Evolvability of proteomes: Predicting protein function in the light of evolution

Inaugural-Dissertation
to obtain the academic degree
Doctor rerum naturalium (Dr. rer. nat.)

submitted to the Department of Biology, Chemistry and Pharmacy
of Freie Universität Berlin

by

David Fournier

from Oullins, France

2014

Time period: October 2009 - March 2014
Supervisor: Dr. Miguel Andrade
Institute: Max Delbrück Center für Molekulare Medizin, Berlin Buch

1st Reviewer: Prof. Dr. Udo Heinemann, Freie Universität
2nd Reviewer: Prof. Dr. Erich Wanker

Date of defense: 04/04/2014

Acknowledgements

First of all, I would like to thank Miguel Andrade for supervising my work, this has been a great time all along these years thinking about these evolutionary scenarios. Thank you not only for the scientific direction but also the great management. Thank you to all the CBDMers, with whom I had very good connections and lots of fun talking with. I have a special thought for all my colleagues who participated to the Marathon Staffel and other running events in Berlin all these years. That was really a lot of fun to be part of this. Moreover, I salute all my colleagues and friends who joined the weekly experimental lab procedure of wine tasting.

I would like to thank Alexandre Budria and Yves Clément for helpful comments and suggestions to the thesis. I thank all collaborators who have participated to the articles I was involved in. I thank especially Prof. Dr. Erich Wanker for providing very interesting data and also for stimulating discussions. Thanks to Alexandre, Prof. Dr. Detlev Ganten and Russ Hodge for their great motivation to organize the seminar on evolutionary medicine. Hopefully there will be more of these conferences in the future and evolution will become part of the medical curriculum as a tool to study diseases and body dysfunctions.

I would like finally to thank all my friends and my family, my mother and my sister who were always very supportive. I dedicate this work to the memory of my father.

Contents

List of figures

List of tables

1. Introduction

1.1. The place of evolutionary theory in modern biology

One of the favorite sentences evolutionary biologists like to quote is the famous statement of Dobzhansky who said that "nothing in biology makes sense, except in the light of evolution". Though we surely agree with this thought, we wonder if this has practical impact for daily lab work of modern biologists. On the surface, evolutionary biology has a defined frame, and is a field of science like another. So what does an immunologist, a specialist of cancer or a biotechnologist could have to do with evolution and a famous theory built by a naturalist of the nineteenth century?

1.1.1. First observation: nothing in experimental biology makes sense, except in the light of evolution

To answer this question, one can first remark that relationship between species is one of the assumptions of most of experimental medicine. Experimentation on animals dates back to antiquity [1] and especially to the work of Galen, who was dissecting apes and dogs in order to understand human pathology. Galen was convinced that there is a common organization between human and these organisms. He stated that to understand human anatomy, the study of wounds is not enough and advised to get complementary information from the dissection of monkeys, whose anatomy is close to ours [2]. His textbooks were famous all along Middle Age until the time of Enlightenment. In the nineteenth century, Claude Bernard relied on the same assumption to perform his medical experiments on livers and kidneys of dogs. Ultimately, he created a new field called experimental medicine or physiology [3]. Today, scientists still experiment under the same paradigm, whether they believe in evolution or not. Modern biology has accumulated an incredible amount of evidence showing the relationship between human and other species at the morphological, physiological, histological, cellular and molecular levels. All aspects of modern biology are therefore tangled with evolutionary pre-assumptions. We sincerely believe in respect to this observation that evolution is today a major actor of biology. As a consequence, many scientists, and especially the “evolution-skeptics”, might be like the Molière’s “bourgeois gentilhomme”, who was doing prose without noticing it: they are handling evolutionary concepts all the time, and are eventually not aware to do so.

