies a tothom!

Mar, moltes gràcies per haver confiat en mi, pel teu suport i per guiar-me en aquest procés d'aprenentatge tant científic com personal. Gràcies per ajudar-me a adquirir el raonament crític i el coneixement necessari per fer ciència, pels consells, per la teva paciència (sobretot en aquest últim mes), per les correccions i per motivar-me a involucrar-me en molts projectes. I pels partits de tennis i de tennis de taula.

Josh, thanks for providing me with the great opportunity to make a short research stay in your group. It was very fruitful, we end up with an article, I learned a lot, I improved my English and I had the opportunity to work in a different environment. It was fun to be in Philly! And of course, many thanks for your help and advices when looking for a post-doc position. Glad to know you Josh.

Robert, a tu et dec molt. Gràcies a les teves classes, i a les d'en Roderic, vaig decidir fer recerca i un doctorat en el camp de la bioinformàtica. Vas fer que una assignatura que a priori pensava que odiaria profundament fos una de les que he disfrutat més durant la carrera. També t'he d'agraïr tant les col·laboracions científiques com la teva ajuda amb les cartes de referència.

Eduardo, muchas gracias por ser mi co-director de tesis y facilitarme los temas burocráticos.

Gràcies als companys de despatx que he tingut durant aquests anys. Agraïr sobretot a aquells que em van ajudar al principi de tot quan vaig arribar i no sabia fer gairebé res: en Loris, en Domènec, en Nico, l'André i la Mireya. A Nico, por ser el único que ha estado des del principio al final, compartiendo trabajo, risas, cafés, asados, Bariloche y Buenos Aires. A Alice, gracias por aportar frescura, cotilleos, charlas sobre política, risas, el despacho no es el mismo desde que te fuiste, ah! y por el calamaro!. Steve for the scientific collaborations, for the coffe-chats, your help with English and also for showing us how is a catalan-scottish wedding. Núria, per l'article de low-complexity i gens nous i per les ciutats visitades aprofitant els congressos. Mireya y Eneritz por todos estos años compartiendo despacho y risas. Núria, Eneritz i Mireya per l'aventura per terres portugueses que ens va portar a passejar-nos disfressades de Pare Noël pels carrers d'Oporto i a fer un tour gastronòmic d'allò més curiós. Núria, Inma i Laia per compartir l'hora de dinar. A Nico, André, Mireya, Steve, Alice, Núria, Inma, Eneritz i Amadís per les cerveses. A l'Alfons i a en Miguel per arreglar tots els problemilles tècnics pacientment. To André for the tennis matches (if we can say match to what we do), for always disapproving my sports preferences, for your Açores advices and of course, for the fantastic arroz caldoso (after so many years waiting for it!). A los chicos del grupo de Toni, en especial a Jaime y Salva, por hacer de los retreats algo divertido, table tennis rules! A en Jaume per ajudar-me amb el PDB i les estructures. Of course I do not want to forget the labmates I had at the University of Philadelphia, for the geeky lunches and the intense board games sessions with too much food and drinks. I would like to specially thank to David (Papi) for his patience and for helping me with the project, for sharing beers and conversations, it was great to work with you! To Etienne for his help when I was looking for housing, and for the crepes! To Alex for being so acid and funny.

A Malena y Nico, por hacerme tia. A Leo y Mitona por las cenas hipercalóricas, por las sesiones de cine i per la gran aventura transiberiana que ens va portar a recórrer més de 8000 km en 1 mes, creuar el contienent asiàtic, atravessar el desert del Gobi, menjar cucs (bé això només alguns) i fer tants canvis horaris que al final no sabíem en quin dia vivíem, però va ser болъшого путешествия. A en Lluís i a en Jesús, pels sopars d'agermanament Fort Pienc-Poblenou. Als xustis (Alba, Alba, Xavi, Marta) pels anys de uni, per les escapades, sopars, festes, per sempre estar disposats a fer el ximple i per la faceta artística. A l'Alba pels cursos de fotografia, pels dies de rebaixes, per les samarretes velles reciclades i per tots aquests anys. A en Pau, la Miriam i l'Anna per P9. A en Ferran, pels sopars. To Philly people, especially to Mac and Karel, who make me feel like I was at home, even when I was several miles away. To Kosta for the Mediterranean style. Als veïnets, per alegrar-me les tardes. Als de la colla de Blanes, per tots aquests anys de riures, tardes al bar, partits del Barça, excursions, etc. A Laura y Belén, por todos estos años. A l'Anna, per Pocatraces (http://pocatraces.blogspot.com).

Als meus pares i germans, sense el seu suport incondicional segur que ara no seria aquí. Gràcies als meus pares per transmetre'm valors, despertar-me la curiositat ja des de ben petita, recolzar-me, animar-me a fer tot allò que em proposo i ensenyar-me a lluitar per allò en que crec. Als meus germans, per aguantar-me estoicamente. Gràcies a la família en general.

I per últim, però el més important, a en Francesc, amb qui portem anys compartint camí, il·lusions i somnis. Per les teves paraules tendres i els teus somriures. Per sorprendre'm. Per estar sempre al meu cantó. Per donarme forces quan a mi se m'acaben. Pel teu suport i ajuda en detalls tècnics. Per compartir les ganes de veure món i acompanyar-me a navegar pel canal Beagle, a visitar els orangutans a Borneo i per tots els viatges que encara ens queden! I com diuen Mishima : Tan breu, com qui sap el que diu .... Simplement, gràcies.

Macarena Toll Riera
Barcelona, January 2012

# Abstract

With the availability of a high number of sequenced genomes the comparative genomics field has experienced a great advance. A wide range of studies that some years ago were unconceivable are now possible. In this thesis we aimed to study evolutionary innovations in mammalian genomes. We chose to centre our studies in mammalian species because at that moment were the genomes with higher quality and also more additional information was available for them, and of course, the inclusion of human species added a point of interest. We wished to give insights into three exciting questions in the field of evolution. First we wanted to assess which is the fraction of mammalian orthologous genes that present lineage-specific deviations in the rate of evolution. We obtained that around 25% of the genes had evidence of accelerations and decelerations specific of a branch and, surprisingly, accelerated cases did not usually overlap with cases of genes experiencing positive selection, showing that tests to detect positive selection are excessively conservative. Secondly, we wanted to deepen into the determinants driving protein evolution, centering on age of origin and structural characteristics. We used protein domains and structures to study them and we mainly found that age of origin seems to be one of the most important determinants. And finally, we investigated the characteristics and mechanisms of origin of a group of very young genes: primate-specific genes. We report that primate-specific genes evolve fast, are short and highly tissue specific. Regarding their mechanism of origin, about 53% of them showed evidence of transposable elements exaptation, 24% of partial or total duplication and surprisingly 5.5% of *de novo* origination from mammalian noncoding regions.

# Resum

Actualment, degut a la disponibilitat d'un gran nombre de genomes seqüenciats, el camp de la genòmica comparativa està experimentant grans avenços. Ara són possibles una àmplia gama d'estudis que fins fa poc eren inimaginables. En aquesta tesi hem volgut estudiar les innovacions evolutives en els genomes de mamífers. Hem escollit centrar l'estudi en mamífers degut a que els seus genomes tenen bona qualitat i hi ha més informació disponible, a més el fet d'incloure l'espècie humana afegeix interès. Ens hem centrat en tres qüestions interessants en el camp de l'evolució. Primer hem volgut determinar quina és la fracció de gens ortòlegs de mamífers que presenten desviacions específiques de llinatge en les tasses evolutives. Hem obtingut que al voltant del 25% dels gens tenen evidencies d'haver estat sotmesos a acceleracions i deceleracions específiques de branca. Hem trobat que sorprenentment, els gens accelerats normalment no solapen amb els gens amb evidencia de selecció positiva, demostrant que els tests emprats per detectar selecció positiva són massa conservadors. En segon lloc, hem aprofundit en quins són els determinants de l'evolució proteica, centrant-nos en l'edat d'origen i en les característiques estructurals. Per estudiar-ho hem utilitzat tant dominis com estructures proteiques i principalment hem trobat que l'edat d'origen és un dels determinants més importants. Finalment, hem investigat les característiques i els mecanismes d'origen d'un grup de gens molt joves: els gens específics de primats. Hem trobat que els gens específics de primats evolucionen ràpid, són curts i específics de teixit. Pel que fa al seu mecanisme d'origen, al voltant d'un 53% dels gens presenten evidencies d'haver-se originat a través de l'exaptació de transposons, 24% a partir de duplicacions parcials o totals i sorprenentment, 5.5% *de novo* a partir de regions no codificants de mamífers.

# Preface

I was very sure that the area in which I wanted to dedicate my PhD research was evolution. Why I became so excited about evolution can be summarized using the famous quote by Theodosius Dobzhansky: *Nothing in Biology Makes Sense Except in the Light of Evolution*.

It is generally accepted that the start of Evolution as a field dates from the mid eighteen century, coinciding with the publication of 'On the Origin of the Species' by Charles Darwin. But it was with the advent of genetics that the natural selection concept become viable. Currently there are several articles that include evolutionary analysis and phylogenies as part of their methodology and results, demonstrating the great success of molecular evolution. The arrival of molecular evolution age has been possible due to important advances that have permitted that now we have available several data that before was inconceivable such as protein sequences, DNA sequences, structures and expression data. And of course, thanks to the improvement of the computers, several of the most important advances have been done in the field of bioinformatics; now we can rapidly align a set of sequences, calculate evolutionary rates or build phylogenetic trees for thousands of proteins. Currently bioinformatics is essential for most of the molecular evolution studies as a huge amount of data is usually availabe.

With the publication of the first fully sequenced genome a new era started in the field of molecular evolution. The collection of completely sequenced genomes goes from bacteria and archaea to virus and eukarya, with an exponentially increasing amount of genomes. The availability of several genomes from different species allows their comparison and provides us with the opportunity to investigate several long-standing evolutionary questions that until few years ago where only approachable theoretically. For example, the analysis and comparison of several genomes showed that genomes are integrated by a mosaic of genes originated at different time points and some of these genes are novel and only found in a particular species. Now it is possible to study the origin and evolution of new genes, which are considered to be major contributors to adaptive evolutionary innovation. Newly created young genes are very exciting because they give us the opportunity to study the action of evolution and natural selection in recent times and also offer us the possibility to study the whole process of gene creation in a short evolutionary scale. We can study how genes and other genomic sequences evolve and how evolution at the molecular level can be linked to adaptation and phenotype changes at the organism level. We can also investigate the importance of natural selection in driving evolution and thus, give insights into the debate between selectionists and neutralists. Besides, several determinants of evolutionary rate have been proposed theoretically but with the advent of the bioinformatics and genomics era we can test them more easily.

At the beginning of this thesis few mammalian genomes were available, but now there are ongoing projects to sequence a huge amount of genomes, for example the Genome 10K Project (Genome 10K Community of Scientists, 2009) which aims to sequence one genome for every vertebrate genus. This project will be very useful for comparative genomics, as the disposal of genomes from closely related species provides the analysis with an increased statistical power and clarity and additionally, subtle evolutionary trends can be revealed. The advent of next generation sequencing permits the sequencing of genomes fast and cheaply, fuelling the research in exciting areas such as experimental evolution and comparative genomics. Currently, 200 years after Darwin's birth, we are living in an unprecedented era to perform evolutionary studies; we have nearly all the tools and data to resolve the unanswered questions.

This thesis focuses on studying several mechanisms related with evolutionary innovation acquisitions in mammalian genomes in the context of the comparative genomics field. The basis of comparative genomics is the comparison of homologous sequences through an alignment to determine to what extent are conserved and infer evolutionary processes. The thesis starts with a brief introduction in which I present some key concepts that I consider essential in the context of the thesis. I divided the introduction in 3 main blocks: models in molecular evolution, protein evolution and origin of new genes, which correspond to the three main areas in which my research could be embraced. The results section is divided in three main chapters. In the first one I present a published article that describes a method to detect lineage-specific deviations in the intensity of natural selection and the results obtained after applying it to a set of mammalian orthologous genes. We obtained the surprising result that around 25% of the orthologous genes in mammals have lineage-specific variations in the intensity of natural selection. In the second chapter I have included two articles that currently are submitted and deal with which are the determinants for protein evolution, specially proposing age as a very important one. The results seems to indicate that, interestingly, the time past since the birth of the gene/structure remains in the 'memory' and determines how it will evolve. The second work presented in that chapter was done during a short stay in the Mathematical Biology Group, leaded by Dr. Joshua B. Plotkin at the University of Pennsylvania. Finally, in the third chapter I include two book chapters and 2 published articles centred on the study of the characteristics and mechanisms of origin of primate orphan genes. This thesis also contains a list of objectives, a chapter with a summary of the main methodology used to reach the results and a general discussion. In the discussion I give insights in how the work presented here has contributed to increase the knowledge in the field, as well as a discussion of some methodological issues that I found while doing the thesis and which are subject to an intense debate. I also include a list of the main findings fruit of my research. Finally, to wrap up the dissertation, I give some ideas related with what I think could be future lines of research on the field. I have also included an annex with my list of publications and congress contributions.

Barcelona, January 2012

# Contents

**Acknowledgments** vii

**Abstract** xi

**Resum** xiii

**Preface** xv

**Contents** xx

**1 Introduction** 1

1.1 Brief history of (molecular) Evolution . . . . . . . . . . . . . . 1

1.2 Models in molecular evolution . . . . . . . . . . . . . . . . . . . 4

1.2.1 Neutralism . . . . . . . . . . . . . . . . . . . . . . . . . 4

1.2.2 Selectionism . . . . . . . . . . . . . . . . . . . . . . . . 10

1.2.3 The debate nowadays . . . . . . . . . . . . . . . . . . . 12

1.3 Protein evolution . . . . . . . . . . . . . . . . . . . . . . . . . . 24

1.3.1 Determinants of protein evolution . . . . . . . . . . . . 24

1.3.2 Protein domains as the unit of evolution . . . . . . . . 31

1.4 Origin of new genes . . . . . . . . . . . . . . . . . . . . . . . . . 34

1.4.1 Mechanisms of new gene formation . . . . . . . . . . . 34

1.4.2 Lineage-specific genes . . . . . . . . . . . . . . . . . . . 42

**2 Methods** 47

2.1 Genomes election . . . . . . . . . . . . . . . . . . . . . . . . . . 47

2.2 Orthologs obtention . . . . . . . . . . . . . . . . . . . . . . . . . 48

2.3 Age assignation . . . . . . . . . . . . . . . . . . . . . . . . . . . 49

2.3.1 BlastP searches . . . . . . . . . . . . . . . . . . . . . . . 49

2.3.2 Domains inference . . . . . . . . . . . . . . . . . . . . . 50

2.4 Alignment process . . . . . . . . . . . . . . . . . . . . . . . . . 51

2.5 Evolutionary rates estimation . . . . . . . . . . . . . . . . . . . 54

3 Results 57

3.1 Variations in the strength of selection in mammals . . . . . . . 59

3.1.1 Lineage-Specific Variation in Intensity of Natural Selection in Mammals . . . . . . . . . . . . . . . . . . . . 61

3.2 Age as a determinant for protein evolution . . . . . . . . . . . 79

3.2.1 The Signature of Time: Younger Domains in Proteins Evolve Faster than Older Ones . . . . . . . . . . . . . . 81

3.2.2 Structure and Age Jointly Influence Rates of Protein Evolution . . . . . . . . . . . . . . . . . . . . . . . . . . 101

3.3 Origin and characteristics of primate orphan genes . . . . . . . 119

3.3.1 Accelerated Evolution of Genes of Recent Origin . . . . 121

3.3.2 Origin of Primate Orphan Genes: A Comparative Genomics Approach . . . . . . . . . . . . . . . . . . . . 139

3.3.3 Evolution of primate orphan proteins . . . . . . . . . . 151

3.3.4 Partial Gene Duplication and the Formation of Novel Genes . . . . . . . . . . . . . . . . . . . . . . . . . . . . 159

4 Discussion 177

4.1 Variations in the strength of selection in mammals . . . . . . . 177

4.2 Age as a determinant for protein evolution . . . . . . . . . . . 181

4.3 Origin and characteristics of primate orphan genes . . . . . . . 189

4.4 Methodological issues . . . . . . . . . . . . . . . . . . . . . . . 195

4.4.1 The importance of good alignments and high quality data . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 196

4.4.2 Difficulties in the methods to detect positive selection and estimate evolutionary rates . . . . . . . . . . . . . . 199

4.4.3 GC-biased gene conversion and positive selection . . . 202

4.4.4 Methodological issues applied to section 3.1.1 . . . . . 205

5 Conclusions 207

6 Future Research 209

7 Annex 213

References 233

# 1

# Introduction

## 1.1 Brief history of (molecular) Evolution

The word evolution comes from the latin verb *evolvere* which is used to designate the action of unroll or unfold. The term started to be used in biology in the seventeenth century in the field of embryology. Although it has always been associated to Darwin, this term only appears once and at the end of his famous text 'On the Origin of the Species' (http://science.jrank.org/pages/7684/Evolution.html).

The general thought is that Evolution as a field started with Darwin. However, some basis where set up before him. Ideas relating the change of the organisms with time could be found from Aristotle to Lamarck. The naturalist Jean-Baptiste Lamarck proposed in 1809 a theory of evolution in which he hypothesized that descendants evolve to more complex and perfect forms. He believed that each lineage has been originated through spontaneous generation and that the acquired traits were inherited by the descendants. During the second half of the eighteen century and the first one of the nineteenth century the evolutionary thinking was spread (Mayr, 2002). In the mid-nineteenth century due to the high amount of fossils found and to the diversity of living organisms scientists realized that species evolve, and started to study this phenomenon. However, modern evolution as we understand today did not start until late nineteenth century, with the publication in 1859 of the book 'On the Origin of the Species' by Charles Darwin. This book came out as a result of his observations on board of the Beagle and posterior work, especially in the field of domestication, once he arrived back to England. In his book he introduces, for the first time, the concept of natural selection to explain the mechanisms driving evolution and the idea of a single common ancestor. Simultaneously, and independently to Darwin, Alfred Wallace (Wallace, 1858) also posed the importance of natural selection (Bowler, 1989). Natural selection theory states that those individuals with more favourable traits have more possibilities to survive, and, consequently, to have descendants. Until then, natural scientists had a more biblical view of the world, and this was the first time that a theory did not rely on religion. It has to be admitted that Darwin, and of course, Wallace, induced a revolution in the scientific knowledge, the world started to be viewed in constant change, instead of static (Mayr, 2002). In fact, Mayr considered it to be the greatest intellectual revolution ever happened. However, Darwin was not capable of demonstrating how features were inherited between generations because he was unaware of George Mendel work (Bowler, 1989). In 1856 Gregor Mendel, while doing experiments with peas, realized that the traits were inherited in a way that he could predict (Mendel, 1901). These experiments provide a mode of inheritance in which selection can operate because until then it was thought that inheritance blended the traits, removing the variation from the population (Hurst, 2009).

In 1900 Hugo de Vries and Carl Correns rediscovered Mendel studies. In 1904, Nuttall (Nuttall and Inchley, 1904) obtained serum proteins of several species and conducted precipitin tests, which consist on mixing sera and antisera from different species. He hypothesized that closer species would give stronger cross-reactions, and this was exactly what he observed. Those experiments allowed him to establish phylogenetic relationships for the first time: apes were the closest to humans, followed by Old World monkeys, New World monkeys and prosimians (Li, 1997; Page and Holmes, 1998). In 1920s R.A. Fisher, J.B.S. Haldane and S. Wright developed the new field of population genetics. The knowledge in population genetics showed that Mendelian genetics was fully consistent with natural selection and gradual evolution, giving rise to the modern evolutionary synthesis. Dobzhansky, Mayr, Huxley and Simpson also contributed significantly to the establishment of evolutionary synthesis, applying the principles of genetics to natural populations (Bowler, 1989). During the evolutionary synthesis development scientists reach a consensus to describe what evolution is: *Evolution is change in the properties of populations of organisms over time*, being population the unit of evolution (Mayr, 2002). The modern evolutionary synthesis became rapidly very popular because it provided an explanation for the history and the diversity observed. With the discovery of the DNA structure in 1953 by Rosalind Franklin, James Watson and Francis Crick (Watson and Crick, 1953) the physical basis of inheritance were demonstrated.

In the fifties a debate was originated between the classical school and the balance school of population genetics regarding the amount of genetic variation found in a population, and this was the seed for the posterior debate between neutralist and selectionists. The new advances in molecular biology techniques gave rise to electrophoresis, which was crucial to show that genetic variation was abundant within species as well as between species (Page and Holmes, 1998). Besides, in 1952, Sanger and colleagues determined, for the first time, the sequence of a protein: insulin. Following insulin, many other sequences were obtained, and consequently, the study of amino acid substitutions between species started. With it started a period of rapid progress (Li, 1997). In 1962, after performing comparative studies of hemoglobins and cytochromes *c*, Zuckerkandl and Pauling (Zuckerkandl and Pauling, 1962) proposed the existence of a molecular clock, which had a great impact in the field. The clock was used to make estimations on the divergence times between species, but it was also used as a strong support for the neutral theory, proposed by Kimura (Kimura, 1968) and King and Jukes (King and Jukes, 1969).

Neutralism was born as a way to give an explanation to the surprisingly high levels of genetic variation found, arguing that the majority of fixed mutations do no have any effect on fitness and are neutral. Thus, since the sixties there are two models to explain how evolution is occurring: in the neutralist model the genetic drift of neutral mutations is the most important process while in the selectionist model the evolution is governed by natural selection acting over advantageous mutations (Page and Holmes, 1998). In 1973, Ohta (Ohta, 1973) proposed the nearly neutral theory, in which she hallmarked the importance of slightly deleterious mutations. Besides, during the sixties and seventies, the high amount of accumulated sequences enhanced studies related with the reconstruction of phylogenetic trees and the development of methods to construct them (Li, 1997).

In the eighties, the development of the polymerase chain reaction (PCR) produced an increase in the number of available sequences for evolutionary analysis. Apart from the improvements in laboratory techniques, it has to be also acknowledged that the progress in molecular evolution and population genetics was partly due to the arrival of high-speed computers (Li, 1997).

With the sequencing of the first genome in 1995, *Haemophilus influenzae*, (Fleischmann et al., 1995), started a new era in molecular biology. Five years after the first draft of the human genome was announced, and shortly after published in Nature (Lander et al., 2001). Since then, there has been a burst of sequencing projects and the number of sequences in the public databases has grown exponentially. With the advent of next generation sequencing, which reduces the costs of genome sequencing and speeds up the whole process, several projects have appeared that aim sequencing a huge number of genomes, such as the 1000 genome project, which objective is to sequence more than 1000 humans of different populations in order to study human genetic variation (Durbin et al., 2010) or the Genome 10K Project which aims the sequencing of 10,000 vertebrate species to gain insights on vertebrate evolution (Genome 10K Community of Scientists, 2009). The high amount of available sequences and the development of new sequencing techniques has opened paths in evolutionary studies that until now were inconceivable.

## 1.2 Models in molecular evolution

There are four basic mechanisms of evolution:

- Mutation: is the raw material for evolution because produces the alteration and variation of the DNA sequences. It is caused by errors during the replication process. Mutation and substitutions can be confounded. A nucleotide mutation is a synonymous or non-synonymous base change in an individual of a population. A substitution is a base change between two populations. If a mutation is fixed in the population, then it becomes a substitution.

- Selection: is the differential capability of genotypes or individuals genetically distinct to survive and reproduce, measured with the fitness. If a mutation lowers the fitness it would be eliminated by purifying selection. On contrary, if it confers a higher fitness it will be fixed through positive selection. Positive selection can be subdivided into directional selection when it favours the fixation of an advantageous allele, and balancing selection, when it acts toward maintaining the polymorphism.

- Genetic drift: is the process that changes, randomly, the allele's frequencies. A source of this random change is for example, the random sampling of the gametes in each generation.

- Migration: different populations have different allele frequencies, and when migration takes place an interchange of genes between the host population and the migrated population is produced, resulting in a change in the allele frequencies.

One of the main questions that are still under debate in the field of molecular evolution is which is the mechanism driving evolution. Two main models have been proposed: selectionism and neutralism. In the first one evolution is governed by natural selection while in the second one genetic drift and neutral mutations are the dominating forces.

### 1.2.1 Neutralism

The neutral theory of molecular evolution was proposed by Kimura (Kimura, 1968) and King and Jukes (King and Jukes, 1969) in the late sixties in an attempt to explain two limitations of selectionism. The first one is the prediction done by selectionists that polymorphic sites should be rare. But, on contrary, with the advent of electrophoresis, a high amount of genetic variation was observed. The second limitation was the Haldane dilemma: for each selective substitution event a selective death should occur, and for this reason, population size should limit the rate of adaptive evolution. But the calculation of the rate of evolution for several proteins was too high to be explained entirely by evolution (Hurst, 2009).

The neutral theory is based on the assumption that most of the mutations are neutral or deleterious. Deleterious mutations decrease the fitness of the individual, and therefore, are eliminated by negative selection. On contrary, neutral mutations do not have any effect on the fitness, and for this reason are not subject to natural selection. Most of them are lost but some become fixed through genetic drift (Li, 1997; Page and Holmes, 1998).

Neutralists and selectionists coincide in the prediction that most mutations are deleterious, but they disagree in the amount of neutral and advantageous mutations (figure 1.1). Neutralists think that chance (genetic drift) plays the main role in the fixation process, while selectionists argue that is necessity (natural selection) what is governing the process (Page and Holmes, 1998). They also disagree in the definition of polymorphic sites, while neutralists see polymorphism as a transient phase to fixation, selectionists argue that polymorphic sites are not transient because they are maintained by selection (Li, 1997).

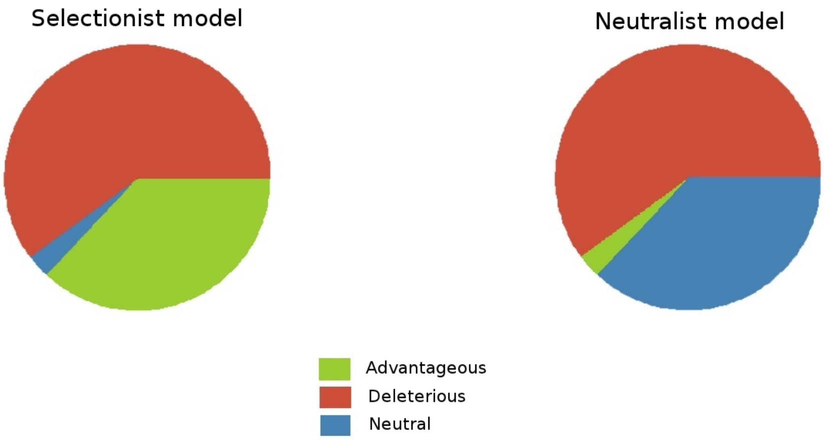


**Figure 1.1:** Comparison between the predicted frequency of each type of mutations arising in a gene using selectionist or neutralist model. Adapted from Page and Holmes (1998)

However, neutralism does not debate the idea of evolution by natural selection; indeed, neutralists think that proteins are well adapted due to the past action of natural selection. However, they argue that the fixation of mutations that confer a selective advantage occurs at a very low frequency, and therefore, most mutations will be rather neutral or deleterious (Page and Holmes, 1998).

One of the best evidences for neutralism is the reported correlation between protein substitution rates and functional constraints. Most of the mutations occurring in highly constrained genes are deleterious and consequently eliminated. On contrary, in lowly constrained genes the majority of the mutations would be neutral and therefore, would not be eliminated, provoking and increase in the substitution rates. Hence, changes in substitution rates could be explained by changes in gene functional constraints (Page and Holmes, 1998).

### The molecular clock

One of the strongest supports for the Neutral theory is the existence of a molecular clock, which was observed for the first time in the early sixties by Zuckerkandl and Pauling (Zuckerkandl and Pauling, 1962). They realized that the number of amino acid changes that has taken place between pairs of globin sequences belonging to different species increased proportionally with the distance separating the compared species (figure 1.2). The molecular clock states that mutations are accumulated approximately at a constant rate, and therefore, that evolution is constant.

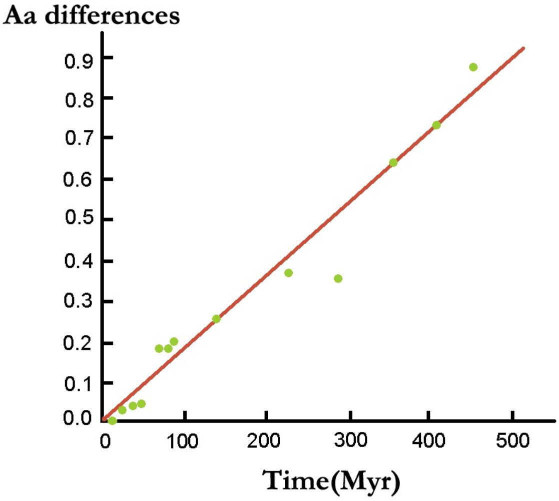


**Figure 1.2:** Constant rate of evolution in the alpha globin protein. Dots represent pairs or groups of species. Modified from Ridley (2004)

Under the Neutral theory the rate of substitution of a neutral mutation is only given by the mutation rate, and it is independent of the population size.

This could be explained as follows:

- The rate of nucleotide substitution at a nucleotide site per year ($k$) in a diploid population of size 2N is given by the number of new mutations arising per year ($\mu$) multiplied by their probability of fixation ($u$).

  $$k\text{=}2\mathrm{N}\mu u$$

- The probability of fixation of a neutral mutation is:

  $$u\text{=}1/2\mathrm{N}$$

- Therefore, the rate of substitution of a neutral mutation is:

  $$k\text{=}(2\mathrm{N})(1/2\mathrm{N})\mu$$

  Which can be simplified into: $k\text{=}\mu$.

On contrary, the substitution rate of selectively advantageous mutations depends on the mutation rate, the population size and the magnitude of the advantage.

- In the presence of selectively advantageous mutations the probability of fixation is given by the magnitude of the selective advantage ($s$) and the effective population size ($\mathrm{N}_e$). When selection is taken into account it has to be used the effective population size, which is the subset of the total population that is able to reproduce, and therefore, are the individuals that could be affected by the action of selection.

  $$u\text{=}2s\mathrm{N}_e/\mathrm{N}$$

- Then, the rate of substitution of an advantageous mutation is given by:

  $$k\text{=}4\mathrm{N}_e s\mu$$

Hence, it is easier to explain the molecular clock from the point of view of neutral theory than of selectionism. To explain it using neutral theory only one variable should be constant, whereas, to explain it by natural selection three variables should be constant, which is more improbable. Therefore, for neutralists, the existence of the molecular clock gives support to the neutral theory because it proves the constant rate of neutral mutation rate (Page and Holmes, 1998). Given the molecular clock, evolution at neutral sites can be used to estimate the mutation rate. Some years after Kimura and Ota (Kimura and Ota, 1974) pointed out that protein evolutionary rate will be maintained constant as long as the tertiary structure and the function of the protein does not change.

### Overdispersion of the clock

The existence of a molecular clock could be tested using the variation found between lineages. Ohta and Kimura (Ota and Kimura, 1971) suggested that if evolution is following a molecular clock, in other words, if it is constant, the fixation of mutations should follow a Poisson distribution. If the variation observed between lineages is higher than the expected under a Poisson process it means that neutral mutations are not being accumulated as a constant clock and it is said that the clock is overdispersed. On contrary, if the data follows a Poisson distribution, the variance in the substitution rate is smaller than the mean substitution rate (the dispersion index, R(t) is not significantly greater than 1), and, consequently, the existence of a molecular clock can be inferred (Page and Holmes, 1998). Just after the formulation of the neutral theory appeared the first examples in which the data did not follow a Poisson process.

### Variations in the substitution rates

Under neutral theory the rate of substitution of a neutral mutation is equal to the mutation rate, thus, variations in the substitution rate could be explained by a variation in the mutation rate. The mutation rate could vary between species mainly due to three lineage effects: differences in generation time, metabolic rate and DNA repair efficiency (Li, 1997; Page and Holmes, 1998).

- Generation time. Is the time of germ-line replication. If we assume that most mutations take place during this process and that the number of cell divisions per generation is similar between species, then, over the same period, species with shorter generation time would have more substitutions than species with longer generation time (Laird et al., 1969; Kohne, 1970). For example, in the mouse lineage the number of nucleotide substitutions fixed per unit of time is the double than in the human lineage. This has been attributed to the shorter generation time of rodents (Waterston et al., 2002).

- Metabolic rate. This hypothesis was proposed in 1993 by Martin and Palumbi (Martin and Palumbi, 1993). During aerobic respiration free-oxygen radicals with mutagenic effects are produced, therefore, species with higher metabolic rates will have higher rates of mutation due to a higher DNA synthesis and also to a higher amount of free-oxygen radicals.

- DNA repair efficiency. Britten (Britten, 1986) proposed that if species present differences in their mechanisms of repairing mutations, then, those species with more efficient mechanism would be evolving at lower rates.

### The nearly neutral theory

To explain the overdispersion of the clock and the lower than expected levels of heterozygosity Ohta proposed in the early seventies the nearly neutral theory (Ohta, 1973). In this theory she states that most nonsynonymous substitutions are slightly deleterious (in the nineties slightly advantageous mutations were incorporated) or nearly neutral. Thus, in the nearly neutral theory the rate of substitutions not only depends on the mutation rate (set to generation time) as in the neutral theory, it also depends on selective coefficient and population size. Substitutions at synonymous sites and in noncoding DNA are considered as strictly neutral, therefore, their substitution rate would be given by the mutation rate and the species generation time (Page and Holmes, 1998).

With the incorporation of slightly deleterious mutations in the theory the size of the population became important. The importance of (negative) selection and genetic drift for the fixation of the allele would rely on the population size. Two classes of mutations can be fixed: effectively neutral and slightly deleterious. Genetic drift has a stronger effect in small population sizes, thus the probability of fixation in those populations is higher. On contrary, purifying selection is higher in large populations, lowering the probability of allele fixation. A mutation will be 'effectively neutral' if it's selective disadvantage is smaller than the inverse of the effective population size ($N_e$), $s << 1/2N_e$, or in other words, if it's substitution rate is similar to neutral mutations. Hence, evolution by effectively neutral mutations is more common in species with smaller population sizes. Similarly, the classification of a mutation as slightly deleterious depends on the effective population size; the same mutation can be slightly deleterious in bacteria (larger populations) but effectively neutral in mammals (smaller population size). For this reason it has been predicted that selection unlikely affect synonymous mutations in species with small population size, but it could affect the codon usage of species with large populations, in which the usage of synonymous codons is biased matching the abundances of tRNA to maximize the rate of protein synthesis, especially in highly expressed genes (Chamary et al., 2006). However, there are several evidences of non-neutral evolution at mammalian synonymous sites, as it will be discussed later (page 201).

Proteins and nonsynonymous substitutions have been reported to fit better into a real time molecular clock than to a generation time one. This incongruence can be explained using the nearly neutral theory: the rate of evolution would be maintained constant as long as there are slightly deleterious mutations and there exists a negative correlation between population size and generation time. Species with large populations tend to be small and to have short generation times (for example mice), while, small population size species have bigger bodies (for example humans) and have long generation times. For this reason, due to the action of genetic drift, a higher number of nearly neutral mutations is fixed in species with small population sizes, but as these species have a long generation time, they have fewer mutations per year. Besides, in large populations the effect of negative selection is stronger and the probability of fixation of the mutations is low, but they have a higher number of mutations per year because their generation time is short, and consequently, they have higher mutation rates. This explains why the rate of evolution is approximately constant per year among species (Page and Holmes, 1998; Eyre-Walker et al., 2002).

Synonymous substitutions depend on generation time, while nonsynonymous rates tend to follow real time. Both types of substitutions show an overdispersion of the clock, but when they are corrected for lineage-specific effects, the overdispersion drops in the case of synonymous, but not in the nonsynonymous, indicating than synonymous substitutions are more prone to vary due to lineage effects (Li, 1997; Page and Holmes, 1998).

### 1.2.2 Selectionism

The bases of selectionism started with Darwin and his book 'On the Origin of Species' (Darwin, 1859). Lately, in the first half of the 20$^{th}$ century Dobzhansky, Huxley, Mayr and Simpson developed the Modern Synthesis or also called, Neo-Darwinism (Koonin, 2009). Until the development of the neutral theory in 1968, selectionism was the only model that existed to explain the evolution.

Under selectionist view, most mutations are deleterious and are removed from the population by negative selection. However, there are some few mutations that confer a selective advantage and are fixed by selection. On contrary to neutralists, selectionists think that neutral mutations are rare. For them the main force governing evolution is natural selection, while mutation and genetic drift have a secondary role. Polymorphisms are understood as a way of maintaining two advantageous alleles in the population through the action of balancing selection.

Gillespie (Gillespie, 1986) used the dispersion index (R(t)) to test if, on contrary to what it was proposed by neutralism, natural selection was the driving force in molecular evolution. He studied the molecular clock in mammals and found that there were periods of stasis and periods with a high number of substitutions possibly indicating a process of adaptation of organisms to changing environments, which favour different mutations at different times. Thus he proposes that the existence of an episodic molecular clock is the evidence that natural selection plays the main role in molecular evolution. However, Ohta (Ohta, 1995) argued that this episodic clock can also be explained by changes in the population size: the periods of high nonsynonymous substitutions are due to the fixation of nearly neutral mutations in small populations (for example, during bottlenecks) by the action of genetic drift (Li, 1997; Page and Holmes, 1998).

There are several evidences demonstrating the action of the natural selection at a molecular level, some of them are the following (Page and Holmes, 1998):

- Convergent evolution. When two unrelated lineages have acquired the same trait. One clear example are the stomach lysozymes present in ruminants and in leaf-eating monkeys. This case is easy to explain using natural selection but is inexplicable using genetic drift.

- Cases in which there are two polymorphisms maintained in the population such as the fast and the slow alleles in the *Adh* gene of *Drosophila* which have different capacities to process alcohol. These polymorphisms are maintained by the action of balancing selection because it has been observed that in the proximity of the polymorphism there are more silent polymorphisms than expected under neutral theory. If the genetic drift was the only force neutral variation would not had been maintained for so long.

- There are many genes which have a higher number of nonsynonymous substitutions than synonymous ones, suggesting the action of positive selection. Several examples are found related with immune system such as the major histocompatibility complex and in genes related with the arms race between host and parasites.

- The presence of linked selection: hitchhiking and background selection. Hitchhiking is the fixation of alleles due to their link to positively selected advantageous alleles. Background selection is the contrary, deleterious alleles are eliminated and with them other alleles that are in linkage disequilibrium.

### 1.2.3 The debate nowadays

Since 40 years ago there has been a debate between selectionists and neutralists (Kimura, 1983; Gillespie, 1991) regarding which is the importance of advantageous mutations for molecular evolution, in other words, whether the fixation of mutations is driven by selection or genetic drift. However, until recently, the data needed to support one or the other point of view was not available (Eyre-Walker, 2006; Eyre-Walker and Keightley, 2009). Nowadays there is a high amount of polymorphism data and several genomes availabe, which could help us to shed light into one of the most important and appealing questions in evolution.

The support for the neutral theory has been obtained from the comparison of distantly related species, while the support for natural selection cames from comparisons involving short periods of time. Hence, one possibility could be that with time the action of natural selection is erased due to the accumulation of neutral mutations. However, the debate is still unresolved (Page and Holmes, 1998). Below I present a brief review with the latest articles related with this subject.

#### The clock and its overdispersion

Kumar and Subramanian (Kumar and Subramanian, 2002) calculated mutation rate between pairs of mammals, and reported mutation-rate differences between all the comparisons: 20% between hamsters and mice, 14% between cows and pigs, 23% between cats and dogs and 22% between humans and Old World Monkeys. Thus, the differences in the mutation rates indicate that a global DNA clock does not exist for mammals. Generation time and other life-history traits are similar between the used pairs; therefore, the source of mutations should be processes that are independent from replication, such as, recombination, repair mechanism and DNA methylation.

Bedford and colleagues (Bedford et al., 2008) have studied the overdispersion of the molecular clock in 3 different groups of species: mammals, *Drosophila* and yeast. They found that while mammalian proteins were highly overdispersed, yeast proteins were little overdispersed and *Drosophila* ones were middle-overdispersed. However, all of them have a substitution pattern with a variance higher than the expected under a Poisson process. They related their results with the effective population size of the species, suggesting that overdispersion could be correlated negatively with the effective population size. They also suggest that as mutational robustness (insensitivity to changes caused by mutations) is more prevalent in populations with large sizes because the strength of selection is greater, species with larger population size are less overdispersed because mutational robustness is stabilizing the molecular clock from the sequence changes. However, they point out that the index of dispersion rather than being a test of the neutral theory is a test of the heterogeneity of sequence evolution because an R(t)>1 is not indicative of the presence of positive selection. Additionally, complex models have suggested that in some occasions neutral substitutions can be overdispersed. Similar results have been obtained using lattice protein simulations instead of real proteins (Bloom et al., 2007).

A detailed analysis has been done using three proteins, glycerol-3-phospate dehydrogenase (GPDH), superoxide dismutase (SOD) and xanthine dehydrogenase (XDH) which had an erratic evolution across time and lineages that took place differently across locus. The authors argued that these observations could be better explained by the action of selection (Rodríguez-Trelles et al., 2001). The majority of the results showed that nonsynonymous substitutions are highly overdispersed, but on contrast, an analysis of nonsynonymous rate variation in mammalian lineages showed that they are not overdispersed. The authors of the study found that most mammalian proteins are under purifying selection which suppresses the variance of evolutionary rates due to other factors. The authors posed that they obtained different result because their data was not biased through the inclusion of a higher amount of hormones, as previous studies, which tend to have larger index of dispersion (Kim and Yi, 2008).

### Slightly deleterious mutations really exist

The majority of the proteins have highly conserved sequences, and this is an evidence of the deleterious effect of most mutations (Piganeau and Eyre-Walker, 2003). It has been estimated that 20% of the nonsynonymous mutations in humans are slightly deleterious (Fay et al., 2001). Nowadays there is no doubt on the real existence of slightly deleterious mutations; they are evidenced by 4 facts (Charlesworth and Eyre-Walker, 2007):

- It has been observed in various species that nonsynonymous mutations segregate at lower allelic frequencies than silent mutations.

- Species with smaller population sizes have comparatively more nonsynonymous substitutions than synonymous ones.

- Regions of the genome with small effective size, such as sex chromosomes, show a higher ratio of nonsynonymous to synonymous substitutions.

- There are several cases in which the ratio of nonysnonymous to synonymous polymorphism is higher than the ratio of nonsynonymous to synonymous substitutions. These cases could be due to the presence of slightly deleterious nonsynonymous mutations, which are rarely fixed but contribute to polymorphism.

Nearly neutral theory predicts that in species with small effective population size there would be a higher fixation of slightly deleterious mutations due to the action of genetic drift. Thus, species with small effective population sizes should have a higher overall ratio of nonsynonymous to synonymous substitutions ($d_N/d_S$). Data from mammalian genomes supports the accumulation of slightly deleterious mutations in species with small population size: $d_N/d_S$ in hominids, which have small population sizes, is higher than in rodents, which have larger population sizes. Therefore, this data is compatible with the nearly neutral theory formulated by Ohta (Ellegren, 2008).

But, if there are slightly deleterious mutations, slightly advantageous mutations should exist due to the back-mutation of the slightly deleterious mutations. In other words, if a slightly deleterious mutation is fixed but afterwards a new mutation occurs in the site that re-establishes the nucleotide that was present before the slightly deleterious mutations, such mutation is slightly advantageous. As explained before, under the nearly neutral model the rate of substitutions is expected to be higher in populations with small effective size. However, if advantageous mutations are added to the model, if there is an increase in the effective population size the substitution rate will be increased temporarily due to the fixation of advantageous mutations that until now where effectively neutral. If only deleterious mutations are taken into account the rate of substitution correlates negatively with population size, independently of expansions in the population size (Gillespie, 1994; Takano-Shimizu, 1999). Charlesworth and Eyre-Walker (Charlesworth and Eyre-Walker, 2007) hypothesize that back-mutations could be detected by comparing the divergence between species that are in island with the divergence in species that are mainland, which should be different due to the differences in the population sizes.

If the colonization if from mainland to island the population size would be contracted, while if the colonization if from island to mainland the population size would be expanded. The rate of evolution should be higher in the second case than in the first one, and this is what they mainly find. They demonstrated statistically that adaptive evolution can take place through back-mutations. Now the question that should be resolved is which fraction of the adaptive evolution is due to back-mutation and which one to changes in the environment.

### Fraction of neutral, advantageous and deleterious mutations

Mutations can be harmful, beneficial or neutral, but rather than these three well defined categories, what we found is a continuous effect of mutations, being, for example, from slightly harmful to highly harmful. In fact, it is difficult to imagine a really neutral mutation, however, what we can imagine are effectively neutral mutations. Effectively neutral mutations do not depend only on the effect of the mutation; they depend also in the effectiveness of the natural selection in the population where the mutation has appeared (Eyre-Walker and Keightley, 2007). There are some estimations of the proportion of nonsynonymous mutations that are neutral in protein-coding sequences, such as 30% in humans (Eyre-Walker and Keightley, 2007), 16% in *Drosophila* (Eyre-Walker, 2002) and 2.8% in enteric bacteria (Charlesworth and Eyre-Walker, 2006). For deleterious mutations it has been estimated that as much as 70% of the amino acid mutations have a strong deleterious effect (rarely are fixed) and less than 10% of the mutations are slightly deleterious (Eyre-Walker et al., 2002). There are some estimates on the amount of advantageous mutations. Using mutagenesis studies it has been confirmed that there is a small amount of advantageous mutations, only 4% of the mutations in the vesicular stomatitis virus were advantageous (Sanjuán et al., 2004), between 0 and 15% in the bacteriophage ΦX174 (Silander et al., 2007), 0% in *Escherichia coli* (Elena et al., 1998) and 6% in *Saccharomyces cerevisiae* (Thatcher et al., 1998). The small fraction of advantageous mutations is not an indicative of their lack of contribution to the evolutionary change, in spite of, they can contribute to it substantially. In fact, more than 15% of the substitutions in *Drosophila melanogaster* are due to advantageous mutations (Eyre-Walker and Keightley, 2007).

The effect of mutations over fitness depend on the population effective size. Experiments using small and large population sizes of the bacteriophage X174 were performed, obtaining that lines with smaller population sizes had lower fitness because they had accumulated more deleterious mutations. Despite of this, these lines had a higher fraction of adaptive mutations than large population size lines (Silander et al., 2007).

### Estimations on the adaptive rates

Several methods based on the McDonald-Kreitman (MK) test (McDonald and Kreitman, 1991) have been developed to estimate the proportion of adaptive substitutions and to test the validity of the neutral theory of molecular evolution (Eyre-Walker, 2006; Eyre-Walker and Keightley, 2009). The basis of this test is that if a mutation is beneficial it would be fixed quickly and therefore, it would contribute more into divergence than into polymorphism. The number of nonsynonymous ($p_N$) and synonymous ($p_S$) polymorphism and nonsynonymous ($d_N$) and synonymous substitutions ($d_S$) (divergence) are compared. If $d_N/d_S$ is greater than $p_N/p_S$ indicates that while synonymous mutations are neutral, some non-synonymous substitutions have been fixed through positive adaptive evolution. If mutations are neutral or strongly deleterious $d_N/d_S$ and $p_N/p_S$ would be almost equal. MK test is mostly robust, except when there is segregation of slightly deleterious nonsynonymous mutations because this type of mutations tends to contribute more to polymorphism than to divergence. Other method that could be used to estimate adaptive substitutions is the rate of nonsynonymous substitutions compared to the rate of synonymous substitutions; if $d_N$ is greater than $d_S$ we can assume that the gene has experienced adaptive evolution (Eyre-Walker, 2006).