1.1.2. Second observation: new solutions to biological problems using concepts from evolutionary biology are emerging

Aside from the consequences it implicitly has for biomedical research, evolution is used on a daily basis as a deliberate tool to solve biological questions not necessarily related to evolution in the first place [4]. The main contribution of evolutionary theory to biology in the last thirty years comes from the possibility to annotate sequences by alignment to annotated sequences stored in public databases. The most common methods to identify domains within protein sequences, including Hidden Markov Models, are based on the assumption that similar protein sequences fold into similar structures. Generally speaking, alignments are practical to find signal, i.e. information, about a given sequence, assuming that sites that are very conserved participate more in the function of the protein than others. As a result, one can predict the impact of a mutation on the protein function. For instance, a conserved site, upon mutation, will be more likely to disrupt the protein function [5]. At a more complex level, information from different sites on the same protein sequence can be used to infer information on the protein. Recently, structures of transmembrane proteins were predicted at a high-resolution level under the assumption that co-evolving amino-acids within a protein are more likely to be interacting with each other than the ones who do not co-evolve [6].

Moreover, in the last twenty years, some very stimulating papers have come from Evolutionary Medicine, a discipline at the interface between medicine and evolutionary biology. This research relies on the use of evolutionary concepts in medicine and seeks for the ultimate (evolutionary) processes shaping diseases [7]. Concepts from the evolutionary theory have indeed become a real asset for medicine in the recent years. We review briefly few examples of practical applications.

A classical example of application of evolutionary concepts to medicine is the management of bacterial resistance to antibiotics [8]. Studying the completely sequenced genomes can help to predict the bacterial strains that are more likely to evolve resistance to a considered drug. With the progressive diminution of costs associated, studies of bacterial genomes will certainly overwhelm phenotype profiling in the identification of resistant strains in a close future. Similarly to the virus phylogeny, the high rate of sequencing now allows to trace the evolution of bacterial resistance in different populations or even at the scale of a hospital [9] and brings the expectation that resistant strains will become easier to detect and to trace in the future. Controlled evolution in the laboratory could help to identify the potential of bacterial strains to increase their pathogenicity and the genetic events involved [10]. Information about bacterial evolution could be used to design drugs that prevent bacterial strains to take the most pathogenic evolutionary route.

Phylogeny of pathogenic strains has shown potential to help identifying the risk for viral strains to evolve higher pathogenicity and cause epidemic outbreaks. The underlying idea is that the strains that have evolved more, in other words the ones that have accumulated the greatest number of mutations, are the most likely to evolve new functions. These potentially emerging

pathogens might thus become more likely to evade immune defenses and cause stronger damages to the human body [11]. As a consequence, phylogeny can help researchers developing vaccines before the apparition of virulent strains, thus limiting the spread of pathogens in human populations.

Now we have solid elements to answer the question asked at the beginning of this introduction; in the light of the examples given in this section, one can say that yes, evolution is highly relevant to biomedical research and will be probably increasingly important in a close future. Aware of this fact, in 2011, we organized a conference on the topic at the Max-Delbrück Center of Berlin (http://cbdm.mdc-berlin.de/~theevolutionworkshop/).

1.2. Brief reminder of major concepts of evolutionary biology

So what are exactly evolution and the theory of evolution? Firstly, one remarks that the concept itself appears rather recently in the history of science. In the philosophy of the Ancient Greece, the universe is perfectly designed and fixed and so are all forms of life. Some philosophers nevertheless exposed theories showing their awareness of the existence of a relationship between all living beings. In his famous treaty on natural history, Aristotle described life as a chain of beings. He classified living beings into twelve categories of increased perfection, from inanimate beings to Plants to Invertebrates to Vertebrates to Humans [1]. Anaximander from Miletus thought that life arises from water, and it transforms to simple organisms and then to the most complex forms [12].

As the influence of the Church was huge during centuries, the perspective of man on nature did not change very much until the eighteenth century. Newton thought that there was uniformity in all living beings, but attributed that to the hand of a creator. Commenting Newton's perspective, Pierre-Louis de Maupertuis came up with the idea that in a remote past, some organisms produced by nature, that he calls monsters, became extinct because they were selected out by environmental conditions while other organisms, more functional, remained. In de Maupertuis' mind, initially, animals formed a continuous chain of relatives, but this chain was later broken due to the death of defective intermediary species [13]. In *Philosophie zoologique*, Jean-Baptiste Lamarck attempted to describe life and its specificity. In his view, simple forms of life can appear spontaneously from matter and therefore are the result of laws of physics, while more complex forms cannot appear spontaneously and are the product of complexification of simpler forms of life [14]. The presence of complex shapes means that evolution has happened, because they could not have emerged spontaneously. Lamarck explained this complexification (or evolution) by stating that parts of a body that are most used are strengthened and passed to their descendants, and the ones that are not used slowly degenerate; but this explanation remained at the stadium of an assumption. The first scientific explanation for the underlying mechanisms of evolution was done for the first time in Charles Darwin's *On the origin of the species* [15]. Darwin explains the origin and relation between species (Figure 1) by a phenomenon called

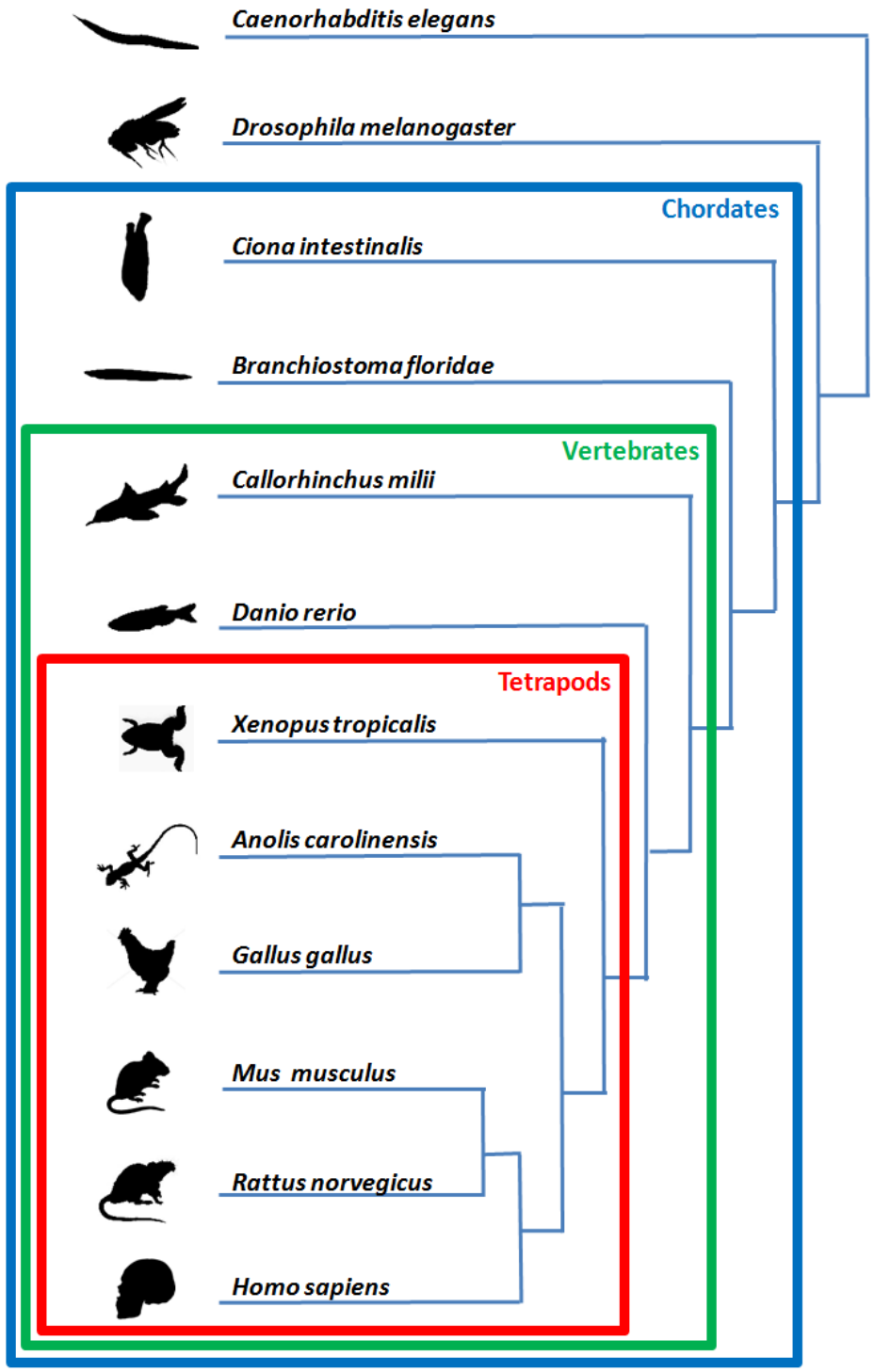


Figure 1. Phylogeny of various Metazoans. We show here well-known species from Chordates, which include *Drosophila* and *Caenorhabditis*, of Vertebrates, which comprise species with vertebrae, but exclude *Ciona* and *Amphioxus*; and finally Tetrapods, which comprise Vertebrates except fishes, coelacanths, and lungfishes (from [16]).

natural selection, a concept that had a huge influence on science and philosophy since the time of its publication.

So, what is natural selection according to Darwin? Evolution consists of three pillars. These are heredity, variation and selection. First, to have a living system evolve, it needs to transmit its characteristics to its offspring. This is what is called **heredity**, the transmission of traits from one organism to a second one that the first one has generated during a process called reproduction. But heredity is not enough for evolution to happen. To change, an organism has to generate **variation**, that is, differences of traits that naturally occur between individuals within a same population or species. Today, this variation is known to be the consequence of mutations happening in the DNA of individuals. The third element for natural selection to happen is **selection**; the concept describes the process that sorts individuals with different traits according to their different reproductive success (commonly referred as "fitness") [17]. Evolution by natural selection happens because some individuals are more likely to reproduce and transmit their traits to the next generation. They have a better fitness and are therefore more likely to generate viable offspring than others. Today, we know that aside from natural selection, another force can also generate evolution, and it is genetic drift. Genetic drift is a process by which random rather than advantageous traits are selected. This process is especially relevant for small populations. In this case, a mutation that generates a non-optimal trait in terms of fitness within an individual has a non-null chance to be fixed in the population, even if it is not advantageous.