Several estimations have been done regarding which fraction of protein-coding sequences has undergone adaptive evolution (table 1.1). The first studies were done in *Drosophila* using the MK test. One study pointed that around 45% of the observed amino acid substitutions were due to the action of positive selection (Smith and Eyre-Walker, 2002). Other study reported that the ratio of divergence was two time higher than the ratio of polymorphism (Fay et al., 2002). Begun and colleagues have found that around 54% of the nonsynonymous differences have been fixed by positive selection in *Drosophila simulans* (Begun et al., 2007a). High values have also been reported in enteric bacteria (Charlesworth and Eyre-Walker, 2006) and virus (Nielsen and Yang, 2003) . On contrary, most of the estimates done in hominoids using MK test have given values close to 0 (Chimpanzee Sequencing and Analysis Consortium, 2005; Zhang and Li, 2005), except the one done by Bustamante et al. (Bustamante et al., 2005), in which they found a fraction around the 6%. Using $d_N/d_S$ tests (branch-site model) Clark and colleagues (Clark et al., 2003) have found that only 0.08% of the hominid genes show evidence of adaptive evolution, a similar fraction was found by Nielsen et al. (Nielsen et al., 2005). In another recent study the authors have performed a scan to detect positive selection in a set of mammalian genomes and they found that 400 genes out of 16,500 presented evidence of positive selection, while 144 more had lineage-specific positive selection (Kosiol et al., 2008). Data from Arabidopsis (Foxe et al., 2008) and yeast (Liti et al., 2009) is similar to the results obtained in humans.

| Organism | Methodology | N adaptive substitutions | Reference |
|---|---|---|---|
| *H. sapiens* | MK | 0% | Zhang et al. (2005) |
| *H. sapiens* | MK | 0-9% | Chimpanzee Sequencing and Analysis Consortium (2005) |
| *H. sapiens* | MK | 6% | Bustamante et al. (2005) |
| *H. sapiens* | MK | 40% | Eyre-Walker and Keightley (2009) |
| *H. sapiens* | $d_N/d_S$ | 0.08% | Clark et al. (2003) |
| *M. musculus* | MK | 57% | Halligan et al. (2010) |
| *D. simulans/D. yakuba* | MK | 45% | Smith and Eyre-Walker (2002) |
| *D. simulans* | MK | 54% | Begun et al. (2007b) |
| *D. melanogaster* | MK | 50% | Eyre-Walker and Keightley (2009) |
| Yeast | MK | 0 | Doniger et al. (2008) |
| *E. coli*/S. enterica | MK | >50% | Charlesworth and Eyre-Walker (2006) |
| *A. thaliana* | MK | 0 | Bustamante et al. (2002) |
| *A. thaliana* | $d_N/d_S$ | 5 | Barrier et al. (2003) |
| HIV | $d_N/d_S$ | 75[a] | Nielsen and Yang (2003) |
| Influenza | $d_N/d_S$ | 85[a] | Nielsen and Yang (2003) |

[a] Proportion of codons showing evidence of adaptive evolution

**Table 1.1:** Estimates of adaptive evolution

Eyre-Walker and Keightley (Eyre-Walker and Keightley, 2009) have recently developed a method to estimate the rate of adaptive molecular evolution taking into account the presence of slightly deleterious mutations. Slightly deleterious mutations should be taken into account because they contribute to polymorphism, but they rarely contribute to divergence because they have little chance of fixation, hence, their exclusion could lead to underestimations. They have tested the method using data from human and *Drosophila*. They estimated that around 30-40% of the human mutations are effectively neutral, while this number is as low as 6% in *Drosophila*. Regarding the adaptive substitutions, there are also big differences between human and *Drosophila*. Around 50% of the substitutions in *Drosophila* are due to the action of positive selection. On the contrary, the fraction of advantageous substitutions in humans is very low. Both of their estimates are in concordance with previous results. These differences in the fraction of advantageous substitutions between *Drosophila* and humans could be explained by their different population size (Eyre-Walker, 2006). However, the authors think that the results in humans could be due to an artefact caused by the existence of a difference between the effective population size of humans and hominids. When they corrected for this artefact they found that around 40% of the human substitutions are adaptive, a number which is close to the value obtained in *Drosophila*.

One of the last articles centred on estimating the fraction of adaptive protein evolution has used wild mice. The authors have used the previously commented method developed by Eyre-Walker and Keightley to estimate the adaptive molecular evolution (Eyre-Walker and Keightley, 2009). In this study it is found that around 57% of the amino acid substitutions in wild mice can be explained by positive selection (Halligan et al., 2010).

Most of the species in which the rate of adaptive substitutions has been calculated present high estimates, bringing the neutral theory into question and suggesting that positive selection plays a more important role than genetic drift in protein evolution. However, there are some exceptions such as humans, yeast and Arabidopsis. Although, yeast and Arabidopsis results could be explained because slightly deleterious mutations have not been taken into account in the calculations. Results in humans could be explained by their small effective population size (Halligan et al., 2010). Alternatively, the results could also be caused by a change in the effective population size (Halligan et al., 2010), in fact, it has been suggested that the effective population size in the ancestral great ape was larger than the actual one (Burgess and Yang, 2008).

Thus, the fraction of genes that have experienced adaptive evolution seems to be correlated with the population size. In fact, in large populations there are more mutations, and selection is more effective if there are more mutations. Consequently, this can mean that species with large population sizes can adapt better to their environment compared to small population size species. But, it can also be that adaptations are due to some strong mutations that are fixed on the population independently of its size, and the excess of adaptive mutations found in populations with large sizes are only acting as slight adjustments (Eyre-Walker, 2006).

Another interesting viewpoint in the study of adaptive protein evolution is the one taken by Shapiro and Alm (Shapiro and Alm, 2008). They developed a methodology called selective signatures to detect lineage-specific deviations in the 'expected' protein evolutionary rate independently of genome and gene-family specific rates. Genome rate variations are due to differences in species characteristics such as generation time and population size, while gene-family variations are related to the specific function of the gene. Their methodology is powerful because it allows detecting selection even if the $d_N/d_S$ values are small. They have performed their study using a set of 30 $\gamma$-proteobacterial genomes. They identified several cases of rapid evolving genes that could be related with an ecological shift, and moreover, observed that pairs of genes with similar selective signatures are more likely to share the same cellular function. Interestingly, they found that the acceleration of gene evolutionary rates could be explained by two factors with a similar weight: positive selection and relaxed negative selection.

### Epistasis

Epistasis is the process by which the effects of one mutation are modified by a posterior mutation, for example, mutations with weak effects can have stronger effects in combination with another mutation, or a destabilizing mutation that is beneficial can be stabilized by a posterior mutation. This factor is closely related with the selectionist-neutralist debate because several lines of evidence suggest that interactions between neutral and beneficial mutations determine evolution. In fact, mutations can be conditionally neutral (or cryptic), meaning that a mutation that in principle is neutral in the background in which arises because it does not alter the fitness, it can latter increase the fitness of subsequent mutations by interacting epistatically. Arising as neutral mutations is very important because allows them to reach a high frequency. Indeed, it has been shown theoretically, using simulated populations of RNAs, that conditionally neutral mutations are more frequent than expected during the adaptive process of a population, and, especially, facilitate adaptation in temporarily stuck populations (Draghi et al., 2011). In a very recent study the authors have found that those populations with higher rates of cryptic variation also have a higher rate of adaptation, indicating that the adaptation to a new environment was facilitated by cryptic genetic variation, which contains pre-adapted new genotypes to changing environments. Epistasis plays a key role in this system because is the combination of mutations what provides the fitness advantage (Hayden et al., 2011). There are also studies in which epistatic interactions are identified. One example is found in the article by Bridgham et al. (Bridgham et al., 2006), in which the authors study how the cortisol-specificity arose in the glucocorticoid receptor. They reported a pair of mutations in which the first one, Leu111Gln did not have much effect, but when followed by Ser106Pro it enhanced the cortisol-specificity.

Thus, summarizing, there are several evidences that epistatic interactions between some neutral mutations facilitate future adaptations. These studies point out the existence of truly neutral mutations and the importance of these neutral mutations for adaptation.

### Robustness and neutral changes

It has been typically assumed that those changes that slightly affect the phenotype or do not affect it would not be important for evolutionary innovations. However, in the last years several examples have been accumulated in which neutral changes fuel innovations. Robustness plays an essential role, the more robust a phenotype, is the greater is the number of mutations that do not affect it (Wagner, 2011). One nice example is the study done by Bloom and colleagues (Bloom et al., 2006b). The authors used the cytochrome P450 to study the acquisition of new enzymatic functions depending on the stability and robustness to mutations of the enzyme. They found that those P450 variants that were more robust and stable were the first ones to acquire new functionalities, as these variants were the only ones able to tolerate the destabilizing mutations that entail the gain of function. Hence, they demonstrated that protein stability promotes evolvability. Similarly, Ferrada and Wagner (Ferrada and Wagner, 2008) observed that more robust proteins, in which most of the mutations are neutral or weak, show a greater functional diversity in the evolutionary history.

### Individual opinions

Despite all the studies performed and all the knowledge acquired in these years that has facilitated the empirical testing of theoretical predictions, there is still controversy regarding the role of neutral and beneficial mutations. There are several studies supporting the selectionist view including those assessing rates of adaptive substitutions in species with large population sizes, however, there is also a number of studies which reinforce the importance of neutral mutations and robustness for adaptation. Below I'm going to summarize briefly some recent opinions and one attempt to reconcile both opposing views.

In a recent commentary Hahn (Hahn, 2008) expresses his view regarding the actual validity of Neutral Theory given the last findings. He thinks that recent studies involving several species have proved that a high number of amino acid substitutions have been fixed by adaptive natural selection, which is contrary to Neutral Theory expectations. He enumerates several predictions done by the Neutral Theory that the actual data does not fit with:

- Genetic diversity should be linearly proportional to population size. However, several measurements have been done across the tree of life and, although population size greatly varies among taxa, the mean difference in nucleotide diversity is only of two orders of magnitude between vertebrates and prokaryotes.

- Positive correlation between divergence and polymorphism because both measures are given by the neutral mutation rate. In fact, what it is observed from data is a negative correlation. Precisely, this is what is expected by selectionist models: if mutations are advantageous, a high fraction of them would be fixed producing a decrease in the polymorphism levels due to hitchhiking.

- No correlation between polymorphism and recombination is expected because the number of neutral mutations is independent of the recombination process. Indeed, a positive correlation between polymorphism and recombination has been reported in several species. This correlation could be explained by selectionists: linked selection (hitchhiking, background selection) is acting across the genome, those regions with high recombination rates can avoid the effects of nearby selection, and therefore, polymorphisms are kept.

Hahn argues that the Neutral Theory has already a set of statistical tools that enables the testing of the hypothesis, but that this is not the case of the Selection Theory. He thinks that in spite of that at first glance the Neutral Theory seems more parsimonious and easier to parameterize, the actual data suggest that selectionist view is more correct. However, a Selection Theory with new estimation methods should be developed.

Eugene Koonin exposed a completely different view in a very recent commentary (Koonin, 2009). He believes, in detriment of selection, that the principal role in evolution is played by non-adaptive processes. He exposes a series of evidences that are against the idea that natural selection is the principal force of evolution:

- The architecture of the genomes varies deeply among species, and this seems to be due to random processes.
- A trend for an increase in complexity could not be observed, thus, it has been hypothesized that the episodes of complexity increase are due to weak purifying selection and not to adaptation ('genomic syndrome' hypothesis).
- The universality of some characteristics of genome evolution could be simply explained by non-selective models.

Michael Lynch has a similar view to Koonin, in fact, he is the responsible of the 'genomic syndrome' hypothesis. He reasoned that genetic changes involving an increase of complexity, such as gene duplication and intron insertion, are slightly deleterious, and therefore, could only be fixed in populations with a small rate of purifying selection. The rate of purifying selection is proportional to the population size, being smaller in populations with small effective sizes or in populations that are suffering a bottleneck. Thus, complexity is acquired as a consequence of an ineffectiveness of the purifying selection, and this is why he coined the term 'genomic syndrome'. Later, these complex features are co-opted for biological functions, and then they are subject to the action of selection. Lynch thinks that the main features of genomes are modelled by non-adaptive evolution and depend on purifying selection, and consequently on the effective population size and mutation rate of the population (Lynch, 2007).

Hans Ellegren does not adopt any of the two positions (Ellegren, 2008). He emphasizes that both theories have data supporting them. The nearly neutral theory is strongly supported by the fact that species with smaller population sizes have higher $d_N/d_S$ values than populations with larger sizes, suggesting the existence of slightly deleterious mutations. However, there is plenty of data suggesting that adaptive evolution has played an important role in 10-40% of all genes in different lineages. He concludes that both selection and genetic drift have a central role in molecular evolution.

Justin Fay (Fay, 2011) argues that contradictory results have been found regarding the evidence of adaptation at the molecular level and he suggests that the estimations done for adaptive substitutions are unreliable. MK tests have been applied in several genomes obtaining disparate results. In a bunch of genomes high levels of adaptation have been reported, questioning the validity of neutral theory. However, in several others, including humans, low levels of adaptations have been found. It could be thought that the effective population size may play an important role in the explanation of those differences, but there are species with large population size, such as yeast and bacteria, for which has not been found a strong signal of selection. He poses that the used models assume independence among sites, however, he emphasizes two examples in which non-independence can produce an overestimation in the positive selection estimation: epistasis and hitchhiking. When epistasis is occurring, the fitness effects at one site depend on the genotype of the other site, and therefore, selective constraints on that site depend on epistatic interactions. Although epistasis has been incorporated into some models, the MK test does not take it into account yet and is not known if it would have an impact on the test. On the other hand, if mutations are slightly deleterious they could be fixed because of the hitchhiking with linked advantageous mutations. Therefore, positive selection can end up with an increase in the deleterious substitution rate. It is not known if these phenomena could lead to an overestimation of the positive selection measured with the MK test. To sum up, he concludes that due to the fact that a consistent pattern is not found when estimating the adaptation among different genomes, other factors, especially non-independence of sites, should be considered when using the MK test before ruling out the neutral theory of molecular evolution.

Andreas Wagner (Wagner, 2008) has tried to reconcile neutralism and selectionism in a network-based hypothesis. He poses that there are several lines of evidence against neutralism, such as the high number of adaptive fixations reported in *Drosophila* or the presence of 'hitchhiking'. But in the other hand he argues that data from protein evolution and molecular engineering highlights the importance of neutral mutations. In fact, there are several examples in enzymes in which neutral mutations facilitate adaptation. Additionally, more robust proteins have a higher functional diversity. For this reason he tries to reconcile all the evidences in a single hypothesis.

Wagner defines neutral networks as a group of connected genotypes that share a phenotype. Those phenotypes that could be adopted by a higher number of sequences would be more robust to mutations. Neutral networks are the base for understanding how Wagner reconciles neutralism and selectionism. He suggests that a genotype can randomly neutrally mutate inside the neutral network of a specific phenotype. But, a beneficial mutation that produces a better phenotype can occur. Once in the new phenotype, a new cycle of neutral mutations could be produced in this new neutral network (figure 1.3). It is important to highlight that the effect of a mutation depends on the context of the previous mutation; the neutrality of mutations can vary depending on the order of occurrence (i.e. mutations in RNA secondary structures). If this model explains how evolution takes place, then, three predictions should be observed: 1) cycles of neutral diversity expansion and selective diversity contraction; 2) pervasive epistatic interactions among mutations; 3) residues can change, they can first evolve neutrally and later be subject to positive selection.

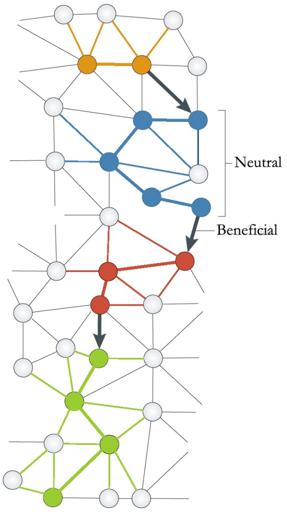


**Figure 1.3:** Neutral network representing cycles of neutral and beneficial mutations. Modified from Wagner (2008)

Thus, he proposes that neutral mutations should be defined as *'A neutral mutation does not change one aspect of a biological system's function in a specific environment and genetic background'*. In the previous definition, neutral mutations were proposed to never affect fitness, however, in this modern vision of neutral mutations the interaction with the environment and with other genes plays an essential role and a mutation that was considered neutral could cause phenotypic effects if the genetic or the environment background are changed. For this reason, neutral mutations can be relevant for innovation and evolvability (Wagner, 2005).

## 1.3 Protein evolution

### 1.3.1 Determinants of protein evolution

It is widely known that there are proteins in genome that evolve very fast, while there are others that evolve very slowly. For example, histones, which are responsible for DNA packaging, are highly conserved and almost do not present changes when comparisons between species are done. On the contrary, proteins related with the major histocompatibility complex, which plays an important role in the immune response, are evolving very fast in order to adapt to the changing adverse conditions. In yeast, evolutionary rates (measured as $d_N$) vary around 1,000-fold between the slowest and fastest evolving proteins (Drummond et al., 2005). Which are the determinants of protein evolution has been debated for several years.

This debate started five decades ago, when it was hypothesized that the evolution of a protein was determined by its level of functional constraint (Ingram, 1961) and its importance for the organism (Wilson et al., 1977). Some years afterwards, Zuckerkandl (Zuckerkandl, 1976) proposed that the evolutionary rate is determined by the proportion of sites in the protein that are involved in specific functions, he named this property functional density. The advent of the genome era has allowed to the scientific community to study in more detail the determinants of protein evolution. Several determinants have been proposed to date such as protein dispensability (measured as the fitness effect of knocking out the gene) (Hirsh and Fraser, 2001), number of protein molecules per cell (Drummond et al., 2006), gene expression (measured as the number of mRNA molecules per cell and the codon adaptation index) (Green et al., 1993; Pál et al., 2001; Wall et al., 2005), protein-protein interactions (Fraser et al., 2002), number of microRNA types and disordered content (Chen et al., 2011), sequence length (Marais and Duret, 2001; Lipman et al., 2002), central role in the interaction network (Hahn and Kern, 2005), age of protein's origin (Albà and Castresana, 2005; Wolf et al., 2009), protein structure, solvent accessibility and pairwise interactions among amino acids (Choi et al., 2007) and folding robustness (Lobkovsky et al., 2010).

Some of these determinants are correlated with each other and Drummond and colleagues (Drummond et al., 2006) made an effort to disentangle them. They did a combined analysis in yeast using 7 predictors that have been previously associated with protein's evolution (dispensability, gene expression level, protein abundance, codon adaptation index, number of protein-protein interactions, gene length and gene's centrality in the interaction network). They used a principal component regression (PCR) analysis and found that a single component accounted for almost half of the variability. This component included gene expression level, codon adaptation index and protein abundance, which are all linked to the number of translational events. Therefore, they suggested that a single determinant is dominating protein's evolutionary rate. But Plotkin and Fraser (Plotkin and Fraser, 2007) claimed that the predictors used presented different levels of noise, and this could be confounding the PCR analysis. When they equalized the noise for each of the predictors they found that there was no evidence to think that protein evolution is determined by only one determinant. Their results showed that many factors such as expression level, protein-protein interaction and gene dispensability may influence evolutionary rates independently. Afterwards, Drummond and Wilke (Drummond and Wilke, 2008) observed covariation between sequence evolution, mRNA level and codon usage in human, mouse, fly, worm, yeast and *E.coli*. Besides, in several studies, genes with high mRNA expression levels have been shown to produce slow evolving proteins in a wide range of taxa (from bacteria to humans). They hypothesize that these observations are indicative of the existence of a selection for structural robustness against mistranslation. Ribosome errors produce misfolded proteins, and those misfolded proteins are very toxic for the cell because can aggregate with other misfolded proteins and form protein-protein aggregations. Misfolded proteins corresponding to highly expressed genes could be very deleterious for a cell, and for this reason its understandable the strong selection to avoid their accumulation Mutations in highly expressed genes that encode for robust proteins may originate less robust proteins and therefore, they are not going to be selected leading to a slow accumulation of changes over time. They have named this hypothesis the Mistranslation-Induced Protein Misfolding (MIM) (Drummond and Wilke, 2008). Studies in *Drosophila* pointed out the importance of translational selection in determining protein evolution, but they found other contributing factors such as tissue bias in expression, gene essentiality, intron number and recombination rate (Larracuente et al., 2008). In a study performed in yeast, the authors have used an integrated probabilistic modelling approach to decipher the correlations between predictors. They found that slowly evolving proteins tend to have lower predicted structural disorder, are involved in specific biological functions (like translation), regulated by a higher number of transcription factors, more abundant, more essential, biased in amino acid composition and enriched for interaction partners. These results agree with previous studies (Xia et al., 2009).

Despite of all these findings, the causes and consequences of the difference in evolutionary rates observed among proteins are still under debate. However, now that we have more hints, gene dispensability is assumed to influence less than expected, protein structure and stability seem to play an important role, and expression level seems to be the strongest determinant of protein evolution found (Pál et al., 2006). Further research in this field is still necessary to resolve one of the main questions in evolution.

### Age as a determinant of protein evolution

Several authors have reported an inverse relationship between protein's age and evolutionary rate, with younger proteins evolving much faster than older ones (Domazet-Loso and Tautz, 2003; Daubin and Ochman, 2004; Subramanian and Kumar, 2004; Albà and Castresana, 2005; Wang et al., 2005; Cai et al., 2006; Luz et al., 2006; Albà and Castresana, 2007; Zhang et al., 2007; Kuo and Kissinger, 2008; Kasuga et al., 2009; Toll-Riera et al., 2009a; Wolf et al., 2009; Cai and Petrov, 2010). This correlation is not species-specific, it has rather been shown to be widely distributed among species. The interrelation between age and protein evolution adds another possible determinat for protein evolutionary rate: age of protein's origin.

The first suggestion of a possible relationship between a protein's age and its rate of evolution dates from 1986, when Doolittle stated that *'some of the most ancient proteins are changing very slowly'* (Doolittle et al., 1986). Several years after, it was observed that orphan genes (genes that do not have recognizable homologues) in *Drosophila* (Domazet-Loso and Tautz, 2003) and bacteria (Daubin and Ochman, 2004) were evolving much faster than nonorphan genes. However, the first authors to report an inverse relationship between gene age and evolutionary rate were Albà and Castresana in 2005 (Albà and Castresana, 2005). In their article they classified human-mouse orthologous pairs in 4 age groups: tetrapods, deuterostomes, metazoans and old according to the range of species in which they found a homolog using BlastP searches (Altschul et al., 1997). They noticed that proteins were evolving differently between groups; the rate of evolution was diminishing with the increase in protein's age. Proteins classified as tetrapods were evolving as much as 4 times faster than proteins classified as old (figure 1.4).

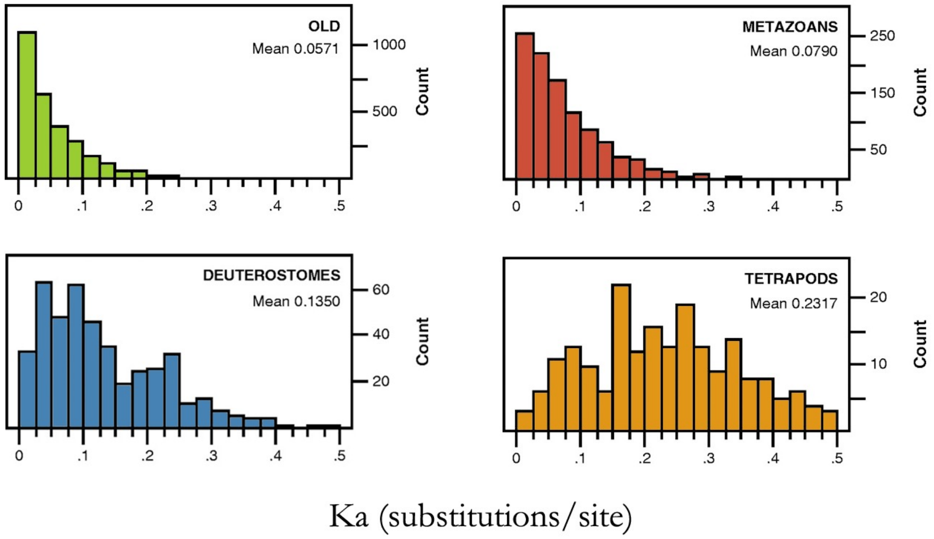


**Figure 1.4:** Distribution of non-synonymous substitutions rates among the 4 age groups. Modified from Albà and Castresana (2005)

They formulated two possible models to explain the observed relationship. In the first model, 'constant constraint model', they hypothesized that the constraints are constant and characteristic for each gene and that the inverse relationship is observed because the youngest genes may be involved in functions that are less evolutionary constrained. The second model, 'increasing constraint model', hypothesizes that when proteins arise they have few selective constraints acting on them, but with time they acquire, gradually, a higher number of functional and structural important sites. Consequently, a higher fraction of their sites are constrained and do not accept mutations, lowering the evolutionary rate.

In a very recent article, Vishnoi and colleagues (Vishnoi et al., 2010), studied evolutionary rates and protein's age among three taxonomic groups: mammals, *Drosophila* and yeast. They found that younger proteins have more variable evolutionary rates values compared to older ones and proposed to extend the 'increasing constraint model' to include a period of high variability in protein's evolutionary rates after their birth.

Additionally, in another recent paper, human-chimpanzee orthologous proteins were classified into ages to decipher which are the forces acting on the observed high rate of protein evolution of lineage-specific genes. The authors found that older genes, compared to younger ones, had less frequent and fewer nonysnynoymous single-nucleotide polymorphisms (SNPs), suggesting that younger genes are under a weaker purifying selection than older genes (Cai and Petrov, 2010).

### Protein structure as a determinant

Protein's tridimensional structure is a key requirement for protein's function. The core of a protein is formed by buried residues and is determinant for the stability of the folded structure. Buried residues are involved in intramolecular interactions and in maintaining protein structure (Franzosa and Xia, 2008). Most mutations occurring in a protein destabilize the structure, and is for this reason that structural constraints could be affecting protein's evolutionary rate and explain the variation observed in evolutionary rate between different proteins (Pál et al., 2006). Structure could be itself a determinant or it could act through other mechanisms, for example, it could play an important role in the selection for structural robustness against mistranslation, which has been suggested to be a key determinant for protein evolution (Franzosa and Xia, 2008).

Several structural descriptors have been studied such as designability, solvent accessibility and protein stability; here I will try to make a short review of the most recent articles.

Highly designable structures are those ones which can be encoded by several sequences; therefore, they can tolerate mutations better. For this reason Bloom and colleagues (Bloom et al., 2006a) hypothesized that their sequences should evolve faster (figure 1.5). They used contact density as a measure of designability because it has been suggested that proteins with higher contact density (and, consequently, higher fraction of buried residues) are more designable (England and Shakhnovich, 2003). They examined the relationship between evolutionary rates and several structural descriptors controlling by gene expression level. The authors found that exposed residues evolve much faster than buried ones, reinforcing the importance of solvent accessibility on the evolution of individual residues. Then, proteins with a higher fraction of buried residues should evolve slower. They tested it and found contact density and fraction of buried sites to be correlated with evolutionary rates measured with $d_N$. They explained this contradictory result as follows: proteins which have a higher fraction of buried residues are more designable and stable, thus, their exposed regions can mutate more freely without destabilizing the protein. They also reported that secondary structure composition does not affect the evolutionary rates. Using a principal component regression analysis they found that structural descriptors could explain between 10 and 12% of the evolutionary rate variation, being protein length and contact density the most important structural descriptors. Therefore, protein structure only accounts for approximately 10% of the evolutionary rate variation, while expression level explained 34%, concluding that expression level is the main determinant, but protein structure also contributes.

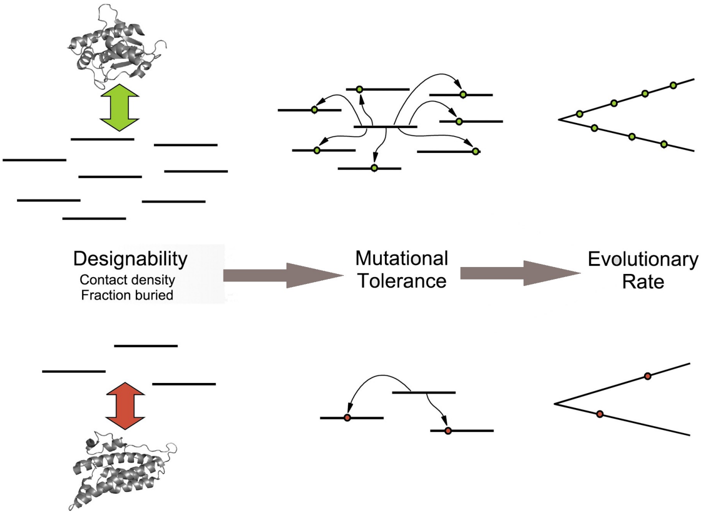


**Figure 1.5:** Relationship between protein's designability and evolutionary rate. Modified from Bloom et al. (2006a)

Lin and colleagues (Lin et al., 2007) used residues solvent accessibility to explain protein evolutionary rate. They claimed that in the study performed by Bloom et al. (Bloom et al., 2006a) there was a bias against disordered proteins. To cover a broader range of proteins, and not only the ones with a PDB structure, they used PDB homologues for the unsolved proteins. They predicted the proportion of exposed residues directly from the amino acid sequences using support vector machine. They found that in those cases with a high alignment length between the sequence and its PDB homologue the fraction of exposed residues was negatively correlated with evolutionary rate, as in Bloom et al (Bloom et al., 2006a). On contrary, when the alignment length was lower and proteins with disordered regions were included, the fraction of exposed residues was positively correlated with the evolutionary rate. Therefore, they suggested that only in well-folded structures the correlation between evolutionary rate and designability can be detected. They proposed that the proportion of solvent-exposed residues in a whole protein is the most important determinant of protein evolution after translational selection. Franzosa and Xia (Franzosa and Xia,

2009) also reported that exposed residues are evolving much faster than buried ones, which are more conserved and have been related with protein stability. In fact, they found a linear relationship between evolutionary rate and solvent exposure. They also claimed the importance of the location of the buried residue, if it is located in a protein-protein interface, it will evolve slower than if not. When they compared evolutionary rates of residues classified in different solvent accessibilities between large-core proteins (higher fraction of buried residues) and small-core proteins (smaller fraction of buried residues) they observed that more accessible residues from large-core proteins were evolving faster than the ones from small-core proteins. Solvent-excluded residues were still constrained implying that they are important for protein stability. This results demonstrate that Bloom and colleagues were correct when they hypothesized that exposed residues belonging to proteins with a very stable core are more prone to mutate explaining the higher evolutionary rates in more designable proteins. In another study, the authors found that the number of intra-protein residue interactions is negatively correlated with amino acid substitution rates. This could be explained by the fact that a mutation in one of these residues will also affect all the residues with which it interacts (Toft and Fares, 2010).

Evolvability is the capacity of the protein to evolve, but most of the proteins have to be folded to function, thus there is selection acting to maintain protein's structure and to avoid destabilizing mutations (Tokuriki and Tawfik, 2009). An analysis of pathogenic mutations revealed that the deleterious effect of around 80% of these mutations is due to effects on protein folding and stability (Yue et al., 2005). Thus, stability can be very important to determine the rate of protein evolution (Tokuriki and Tawfik, 2009). Bloom and colleagues (Bloom et al., 2006a) studied the relationship between protein stability and evolvability and found that more stable proteins can better tolerate functionally beneficial mutations that destabilize the structure, and therefore, the most stable proteins are the ones that are more evolvable. Mutations increasing the stability of the proteins are neutral because they are not involved in protein function, but they are crucial to allow destabilizing mutations that are beneficial for the functionality of the protein. This phenomenon in which the combination of mutations is more beneficial than a unique mutation is known as positive epistasis (Kryazhimskiy et al., 2011).

Wolf and colleagues (Wolf et al., 2008) have used a protein domain approach to decipher the contribution of expression level, protein domain structure and function into protein evolutionary rate. They assumed that domains forming a protein are translated at the same rate, and therefore, under the hypothesis of the existence of a selection for structural robustness against mistranslation, domains inside a protein are expected to evolve at more similar rates than the same domains belonging to different protein. Despite of being translated at the same rate, they found that domains belonging to the same protein did not evolve at the same rates, and this is due to the differences existing in the intrinsic constraints between domains, although, homogenization of their evolutionary rates could be observed. Therefore, they concluded that the rate of protein sequence evolution depends on the intrinsic misfolding robustness (given by the structure, the stability and the designability of the domain) and on the selection to avoid misfolding.

Lastly, in a very recent article (Lobkovsky et al., 2010) the authors defined protein's fitness as a function of the number of misfolded proteins that are produced to get the needed protein abundance. They used lattice models to study protein folding. The results obtained from their model suggest that protein evolution is governed by the protein folding physics. Therefore purifying selection plays an important role against those mutations that produce misfolded proteins and selection for function, believed until recently to be key for protein evolution, plays a more secondary role.

### 1.3.2 Protein domains as the unit of evolution

#### Domain properties

Most proteins are formed by modules called domains. Domains are protein fragments found in several proteins, which have a function, an evolutionary history and can fold independently, thus, they can be considered the main units of evolution (Vogel et al., 2004a; Ekman et al., 2005; Itoh et al., 2007). Proteins can be formed by one domain (single-domain proteins) or by several domains (multi-domain proteins); their sequential order in a protein is known as domain architecture (Fong et al., 2007). Domains have an average length of 120 amino acids, hence, they can occupy a large fraction of protein's sequence length. Consequently, a change in the domain architecture can be very important at the protein level (Buljan and Bateman, 2009). In fact, domains contained in the protein and their interactions have been found to determine the function of the proteins (Vogel et al., 2004a). Single-domain proteins containing domains belonging to the same family have been estimated to have a similar function in 67% of the cases, whereas this number decreases to 35% when two-domain proteins share only one domain. Therefore, changes in the domain content may alter drastically protein's function. Domains are found in the three domains of life (archaea, bacteria and eukarya). However, it has been reported that multicellular eukaryotes have a higher number of domains per protein and a higher variety of domains architectures than unicellular eukaryotes and prokaryotes (Apic et al., 2001; Ekman et al., 2005; Itoh et al., 2007). Domains, and their combinations, have been claimed to have played an important role in the evolution of multicellularity and in the acquisition of the new functions that it implies (i.e. cell adhesion, cell differentiation, cell communication) (Patthy, 2003; Itoh et al., 2007; Buljan and Bateman, 2009; Buljan et al., 2010).

There are domains that are widespread across the three domains of life, while there are other that have appeared more recently and are only present, for example, in eukarya. In a very interesting study the authors observed that most of the human domains (around 60%) are shared across a wide range of species, indicating that they have an ancestral origin. Additionally, they observed that few domains have been originated at higher nodes of the tree, suggesting that proteins have mainly evolved by reusing and combining ancestral domains, rather than creating new ones (Pal and Guda, 2006). Thus, few domains have been originated in the chordate lineages, although, they have been related with key chordate novelties such as immune, nervous and skeletal system functions (Lander et al., 2001). The contrary pattern has been observed for domain combinations, although several domains are common across the three domains of life, as few as 5% of the two-domain combinations are shared (Apic et al., 2001; Vogel et al., 2004a,b). Therefore, the reuse of domains, rather than the creation of new ones speeds up the evolution of the cellular complexity (Moore et al., 2008).

Domains can be gained into proteins through various mechanisms: gene fusion, exon extension, exon rebombination, intron recombination and retroposition (Marsh and Teichmann, 2010) (figure 1.6).

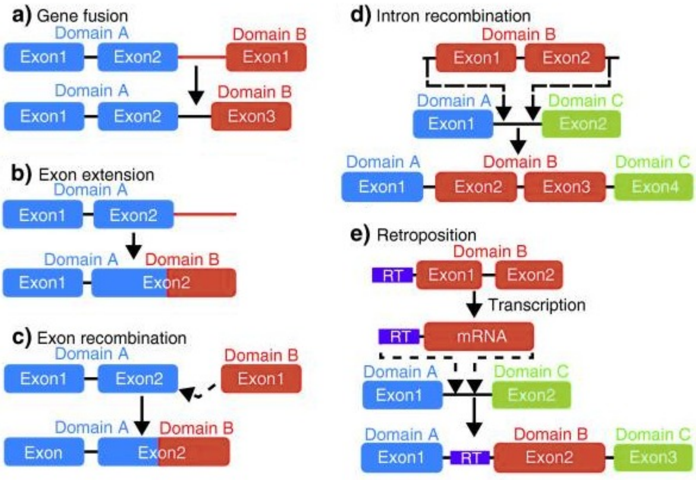


**Figure 1.6:** Mechanisms of protein domain gain. Modified from Marsh and Teichmann (2010)

It has been always thought that exon shuffling (or intronic recombination) plays an essential role in domain acquisition (Patthy, 2003; Vogel et al., 2004b; Vibranovski et al., 2005; Moore et al., 2008) but in a very recent study using animal proteins (Buljan et al., 2010) it has been reported that the most frequent process is domain acquisition through gene fusion, with a possible mediation of non-allelic homologous recombination. In fact, only 10% of the domain gains were surrounded by introns of symmetric phase and were located in the middle of the protein, leading to think that intronic recombination is not the main contributor to domain gain, although it could have played an important role in the evolution of some multi-domain proteins (Buljan et al., 2010). Multi-domain proteins evolve mainly by stepwise insertion of single domains at the protein termini (Buljan et al., 2010), except the cases of tandem domain duplications (Björklund et al., 2005). Protein terminus are flexible, charged and are located at the surface of proteins. Thus, it is less probable that the addition of a domain in here disrupts the structure of the protein (Buljan and Bateman, 2009).

### Domain classification

There are two types of databases for domain classification corresponding to two different domain definitions, the structural definition in which domains are defined as 'independently folding units' and the evolutionary based definition in which domains are 'independently evolving units'. Structural based databases use tri-dimensional information, grouping together those domains with similar tri-dimensional architectures (Müller et al., 2002). In contrast, evolutionary databases are based on the conservation of the primary sequence (Pal and Guda, 2006). The evolutionary based databases contain a higher number of domains than the structure based databases (Ekman et al., 2005; Pal and Guda, 2006). However, usually those two definitions coincide in their classification of domains into families (Ekman et al., 2005).

Structurally based databases include SCOP (Structural Classification of Proteins) (Murzin et al., 1995), CATH (Protein Structure Classification) (Orengo et al., 1997) and FSSP (Families of Structurally Similar Proteins) (Holm and Sander, 1994). The SCOP database (Murzin et al., 1995) is created by a combination of automated methods and manual inspection. It describes the structural and evolutionary relationships between proteins with known folds. The CATH database (Orengo et al., 1997) is also manually curated and provides a classification of protein domain structures. Structural domains are assigned into homologous superfamilies, which consist of families of domains that are evolutionarily related. The FSSP database (Holm and Sander, 1994) contains structural alignments for proteins present in the Protein Data Bank (PDB), which is a database that stores information about structures, nucleic acids and complex assemblies determined experimentally.

Evolutionary based databases include SMART (Simple Modular Architecture Research Tool) (Schultz, 1998), ProDom (Protein Domain Database) (Sonnhammer and Kahn, 1994) and Pfam (Protein Family Database) (Sonnhammer et al., 1997). The SMART database contains domain families involved in signalling, extracellular and associated with chromatin processes (Schultz, 1998). ProDom domain families are generated automatically from the Uniprot Knowledge Database (Corpet et al., 2000). The Pfam database is a collection of protein domains and families. Each family is represented by multiple sequence alignments and by profile-Hidden Markov Models (HMM). In Pfam-A the entries are manually curated, while in Pfam-B entries are generated automatically. Libraries containing the HMM profiles can be downloaded allowing the users to assign domains to their proteins (Sonnhammer et al., 1997).

## 1.4 Origin of new genes

### 1.4.1 Mechanisms of new gene formation

One of the subjects that are more appealing for scientists and has been intriguing them since long is how all the genes present in the genomes have been originated. As early as in the thirties, Haldane (Haldane, 1932) and Muller (Muller, 1935) proposed, for first time, gene duplication as a mechanism for new gene creation. Several other mechanisms have been proposed since then, such as retroposition, exaptation from mobile elements, lateral gene transfer, gene fusion/fission and *de novo* origination (Long et al., 2003). The most appropriate subset to study new gene formation is the one formed by young genes. Young genes are at their initial stages of evolution, thus, the details related with their origin are still present. These details are erased progressively over time; therefore, ancient genes contain little information about this aspect (Long et al., 2003). The study of new gene origination is also very interesting because new genes have been considered to be very important for adaptive evolutionary innovations.

In here I have centred my attention in protein coding genes, a review focusing on non-coding genes has been recently published by Kaessman (Kaessmann, 2010).

### Gene duplication

Gene duplication is considered the main mechanism acting in the creation of new genes and is also the best studied. Duplications could occur at different levels. They could be produced at a genome level, being then a whole genome duplication, an example is the duplication occurred at the base of vertebrates (Kasahara, 2007). If the whole genome duplication is retained it gives raise to polyploidy genomes (Van de Peer et al., 2009). Duplications could also be segmental, implying long stretches of DNA (>1Kb) with high similarity (90-98%) (Eichler, 2001). In fact, it has been reported that 5.2% of the human genome is covered by segmental duplications (Bailey et al., 2002). And, lastly, duplications could be affecting only a gene (figure 1.7), creating tandem or dispersed duplicates. The duplication can affect the totality of the sequence or it can rather affect only a fragment of the sequence, being then, a partial duplication (Katju and Lynch, 2006). The molecular mechanisms proposed to be involved in duplication are non-allelic homologous recombination, transposon-mediated transposition and illegitimate recombination. The two first mechanisms require the presence of homology (Zhou et al., 2008a).

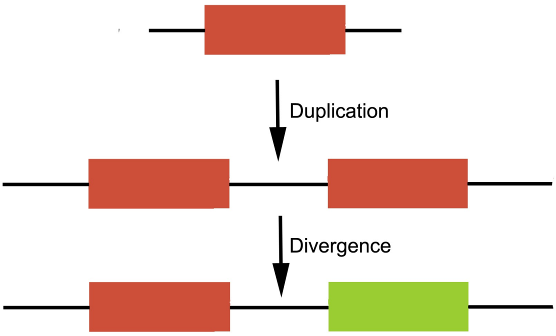


**Figure 1.7:** Schema representing the total duplication of a gene. Adapted from Long et al. (2003)

Gene duplication is a major source of evolutionary novelty. It has been associated with functional diversification, increased coding sequence evolutionary rates (Lynch and Conery, 2000; Scannell and Wolfe, 2008) and with a higher tissue expression divergence (Gu et al., 2002; Makova and Li, 2003; Farré and Albà, 2010). There are several examples of gene families that have been expanded through gene duplication, for example olfactory receptors in mouse (Waterston et al., 2002) and humans (Gilad et al., 2005) and KRAB-associated zinc-finger in primat such as the acquisition of the trichromatic color vision. The three colour pigments, green, red and blue are encoded by three different genes, one autosomal and two located on the chromosome X (red and green). It has been hypothesized that red and green pigments have been originated by a duplication occurring in Old World Monkeys after their divergence from New World Monkeys (Nathans et al., 1986).

Duplicated genes arise at very high rate, on average, 0.01 duplicates arise per gene per million years. After the duplication the duplicated copy could have different fates. The most frequent fate is the silencing of the copy after approximately 4 million years due to the accumulation of degenerative mutations (Lynch and Conery, 2000). However, sometimes the two copies survive. The duplicated copy can acquire beneficial mutations and consequently gain a novel function with respect to the parental gene (neofunctionalization), while the parental preserves its original function (Ohno, 1970). The duplicated copy could also be retained due to the split of the original function between the two gene copies (subfuctionalization) (Hughes, 1994). Finally, if an increase of dosage of a particular gene is beneficial, the duplication may be fixed by positive selection maintaining the same gene structure and function than the parental gene (Kondrashov et al., 2002).

Anyhow, duplicated genes are not always identical to their parental gene. They could have arisen through a partial duplication instead of a complete duplication or they could recruit additional sequences and form a chimeric structure. Examples of chimeric structures have been reported in *Caenorhabditis elegans* (Katju and Lynch, 2003, 2006) and *Drosophila* (Zhou et al., 2008a; Chen et al., 2010). Surprisingly, only 40% of the new duplicates in *Caenorhabditis elegans* arose from complete gene duplications (Katju and Lynch, 2003). Partially duplicated genes and chimeric genes are expected to adopt immediately a new function respect to the parental gene. Consequently, would have more chances to be preserved in comparison with complete duplicated genes which are redundant to their parental, having a higher probability to accumulate deleterious mutations(Patthy, 1999; Zhou et al., 2008a). An example of partially duplicated chimeric gene is the *Hun* gene in *Drosophila*, which is located on chromosome X. *Hun* has duplicated partially from *Bällchen* gene, which is in chromosome 3R. *Hun* lacks 3′ coding sequence with respect to *Bällchen*, but has gained 33 amino acid from a nearby intergenic sequence. *Bällchen* is expressed ubiquitously, while *Hun* has testes-specific expression (Arguello et al., 2006).

It is not well know how the duplicated copy acquires the new function, it could be due to the action of positive selection or to the fixation of neutral mutations. A nice example of positive selection acting on a ribonuclease duplicated gene is the one reported by Zhang et al. (Zhang et al., 2002). Douc langurs (*Pygathrix nemaeus*) are a type of monkeys that eat leaves instead of fruits and insects. The leaves are fermented by symbiotic bacteria in the foregut. Pancreatic ribonucelase 1B (RNASE1B) was originated around 4.2 million years ago by a duplication of RNASE1, which is an enzyme used to digest the bacteria present in the small intestine. The evolutionary rates of the RNASE1B are very high (Ka/Ks ratio= 4.03), on the contrary, the paralagous copy, RNASE1, does not show changes. Moreover, most of the substitutions imply the gain of negative charge, which leads to the reduction of the optimal pH for RNASE1B. This could be related with an increase of the digestive system efficiency because the pH in the small intestine of douc langurs is low. The entire process could be explained by a first phase of reduced selective constraints just after the duplication, in which changes are incorporated. In a second phase these mutations change the function of the duplicated copy, which is totally established after a burst of positive selection.

### Retroposition

Retroposition is a RNA-based duplication. Copies are created through the reverse transcription of a mature messenger RNA (mRNA) from a parental gene and the DNA copy is lately inserted into the genome (figure 1.8). Due to the mechanism of reverse transcription the retrocopies do not contain introns. Additionally, the copied fragment is usually lacking the promoter region, and for this reason most of the copies are pseudogenized. However, it has been observed that some retrocopies could adopt regulatory sequences that are nearby (Long et al., 2003; Kaessmann, 2010). In comparison with DNA duplicates, retrogenes frequently show new expression patterns, new genomic locations and new functions (Kaessmann et al., 2009). Several recent human genes have been shown to be originated from retroposition, most of them are expressed in testis and are involved in spermatogenesis (Marques et al., 2005).

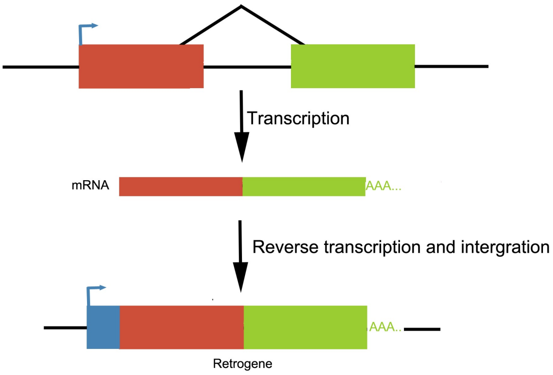


**Figure 1.8:** A retrocopy arises from the reverse transcription and integration of an mRNA from a parental gene. The new retrogen should acquire a TSS and a promoter to be functional. Adapted from Kaessmann (2010)

## Exaptation from mobile elements

Mobile elements are DNA or RNA sequences that can be inserted in other regions of the genome (figure 1.9). There are several types of mobile elements, such as SINE, LINE and LTR. Two mouse genes have been described to have the major part of their sequence overlapped by transposable elements (Nekrutenko and Li, 2001). Moreover, several primate orphan genes have a significant fraction of their sequence covered by transposable elements, mostly Alu elements, which are a type of SINE found only in primates (Toll-Riera et al., 2009b). In fact, it has been shown that the integration of mobile elements to generate novel functions is common among nuclear genes (Nekrutenko and Li, 2001).