We now summarize the concept of natural selection. It can be described as follows. If an individual shows phenotypic traits that give him an advantage upon the other individuals of the population, it will produce more descendants, and if the trait is transmitted from generation to generation, its carriers will see their frequency increase in time.

Since the end of nineteenth century, the theory of natural selection has been refined with concepts from genetics and is consolidated every day by accumulation of more biological data. Since the earliest characterization of genes and with the regular publications of genomes of diverse species in the last decade, the evidence for the relatedness between all forms of life has grown dramatically. The theory is proven beyond any shadow of a doubt.

1.3. Evolvability and robustness of living systems

1.3.1. At the level of genomes

A major gap in Darwin's theory concerns the origins of new phenotypes in the population. What is the source of variation? What is the mechanistic explanation behind the – apparently – spontaneous emergence of new traits in living systems? Since the publication of Darwin's opus, we know that variation is created by mutations and chromosomal events occurring in the DNA at specific positions that we call genes (though it can also happen outside, in regulatory sequences). In the seventies, Ohno described genes to appear by duplication of older genes. As a

consequence, in the pre-genome era, scientists were expecting that most of the diversity observable in living systems could be explained by screening the genetic differences between species. The publication of genomes of various species in the last two decades turned this expectation down. Today, by looking at genomic data, we know that the difference in terms of number of genes between human and say, *Drosophila* is much lower than was expected: flies display about 13,600 genes and human "only" 20,000! Moreover, a majority of protein sequences encoded within genomes are widely conserved among all forms of life. For instance, mice have a genome that presents 79% of protein sequences conserved outside of chordates, and 52% in non-animals [18]. As a consequence, diversity of shapes among forms of life cannot be explained by the sole description of gene sequences [19]. Variation has therefore to rise from a simple (though voluminous) vocabulary of about 20,000 genes, to build up an organism during development.

1.3.2. At higher degrees of cellular complexity

Since the foundation of molecular biology, complexity of phenotypic traits is viewed to come from regulatory mechanisms that can modulate the expression of genes topically and temporally. Francois Jacob stated that "novelties come from previously unseen association of old material"; "to create is to recombine" in a famous article [20]. In the sixties, he hypothesized that genes are regulated by DNA or proteins that bind specific sequences close to the gene location itself. Today, with the advancement of cellular biology and embryology, we know that regulation of gene expression happens at several levels, from genes to cells to tissues to organs. Theoretically, a genome is the same in all cells. The building blocks of genomes, the genes, and cellular pathways are basically the same in all types of cells. These blocks are conserved, but at the same time they can easily be combined to promote new function, and therefore variation. For instance, it has been shown that the shape of the beak of two different finches differs by the expression of only one gene in a particular tissue at a particular moment. In the developing beak of *Geospiza magnirostris*, which displays a large beak, so-called "primordia" cells express the gene coding for protein Bmp in the area of the developing beak at an earlier stage of embryogenesis, and at a higher level, than in *Geospiza difficilis*, which show forceps-shaped beaks [21]. The large beak of *G. magnirostris* is more suitable to crack nuts. It was shown that injecting Bmp protein in the developing embryo of a finch with forceps-like beak at the right time and the right place made the beak to turn into nut-cracker style. The simple expression of one gene at a certain time induced dramatic variation of a species' phenotypic trait. Though the core components that are cells and gene expression remain the same, their relatively "evolutionary-cheap" new combination promotes a radical change for the organism's biology. This property of living systems to evolve somehow easily from conserved core processes has been called evolvability, another word for the potential of living systems to evolve [22].

1.4. Strategies that promote evolvability

In this work, we decided to focus on two processes that provide living systems with the potential to evolve. One of them is innovation by gene duplication, with a special focus on duplications leading to repeats within proteins. Gene duplication is thought to be the primer motor of emergence of new genes. The other process is the use of existing proteins expressed in a particular tissue, to build a new physiological system. These two aspects of evolvability (among many others) are somehow related because gene duplication favors the selection of mutations in redundant systems and promotes the emergence of new functions within cells.

1.4.1. First strategy: Innovation by gene duplication

1.4.1.1. Concept of gene duplication

As well as complex species did not emerge spontaneously but rather from simpler forms of life, proteins are not built *de novo* but derive from ancient proteins. Probability for protein emergence from scratch in a complex organism with a huge amount of constraints is almost zero [20]. New sequences emerged by gene duplication, the process of copying a gene one or several times. Duplication prevents the old function to be corrupted while evolution can happen on the new copy of the gene. [23]. Gene duplication was shown to occur in Metazoans at a rate of once per gene per 100 Million Year [24]. There are two main forces that lead to such an event: unequal crossing over and retroposition [25]. Unequal crossing-over involves two homologous regions on two different chromosomes, so the duplication happens close to the gene, while retroposition is an insertion happening randomly in the genome. Moreover, entire genomes can be duplicated in one event. Several taxa are well-known for having experienced this type of genome-wide duplication events, such as the lancelet *Branchiostoma floridae* [26], or the teleost fish lineage [27].

1.4.1.2. Emergence of protein repeats

Proteins are primarily characterized by their sequence. It has to be translated from DNA into a 3D protein molecule and experience post-translational modifications. The most common features present in proteins are the domains, regions of the protein that fold in compact units. Domains are usually implicated in specific protein functions such as the catalytic activity of an enzyme or the interaction with another protein. Some of these domains are made of a series of similar motives called repeats. Repeats constitute 14% of all known proteomes [28]. A repeat is an amino-acid or a group of amino acids that occurs several times in a protein. Poly-glutamine (polyQ) is a well-known studied example of single amino acid repetition, whose expansion in diverse proteins is known to cause degenerative diseases. Repeats are therefore of medical interest and their study is both important for structure determination and for understanding the mechanisms of diseases. The most common repeats are repetitions of single residues, while repetition of longer fragments are less frequent and can form secondary structures like alpha-