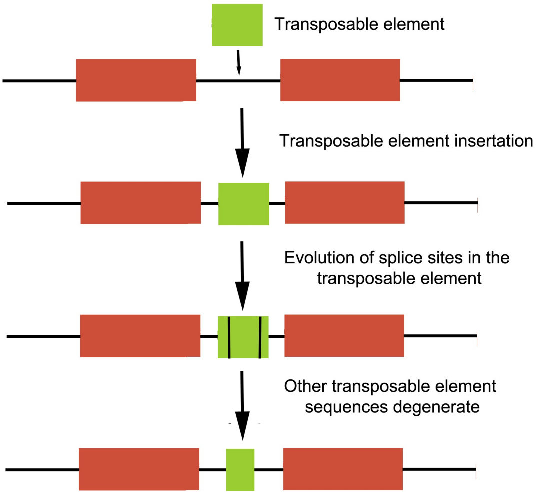

Transposable element
Transposable element insertation
Evolution of splice sites in the transposable element
Other transposable element sequences degenerate

**Figure 1.9:** Mechanism of exaptation from transposable elements. Adapted from Long et al. (2003)

### *De novo* origination

*De novo* origination implies the birth of a new gene from noncoding genomic regions (introns, intergenic regions, untranslated 5′ or 3′ region). This mechanism was thought to be rare (Long et al., 2003), but recent studies have reported examples in several species such as primates (Toll-Riera et al., 2009a), mouse (Heinen et al., 2009), *Drosophila* (Levine et al., 2006; Zhou et al., 2008a) and *Saccharomyces cerevisiae* (Cai et al., 2008). Recent studies have revealed that almost all the genome is found in primary transcripts (Birney et al., 2007), and this finding could be the key to explain how genes are originated from scratch. Inside these primary transcripts, short ORFs could be present, and if they acquire nearby regulatory regions and are translated into peptides could give rise to a new function. If this new function is advantageous, the new gene would be fixed in the population (figure 1.10).

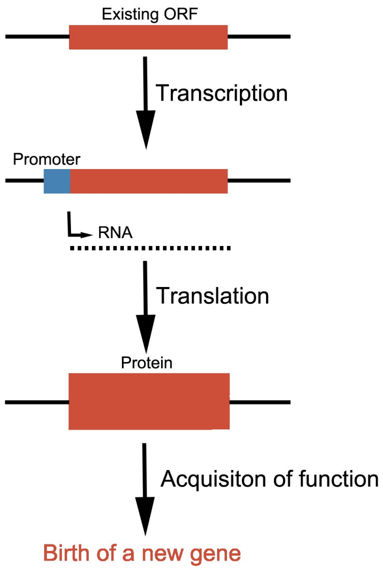


**Figure 1.10:** Diagram representing how a new gene is originated *de novo*

### Horizontal gene transfer

Horizontal gene transfer (HGT) is the process that leads to the transfer of genetic material between two unrelated species (figure 1.11). HGT is more frequent among prokaryote species and it has played a very important role in the evolution of some of their particularities (Boucher et al., 2003). However, there are cases of described HGT between endosymbiotic bacteria and their multicellular eukaryotes hosts, such as gene transference from *Wolbachia pipientis* to four insects and to four nematode species (Hotopp et al., 2007).

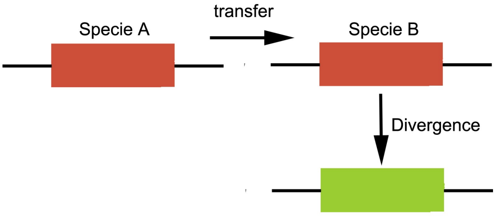


**Figure 1.11:** Schema representing the horizontal gene transfer between species. Adapted from Long et al. (2003)

### Gene fusion/fission

Two genes that are side by side in the genome can be fused into one (figure 1.12). This mechanism has been observed mainly in prokaryotes, however, some examples of fused genes could be found in the human lineage, such as the *KUA-UEV* gene (Thomson et al., 2000). Fission is the contrary process, one gene is split into two genes (figure 1.12) (Long et al., 2003).

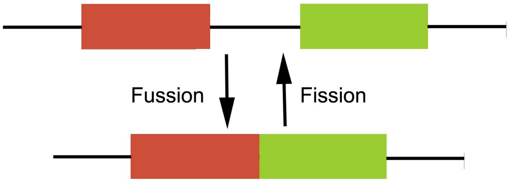


**Figure 1.12:** Diagram representing the gene fusion/fission mechanism. Adapted from Long et al. (2003)

It has to be taken into account that another possible mechanism of new gene formation is the one that involves several of the mechanisms mentioned above. One well-known example is the *jingwei* gene, found in the African *Drosophila* species. This gene was originated through a combination of duplication and retroposition of two different genes (Long et al., 2003).

Zhou and colleagues (Zhou et al., 2008a), in a recent study, have made an effort to assess the magnitude of the contribution of each of these mechanism of origin in *Drosophila*. They have studied young genes in the *Drosophila melanogaster* species subgroup. They found that the most common mechanism was gene duplication, being tandem duplication responsible for the origin of most of the duplicates limited to single species, whereas dispersed duplication was more common in the origin of genes shared by multiple species. Surprisingly, as many as 11.9% of the new genes were originated from noncoding sequences, and 10% through retroposition.

In a very recent article, Capra and colleagues (Capra et al., 2010) studied an interesting aspect of new gene origination that had not yet been tackled in detail, they compared several aspects in genes originated by duplication and in genes not originated by duplication (named novel genes). The evolutionary pressures over genes originated by duplication and over novel genes should be different due to the fact that, on contrary to novel genes, duplicated genes arise functionally and structurally well-formed. They showed that although duplicated genes are initially more integrated into cellular networks, both types of new genes gain function and interactions with time, but novel genes do it more rapidly than duplicated genes. Additionally, novel genes were found to increase in length by the incorporation of transposable elements or surrounding sequences. This increase in length could be related with the rapid gain of function and interactions experienced by novel genes. Duplicated genes were more centrally located in the network and related with environment interaction, whereas novel genes did not show any bias. Strikingly, they found that genes tended to interact with genes similar in age and mechanism of origin. Thus, the type of mechanism of origin seems to play an important role in the gene's subsequent evolution.

### 1.4.2 Lineage-specific genes

The number of genes contained in a genome is different among the sequenced genomes, indicating that mechanisms of gene birth and loss are frequent (Long et al., 2003). The sequencing of several genomes revealed the presence of a set of genes with no homologs in other genomes; these genes have been named 'orphan' genes. The fraction of genes classified as orphan is not insignificant, they represent a substantial fraction of every genome: around 14% of the genes in 60 fully sequenced microbial genomes (Siew and Fischer, 2003) and between 20 and 29% in *Drosophila* (Domazet-Loso and Tautz, 2003; Clark et al., 2007). Orphan genes are an extreme subgroup of lineage-specific genes, which are genes that are only present in some nodes of a phylogeny. Lineage-specific genes and orphan genes have been studied in several species, such as primates (Toll-Riera et al., 2009a), *Drosophila* (Domazet-Loso and Tautz, 2003), insects (Zhang et al., 2007), apicomplexan parasites (Kuo and Kissinger, 2008), ascomycotan fungi (figure 1.13) (Cai et al., 2006), rice (Guo et al., 2007) and in bacterial and archaeal genomes (Wilson et al., 2007).

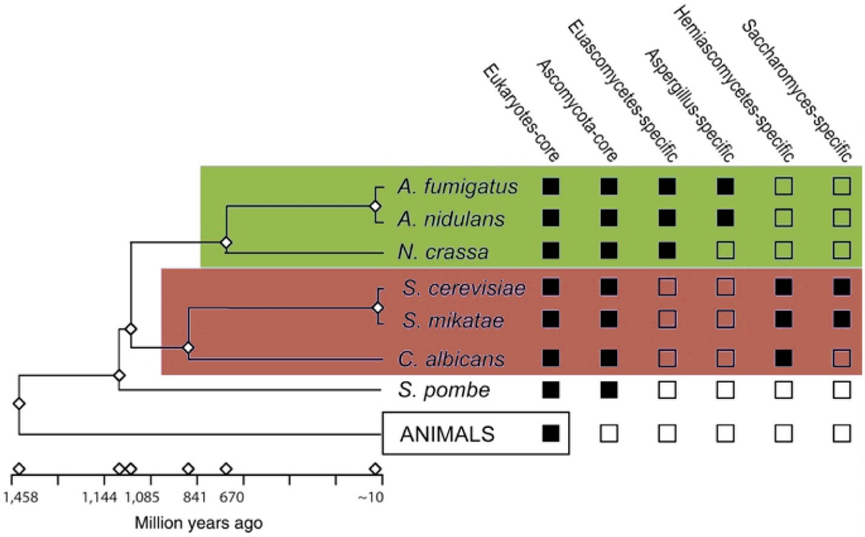


**Figure 1.13:** Classification in several lineage-specific groups in ascomycotan fungi. Modified from Cai et al. (2006)

Lineage-specific and orphan genes are likely involved in important species-specific adaptive processes and their study could contribute to unravel key recent adaptive processes. They have been related with important physiological adaptations, such as vomeronasal receptors and casein milk proteins. The sequencing of the chicken genome revealed the absence of the genes encoding for vomeronasal receptors and casein milk proteins, indicating that the evolution of the vomeronasal organ and the mammary glands took place in the mammalian clade (International Chicken Genome Sequencing Consortium, 2004). Additionally, several lineage-specific genes have been found to be related with defence against pathogens, such as dermcidin in primates (Toll-Riera et al., 2009a) and surface antigens in apicomplexan parasites (Kuo and Kissinger, 2008). Insect-specific proteins have been implicated with a role in communication and adaptation to the environment, stress and immune response (Zhang et al., 2007). Interestingly, it has been noticed in rice that more orphan genes are expressed under environmental pressure (injury and hormone treatment) than non-orphan genes, indicating that the first ones are much more evolutionary flexible (Guo et al., 2007).

However, lineage-specific genes are usually poorly annotated and for most of them the function is not well known (Kuo and Kissinger, 2008). The fact that they are very recent and that until their birth the organism has managed to live without them lead to the scientists to suppose that they are involved in non-essential and secondary functions. In spite of this, a surprisingly high fraction of them are essential. In a very recent and interesting paper Chen and colleagues (Chen et al., 2010) reported that new young genes in *Drosophila* became essential after a short period of time. In their study, the authors first identified 566 new young genes in *Drosophila melanogaster* (less than 34 million years old) and then designed RNA interference lines to knockdown them. Outstandingly, they found that 30% of these young genes were lethal, *Drosophila* could not survive without them. To have a positive control they chose randomly a similar number of old genes and they obtained that 35% of them were essential, being this number not statistically different from the fraction of essential genes found among young genes. Additionally, they found that these young genes were mainly originated by duplication and that they had higher evolutionary rates than their parental gene, indicating the action of positive selection or a relaxation of the functional constraints. They hypothesized that new genes are integrated into existing pathways and due to the action of mutation and selection they are optimized, quickly becoming essential for the viability of the organism.

It has been reported that in primates (Toll-Riera et al., 2009a), mammals (Albà and Castresana, 2005), yeast (Cai et al., 2006), *Drosophila* (Domazet-Loso and Tautz, 2003), *Escherichia coli* (Daubin and Ochman, 2004) and rice (Guo et al., 2007) orphan genes evolve faster and are shorter than non-orphan genes. The high evolutionary rate found in *Drosophila* orphan genes has shed light into a possible mechanism for their creation. Domazet-Loso and Tautz (Domazet-Loso and Tautz, 2003) hypothesized that gene duplication is the first step in orphan gene's creation (figure 1.14). The duplicated copy can follow two possible paths, it can be pseudogenized and subsequently lost or can be kept with a redundant or a related function. Due to the fact that the ancestral function is performed by the parental gene, the duplicated gene is not under selective pressure and consequently is able to evolve quickly, loosing its similarity with the parental gene. Lately, the duplicated gene might be recruited into a new pathway, and firstly, it would evolve again under fast adaptive evolution and once it has achieved the new function would start to evolve slowly in order to retain the newly acquired function. Due to these episodes of fast evolution, the similarity between the duplicated copy and the parental gene could not be longer detected, explaining why those genes have been classified as orphans. Of course, orphan genes could have been originated by other mechanisms that do not imply significant homology with other genes, such as exaptation from transposable elements and *de novo* origination (Toll-Riera et al., 2009b).

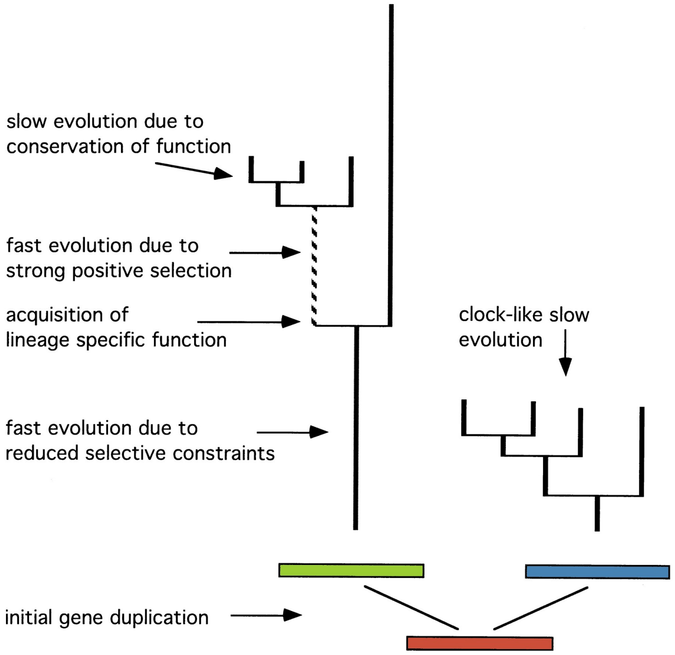

slow evolution due to conservation of function
fast evolution due to strong positive selection
acquisition of lineage specific function
fast evolution due to reduced selective constraints
initial gene duplication
clock-like slow evolution

**Figure 1.14:** Model for orphan gene evolution by gene duplication. Modified from Domazet-Loso and Tautz (2003)

To identify orphan genes one should use comparative genomics. Although orphan genes are easy to define as genes that do not have homologues in other genomes, they are hard to identify. The genes identified as orphan depend on the detection method and on the set of genomes used. The methods to detect orphan genes relies on similarity searches, and the preferred one is BLAST, which has been shown to pick up most of the remote homologues (Tautz and Domazet-Lošo, 2011).

# 2

# Methods

## 2.1 Genomes election

In the articles included in this thesis we have mainly used mammalian genomes, specially *Homo sapiens, Macaca mulatta, Mus musculus, Rattus norvergicus, Canis familiaris* and *Bos taurus* (figure 2.1).

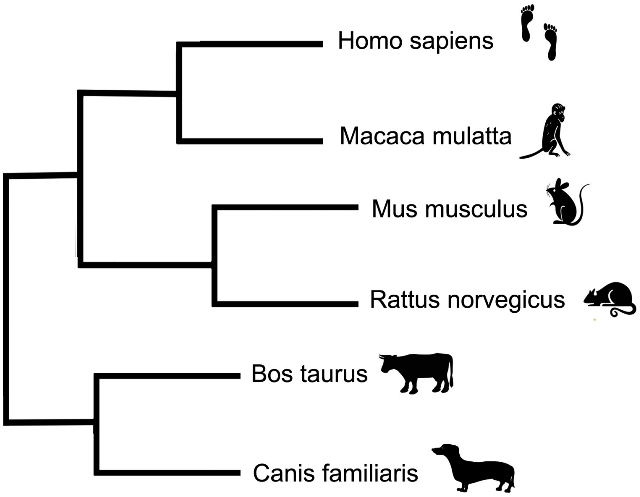


**Figure 2.1:** Phylogeny of the species used in this thesis

We have chosen those species because they have an optimal divergence time to calculate the evolutionary rates at DNA level. The relationship between sequence differences and time since species divergence is not linear (figure 2.2) because multiple substitutions can occur at the same site with the increase on time. It is said that sequences become saturated with time. If the species are too close (i.e. human and chimpanzee), the number of nonsynonymous substitutions that have occurred since their divergence is very small. Consequently, the errors in substitution rate estimates can be very high, as a small change in the number of inferred substitutions will lead to a huge difference in the estimated substitution rate. Another inconvenient of using very closely related species is that some observed differences could be due to polymorphisms instead of substitutions derived from mutations. On the contrary, if species are too distant sites will be saturated because most of the changes had already occurred and, thus, the new substitution overwrites the old one. Hence, the number of substitutions would be underestimated.

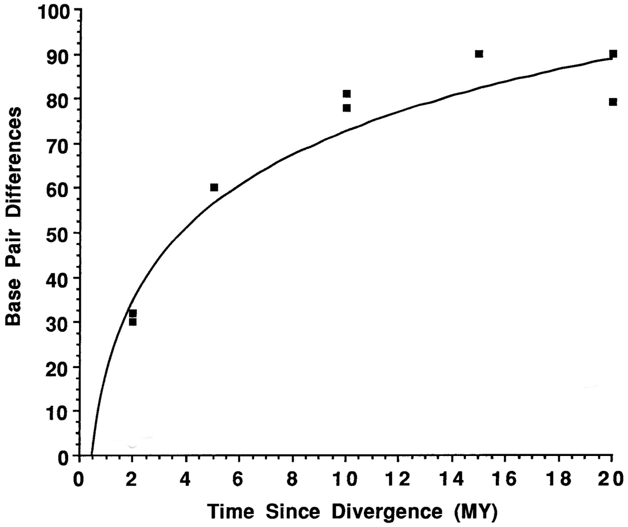


**Figure 2.2:** Number of nucleotide substitutions between pairs of COII mitochondrial gene in bovids against the estimated time of divergence in MY. Modified from Janecek et al. (1996)

## 2.2 Orthologs obtention

In all the studies presented in this thesis we obtained the orthologs defined in Ensembl (Flicek et al., 2011), using the Biomart tool. We used one-to-one orthologs, therefore we discarded all the paralogous genes.

The ideal scenario to calculate orthology would be to perform all the possible combinations between all the transcripts of all genes present in the sequenced genomes and then choose the most similar ones. However, this is computationally impossible, and approximations have to be used. In Ensembl orthologs are defined using similarity searches of the longest available transcript per gene (Vilella et al., 2009). Despite this approximation generally works correctly, there are some cases that can lead to erroneous assignments. For example, if a gene in humans has 5 transcripts and in cow it only has 1 transcript, the ortholog pair would be the longest transcript from human and the transcript from cow. However, there is the possibility that the transcript from cow is more similar to one of the other transcripts from human. The genomes that have a higher quality, human and mouse are precisely the ones having a higher number of defined transcripts per gene.

## 2.3 Age assignation

An age is assigned to a protein based on the phylogenetic distribution of its homologues. In the work presented in this thesis I have used two different pipelines to assign an age: BlastP searches and domain inference, which basically differ in the methodology used to assign the homologues. In both cases the first step is to select the list of species that would be used for the classification. We usually use around 15 species distributed in four age groups: Eukarya, Metazoans, Vertebrates and Mammals. If fewer species are used there is not enough power for age classification, whereas, if too many species are used the process is very slow.

The age assignation procedure is also used to identify orphan/lineage-specific genes. This type of genes is found in a restricted group of species. For example, primate-specific genes could be defined as those present in human, chimp and macaque but absent from other mammalian or vertebrate species.

### 2.3.1 BlastP searches

Once we have the set of proteins we want to classify by age we choose the list of proteomes we want to use. Then, for each protein we perform a BlastP (Altschul et al., 1997) search against each of the proteomes. BlastP is a program that uses a protein query to search into a protein database for the more similar sequences. We used an e-value cut-off of $10^{-4}$, and, consequently, we considered that a hit with an e-value at least of $10^{-4}$ is indicative of the existence of a homolog in that species (Albà and Castresana, 2005). The election of the e-value is a matter of finding a correct balance, stringent e-values have the problem of not finding possible remote-homologues and, less stringent ones can produce spurious hits.

According to the range of species in which homologs could be found, the protein is classified in one age group or in another. For example, in one of the works presented here (page 139) (Toll-Riera et al., 2009a) we classified human proteins in 4 age groups: primates, mammals, vertebrates and eukarya. If the human protein had BlastP hits in *Pan troglodytes* and *Macaca mulatta* it was classified as Primates. If it had hits in *Mus musculus*, *Rattus* *norvegicus*, *Bos Taurus*, *Canis familiaris* and primate species it was classified as Mammals. If the protein had homolgues in the primates, mammalian species and in *Gallus gallus*, *Takifugu rubripes*, *Danio rerio* and *Xenopus tropicalis* it was classified as Vertebrates. Finally, if the protein was found in all the previous cited species and in *Drosophila melanogaster*, *Caenorhabditis elegans*, *Arabidopsis thaliana*, *Saccharomyces cerevisiae*, and *Schizosaccharomyces pombe* it was classified as Eukarya (figure 2.3).

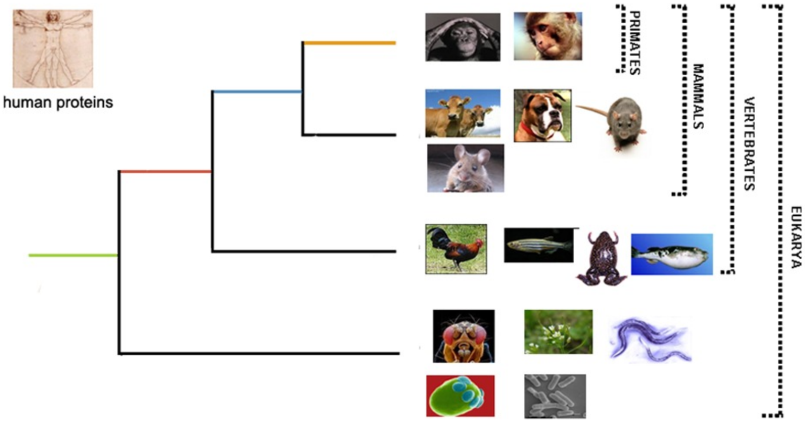


**Figure 2.3:** Schema representing how human proteins are classified into age groups according to the rank of species in which homologues could be found

However, the criteria to classify proteins into ages can be more relaxed, and this is what we did in other works presented in this dissertation (pages 81, 101). Instead of requiring finding BlastP hits for all the species in a group, we only required to find at least one hit for each group. Then, using the above example, if we have a human protein with homologues in mouse and fugu, it would be classified as vertebrate.

### 2.3.2 Domains inference

BlastP is commonly used to assign an age to proteins but it has been argued that this method suffers from circularity. Circularity is claimed due to the hypothetically lack of power of BlastP to detect homologs for fast-evolving proteins, and, therefore, those fast-evolving proteins would tend to be classified as young (Elhaik et al., 2006). However, Albà and Castresana (Albà and Castresana, 2007) have used simulations of protein evolution to demonstrate that this methodology is not affected by a Blast artefact, which only seems to happen in those rare cases in which there are extremely fast-evolving sequences and in those sequences that are evolving homogenously, which is not the case for the majority of the mammalian proteins.

However, we have come up with a new methodology to classify proteins in ages that does not rely on BlastP searches. Instead of using the whole protein to classify in ages, we used the smallest functional unit of a protein: the domain. Protein three-dimensional structures are more conserved than sequences (Wolf et al., 1999; Kinch and Grishin, 2002; Ponting and Russell, 2002), and, additionally, hidden markov models could be used to identify the domains, allowing us to look further back on time facilitating the identification of remote homologues. Although the circularity problem could still exist, with this new methodology we can perform more sensitive searches and consequently, reduce the problems in homology ascertainment.

The first step is to identify the domains present in the proteins that we want to classify in ages. To do it we used the Hmmscan program (previously named Hmmpfam), which is part of the HMMER package (Eddy, 1998). It performs searches of a single sequence against a library of hidden markov models (HMM) using a specific e-value cut-off. We used a library of HMM of domains obtained from the Pfam database (Finn et al., 2008).

Once we have assigned the domains to the proteins we have to designate an age to each of these domains. To do it we first have to choose a list of proteomes and, then, to assign an age to the domain we simply perform Hmmscan searches of the domain against those proteomes in order to know in which species the domain is present. According to the rank of species in which we find the domain we establish the age of the domain (as in the BlastP methodology). Finally, proteins are classified in age according to the oldest domain they contain.

## 2.4 Alignment process

Sequence alignments are one of the most used techniques in bioiformatics to identify residues derived from the same ancestor. When several sequences have to be compared a multiple alignment should be built. There are several softwares to compute multiple alignments (table 2.1). The classical alignment program is ClustalW, but in the last year several other more refined programs have been developed.

Most of the nowadays softwares use a method known as 'progressive algorithm' (figure 2.4), which is an agglomerative procedure. In a first step all pairwise sequence comparisons are performed to obtain a distance matrix with the percent identity. Then a clustering algorithm (NJ or UPGMA)

| Tools | Webpage |
|---|---|
| CLUSTALW | http://www.clustal.org |
| DIALIGN | http://bibiserv.techfak.uni-bielefeld.de/dialign |
| MAFFT | http://align.bmr.kyushu-u.ac.jp/mafft/software |
| MUMMALS | http://prodata.swmed.edu/mummals |
| MUSCLE | http://www.drive5.com/muscle |
| PRALINE | http://zeus.cs.vu.nl/programs/pralinewww/ |
| PRANK+F | http://www.ebi.ac.uk/goldman-srv/prank/prank |
| PRIME | http://prime.cbrc.jp |
| ProbAlign | http://probalign.njit.edu |
| PROBCONS | http://probcons.stanford.edu |
| ProDA | http://proda.stanford.edu |
| PROMALS | http://prodata.swmed.edu/promals |
| SPEM | http://sparks.informatics.iupui.edu |
| T-Coffee, M-Coffee, 3D-Coffee | http://www.tcoffee.org |

**Table 2.1:** Multiple alignment softwares. Modified from Do and Katoh (2008)

is applied into the distance matrix to obtain a guide tree. What the agglomerative algorithm does is to follow the guide tree topology from the leaf to the root aligning every pair of sequences in each node using the Needleman and Wunsch or the Viterbi algorithm. Therefore, the alignment is built up adding sequences progressively according to the guide tree order (Kemena and Notredame, 2009). The problem is that errors made at each pairwise alignment can not be repaired and are accumulated. Two techniques are used to minimize those errors: iterative refinement and consistency scoring. The iterative refinement consists in dividing repeatedly the aligned sequences into sub-alignments which are then realigned. MAFFT and MUSCLE use this technique. The consistency scoring is for example used in T-Coffee and it takes into account the information of how two sequences can be aligned with regard to other sequences (Pei, 2008).

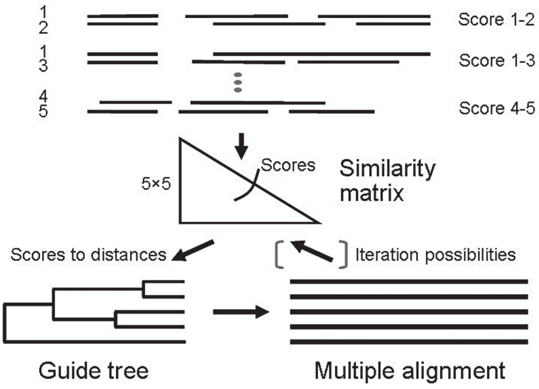
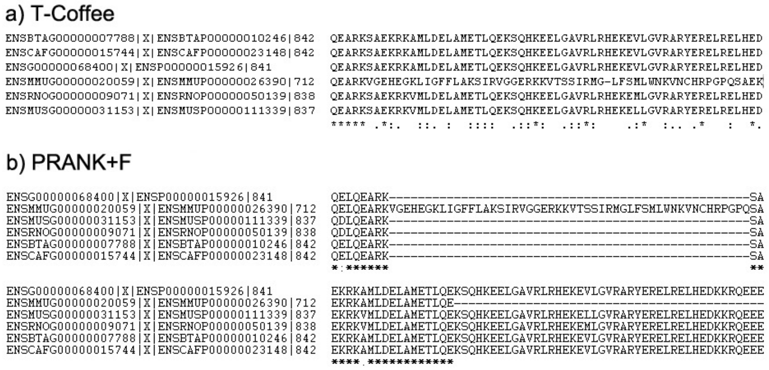


**Figure 2.4:** Schema representing the progressive algorithm. Taken from Pirovano and Heringa (2008)

Software selection depends on the sequences characteristics. In general, if the sequences to align are less than 20 and are homologous with a high percent identity (>40%) most current softwares will perform correctly. When choosing the alignment program it is important to take into account the accuracy and the computational cost. For example, MAFFT and MUSCLE are softwares that perform the alignment fast, reducing the computational cost. Also it is important the length of the alignment, CLUSTALW and MAFFT are the most efficient for longer sequences, but for example T-Coffee can not deal well with too long sequences. If the sequence identity is low (less than 35%) it has been proved that the best tools are T-Coffee, PROBCONS and MAFFT (Do and Katoh, 2008).

In the articles presented in this thesis three softwares have been used depending on the characteristics of the study: MAFFT(Katoh et al., 2002) , T-Coffee (Notredame et al., 2000) and PRANK+F (Löytynoja and Goldman, 2008). PRANK+F is a very recent algorithm which introduces gaps according to the phylogeny of the sequences and is particularly useful for evolutionary studies (figure 2.5). In evolutionary studies a gap is preferred over non similar and possibly non-homologous aligned positions because overaligned positions could produce spurious results when calculating evolutionary rates. However, most of the aligners create too few insertions and too many deletions and substitutions, resulting in very compact alignments. Thus, the novelty of the PRANK+F algorithm is the treatment of insertions and deletions as two distinct evolutionary events. PRANK+F has the disadvantage that is slow and that a phylogenetic tree is required. To perform pairwise alignments we have mostly used T-Coffee, except some cases in which the sequences were so long that T-Coffee could not deal with them and in those cases MAFFT was used.


a) T-Coffee

```
ENSBTAG00000007788|X|ENSBTAP00000010246|842  QEARKSAEKRKAMLDELAMETLQEKSQHKEELGAVRLRHEKEVLGVRARYERELRELHED
ENSCAFG00000015744|X|ENSCAFP00000023148|842  QEARKSAEKRKAMLDELAMETLQEKSQHKEELGAVRLRHEKEVLGVRARYERELRELHED
ENSG00000068400|X|ENSP00000015926|841        QEARKSAEKRKAMLDELAMETLQEKSQHKEELGAVRLRHEKEVLGVRARYERELRELHED
ENSMMUG00000020059|X|ENSMMUP00000026390|712  QEARKVGEHEGKLIGFFLAKSIRVGGERKKVTSSIRMG-LFSMLWNKVNCHRPGPQSAEK
ENSRNOG00000009071|X|ENSRNOP00000050139|838  QEARKSAEKRKVMLDELAMETLQEKSQHKEELGAVRLRHEKEMLGVRARYERELRELHED
ENSMUSG00000031153|X|ENSMUSP00000111339|837  QEARKSAEKRKVMLDELAMETLQEKSQHKEELGAVRLRHEKELLGVRARYERELRELHED
                                             ***** .*:.  ::. :  ::::  .::*:  .::*:    .:*  :.. .*   :  *.
```

b) PRANK+F

```
ENSG00000068400|X|ENSP00000015926|841        QELQEARK--------------------------------------------------SA
ENSMMUG00000020059|X|ENSMMUP00000026390|712  QELQEARKVGEHEGKLIGFFLAKSIRVGGERKKVTSSIRMGLFSMLWNKVNCHRPGPQSA
ENSMUSG00000031153|X|ENSMUSP00000111339|837  QDLQEARK--------------------------------------------------SA
ENSRNOG00000009071|X|ENSRNOP00000050139|838  QDLQEARK--------------------------------------------------SA
ENSBTAG00000007788|X|ENSBTAP00000010246|842  QELQEARK--------------------------------------------------SA
ENSCAFG00000015744|X|ENSCAFP00000023148|842  QELQEARK--------------------------------------------------SA
                                             *:******                                                  **

ENSG00000068400|X|ENSP00000015926|841        EKRKAMLDELAMETLQEKSQHKEELGAVRLRHEKEVLGVRARYERELRELHEDKKRQEEE
ENSMMUG00000020059|X|ENSMMUP00000026390|712  EKRKAMLDELAMETLQE-------------------------------------------
ENSMUSG00000031153|X|ENSMUSP00000111339|837  EKRKVMLDELAMETLQEKSQHKEELGAVRLRHEKELLGVRARYERELRELHEDKKRQEEE
ENSRNOG00000009071|X|ENSRNOP00000050139|838  EKRKVMLDELAMETLQEKSQHKEELGAVRLRHEKEMLGVRARYERELRELHEDKKRQEEE
ENSBTAG00000007788|X|ENSBTAP00000010246|842  EKRKAMLDELAMETLQEKSQHKEELGAVRLRHEKEVLGVRARYERELRELHEDKKRQEEE
ENSCAFG00000015744|X|ENSCAFP00000023148|842  EKRKAMLDELAMETLQEKSQHKEELGAVRLRHEKEVLGVRARYERELRELHEDKKRQEEE
                                             ****.************
```

**Figure 2.5:** Example of a protein alignment using T-Coffee and PRANK+F. PRANK+F reaches a higher number of identities allowing a long insertion

## 2.5 Evolutionary rates estimation

In comparative genomics we often want to quantify the speed at which amino acid changes are accumulated in the course of evolution, or in other words, the evolutionary rate. One standard way to measure the evolutionary rates is to use the number of nonsynonymous subtitutions per nonsynonymous site divided by the number of synonymous substitutions per synonymous site ($d_N/d_S$), based on alignments. Synonymous substitutions ($d_S$) are assumed to be selectively neutral and for this reason $d_S$ is used as a measure of the mutation rate. However it is now known that synonymous sites are not as neutral as previously thought and they are also under the action of selection. If we assume that synonymous substitutions are not affected by selection, we can use them as a neutral background to infer positive selection at nonsynonymous sites. By using $d_N/d_S$, $d_N$ is normalized by $d_S$ to account for local variations on the mutation rate. If $d_N > d_S$ adaptive evolution can be claimed, $d_N = d_S$ indicates that selection is absent, and finally, if $d_N < d_S$ the action of negative selection could be inferred, implying that the protein is subject to important functional constraints (Ellegren, 2008).

The two most common methods to calculate evolutionary rates are the Hyphy Package (Pond et al., 2005) and the PAML (Phylogenetic Analysis by Maximum Likelihood) package (Yang, 2007). In this dissertation we have used the PAML package, which uses maximum likelihood to perform phylogenetic analysis on DNA and protein sequences. Inside the PAML package there are several programs, and we have used codeml, which estimates the synonymous and nonsynonymous substitution rates in protein-coding DNA sequences. Additionally, it can also be used to detect positive selection. There are several models implemented in codeml:

- Basic model: all the lineages share a common $d_N/d_S$ value and a global clock is assumed.

- Branch models: the $d_N/d_S$ ratio is allowed to vary among branches. In the free-ratio model a $d_N/d_S$ value is estimated for each lineage, rates are free to vary from branch to branch (no global clock is assumed). However, instead of having a ratio for each branch it could also be specified how many ratios are wanted, and for which branches. Positive selection is detected in a specific branch only if the average $d_N/d_S$ ratio over all sites is significantly greater than 1.

- Site models: allow the $d_N/d_S$ ratio to vary among sites. This test would only detect positive selection if the average $d_N/d_S$ ratio in all the branches of the tree is greater than 1.

- Branch-site models: in this type of models the $d_N/d_S$ ratio is allowed to vary across branches and also among sites. It is very useful if one wants to detect positive selection which is only affecting few sites in some specific lineages.

In the articles included in this dissertation we have used branch models, specifically free-ratio models to calculate the $d_N/d_S$ values in a general way. $d_N$, $d_S$, $d_N/d_S$, the number of nonsynonymous substitutions and the number of synonymous substitutions could be obtained. Although $d_N/d_S$ $>$1 is generally indicative of the presence of positive selection, this is hardly found because positive selection mostly acts in a few sites and for a short period of time in a specific lineage. For this reason when we wanted to detect positive selection we used a branch-site model (type A), which is more sensitive. In this test the branch for which we want to detect positive selection is named foreground (and is has to be labelled in the tree), and the rest of branches are background branches. The test assumes four site classes (table 2.2):

- Site class 0: codons that evolve under purifying selection in all lineages.
- Site class 1: codons that evolve neutrally along the tree
- Site class 2a: codons that evolve under positive selection on the foreground branches, but under purifying selection on background branches.
- Site class 2b: codons that evolve under positive selection on the foreground branches, but neutrally in the background branches.

| Site Class | Proportion | Background | Foreground |
|---|---|---|---|
| 0 | P$_0$ | 0 $<\omega_0<$1 | 0 $<\omega_0<$1 |
| 1 | P$_1$ | $\omega_1$ =1 | $\omega_1$ =1 |
| 2a | (1- p$_0$ - p$_1$) p$_0$/( p$_0$ + p$_1$) | 0 $<\omega_0<$1 | $\omega_2 \geq$ 1 |
| 2b | (1- p$_0$ - p$_1$) p$_1$/( p$_0$ + p$_1$) | $\omega_1$ = 1 | $\omega_2 \geq$ 1 |

**Table 2.2:** Branch-site Model. Taken from Fletcher and Yang (2010)

This is the alternative hypothesis, and the null hypothesis is the same model but with the $d_N/d_S$ ratio ($\omega_2$) in the foreground branches equal to 1. To determine which of the two models is more likely to be statistically significant a likelihood ratio test statistic is computed as twice the log likelihood difference between both models ($2\Delta lnL$) and then the result is compared with the $\chi^2$ distribution with one degree of freedom. Therefore, positive selection is inferred in the foreground branch if the likelihood of the alternative value is statistically higher than the one for the null model. Interestingly, if the null model is rejected, the test also indicates which specific amino acids are under positive selection, using a Bayes empirical Bayes (BEB) approach (Fletcher and Yang, 2010).

To sum up, to calculate the evolutionary rates, the first step is to define the set of homologous sequences we want to compare. It is important to choose a set with intermediate divergence, maximizing the power of the test. Once we have chosen the sequences, we have to align them at the protein level. Then the alignment is converted into the corresponding codon alignment and passed to PAML to infer the evolutionary rates. Except for pairwise estimations, PAML also needs a phylogenetic tree, which should be unrooted unless when the molecular clock is assumed. The evolutionary rate estimates should pass some quality filters to ensure their robustness. The length of the sequence used to calculate the evolutionary rates should be at least of 60 amino acids because codeml performs better with longer sequences. Filters in $d_N$ and $d_S$ should also be set up, $d_S$ values higher than 2 should be discarded because they could be indicative of saturated sequences. Very low $d_S$ values (<0.01) should also be eliminated because do not allow reliable estimates of the $d_N/d_S$ ratio. Too high $d_N$ values (depending on the divergence of the species used) should not be included because could be indicative of non *bona fide* orthologs. And finally, correction for multiple hypothesis testing should also be considered when a biological hypothesis is lacking and several tests to detect positive selection are performed. In the article entitled *Lineage-specific variation in the intesity of natural selection in mammals* (page 61) (Toll-Riera et al., 2010) we have used the q-value test, which gives a false discovery rate and has an R library available (Storey and Tibshirani, 2003).

# 3

# Results

## 3.1 Variations in the strength of natural selection across the mammalia phylogeny

In the early sixties Zuckerkandl and Pauling observed that the number of amino acid differences between a pair of orthologous proteins was proportional to the time passed since their divergence. This observation lead to the formulation of the molecular clock hypothesis, in which it is proposed that protein coding genes evolve at a constant rate. Lately, it was proposed that this rate would remain constant as long as the proteins do not suffer functional or structural changes. The molecular clock is one of the evidences that neutralists claim to prove the Neutral Theory.

This chapter includes one published article that studies lineage-specific deviations in the intensity of natural selection in a dataset of mammalian orthologous genes, because, as commented previously, could be indicative of functional or structural changes. Those deviations are identified using a new methodology that is able to detect them by building a species tree and comparing it to a gene tree. In this article we do not get into the neutralist-selectionist debate.

### 3.1.1 Lineage-Specific Variation in Intensity of Natural Selection in Mammals

**Authors:** Macarena Toll-Riera, Steve Laurie and M. Mar Albà

**Published in:** Molecular Biology and Evolution. 2011 Jan; 28(1):383-98

**Full text:** http://mbe.oxfordjournals.org/content/28/1/383.abstract

**Summary**

The molecular clock hypothesis states that protein-coding genes evolve at an approximately constant rate. However, this is only expected to be true as long as the function and the tertiary structure of the molecule remain unaltered. An important implication of this statement is that significant deviations in the rate of evolution of a gene with respect to the species clock are likely to reflect functional and/or structural alterations. Here, we present a method to identify such deviations and apply it to a data set of 2,929 high-quality coding sequence alignments corresponding to one-to-one orthologous genes from six mammalian species-human, macaque, mouse, rat, cow, and dog. Deviated branches are defined as those that present significant alterations in both the rate of nonsynonymous substitutions ($d_N$) and the selective pressure ($d_N/d_S$). Strikingly, we find that as many as 24.5% of the genes show branch-specific deviations in $d_N$ and $d_N/d_S$, though this is a relatively well-conserved set of genes. Around half of these genes show branch-specific acceleration of evolutionary rates. Positive selection (PS) tests based on divergence data only identify 17.7% of the accelerated branches. Failure to identify PS in accelerated branches with an excess of radical amino acid replacements suggests that these tests are conservative. Interestingly, genes with accelerated branches are significantly enriched in neural proteins, indicating that this type of protein might play a more important role than previously thought in species diversification, although they are generally not detected by PS tests. We discuss in detail several examples of genes that show lineage-specific evolutionary rate acceleration and are involved in synaptic transmission, chemosensory perception, and ubiquitination

## 3.2 Age as a determinant for protein evolution

Several determinants for protein evolution have been proposed such as protein-protein interactions, structural properties, gene expression and dispensability. However, which determinants are driving evolution is still under debate.

In this section two submitted articles are presented. In the first article we have studied protein age as a possible determinant of protein evolution. We have centred on how human protein domains classified in three different age groups evolve, including the comparison between domains classified in different age groups that belong to the same protein. The comparison of evolutionary rates of domains found inside the same protein permits us to control for two other proposed determinants, function and expression. The second article was mainly done during a short-stay in Dr. Joshua B. Plotkin group, at the University of Pennsylvania. In this work we analyzed the interaction of age and structural characteristics in an attempt to find an explanation for the described inverse relationship between age and evolutionary rate.

### 3.2.1 The Signature of Time: Younger Domains in Proteins Evolve Faster than Older Ones

**Authors:** Macarena Toll-Riera and M.Mar Albà

**Published in:** Submitted

**Summary**

The causes behind the wide variation in the evolutionary rates exhibited by different genes have puzzled scientists for decades. It has been observed that recently-formed lineage-specific genes evolve significantly faster than older genes, pointing to the intriguing possibility that the time past since the gene originated strongly influences its current evolutionary rate. Moreover, it has been observed than young genes tend to be shorter than older ones, which may imply that genes tend to increase in length over time. In order to further understand how the age of a sequence impacts its subsequent evolution, we have turned our attention to the evolutionary rates of protein domains. As domains of different age can combine in the same protein, this analysis has the advantage that we can automatically control for other factors that have been related to evolutionary rate variation, such as gene expression level and protein functional class. We report that the age of a domain strongly correlates with its evolutionary rate (measured as $d_N/d_S$) both in mammals and flies. Importantly, young domains evolve significantly faster than older domains even when located in the same protein, providing the strongest evidence to date that the time of origin of a sequence carries a signature of its evolutionary rate that is independent of any properties at the whole gene level. We also show that, in mammalian proteins, novel domains tend to be incorporated at the protein N-terminus, resulting in an increase in protein sequence length.

# The Signature of Time: Younger Domains in Proteins Evolve Faster than Older Ones

Macarena Toll-Riera [1] and M.Mar Albà [1,2,*]

**Abstract**

The causes behind the wide variation in the evolutionary rates exhibited by different genes have puzzled scientists for decades. It has been observed that recently-formed lineage-specific genes evolve significantly faster than older genes, pointing to the intriguing possibility that the time past since the gene originated strongly influences its current evolutionary rate. Moreover, it has been observed than young genes tend to be shorter than older ones, which may imply that genes tend to increase in length over time. In order to further understand how the age of a sequence impacts its subsequent evolution, we have turned our attention to the evolutionary rates of protein domains. As domains of different age can combine in the same protein, this analysis has the advantage that we can automatically control for other factors that have been related to evolutionary rate variation, such as gene expression level and protein functional class. We report that the age of a domain strongly correlates with its evolutionary rate (measured as $d_N/d_S$) both in mammals and flies. Importantly, young domains evolve significantly faster than older domains even when located in the same protein, providing the strongest evidence to date that the time of origin of a sequence carries a signature of its evolutionary rate that is independent of any properties at the whole gene level. We also show that, in mammalian proteins, novel domains tend to be incorporated at the protein N-terminus, resulting in an increase in protein sequence length.

KEYWORDS: GENE AGE, PROTEIN DOMAIN, EVOLUTIONARY RATE
RUNNING TITLE: SEQUENCE AGE AND EVOLUTIONARY RATE

## 1 Introduction

It has long been noticed that different genes evolve at markedly different rates, an observation initially attributed to differences in the selective constraints affecting proteins performing different cellular functions (Zuckerkandl, 1976; Wilson et al., 1977; Doolittle et al., 1986). The advent of genomic sequences and high-throughput functional experimentation has fuelled the search for universal factors that may explain gene evolutionary rate variation, including the number of protein-protein interactions, gene dispensability (Hirsh and Fraser, 2001; Wall et al., 2005), gene expression level (Pál et al., 2001; Drummond et al., 2005) and gene tissue expression breadth (Duret and Mouchiroud, 2000; Zhang and Li, 2004). Among these factors, possibly the strongest correlate is gene expression level (Krylov et al., 2003; Drummond et al., 2006). It has been argued that the fact that highly expressed genes evolve slowly is due to the increased selective pressure to prevent their misfolding and thus, avoid their accumulation in the cell, which is very toxic (Drummond and Wilke, 2008).

Another important factor correlating with gene evolutionary rate is gene age: recently emerged genes tend to evolve more rapidly than older genes in a wide variety of organisms, including *Drosophila* (Domazet-Loso and Tautz, 2003; Wolf et al., 2009), bacteria (Daubin and Ochman, 2004), mammals (Albà and Castresana, 2005; Luz et al., 2006; Toll-Riera et al., 2009; Wolf et al., 2009; Cai and Petrov, 2010), fungi (Cai et al., 2006) and *Plasmodium* (Kuo and Kissinger, 2008).Therefore, it seems that the time of origin of a gene bears a signature that strongly affects gene's evolutionary rate. It is important to note that gene age and level of expression are not independent factors: young genes are generally expressed at lower

[1] Evolutionary Genomics Group, Research Programme on Biomedical Informatics (GRIB), Hospital del Mar Research Institute (IMIM) - Universitat Pompeu Fabra (UPF), Barcelona, Spain
[2] Catalan Institution for Research and Advanced Studies (ICREA), Barcelona, Spain
* To whom correspondence should be addressed

levels and in fewer tissues than older genes (Toll-Riera et al., 2009; Cai and Petrov, 2010). In addition, regression models including both gene age and gene expression level have failed to identify a dominant factor (Cai and Petrov, 2010).