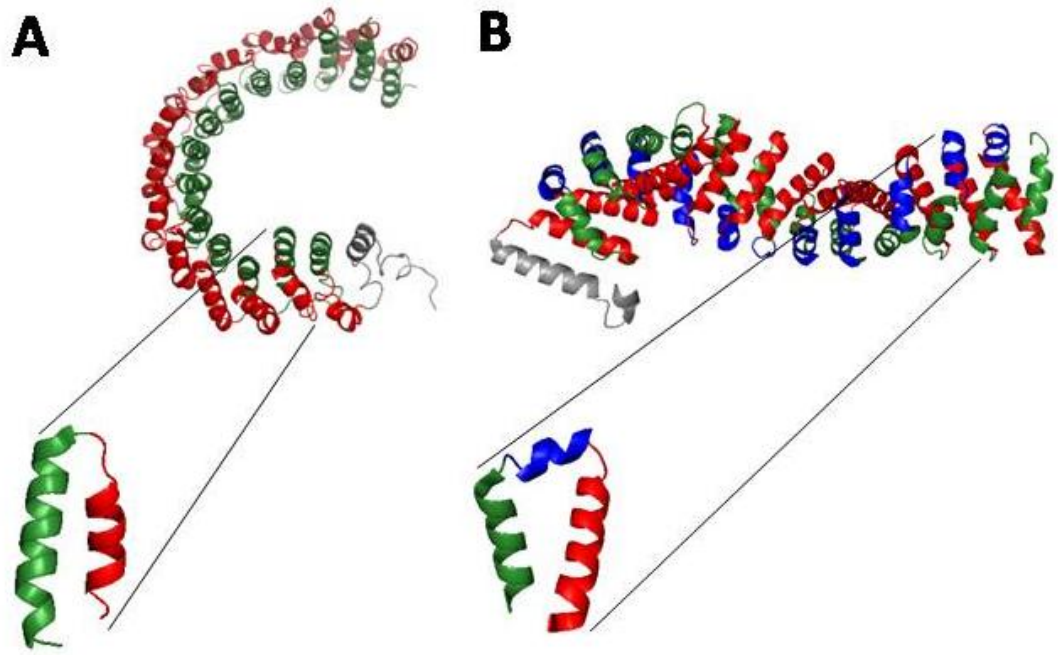


Figure 2. Representation of an alpha-solenoid protein, the regulatory subunit of PP2A. Alpha-solenoids are made of repeated motifs forming a flexible rod. One repeat is constituted by two alpha-helices disposed in anti-parallel fashion, separated by a small protein coil. A. Structure of the regulatory subunit of protein phosphatase 2A (PDB ID 2IAE, chain A [29]). The domain is an alpha-solenoid constituted of HEAT repeats (one repeat being made of two helices, here depicted in green and red). B. Armadillo repeats comprise two anti-parallel too, but they are interspaced by a small helix (here in blue; structure of Beta-Catenin [30], PDB ID 2Z6H).

helices or beta-sheets, that can assemble into domains [31] (Figure 2). When repetitions are made of fragments, while their structures remain similar, different degrees of similarity might be observed at the level of sequence, that sometimes are very low making very difficult the detection of repeats by sequence analysis.

Protein domains are usually well-conserved among living beings, because mutations might disrupt their compact structure needed to achieve their specialized function. In contrast, domains formed by repeats are generally less conserved. This probably comes from the fact that their structure is more flexible than the one of domains. They are typically associated to protein disorder [32] or high flexibility [33], which could explain their low conservation [34].

While protein domains emerge by duplication of proteins or protein fragments [35], the evolution of domains formed by repeats is particular due to their repetitive nature. Let's take the example of HEAT-repeats, which are about 39 amino-acids long, contain two anti-parallel alpha-helices and stack to form sorts of tunnel-shaped extended regions [36] (Figure 2). While it could have been possible to evolve these structures by producing secondary structures from scratch, it seems

more plausible that tandem repeats of fewer than 50 residues emerged by intra-gene duplication, by copying an original unique motif several times [34]. HEAT repeats are known to be involved in protein-protein interactions (PPIs) [37]. Their multiplication is likely to increase the surface of the protein, and therefore expand their functional landscape, as longer proteins on average are more promiscuous.

1.4.2. Second strategy: Evolving new physiological compartments

Properties of evolvability and robustness of living systems were clearly key elements in the evolution and diversification of metazoans from their last common ancestor. 540 million years ago started a time of great innovation among the animal kingdom, which saw an explosion of diversity of shapes [38]. The core components of several phyla appeared at that time. It was the time were cnidarians, insects and chordates emerged. The diversity of shapes has been correlated with the colonization of new environmental niches that started at that time [22]. Kirschner and Gerhart say that, in order to produce such diverse phylum, metazoans had to come "from a predecessor with great evolvability of compartmental body organization" [22]. Starting from a simple ancestral body plan, but flexible to change, due to some inner properties, rendered the emergence of new shapes cheap to do. Which processes allowed such diversifications to be possible?

1.4.2.1. Compartmentalization in metazoans

One way to innovate, rather than completely change a system, is to add modules to an already existing system. The same basic cellular elements can be used differently to participate in different functions, tissues, or groups of cells. With time, these elements will diverge enough to become separate modules within the organism's body, namely a tissue or an organ. This happens when living beings evolve to a higher degree of complexity. One of the ways to create modules is to express different genes at different places in the body during development, which results in apparition of a new morphology and a new function at a defined place in the embryo. This is the case of the expression of developmental genes in *Drosophila*. Several proteins determine the anterior/posterior axis of the larva; two of them, *bicoid* and *hunchback*, are more expressed in the anterior part of the embryo and two others, *nanos* and *caudal*, are expressed mostly in the posterior part [39]. Such patterns of gene expression help to determine the fate of cells within metazoans.

Compartmentalization can lead to production of different organs from the same embryonic elements. Differences in gene expression lead to very different morphologies of vertebrae in human. Bone-forming cells are present in all vertebrae. Nevertheless, they will produce a rib in the thoracic region, but they will not produce any appendage in the cervical region. These two different patterns depend on the expression of different combination of Hox genes happening in the same type of embryonic cells, but located at different position in the body [40].

In conclusion, compartmentalization is a good illustration of the difference between robustness and evolvability, and of their necessary relationship. While the basic components of cells such as genes or biochemical pathways remain the same, some subtle tunings can change their function and make them totally different from one organ to another. Basic cell functions such as DNA replication, translation, and phosphorylation are the robust part of the system but their modulation creates opportunity for change and consequently increases the potential of the system to evolve.

1.4.2.2. Evolution of physiological systems

Through evolution new functions emerge from already existing ones, for example enzymes whose mutation might give birth to a new catalytic activity. But anatomical and physiological features can also be the result of evolutionary processes.

Physiology is the science that studies the regulation of high-level processes within the human body. In simple living systems, such as bacteria, the biochemical pathways necessary to maintain the integrity of the organism, such as nutrition, response to stress or reproduction, are all the product of the same cell; and though these basic pathways are similar to those present in complex organisms, they have a limited range of action. In metazoans, cells are enclosed in a particular tissue that is devoted to a unique function of the human body. Red blood cells for instance are specialized in the delivery of oxygen to all the tissues of the organism. Physiological systems have a defined spatial distribution and are the result of specialization of cells during development.

One can wonder how physiological systems have emerged. Evolutionary theory teaches that living beings evolve by slow changes. New functions are built on the foundation of already existing functions, instead of appearing from scratch. Usually, the building bricks are the same but the way they are used is different. As an example, the basic components of neurons are very similar in the brain and in target organs such as muscles, but each type of neuron, central and peripheral, has evolved a specialized morphology and physiology adapted to its particular function.

One of the most famous examples of evolution of a new physiological system is the transition from the fin of fishes to the limb of terrestrial vertebrates. While comparative anatomy has described the morphological steps that link the two anatomical features, we now know that this evolutionary innovation was rendered possible by genetic innovation, probably from the evolution of a limited set of genes [41]. Another example of physiological evolution are the chromaffin cells of adrenal glands, which deliver adrenaline to the blood circulation. From a developmental point of view, they are homologous to post-sympathetic neurons [42]. This means that at some point in the past some ancestral neural tissue evolved to become specialized in the production of adrenaline.