All previous studies on the influence of gene age on evolutionary rates have used complete protein sequences, but a more powerful approach is to use protein domains. Domains are discrete protein regions that are present in several proteins, have specific functions, can fold independently and, very importantly, have been formed at different times (Lander et al., 2001; Müller et al., 2002; Vogel et al., 2004; Ekman et al., 2005; Choi and Kim, 2006; Pal and Guda, 2006). Typically, the age of a protein sequence is determined by BLASTP sequence similarity searches against the proteomes derived from fully sequenced genomes (Tautz and Domazet-Lošo, 2011). Using protein evolution simulations it has been shown that this method recovers most of the homologues in distant species even if the proteins are evolving at the highest observed rates (Albà and Castresana, 2007). To classify domains in different age classes we can benefit from the existing libraries of domain-specific hidden markov models, which increases the depth of homology detection (Eddy, 1998). More importantly, as domains of different age may combine in the same protein, we can directly examine the influence of the time of formation on the rate of evolution, independently of gene expression level or protein functional class (Wolf et al., 2008; Zhou et al., 2008). Here we quantify for the first time the influence of the age of a domain in the pace of sequence evolution. We show that proteins are heterogeneous entities formed by sequences added at different times that evolve according to their age.

## 2 Results

### 2.1 Young protein domains evolve faster than older ones

We identified all known protein domains in a large set of human proteins with 1:1 orthologues in mouse (15,630 genes) using domain-specific hidden markov models (HHMs) from the Pfam database (Finn et al., 2008). Subsequently, we determined the phylogenetic distribution of each domain by performing HHM-based searches against 15 additional eukaryotic species. We classified 21,730 human domain occurrences, corresponding to 3,473 different Pfam domain types, into one of the three age classes: 'Mammalian', 'Vertebrate' and 'Old' (see Materials and Methods). The most common domains in each category are shown in S1. Domains of different age were on average very similar in length (S2) but showed very remarkable differences in the non-synonymous to synonymous substitution rate ratio ($d_N/d_S$) (Figure 1, S3). Younger domains showed significantly higher $d_N/d_S$ values than older ones, indicating that they are evolving more rapidly (Kolmogorov-Smirnov test, $p<10^{-5}$). The results did not vary significantly when we used the median $d_N/d_S$ for each domain type as the representative domain $d_N/d_S$ value, which eliminated possible biases caused by very abundant domains (S4). Additionally, the inverse relationship between domain age and evolutionary rate was maintained when a wide range of E-value cut-offs was used for the classification of domains in different age classes (S5 and S6).

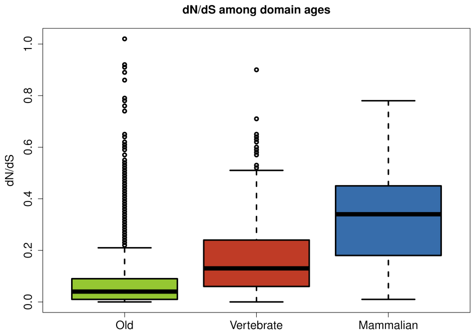


**Figure 1.** Non-synonymous to synonymous substitution rate ratio ($d_N/d_S$) for protein domains of different age. Old: 12,076 domains, Vertebrate: 521 domains, Mammalian: 47 domains. $d_N$ and $d_S$ were calculated for human and mouse orthologous genes, domains with unreliable $d_N$ or $d_S$ estimates were not considered (domain length <60 amino acids or $d_N$ >0.5 or $d_S$ >2). Differences between any pair of age classes were statistically significant (Kolmogorov-Smirnov test, p-value$<10^{-5}$). The area within the box contains 50% of the data; horizontal line is the median; outliers (5%) are represented as small circles.

## 2.2 When the young and the old combine in a single protein

Domains of different age are sometimes found in the same protein. In this situation, do they maintain their characteristic age-related evolutionary rates? To answer this question we focused on the 330 proteins containing both Old and Vertebrate domains (only a few proteins contained Mammalian domains combined with domains of a different age). Interestingly, we found that, in general, Vertebrate domains continued to evolve significantly faster than Old domains (Wilcoxon test, p$<10^{-5}$) (Figure 2). Besides, out of 174 domain pairwise comparisons, 141 showed higher $d_N/d_S$ values for the Vertebrate domain, compared with only 27 for the Old domain (the remaining 6 cases did not show any significant differences, binomial test, p-value >0.01). Furthermore, the relative difference in $d_N/d_S$ values tended to be much larger in pairs in which the Vertebrate domain evolved faster than when the Old domain evolved faster (Figure 3). In conclusion, the age-related differences in domain evolutionary rates were essentially maintained when the domains were found in the same protein.

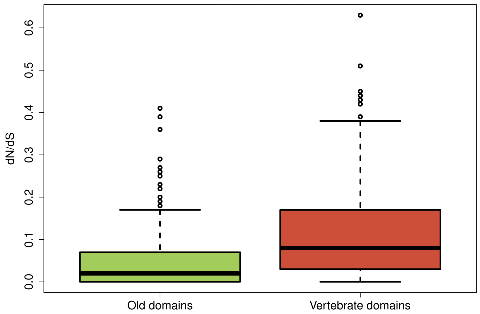


**Figure 2.** Distribution of non-synonymous to synonymous ($d_N/d_S$) substitution rates for domains of different age combined in the same protein. $d_N/d_S$ values were calculated in the domains found in 330 human and mouse 1:1 orthologous proteins. Differences in $d_N/d_S$ between Old and Vertebrate domains were statistically highly significant (Wilcoxon test, p$<10^{-5}$).

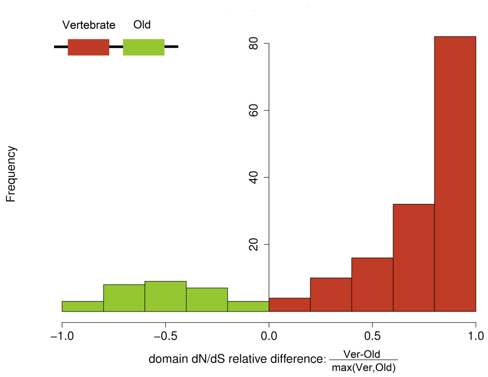


**Figure 3.** Differences in $d_N/d_S$ between Vertebrate and Old domains located in the same protein. N=174 pairs of Old and Vertebrate domains; domain $d_N/d_S$ relative difference: non-synonymous to synonymous substitution rate ratio ($d_N/d_S$) of the Vertebrate domain minus the $d_N/d_S$ of the Old domain divided by the higher of the two.

The 330 human proteins containing both Old and Vertebrate domains is an ideal set to test if we have underestimated the age of the domains classified as Vertebrates. Those proteins showed homology - through the Old domain - to proteins in non-vertebrate species. But did they contain any traces of the domain we had classified as Vertebrate? To be able to address this question we generated pairwise alignments of the human protein and the first three BLASTP hits in *D.melanogaster* and *C.elegans*, using Prank+F (Löytynoja and Goldman, 2008). In contrast to the Old domain, which was well conserved in the non-vertebrate species, the percent identity of the region corresponding to the Vertebrate domain was very low, peaking at about 10%, and, in fact, the majority of proteins showed less than 20% identity (S7). This is consistent with most Vertebrate domains being genuinely vertebrate-specific. When we used the 20% identity cut-off to keep true vertebrate-specific domains we found that the percent identity distribution of the domains classified as Vertebrate was equivalent to the one obtained when the domains were shuffled randomly, denoting complete lack of homology (S8). In contrast, the percent identity of the Old domains was much higher than for randomly shuffled sequences, as expected under homology. Restricting the analysis to proteins with less than 20% identity (285 proteins), or to proteins with less than 10% identity in the Vertebrate domain region (69 proteins), did not significantly alter the differences in $d_N/d_S$ in Old versus Vertebrate domains (S9 and S10, respectively), indicating that the results are highly robust.

Figure 4 shows several examples of Old and Vertebrate domains combined in the same protein. The first example is the heat shock transcription factor 1, a key regulator of the expression of heat shock proteins (Anckar and Sistonen, 2011). The vertebrate-specific transactivator domain has a $d_N/d_S$ of 0.1 compared to 0.002 for the phylogenetically well-conserved DNA-binding domain. Platelet-derived growth factor alpha polypeptide is important for the formation of oligodendrocytes in the central nervous system (Frost et al., 2009). The N-terminal region is only found in vertebrates and is currently evolving two orders of magnitude more rapidly than the oldest part. Finally, the progesterone receptor contains two old domains, a C4 type Zinc finger that acts as a DNA binding domain, and a hormone-binding domain that recognizes the hormone, as well as a vertebrate-specific domain that mediates receptor-specific interactions (Wardell et al., 2002). Again, the latter domain is evolving much faster than the other two domains.

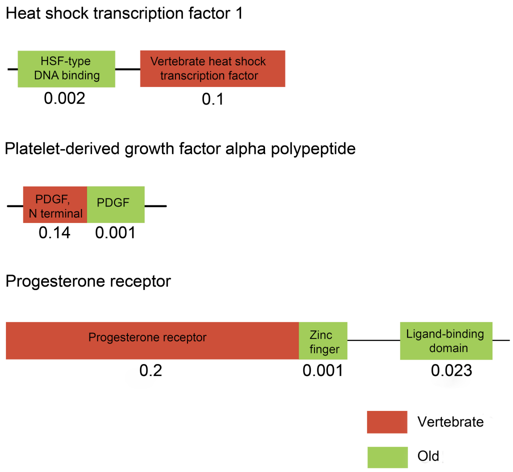


**Figure 4.** Proteins containing domains of different age. Ensembl protein identifiers: Heat shock transcription factor 1: ENSP00000332698; platelet-derived growth factor alpha polypeptide: ENSP00000346508; progesterone receptor: ENSP00000325120. Non-synonymous to synonymous ($d_N/d_S$) substitution rate values are indicated below each domain

## 2.3 Protein length increase

If we date the age of a protein using the oldest domain it contains, or using BLASTP searches if the protein has no known domains, we observe that older proteins tend to be longer than younger proteins (Table 1). This is not due to older domains being shorter (S2) but it is related to older proteins containing, on average, a higher number of domains. This seems to suggest that, as proteins get older, they tend to increase their length through the gain of new domains. Is this model supported by the data?

| | Age | N proteins | Domains/prot. | Length protein* | dN/dS$^{a}$* |
|---|---|---|---|---|---|
| Proteins with domains | | | | | |
| | Old | 1039 | 1.91 (1) | 616.3 (473) | 0.11 (0.08) |
| | Vertebrate | 473 | 1.15 (1) | 394.3 (269) | 0.21 (0.18) |
| | Mammalian | 62 | 1.02 (1) | 267.9 (163) | 0.35 (0.33) |
| Proteins without domains | | | | | |
| | Old | 1816 | NA | 654.6 (501) | 0.15 (0.12) |
| | Vertebrate | 851 | NA | 449.0 (319) | 0.21 (0.18) |
| | Mammalian | 358 | NA | 308.2 (214.5) | 0.39 (0.31) |

**Table 1.** Evolutionary properties of human proteins of different age. Mean and median (in brackets) are shown. $^{a}d_N/d_S$ was calculated for 10,636 Old, 416 Vertebrate and 40 Mammalian proteins with domains, and for 1,740 Old, 784 Vertebrate and 274 Mammalian proteins without domains. *Kolmogorov-Smirnov test $p<10^{-5}$ in all pairwise comparisons

We used again the set of 330 proteins containing both Old and Vertebrate domains to examine this question. In most cases, the Old domain could be found both in combination with the Vertebrate domain but also in other configurations (alone or in combination with other Old domains). In contrast, most Vertebrate domains in these proteins were only found as part of a combination with the Old domain (Figure 5a). Therefore, in most cases, the Vertebrate domain had been formed in the context of an existing, older, protein. Less often, both the Old and the Vertebrate domain could be found separately as well as in combination (Figure 5b). These cases were compatible with domain fusion. For comparison, Vertebrate domains not found combined with any Old domain were nearly twice as abundant as Vertebrate domains born in the context of an existing protein (Figure 5c). These domains represent genes formed *de novo*.

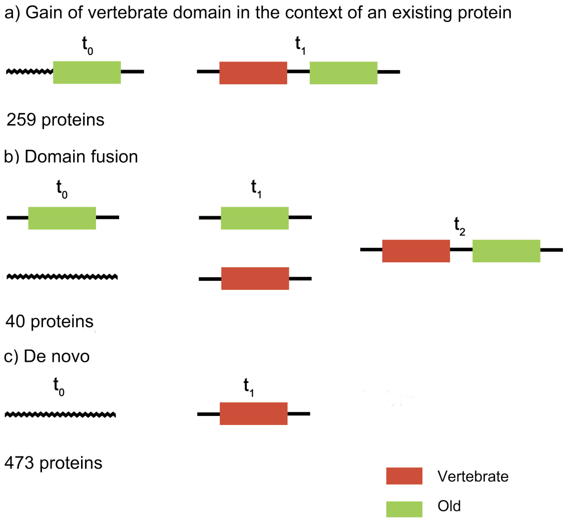


**Figure 5.** Evolutionary scenarios for proteins containing Vertebrate domains. a) The Old domain can be found in combination with the Vertebrate domain as well as separate, but the Vertebrate domain is always found in combination with the Old domain; 118 different Vertebrate domains, 148 different Old domains. b) Both Old and Vertebrate domains can be found combined as well as separate; 11 different Vertebrate domains, 23 different Old domains. c) Vertebrate domains that are never found in combination with Old domains; 234 different Vertebrate domains.

Where are the domains preferentially gained in existing proteins? Both in Old proteins with two domains and in Old proteins with more than two domains we found a strong bias for the incorporation of the Vertebrate domains in the N-terminus (chi-square test, $p<10^{-5}$, Figure 6). Taken together, these data indicates that proteins tend to increase their length over time through the incorporation of novel domains, and that, in mammalian proteins, this happens by extensions of the coding sequence at the 5'end of genes.

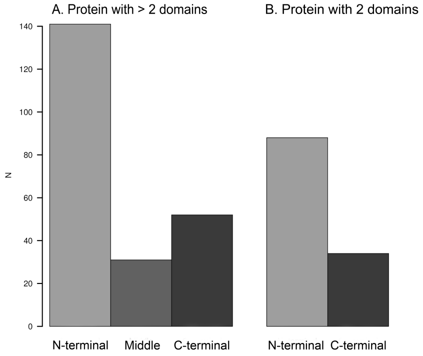


**Figure 6.** Relative position of the Vertebrate domain in proteins combining Old and Vertebrate domains. A. Proteins containing more than 2 domains. B. Proteins containing 2 domains.

## 3 Discussion

An inverse relationship between gene age and evolutionary rate has been observed in different lineages and seems to be universal (Domazet-Loso and Tautz, 2003; Albà and Castresana, 2005; Cai et al., 2006; Wolf et al., 2009). However, so far, the definition of the age of the gene has remained linked to the power to detect homologues via pairwise sequence comparisons, which may potentially hinder the identification of distant homologues of rapidly evolving proteins (Elhaik et al., 2006). Even though sequence simulations show that this is unlikely to be the case, at least for relatively young proteins (Albà and Castresana, 2007), some researchers prefer to employ the term apparent age to indicate the point at which homologues are no longer detected (Wolf et al., 2009). The use of hidden markov models based on conserved sequence domains improves homology detection. Using these models we have found that the time past since a sequence has originated dominates its pace of evolution to an unsuspected high degree: when young and old domains are found in the same protein they maintain their characteristic evolutionary rate differences. Although we have based our study on human proteins, this trend is likely to be general, as we have detected a similar inverse relationship between protein domain age and evolutionary rate in *D.melanogaster* (S11,S12).

Most studies searching for factors influencing protein evolutionary rates have taken complete proteins as the evolutionary units (Duret and Mouchiroud, 2000; Hirsh and Fraser, 2001; Pál et al., 2001; Krylov et al., 2003; Zhang and Li, 2004; Wall et al., 2005; Drummond et al., 2006). An exception is a study by Wolf and colleagues, in which they investigated whether domains located in the same protein had more homogeneous evolutionary rates than when the same domains were found in different proteins, as would be expected if the level of expression of a gene was a strong determinant of evolutionary rate (Wolf et al., 2008). They concluded that rates do indeed became more similar when the domains were in the same protein, but the differences between domains were not completely erased, indicating that other factors, intrinsic to the domains, were also important. Our analysis shows that young domains continue to evolve very rapidly compared to old domains when located in the same protein (Figure 2, 3, S3, S4), supporting a model in which the expression level has little or no influence on evolutionary rates. As the vast majority of these domains do not exist in a different context, a direct comparison of the same domain alone or in combination, analogous to the one in Wolf et al. (Wolf et al., 2008), was not applicable here.

The relationship between domain age and evolutionary rate (measured as $d_N/d_S$) is not linear but shows an exponential decrease from youngest to oldest, until the rate eventually stabilizes (Figure 1, S11). The latter is indicated by the lack of significant rate differences between Metazoan and Eukarya domains, independently of where we measure them in human and mouse orthologues, or in *D.melanogaster* and *D.simulans* orthologues. What are the implications of the differences in the evolutionary rates of young and old sequences? One hypothesis is that the strength of purifying selection increases over time, as more residues become co-opted for function (Albà and Castresana, 2005). Another hypothesis, which is not incompatible with the first one, is that adaptive mutations become increasingly saturated over time (Goodman et al., 1975; Hartl et al., 1985). Cai and Petrov have recently reported that relaxed purifying selection, and to a lesser extent adaptive selection, may explain the high evolutionary rates of human young genes (Cai and Petrov, 2010). It has also been observed that younger proteins also show more variable rates than older ones (Vishnoi et al., 2010), indicating that it seems possible that the function of younger proteins tends to change more readily over time than that of older proteins as higher rates would facilitate functional shifts.

The largest fraction of young domains is not found in combination with older domains but corresponds to novel genes (Figure 5c). We observed that these genes typically contain a single domain (Table 1), which is consistent with results from previous studies focusing on domain combinations (Pal and Guda, 2006; Ekman et al., 2007). In addition, many novel genes do not contain any known domains. Capra and colleagues also found younger proteins to be less covered by Pfam domains than average (Capra et al., 2010). Lineage-specific proteins, even if very young, can play essential functions (Chen et al., 2010), so the lack of annotation of domains in young proteins probably reflects our poor understanding of such proteins rather than a lack of functionality.

Novel domains can be gained by several mechanisms, such as gene fusion, exon extension, recombination and retrotransposition (Björklund et al., 2005; Ekman et al., 2007; Marsh and Teichmann, 2010). It has been reported that domain architecture in all branches of life tends to gain complexity over time, with a preponderance of fusion events (Fong et al., 2007; Buljan et al., 2010). In the case of proteins combining old and young domains, we find that the most common scenario is the formation of novel domains inside existing proteins, followed by domain fusion, predominantly at the N-terminus. These observations, together with the finding that old proteins tend to be longer than younger ones (Albà and Castresana, 2005; Toll-Riera et al., 2009; Wolf et al., 2009; Capra et al., 2010), support a model in which proteins tend to increase their length and domain complexity over time.

In conclusion, this works demonstrate that the time of origin of a sequence is one of the strongest determinants of its current evolutionary rate. Novel domains evolve much faster than older domains, even if found in the same protein. Therefore proteins need to be considered heterogeneous entities in which sequences formed at different times maintain their characteristic evolutionary signature.

# 4 Material and Methods

## 4.1 Protein domain identification

We obtained 15,630 one-to-one orthologous human and mouse genes using version 56 of Ensembl (Hubbard et al., 2009). We took the protein corresponding to the longest coding transcript for each gene as the representative one, as defined in Ensembl.

We used Hmmpfam (HMMER 2.3.2) (Eddy, 1998) to identify all known protein domains. We employed the Pfam_ls (version 23) (Finn et al., 2008) library of domains, which contains 10,340 hidden markov models (HMMs), and an E-value cut-off of $10^{-5}$. We used an in-house Perl program to parse the Hmmpfam results and to assign the domains to the proteins. We identified 3,482 different domains in a set of 14,784 human proteins with 1:1 orthologs in mouse.

## 4.2 Determination of the age of protein domains

To classify human domains into age groups we used the following classes: mammals (*Mus musculus*, *Rattus norvegicus*, *Bos Taurus*), non-mammalian vetebrates (*Danio rerio*, *Gallus gallus*, *Takifugu rubripes*, *Xenopus tropicalis*), other metazoans (*Anopheles gambiae*, *Caenorhabditis elegans*, *Ciona intestinalis*, *Drosophila melanogaster*) and other eukaryotes (*Arabidopsis thaliana*, *Oryza sativa*, *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*). We assigned an age group to each domain following the rank of species in which a domain was found. For example if a human domain was found in at least one mammalian species but in none of the other vertebrate, metazoan or eukaryotic species it was classified as Mammalian. We classified 2,294 different human domains as Eukarya, 745 as Metazoan, 369 as Vertebrate and 65 as Mammalian. The Eukarya and Metazoan groups were both considered Old (>500 Mya) and the Vertebrate and Mammalian Young (<500 Mya). The classification was robust for a range of Hmmpfam searches E-value cut-offs (Supplemental file S5 and S6).

## 4.3 Determination of the age of proteins

Domain ages were used to classify proteins in ages according to the age of the oldest domain they contained. We obtained 11,039 proteins classified as Old, 473 as Vertebrate and 62 as Mammalian.

The dataset contained 3,088 proteins that did not have any domain. For these proteins we used BLASTP sequence similarity searches (Altschul et al., 1997) against the genomes listed before to classify them in age groups (E-value $<10^{-4}$). Following this procedure we obtained 932 proteins classified as Eukarya, 884 in Metazoan, 851 in Vertebrate and 358 in Mammalian.

## 4.4 Estimation of evolutionary rates

We aligned orthologous domains from human and mouse protein sequences using T-coffee (Notredame et al., 2000). To make sure that we were aligning orthologous domains, we only used orthologues in which the domain structure was completely conserved between the two species, obtaining a total of 18,193 human-mouse alignments. Subsequently, we obtained nucleotide coding sequence domain alignments based on the T-coffee protein alignments using an in-house Perl program. Alignments at protein and coding sequence level were performed following the same procedure as for domains.

We estimated non-synonymous ($d_N$) and synonymous ($d_S$) substitution rates in human and mouse orthologues. For each domain and protein alignment, we estimated the number of non-synonymous substitutions per non-synonymous site ($d_N$), the number of synonymous substitutions per synonymous site ($d_S$), and the $d_N/d_S$ ratio, using maximum likelihood as implemented in the codeml program of the PAML software package (Yang, 2007). The estimations were performed pairwisely, using runmode =-2. Under this condition the program automatically sets model=0 and nssites =0.

Domains shorter than 60 amino acids and/or with a $d_N > 0.5$ and/or $d_S > 2$ were discarded to ensure robustness in the evolutionary rates estimation. After the filtering process we ended up with 12,647 different human domains with $d_N$ and $d_S$ data. Eukarya and Metazoan showed a very similar evolutionary rate distribution in both cases, and were merged into a single group (Old) for the rest of the analysis.

### 4.5 Comparisons of evolutionary rates from pairs of domains located in the same protein

We compared the non-synonymous to synonymous substitution rates ($d_N/d_S$) of pairs of Old and Vertebrate domains located in the same protein (330 proteins, Figure 2). We computed the difference in $d_N/d_S$ of the Vertebrate domain minus the $d_N/d_S$ of the Old domain and divided it by the higher $d_N/d_S$ of the two (Figure 3). To determine if the difference in the estimated number of non-synonymous substitutions over synonymous substitutions was statistically different between Old and Vertebrate domains we applied a binomial test comparing the total number of non-synonymous substitutions and synonymous substitutions between the two age groups.

### 4.6 Statistical Tests and Graphics

The R statistical software package (R Development Core Team, 2008) was used to perform all statistical tests and generate graphics.

## 5 Acknowledgements

We would like to thank Nicolás Bellora and the Evolutionary Genomics Group members for fruitful discussions. We received financial support from Ministerio de Educación (FPU to M.T.-R.), Ministerio de Innovación y Tecnología (BIO2009-08160) and Institució Catalana de Recerca i Estudis Avanats (M.M.A.).

# 6 Supplementary Information

| Domain | Number of domain occurrences | PF id |
|---|---|---|
| **Old** | | |
| 7 transmembrane receptor (rhodopsin family) | 385 | PF00001 |
| Protein kinase domain | 339 | PF00069 |
| Zinc finger, C2H2 type | 280 | PF00096 |
| PH domain | 181 | PF00169 |
| Homeobox domain | 175 | PF00046 |
| RNA recognition motif | 139 | PF00076 |
| Zinc finger, C3HC4 type | 122 | PF00097 |
| PDZ domain | 120 | PF00595 |
| SH3 domain | 115 | PF00018 |
| Immunoglobulin I-set domain | 114 | PF07679 |
| **Vertebrate** | | |
| KRAB box | 74 | PF01352 |
| SCAN domain | 33 | PF02023 |
| S-100/ICaBP type calcium binding domain | 17 | PF01023 |
| Small cytokines (intecrine/chemokine) interleukin-8 like | 17 | PF00048 |
| Mammalian taste receptor protein | 11 | PF05296 |
| Protein of unknown function | 11 | PF04826 |
| u-Par/Ly-6 domain | 11 | PF00021 |
| **Mammalian** | | |
| Transcription elongation factor A | 4 | PF06137 |
| ”Intracellular adhesion molecule, N-terminal domain” | 4 | PF03921 |
| Cornifin (SPRR) family | 3 | PF02389 |

**Table S1.** Most common domains in each age class

| Age domain | N | Median domain length | Mean domain length |
|---|---|---|---|
| Old | 20735 | 101 | 145.8 |
| Vertebrate | 916 | 102 | 157.2 |
| Mammals | 79 | 111 | 162 |

**Table S2.** Domain sequence length statistics for domains of different age

| Age | N domain occurrences | N domain types | Average | $d_N/d_S$ | $d_N$ | $d_S$ |
|---|---|---|---|---|---|---|
| Old | 12076 | 2586 | Mean | 0.07 | 0.05 | 0.75 |
| | | | Median | 0.04 | 0.03 | 0.66 |
| Vertebrate | 521 | 268 | Mean | 0.17 | 0.12 | 0.8 |
| | | | Median | 0.13 | 0.1 | 0.74 |
| Mammalian | 47 | 37 | Mean | 0.33 | 0.24 | 0.8 |
| | | | Median | 0.34 | 0.25 | 0.76 |

**Table S3.**Relationship between evolutionary rates and protein domain age. Non-synonymous ($d_N$) and synonymous ($d_S$) substitution rates, calculated for human and mouse orthologous sequenes, corresponding to domains classified in different age classes (Old, Vertebrate, Mammalian). $d_N$ and $d_S$ statistics are calculated for all domain ocurrences. N domain types refer to the number of non-redundant domains. The number of domains analyzed is lower than in S2 because we filtered out domains with unreliable $d_N$ or $d_S$ estimates (domains shorter than 60 amino acids or with $d_N > 0.5$ or with $d_S > 2$)

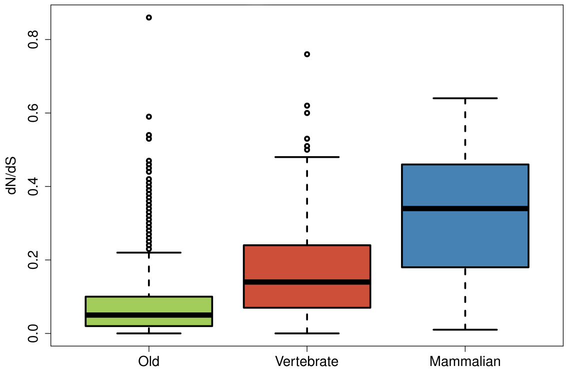


**Figure S4.** Distribution of the median values corresponding to the non-synonymous to synonymous ($d_N/d_S$) substitution rates for each domain type. The area within the box contains 50% of the data; horizontal line is the median. Outliers (5%) are represented as small circles. Differences between pairs of groups are highly significant (Kolmogorov-Smirnov test, p<$10^{-5}$).

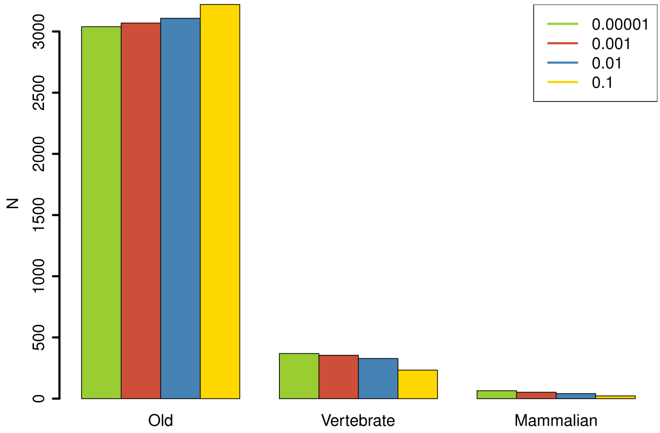


**Figure S5.** Number of domains classified as Old, Vertebrate and Mammalian, depending on the E-value cutoff employed for the identification of domains in different proteomes using searches with the Pfam library (see main manuscript file).

| E-value | Age | N | N after filtering | Average $d_N/d_S$ | Median $d_N/d_S$ |
|---|---|---|---|---|---|
| 0.00001 | Old | 3039 | 2586 | 0.07 | 0.05 |
| | Vert | 369 | 268 | 0.17 | 0.14 |
| | Mam | 65 | 37 | 0.33 | 0.34 |
| 0.001 | Old | 3069 | 2605 | 0.07 | 0.05 |
| | Vert | 355 | 256 | 0.18 | 0.14 |
| | Mam | 53 | 29 | 0.32 | 0.34 |
| 0.01 | Old | 3108 | 2629 | 0.07 | 0.05 |
| | Vert | 328 | 256 | 0.18 | 0.14 |
| | Mam | 41 | 21 | 0.31 | 0.34 |
| 0.1 | Old | 3321 | 2702 | 0.07 | 0.05 |
| | Vert | 233 | 175 | 0.2 | 0.15 |
| | Mam | 23 | 14 | 0.3 | 0.37 |

**Table S6.** Non-synonymous to synonymous ($d_N/d_S$) substitution rates for domains classified in different age classes defined using different Hmmpfam E-value cut-offs. N after filtering refers to the number of domain types left after filtering cases with no reliable $d_N$ and/or $d_S$ measurements (see Materials and Methods in main manuscript text)

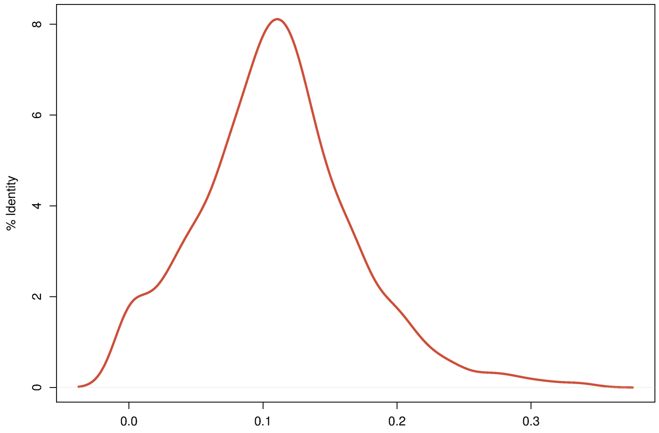


**Figure S7.** Percent identity of the region corresponding to the Vertebrate domain in 330 proteins containing both Old and Vertebrate domains. Pairwise complete protein sequence alignments between human and *C.elegans* and human and *D.melanogaster* homologs were generated with Prank+F. The first BLASTP hit in *C.elegans* or *D.melanogaster* was taken for generating this figure, although similar results were obtained for the second and third hits.

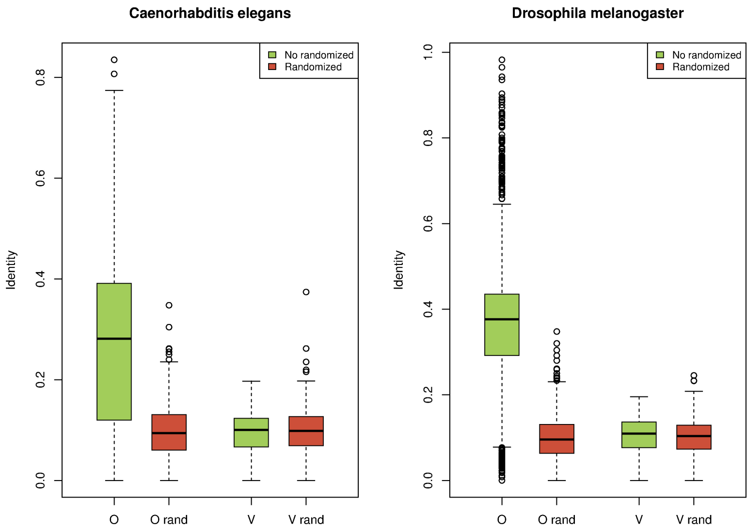


**Figure S8.** Comparison of the percent identity of human domains classified as Old or Vertebrate in complete protein alignments with *C. elegans* and *D. melanogaster* homologues. The data corresponds to 285 different proteins in which the region corresponding to the Vertebrate domain shows less than 20 percent identity. The first BLASTP hit in *C.elegans* or *D.melanogaster* was taken, although similar results were obtained for the second or third hits. O: Old domains; O rand: random shuffle of old domains; V: Vertebrate domains; V rand: random shuffle of vertebrate domains. Differences between O and O rand are highly significant in both comparisons (Wilcoxon test, p $<10^{-15}$). Differences between V and V rand are not significant in any of the two comparisons (Wilcoxon test, p $>0.05$).

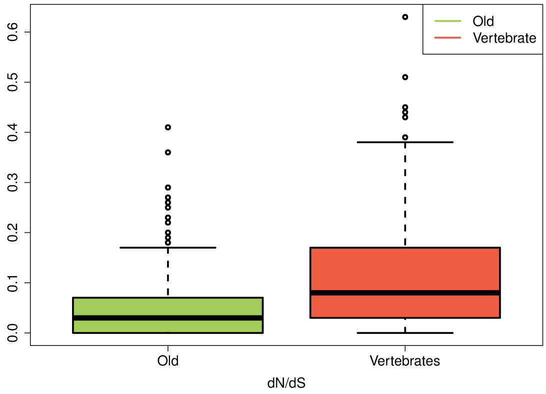


**Figure S9.** Distribution of non-synonymous to synonymous ($d_N/d_S$) substitution rates for domains of different age combined in the same protein. This data correspond to 285 proteins in which the Vertebrate domain showed less than 20% identity with D.melanogaster or C.elegans homologs. Differences in $d_N/d_S$ between Old and Vertebrate domains were statistically highly significant (Wilcoxon test, p$<10^{-5}$).

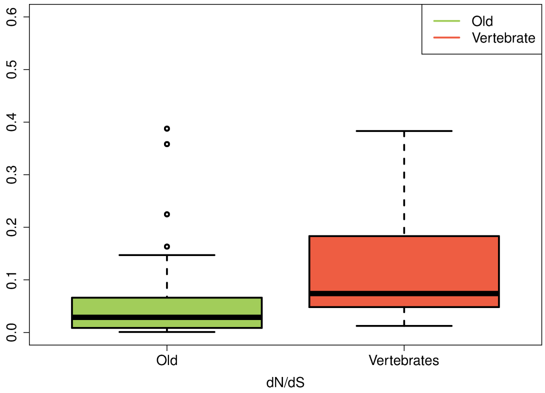


**Figure S10.** Distribution of non-synonymous to synonymous ($d_N/d_S$) substitution rates for domains of different age combined in the same protein. This data correspond to 69 proteins in which the Vertebrate domain showed less than 10% identity with *D.melanogaster* or *C.elegans homologs*. Differences in $d_N/d_S$ between Old and Vertebrate domains were statistically highly significant (Wilcoxon test, $p<10^{-5}$).

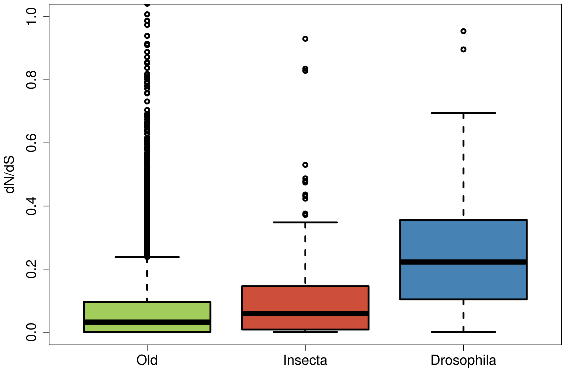


**Figure S11.** Distribution of non-synonymous to synonymous ($d_N/d_S$) values for *D.melanogaster* protein domains classified in different age groups. Differences between pairs of groups were significant (Kolmogorov-Smirnov test, $p<10^{-5}$). We obtained 11,013 one to one orthologous genes from *D.melanogaster* and *D.simulans*, using Ensembl (Hubbard et al., 2009). Using domain searches in other proteomes we classified D.melanogaster domains in the following groups: Drosophilids (*D. simulans*, *D. yakuba*, *D. erecta*, *D. pseudobscura*, *D. viriliae*, *D. Grishewi*), non-Drosophila Insecta (*Anopheles gambiae*, *Apis mellifera*, *Acyrthosiphon pisum*), other metazoans (*Takifugu rubripes*, *Homo sapiens*, *Ciona intestinalis*, *C.elegans*) and other eukaryotes (*Arabidopsis thaliana*, *Oryza sativa*, *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*). These proteomes were downloaded from Ensembl (Hubbard et al., 2009) and Uniprot (Jain et al., 2009). We classified 1,994 different *D.melanogaster* domains as Eukarya, 564 as Metazoan, 30 as Insecta and 22 as Drosophila.

| Age | N domain occurrences | N Domain types | Average | $d_N/d_S$ | $d_N$ | $d_S$ |
|---|---|---|---|---|---|---|
| Old | 6844 | 2185 | Mean | 0.138 | 0.012 | 0.151 |
| | | | Median | 0.032 | 0.004 | 0.138 |
| Insecta | 278 | 25 | Mean | 0.114 | 0.016 | 0.177 |
| | | | Median | 0.059 | 0.009 | 0.158 |
| Drosophila | 66 | 16 | Mean | 0.274 | 0.036 | 0.147 |
| | | | Median | 0.222 | 0.027 | 0.132 |

**Table S12.** Relationship between evolutionary rates and protein domain age in *D.melanogaster proteins*. Non-synonymous ($d_N$) and synonymous ($d_S$) substitution rates, calculated for *D.melanogaster* and *D.simulans*, corresponding to domains classified in different age classes (Old, Insecta, Drosophila). $d_N$ and $d_S$ statistics are calculated for all domain ocurrences. N domain types refer to the number of non-redundant domains

## 7 References

Albà, M. M. and Castresana, J. (2005). Inverse relationship between evolutionary rate and age of mammalian genes. *Molecular biology and evolution*, 22(3):598–606.

Albà, M. M. and Castresana, J. (2007). On homology searches by protein Blast and the characterization of the age of genes. *BMC evolutionary biology*, 7:53.

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research*, 25(17):3389–402.

Anckar, J. and Sistonen, L. (2011). Regulation of HSF1 Function in the Heat Stress Response: Implications in Aging and Disease. *Annual review of biochemistry*, 80:1089–115.

Björklund, A. K., Ekman, D., Light, S., Frey-Skött, J., and Elofsson, A. (2005). Domain rearrangements in protein evolution. *Journal of molecular biology*, 353(4):911–23.

Buljan, M., Frankish, A., and Bateman, A. (2010). Quantifying the mechanisms of domain gain in animal proteins. *Genome biology*, 11(7):R74.

Cai, J. J. and Petrov, D. A. (2010). Relaxed purifying selection and possibly high rate of adaptation in primate lineage-specific genes. *Genome biology and evolution*, 2:393–409.

Cai, J. J., Woo, P. C. Y., Lau, S. K. P., Smith, D. K., and Yuen, K.-Y. (2006). Accelerated evolutionary rate may be responsible for the emergence of lineage-specific genes in ascomycota. *Journal of molecular evolution*, 63(1):1–11.

Capra, J. A., Pollard, K. S., and Singh, M. (2010). Novel genes exhibit distinct patterns of function acquisition and network integration. *Genome biology*, 11(12):R127.

Chen, S., Zhang, Y. E., and Long, M. (2010). New Genes in Drosophila Quickly Become Essential. *Science*, 330(6011):1682–1685.

Choi, I.-G. and Kim, S.-H. (2006). Evolution of protein structural classes and protein sequence families. *Proceedings of the National Academy of Sciences of the United States of America*, 103(38):14056–61.

Daubin, V. and Ochman, H. (2004). Bacterial genomes as new gene homes: the genealogy of ORFans in E. coli. *Genome research*, 14(6):1036–42.

Domazet-Loso, T. and Tautz, D. (2003). An evolutionary analysis of orphan genes in Drosophila. *Genome research*, 13(10):2213–9.

Doolittle, R. F., Feng, D. F., Johnson, M. S., and McClure, M. A. (1986). Relationships of human protein sequences to those of other organisms. *Cold Spring Harbor symposia on quantitative biology*, 51 Pt 1:447–55.

Drummond, D. A., Bloom, J. D., Adami, C., Wilke, C. O., and Arnold, F. H. (2005). Why highly expressed proteins evolve slowly. *Proceedings of the National Academy of Sciences of the United States of America*, 102(40):14338–43.

Drummond, D. A., Raval, A., and Wilke, C. O. (2006). A single determinant dominates the rate of yeast protein evolution. *Molecular biology and evolution*, 23(2):327–37.

Drummond, D. A. and Wilke, C. O. (2008). Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell*, 134(2):341–52.

Duret, L. and Mouchiroud, D. (2000). Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate. *Molecular biology and evolution*, 17(1):68–74.

Eddy, S. R. (1998). Profile hidden Markov models. *Bioinformatics (Oxford, England)*, 14(9):755–63.

Ekman, D., Björklund, A. K., and Elofsson, A. (2007). Quantification of the elevated rate of domain rearrangements in metazoa. *Journal of molecular biology*, 372(5):1337–48.

Ekman, D., Björklund, A. K., Frey-Skött, J., and Elofsson, A. (2005). Multi-domain proteins in the three kingdoms of life: orphan domains and other unassigned regions. *Journal of molecular biology*, 348(1):231–43.

Elhaik, E., Sabath, N., and Graur, D. (2006). The ”inverse relationship between evolutionary rate and age of mammalian genes” is an artifact of increased genetic distance with rate of evolution and time of divergence. *Molecular biology and evolution*, 23(1):1–3.

Finn, R. D., Tate, J., Mistry, J., Coggill, P. C., Sammut, S. J., Hotz, H.-R., Ceric, G., Forslund, K., Eddy, S. R., Sonnhammer, E. L. L., and Bateman, A. (2008). The Pfam protein families database. *Nucleic acids research*, 36(Database issue):D281–8.

Fong, J. H., Geer, L. Y., Panchenko, A. R., and Bryant, S. H. (2007). Modeling the evolution of protein domain architectures using maximum parsimony. *Journal of molecular biology*, 366(1):307–15.

Frost, E. E., Zhou, Z., Krasnesky, K., and Armstrong, R. C. (2009). Initiation of oligodendrocyte progenitor cell migration by a PDGF-A activated extracellular regulated kinase (ERK) signaling pathway. *Neurochemical research*, 34(1):169–81.

Goodman, M., Moore, G. W., and Matsuda, G. (1975). Darwinian evolution in the genealogy of haemoglobin. *Nature*, 253(5493):603–8.

Hartl, D. L., Dykhuizen, D. E., and Dean, A. M. (1985). Limits of adaptation: the evolution of selective neutrality. *Genetics*, 111(3):655–74.

Hirsh, A. E. and Fraser, H. B. (2001). Protein dispensability and rate of evolution. *Nature*, 411(6841):1046–9.

Hubbard, T. J. P., Aken, B. L., Ayling, S., Ballester, B., Beal, K., Bragin, E., Brent, S., Chen, Y., Clapham, P., Clarke, L., Coates, G., Fairley, S., Fitzgerald, S., Fernandez-Banet, J., Gordon, L., Graf, S., Haider, S., Hammond, M., Holland, R., Howe, K., Jenkinson, A., Johnson, N., Kahari, A., Keefe, D., Keenan, S., Kinsella, R., Kokocinski, F., Kulesha, E., Lawson, D., Longden, I., Megy, K., Meidl, P., Overduin, B., Parker, A., Pritchard, B., Rios, D., Schuster, M., Slater, G., Smedley, D., Spooner, W., Spudich, G., Trevanion, S., Vilella, A., Vogel, J., White, S., Wilder, S., Zadissa, A., Birney, E., Cunningham, F., Curwen, V., Durbin, R., Fernandez-Suarez, X. M., Herrero, J., Kasprzyk, A., Proctor, G., Smith, J., Searle, S., and Flicek, P. (2009). Ensembl 2009. *Nucleic acids research*, 37(Database issue):D690–7.

Jain, E., Bairoch, A., Duvaud, S., Phan, I., Redaschi, N., Suzek, B. E., Martin, M. J., McGarvey, P., and Gasteiger, E. (2009). Infrastructure for the life sciences: design and implementation of the UniProt website. *BMC bioinformatics*, 10:136.

Krylov, D. M., Wolf, Y. I., Rogozin, I. B., and Koonin, E. V. (2003). Gene loss, protein sequence divergence, gene dispensability, expression level, and interactivity are correlated in eukaryotic evolution. *Genome research*, 13(10):2229–35.

Kuo, C.-H. and Kissinger, J. C. (2008). Consistent and contrasting properties of lineage-specific genes in the apicomplexan parasites Plasmodium and Theileria. *BMC evolutionary biology*, 8:108.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., and et al. (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921.

Löytynoja, A. and Goldman, N. (2008). Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. *Science (New York, N.Y.)*, 320(5883):1632–5.

Luz, H., Staub, E., and Vingron, M. (2006). About the interrelation of evolutionary rate and protein age. *Genome informatics. International Conference on Genome Informatics*, 17(1):240–50.

Marsh, J. A. and Teichmann, S. A. (2010). How do proteins gain new domains? *Genome biology*, 11(7):126.

Müller, A., MacCallum, R. M., and Sternberg, M. J. E. (2002). Structural characterization of the human proteome. *Genome research*, 12(11):1625–41.

Notredame, C., Higgins, D. G., and Heringa, J. (2000). T-Coffee: A novel method for fast and accurate multiple sequence alignment. *Journal of molecular biology*, 302(1):205–17.

Pál, C., Papp, B., and Hurst, L. D. (2001). Highly expressed genes in yeast evolve slowly. *Genetics*, 158(2):927–31.

Pal, L. R. and Guda, C. (2006). Tracing the origin of functional and conserved domains in the human proteome: implications for protein evolution at the modular level. *BMC evolutionary biology*, 6:91.

R Development Core Team (2008). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna (Austria).

Tautz, D. and Domazet-Lošo, T. (2011). The evolutionary origin of orphan genes. *Nature Reviews Genetics*, 12(10):692–702.

Toll-Riera, M., Bosch, N., Bellora, N., Castelo, R., Armengol, L., Estivill, X., and Albà, M. M. (2009). Origin of primate orphan genes: a comparative genomics approach. *Molecular biology and evolution*, 26(3):603–12.

Vishnoi, A., Kryazhimskiy, S., Bazykin, G. a., Hannenhalli, S., and Plotkin, J. B. (2010). Young proteins experience more variable selection pressures than old proteins. *Genome research*, 20(11):1574–81.

Vogel, C., Bashton, M., Kerrison, N. D., Chothia, C., and Teichmann, S. A. (2004). Structure, function and evolution of multidomain proteins. *Current opinion in structural biology*, 14(2):208–16.

Wall, D. P., Hirsh, A. E., Fraser, H. B., Kumm, J., Giaever, G., Eisen, M. B., and Feldman, M. W. (2005). Functional genomic analysis of the rates of protein evolution. *Proceedings of the National Academy of Sciences of the United States of America*, 102(15):5483–8.

Wardell, S. E., Boonyaratanakornkit, V., Adelman, J. S., Aronheim, A., and Edwards, D. P. (2002). Jun dimerization protein 2 functions as a progesterone receptor N-terminal domain coactivator. *Molecular and cellular biology*, 22(15):5451–66.

Wilson, A. C., Carlson, S. S., and White, T. J. (1977). Biochemical evolution. *Annual review of biochemistry*, 46:573–639.

Wolf, M. Y., Wolf, Y. I., and Koonin, E. V. (2008). Comparable contributions of structural-functional constraints and expression level to the rate of protein sequence evolution. *Biology direct*, 3:40.

Wolf, Y. I., Novichkov, P. S., Karev, G. P., Koonin, E. V., and Lipman, D. J. (2009). The universal distribution of evolutionary rates of genes and distinct characteristics of eukaryotic genes of different apparent ages. *Proceedings of the National Academy of Sciences of the United States of America*, 106(18):7273–80.

Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Molecular biology and evolution*, 24(8):1586–91.

Zhang, L. and Li, W.-H. (2004). Mammalian housekeeping genes evolve more slowly than tissue-specific genes. *Molecular biology and evolution*, 21(2):236–9.

Zhou, Q., Zhang, G., Zhang, Y., Xu, S., Zhao, R., Zhan, Z., Li, X., Ding, Y., Yang, S., and Wang, W. (2008). On the origin of new genes in Drosophila. *Genome research*, 18(9):1446–55.

Zuckerkandl, E. (1976). Evolutionary processes and evolutionary noise at the molecular level. I. Functional density in proteins. *Journal of molecular evolution*, 7(3):167–83.

### 3.2.2 Structure and Age Jointly Influence Rates of Protein Evolution

**Authors:** Macarena Toll-Riera, David Bostick, M.Mar Albà and Joshua B. Plotkin

**Published in:** Submitted

**Summary**

What factors determine a protein's rate of evolution are still under debate. Especially unclear is the relative role of intrinsic factors of present-day proteins versus historical factors such as protein age. Here we study the interplay of structural properties and evolutionary age, as determinants of protein evolutionary rate. We use a large set of one-to-one orthologs between human and mouse proteins, with mapped PDB structures. We report that previously observed structural correlations also hold within each age group - including relationships between solvent accessibility, designabililty, and evolutionary rates. However, age also plays a crucial role: age modulates the relationship between solvent accessibility and rate, and younger proteins, despite of being less designable, are evolving faster than older proteins. We show that previously reported relationships between age and rate cannot be explained by structural biases among age groups. Finally, we introduce a knowledge-based potential function to study the stability of proteins through large-scale computation. We find that older proteins are more stable for their native structure, and also more robust to mutations, than younger ones. Our results underscore that several determinants, both intrinsic and historical, can interact to determine rates of protein evolution.

# Structure and Age Jointly Influence Rates of Protein Evolution

Macarena Toll-Riera [1,2] , David Bostick [2] , M.Mar Albà [1,3*] and Joshua B. Plotkin [2*]

**Abstract**

What factors determine a protein's rate of evolution are still under debate. Especially unclear is the relative role of intrinsic factors of present-day proteins versus historical factors such as protein age. Here we study the interplay of structural properties and evolutionary age, as determinants of protein evolutionary rate. We use a large set of one-to-one orthologs between human and mouse proteins, with mapped PDB structures. We report that previously observed structural correlations also hold within each age group - including relationships between solvent accessibility, designabililty, and evolutionary rates. However, age also plays a crucial role: age modulates the relationship between solvent accessibility and rate, and younger proteins, despite of being less designable, are evolving faster than older proteins. We show that previously reported relationships between age and rate cannot be explained by structural biases among age groups. Finally, we introduce a knowledge-based potential function to study the stability of proteins through large-scale computation. We find that older proteins are more stable for their native structure, and also more robust to mutations, than younger ones. Our results underscore that several determinants, both intrinsic and historical, can interact to determine rates of protein evolution.

## 1 Introduction

It is well known that protein evolutionary rates are not homogeneous, with as much variation within an organism as between organisms. In fact, evolutionary rates vary as much as 1,000-fold among the proteins in the yeast *S. cerevisiae* (Drummond et al., 2005). Therefore, there has been longstanding interest in deciphering the causes of this variation, with a large literature of theoretical and empirical studies alike.

Numerous possible determinants for protein evolutionary rate have been proposed, such as protein dispensability (Hirsh and Fraser, 2001), number of mRNA molecules per cell (Green et al., 1993; Pál et al., 2001), number of protein molecules per cell (Drummond et al., 2006), codon adaptation index (Pál et al., 2001; Wall et al., 2005), number of protein-protein interactions (Fraser et al., 2002), sequence length (Marais and Duret, 2001; Lipman et al., 2002), role in the interaction network (Hahn and Kern, 2005), and structural properties such as solvent accessibility and folding robustness (Bloom et al., 2006a; Franzosa and Xia, 2009; Lobkovsky et al., 2010). Some of the proposed determinants are correlated with one another, which makes the identification of causal factors difficult. For this reason Drummond and colleagues (Drummond et al., 2006) tried to disentangle these factors by performing a

[1] Evolutionary Genomics Group, Research Programme on Biomedical Informatics (GRIB), Hospital del Mar Research Institute (IMIM) - Universitat Pompeu Fabra (UPF), Barcelona, Spain

[2] Department of Biology, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA

[3] Catalan Institution for Research and Advanced Studies (ICREA), Barcelona, Spain

* To whom correspondence should be addressed

Corresponding authors contact information:
**M.Mar Albà**
ICREA Research Professor, Research Programme on Biomedical Informatics (GRIB), Hospital del Mar Research Institute (IMIM) - Universitat Pompeu Fabra (UPF), Barcelona 08003, Spain
e-mail: malba@imim.es
**Joshua B. Plotkin**
Department of Biology, University of Pennsylvania, 433 S. University Ave, Philadelphia PA 19104
e-mail: jplotkin@sas.upenn.edu

principal component regression (PCR) analysis. They found that a single component, which included codon adaptation index, protein abundance and gene expression level, accounted for nearly half of the observed variability in protein's evolution. Nonetheless, those expression-related factors have been measured with less noise than other possible factors, which further complicates even the principal component regression (Plotkin and Fraser, 2007). In related work, Drummond and Wilke (Drummond and Wilke, 2008) observed covariation between sequence evolution, codon usage and mRNA level among a broad range of species. They suggested there may be selection for robustness against mistranslation, since mistranslation-induced misfolding would be more deleterious for highly expressed proteins.

A protein's three-dimensional structure may also be a key factor in determining its evolutionary rate. The core of a protein is mostly formed by buried residues, which often play a crucial role in the stability of the folded structure (Franzosa and Xia, 2008). Most mutations in the core of a protein tend to destabilize the protein (Pál et al., 2006), and it known that exposed residues evolve faster than buried ones (Goldman et al., 1998; Mirny and Shakhnovich, 1999; Bustamante et al., 2000; Bloom et al., 2006a; Conant and Stadler, 2009; Franzosa and Xia, 2009). In fact, the more general relationship between solvent exposure and evolutionary rate is linear and very strong (Franzosa and Xia, 2009). Given these results, we might expect those proteins with a higher fraction of exposed residues to evolve faster; but, surprisingly, Bloom and others found the contrary pattern (Bloom et al., 2006a; Franzosa and Xia, 2009). Bloom et al explained this incongruence using protein designability, defined roughly as the number of sequences than can fold into a structure. Since a higher number of sequences can fold into highly designable structures, designable structures are more tolerant to mutations and hence, evolve faster. As designability has been related to contact density (England and Shakhnovich, 2003) and contact density is highly correlated with the fraction of buried residues, the authors hypothesize that highly designable proteins have a higher fraction of buried residues and consequently have stable core, allowing the exposed residues to freely mutate without compromising stability. In fact, Franzosa and Xia (Franzosa and Xia, 2009) have demonstrated how large-core proteins (which are the ones having an overall low solvent exposure value) have low solvent exposure values but high $d_N/d_S$, specially observing that highly exposed residues in large-core proteins are evolving faster than in small-core proteins. Also, proteins with a higher contact density tend to evolve more rapidly – in fly, yeast, *E.coli* and human (Zhou et al., 2008). Additionally, highly designable proteins have been shown to evolve more functional innovations (Ferrada and Wagner, 2008). Bloom and colleagues (Bloom et al., 2006a) have also carried out a PCR analysis showing that the component measuring expression level could explain around 34% of the rate variation, whereas structural characteristics explained approximately the 10% of the rate variation. Other structural properties are also correlated with evolutionary rates, such as the number of intra-protein residue interactions, which tend to reduce rates of evolution (Toft and Fares, 2010). Structure itself could be a determinant of protein evolution, or indeed, could be acting through other mechanisms, for example, it could play a crucial role in the selection for structural robustness against mistranslation in highly expressed proteins, which has already been shown to be a key determinant of protein evolution (Bloom et al., 2006a).

Quite aside from the factors discussed above, which are intrinsic to the properties of a protein in an organism today, studies have also shown that the age of a protein, which depends on its evolutionary history, is also correlated with evolutionary rates (Albà and Castresana, 2005; Wolf et al., 2009; Vishnoi et al., 2010). In particular, an inverse relationship between age and evolutionary rate has been widely observed (Domazet-Loso and Tautz, 2003; Albà and Castresana, 2005; Wolf et al., 2009), suggesting that a protein's evolution could be shaped in part by its evolutionary origin. This relationship has been reported in a broad range of organisms: primates (Toll-Riera et al., 2009), mammals (Albà and Castresana, 2005), *Drosophila* (Domazet-Loso and Tautz, 2003; Wolf et al., 2009), *Plasmodium* (Kuo and Kissinger, 2008), fungi (Cai et al., 2006) and bacteria (Daubin and Ochman, 2004).

Despite all these findings, what factors determine a protein's evolutionary rate are still under debate - and the relative role of intrinsic factors of present-day proteins, versus historical factors such as protein age, remains poorly characterized. Here we explored the interplay between two very different factors: a protein's age and its structural properties. Our objective is to determine whether structural biases among age groups could explain the reported differences in evolutionary rates with age (Albà and Castresana, 2005; Wolf et al., 2009). To do so we used a dataset of human proteins with homologues in mouse for which we were able to map a PDB structure. Age was assigned to each PDB structure and then structural properties (solvent exposure, designability, stability and secondary structure) were calculated among the PDB structures classified in the age groups. We found that differences in evolutionary rates previously observed among age groups could not be explained due to differences in the structural properties among age groups; similarly, differences in rates correlated with structural differences cannot be entirely explained by the age of the PDB structure, although a marginal influence of age is observed. Our results therefore reinforce the idea that there is not a single determinant of evolutionary rate, and that both intrinsic present-day properties as well as evolutionary age independently contribute to differential rates of protein evolution.

## 2 Results

### 2.1 Interactions between age and structural determinants of evolutionary rates

It has been widely argued that both protein structure and protein age play important roles as determinants of protein evolution. However, how structure and protein age are related has not been yet studied. We have found an interesting interplay between structure and age: a set of structural characteristics that are correlated with evolutionary rates, but in a manner that depends on protein age.

#### 2.1.1 Linear relationship between solvent-accessibility and evolutionary rate

We calculated the relative solvent accessibility (RSA) for each residue in every PDB structure that mapped to human proteins (406,970 residues in total, across 2,595 PDB structures). We apportioned the RSA values into 20 bins and we concatenated all the residues within each bin to calculate the evolutionary rate (measured as $d_N$) of residues as a function of accessibility. We found a strong correlation between solvent accessibility and $d_N$ (Pearson correlation: 0.971, p-value=1.179 e$^{-12}$) in mammals (supplementary file S1), which is similar to the linear correlation between evolutionary rate and solvent accessibility previously reported in yeast (Franzosa and Xia, 2009), suggesting that this relationship is an universal trend.

Additionally, we separated the PDB structures according to their age (i.e. the youngest proteins, which originated in Vertebrates, the medium-aged proteins which originated in Metazoans, and the oldest proteins which originated in Eukaryotes) we found a similar correlation between accessibility and evolutionary rate within each age group (Pearson correlation $>$0.94 and p-value $<10^{-10}$ in all the age groups) (figure 1). But, interestingly, the slope is different among age groups: the younger proteins show a more dramatic influence of solvent accessibility on evolutionary rate. For the linear model $d_N$ $\sim$RSA, the slope in Eukarya is 0.0025; for Metazoans and Vertebrates, it is 0.003 and 0.006, respectively. We also considered an interaction term of RSA with age ($d_N$ $\sim$RSA + RSA*age + age) in all the possible pairwise comparisons between age groups, in order to assess the importance of age. The interaction was generally significant (Eukarya vs Metazoans: 0.11, Eukarya vs Vertebrates: 1.70e$^{-07}$, Metazoans vs Vertebrates: 4.73e$^{-06}$) supporting the notion that age plays a role in shaping the relationship between solvent accessibility and evolutionary rate.

#### 2.1.2 Fraction of residues exposed and designability

Given the linear relationship between solvent accessibility and evolutionary rates one expects to find that those structures containing a higher number of exposed residues would be evolving faster. But Bloom and colleagues (Bloom et al., 2006a) have found exactly the contrary: the fraction of buried residues in a protein is positively correlated with its evolutionary rate ($d_N$). Bloom et al explained this incongruence using the concept of protein designability, as discussed above. However, there are discrepancies regarding the relationship between contact density (or fraction of buried residues) and evolutionary rate. Shakhnovich (Shakhnovich, 2006) found, in yeast and *C.elegans*, a negative correlation and Lin and colleagues (Lin et al., 2007) found a negative correlation when they used predictions based on support-vector machine and no correlation when they calculated the fraction of buried residues directly from the crystal structures. These discrepancies are probably due to methodological differences between studies. Here we have been more stringent than in earlier studies, using 99% sequence identity to assign structure as compared with the 40% criteria used in Zhou et al (Zhou et al., 2008).

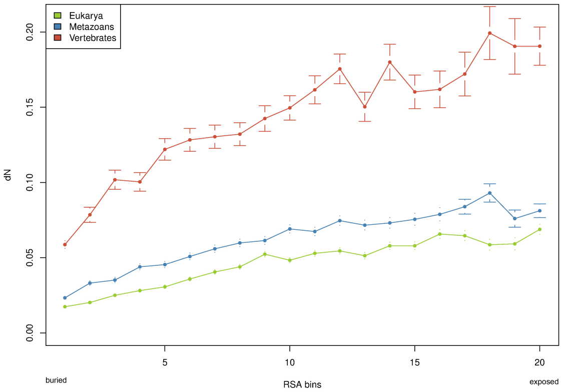


**Figure 1.** Linear relationship between solvent accessibility and $d_N$ in Eukarya, Metazoans and Vertebrates age groups. Eukarya: Pearson correlation: 0.957, p-value= 4.477e$^{-11}$; Metazoans: Pearson correlation: 0.950, p-value =1.445e$^{-10}$; Vertebrates: Pearson correlation: 0.941, p-value= 7.005e$^{-10}$. Errors bars indicate the standard error for the dN calculation.

We tested the impact of designability in the context of PDB structures classified by their age of origin. We first calculated the evolutionary rate ($d_N$) of each PDB structure as well as the fraction of residues exposed (exposed residues/(buried+exposed residues) *100). We found that the oldest Eukaryotic PDB structures were evolving the slowest, followed by Metazoans and then Vertebrates (Wilcoxon tests, p-value $<10^{-3}$ in all the pairwise comparisons), which confirms the inverse relationship between protein age and evolutionary rate that has been reported previously (Albà and Castresana, 2005) (supplementary file S2). Besides, older folds have been previously reported to be more conserved than younger ones (Wong and Frishman, 2006). At the same time, we found that younger PDB structures have a significantly higher fraction of exposed residues than older ones (Wilcoxon tests, p-value $<10^{-3}$ in all the pairwise comparisons) (figure 2), despite the fact that the younger PDB structures evolve faster. This is contradictory with what has been found in Bloom et al. (Bloom et al., 2006a) and Franzosa et al. (Franzosa and Xia, 2009).

In an effort to disentangle this contradictory result we obtained for each age group the fastest ($d_N/d_S$ >0.1) and the slowest evolving PDB structures ($d_N/d_S$ = 0.001 in Eukarya and Metazoan and $d_N/d_S$ <0.1 in Vertebrates) and we checked their fraction of exposed residues. Within the three age groups we found that the fastest evolving PDB structures had a higher fraction of buried residues than the slowest ones (Wilcoxon test, Eukarya: p-value= 2.697e$^{-07}$, Metazoans: p-value= 0.004, Vertebrates: p-value= 0.05). Furthermore, among the fastest evolving PDB structures, the younger ones had a lower fraction of buried residues than the older ones (Wilcoxon test, Eukarya vs Metazoans: p-value=2.765e$^{-05}$, Eukarya vs Vertebrates: p-value=2.140e$^{-10}$, Metazoans vs Vertebrats: p-value=0.0008). Thus, while the impact of designability on evolutionary rate holds within each age class, it does not hold between age groups. Therefore, our results in part confirm those of Bloom et al. (Bloom et al., 2006a), at least within each age class, but they also suggest that protein age has a stronger overall effect on evolutionary rate than designability does.

### 2.2 Protein age, stability, and mutational robustness

An important, related question is whether protein stability depends on protein age. To quantify stability for the large set of proteins used in this study, we used a well-known coarse-grained four-body knowledge-based potential function (see Materials and Methods), described by Gan (Gan et al., 2001) and Krishnamoorthy (Krishnamoorthy and Tropsha, 2003). This potential has been shown to successfully score stability changes due to both mutational and structural protein alterations in

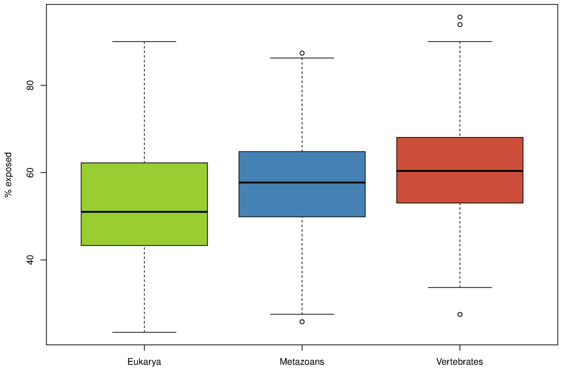


**Figure 2.** Percentage of residues exposed in PDB structures classified in 3 age groups: Eukarya, Metazoans and Vertebrates. Wilcoxon tests were performed to assess the significance of the difference: Eukarya vs Metazoans:p-value $<2.2e^{-16}$ , Eukarya vs Vertebrate:p-value $<2.2e^{-16}$ , Metazoans vs Vertebrates: p-value = 0.0005

a manner consistent with free energy changes derived from unfolding experiments (Deutsch and Krishnamoorthy, 2007; Carter et al., 2001). Thus, for convenience in what follows, we refer to the score of a given protein (conformation + sequence) as $\Delta$G (analogous to the free energy of folding: lower $\Delta$G implies greater stability). To validate our implementation of this potential, we tested its ability to distinguish native from misfolded decoy protein conformations (i.e., physically reasonable alternative protein conformations generated computationally from a native structure) taken from a standard database (Samudrala and Levitt, 2000).Our implementation of the score ranked native structures among their decoys in a manner consistent with (in some cases, more favorably than) previous work (Krishnamoorthy and Tropsha, 2003) (supplementary file S3).

As a secondary validation of our stability scoring function, we re-considered the correlation between RSA and evolutionary rate, described above. Given this empirical correlation, we should expect that mutations with a higher impact on the stability of the protein would tend to occur in the residues that are more buried. To test this computationally, for every protein in our PDB data set, we mutated each residue to a randomly selected residue while holding all other residue identities fixed. Then, we classified each residue in a bin according to the impact of the mutation on the stability score relative to the native sequence (using the absolute value, $\Delta\Delta$G, where $|\Delta\Delta G|=|\Delta G(\text{native})-\Delta G(\text{mutant})|$ – larger values imply greater absolute perturbations to the stability). We found that the residues with less solvent accessibility exhibited significantly greater impacts on computed stability when mutated, in accordance with expectation (supplementary file S4).

We used the potential function to score the overall stability, measured as $\Delta$G, for each PDB structure. To control for any length dependence in the score (a correlation between length and contact density has already been reported (Bloom et al., 2006a), we binned the lengths of all proteins to obtain a set of structures with the exact same length distribution within each age class. In doing so, however, we were not able to retain enough Vertebrate PDB structures for further analysis, and so restricted our comparisons to Eukarya and Metazoans. When we compared $\Delta$G amongst Eukarya and Metazoans, paired by length bin, we found that Eukaryotic structures are more stable on average (Wilcoxon-paired test, p-value <0.01, Eukarya median: -90.74, Metazoan median: -85.08). This suggests that older proteins are more stable, on average, than younger proteins.

Furthermore, we studied how mutational robustness might vary with protein age. To estimate robustness we simulated random amino-acid mutations in 2% of the residues of each PDB structure, and we repeated this process 1000 times for each structure (supplementary file S5). We then used two measures, Z-score and Rank, to assess how robust the native structure is to mutation. The Z-score was calculated for each protein as the protein's stability score minus the mean score for the population of 1000 mutated structures divided by the its standard deviation, $\sigma$, $(Z=(\Delta G-\langle\Delta G\rangle)/\sigma)$. Younger PDB structures were significantly less robust to mutations (higher Z-score) than older proteins (Wilcoxon test, Eukarya vs Metazoans p-value $<10^{-15}$, Eukarya vs Vertebrates p-value $<10^{-14}$, Metazoans vs Vertebrates p-value= 0.131). We also computed the rank of each native protein score within the population of 1000 mutant scores; and we found the same trend: the native sequence-structure compatibility of younger proteins was significantly less robust (higher rank) than that of older proteins (Wilcoxon test, p-value $<10^{-7}$ in all the pairwise comparisons) (figure 3). Similar results were obtained when we increased the mutation rate to 10% of residues within each PDB structure (data not shown).

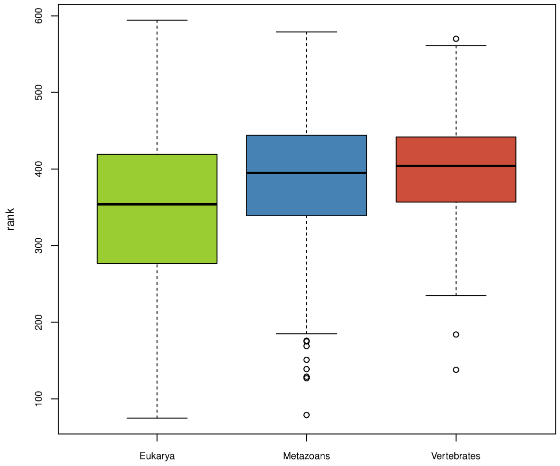


**Figure 3.** Rank of the stability score of wildtype protein sequence among 1000 mutated sequences in 3 age groups: Eukarya, Metazoans and Vertebrates. Wilcoxon tests were performed to assess the significance of the difference: Eukarya vs Metazoans:p-value $<1.684e^{-14}$ , Eukarya vs Vertebrate:p-value $<2.2e^{-16}$ , Metazoans vs Vertebrates:p-value = $1.119e^{-8}$). Low rank suggests that the native structure is relatively robust to mutations.

More designable proteins are generally more stable (Wingreen et al., 2003) and have a higher fraction of buried residues, which may lead to a more robust protein core. It has been shown that stability generally enhances tolerance to mutations - more beneficial mutations are accepted because they do not destabilize the native structure (Bloom et al., 2005, 2006b). Thus, our results on the greater stability and robustness of older proteins generally concord with earlier notions of designability and mutational tolerance.

## 2.3 Protein Age and secondary structure

We also investigated the relationships between protein age, secondary structure classification, and evolutionary rates. We classified each residue in every PDB structure according to the type of secondary structure in which it participates as well as according to whether it is buried (RSA <25%) or exposed (RSA >25%) as in Bloom et al. (Bloom et al., 2006a). Each residue was mapped to one of four secondary structure categories by DSSP (Kabsch and Sander, 1983): helix (class H in DSSP), sheet (class E in DSSP), turn (classes S and T), coil (classes B, G, I and '.'). Evolutionary rates within each structural category were computed by concatenating, for each PDB structure, all the residues classified in a given structural category and comparing those residue positions to homologous positions in mouse.

Generally, we found that exposed residues evolved faster than buried ones (Wilcoxon test, p-value <0.01) and that residues classified as helix evolve slower (Wilcoxon test, p-value <0.01) than the residues classified in other categories (supplementary file S6). More importantly, when we separated the secondary structures and solvent accessibility according to age group we found that the younger structures were evolving faster than the older ones (Wilcoxon test, table 1, figure 4 ) within each structural category. This implies that differences in the frequency of structural categories by age class cannot explain the previously reported inverse relationship between protein age and evolutionary rate (Albà and Castresana, 2005). Thus, this analysis supports the important role for protein age in shaping evolutionary rates, above and beyond the influence of solvent accessibility and secondary structure.

| Secondary structure | Age | $d_N/d_S$ | $d_N$ |
|---|---|---|---|
| Helix | Eukarya-Metazoan | 0.929 | 0.6 |
| | Eukarya-Vertebrates | $5.286e^{-06}$ | $5.188e^{-06}$ |
| | Metazoans-Vertebrates | $4.771e^{-05}$ | $5.74e^{-05}$ |
| Sheet | Eukarya-Metazoan | 0.048 | 0.009 |
| | Eukarya-Vertebrates | $2.737e^{-08}$ | $2.521e^{-09}$ |
| | Metazoans-Vertebrates | $3.057e^{-05}$ | $4.129e^{-05}$ |
| Turn | Eukarya-Metazoan | 0.4841 | 0.205 |
| Coil | Eukarya-Metazoan | 0.001 | 0.0002 |
| | Eukarya-Vertebrates | $3.070e^{-05}$ | $3.542e^{-06}$ |
| | Metazoans-Vertebrates | 0.01 | 0.005 |
| Exposed | Eukarya-Metazoan | 0.132 | 0.01 |
| | Eukarya-Vertebrates | $2.681e^{-16}$ | $<2.2e^{-16}$ |
| | Metazoans-Vertebrates | $7.402e^{-13}$ | $4.318e^{-12}$ |
| Buried | Eukarya-Metazoan | 0.066 | 0.005 |
| | Eukarya-Vertebrates | $<2.2e^{-16}$ | $<2.2e^{-16}$ |
| | Metazoans-Vertebrates | $3.713e^{-12}$ | $4.207e^{-12}$ |

**Table 1.** Comparisons between the 3 age classes in each secondary structure and solvent accessibility types

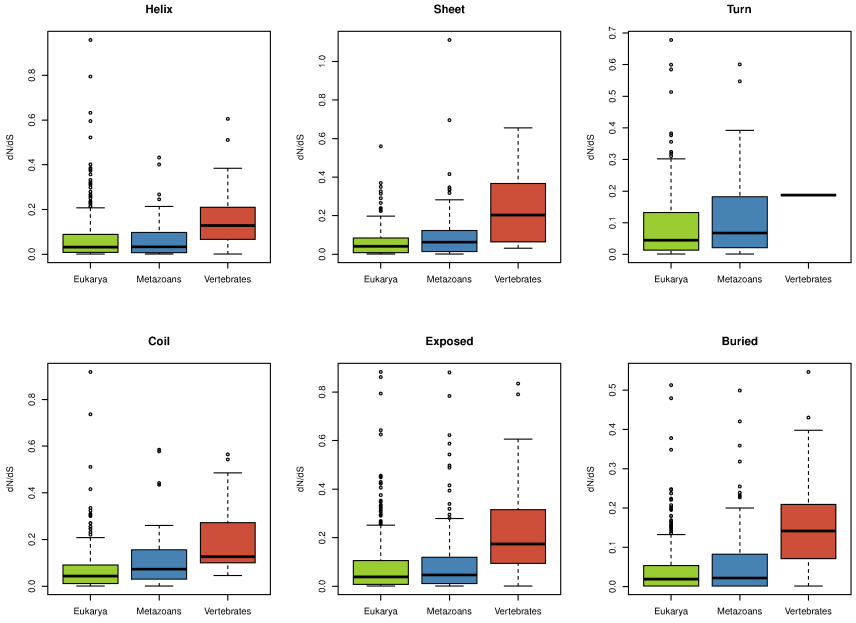


**Figure 4.** Evolutionary rates by age and secondary structure/solvent accessibility categories. An inverse correlation between the age of the protein and evolutionary rate occurs within each structural category. Wilcoxon tests were performed (see table 1).

## 3 Discussion

Interactions among various determinants of protein evolution are not well understood despite several decades of investigation. In this work, we have studied two types of proposed determinants: structural properties intrinsic to present-day proteins, and protein age. We found that several well-known relationships between structural properties and evolutionary rate that had previously been reported irregardless of age also hold within each age class: residues with high solvent accessibility evolve more quickly (Goldman et al., 1998; Mirny and Shakhnovich, 1999; Bustamante et al., 2000; Bloom et al., 2006a; Conant and Stadler, 2009; Franzosa and Xia, 2009), while proteins with a larger fraction of exposed residues evolve more slowly (Bloom et al., 2006a; Franzosa and Xia, 2009). At the same time, the age of a protein can modulate the effect of such structural properties on evolutionary rates - e.g. the strength of relationship between solvent accessibility and evolutionary rate depends on the age of the protein in which the residue is found. We also studied secondary structures of proteins, and we confirmed that the typical inverse relationship between protein age and evolutionary rate holds within each structural class of residues. This implies that differences in the frequency of structural categories by age class cannot explain the previously reported inverse relationship between age and rate. Finally, we introduced a knowledge-based potential to study the relationships between protein age and stability. We found that older proteins are more stable, on average, than younger proteins, and that older structures are also more robust to mutation than younger structures.

Our results provide a more nuanced view on the determinants of protein evolutionary rates. Whereas some determinants of rates hold within each age class, age can nonetheless modulate these effects. And other relationships that hold irregardless of age (such as, proteins with a greater fraction of exposed residues evolve more slowly) cannot explain differences in rates between age classes.

Our analyses certainly suffer from several drawbacks. Most important, we were able to map a structure to only 14% of the one-to-one orthologous proteins between human and mouse, and this fraction would be even smaller if we had chosen other species. Despite the increase in solved structures over the past few years, the number of mapped structures is still a small fraction of known proteins. Additionally, we have to bear in mind that there are biases in the type of proteins that enjoy solved structures. For example disordered regions are poorly represented in PDB, as they are difficult to crystallize. Younger proteins are enriched in low-complexity regions (Toll-Riera et al., 2011; Simon and Hancock, 2009), many of which are expected to be disordered (Simon and Hancock, 2009). How this adds to the differences in evolutionary rates between age classes is an aspect that remains to be studied.

Choi and Kim (Choi and Kim, 2006) have reported that old proteins are longer and have more complex tertiary structures ($\alpha/\beta$) than younger proteins, hypothesizing that proteins tend to become more complex in their structure along their evolutionary history. Our results also give insights on the evolution of protein structural characteristics, as we have found that older structures are more designable, stable and robust to mutations than younger ones. These findings suggest that structures may acquire stability and robustness to mutations with time. However, these findings also raise new questions. Since stability increases a protein's tolerance to mutations (Bloom et al., 2006b) we might expect that younger structures would be evolving slowly due to the destabilizing effect of mutations. But we find them to evolve fast. One possible explanation is that previous studies have assumed proteins are generally under the same degree of selection, regardless of age. But some of our results might be due to differential strengths of selection in old versus young proteins. We hypothesize that younger sequences mapped to PDB may be experiencing strong positive selection for stabilizing mutations, which explains their higher rates of evolution; whereas older protein are already stable and robust, and thus lack this type of positive selection. Thus, we propose that with time structures acquire stability and designability through the fixation of adaptive mutations. Using single nuclotide polymorphisms (SNP) data Cai and Petrov have found some evidence for increased positive selection in primate-specific genes, although they have also reported that relaxed negative selection is likely to be more important in young genes than in older genes (Cai and Petrov, 2010). In conclusion, our results reinforce the idea that protein evolution is not explained by a single determinant, but rather by the interplay of many determinants, including even factors that are not intrinsic to the present-day protein but depend on evolutionary age.

# 4 Material

## 4.1 Datasets

13494 orthologous one-to-one between *Homo sapiens* and *Mus musculus* were obtained from Ensembl (version 62) (Flicek et al., 2011). In order to assign a known structure to our proteins we performed BlastP searches (Altschul et al., 1997) between the structures deposited in the Protein Data Bank (Berman et al., 2000) and our dataset of human proteins with orthologous in mouse. We only kept those hits with an identity at least of 99%. If several hits were overlapping we chose the one that is closer to the human protein. Afterwards we applied a strong filtering process in which we discarded 506 PDB structures because they were shorter than 50 amino acids, they had discontinuous positions or they do not have C correctly annotated. We finally obtained 1899 proteins with a PDB structure mapped to them, covering a total of 2145 structures.

For each human protein region with a structure assigned we recorded the information regarding to the solvent-accessibility and the secondary structure. The information for the secondary structure and for solvent accessibility was obtained from the DSSP files (downloaded from http://srs.ebi.ac.uk/srsbin/cgi-bin/wgetz?-page+LibInfo+-lib+DSSP). We only recorded those positions in which there was the same amino acid in the human protein and in the PDB structure. Residues were classified in 4 secondary structures based on the DSSP (Kabsch and Sander, 1983) assignation for the residue: helix (class H in DSSP), sheet (class E in DSSP), turn (classes S and T) and coil (classes B, G, I and '.'), as in Bloom et al. (Bloom et al., 2006a). For each residue we calculated the solvent-accessibility as the RSA (relative solvent accessibility) which was obtained normalizing the accessibility obtained from DSSP by the reference solvent-accessible surface areas (ASA) of each amino acid. ASA is calculated for residue X in an extended Gly-X-Gly peptide; ASA values were obtained from Miller et al. (Miller et al., 1987). Residues were classified as buried if the RSA value was lower than 25% and as exposed if it was higher than 25%, as in Bloom et al. (Bloom et al., 2006a). Additionally we binned the RSA values in 20 bins, and we classified each residue in one of these RSA bins.

The fraction of exposed residues for a given PDB was calculated dividing the number of residues classified as exposed by the sum of the number of exposed and buried residues.

## 4.2 Age assignation

For each PDB structure we used BlastP searches with an e-value cut-off of $10^{-4}$ against several genomes to asses the presence of homologues. We used the following age classes: mammals (*Mus musculus*, *Rattus norvegicus*), non-mamalian vertebrates (*Gallus gallus*, *Xenopus tropicalis*, *Danio rerio*, *Takifugu rubripes*), other metazoans (*Ciona intestinalis*, *Drosophila melanogaser*, *Anopheles gambiae*, *Caenorhabditis elegans*) and other eukaryotes (*Schizosaccharomyces pombe*, *Saccharomyces cerevisiae*, *Oryza sativa*, *Arabidopsis thaliana*). Then, an age is assigned to each PDB chain according to the phylogenetic width of its homologues. We obtained 1157 PDB structures classified as eukarya, 725 as metazoan, 253 as vertebrate and 25 as mammals. As very few PDB structures were classified as mammals they were discarded for the analysis.

## 4.3 Evolutionary rates estimation

To estimate the evolutionary rates we only used those PDB structures in which the corresponding region in the human protein had at least 50% identity with its syntenic region in mouse. Pairwise alignments for the protein region corresponding to the PDB structure in human and in mouse were performed using T-Coffee (Notredame et al., 2000) and subsequently we obtained the nucleotide coding sequence alignment using an in-house Perl program.

To perform the secondary structure and the solvent-accessibility analysis we concatenated for each PDB region in the protein all the residues that were sharing the same type of secondary structure/solvent-accessibility, as long as the amino acid position in the protein was exactly the same as in the PDB structure. Then, for example, for a given protein region with a mapped PDB structure, we concatenated all the residues that were classified as helix and we took also the corresponding residues in mouse (as long as the mouse region homologous to human and human had at least a 50% of identity, which was accomplished in the majority of the cases), therefore, we constructed two new orthologous sequences with information corresponding only to one type of structure, helix in this case. These new sequences were aligned using T-coffee and realigned afterwards at nucleotide coding sequence level.

We additionally concatenated all the PDB residues classified in the same RSA bin and also all the residues that were classified in the same RSA bin and in the same age. The corresponding residue in mouse was also obtained. By doing that we obtained very long orthologous sequences that were aligned using MAFFT (Katoh et al., 2002).

To estimate the evolutionary rates we calculated the number of non-synonymous substitutions per non-synonymous site ($d_N$), the number of synonymous substitutions per synonymous site ($d_S$) and the $d_N/d_S$ ratio using the codeml program, which is inside the PAML software packages (Yang, 2007).

Several filters have been applied to the evolutionary rates estimations to ensure their robustness. Sequences shorter than 60 amino acids were discarded, as well as sequences with dN >0.5 and/or dS >2 which could be indicative of a lack of homology and of the presence of sequence saturation respectively.

### 4.4 Stability computations

To calculate the stability of the PDB structures we used a knowledge based potential, described by Gan (Gan et al., 2001) and Krishnamoorthy (Krishnamoorthy and Tropsha, 2003), that was trained on a nonredundant set of 3,425 X-ray protein structures downloaded from the PISCES database (Wang and Dunbrack, 2003) maintained by the Dunbrack laboratory. This set of proteins represented a subset of a list of 4,944 PDB chains that met strict parsing criteria (Krishnamoorthy and Tropsha, 2003). Each chain in the set shares no more than 25% sequence identity with any other chain, was resolved to <2.0 Angstroms, and solved with an R-factor of 0.25 or better. This type of potential has been widely validated (Deutsch and Krishnamoorthy, 2007; Masso et al., 2006).

We did two rounds of point mutations. In the first round we introduced 1 random mutation with random placement along the sequence for every 50 amino acids in the protein; in the second round, 1 random mutation with random placement along the sequence for every 10 amino acids. We repeated this process 1000 times for each PDB structure, obtaining 1000 mutated structures. For those structures obtained by NMR spectroscopy we used the first structural model presented in the PDB file. Then, we assessed the stability for the native PDB structure and mutated sequence using the potential, obtaining the measure, $\Delta$G, which describes the stability - lesser values imply more stability. We also calculated the destabilizing effect of mutations (robustness) using Z-score and Rank measures. The Z-score for a protein structure with specified sequence is calculated as ($Z=(\Delta G-\langle\Delta G\rangle)/\sigma$) , where $\langle\Delta G\rangle$ is the average stability score and $\sigma$ is the standard deviation in $\Delta$G derived from the 1000 mutated structures. The rank of the native sequence in these experiments is defined as the enumerated position of the native $\Delta$G value in the sorted list - from lowest (most stable) to highest (least stable) - of $\Delta$G values from the 1000 mutated structures.

To control for any possible dependence of the knowledge based potential score on protein length, we binned the PDB structures in our data set by length when comparing native $\Delta$G values for the proteins classified by age. In doing so, we ensure that our comparisons of stability across age grouped proteins are unbiased by protein length. Due to this binning, we lacked sufficient data to perform these comparisons for the representative Vertebrate PDB structures.

# 5 Supplementary Information

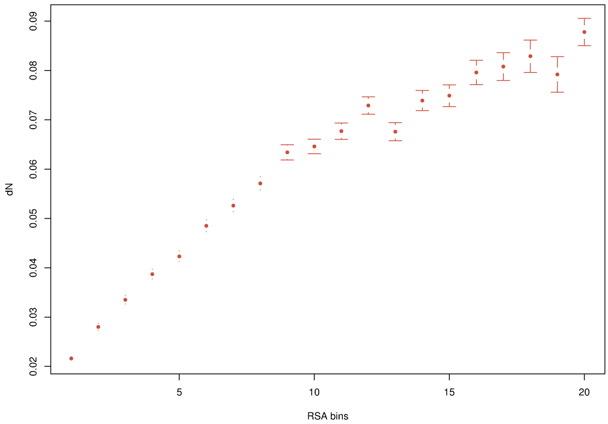


**Figure S1.** Linear correlation between $d_N$ and solvent accessibility (RSA) (Pearson correlation: 0.971, p-value=1.179 e$^{-12}$). RSA was separated in 20 bins and residues classified in the same bin were concatenated for all the PDBs to calculate the evolutionary rates.

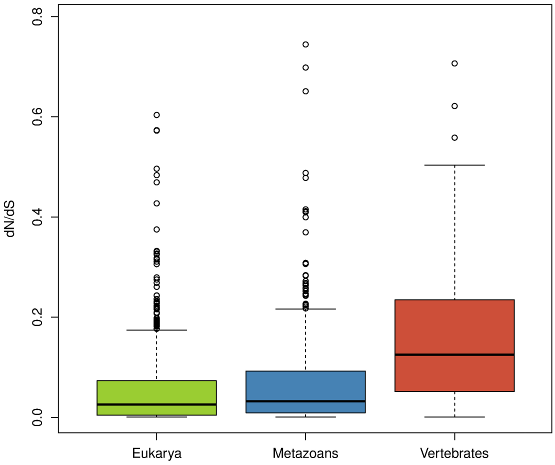
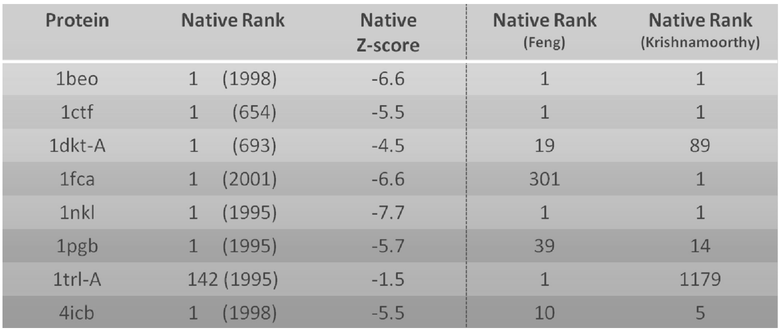


**Figure S2.** Evolutionary rates (measuared as $d_N/d_S$) in the three age groups: Eukarya, Metazoans, Vertebrates. The differences are significant in all pairwise comparisons (wilcoxon tests, Eukarya vs Metazoans:p-value = 0.004 , Eukarya vs Vertebrates:p-value <2.2e$^{-16}$ , Metazoans vs Vertebrates: p-value <2.2e$^{-16}$ ).



| Protein | Native Rank | Native Z-score | Native Rank (Feng) | Native Rank (Krishnamoorthy) |
|---|---|---|---|---|
| 1beo | 1 (1998) | -6.6 | 1 | 1 |
| 1ctf | 1 (654) | -5.5 | 1 | 1 |
| 1dkt-A | 1 (693) | -4.5 | 19 | 89 |
| 1fca | 1 (2001) | -6.6 | 301 | 1 |
| 1nkl | 1 (1995) | -7.7 | 1 | 1 |
| 1pgb | 1 (1995) | -5.7 | 39 | 14 |
| 1trl-A | 142 (1995) | -1.5 | 1 | 1179 |
| 4icb | 1 (1998) | -5.5 | 10 | 5 |

**Figure S3.** Structure Recognition: Discrimination of Native from Decoy Structures. Comparison of the performance of our potential (Native rank) with the performance of the potential derived by Feng (Feng et al., 2007) and Krishnamoorthy (Krishnamoorthy and Tropsha, 2003).

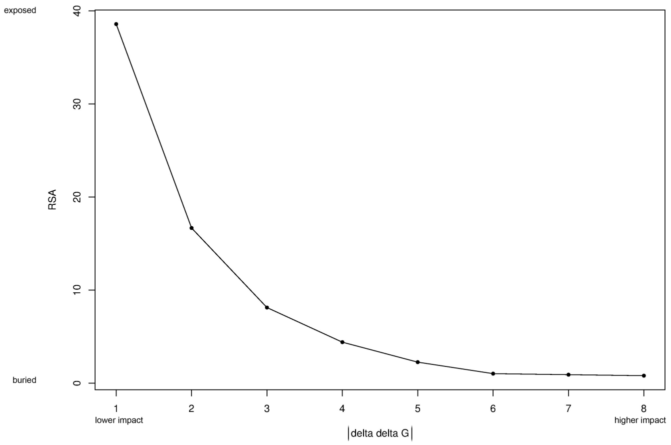


**Figure S4.** Mutations with a higher impact tend to occur in more buried residues. Differences between delta delta G are highly significative (wilcoxon test, p-value <2.2 e$^{-16}$) exceptuating the comparison between bin 6 and 7 and bin 7 and 8.

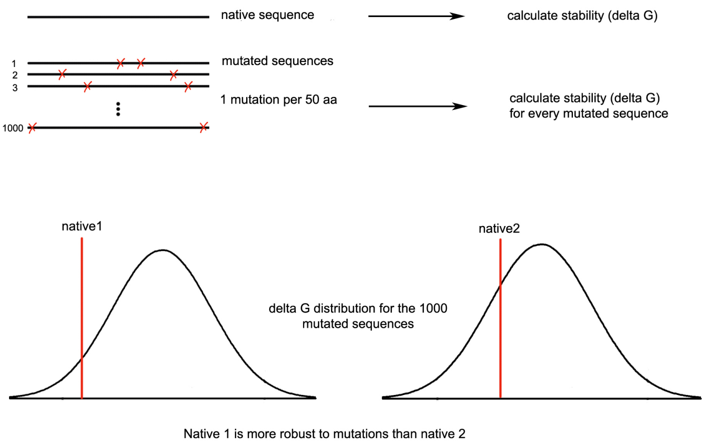

native sequence
calculate stability (delta G)
1
2
3
1000
mutated sequences
1 mutation per 50 aa
calculate stability (delta G)
for every mutated sequence
native1
native2
delta G distribution for the 1000
mutated sequences
Native 1 is more robust to mutations than native 2

**Figure S5.** Diagram representing the pipline done to assess PDB's robustness against point mutations.

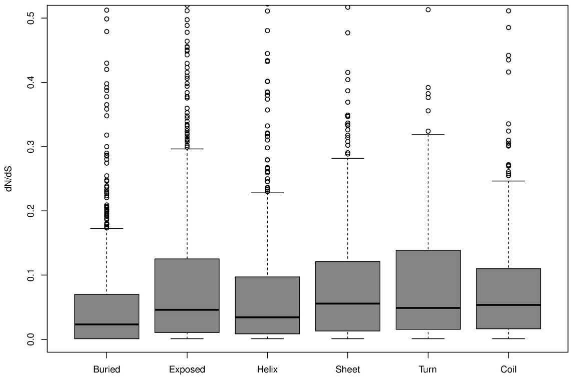


**Figure S6.** Residues classified in structural classes (Helix, Sheet, Turn and Coil) and solvent accessibility properties (Buried, Exposed). Two trends could be observed 1) exposed residues evolve faster than buried ones (wilcoxon test, p-value <0.01), 2) helix structure is evolving slower than the other types of secondary structures (wilcoxon test, p-value <0.01).

## 6 References

Albà, M. M. and Castresana, J. (2005). Inverse relationship between evolutionary rate and age of mammalian genes. *Molecular biology and evolution*, 22(3):598–606.

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research*, 25(17):3389–402.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000). The Protein Data Bank. *Nucleic acids research*, 28(1):235–42.

Bloom, J. D., Drummond, D. A., Arnold, F. H., and Wilke, C. O. (2006a). Structural determinants of the rate of protein evolution in yeast. *Molecular biology and evolution*, 23(9):1751–61.

Bloom, J. D., Labthavikul, S. T., Otey, C. R., and Arnold, F. H. (2006b). Protein stability promotes evolvability. *Proceedings of the National Academy of Sciences of the United States of America*, 103(15):5869–74.

Bloom, J. D., Silberg, J. J., Wilke, C. O., Drummond, D. A., Adami, C., and Arnold, F. H. (2005). Thermodynamic prediction of protein neutrality. *Proceedings of the National Academy of Sciences of the United States of America*, 102(3):606–11.

Bustamante, C. D., Townsend, J. P., and Hartl, D. L. (2000). Solvent accessibility and purifying selection within proteins of Escherichia coli and Salmonella enterica. *Molecular biology and evolution*, 17(2):301–8.

Cai, J. J. and Petrov, D. A. (2010). Relaxed purifying selection and possibly high rate of adaptation in primate lineage-specific genes. *Genome biology and evolution*, 2:393–409.

Cai, J. J., Woo, P. C. Y., Lau, S. K. P., Smith, D. K., and Yuen, K.-Y. (2006). Accelerated evolutionary rate may be responsible for the emergence of lineage-specific genes in ascomycota. *Journal of molecular evolution*, 63(1):1–11.

Carter, C. W., LeFebvre, B. C., Cammer, S. A., Tropsha, A., and Edgell, M. H. (2001). Four-body potentials reveal protein-specific correlations to stability changes caused by hydrophobic core mutations. *Journal of molecular biology*, 311(4):625–38.

Choi, I.-G. and Kim, S.-H. (2006). Evolution of protein structural classes and protein sequence families. *Proceedings of the National Academy of Sciences of the United States of America*, 103(38):14056–61.

Conant, G. C. and Stadler, P. F. (2009). Solvent exposure imparts similar selective pressures across a range of yeast proteins. *Molecular biology and evolution*, 26(5):1155–61.

Daubin, V. and Ochman, H. (2004). Bacterial genomes as new gene homes: the genealogy of ORFans in E. coli. *Genome research*, 14(6):1036–42.

Deutsch, C. and Krishnamoorthy, B. (2007). Four-body scoring function for mutagenesis. *Bioinformatics (Oxford, England)*, 23(22):3009–15.

Domazet-Loso, T. and Tautz, D. (2003). An evolutionary analysis of orphan genes in Drosophila. *Genome research*, 13(10):2213–9.

Drummond, D. A., Bloom, J. D., Adami, C., Wilke, C. O., and Arnold, F. H. (2005). Why highly expressed proteins evolve slowly. *Proceedings of the National Academy of Sciences of the United States of America*, 102(40):14338–43.

Drummond, D. A., Raval, A., and Wilke, C. O. (2006). A single determinant dominates the rate of yeast protein evolution. *Molecular biology and evolution*, 23(2):327–37.

Drummond, D. A. and Wilke, C. O. (2008). Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell*, 134(2):341–52.

England, J. L. and Shakhnovich, E. I. (2003). Structural determinant of protein designability. *Physical review letters*, 90(21):218101.

Feng, Y., Kloczkowski, A., and Jernigan, R. L. (2007). Four-body contact potentials derived from two protein datasets to discriminate native structures from decoys. *Proteins*, 68(1):57–66.

Ferrada, E. and Wagner, A. (2008). Protein robustness promotes evolutionary innovations on large evolutionary time-scales. *Proceedings. Biological sciences / The Royal Society*, 275(1643):1595–602.

Flicek, P., Amode, M. R., Barrell, D., Beal, K., Brent, S., Chen, Y., Clapham, P., Coates, G., Fairley, S., Fitzgerald, S., and et al. (2011). Ensembl 2011. *Nucleic acids research*, 39(Database issue):D800–6.

Franzosa, E. and Xia, Y. (2008). Structural perspectives on protein evolution.

Franzosa, E. A. and Xia, Y. (2009). Structural determinants of protein evolution are context-sensitive at the residue level. *Molecular biology and evolution*, 26(10):2387–95.

Fraser, H. B., Hirsh, A. E., Steinmetz, L. M., Scharfe, C., and Feldman, M. W. (2002). Evolutionary rate in the protein interaction network. *Science (New York, N.Y.)*, 296(5568):750–2.

Gan, H. H., Tropsha, A., and Schlick, T. (2001). Lattice protein folding with two and four-body statistical potentials. *Proteins*, 43(2):161–74.

Goldman, N., Thorne, J. L., and Jones, D. T. (1998). Assessing the impact of secondary structure and solvent accessibility on protein evolution. *Genetics*, 149(1):445–58.

Green, P., Lipman, D., Hillier, L., Waterston, R., States, D., and Claverie, J. M. (1993). Ancient conserved regions in new gene sequences and the protein databases. *Science (New York, N.Y.)*, 259(5102):1711–6.

Hahn, M. W. and Kern, A. D. (2005). Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. *Molecular biology and evolution*, 22(4):803–6.

Hirsh, A. E. and Fraser, H. B. (2001). Protein dispensability and rate of evolution. *Nature*, 411(6841):1046–9.

Kabsch, W. and Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12):2577–637.

Katoh, K., Misawa, K., Kuma, K.-i., and Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic acids research*, 30(14):3059–66.

Krishnamoorthy, B. and Tropsha, A. (2003). Development of a four-body statistical pseudo-potential to discriminate native from non-native protein conformations. *Bioinformatics (Oxford, England)*, 19(12):1540–8.

Kuo, C.-H. and Kissinger, J. C. (2008). Consistent and contrasting properties of lineage-specific genes in the apicomplexan parasites Plasmodium and Theileria. *BMC evolutionary biology*, 8:108.

Lin, Y.-S., Hsu, W.-L., Hwang, J.-K., and Li, W.-H. (2007). Proportion of solvent-exposed amino acids in a protein and rate of protein evolution. *Molecular biology and evolution*, 24(4):1005–11.

Lipman, D. J., Souvorov, A., Koonin, E. V., Panchenko, A. R., and Tatusova, T. A. (2002). The relationship of protein conservation and sequence length. *BMC evolutionary biology*, 2:20.

Lobkovsky, A. E., Wolf, Y. I., and Koonin, E. V. (2010). Universal distribution of protein evolution rates as a consequence of protein folding physics. *Proceedings of the National Academy of Sciences of the United States of America*, 107(7):2983–8.

Marais, G. and Duret, L. (2001). Synonymous codon usage, accuracy of translation, and gene length in Caenorhabditis elegans. *Journal of molecular evolution*, 52(3):275–80.

Masso, M., Lu, Z., and Vaisman, I. I. (2006). Computational mutagenesis studies of protein structure-function correlations. *Proteins*, 64(1):234–45.

Miller, S., Janin, J., Lesk, A. M., and Chothia, C. (1987). Interior and surface of monomeric proteins. *Journal of molecular biology*, 196(3):641–56.

Mirny, L. A. and Shakhnovich, E. I. (1999). Universally conserved positions in protein folds: reading evolutionary signals about stability, folding kinetics and function. *Journal of molecular biology*, 291(1):177–96.

Notredame, C., Higgins, D. G., and Heringa, J. (2000). T-Coffee: A novel method for fast and accurate multiple sequence alignment. *Journal of molecular biology*, 302(1):205–17.

Pál, C., Papp, B., and Hurst, L. D. (2001). Highly expressed genes in yeast evolve slowly. *Genetics*, 158(2):927–31.

Pál, C., Papp, B., and Lercher, M. J. (2006). An integrated view of protein evolution. *Nature reviews. Genetics*, 7(5):337–48.

Plotkin, J. B. and Fraser, H. B. (2007). Assessing the determinants of evolutionary rates in the presence of noise. *Molecular biology and evolution*, 24(5):1113–21.

Samudrala, R. and Levitt, M. (2000). Decoys 'R' Us: a database of incorrect conformations to improve protein structure prediction. *Protein science : a publication of the Protein Society*, 9(7):1399–401.

Shakhnovich, B. E. (2006). Relative contributions of structural designability and functional diversity in molecular evolution of duplicates. *Bioinformatics (Oxford, England)*, 22(14):e440–5.

Simon, M. and Hancock, J. M. (2009). Tandem and cryptic amino acid repeats accumulate in disordered regions of proteins. *Genome biology*, 10(6):R59.

Toft, C. and Fares, M. A. (2010). Structural Calibration of the Rates of Amino Acid Evolution in a Search for Darwin in Drifting Biological Systems. *Molecular biology and evolution*, 27(10):2375–85.

Toll-Riera, M., Bosch, N., Bellora, N., Castelo, R., Armengol, L., Estivill, X., and Albà, M. M. (2009). Origin of primate orphan genes: a comparative genomics approach. *Molecular biology and evolution*, 26(3):603–12.

Toll-Riera, M., Radó-Trilla, N., Martys, F., and Albà, M. M. (2011). Role of Low-Complexity Sequences in the Formation of Novel Protein Coding Sequences. *Molecular biology and evolution*.

Vishnoi, A., Kryazhimskiy, S., Bazykin, G. a., Hannenhalli, S., and Plotkin, J. B. (2010). Young proteins experience more variable selection pressures than old proteins. *Genome research*, 20(11):1574–81.

Wall, D. P., Hirsh, A. E., Fraser, H. B., Kumm, J., Giaever, G., Eisen, M. B., and Feldman, M. W. (2005). Functional genomic analysis of the rates of protein evolution. *Proceedings of the National Academy of Sciences of the United States of America*, 102(15):5483–8.

Wang, G. and Dunbrack, R. L. (2003). PISCES: a protein sequence culling server. *Bioinformatics (Oxford, England)*, 19(12):1589–91.

Wingreen, N., Li, H., and Tang, C. (2003). Designability and Thermal Stability of Protein Structures. *Polymer*, 45(2):12.

Wolf, Y. I., Novichkov, P. S., Karev, G. P., Koonin, E. V., and Lipman, D. J. (2009). The universal distribution of evolutionary rates of genes and distinct characteristics of eukaryotic genes of different apparent ages. *Proceedings of the National Academy of Sciences of the United States of America*, 106(18):7273–80.

Wong, P. and Frishman, D. (2006). Fold designability, distribution, and disease. *PLoS computational biology*, 2(5):e40.

Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Molecular biology and evolution*, 24(8):1586–91.

Zhou, T., Drummond, D. A., and Wilke, C. O. (2008). Contact density affects protein evolutionary rate from bacteria to animals. *Journal of molecular evolution*, 66(4):395–404.

## 3.3 Origin and characteristics of primate orphan genes

Genomes are composed by genes originated at different time points, and every genome contains genes that are unique. Those genes are named orphan genes. Despite that orphan genes have been proposed to be involved in lineage-specific adaptive processes little is known about them.

This chapter includes 4 documents, the first one is a book chapter that reproduces the work done by Albà and Castresana (Albà and Castresana, 2005) but using more recent mammalian genes, primate orphan genes, and also provides a review on the subject. The second article is the main work of the chapter and presents the first study done on the mechanisms of origin of primate orphan genes. The third article is a review on the mechanisms of origin of primate orphan genes and extends a little bit the exaptation of transposable elements as a mechanism of origin. And finally, the last article is a book chapter about the role of gene duplication in the formation of orphan genes, highlighting the importance of partial gene duplication.

### 3.3.1 Accelerated Evolution of Genes of Recent Origin

**Authors:** Macarena Toll-Riera, Jose Castresana and M. Mar Albà

**Published in:** Evolutionary Biology from Concept to Application (Chapter 3). Ed. Pontarotti. Springer: Berlin

**Full text:** http://www.springerlink.com/content/m85w5421t3x0xm22/

**Summary**

The gene content of any genome is a rich mosaic of genes that have originated at different times during evolution. Among the most interesting properties related to gene age is the fact that younger genes tend to show accelerated evolutionary rates with respect to older genes. Here, we use a large number of closely related mammalian genomes to gain further insights into the relationship between gene age and evolutionary rate. We define a group of primate-specific genes that are absent from 11 non-primate mammalian genomes as well as from other eukaryotic genomes. These genes, of very recent origin, show the highest evolutionary rate and the shortest protein length. We discuss how these results may shed light on understanding the proposed mechanisms for the origin of lineage-specific, novel genes

Toll-Riera M, Castresana J, Alba M.M. Accelerated Evolution of Genes of Recent Origin. In: Pierre Pontarotti, Book editor. Evolutionary Biology from Concept to Application. France: Springer; 2008. p.183-90.

### 3.3.2 Origin of Primate Orphan Genes: A Comparative Genomics Approach

**Authors:** Macarena Toll-Riera, Nina Bosch, Nicolas Bellora, Robert Castelo, Lluis Armengol, Xavier Estivill and M. Mar Albà

**Published in:** Molecular Biology and Evolution. 2009 Mar; 26(3):603-12

**Full text:** http://mbe.oxfordjournals.org/content/26/3/603.abstract

**Summary**

Genomes contain a large number of genes that do not have recognizable homologues in other species and that are likely to be involved in important species-specific adaptive processes. The origin of many such ″orphan″ genes remains unknown. Here we present the first systematic study of the characteristics and mechanisms of formation of primate-specific orphan genes. We determine that codon usage values for most orphan genes fall within the bulk of the codon usage distribution of bona fide human proteins, supporting their current protein-coding annotation. We also show that primate orphan genes display distinctive features in relation to genes of wider phylogenetic distribution: higher tissue specificity, more rapid evolution, and shorter peptide size. We estimate that around 24% are highly divergent members of mammalian protein families. Interestingly, around 53% of the orphan genes contain sequences derived from transposable elements (TEs) and are mostly located in primate-specific genomic regions. This indicates frequent recruitment of TEs as part of novel genes. Finally, we also obtain evidence that a small fraction of primate orphan genes, around 5.5%, might have originated de novo from mammalian noncoding genomic regions

### 3.3.3 Evolution of primate orphan proteins

**Authors:** Macarena Toll-Riera, Robert Castelo, Nicolas Bellora and M. Mar Albà

**Published in:** Biochemical Society Transactions. 2009 Aug;37(Pt 4):778-82

**Full text:** http://www.biochemsoctrans.org/bst/037/bst0370778.htm

**Summary**

Genomes contain a large number of genes that do not have recognizable homologues in other species. These genes, found in only one or a few closely related species, are known as orphan genes. Their limited distribution implies that many of them are probably involved in lineage-specific adaptive processes. One important question that has remained elusive to date is how orphan genes originate. It has been proposed that they might have arisen by gene duplication followed by a period of very rapid sequence divergence, which would have erased any traces of similarity to other evolutionarily related genes. However, this explanation does not seem plausible for genes lacking homologues in very closely related species. In the present article, we review recent efforts to identify the mechanisms of formation of primate orphan genes. These studies reveal an unexpected important role of transposable elements in the formation of novel protein-coding genes in the genomes of primates.

### 3.3.4 Partial Gene Duplication and the Formation of Novel Genes

**Authors:** Macarena Toll-Riera, Steve Laurie, Núria Radó-Trilla and M. Mar Albà

**Published in:** Gene Duplication (Book 1), Intech

**Full text:** http://www.intechopen.com/books/gene-duplication/partial-gene-duplication-and-the-formation-of-novel-genes

Toll-Riera M, Laurie S, Radó-Trilla N, Alba M.M. Partial Gene Duplication and the Formation of Novel Genes. In: Felix Friedberg, Book editor. Gene Duplication. Croatia: InTech; 2011. p.95-110.

# 4

# Discussion

In this section I am going to go through the main topics covered in the articles and discuss them in the context of the general knowledge in the area, trying to give insights about how they have contributed into the field. I have also included a section in which I am going to describe the main pitfalls I found while doing this thesis, putting special emphasis on some methodological issues, such as the importance of a correct alignment, the problems derived from bad quality sequences and the reliability of the evolutionary rate estimations. This subject is currently a hot topic.

## 4.1 Variations in the strength of selection in mammals

In the article entitled *Lineage-specific variation in Intensity of Natural Selection in Mammals* we have centred our attention in developing a methodology that allows us to identify lineage-specific variations in evolutionary rates. Our objective was developing the methodology as well as assessing how frequent are changes in the selective pressures among mammalian orthologous genes. Until date it was still not studied if orthologous mammalian genes are evolving all in the same fashion or if there are lineage-specific variations, which could be leading to species-specific functional divergence.

The molecular clock hypothesis (Zuckerkandl and Pauling, 1962) predicts that genes accumulate changes at an approximately constant rate. Lately it was pointed out that the molecular clock hypothesis stands as long as the function and tertiary structure of the protein remains unaltered (Kimura and Ota, 1974). Several examples have been reported in which the assumption of a constant rate of evolution is not accomplished when comparing evolutionary rates between species (Ohta and Ina, 1995; Ayala, 2000; Arbiza et al., 2006; Bedford et al., 2008). It is mainly due to two reasons:

- Species specific characteristics: there are differences between species regarding their generation time, DNA replication fidelity and mechanisms of error repair, historical population size, etc. One clear example is the differences in the generation time between mouse and human, which leads to a difference in the number of fixed nucleotide substitutions when the two species are compared, being almost double in mouse (Waterston et al., 2002).

- Changes in the type and strength of selection affecting a gene in a particular lineage.

We were interested in this second reason, and therefore we thought a methodology that allowed us to quantify it. We had mainly to take into account two effects to be able to identify lineage-specific rate variations:

- Species specific characteristics: differences between species in some of their characteristics such as generation time, error repair and population size.
- Protein family: the specific function of the protein, as it could be influencing its evolution. For example, immune response genes are known to evolve fast in order to adapt to changing environments. On contrary, ribosomal proteins or histones are evolving very slowly.

To perfomr this study we used a dataset comprised by one-to-one orthologs between six mammalian species. Briefly, what we did is to obtain a reference tree by concatenating randomly 150 human proteins of the dataset (and its orthologs). This reference tree contains the information regarding the typical species-specific evolutionary rate given their characteristics, such as generation time and population size. Then, for each protein's observed tree be wanted to asses if the branches were evolving differently from the expected. We did that calculating the expected tree by maintaining the total branch length constant (to account for the global rate of evolution of the specific gene) but with relative branch lengths (to take into account the species-specific differences represented by the reference tree). Finally, we compared the observed tree with the expected tree and if branches were evolving significantly faster were classified as accelerated and if, on the contrary, were evolving significantly slower were classified as decelerated. This methodology is very similar to the one used by Shapiro and Alm (Shapiro and Alm, 2008) but they did not use $d_N/d_S$ because they were working with $\gamma$-proteobacteria, which implies a considerable high divergence time and, consequently, saturation of synonymous substitutions. Instead of $d_N$, $d_S$ and $d_N/d_S$ estimates they have used branch-specific rates of amino acid substitutions. However, they also provided $d_N/d_S$ estimates. To distinguish positive selection from relaxation of selective constraints they used polymorphism data from *Escherichia coli* to perform a MK test, whereas, we used branch-site tests implemented in the codeml package. In this study we applied sever filters that reduced our dataset considerably, but which increased our confidence that we had high quality data. The methodological issues from this article will be explained in more detail in section 4.4.

At an initial step we checked if our dataset was explained by a model in which all branches were sharing the same $d_N/d_S$ (one-ratio) or wether the data fitted better in a model in which $d_N/d_S$ was estimated for each branch (free-ratio). As we expected, we found that most trees followed a nonclock like behaviour and this could be explained basically by differences in species-specific traits. For example, $d_N/d_S$ was higher in primates that in rodents, and this could be explained by the smaller effective population size of primates, as purifying selection is less effective in smaller populations. I would like to remark that these results are independent of the neutralists versus selectionists debate, as they only tell us that a global common clock for all the species does not exist, but it still could exist a species-specific clock. This is related with the results obtained by Kumar and Subramanian (Kumar and Subramanian, 2002), in which they reported differences in the mutation rates across mammals, which means that they could not find a global DNA clock for all mammals. As they were comparing species with similar characteristics (generation time, life-history traits), the differences seen in mutation rates should be due to replication-independent processes.

Surprisingly, we found that 24.5% of the genes showed branch-specific deviations (accelerations or decelerations), demonstrating a high evolutionary flexibility on the selective pressures, changing in different lineages and among different periods of times. We have to keep in mind that the used dataset are one-to-one orthologs, so is a well-conserved set which excludes duplicates, which are usually under functional diversification. Most trees had only one deviated branch, and the total number of accelerated branches was similar than the number of decelerated ones. When we compared with a Poisson distribution we found that we had an excess of trees presenting two or more deviated branches, these trees probably belong to genes which easily undergo changes in the selective pressures.

The values have been normalized, thus, an accelerated branch has not to be mandatorily associated with a high $d_N/d_S$, and therefore associated with positive selection, it only indicates that a specific branch of the tree is evolving significantly faster than the others. To infer positive selection or relaxation of selective constraints we applied the branch-site test implemented in the Paml package (Yang, 2007). Although we find that accelerated branches were enriched in positive selection (18% of accelerated branches were significant for the positive selection test), there was little overlap between both groups, and in addition, cases in which the accelerated branch had a $d_N/d_S$ >1 no positive selection was detected. This is because positive selection tests and our methodology measure two different aspects of gene evolution, evidence of adaptation and species-specific deviations, respectively. A gene could be positively selected, but this does not mean that there should be differences in how the different lineages evolve; all the species could be evolving in the same direction, for example, in the case of an immune system gene. Is for this reason that the group of positively selected genes and the group of genes presenting accelerated branches are not expected to totally overlap. Nevertheless, accelerated branches could be also explained due to a relaxation of the selective constraints. Additionally, it could also be a problem of the branch-sites test used to detect positive selection which is too conservative and is not able to detect positive selection in sequences which are overall very conserved but have some branch-specific changes. As our methodology takes into account the general evolution of the gene, it can identify those cases. One example is the glutamate receptor subunit 3A (GRIN3A), which is involved in the synaptic transmission. This gene is highly conserved among the studied mammals, but it has seven human-specific radical changes, which could have important functional consequences; it is detected as accelerated by our methodology, but not as positively selected. In Shapiro and Alm (Shapiro and Alm, 2008) study they found that around half of the fast-evolving genes could be explained by the action of positive selection and the other half could be due to relaxation of selective constraints. Therefore, they found a much higher fraction of positive selection than us. This could be due to differences in the studied species, mammals have smaller population sizes than bacteria and hence, the efficacy of selection is expected to be lower.

Importantly, we found that accelerated branches had more radical changes than nondeviated ones, and that they had an enrichment in replacements involving gain or loss of charge, which are the changes expected to be more critical for protein's function. Studies looking for positively selected genes have mainly found an enrichment in functions related with immunity and chemosensory perception which reinforces the idea that deviations from the molecular clock are non random. In here we find enrichment in proteins involved in neural functions and in the interaction with the environment. It has been not previously reported the existence of an excess of functional changes in neural functions, and this could be explained by the lack of sensitivity of the branch-site test to detect positive selection in sequences with few but radical changes, as the case of GRIN3A explained before.

We found a similar number of decelerated branches than accelerated ones. Decelerated branches are more difficult to interpret than accelerated branches because in these cases the functional constraints have increased in some specific branches. One possibility is that some positions have been coopted for a novel function, such as interacting with another protein;

therefore these residues would now be under negative selection not to change.

One important conclusion of this work is that well-conserved core orthologous proteins are not static entities, they can be rather playing an important role in the adaptation of organisms to new environments, which has always been attributed to the acquisition of new genes (Shapiro and Alm, 2008). This methodology has the inconvenient that it can only be used between closely related core orthologous genes to avoid the saturation of the synonymous changes. But as reported by Shapiro and Alm and by us, it has the advantage of detecting those cases in which selection is acting only on a few amino acids and in genes with a low $d_N/d_S$, as in the example of GRIN3A, where the branch-site test seems to fail. However, if all branches are evolving with elevated $d_N/d_S$ and in the same fashion, the method presented here is not going to be informative, but the branch-site test could still be able to detect positive selection.

I would like to point out that whether it exists a molecular clock or not was not the purpose of the article and, as I have reviewed in the introduction, there are plenty of studies focusing on revealing the fraction of proteins under adaptive evolution among several genomes. In fact, we are not studying positive selection in detail, because those trees with no deviated branches can still be under positive selection, and we are focusing our analysis on trees presenting deviated branches. Additionally, for those trees with accelerated branches which are not significant for the positive selection test we can not discern if they are accelerated due to positive selection but we have not been able to detect it due to the conservativeness of the test or they are accelerated due to the relaxation of selective constraints. Trees with accelerated branches could be explained by the neutral theory as well as by selectionists. Under the neutral theory they can be explained because they imply a change of function or protein tertiary structure, while selectionists can explain acceleration due to the action of positive selection.

## 4.2 Age as a determinant for protein evolution

It is widely known that different proteins are evolving at different rates, and differences are over three orders of magnitude. Several determinants have been proposed to drive protein evolution such as protein dispensability (Hirsh and Fraser, 2001), gene expression (Green et al., 1993; Pál et al., 2001; Wall et al., 2005),protein-protein interactions (Fraser et al., 2002), age of protein's origin (Albà and Castresana, 2005; Wolf et al., 2009), protein structure, solvent accessibility and pairwise interactions among amino acids (Choi et al., 2007). Bloom and colleagues (Bloom et al., 2006a) performed a PCR analysis and reported that the main contributors are expression, which explains around 40% of the variation and, to a lesser extent, protein structure, which contributes to around 10%. Gene expression, measured as mRNA level and codon usage bias, has been observed to covary with sequence evolution from human to bacteria. The mechanism proposed to act is selection against the toxicity of misfolded proteins, a hypothesis named Mistranslation-Induced Protein Misfolding (MIM) (Drummond and Wilke, 2008).

Now, with the work presented in section 3.2.1 and section 3.2.2, we provide strong evidence that protein age is an important and independent determinant of the evolutionary rate. We also give insights into how protein age interplays with protein structure.

Albà and Castresana (Albà and Castresana, 2005) were the first to report an inverse relationship between protein's age and evolutionary rate, meaning that younger genes are evolving much faster than older ones, suggesting that protein age could be considered another determinant of protein's evolutionary rate. They observed this relationship in mammals, but it has also been verified in a wide range of species such as primates (Toll-Riera et al., 2009a), fungi (Cai et al., 2006), *Drosophila* (Domazet-Loso and Tautz, 2003; Wolf et al., 2009), *Plasmodium* (Kuo and Kissinger, 2008) and bacteria (Daubin and Ochman, 2004), therefore, it seems to be universal. However, it is also well-known that young genes are expressed at lower levels and in a fewer number of tissues than older ones (Cai et al., 2006; Toll-Riera et al., 2009a). Consequently, the role of protein age had not been separated from the role that expression might play in determining protein evolution.

Moreover, it has been argued that the observed relationship between protein's age and evolutionary rate could be due to an artefact as a consequence of a circularity problem caused by the BlastP failure to detect distant homologues of rapidly evolving proteins (Elhaik et al., 2006). Nonetheless, Albà and Castresana (Albà and Castresana, 2007) simulated protein sequences corresponding to old genes (Eukarya) and evolved them using the evolutionary rates observed in mammalian genes. Then they assessed if they were able to classify them correctly using BlastP searches. As the simulated sequences were old, all the sequences not classified as Eukarya would be due to misclassifications produced by Blast lack of power to detect remote homologues. They obtained that a very small fraction of the sequences suffer from misclassification, therefore is highly unlikely that the observed relationship between protein age and protein evolution is the product of a BlastP artefact.

In section 3.2.1 we have gone one step further in the study of protein age as a determinant of protein evolution. We have tried to overcome the two pitfalls seen in previous studies: the possible BlastP artefact and the dependence between protein's age and expression level. For this reason we have employed a new approach, we have used protein domains instead of proteins. Domains are protein fragments that are found in several proteins. They are considered the units of evolution because they can fold independently, have a function and they have their own evolutionary history (Vogel et al., 2004a; Ekman et al., 2005; Itoh et al., 2007).

The use of protein domains has two main advantages regarding the previous methodology. First, it allows the use of hidden markov models (HMM) to perform domain searches across the genomes to identify homologues. The use of HMM is an improvement over BlastP because is more sensitive and facilitates the identification of distant homologues. Of course, some sceptics can still argue that some remote homologues could still be missed. And secondly, we can compare the evolutionary rates between domains found in the same protein but that differ in their origination time. When doing that other factors suggested to determine protein's evolution, such as protein function and gene expression level can be controlled for.

Accordingly to previous results, we find that the inverse correlation between age and evolutionary rate also holds at the domain level, both in mammals and in *Drosophila*. But the advantage of using domains is that we can test for that correlation inside proteins containing young and old domains. Indeed, we found that young domains are evolving much faster than older domains also when they belong to the same protein. Interestingly, there is only one other study measuring the evolutionary rates at domain level. In this study the authors wanted to test the hypothesis that domains located in multidomain proteins would have more homogeneous rates because they are translated at the same rate than the same domains when found in different proteins. They found the expected homogenization of evolutionary rates of domains belonging to the same protein. However they still found significant differences between those domains, indicating that expression alone could not account for the observed variation in their evolutionary rates (Wolf et al., 2008). The authors suggested that this second player could be domain-specific structural and functional constraints, but, indeed, it can certainly be caused by the age effect we are reporting.

Additionally we gave some insights on how young domains are gained into existing old proteins. We observed that older proteins are longer and have a higher number of domains than younger proteins. We found evidence that young domains found in old proteins are mostly formed in the context of an existing old protein and less frequently are due to domain fusion between an old and a new domain. New young domains are overrepresented in the N-terminus of the protein, suggesting that novel domains, at least in mammalian proteins, are acquired mostly by an extension of the coding sequence in the 5′ region of the genes. Terminal parts of proteins are flexible, charged and are located at the surface of proteins, and for this reason the addition of a new domain in those regions is less likely to disrupt the structure of the protein (Buljan and Bateman, 2009). Therefore, our results seem to indicate that proteins increase their length and complexity progressively by the acquisition of younger domains, mainly in the N-terminus region. This result agrees with the study performed by Choi and Kim (Choi and Kim, 2006) in which it was found that young proteins are usually short and adopt simple structures: $\alpha$, $\beta$ and $\alpha$+$\beta$, middle-age proteins increase in length and can also adopt more complex $\alpha/\beta$ structures and the oldest proteins are even longer and most of them have $\alpha/\beta$ structures. However, it does not agree with some studies in which it is proposed that the most important mechanisms involved in novel arrangements are terminal loss of domains and fusion of existing genes (Bornberg-Bauer et al., 2010).

Nontheless, most of the domains classified as young are not found in combination with older domains, they commonly form single domain young proteins. Besides, there is also an important fraction of young proteins that do not contain any domain, which is probably an evidence of our current lack of understanding of this type of proteins. It has been observed in various studies (Lander et al., 2001; Pal and Guda, 2006; Yang and Bourne, 2009), and also in this, that there is a higher fraction of domains with an old origin than domains with a vertebrate origin. This observation could be partly due to the lack of knowledge we have of young domains compared to the well characterized old domains (Pal and Guda, 2006; Yang and Bourne, 2009). However, the general trend observed is that in spite of creating new domains, younger proteins have evolved through the acquisition and combination of already existing domains (Patthy, 2003; Pal and Guda, 2006; Ekman et al., 2007; Yang and Bourne, 2009). Yang et al. (Yang and Bourne, 2009) have inferred that only 831 new domains have been created after the emergence of the first eukaryotic cell. This number represents less than the 25% of the domains. On the contrary, the number of domain combinations shows an inverse trend, only 4% of the combinations have been created in the root of the tree. One clear example is the Zinc finger. Zinc finger is one of the most common domains in our dataset, but it has also been reported to be one of the most abundant domains in human (Lander et al., 2001; Müller et al., 2002). Additionally, it has been shown to be very robust to mutations, which probably accounts for ist versatility (Wagner, 2011). In fact, zinc finger superfamily is highly expanded in vertebrates (Müller et al., 2002; Vogel and Chothia, 2006), and in vertebrates a novel combination with KRAB and SCAN domains has been created (Lander et al., 2001). Moreover, the most common combination of two domains in humans is the one formed by zinc finger and KRAB box (Pal and Guda, 2006), which are, precisely, one of the most abundant domains in our old and vertebrate groups respectively. Therefore, the reuse of domains rather than the creation of new ones speeds up the evolution of cellular complexity (Moore et al., 2008).

We have made an effort to demonstrate that young domains are not being misclassified. We performed BlastP searches of the proteins containing an old and a young domain against *Drosophila melanogaster* and *Caenorhabditis elegans* to find homologues in those proteomes. We should find a hit in those genomes through the old domain and consequently, if a remote homologue to the vertebrate domain exists (which was not detected previously), it should be now found in the corresponding region in *D. melanogaster* and *C.elegans*. For each protein we kept the best BlastP hit and we performed pairwise alignments between it and the query to compare the percent identity found a) in vertebrate domains and the corresponding region in *D. melanogaster* and *C.elegans*, b) in old domains and their corresponding region in the same two species and, c) in old/vertebrate domains randomized and the corresponding region in the two species. We found that the distribution of the percent identities found when vertebrate domains were randomized was not statistically different from non randomized domains. In contrast, strong and statistically significant differences were found in the comparison of old nonrandomized domains and the randomized ones. Therefore, no evidence of the vertebrate domain could be detected in fly and worm, and is for this reason that we claim that we have correctly identified remote homologues in the vast majority of the cases. Protein length has also been related with evolutionary rates (Lipman et al., 2002) and it has also been argued that shorter sequences are not easily detected at long evolutionary distances, but in here, there are no differences in the protein's domain length across age classes. Finally, one could also argue that the relationship between protein age and evolutionary rate could be explained by the fact that younger genes have been originated recently, and as duplication is the main mechanism of origin the high rates could be due to the high divergence experienced after duplication events. This is not the case, first, because we are working at domain level and, second, because we are using one-to-one orthologous proteins.

To sum up, the work presented in section 3.2.1 is the strongest evidence until date that age is determining protein's evolution. This new approach has the advantage that as domains found in the same protein are expressed at the same rate and share the same protein function other previously suggested determinants, such as expression level and protein function, are controlled for. How age is influencing evolution could be explained by two hypotheses, 1) with time most adaptive mutations have already occurred and for this reason adaptive mutations become saturated or 2) with time a higher fraction of residues are implied in a function, increasing the strength of purifying selection. This work shows the high modularity of proteins, multi-domain proteins are a set of functional pieces originated at different time points and evolving at different rates according to it. Thus, independently of the core function of a protein, its older domain will be, essentially, evolving slower than its younger domain.

In section 3.2.2 we have extended the research performed in the previous section (3.2.1) by studying how protein structure is related with age. Various structural properties have been studied when trying to identify the determinants of protein evolution, such as solvent exposure (Franzosa and Xia, 2009; Lin et al., 2007), contact density (Bloom et al., 2006a; Zhou et al., 2008b) and designability (Bloom et al., 2006a). A principal component regression (PCR) analysis has pointed out that around 10% of the rate variation could be explained by structural characteristics (Bloom et al., 2006a). However, there is also the doubt of whether structure acts as a determinant itself or whether it works through the interplay with other mechanisms. In here we aimed to give insights into the possible interplay between age and structure and determine if the relationship found between evolutionary rates and age (Albà and Castresana, 2005; Wolf et al., 2009) could be explained by structural biases among age groups. To perform this study we have moved on to PDB structures, mapping them to one-to-one orthologous proteins between human and mouse and then assigning an age to each structure using BlastP similarity searches (Altschul et al., 1997).

It is widely known that buried residues tend to be conserved (Goldman et al., 1998; Mirny and Shakhnovich, 1999; Bustamante et al., 2000; Bloom et al., 2006a; Conant and Stadler, 2009), and, in fact, a direct relationship between solvent accessibility and evolutionary rate has been found in yeast (Franzosa and Xia, 2009). Accordingly, we also report, for the first time, this relationship in human, suggesting that it is an universal trend. We also found this strong direct relationship when we take into account the age of the structure, but, interestingly, the slope varies between ages, suggesting that the relationship is slightly governed by the age of origin of the structure, as the differences in evolutionary rates between buried and exposed residues are more abrupt in younger structures than in older ones. Besides, contact density (measured using several indicators, such as fraction of buried residues) has been proposed to be a measure of designability, being more designable those proteins with a higher contact density (England and Shakhnovich, 2003). More designable structures are expected to evolve faster (Bloom et al., 2006a), and indeed, proteins with higher contact density tended to evolve rapidly in fly, yeast, *E.coli*, and human (Zhou et al., 2008b). It has been hypothesized that proteins with a larger core (larger number of buried residues) evolve faster that proteins with smaller core because larger cores permit surface exposed residues to vary more freely (Bloom et al., 2006a; Franzosa and Xia, 2009). However, there are discrepancies regarding the relationship between contact density and evolutionary rate. Shakhnovich (Shakhnovich, 2006) found, in *C.elegans* and yeast, a negative correlation, and Lin and colleagues (Lin et al., 2007) found a negative correlation when they used support-vector machine predictions and no correlation when they used crystal structures to calculate the fraction of buried residues. These discrepancies are probably due to methodological differences between studies. In here we inspected how protein designability is influenced by age. We have been much more severe that precedent studies. We used a 99% sequence identity to assign structure, compared with, for example, the 40% identity used in Zhou et al (Zhou et al., 2008b). We found that old structures have a higher fraction of buried residues - are more designable, but, evolve slower than younger ones. Older folds have been previously reported to be more conserved than younger ones (Wong and Frishman, 2006). Notwithstanding, we observed that, in each age group, the fastest evolving structures tended to have a larger fraction of buried residues. Similarly, we detected a positive correlation between $d_N$ and the fraction of buried residues in Eukarya and Metazoans. Thus, our results suggest that the age effect is much stronger than the designability effect on evolutionary rates, and for this reason we can only detect the correlation between evolutionary rate and fraction of buried residues within each age group. More designable proteins are generally more stable (Wingreen et al., 2003) and have a higher fraction of buried residues, which possibly, leads to a more robust protein core. Stability enhances tolerance to mutations, as the probability of a mutation to disrupt the native structure is smaller (Bloom et al., 2005, 2006b). And precisely this is what we found among our structures, older structures are more stable and tolerant to mutations than younger ones.

Thus, by now, we have seen that differences in the designability across age groups could not explain the previously reported inverse relationship between protein age and evolutionary rate (Domazet-Loso and Tautz, 2003; Albà and Castresana, 2005; Wolf et al., 2009) but is still possible that this relationship is caused by biases in the secondary structure and solvent accessibility between age groups. To study this possibility we classified all the residues in 4 secondary structural classes (helix, sheet, turn and coil) and into two solvent accessibility classes (buried, exposed). Interestingly, when we classified the residues belonging to each of these six structural classes according to the age of their structure we found that inside each structural class the inverse relationship between evolutionary rate and age was still detected (Albà and Castresana, 2005; Wolf et al., 2009), which indicates that biases in the secondary structure/solvent accessibility cannot explain the observed differences in evolutionary rates between age groups.

Choi and Kim (Choi and Kim, 2006), as explained above, reported that older proteins are mainly formed by more complex structures ($\alpha/\beta$) than younger proteins. We classified our proteins according to SCOP (Murzin et al., 1995) and we observed similar results to Choi and Kim, Vertebrate proteins were mainly $\alpha$ proteins, Metazoans $\alpha$ and $\beta$ and Eukarya were $\alpha$, $\beta$ and $\alpha/\beta$. Aside from validating Choi and Kim results, our data gives insights into the temporal aspects of structural properties: older structures are more stable, robust to mutations and designable than younger ones, suggesting that stability and robustness to mutations are acquired with time. However, some of our data seems incongruent. We find that younger structures are evolving fastly, despite having a higher fraction of exposed residues (less designable) and being less robust to mutations. What we expected to find, taking into account structural properties, is younger structures evolving slowly because their lower stability and robustness causes that most of the mutations have a destabilizing effect. This contradiction could be explained if we consider that proteins belonging to different ages are under different degrees of selection. In particular, younger structures would be experiencing a strong positive selection to fix stabilizing mutations, explaining their higher evolutionary rates when compared with older structures. It was hypothesized by Albà and Castresana (Albà and Castresana, 2005) that perhaps young proteins were evolving more rapidly than older ones because they were subject to lower selective contraints, that is, they could accept more mutations without compromising function or structure. However, our results show that, at least with regards to structure, in younger proteins a higher proportion of the mutation should be deleterious (destabilizing), invalidating the above mentioned hypothesis. The data instead supports the idea that young proteins may incorporate beneficial mutations that increase stability and robustness to mutations. Nonetheless this only accounts for young proteins with a cristalized structure. It is known that young proteins are enriched in low complexity regions (LCR) (Simon and Hancock, 2009; Toll-Riera et al., 2011b). LCR are generally disordered (Simon and Hancock, 2009), thus, difficult to cristalize. Then, the hypothesis that younger proteins evolve faster due to the lower selective constraints can still hold for young proteins enriched in LCR which normally are unestructured and hence do not have the limitation of the distabilizing mutations.

The data presented in these two articles strongly suggest that protein sequence age is one of the strongest determinants for driving protein evolution, as its effect remains even when several other proposed determinants are controlled for. However, it also shows that we should not try to simplify the picture and propose only one determinant, as there is increasing data that suggest the several determinants such as gene expression and structural properties can also play a role, and especially, age and structural characteristics seems to interplay.

## 4.3 Origin and characteristics of primate orphan genes

With the sequencing of several genomes it has been noticed that the number of genes varies across related organisms, indicating the existence of mechanisms of gene birth and gene loss. Therefore, one very intriguing question is how genes have originated. In most genes the clues regarding to their mechanism of origin have been erased by the passage of time. This is the reason why the best set to study gene birth is the one formed by young genes. And the newest genes by definition are orphan/lineage-specific genes, which are those genes that are only found in one species or a set of closely related species, but lack homologues in other species (Fischer and Eisenberg, 1999). Orphan genes have been reported in several organisms, such as in mouse (Waterston et al., 2002), in which they are 14% of the genes, or in *Drosophila melanogaster* (Zhang et al., 2007), in which as many as 18% of the genes have been found to lack homologues in other insects. Indeed, we have determined that around 3% of the human genes are primate-specific, as no homologues could be found in more distant species (Toll-Riera et al., 2008).

Gene duplication and regulatory evolution have been proposed to be the main players in the diversification of genomes and species, and for this reason evolution has been said to be a tinkerer because instead of creating novelty from scratch it generates novelty by coping and altering existing structures (Tautz and Domazet-Lošo, 2011). However, althought orphan/lineage-specific genes have been poorly studied due to their limited distribution, they have been proposed to drive morphological speciation, facilitating the adaptation of the organism to the changing conditions and species-specific developmental patterns (Khalturin et al., 2009; Kaessmann, 2010; Tautz and Domazet-Lošo, 2011). And now, for the first time, thanks to the high number of available genomes we have the ideal framework to study them. The study of the mechanism of origin of young genes has the advantage that we can study two phenomena at the same time: which are the most frequent mechanisms for gene birth and which are the properties of orphan genes.

The first proposed mechanism for gene formation was gene duplication in the early thirties (Haldane, 1932; Muller, 1935). Later, several other mechanisms have been reported such as exon shuffling, gene fusion, lateral gene transfer, exonization from transposable elements and *de novo* formation from noncoding regions, being gene duplication the most common (Long et al., 2003). Although there is a broad knowledge about which are the mechanisms for gene formation, the contribution of each of them into the birth of orphan/young genes has not been addressed until very recently. By the date only two articles have attempted to figure out the frequency of occurrence of each mechanism. In the first study, performed by Zhou and colleagues (Zhou et al., 2008a), the authors studied the origin of genes in *Drosophila*. And the second article is the main work presented in this chapter, entitled *Origin of primate orphan genes: a comparative genomics approach*.

To perform the study we first had to define a set of orphan genes. We decided to use primate-specific genes; genes were classified in that group if they had homologs in macaque and in chimpanzee, but not in 13 additional non-primate genomes. We obtained a total of 270 human genes classified as primate-specific. Homologues were detected using BlastP searches (Altschul et al., 1997), the pipeline used is described in detail in the methods section of this thesis (section 2.3.1). Although, as previously mentioned, there has been some debate regarding the validity of this methodology (Albà and Castresana, 2007; Elhaik et al., 2006), the use of sequence similarity searches is now widely accepted (Tautz and Domazet-Lošo, 2011). To investigate the possible mechanisms of origin of this type of young genes we performed several analyses: BlastP searches against all human proteins to identify putative paralogs, inspection of the syntenic genomic regions in several mammalian genomes and similarity searches with transposable elements. We have mainly found three mechanisms of origin in primate orphan genes. Around 53% of them have been originated by the exaptation of transposable elements (TE), 24% by gene duplication and about 5.5% from non-coding sequences. 59% of the duplicated genes showed evidences of having arisen by unequal crosser over, 14% by retrotransposition and for the remaining ones we could not discern between both mechanisms. Zhou and colleagues (Zhou et al., 2008a) also studied young genes, but in their case the genes were from the *Drosophila melanogaster* species subgroup. They reported 4 mechanisms of origin: tandem duplication, dispersed duplication, retroposition and *de novo* origination, being gene duplication the most common one. Similarly to us, they also found that *de novo* origination played a more important role than previously thought, finding that around 12% of the genes had originated by this mechanism. Finally, around 10% of their genes had arisen from retroposition. In our study retropostion events are included inside the duplication category, as retroposition is a RNA-based duplication.

A high number of our orphan genes contained TE sequences, we found two common scenarios: a) the exon boundaries of the gene coincided with the TE, which is indicative of exonization of the TE; b) the TE was embedded into an exon. These processes are facilitated because there are potential splice sites in some TEs (Nekrutenko and Li, 2001). However, there were several cases in which other exons of the protein did not contain TEs, indicating the possibility of several additional mechanisms involved in the formation of the gene. This important role of TEs in the formation of orphan genes is surprising, as until now it was estimated that only 4% of the human genes contained TEs (Nekrutenko and Li, 2001). However, TEs, especially Alus, are essential for exon creation in primates (Corvelo and Eyras, 2008). Accordingly, in spite that LINEs (long interspersed transposable elements) are the most frequent TEs in human genome (Lander et al., 2001), the most common TEs in our dataset of primate orphan genes were SINEs (short interspersed transposable elements), the family to which Alu belongs. Alus are primate-specific. Recently, it has been identified a human-specific gene, *FLJ33706*, for which Alus contributed significantly to its formation. There is data that suggests that this gene could be involved in Alzheimer's disease and in nicotine addiction (Li et al., 2010).

Genes originated by gene duplication are not genuine orphans, as they have paralogous that are conserved outside mammals. Those genes could have been initially classified as orphans due to the high divergence experienced, which has masked the similarity (Domazet-Loso and Tautz, 2003). One clear example of highly divergent copy is dermcidin, which is just next to lacritin, it has a similar exonic structure but the two genes have diverged so much that their similarity is very hard to detect. However, it is also possible that homologues could not have been detected because the gene has been formed by a partial duplication instead of a total duplication, which makes homology detection difficult, especially if the duplicated region is short and highly divergent. Partial duplications are more frequent that what it was thought, they have been reported to be involved in 60% of the gene duplications occurred in *C.elegans* (Katju and Lynch, 2006), and they also represent an important fraction of the duplication events in *Drosophila* (Zhou et al., 2008a; Chen et al., 2010). Partially duplicated genes can subsequently adopt surrounding genomic sequences and form chimeric gene structures. The recruited sequences can come from repetitive elements, intronic or intergenic sequences or from the coding regions of other genes. Hence, thanks to these newly acquired sequences, they have more chances to adopt a function immediately after the duplication compared to total duplicated genes, and thus, they have more chances to be retained (Zhou et al., 2008a). Interestingly, it has been seen in *Drosophila* that the fraction of genes arising from a complete duplication diminishes with the gene age, indicating, as commented before, that functional redundant copies have fewer chances to be fixed (Zhou et al., 2008a). In our study of the mechanisms of formation of orphan genes, we have detected that 24% of the cases could have arisen from gene duplication (Toll-Riera et al., 2009a). As we suspected that a high fraction of them could be due to partial duplications we searched for some examples and we presented them in the book chapter found in the section 3.3.4 of this thesis: *Partial gene duplication and the formation of novel genes*. One very interesting example is the FAM9 family. This family is composed by three genes, FAM9A, FAM9B and FAM9C and it has been suggested to be involved in the meiotic prophase. After the duplication, one of the copies, FAM9A, has suffered an expansion of a low complexity region, which could be the responsible of its differential localization in the cell. In all the examples we find that the partial duplicated copy is evolving much faster than the parental one (Toll-Riera et al., 2011a). Consistently, several previous studies have found that duplicated genes evolve faster than non-duplicated ones (Lynch and Conery, 2000; Kondrashov et al., 2002; Scannell and Wolfe, 2008).

The origin of genes from non-coding sequences has been previously reported in *Drosophila* (Levine et al., 2006; Zhou et al., 2008a) and in yeast (Cai et al., 2008) and later we found cases in primate-specific genes (Toll-Riera et al., 2009a). After the publication of our article several additional examples were described in *Drosophila* (Chen et al., 2010), mouse (Heinen et al., 2009), human (Knowles and McLysaght, 2009; Wu et al., 2011), *Plasmodium vivax* (Yang and Huang, 2011) and rice (Xiao et al., 2009), highlighting the importance of this mechanism. The current technology has revealed that nearly all the genome is transcribed; therefore, it seems feasible to think that some short ORFs could be translated into peptides that would be then tested by natural selection and retained if they are advantageous (Toll-Riera et al., 2009a). In a very recent article (Wilson and Masel, 2011) the authors were curious about how novel genes could have evolved from noncoding sequences and studied noncoding transcripts associated with ribosomes in *Saccharomyces cerevisiae*. They found that there was a high number of these transcripts, and a significant proportion of them had ribosomal densities similar to the ones found in coding genes. However, this association was not due to unannoted protein-coding genes. The authors suggested that the results demonstrate the plausibility of *de novo* origin of genes and hypothesized that first noncoding regions are translated at low rates and then they experience a strong selective process in which deleterious polypetides are eliminated. One intriguing question is how noncoding regions can originate foldable sequences. Bornberg-Bauer and colleagues (Bornberg-Bauer et al., 2010) argue that it is imaginable that *de novo* genes acquire a fold because intergenic and genic sequences have similar nucleotide composition, additionally nearly random sequences are very probable functional and also, the folding of fragments might be stabilized by an existing scaffold. Besides, it is known that proteins do not always fold in a unique stable fold, they are very flexible and dynamic and they can assume different structures depending on the ligand. Furthermore, disordered regions have been seen to be very important for protein-interaction. Thus, proteins can tolerate mutations without altering the function, facilitating the action of selection (Bornberg-Bauer et al., 2010).

As suggested by Tautz and Domazet-Loso, genes classified in our study as TE exaptations (142) could be also classified as *de novo*, because most of them contain sequence that do not come from TEs (Tautz and Domazet-Lošo, 2011). Therefore, with this simplification of the classification, only one-quarter of the classified primate orphan genes would have been originated from duplication, while three-quarters would have evolved *de novo*. We have to bear in mind that all the surveys done have been very conservative, and thus, it is very probable that the number of *de novo* genes is an underestimation (Tautz and Domazet-Lošo, 2011). Until now it was always said that proteins arise mainly by tinkering, because existing sequences are used via recruiting and adapting fragments of neighbour DNA or making modular rearrangements of already existing domain combinations (Bornberg-Bauer et al., 2010). However, the high number of orphan genes that show evidence of *de novo* origination demonstrate that tinkering is not the only source of novelty, *de novo* gene emergence also plays an important role in the acquisition of novelties (Tautz and Domazet-Lošo, 2011).

Besides, we have also reported that primate orphan genes have differential characteristics when compared to genes of wider phylogenetic distribution: they evolve faster, they are expressed in a tissue-specific manner and they are usually short. Orphan genes have also been reported to evolve faster in mammals (Albà and Castresana, 2005), yeast (Cai et al., 2006), *Drosophila* (Domazet-Loso and Tautz, 2003; Chen et al., 2010), *Escherichia coli* (Daubin and Ochman, 2004) and rice (Guo et al., 2007). Interestingly, the evolutionary rates of orphan genes in primates (calculated using orthologs between human and macaque) (Toll-Riera et al., 2008, 2009a) showed higher values than orphans genes in mammals (calculated using orthologs between human and mouse) (Albà and Castresana, 2005). This could be possible explained by the reduced purifying selection experienced by hominids compared to rodents as a consequence of their smaller population size (Gibbs et al., 2007). The higher evolutionary rates experienced by orphan genes could be due to the action of positive selection or to the relaxation of the selective constraints, being the later option the most frequent in primate lineage-specific genes (Cai and Petrov, 2010). The shorter size could be due to three possibilities. The first one is that it is a consequence of partial duplication. Secondly, the reduced size is also compatible with *de novo* origination as it seems more plausible that short open reading frames, rather than longer ones, arise from noncoding regions of the genome (Toll-Riera et al., 2008). Lastly, it could be also explained because they are very young and they are still not well-formed. In fact, in the article entitled *The signature of time: younger domains in proteins evolve faster than older ones*, presented in section 3.2.1, we have demonstrated how proteins increase in length with time by the addition/formation of young domains in the 5′ region. Similarly, it has also been reported that with time proteins tend to increase in length and to fold into more complex $\alpha/\beta$ structures (Choi and Kim, 2006).

Little is known about the function of orphan/lineage specific genes, but they have been related with defence against pathogens in primates (Toll-Riera et al., 2009a) and in apicomplexan parasites (Kuo and Kissinger, 2008). In insects they have been implicated in stress, immune response, communication and adaptation to the environment (Zhang et al., 2007) and they have been reported to be more expressed under environmental pressure in rice (Guo et al., 2007).

The interest for orphan/lineage-specific genes has dramatically increased in the last years, one very clear sign is the high number of citations that the main work presented in this chapter has received (31 citations). There are two very recent works which I think that are particularly interesting. In the work presented by Chen and colleagues (Chen et al., 2010) the authors have studied young genes in *Drosophila* in order to evaluate if they are dispensable, as it is expected given that until their birth the organism has managed to live without them. However, surprisingly, they reported that when they knock out (KO) them using RNA interference, 30% of the KOs were lethal, a very similar fraction was found in the control dataset. Additionally, they reported that those genes have mainly arisen by gene duplication and have experienced high evolutionary rates. These results highlight the essentiality of young genes and suggest that young genes are rapidly integrated into existing pathways. I would like to notice that most of the young genes found in this study arise from duplication, therefore, they are not true orphans. However, there were 16 genes originated *de novo*, which could be considered orphans, and 3 of them showed evidences of lethality, reinforcing, thus, the idea of the essentiality of orphan genes. Another interesting study is the one performed by Capra et al. (Capra et al., 2010) in which they bring up the question of whether there are differences in function acquisition and network integration among genes originated by different mechanisms, given that, for example, genes created through duplication are well formed at birth while genes originated *de novo* are not. They compared genes originated by duplication with those that did not (named novel genes by the authors) finding that, initially, duplicated genes are more integrated into the network, but with time, novel genes gain function and interactions more rapidly. Consistently, they also reported that novel genes increase in length by gaining sequences from the surroundings and from TE, hypothesizing that this gain in sequence could be the cause of their rapid gain of function and interactions. Interestingly, they found that genes tended to interact with those genes originated by the same mechanisms. Therefore, this article shows that the mechanism by which the genes have originated has an influence in their subsequent evolution and integration in cellular networks.

Additionally, a study analyzing the mechanism of origin of lineage-specific genes in *Arabidopsis thaliana* has been recently published(Donoghue et al., 2011). The authors reported that almost 25% of those genes have arisen from gene duplication, 10% showed evidence of transposon exaptation and around half had alignments to intergenic regions in *Arabidopsis lyrata*, suggesting two possible scenarios: *de novo* origination or differential retention and loss. Consistently with our study, the authors also reported high tissue specificity in lineage-specific genes. Among lineage-specific genes they also found an enrichment for genes involved in stress response, reinforcing the idea that lineage-specific genes are vital for the adaptation to new environments.

However, there is controversy over the real existence of orphan genes. There are some authors that claim that most human ORFs which are not conserved among mammals are spurious (Clamp et al., 2007). For this reason we paid exceptional care when we defined our dataset of primate-orphan genes. First, we only considered those human orphan genes with defined orthologs in *Pan troglodytes* and *Macaca mulatta*. Second, we verified that the codon usage in human orphan genes was similar to the one found in human protein-coding genes and was different from the one found in a set of noncoding RNAs and in the noncoding frames of the orphan genes. Third, most of them had expression data. Fourth, the characteristics of the subset of orphan genes experimentally validated were similar to the non-validated orphan genes. And finally, a high fraction of the orphan genes belonged to protein families which included non-primate homologues.

To sum up, recent work in the field of orphan genes and their mechanism of origin has shed light into which are the sources of novelty and add orphan genes into the previous proposed drivers of innovations: gene duplication and regulatory evolution. Interestingly, *de novo* origin of genes seems to play a more important role than previously thought, whereas, the importance of gene duplication could have been overestimated.

## 4.4 Methodological issues

In this section I am going to report some important methodological pitfalls that I found while doing the thesis as well as other issues that should be taken into account when doing genome-wide comparative genomics studies.

### 4.4.1 The importance of good alignments and high quality data

A high fraction of the current studies in the comparative genomics field use alignments as a starting point and therefore, most of the results rely on them. It is impossible to know the true alignment, but approximations could be obtained by using one of the several available aligners. However, alignments are not usually questioned once done, they are treated as observations, and the analysis goes on without controlling for potential misalignment-related errors. The importance of the errors derived from the alignment depends on the objective of the study, for example, in a phylogenetic study where the gene is carefully chosen the alignment errors would be minimal. However, in a comparative genomics study using thousands of genes, the analysis is automated and repeated several times, and consequently the alignment can not be carefully revised, and, therefore alignment uncertainties can be important (Wong et al., 2008; Markova-Raina and Petrov, 2011). Most of the analysis methods were designed for carefully constructed single alignments, but now the same methods are applied to large datasets, in which the alignments are automatically done without manual inspection.

Wong and colleagues (Wong et al., 2008) used yeast species to assess how uncertainties in the alignments affect evolutionary analysis. One of the first steps on most analysis is the identification of orthologues, and it is really important that it is done in a correct way because aligners do not question it, they try to align everything. The authors performed alignments using seven different programs (ClustalW, Muscle, T-Coffee, Dialign 2, Mafft, Dca and ProbCons) and then they performed pylogenetic and positive selection analysis. They reported that, as expected, both types of analysis are sensitive to the aligner used. They found that nearly half of the used ORFs differ in the trees depending on the used aligner. The substitution rate estimations did not differ greatly among aligners, however, 28% of the positively selected sites were sensitive to the aligner used. The authors argued that this is not due to the aligners, but to the underlying variability in the processes of insertion, deletion and substitutions of some particular ORFs, that make them more difficult to align. Most of the studies are well designed and apply severe filters but the problem relies in the fact that the methods used for the analysis do not take into account alignment uncertainties. Additionally, usually the most interesting genes are the ones that have diverged the most, and, precisely, those ones are the most difficult to align. The authors also believe that alignment uncertainties could not be resolved by discarding genes or fragments of them because they are informative. For example, discarding positions with gaps excludes positions in which a insertion occurred in the other lineage. Moreover, when they performed the phylogenetic analysis discarding the gapped positions from the alignment they still found differences among aligners. In statistics the parameter uncertainty is usually treated as a random variable, for this reason the authors propose to treat alignments as random variables and infer the posterior parameters taking into account the different alignments according to their probability.

The branch-site model test, implemented in the PAML package, is specially designed to detect episodic positive selection on a few sites and in particular lineages. The authors of this test initially tested it under idealized conditions, but as these conditions are never reached in common studies, they tested it again but taking into account insertions, deletions and alignment errors (Fletcher and Yang, 2010). They reported that if the alignment is correct, the presence of insertions and deletions does not cause false positives in the test. They used 4 different aligners: Prank+F, Muscle, Mafft and ClustalW, and they observed that the amount of false positives brutally increased with the presence of alignment errors. When the divergence between sequences decreased, the false-positive rate also decreased. Although Prank+F outperformed the other aligners, it still produced a too high rate of false positives. The alignment accuracy they found is as follows: Prank+F >Muscle & Mafft >ClustalW. This differential accuracy is caused because ClustalW, Muscle and Mafft do not deal correctly with insertions, as they penalize multiple times the same insertion event during the progressive alignment algorithm (Löytynoja and Goldman, 2008). The authors also tried to remove gaps from the alignments before applying the branch-site test, but false positives were only slightly reduced. The use of Prank+F significantly reduced the fraction of positive selected sites estimated in two previous studies using other aligners. Prank+F is working better than other aligners because it does a more correct handling of gaps. By introducing more gaps it does not place non-homologous codons in the same column as frequently as the other algorithms. This becomes evident when alignment lengths are compared, Mafft, Muscle and ClustalW produce shorter alignments than Prank+F (Fletcher and Yang, 2010). When nonhomologous codons are placed in the same column the branch-site test would be misleading because it would interpret it as an excessive amount of amino acid changes at those sites.

Several articles have attempted to estimate the fraction of positively selected genes in the genomes, for example, estimates in human vary in three orders of magnitude, ranging from 0.02% to 8.7%. Schneider and colleagues (Schneider et al., 2009) have centred on assessing the number of inferred positive selected genes depending on the applied filters. They used orthologous protein-coding genes from human, chimpanzee, macaque, mouse, rat, dog and cow. They performed the alignments using the Darwin multiple sequence alignment package and used branch-site model A of the PAML package to identify positive selection. They found that the fraction of genes predicted to experience positive selection was higher in the set of genes with lower coverage, with inferred annotation status and in the set containing alignments with ambiguities. These effects are not cumulative, but all inflate the number of genes with positive selection evidence. Therefore, their results indicate that the proportion of positively selected genes increase with the decrease of the quality of the sequence. When only good quality sequences were used, the proportion of positively selected genes decreased dramatically. Hence, one should be cautious when inferring positive selection. Mallick and colleagues performed a similar study (Mallick et al., 2009). They based their work in previous articles reporting more positive selection on the chimpanzee lineage than in the human lineage (Bakewell et al., 2007; Gibbs et al., 2007) because they suspected that those results could be an artefact caused by the lower quality of the chimpanzee sequence, a fact that is worsen by the small divergence between both species. Despite of most of the bases being correct, if some errors are clustered in some specific codons an artefact signal for positive selection can appear. As genome scans take into account thousands of genes, some of the positively selected genes could be false positives. They reexamined the cases of genes detected to be positively selected using a new bioinformatics approach that generated high reliable aligned bases with the cost of losing some exon coverage. As they suspected, they were not able to replicate the results, specially the ones from Bakewell and colleagues. This was mostly because a high fraction of the sites experiencing positive selection fall into low sequence quality regions. Therefore, the quality filters applied in those studies seem to not be enough to discard a reasonable fraction of false positives. Hence, the results presented by Mallick et al.(Mallick et al., 2009) go against the hypothesis that positive selection has been more effective in chimpanzees than in humans and reinforces the idea that strong filters should be applied in order to reduce the rate of false positives and to avoid wrong conclusions. They also suggest that it is a good idea to resequence some of the loci predicted to be positive selected to validate the bioinformatics pipeline.

All the previous reports about difficulties when inferring positive selection where done using branch-site models. Markova-Raina and Petrov (Markova-Raina and Petrov, 2011) used six different popular aligners to assess their sensitivity and rate of false positive when positive selection is estimated based on site-specific divergence models. The chosen aligners where: Prank+F, T-Coffee, ClustalW, ProbCons, Amap and Muscle and they used data from the 12 *Drosophila* genomes. The number of positively selected genes varied as much as 60% depending on the used aligner. The situation did not improve when the regions with gaps were removed from the alignments. They also tried to use quality controls and more closely related species, but even doing that, half of the genes inferred to be under positive selection were possible false positives. They reported, as in the previous commented studies, that Prank+F, which considers evolutionary information, was the best performing aligner, although, it had still a too high false positive rate (50-55%). They also found that the number of positive selection events was very low in those alignments that were consistent among all the aligners. The detection of Gene Ontology (GO) terms under or overrepresented was also affected by the aligner choice. They listed the most common causes of misalignments: bad annotation in CDS start and end, alternative splicing, incorrectly annotated intron positions, repeats and differential annotation of exon boundaries.

### 4.4.2 Difficulties in the methods to detect positive selection and estimate evolutionary rates

There are a high number of studies inferring positive selection (Arbiza et al., 2006; Bakewell et al., 2007; Clark et al., 2003; Kosiol et al., 2008). However, several difficulties have been suggested.

Independently from the alignment and quality issues there are also some difficulties when inferring positive selection regarding the methods used. Nozawa and colleagues (Nozawa et al., 2009) performed computer simulations to study the reliability of the methods to estimate positive selection using a set of vertebrate vision genes. They showed that branch-site methods give false prediction of positive selection when the used foreground branch has a small number of nucleotide substitutions. Besides, they also reported that there are differences depending on the method used, HyPhy or PAML. They also found wrongly identified positively selected sites when multiple nonsynonymous substitutions took place in the same codon. Most statistical methods search for codons with high $d_N/d_S$ values to infer positive selection and adaptation. However, when experimentally determined functional changes are examined, most of them do not show a high $d_N/d_S$ value, because, normally, a functional change occurs by the replacement of one amino acid by another one in one or a few positions. Therefore, the current methodology often identifies false positively selected sites. Ideally, some experimental confirmation, such as site-directed mutagenesis, should be performed before inferring adaptive evolution. However, the authors highlight that the use of $d_N/d_S$ to measure the selection acting on an entire gene (evolutionary rate) is valid, as the ratio is calculated using the average rates of nonsynonymous and synonymous substitutions for the whole gene.

Hughes (Hughes, 2007) is very sceptic about the studies trying to infer positive selection. He thinks that the current statistical approaches are flawed. He poses that McDonald-Kreitman test has problems when there are changes over time in the rate of synonymous substitutions, when there is recombination and also, is not capable of distinguishing between positive selection and relaxation of purifying selection, which could be important when species have suffered a bottleneck (during bottleneck purifying selection is less effective in removing slightly deleterious mutations). He thinks that the situation does not improve with the use of methods that use phylogenies to infer the pattern of nucleotide changes at codons because for most genes under positive selection the phylogeny is usually very difficult to reconstruct. Additionally, these methods assume that if a codon has a $d_N > d_S$ is indicative of having experienced positive selection. However, Hughes argues that this could also be due to chance. And another important problem that he suggests is that those methods are applied without any a priori biological hypothesis, and frequently, the patterns of those codons predicted to be positively selected could be rather due to a relaxation of purifying selection or to the absence of synonymous substitutions. Therefore, statistical evidence of positive selection for a gene can not be used as a proof of adaptive evolution.

Natural selection is inferred when it favours repeated changes at the amino-acid level, therefore $d_N > d_S$. One of the most well-known examples is the vertebrate major histocompatibility complex (MHC). But this example is unique because it was discovered after a biological hypothesis was set. Additionally, positive selection only occurs repeatedly in the peptide-binding-region due to a co-evolutionary process with pathogens. Therefore, there is no reason to think that the pattern of selection experienced by MHC could be generalized. On the contrary, there are evidences that a single amino-acid substitution, instead of a serie of substitutions, may produce an adaptive phenotype. This is illustrated in the beach mice, where a single amino-acid replacement in the melanocortin-1 receptor changes the coat color to be more similar to the sand (Hoekstra et al., 2006). Hughes argues that this particular example would have been missed by the current statistical methods to detect positive selection, and he notices that there are no statistical tests to detect adaptive evolution involving a single nonsynonymous substitution (Hughes, 2007).

Hughes and Friedman (Hughes and Friedman, 2008), after expressing their worries about positive selection tests, performed a study using empirical data of seven species of mammals in which they calculated the number of synonymous and nonsynonymous substitutions and applied the branch-site test. First they showed, using probability theory, that when branch lengths are very short it is very probable that $d_S$ is nearly 0 or 0, therefore, there will be a nonzero probability that $d_N > d_S$ by chance, also in the presence of strong purifying selection. Using a dataset of mammalian orthologs they found that the occurrence of genes with $d_N > d_S$ was really more frequent in the shorter branches: the primate ones. Thus, for most cases the explanation should be a high stochastic error on short branches rather than positive selection. When they used the branch-site method implemented in the PAML package they observed that positive selection was usually associated with very low $d_S$ and very high $d_N$ values, compared to other branches of the trees, suggesting that chance plays an important role. Additionally, for most of the codons identified to be evolving under positive selection, no synonymous changes were detected, suggesting that it is probable that the high $d_N/d_S$ ratio is due to low $d_S$ rather to a high $d_N$. The low $d_S$ could be explained by the action of purifying selection on the synonymous sites. Some cases have been reported. Thus, the action of purifying selection over synonymous sites should be discarded before inferring positive selection (Chamary et al., 2006).

*Microcephalin* and ASPM are two genes that play a role in adult brain size in humans. Positive selection studies have detected elevated $d_N$ values in both of them, which has lead the scientists to hypothesize that is due to adaptive evolution for increased brain size in primates (Evans et al., 2004; Kouprina et al., 2004). There is no biological evidence to hypothesize this, thus, before claiming positive selection, the other possibility, relaxation of purifying selection, should be ruled out. Hughes and Friedman did not find evidences of the action of positive selection in those genes. The contrary situation is found in immune genes. It is known that immune genes have accelerated rates of non-synonymous substitutions, and there is a biological hypothesis behind it. However, neither the $d_N/d_S$ or the branch-site methods were able to detect it. The authors pose that these examples demonstrate that the methods mainly detect statistical artefacts, rather than adaptive evolution.

As explained in the methods section (page 47), it is very important to choose the correct genomes to avoid the mutational saturation of the sequences. However, it has been reported that even in high quality genes, there is a strong positive correlation between the number of positively selected genes and branch length, indicating that despite the effort done in the methods to take into account the saturation in synonymous substitutions, they are still underestimated, which causes the overestimation of the $d_N/d_S$ ratio (Schneider et al., 2009).

The last but not the least important question regarding evolutionary rate estimations and the detection of positive selection is the assumption that synonymous substitutions are neutral. There are several examples of non-neutral evolution of synonymous substitutions. For example, in bacteria, yeast, flies, worms and plants the usage of synonymous codons has been found to be biased, especially in highly expressed genes. However, as mammalian species have small population sizes, it has been always assumed that synonymous mutations are 'effectively neutral'. But there is some data supporting a weak relationship between codon usage and gene expression in mammals. Also selection seems to act when synonymous mutations affect the mRNA stability or when they disrupt the splicing process by altering intron removal. Synonymous mutations have been proposed additionally to affect protein folding and RNA editing. Thus, although most of the synonymous mutations are neutral, it has been estimated that selection can operate in as many as 40% of them (Chamary et al., 2006).

Several authors call into question that the $d_N > d_S$ pattern is a signature of positive selection, and thus, that it can be used to discover genes that have experienced positive selection in the past. They argue that most of the codons showing $d_N > d_S$ are due to chance. Therefore, we should be cautious when extracting conclusions from results from genome-wide scans for positive selection, unless there is biological evidence and experimental validation behind them. However, it is important to study positive selection because adaptive changes are essential to understand species differences and evolutionary innovations. The key is to combine phylogenetic analyses with well designed analysis and experimental validation.

### 4.4.3 GC-biased gene conversion and positive selection

When accelerated evolution is detected positive selection is usually claimed. However, there are other known processes that can lead to an accelerated evolution and one clear example is GC-biased gene conversion (gBGC). Gene conversion is the nonreciprocal transfer of genetic information between homologous sequences, and is involved in meiotic recombination. gBGC is a process associated with recombination that produces a biased fixation of G and C nucleotides over A and T, thus, an AT/CG heterozygote will have more gametes carrying G or C than A or T (Galtier and Duret, 2007). The proposed mechanism is a bias towards the incorporation of G/C nucleotides during the repair of mismatches in the heteroduplex DNA intermediates formed in meiotic recombination (Duret and Galtier, 2009). In fact, it has been observed in mammals that the repair of DNA mismatches in mitotic cells is highly GC-biased, probably reflecting an adaptation to deal with the hypermutability of methylated cytosines (Duret and Galtier, 2009). gBGC influences GC content dynamics in the mammalian genome (Galtier et al., 2009; Ratnakumar et al., 2010). In primates there are two lines of evidence that show that gBGC could be influencing genome evolution. Firstly, the rate of AT → GC nucleotide substitutions is strongly influenced by long-term average recombination (Meunier and Duret, 2004). And, secondly, using polymorphism data it was observed that AT → GC mutations segregate at higher frequency that GC → AT (Webster and Smith, 2004), and this bias was found to be higher in regions experiencing high recombination (Spencer, 2006). Human recombination typically occurs in hotspots that are not conserved between human and chimpanzee, indicating that they have a very short evolutionary lifespan (Galtier and Duret, 2007; Ratnakumar et al., 2010).

Besides, gBGC could result in the fixation of slightly deleterious AT → GC substitutions in functional sites, and this, could be erroneously attributed to positive selection. In fact, gBGC is equivalent to directional selection because GC → AT mutations have more chances to be passed to the next generation and be fixed. As recombination usually takes places in hotspots and hostpots have short lifespan, gBGC is going to cause local and transient bursts of substitutions, similar to selection (Galtier and Duret, 2007). Moreover, it can also avoid the fixation of advantageous GC → AT substitutions (Galtier et al., 2009). Additionally, it has been shown that gBGC can cause an increase in the $d_N/d_S$ ratio because nonsynoymous codon positions have lower GC content than synonymous codon positions. Hence, more AT → GC substitutions can occur on nonsynonymous sites, increasing the $d_N$ respect the $d_S$ (Ratnakumar et al., 2010). There are three features that allow discerning between positive selection and gBGC (Ratnakumar et al., 2010) (table 4.1):

- gBGC produces AT → GC bias, but not selection
- gBGC operates on functional sites but also in flanking neutral sites, while selection only operates on functional sites.
- gBGC is associated with regions with high male recombination, but not selection.

In order to know how gBGC affects positive selection, Ratnakumar and colleagues (Ratnakumar et al., 2010) analyzed the data arising from a scan for positive selection in primates. They found that the fastest evolving human and chimpanzee genes had elevated recombination rates, were closer to recombination hotspots, enriched in subtelomeric regions and have elevated levels of male recombination. They also reported that genes identified to be positively selected using branch-site methods have an elevated GC content, also in the flanking non-coding regions, results that are highly consistent with a regional effect of gBGC. 14% of the genes belonging to primates were located in the high GC category, and this fraction increased to 22% when shorter branches were taken into account. Those genes are candidates to be subject to gBGC and not to positive selection. The higher presence in shorter branches is consistent with the short lifespan of recombination

| Criterion | gBGC | Adaptation | Test |
|---|---|---|---|
| Target sites | All | Functional sites | Reject selection if non-functional sites are involved |
| Substitution pattern | GC biased | No systematic bias | Reject selection in case of strong bias; reject gBGC in case of AT bias |
| Relationship with recombination | Strong | Weak | Favour gBGC if substitution hotspots are concentrated in regions of high recombination |
| Selective sweep | Yes, no hitchhiking | Yes, with hitchhiking | Reject gBGC if coalescence based neutrality tests applied to flanking regions are positive |

**Table 4.1:** Criterions to discern between gBGC and adaptation. Adapted from Galtier and Duret (2007)

hotspots. Surprisingly, they found that theoretical modelling indicates that in some particular cases, gBGC can produce $d_N/d_S$ values as high as 2. Therefore, their results showed that gBGC is affecting the evolution of primate coding sequences, and thus, could be confounding positive selection tests. However, genome scans taking into account long evolutionary times should be robust to the transient episodes of gBGC, then, being gBGC mainly a problem when positive selected genes are tried to be identified in short branches, for example, in the search of human-specific adaptations. In a similar study the authors have studied accelerated primate exons (Galtier et al., 2009). More acceleration events were found in chimpanzee and in the ancestral branches (human-chimpanzee, human-chimpanzee-orangutan) and they argued that this could be as a result of gBGC being more efficient in larger populations. The percentage of AT → GC changes was higher in accelerated branches, being 19 episodes significantly GC-biased at 1% level. Those exons tended to be located in high-recombining regions, had a higher median crossover rate in the male germline and the synonymous changes were biased towards GC. All these results are consistent with the action of gBGC, rather than positive selection.

Several studies have searched for highly conserved noncoding elements across vertebrates but divergent in humans, reporting several human accelerated regions (HARs). These acceleration specific to human had been interpreted in adaptive terms. However, when those regions are inspected in detail it is revealed that most of the changes are AT → GC, being this pattern extended into flanking regions, and most of those regions are located in high-recombining regions of the genome. Thus, all the evidences suggest that an important fraction of HARs are functional regions under negative selection that have experienced strong gBGC due to their location in recombination hotspots (Galtier and Duret, 2007).

Therefore, protein evolution is influenced by gBGC, which can lead to accelerated evolution through the fixation of AT → GC mutations in conserved exons. Thus, gBGC can counteract purifying selection and promote the fixation of deleterious amino acid mutations mainly in recombination hotspots. Is for this reason that recombination hotspots have been said to be the Achilles' heels of our genome (Duret and Galtier, 2009). The fixation of deleterious alleles caused by gBGC can be followed by positively selected compensatory substitutions to restore protein function (Galtier et al., 2009). In conclusion, one should bare in mind that gBGC can be corrupting positive selection tests, and, hence, selective hypothesis should only be formulated after neutral and gBGC models have been discarded.

### 4.4.4 Methodological issues applied to section 3.1.1

In the article enclosed in section 3.1.1 (Toll-Riera et al., 2010) we took into account several of the methodological issues commented above. In an attempt to minimize the number of incorrectly aligned homologous positions and in order to not overestimate the number of nonsynonymous substitutions we performed multiple alignments using Prank+F. We used the set of mammalian genomes that was best characterized at the time. However, as we were aware that the genomes had different degrees of gene annotation reliability we decided to apply a set of rigorous filters:

- We discarded those orthologous gene families that included a sequence with ambiguous amino acids.

- We eliminated orthologous gene families with one of the sequences being shorter than half the length of the longest sequence.

- We discarded very short alignments.

- We discarded trees with branches with a $d_S > 2$ because they could indicate saturation of the synonymous substitutions. We also discarded those trees that had $d_N > 2$ because they could be due to the inclusion of non *bona fide* orthologs. Finally, we also discarded those cases with $d_S < 0.01$ as they corrupted the estimates of $d_N/d_S$ and could be indicative of the action of purifying selection at the synonymous sites.

- To avoid the alignment of non-orthologous exons caused by the incompleteness of transcript annotations and to the presence of incorrectly annotated genes, we discarded those alignments in which the region orthologous to an specific exon had an overall sequence similarity smaller than 50%.

As we did not have any prior biological hypothesis on the presence of positive selection in particular genes, and several branches were tested for positive selection, we applied the q-value test to correct for multiple testing.

Despite all the applied filters, the proportion of genes with a significant signal of positive selection varied among branches, being macaque the branch with a higher fraction. As there is no biological reason to expect more positively selected genes in the macaque genome, this higher number is possible due to the low quality of the macaque genome assembly. We also observed that genes classified as positively selected using the branch-site test tended to have higher $d_N/d_S$ values than average, but this was not the case for the genes that were detected to be accelerated using our methodology. This is telling us that adaptive changes taking place in very slowly evolving genes will not usually be detected with the branch-site tests in PAML. One very clear example is the GRIN3A, which is a very well conserved protein that, in spite of having seven human-specific nonsynonymous changes, is not detected as positively selected with the branch-site test.

Thus, although we have paid special attention to the methodology used, there are still some results for which the most reliable explanation is a methodological issue. Branch-site tests are very popular and widely used, but we have evidence that they are not detecting several events of evolutionary rate acceleration that could be related with lineage-specific functions. Thus, an effort should be made to improve the quality of the genome annotations and the sensibility of the branch-site tests.

# 5

# Conclusions

1. We have developed a novel method that allows the identification of lineage-specific variation in the intensity of natural selection. Quality filters should be employed in genome-wide studies to reduce the probability of obtaining spurious results.

2. Around 25% of the one-to-one orthologous genes in six mammalian species show branch-specific evolutionary rate deviations.

3. Genes showing branch-specific evolutionary rate acceleration are enriched in neural proteins, suggesting that they play an important role in species diversification.

4. Younger protein domains evolve significantly faster than older domains in mammals and flies, confirming the previously observed inverse relationship between protein age and evolutionary rate. This correlation is also found in proteins containing domains classified in different age groups, demonstrating that age is a key determinant to explain protein evolution.

5. Older proteins are longer and have a higher number of domains than younger proteins. Young, vertebrate-specific, domains are usually gained at the N-terminus of older proteins.

6. There is a significant positive linear relationship between residue solvent accessibility and evolutionary rate in mammalian protein structures. When we compare residues with similar solvent accessibility located in proteins of different age we observe that they tend to evolve faster in younger proteins. Younger protein structures have a higher fraction of solvent-exposed residues and are evolving faster than older ones. However, in each age group, proteins with a higher number of buried residues evolve more rapidly, probably because they are more designable.

7. For residues that belong to the same solvent accessibility (buried, exposed) or secondary structure (helix, sheet, turn, coil) group, there is an inverse relationship between the age of the structure and its

evolutionary rate, showing that biases in the secondary structure and solvent accessibility properties can not explain the observed differences in evolutionary rates between age classes.

8. Old structures are more stable and robust to mutations than young structures; therefore, structures may acquire stability and robustness over time.

9. Around 3% of human genes are primate-specific. Those primate-specific genes have differential characteristics; they generally evolve very fast, are shorter and are highly tissue-specific.

10. Around 24% of the primate-specific genes have arisen from gene duplication (including partial duplication), 53% of them from the exaptation of transposable elements and around 5.5% of the genes *de novo* from noncoding mammalian genomic regions.

# 6

# Future Research

In this section I am going to briefly mention some ideas to follow the research presented in this thesis. I am also going to give some insights about what could be done in the field with the current advances in sequencing technology.

In the first chapter of the results section a new methodology that allows the identification of lineage-specific deviations in the evolutionary rates is presented. The work was performed in mammalian, obtaining that around 25% of the studied orthologous genes, a well-conserved set, showed branch specific deviations. One intriguing question is if that fraction is related with the species used or if it is a general trend. The number of sequenced genomes has brutally increased in the last years, thus, the same study could be performed using several different groups of species such as insects, vertebrates, and even more interesting, could also be applied to extremophile Achaea or bacteria species and compare all the fractions of deviated genes obtained in those species and try to relate them with the environment and with adaptive mechanisms. Besides, other interesting research lines that could be included in the framework of this chapter are related with the development of improved methodologies to identify positive selection even when a small number of changes are involved.

Although a significant effort has been done to decipher which are the determinants for protein evolution it is still not clear if there is only one determinant governing protein evolution or if protein evolution is determined by an interplay of several factors. Structure could be a key factor, but there are still relatively few structures solved. Nonetheless, an article has appeared recently that reports the use of an X-ray free-electron laser to determine structures without the need to crystallize them (Barty et al., 2011). Besides, a justification should be found to explain why protein age seems to be playing such an important role in determining protein evolution. We have hypothesized that positive selection for stabilizing mutations can be playing a key role, but this should be tested.

In the last years the study of orphan genes has become very popular; but there are still some questions that remain to be answered. Orphan genes have typically been associated with high evolutionary rates, however, there are also orphan genes evolving slowly, and little attention has been paid to them. It would be very interesting to perform a study comparing the characteristics of orphan genes situated in the two tails of the distribution of the evolutionary rates. Do they differ in their functions? In the expression pattern? Or in their mechanism of origin? As mentioned by Domazet-Loso and Tautz (Tautz and Domazet-Lošo, 2011) it would be also very exciting to study the protein structures of orphan genes, do they fold into known protein folds? can they be crystallized? Are they unstructured? The study of their structural properties could give insights to understand how a *de novo* sequence could have evolved to become functional as well as to test if protein folds converge. A known function is missing for most of the orphan genes, and to understand which role orphan genes play in adaptive innovations it is essential to determine it. Moreover, which are the mechanisms of origin of orphan genes has only been studied in depth in few species, and therefore, general conclusions about their contribution have to wait until more species are scanned and more hypothetical mechanisms are studied in detail. Besides, several questions regarding the proposed mechanisms remain to be answered, for example, how frequent is partial gene duplication? Is is more frequent than complete gene duplication?

The tremendous advances in the field of genomics with the advent of next generation sequencing techniques open a wide range of opportunities for evolutionary studies. One of the most interesting fields is experimental evolution, which is the ideal framework to test hypothesis and theories of evolution. In those experiments population replicates are propagated for several generations in the laboratory under different controlled conditions. In this controlled environment evolution can be directly observed while populations adapt to new environmental conditions or suffer changes due to genetic drift. Afterwards, the mutations that caused the adaptation can be identified by sequencing the evolved stains and comparing them to the ancestral one. As several generations are needed, the most commonly used species are microorganisms (Buckling et al., 2009). With the use of experimental evolution some long-standing hypothesis have been tested, such as the accelerated evolution experienced when host and parasites coevolve (Paterson et al., 2010). The use of experimental evolution can also shed light into a modern version of the neutralism-selectionist debate. For example, a very interesting and recent study has demonstrated how conditionally neutral (or cryptic) mutations can facilitate rapid posterior adaptation, highlighting the role of epistasis and robustness in adaptive evolution (Hayden et al., 2011). Several other exciting questions related with robustness, evolvability and epistasis could be addressed taking advantage of the combination of experimental evolution and next generation sequencing, for example, are there really neutral mutations? Does the fraction of neutral and beneficial mutations depend on the molecule?

Until now, the best studied organisms have been model organisms, such as mouse, fly and *C. elegans*. Fortunately, the advances in sequencing technology give us also the opportunity to sequence and study in detail nonmodel organisms. This could facilitate and increase our knowledge about fascinating questions such as how some species can survive and adapt to extreme conditions or why some other species seem almost not to evolve, such as the coelacanth.

# 7

# Annex

## Peer-reviewed publications

**Toll-Riera M**, Bostick D, Albà M.M, Plotkin J.B. Structure and age jointly influence rates of protein evolution. (Submitted)

**Toll-Riera M**, Albà M.M. The signature of time: younger domains in proteins evolve faster than older ones. (Submitted)

Laurie S, **Toll-Riera M**, Radó-Trilla N, Albà M.M. (2012). Sequence shortening in the rodent ancestor.*Genome Res*, 2012 Jan 9

**Toll-Riera M**, Radó-Trilla N, Martys F, Albà M.M. (2011). Role of low-complexity sequences in the formation of novel protein coding sequences. *Mol Biol Evol*. 2011, Dec 8.

**Toll-Riera M**, Laurie S, Albà M.M. (2011). Lineage-specific variation in intensity of natural selection in mammals. *Mol Biol Evol*, 28(1):383-98

Mularoni L, Ledda A, **Toll-Riera M**, Albà M.M. (2010). Natural selection drives the accumulation of amino acid tandem repeats in human proteins. *Genome Res*, 20(6):745-54

**Toll-Riera M**, Castelo R, Bellora N, Albà M.M. (2009). Evolution of primate orphan proteins. *Biochem Soc Trans*, 37(Pt 4):778-82. Review

**Toll-Riera M**, Bosch N, Bellora N, Castelo R, Armengol L, Estivill X, Albà M.M. (2009). Origin of primate orphan genes: a comparative genomics approach. *Mol Biol Evol*, 26(3):603-12

Mosquera M, Llorente M, Riba D, Estebaranz F, Gonzalez-Brao M, Lorenzo C, Sanmarti N, **Toll M**, Carbonell E, Feliu O. (2006). Ethological study of manual laterality in naturalistic housed chimpanzees (Pan troglodytes) from the Mona Foundation Sanctuary (Girona, Spain). *Laterality*, 12:19-30

## Book chapters

**Toll-Riera, M.**,Laurie, S., Radó-Trilla, N., Albà, M.M.(2011). Partial gene duplication and the formation of novel genes. In *Gene Duplication / Book 1*, Intech

**Toll-Riera, M.**, Castresana, J., Albà, M.M. (2008). Accelerated Evolution of Genes of Recent Origin. In *Evolutionary Biology from Concept to Application*. Ed. Pontarotti. Springer: Berlin

Mularoni, L., **Toll-Riera, M.**, Albà, M.M. (2008). Comparative Genetics of Trinucleotide Repeats in the Human and Ape Genomes. In *Encyclopedia of Life Sciences* John Wiley & Sons Ltd: Chichester

## Conference attendance

9th CRG Annual Symposium: Medical Genome Sequencing: Understanding the Genomes of Disease. Parc de Recerca Biomèdica de Barcelona (PRBB), October 28th- 29th, 2010. Barcelona

The Encode Project ten years after the human genome sequence. Parc de Recerca Biomèdica de Barcelona (PRBB). July 20th , 2010. Barcelona

Society for Molecular Biology & Evolution (SMBE) Annual Meeting. July 3rd- 8th, 2010. Lyon, France
Poster: **Toll-Riera M.**, Albà M.M. Evolution of protein domains in vertebrate proteins
Oral communication: **Toll-Riera M.**, Laurie S., Albà M.M.. Non-clock evolution of mammalian proteins

Evolution: the molecular landscape. Cold Spring Harbor Laboratory. May 27th- June 1st, 2009. New York, USA.
Poster: **Toll-Riera M.**, Laurie S., Albà M.M. Changes in the evolutionary rates along the mammalian phylogeny.

Protein evolution-sequences, structures and systems. Wellcome Trust Conference Center. January 26th- 27th, 2009. Hinxton, UK
Poster and oral communication: **Toll-Riera M.**, Bosch N., Bellora N., Castelo R., Armengol Ll., Estivill X., Albà M.M. Evolution of primate orphan proteins

Society for Molecular Biology & Evolution (SMBE) Annual Meeting. Barcelona. June 5th-8th, 2008.

Poster: **Toll-Riera M.**, Bosch N., Bellora N., Castelo R., Armengol Ll., Estivill X., Albà M.M. How do novel genes arise? Insights from mammalian genome sequence comparisons

5th annual meeting of the Spanish Primatology Association ”Primatology at 21th century”. September, 16th- 20th, 2003
Poster: M. Llorente Espino, O. Feliu Olleta, E. Carbonell i Roura, M. Mosquera Martínez, D. Riba Cano, C. Lorenzo Merino, V. Celiberti, I. álvarez de Quevedo i Gispert, M. Martínez Masagué, **M. Toll i Riera**, F. Estebaranz i Sánchez, L. Dotras Navarro, N. Sanmartí Boixeda, M. Trueba Gutiérrez, M. González Brao, M. Puig McLean, D. Santos Fita, I. Martín Hurtado y A. Sevilla Garreta. Proyecto Mona: rehabilitación, resocialización y deshumanización de primates decomisados

# References

Albà, M. M. and Castresana, J. (2005). Inverse relationship between evolutionary rate and age of mammalian genes. *Molecular biology and evolution*, 22(3):598–606.

Albà, M. M. and Castresana, J. (2007). On homology searches by protein Blast and the characterization of the age of genes. *BMC evolutionary biology*, 7:53.

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research*, 25(17):3389–402.

Apic, G., Gough, J., and Teichmann, S. A. (2001). Domain combinations in archaeal, eubacterial and eukaryotic proteomes. *Journal of molecular biology*, 310(2):311–25.

Arbiza, L., Dopazo, J., and Dopazo, H. (2006). Positive selection, relaxation, and acceleration in the evolution of the human and chimp genome. *PLoS computational biology*, 2(4):e38.

Arguello, J. R., Chen, Y., Yang, S., Wang, W., and Long, M. (2006). Origination of an X-linked testes chimeric gene by illegitimate recombination in Drosophila. *PLoS genetics*, 2(5):e77.

Ayala, F. J. (2000). Neutralism and selectionism: the molecular clock. *Gene*, 261(1):27–33.

Bailey, J. A., Gu, Z., Clark, R. A., Reinert, K., Samonte, R. V., Schwartz, S., Adams, M. D., Myers, E. W., Li, P. W., and Eichler, E. E. (2002). Recent segmental duplications in the human genome. *Science (New York, N.Y.)*, 297(5583):1003–7.

Bakewell, M. A., Shi, P., and Zhang, J. (2007). More genes underwent positive selection in chimpanzee evolution than in human evolution. *Proceedings of the National Academy of Sciences of the United States of America*, 104(18):7489–94.

Barrier, M., Bustamante, C. D., Yu, J., and Purugganan, M. D. (2003). Selection on rapidly evolving proteins in the Arabidopsis genome. *Genetics*, 163(2):723–33.

Barty, A., Caleman, C., Aquila, A., Timneanu, N., Lomb, L., White, T. A., Andreasson, J., Arnlund, D., Bajt, S., Barends, T. R. M., Barthelmess, M., Bogan, M. J., Bostedt, C., Bozek, J. D., Coffee, R., Coppola, N., Davidsson, J., DePonte, D. P., Doak, R. B., Ekeberg, T., Elser, V., Epp, S. W., Erk, B., Fleckenstein, H., Foucar, L., Fromme, P., Graafsma, H., Gumprecht, L., Hajdu, J., Hampton, C. Y., Hartmann, R., Hartmann, A., Hauser, G., Hirsemann, H., Holl, P., Hunter, M. S., Johansson, L., Kassemeyer, S., Kimmel, N., Kirian, R. A., Liang, M., Maia, F. R. N. C., Malmerberg, E., Marchesini, S., Martin, A. V., Nass, K., Neutze, R., Reich, C., Rolles, D., Rudek, B., Rudenko, A., Scott, H., Schlichting, I., Schulz, J., Seibert, M. M., Shoeman, R. L., Sierra, R. G., Soltau, H., Spence, J. C. H., Stellato, F., Stern, S., Strüder, L., Ullrich, J., Wang, X., Weidenspointner, G., Weierstall, U., Wunderer, C. B., and Chapman, H. N. (2011). Self-terminating diffraction gates femtosecond X-ray nanocrystallography measurements. *Nature Photonics*, 6(1):35–40.

Bedford, T., Wapinski, I., and Hartl, D. L. (2008). Overdispersion of the molecular clock varies between yeast, Drosophila and mammals. *Genetics*, 179(2):977–84.

Begun, D. J., Holloway, A. K., Stevens, K., Hillier, L. W., Poh, Y.-P., Hahn, M. W., Nista, P. M., Jones, C. D., Kern, A. D., Dewey, C. N., Pachter, L., Myers, E., and Langley, C. H. (2007a). Population genomics: whole-genome analysis of polymorphism and divergence in Drosophila simulans. *PLoS biology*, 5(11):e310.

Begun, D. J., Lindfors, H. A., Kern, A. D., and Jones, C. D. (2007b). Evidence for de novo evolution of testis-expressed genes in the Drosophila yakuba/Drosophila erecta clade. *Genetics*, 176(2):1131–7.

Birney, E., Stamatoyannopoulos, J. A., Dutta, A., Guigó, R., Gingeras, T. R., Margulies, E. H., Weng, Z., Snyder, M., Dermitzakis, E. T., Thurman, R. E., and et al. (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, 447(7146):799–816.

Björklund, A. K., Ekman, D., Light, S., Frey-Skött, J., and Elofsson, A. (2005). Domain rearrangements in protein evolution. *Journal of molecular biology*, 353(4):911–23.

Bloom, J. D., Drummond, D. A., Arnold, F. H., and Wilke, C. O. (2006a). Structural determinants of the rate of protein evolution in yeast. *Molecular biology and evolution*, 23(9):1751–61.

Bloom, J. D., Labthavikul, S. T., Otey, C. R., and Arnold, F. H. (2006b). Protein stability promotes evolvability. *Proceedings of the National Academy of Sciences of the United States of America*, 103(15):5869–74.

Bloom, J. D., Raval, A., and Wilke, C. O. (2007). Thermodynamics of neutral protein evolution. *Genetics*, 175(1):255–66.

Bloom, J. D., Silberg, J. J., Wilke, C. O., Drummond, D. A., Adami, C., and Arnold, F. H. (2005). Thermodynamic prediction of protein neutrality. *Proceedings of the National Academy of Sciences of the United States of America*, 102(3):606–11.

Bornberg-Bauer, E., Huylmans, A.-K., and Sikosek, T. (2010). How do new proteins arise? *Current opinion in structural biology*, 20(3):390–6.

Boucher, Y., Douady, C. J., Papke, R. T., Walsh, D. A., Boudreau, M. E. R., Nesbø, C. L., Case, R. J., and Doolittle, W. F. (2003). Lateral gene transfer and the origins of prokaryotic groups. *Annual review of genetics*, 37:283–328.

Bowler, P. J. (1989). *Evolution: The history of an idea*. University edition.

Bridgham, J. T., Carroll, S. M., and Thornton, J. W. (2006). Evolution of hormone-receptor complexity by molecular exploitation. *Science (New York, N.Y.)*, 312(5770):97–101.

Britten, R. J. (1986). Rates of DNA sequence evolution differ between taxonomic groups. *Science (New York, N.Y.)*, 231(4744):1393–8.

Buckling, A., Craig Maclean, R., Brockhurst, M. A., and Colegrave, N. (2009). The Beagle in a bottle. *Nature*, 457(7231):824–9.

Buljan, M. and Bateman, A. (2009). The evolution of protein domain families. *Biochemical Society transactions*, 37(Pt 4):751–5.

Buljan, M., Frankish, A., and Bateman, A. (2010). Quantifying the mechanisms of domain gain in animal proteins. *Genome biology*, 11(7):R74.

Burgess, R. and Yang, Z. (2008). Estimation of hominoid ancestral population sizes under bayesian coalescent models incorporating mutation rate variation and sequencing errors. *Molecular biology and evolution*, 25(9):1979–94.

Bustamante, C. D., Fledel-Alon, A., Williamson, S., Nielsen, R., Todd Hubisz, M., Glanowski, S., Tanenbaum, D. M., White, T. J., Sninsky, J. J., Hernandez, R. D., Civello, D., Adams, M. D., Cargill, M., and Clark, A. G. (2005). Natural selection on protein-coding genes in the human genome. *Nature*, 437(7062):1153–1157.

Bustamante, C. D., Nielsen, R., Sawyer, S. A., Olsen, K. M., Purugganan, M. D., and Hartl, D. L. (2002). The cost of inbreeding in Arabidopsis. *Nature*, 416(6880):531–4.

Bustamante, C. D., Townsend, J. P., and Hartl, D. L. (2000). Solvent accessibility and purifying selection within proteins of Escherichia coli and Salmonella enterica. *Molecular biology and evolution*, 17(2):301–8.

Cai, J., Zhao, R., Jiang, H., and Wang, W. (2008). De novo origination of a new protein-coding gene in Saccharomyces cerevisiae. *Genetics*, 179(1):487–96.

Cai, J. J. and Petrov, D. A. (2010). Relaxed purifying selection and possibly high rate of adaptation in primate lineage-specific genes. *Genome biology and evolution*, 2:393–409.

Cai, J. J., Woo, P. C. Y., Lau, S. K. P., Smith, D. K., and Yuen, K.-Y. (2006). Accelerated evolutionary rate may be responsible for the emergence of lineage-specific genes in ascomycota. *Journal of molecular evolution*, 63(1):1–11.

Capra, J. A., Pollard, K. S., and Singh, M. (2010). Novel genes exhibit distinct patterns of function acquisition and network integration. *Genome biology*, 11(12):R127.

Chamary, J. V., Parmley, J. L., and Hurst, L. D. (2006). Hearing silence: non-neutral evolution at synonymous sites in mammals. *Nature reviews. Genetics*, 7(2):98–108.

Charlesworth, J. and Eyre-Walker, A. (2006). The rate of adaptive evolution in enteric bacteria. *Molecular biology and evolution*, 23(7):1348–56.

Charlesworth, J. and Eyre-Walker, A. (2007). The other side of the nearly neutral theory, evidence of slightly advantageous back-mutations. *Proceedings of the National Academy of Sciences of the United States of America*, 104(43):16992–7.

Chen, S., Zhang, Y. E., and Long, M. (2010). New Genes in Drosophila Quickly Become Essential. *Science*, 330(6011):1682–1685.

Chen, S. C.-C., Chuang, T.-J., and Li, W.-H. (2011). The relationships among microRNA regulation, intrinsically disordered regions, and other indicators of protein evolutionary rate. *Molecular biology and evolution*, 28(9):2513–20.

Chimpanzee Sequencing and Analysis Consortium (2005). Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature*, 437(7055):69–87.

Choi, I.-G. and Kim, S.-H. (2006). Evolution of protein structural classes and protein sequence families. *Proceedings of the National Academy of Sciences of the United States of America*, 103(38):14056–61.

Choi, S. C., Hobolth, A., Robinson, D. M., Kishino, H., and Thorne, J. L. (2007). Quantifying the impact of protein tertiary structure on molecular evolution. *Molecular biology and evolution*, 24(8):1769–82.

Clamp, M., Fry, B., Kamal, M., Xie, X., Cuff, J., Lin, M. F., Kellis, M., Lindblad-Toh, K., and Lander, E. S. (2007). Distinguishing protein-coding and noncoding genes in the human genome. *Proceedings of the National Academy of Sciences of the United States of America*, 104(49):19428–33.

Clark, A. G., Eisen, M. B., Smith, D. R., Bergman, C. M., Oliver, B., Markow, T. A., Kaufman, T. C., Kellis, M., Gelbart, W., Iyer, V. N., and et al. (2007). Evolution of genes and genomes on the Drosophila phylogeny. *Nature*, 450(7167):203–18.

Clark, A. G., Glanowski, S., Nielsen, R., Thomas, P. D., Kejariwal, A., Todd, M. A., Tanenbaum, D. M., Civello, D., Lu, F., Murphy, B., Ferriera, S., Wang, G., Zheng, X., White, T. J., Sninsky, J. J., Adams, M. D., and Cargill, M. (2003). Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. *Science (New York, N.Y.)*, 302(5652):1960–3.

Conant, G. C. and Stadler, P. F. (2009). Solvent exposure imparts similar selective pressures across a range of yeast proteins. *Molecular biology and evolution*, 26(5):1155–61.

Corpet, F., Servant, F., Gouzy, J., and Kahn, D. (2000). ProDom and ProDom-CG: tools for protein domain analysis and whole genome comparisons. *Nucleic acids research*, 28(1):267–9.

Corvelo, A. and Eyras, E. (2008). Exon creation and establishment in human genes. *Genome biology*, 9(9):R141.

Darwin, C. (1859). *On the Origin of Species*. London, murray edition.

Daubin, V. and Ochman, H. (2004). Bacterial genomes as new gene homes: the genealogy of ORFans in E. coli. *Genome research*, 14(6):1036–42.

Do, C. B. and Katoh, K. (2008). Protein multiple sequence alignment. *Methods in molecular biology (Clifton, N.J.)*, 484:379–413.

Domazet-Loso, T. and Tautz, D. (2003). An evolutionary analysis of orphan genes in Drosophila. *Genome research*, 13(10):2213–9.

Doniger, S. W., Kim, H. S., Swain, D., Corcuera, D., Williams, M., Yang, S.-P., and Fay, J. C. (2008). A catalog of neutral and deleterious polymorphism in yeast. *PLoS genetics*, 4(8):e1000183.

Donoghue, M. T., Keshavaiah, C., Swamidatta, S. H., and Spillane, C. (2011). Evolutionary origins of Brassicaceae specific genes in Arabidopsis thaliana. *BMC evolutionary biology*, 11:47.

Doolittle, R. F., Feng, D. F., Johnson, M. S., and McClure, M. A. (1986). Relationships of human protein sequences to those of other organisms. *Cold Spring Harbor symposia on quantitative biology*, 51 Pt 1:447–55.

Draghi, J. A., Parsons, T. L., and Plotkin, J. B. (2011). Epistasis Increases the Rate of Conditionally Neutral Substitution in an Adapting Population. *Genetics*.

Drummond, D. A., Bloom, J. D., Adami, C., Wilke, C. O., and Arnold, F. H. (2005). Why highly expressed proteins evolve slowly. *Proceedings of the National Academy of Sciences of the United States of America*, 102(40):14338–43.

Drummond, D. A., Raval, A., and Wilke, C. O. (2006). A single determinant dominates the rate of yeast protein evolution. *Molecular biology and evolution*, 23(2):327–37.

Drummond, D. A. and Wilke, C. O. (2008). Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell*, 134(2):341–52.

Durbin, R. M., Altshuler, D. L., Abecasis, G. R., Bentley, D. R., Chakravarti, A., Clark, A. G., Collins, F. S., De La Vega, F. M., Donnelly, P., Egholm, M., and et al. (2010). A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–1073.

Duret, L. and Galtier, N. (2009). Biased gene conversion and the evolution of mammalian genomic landscapes. *Annual review of genomics and human genetics*, 10:285–311.

Eddy, S. R. (1998). Profile hidden Markov models. *Bioinformatics (Oxford, England)*, 14(9):755–63.

Eichler, E. E. (2001). Segmental duplications: what's missing, misassigned, and misassembled–and should we care? *Genome research*, 11(5):653–6.

Ekman, D., Björklund, A. K., and Elofsson, A. (2007). Quantification of the elevated rate of domain rearrangements in metazoa. *Journal of molecular biology*, 372(5):1337–48.

Ekman, D., Björklund, A. K., Frey-Skött, J., and Elofsson, A. (2005). Multi-domain proteins in the three kingdoms of life: orphan domains and other unassigned regions. *Journal of molecular biology*, 348(1):231–43.

Elena, S. F., Ekunwe, L., Hajela, N., Oden, S. A., and Lenski, R. E. (1998). Distribution of fitness effects caused by random insertion mutations in Escherichia coli. *Genetica*, 102-103(1-6):349–58.

Elhaik, E., Sabath, N., and Graur, D. (2006). The "inverse relationship between evolutionary rate and age of mammalian genes" is an artifact of increased genetic distance with rate of evolution and time of divergence. *Molecular biology and evolution*, 23(1):1–3.

Ellegren, H. (2008). Comparative genomics and the study of evolution by natural selection. *Molecular ecology*, 17(21):4586–96.

England, J. L. and Shakhnovich, E. I. (2003). Structural determinant of protein designability. *Physical review letters*, 90(21):218101.

Evans, P. D., Anderson, J. R., Vallender, E. J., Choi, S. S., and Lahn, B. T. (2004). Reconstructing the evolutionary history of microcephalin, a gene controlling human brain size. *Human molecular genetics*, 13(11):1139–45.

Eyre-Walker, A. (2002). Changing effective population size and the McDonald-Kreitman test. *Genetics*, 162(4):2017–24.

Eyre-Walker, A. (2006). The genomic rate of adaptive evolution. *Trends in ecology & evolution (Personal edition)*, 21(10):569–75.

Eyre-Walker, A. and Keightley, P. D. (2007). The distribution of fitness effects of new mutations. *Nature reviews. Genetics*, 8(8):610–8.

Eyre-Walker, A. and Keightley, P. D. (2009). Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change. *Molecular biology and evolution*, 26(9):2097–108.

Eyre-Walker, A., Keightley, P. D., Smith, N. G. C., and Gaffney, D. (2002). Quantifying the slightly deleterious mutation model of molecular evolution. *Molecular biology and evolution*, 19(12):2142–9.

Farré, D. and Albà, M. M. (2010). Heterogeneous patterns of gene-expression diversification in mammalian gene duplicates. *Molecular biology and evolution*, 27(2):325–35.

Fay, J. C. (2011). Weighing the evidence for adaptation at the molecular level. *Trends in genetics : TIG*, 27(9):343–9.

Fay, J. C., Wyckoff, G. J., and Wu, C. I. (2001). Positive and negative selection on the human genome. *Genetics*, 158(3):1227–34.

Fay, J. C., Wyckoff, G. J., and Wu, C.-I. (2002). Testing the neutral theory of molecular evolution with genomic data from Drosophila. *Nature*, 415(6875):1024–6.

Ferrada, E. and Wagner, A. (2008). Protein robustness promotes evolutionary innovations on large evolutionary time-scales. *Proceedings. Biological sciences / The Royal Society*, 275(1643):1595–602.

Finn, R. D., Tate, J., Mistry, J., Coggill, P. C., Sammut, S. J., Hotz, H.-R., Ceric, G., Forslund, K., Eddy, S. R., Sonnhammer, E. L. L., and Bateman, A. (2008). The Pfam protein families database. *Nucleic acids research*, 36(Database issue):D281–8.

Fischer, D. and Eisenberg, D. (1999). Finding families for genomic ORFans. *Bioinformatics (Oxford, England)*, 15(9):759–62.

Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J. F., Dougherty, B. A., and Merrick, J. M. (1995). Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science (New York, N.Y.)*, 269(5223):496–512.

Fletcher, W. and Yang, Z. (2010). The Effect of Insertions, Deletions and Alignment Errors on the Branch-Site Test of Positive Selection. *Molecular biology and evolution*, 27(10):2257–67.

Flicek, P., Amode, M. R., Barrell, D., Beal, K., Brent, S., Chen, Y., Clapham, P., Coates, G., Fairley, S., Fitzgerald, S., and et al. (2011). Ensembl 2011. *Nucleic acids research*, 39(Database issue):D800–6.

Fong, J. H., Geer, L. Y., Panchenko, A. R., and Bryant, S. H. (2007). Modeling the evolution of protein domain architectures using maximum parsimony. *Journal of molecular biology*, 366(1):307–15.

Foxe, J. P., Dar, V.-u.-N., Zheng, H., Nordborg, M., Gaut, B. S., and Wright, S. I. (2008). Selection on amino acid substitutions in Arabidopsis. *Molecular biology and evolution*, 25(7):1375–83.

Franzosa, E. and Xia, Y. (2008). Structural perspectives on protein evolution.

Franzosa, E. A. and Xia, Y. (2009). Structural determinants of protein evolution are context-sensitive at the residue level. *Molecular biology and evolution*, 26(10):2387–95.

Fraser, H. B., Hirsh, A. E., Steinmetz, L. M., Scharfe, C., and Feldman, M. W. (2002). Evolutionary rate in the protein interaction network. *Science (New York, N.Y.)*, 296(5568):750–2.

Galtier, N. and Duret, L. (2007). Adaptation or biased gene conversion? Extending the null hypothesis of molecular evolution. *Trends in genetics : TIG*, 23(6):273–7.

Galtier, N., Duret, L., Glémin, S., and Ranwez, V. (2009). GC-biased gene conversion promotes the fixation of deleterious amino acid changes in primates. *Trends in genetics : TIG*, 25(1):1–5.

Genome 10K Community of Scientists (2009). Genome 10K: a proposal to obtain whole-genome sequence for 10,000 vertebrate species. *The Journal of heredity*, 100(6):659–74.

Gibbs, R. A., Rogers, J., Katze, M. G., Bumgarner, R., Weinstock, G. M., Mardis, E. R., Remington, K. A., Strausberg, R. L., Venter, J. C., and Wilson, R. K. e. a. (2007). Evolutionary and Biomedical Insights from the Rhesus Macaque Genome. *Science*, 316(5822):222–234.

Gilad, Y., Man, O., and Glusman, G. (2005). A comparison of the human and chimpanzee olfactory receptor gene repertoires. *Genome research*, 15(2):224–30.

Gillespie, J. (1991). *The Causes of Molecular Evolution*. Oxford University Press, Oxford.

Gillespie, J. H. (1986). Natural selection and the molecular clock. *Molecular biology and evolution*, 3(2):138–55.

Gillespie, J. H. (1994). Substitution processes in molecular evolution. III. Deleterious alleles. *Genetics*, 138(3):943–52.

Goldman, N., Thorne, J. L., and Jones, D. T. (1998). Assessing the impact of secondary structure and solvent accessibility on protein evolution. *Genetics*, 149(1):445–58.

Green, P., Lipman, D., Hillier, L., Waterston, R., States, D., and Claverie, J. M. (1993). Ancient conserved regions in new gene sequences and the protein databases. *Science (New York, N.Y.)*, 259(5102):1711–6.

Gu, Z., Nicolae, D., Lu, H. H.-S., and Li, W. H. (2002). Rapid divergence in expression between duplicate genes inferred from microarray data. *Trends in genetics : TIG*, 18(12):609–13.

Guo, W.-J., Li, P., Ling, J., and Ye, S.-P. (2007). Significant comparative characteristics between orphan and nonorphan genes in the rice (Oryza sativa L.) genome. *Comparative and functional genomics*, page 21676.

Hahn, M. W. (2008). Toward a selection theory of molecular evolution. *Evolution; international journal of organic evolution*, 62(2):255–65.

Hahn, M. W. and Kern, A. D. (2005). Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. *Molecular biology and evolution*, 22(4):803–6.

Haldane, J. (1932). *The causes of evolution.* . Longmans and Green, London.

Halligan, D. L., Oliver, F., Eyre-Walker, A., Harr, B., and Keightley, P. D. (2010). Evidence for pervasive adaptive protein evolution in wild mice. *PLoS genetics*, 6(1):e1000825.

Hayden, E. J., Ferrada, E., and Wagner, A. (2011). Cryptic genetic variation promotes rapid evolutionary adaptation in an RNA enzyme. *Nature*, 474(7349):92–5.

Heinen, T. J. A. J., Staubach, F., Häming, D., and Tautz, D. (2009). Emergence of a new gene from an intergenic region. *Current biology : CB*, 19(18):1527–31.

Hirsh, A. E. and Fraser, H. B. (2001). Protein dispensability and rate of evolution. *Nature*, 411(6841):1046–9.

Hoekstra, H. E., Hirschmann, R. J., Bundey, R. A., Insel, P. A., and Crossland, J. P. (2006). A single amino acid mutation contributes to adaptive beach mouse color pattern. *Science (New York, N.Y.)*, 313(5783):101–4.

Holm, L. and Sander, C. (1994). The FSSP database of structurally aligned protein fold families. *Nucleic acids research*, 22(17):3600–9.

Hotopp, J. C. D., Clark, M. E., Oliveira, D. C. S. G., Foster, J. M., Fischer, P., Torres, M. C. M. n., Giebel, J. D., Kumar, N., Ishmael, N., Wang, S., Ingram, J., Nene, R. V., Shepard, J., Tomkins, J., Richards, S., Spiro, D. J., Ghedin, E., Slatko, B. E., Tettelin, H., and Werren, J. H. (2007). Widespread lateral gene transfer from intracellular bacteria to multicellular eukaryotes. *Science (New York, N.Y.)*, 317(5845):1753–6.

Hughes, A. L. (1994). The evolution of functionally novel proteins after gene duplication. *Proceedings. Biological sciences / The Royal Society*, 256(1346):119–24.

Hughes, A. L. (2007). Looking for Darwin in all the wrong places: the misguided quest for positive selection at the nucleotide sequence level. *Heredity*, 99(4):364–73.

Hughes, A. L. and Friedman, R. (2008). Codon-based tests of positive selection, branch lengths, and the evolution of mammalian immune system genes. *Immunogenetics*, 60(9):495–506.

Hurst, L. D. (2009). Fundamental concepts in genetics: genetics and the understanding of selection. *Nature reviews. Genetics*, 10(2):83–93.

Ingram, V. M. (1961). Gene evolution and the haemoglobins. *Nature*, 189:704–8.

International Chicken Genome Sequencing Consortium (2004). Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature*, 432(7018):695–716.

Itoh, M., Nacher, J. C., Kuma, K.-i., Goto, S., and Kanehisa, M. (2007). Evolutionary history and functional implications of protein domains and their combinations in eukaryotes. *Genome biology*, 8(6):R121.

Janecek, L. L., Honeycutt, R. L., Adkins, R. M., and Davis, S. K. (1996). Mitochondrial gene sequences and the molecular systematics of the artiodactyl subfamily bovinae. *Molecular phylogenetics and evolution*, 6(1):107–19.

Kaessmann, H. (2010). Origins, evolution and phenotypic impact of new genes. *Genome Research*, 20(10):1313–26.

Kaessmann, H., Vinckenbosch, N., and Long, M. (2009). RNA-based gene duplication: mechanistic and evolutionary insights. *Nature reviews. Genetics*, 10(1):19–31.

Kasahara, M. (2007). The 2R hypothesis: an update. *Current opinion in immunology*, 19(5):547–52.

Kasuga, T., Mannhaupt, G., and Glass, N. L. (2009). Relationship between phylogenetic distribution and genomic features in Neurospora crassa. *PloS one*, 4(4):e5286.

Katju, V. and Lynch, M. (2003). The structure and early evolution of recently arisen gene duplicates in the Caenorhabditis elegans genome. *Genetics*, 165(4):1793–803.

Katju, V. and Lynch, M. (2006). On the formation of novel genes by duplication in the Caenorhabditis elegans genome. *Molecular biology and evolution*, 23(5):1056–67.

Katoh, K., Misawa, K., Kuma, K.-i., and Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic acids research*, 30(14):3059–66.

Kemena, C. and Notredame, C. (2009). Upcoming challenges for multiple sequence alignment methods in the high-throughput era. *Bioinformatics (Oxford, England)*, 25(19):2455–65.

Khalturin, K., Hemmrich, G., Fraune, S., Augustin, R., and Bosch, T. C. G. (2009). More than just orphans: are taxonomically-restricted genes important in evolution? *Trends in genetics : TIG*, 25(9):404–13.

Kim, S.-H. and Yi, S. V. (2008). Mammalian nonsynonymous sites are not overdispersed: comparative genomic analysis of index of dispersion of mammalian proteins. *Molecular biology and evolution*, 25(4):634–42.

Kimura, M. (1968). Evolutionary rate at the molecular level. *Nature*, 217(5129):624–6.

Kimura, M. (1983). *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge.

Kimura, M. and Ota, T. (1974). On some principles governing molecular evolution. *Proceedings of the National Academy of Sciences of the United States of America*, 71(7):2848–52.

Kinch, L. N. and Grishin, N. V. (2002). Evolution of protein structures and functions. *Current opinion in structural biology*, 12(3):400–8.

King, J. L. and Jukes, T. H. (1969). Non-Darwinian evolution. *Science (New York, N.Y.)*, 164(881):788–98.

Knowles, D. G. and McLysaght, A. (2009). Recent de novo origin of human protein-coding genes. *Genome research*, 19(10):1752–9.

Kohne, D. E. (1970). Evolution of higher-organism DNA. *Quarterly reviews of biophysics*, 3(3):327–75.

Kondrashov, F. A., Rogozin, I. B., Wolf, Y. I., and Koonin, E. V. (2002). Selection in the evolution of gene duplications. *Genome biology*, 3(2):RESEARCH0008.

Koonin, E. V. (2009). Darwinian evolution in the light of genomics. *Nucleic acids research*, 37(4):1011–34.

Kosiol, C., Vinar, T., da Fonseca, R. R., Hubisz, M. J., Bustamante, C. D., Nielsen, R., and Siepel, A. (2008). Patterns of positive selection in six Mammalian genomes. *PLoS genetics*, 4(8):e1000144.

Kouprina, N., Pavlicek, A., Mochida, G. H., Solomon, G., Gersch, W., Yoon, Y.-H., Collura, R., Ruvolo, M., Barrett, J. C., Woods, C. G., Walsh, C. A., Jurka, J., and Larionov, V. (2004). Accelerated evolution of the ASPM gene controlling brain size begins prior to human brain expansion. *PLoS biology*, 2(5):E126.

Kryazhimskiy, S., Dushoff, J., Bazykin, G. A., and Plotkin, J. B. (2011). Prevalence of Epistasis in the Evolution of Influenza A Surface Proteins. *PLoS Genetics*, 7(2):e1001301.

Kumar, S. and Subramanian, S. (2002). Mutation rates in mammalian genomes. *Proceedings of the National Academy of Sciences of the United States of America*, 99(2):803–8.

Kuo, C.-H. and Kissinger, J. C. (2008). Consistent and contrasting properties of lineage-specific genes in the apicomplexan parasites Plasmodium and Theileria. *BMC evolutionary biology*, 8:108.

Laird, C. D., McConaughy, B. L., and McCarthy, B. J. (1969). Rate of fixation of nucleotide substitutions in evolution. *Nature*, 224(5215):149–54.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., and et al. (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921.

Larracuente, A. M., Sackton, T. B., Greenberg, A. J., Wong, A., Singh, N. D., Sturgill, D., Zhang, Y., Oliver, B., and Clark, A. G. (2008). Evolution of protein-coding genes in Drosophila. *Trends in genetics : TIG*, 24(3):114–23.

Levine, M. T., Jones, C. D., Kern, A. D., Lindfors, H. A., and Begun, D. J. (2006). Novel genes derived from noncoding DNA in Drosophila melanogaster are frequently X-linked and exhibit testis-biased expression. *Proceedings of the National Academy of Sciences of the United States of America*, 103(26):9935–9.

Li, C.-Y., Zhang, Y., Wang, Z., Zhang, Y., Cao, C., Zhang, P.-W., Lu, S.-J., Li, X.-M., Yu, Q., Zheng, X., Du, Q., Uhl, G. R., Liu, Q.-R., and Wei, L. (2010). A human-specific de novo protein-coding gene associated with human brain functions. *PLoS computational biology*, 6(3):e1000734.

Li, W.-H. (1997). *Molecular Evolution*. Sinauer as edition.

Lin, Y.-S., Hsu, W.-L., Hwang, J.-K., and Li, W.-H. (2007). Proportion of solvent-exposed amino acids in a protein and rate of protein evolution. *Molecular biology and evolution*, 24(4):1005–11.

Lipman, D. J., Souvorov, A., Koonin, E. V., Panchenko, A. R., and Tatusova, T. A. (2002). The relationship of protein conservation and sequence length. *BMC evolutionary biology*, 2:20.

Liti, G., Carter, D. M., Moses, A. M., Warringer, J., Parts, L., James, S. A., Davey, R. P., Roberts, I. N., Burt, A., Koufopanou, V., Tsai, I. J., Bergman, C. M., Bensasson, D., O'Kelly, M. J. T., van Oudenaarden, A., Barton, D. B. H., Bailes, E., Nguyen, A. N., Jones, M., Quail, M. A., Goodhead, I., Sims, S., Smith, F., Blomberg, A., Durbin, R., and Louis, E. J. (2009). Population genomics of domestic and wild yeasts. *Nature*, 458(7236):337–41.

Lobkovsky, A. E., Wolf, Y. I., and Koonin, E. V. (2010). Universal distribution of protein evolution rates as a consequence of protein folding physics. *Proceedings of the National Academy of Sciences of the United States of America*, 107(7):2983–8.

Long, M., Betrán, E., Thornton, K., and Wang, W. (2003). The origin of new genes: glimpses from the young and old. *Nature reviews. Genetics*, 4(11):865–75.

Löytynoja, A. and Goldman, N. (2008). Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. *Science (New York, N.Y.)*, 320(5883):1632–5.

Luz, H., Staub, E., and Vingron, M. (2006). About the interrelation of evolutionary rate and protein age. *Genome informatics. International Conference on Genome Informatics*, 17(1):240–50.

Lynch, M. (2007). *The origins of Genome Architecture*. Sunderland, MA.

Lynch, M. and Conery, J. S. (2000). The evolutionary fate and consequences of duplicate genes. *Science (New York, N.Y.)*, 290(5494):1151–5.

Makova, K. D. and Li, W.-H. (2003). Divergence in the spatial pattern of gene expression between human duplicate genes. *Genome research*, 13(7):1638–45.

Mallick, S., Gnerre, S., Muller, P., and Reich, D. (2009). The difficulty of avoiding false positives in genome scans for natural selection. *Genome research*, 19(5):922–33.

Marais, G. and Duret, L. (2001). Synonymous codon usage, accuracy of translation, and gene length in Caenorhabditis elegans. *Journal of molecular evolution*, 52(3):275–80.

Markova-Raina, P. and Petrov, D. (2011). High sensitivity to aligner and high rate of false positives in the estimates of positive selection in the 12 Drosophila genomes. *Genome Research*, pages gr.115949.110–.

Marques, A. C., Dupanloup, I., Vinckenbosch, N., Reymond, A., and Kaessmann, H. (2005). Emergence of young human genes after a burst of retroposition in primates. *PLoS biology*, 3(11):e357.

Marsh, J. A. and Teichmann, S. A. (2010). How do proteins gain new domains? *Genome biology*, 11(7):126.

Martin, A. P. and Palumbi, S. R. (1993). Body size, metabolic rate, generation time, and the molecular clock. *Proceedings of the National Academy of Sciences of the United States of America*, 90(9):4087–91.

Mayr, E. (2002). *What Evolution is*. London, phoenix edition.

McDonald, J. H. and Kreitman, M. (1991). Adaptive protein evolution at the Adh locus in Drosophila. *Nature*, 351(6328):652–4.

Mendel, G. (1901). Experiments in plant hybridization. *Journal of the Royal Horticultural Society*, 26:1–32.

Meunier, J. and Duret, L. (2004). Recombination drives the evolution of GC-content in the human genome. *Molecular biology and evolution*, 21(6):984–90.

Mirny, L. A. and Shakhnovich, E. I. (1999). Universally conserved positions in protein folds: reading evolutionary signals about stability, folding kinetics and function. *Journal of molecular biology*, 291(1):177–96.

Moore, A. D., Björklund, A. K., Ekman, D., Bornberg-Bauer, E., and Elofsson, A. (2008). Arrangements in the modular evolution of proteins. *Trends in biochemical sciences*, 33(9):444–51.

Müller, A., MacCallum, R. M., and Sternberg, M. J. E. (2002). Structural characterization of the human proteome. *Genome research*, 12(11):1625–41.

Muller, H. (1935). The origination of chromatin deficiencies as minute deletions subject to insertion elsewhere. *Genetica*, 17:237–252.

Murzin, A. G., Brenner, S. E., Hubbard, T., and Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *Journal of molecular biology*, 247(4):536–40.

Nathans, J., Thomas, D., and Hogness, D. S. (1986). Molecular genetics of human color vision: the genes encoding blue, green, and red pigments. *Science (New York, N.Y.)*, 232(4747):193–202.

Nekrutenko, A. and Li, W. H. (2001). Transposable elements are found in a large number of human protein-coding genes. *Trends in genetics : TIG*, 17(11):619–21.

Nielsen, R., Bustamante, C., Clark, A. G., Glanowski, S., Sackton, T. B., Hubisz, M. J., Fledel-Alon, A., Tanenbaum, D. M., Civello, D., White, T. J., J Sninsky, J., Adams, M. D., and Cargill, M. (2005). A scan for positively selected genes in the genomes of humans and chimpanzees. *PLoS biology*, 3(6):e170.

Nielsen, R. and Yang, Z. (2003). Estimating the distribution of selection coefficients from phylogenetic data with applications to mitochondrial and viral DNA. *Molecular biology and evolution*, 20(8):1231–9.

Notredame, C., Higgins, D. G., and Heringa, J. (2000). T-Coffee: A novel method for fast and accurate multiple sequence alignment. *Journal of molecular biology*, 302(1):205–17.

Nozawa, M., Suzuki, Y., and Nei, M. (2009). Reliabilities of identifying positive selection by the branch-site and the site-prediction methods. *Proceedings of the National Academy of Sciences of the United States of America*, 106(16):6700–5.

Nuttall, G. H. and Inchley, O. (1904). An improved Method of measuring the amount of Precipitum in connection with Tests with Precipitating Antisera. *The Journal of hygiene*, 4(2):201–6.

Ohno, S. (1970). *Evolution by gene duplication*. Springer-Verlag, New York.

Ohta, T. (1973). Slightly deleterious mutant substitutions in evolution. *Nature*, 246(5428):96–8.

Ohta, T. (1995). Synonymous and nonsynonymous substitutions in mammalian genes and the nearly neutral theory. *Journal of molecular evolution*, 40(1):56–63.

Ohta, T. and Ina, Y. (1995). Variation in synonymous substitution rates among mammalian genes and the correlation between synonymous and nonsynonymous divergences. *Journal of molecular evolution*, 41(6):717–20.

Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B., and Thornton, J. M. (1997). CATH–a hierarchic classification of protein domain structures. *Structure (London, England : 1993)*, 5(8):1093–108.

Ota, T. and Kimura, M. (1971). On the constancy of the evolutionary rate of cistrons. *Journal of molecular evolution*, 1(1):18–25.

Page, R. D. M. and Holmes, E. C. (1998). *Molecular Evolution. A Phylogenetic Approach*. Oxford, blackwell edition.

Pál, C., Papp, B., and Hurst, L. D. (2001). Highly expressed genes in yeast evolve slowly. *Genetics*, 158(2):927–31.

Pál, C., Papp, B., and Lercher, M. J. (2006). An integrated view of protein evolution. *Nature reviews. Genetics*, 7(5):337–48.

Pal, L. R. and Guda, C. (2006). Tracing the origin of functional and conserved domains in the human proteome: implications for protein evolution at the modular level. *BMC evolutionary biology*, 6:91.

Paterson, S., Vogwill, T., Buckling, A., Benmayor, R., Spiers, A. J., Thomson, N. R., Quail, M., Smith, F., Walker, D., Libberton, B., Fenton, A., Hall, N., and Brockhurst, M. A. (2010). Antagonistic coevolution accelerates molecular evolution. *Nature*, 464(7286):275–8.

Patthy, L. (1999). Genome evolution and the evolution of exon-shuffling–a review. *Gene*, 238(1):103–14.

Patthy, L. (2003). Modular assembly of genes and the evolution of new functions. *Genetica*, 118(2-3):217–31.

Pei, J. (2008). Multiple protein sequence alignment. *Current opinion in structural biology*, 18(3):382–6.

Piganeau, G. and Eyre-Walker, A. (2003). Estimating the distribution of fitness effects from DNA sequence data: implications for the molecular clock. *Proceedings of the National Academy of Sciences of the United States of America*, 100(18):10335–40.

Pirovano, W. and Heringa, J. (2008). Multiple sequence alignment. *Methods in molecular biology (Clifton, N.J.)*, 452:143–61.

Plotkin, J. B. and Fraser, H. B. (2007). Assessing the determinants of evolutionary rates in the presence of noise. *Molecular biology and evolution*, 24(5):1113–21.

Pond, S. L. K., Frost, S. D. W., and Muse, S. V. (2005). HyPhy: hypothesis testing using phylogenies. *Bioinformatics (Oxford, England)*, 21(5):676–9.

Ponting, C. P. and Russell, R. R. (2002). The natural history of protein domains. *Annual review of biophysics and biomolecular structure*, 31:45–71.

Ratnakumar, A., Mousset, S., Glémin, S., Berglund, J., Galtier, N., Duret, L., and Webster, M. T. (2010). Detecting positive selection within genomes: the problem of biased gene conversion. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 365(1552):2571–80.

Ridley, M. (2004). *Evolution*. Wiley-Blackwell.

Rodríguez-Trelles, F., Tarrío, R., and Ayala, F. J. (2001). Erratic overdispersion of three molecular clocks: GPDH, SOD, and XDH. *Proceedings of the National Academy of Sciences of the United States of America*, 98(20):11405–10.

Sanjuán, R., Moya, A., and Elena, S. F. (2004). The distribution of fitness effects caused by single-nucleotide substitutions in an RNA virus. *Proceedings of the National Academy of Sciences of the United States of America*, 101(22):8396–401.

Scannell, D. R. and Wolfe, K. H. (2008). A burst of protein sequence evolution and a prolonged period of asymmetric evolution follow gene duplication in yeast. *Genome research*, 18(1):137–47.

Schneider, A., Souvorov, A., Sabath, N., Landan, G., Gonnet, G. H., and Graur, D. (2009). Estimates of positive darwinian selection are inflated by errors in sequencing, annotation, and alignment. *Genome biology and evolution*, 2009:114–8.

Schultz, J. (1998). SMART, a simple modular architecture research tool: Identification of signaling domains. *Proceedings of the National Academy of Sciences*, 95(11):5857–5864.

Shakhnovich, B. E. (2006). Relative contributions of structural designability and functional diversity in molecular evolution of duplicates. *Bioinformatics (Oxford, England)*, 22(14):e440–5.

Shapiro, B. J. and Alm, E. J. (2008). Comparing patterns of natural selection across species using selective signatures. *PLoS genetics*, 4(2):e23.

Siew, N. and Fischer, D. (2003). Analysis of singleton ORFans in fully sequenced microbial genomes. *Proteins*, 53(2):241–51.

Silander, O. K., Tenaillon, O., and Chao, L. (2007). Understanding the evolutionary fate of finite populations: the dynamics of mutational effects. *PLoS biology*, 5(4):e94.

Simon, M. and Hancock, J. M. (2009). Tandem and cryptic amino acid repeats accumulate in disordered regions of proteins. *Genome biology*, 10(6):R59.

Smith, N. G. C. and Eyre-Walker, A. (2002). Adaptive protein evolution in Drosophila. *Nature*, 415(6875):1022–4.

Sonnhammer, E. L., Eddy, S. R., and Durbin, R. (1997). Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins*, 28(3):405–20.

Sonnhammer, E. L. and Kahn, D. (1994). Modular arrangement of proteins as inferred from analysis of homology. *Protein science : a publication of the Protein Society*, 3(3):482–92.

Spencer, C. C. A. (2006). Human polymorphism around recombination hotspots. *Biochemical Society transactions*, 34(Pt 4):535–6.

Storey, J. D. and Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences of the United States of America*, 100(16):9440–5.

Subramanian, S. and Kumar, S. (2004). Gene expression intensity shapes evolutionary rates of the proteins encoded by the vertebrate genome. *Genetics*, 168(1):373–81.

Takano-Shimizu, T. (1999). Local recombination and mutation effects on molecular evolution in Drosophila. *Genetics*, 153(3):1285–96.

Tautz, D. and Domazet-Lošo, T. (2011). The evolutionary origin of orphan genes. *Nature Reviews Genetics*, 12(10):692–702.

Thatcher, J. W., Shaw, J. M., and Dickinson, W. J. (1998). Marginal fitness contributions of nonessential genes in yeast. *Proceedings of the National Academy of Sciences of the United States of America*, 95(1):253–7.

Thomson, T. M., Lozano, J. J., Loukili, N., Carrió, R., Serras, F., Cormand, B., Valeri, M., Díaz, V. M., Abril, J., Burset, M., Merino, J., Macaya, A., Corominas, M., and Guigó, R. (2000). Fusion of the human gene for the polyubiquitination coeffector UEV1 with Kua, a newly identified gene. *Genome research*, 10(11):1743–56.

Toft, C. and Fares, M. A. (2010). Structural Calibration of the Rates of Amino Acid Evolution in a Search for Darwin in Drifting Biological Systems. *Molecular biology and evolution*, 27(10):2375–85.

Tokuriki, N. and Tawfik, D. S. (2009). Stability effects of mutations and protein evolvability. *Current opinion in structural biology*, 19(5):596–604.

Toll-Riera, M., Bosch, N., Bellora, N., Castelo, R., Armengol, L., Estivill, X., and Albà, M. M. (2009a). Origin of primate orphan genes: a comparative genomics approach. *Molecular biology and evolution*, 26(3):603–12.

Toll-Riera, M., Castelo, R., Bellora, N., and Albà, M. M. (2009b). Evolution of primate orphan proteins. *Biochemical Society transactions*, 37(Pt 4):778–82.

Toll-Riera, M., Castresana, J., and Albà, M. M. (2008). Accelerated evolution of genes of recent origin. In Pantarotti, editor, *Evolutionary Biology: from Concept to Application*. Springer.

Toll-Riera, M., Laurie, S., and Albà, M. M. (2010). Lineage-specific Variation in Intensity of Natural Selection in Mammals. *Molecular biology and evolution*, 28(1):383–98.

Toll-Riera, M., Laurie, S., Rado-Trilla, N., and Albà, M. (2011a). Partial gene duplication and the formation of novel genes. In Friedberg, F., editor, *Gene Duplication*, chapter 6. Intech.

Toll-Riera, M., Radó-Trilla, N., Martys, F., and Albà, M. M. (2011b). Role of Low-Complexity Sequences in the Formation of Novel Protein Coding Sequences. *Molecular biology and evolution*.

Van de Peer, Y., Maere, S., and Meyer, A. (2009). The evolutionary significance of ancient genome duplications. *Nature reviews. Genetics*, 10(10):725–32.

Vibranovski, M. D., Sakabe, N. J., de Oliveira, R. S., and de Souza, S. J. (2005). Signs of ancient and modern exon-shuffling are correlated to the distribution of ancient and modern domains along proteins. *Journal of molecular evolution*, 61(3):341–50.

Vilella, A. J., Severin, J., Ureta-Vidal, A., Heng, L., Durbin, R., and Birney, E. (2009). EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome research*, 19(2):327–35.

Vishnoi, A., Kryazhimskiy, S., Bazykin, G. a., Hannenhalli, S., and Plotkin, J. B. (2010). Young proteins experience more variable selection pressures than old proteins. *Genome research*, 20(11):1574–81.

Vogel, C., Bashton, M., Kerrison, N. D., Chothia, C., and Teichmann, S. A. (2004a). Structure, function and evolution of multidomain proteins. *Current opinion in structural biology*, 14(2):208–16.

Vogel, C., Berzuini, C., Bashton, M., Gough, J., and Teichmann, S. A. (2004b). Supra-domains: evolutionary units larger than single protein domains. *Journal of molecular biology*, 336(3):809–23.

Vogel, C. and Chothia, C. (2006). Protein family expansions and biological complexity. *PLoS computational biology*, 2(5):e48.

Wagner, A. (2005). Robustness, evolvability, and neutrality. *FEBS letters*, 579(8):1772–8.

Wagner, A. (2008). Neutralism and selectionism: a network-based reconciliation. *Nature reviews. Genetics*, 9(12):965–74.

Wagner, A. (2011). *The origins of evolutionary innovations: a theory of transformative change in living systems*. Oxford University Press.

Wall, D. P., Hirsh, A. E., Fraser, H. B., Kumm, J., Giaever, G., Eisen, M. B., and Feldman, M. W. (2005). Functional genomic analysis of the rates of protein evolution. *Proceedings of the National Academy of Sciences of the United States of America*, 102(15):5483–8.

Wallace, A. R. (1858). On the Tendency of Varieties to Depart Indefinitely From the Original Type.

Wang, W., Zheng, H., Yang, S., Yu, H., Li, J., Jiang, H., Su, J., Yang, L., Zhang, J., McDermott, J., Samudrala, R., Wang, J., Yang, H., Yu, J., Kristiansen, K., Wong, G. K.-S., and Wang, J. (2005). Origin and evolution of new exons in rodents. *Genome research*, 15(9):1258–64.

Waterston, R. H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J. F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., and et al. (2002). Initial sequencing and comparative analysis of the mouse genome. *Nature*, 420(6915):520–62.

Watson, J. D. and Crick, F. H. (1953). The structure of DNA. *Cold Spring Harbor symposia on quantitative biology*, 18:123–31.

Webster, M. T. and Smith, N. G. C. (2004). Fixation biases affecting human SNPs. *Trends in genetics : TIG*, 20(3):122–6.

Wilson, A. C., Carlson, S. S., and White, T. J. (1977). Biochemical evolution. *Annual review of biochemistry*, 46:573–639.

Wilson, B. A. and Masel, J. (2011). Putatively Noncoding Transcripts Show Extensive Association with Ribosomes. *Genome biology and evolution*, 3:1245–52.

Wilson, G. A., Feil, E. J., Lilley, A. K., and Field, D. (2007). Large-scale comparative genomic ranking of taxonomically restricted genes (TRGs) in bacterial and archaeal genomes. *PloS one*, 2(3):e324.

Wingreen, N., Li, H., and Tang, C. (2003). Designability and Thermal Stability of Protein Structures. *Polymer*, 45(2):12.

Wolf, M. Y., Wolf, Y. I., and Koonin, E. V. (2008). Comparable contributions of structural-functional constraints and expression level to the rate of protein sequence evolution. *Biology direct*, 3:40.

Wolf, Y. I., Brenner, S. E., Bash, P. A., and Koonin, E. V. (1999). Distribution of protein folds in the three superkingdoms of life. *Genome research*, 9(1):17–26.

Wolf, Y. I., Novichkov, P. S., Karev, G. P., Koonin, E. V., and Lipman, D. J. (2009). The universal distribution of evolutionary rates of genes and distinct characteristics of eukaryotic genes of different apparent ages. *Proceedings of the National Academy of Sciences of the United States of America*, 106(18):7273–80.

Wong, K. M., Suchard, M. A., and Huelsenbeck, J. P. (2008). Alignment uncertainty and genomic analysis. *Science (New York, N.Y.)*, 319(5862):473–6.

Wong, P. and Frishman, D. (2006). Fold designability, distribution, and disease. *PLoS computational biology*, 2(5):e40.

Wu, D.-D., Irwin, D. M., and Zhang, Y.-P. (2011). De Novo Origin of Human Protein-Coding Genes. *PLoS Genetics*, 7(11):e1002379.

Xia, Y., Franzosa, E. A., and Gerstein, M. B. (2009). Integrated assessment of genomic correlates of protein evolutionary rate. *PLoS computational biology*, 5(6):e1000413.

Xiao, W., Liu, H., Li, Y., Li, X., Xu, C., Long, M., and Wang, S. (2009). A rice gene of de novo origin negatively regulates pathogen-induced defense response. *PloS one*, 4(2):e4603.

Yang, S. and Bourne, P. E. (2009). The evolutionary history of protein domains viewed by species phylogeny. *PloS one*, 4(12):e8378.

Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Molecular biology and evolution*, 24(8):1586–91.

Yang, Z. and Huang, J. (2011). De novo origin of new genes with introns in Plasmodium vivax. *FEBS letters*, 585(4):641–4.

Yue, P., Li, Z., and Moult, J. (2005). Loss of protein structure stability as a major causative factor in monogenic disease. *Journal of molecular biology*, 353(2):459–73.

Zhang, G., Wang, H., Shi, J., Wang, X., Zheng, H., Wong, G. K.-S., Clark, T., Wang, W., Wang, J., and Kang, L. (2007). Identification and characterization of insect-specific proteins by genome data analysis. *BMC genomics*, 8:93.

Zhang, J., Nielsen, R., and Yang, Z. (2005). Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Molecular biology and evolution*, 22(12):2472–9.

Zhang, J., Zhang, Y.-p., and Rosenberg, H. F. (2002). Adaptive evolution of a duplicated pancreatic ribonuclease gene in a leaf-eating monkey. *Nature genetics*, 30(4):411–5.

Zhang, L. and Li, W.-H. (2005). Human SNPs reveal no evidence of frequent positive selection. *Molecular biology and evolution*, 22(12):2504–7.

Zhou, Q., Zhang, G., Zhang, Y., Xu, S., Zhao, R., Zhan, Z., Li, X., Ding, Y., Yang, S., and Wang, W. (2008a). On the origin of new genes in Drosophila. *Genome research*, 18(9):1446–55.

Zhou, T., Drummond, D. A., and Wilke, C. O. (2008b). Contact density affects protein evolutionary rate from bacteria to animals. *Journal of molecular evolution*, 66(4):395–404.

Zuckerkandl, E. (1976). Evolutionary processes and evolutionary noise at the molecular level. I. Functional density in proteins. *Journal of molecular evolution*, 7(3):167–83.

Zuckerkandl, E. and Pauling, L. (1962). Molecular diseases, evolution and genic heterogeneity. In Kasha, M. and Pullman, B., editors, *Horizons in biochemistry*, pages 189–225. New York, academic p edition.

Beagle Channel, November 2009

There is grandeur in this view of life, with its several powers, having been originally breathed into a few forms or into one; and that, whilst this planet has gone cycling on according to the fixed law of gravity, from so simple a beginning endless forms most beautiful and most wonderful have been, and are being, evolved

Charles Darwin
*Origin of Species (1859)*