Sometimes a complete new physiological system emerges from the expansion of another. Lungs are such an example. They are believed to have originally derived from digestive tract in fishes that lived in freshwater. The water was stagnant and poor in oxygen. At some time, they acquired the capacity to swallow air and increase oxygen level in their blood by making it cross the wall of the esophagus. The wall of the esophagus later expanded to give birth to a more complex structure A diverticula first emerged from the esophagus, and later increased the air/blood surface by developing many cavities and recruiting many small vessels eventually leading to the modern Tetrapod lung [43].

As shown by these different examples, new tissues or physiological systems with new function can evolve by specialization of already existing features. We believe that by studying the evolutionary history of physiological systems, such as the regulation of hypertension, by giving hints on where they come from, might be important to fully understand how they function and eventually dysfunction.

1.5. Methods to study protein mutations using structural and evolutionary information

1.5.1. Mutations in the context of genetic diseases

Mutations are genetic events happening at a frequency of $5x10^{-5}$ in mammalian cells [44]. Mutations mostly arise as a consequence of damages caused to DNA. In order to repair these damages, eukaryotic cells have a set of specialized polymerases. During replication, they synthesize the DNA molecule opposing the faulty sites in a process called translesion synthesis. Some of these polymerases are very accurate and produce a DNA strand faithful to the original sequence, while others are prone to inserting a faulty nucleotide in the sequence and therefore cause a mutation [45]. If the mutation happens in a coding sequence, it can have impact on the structure and function of the corresponding protein.

In this case, the protein is shaped correctly, but the local variation introduced by the site mutation prevents the protein to function properly. For instance, mutations introduced in a binding site of an enzyme could prevent the ligand to bind and the catalysis to happen. One example is a common mutation in an enzyme called phenylalanine hydroxylase, which participates in the catabolism of phenylalanine into tyrosine. The mutation occurs in the binding site of the enzyme and induces an accumulation of phenylalanine in the body of the patients bearing the mutation [46]. Other diseases caused by mutations are the ones impairing the functioning of the machinery that helps to repair DNA. Mutations in various proteins involved in nucleotide excision repair lead to genetic diseases such as Xeroderma pigmentosum. Such mutations prevent the proper nucleotides to be inserted where DNA was damaged, for instance with the action of UV light. As a result, mutations accumulate and can result into skin cancers.

1.5.2. Tools to explore the effect of mutations on protein structure and function

1.5.2.1. Sequence alignment

To predict the outcome of a mutation, one can study its conservation across species. In theory, if a region is important for a cellular function, it will be more conserved than average. An example of such conserved regions is the binding site of enzymes. Mutations will impair the efficiency of the enzyme by disturbing its catalytic activity, with potential dramatic consequences for the fitness of the individual. The mutant individuals will be discarded by natural selection. Consequently, the sequence of the enzyme will not vary much during the course of evolution. Thus, conservation in a multiple sequence alignment can be used to predict the outcome of a mutation and be helpful in the context of mutations associated to genetic diseases.

1.5.2.2. Prediction tools

Aside from manual alignments that depend on the user's personal technical skills, there have been several tools developed in the recent years to help deciding in a standardized way whether a mutation might be deleterious or not [47]. They rely on different assumptions and their results therefore may be conflicting at times. Prediction by the tool MutationAssessor is based on protein alignments [5]. Its algorithm scores mutations according to the pattern of conservation at the residue site considered. Again, the more conserved a site, the more disruptive the mutation would be. PolyPhen-2 (Prediction of functional effects of human non-synonymous SNPs) [48] is also based on multiple sequence alignments, but using the assumption that sites showing a higher nucleotide diversity in the human population are more likely to mutate. Differently, SDM (Site Directed Mutator) computes the impact of a mutation on the structure of the protein in terms of thermodynamics [49]. The variation of free energy of the molecule upon mutation is calculated. The smaller the variation, the more neutral the mutation will be. For such a computation, a 3D structure of the protein is required.

1.5.2.3. Protein visualization

Finally, screening the protein structure might help to get an idea on the impact of mutation. For this, one can use a representation tool that allows simulating mutagenesis on 3D models of proteins, such as PyMOL (PyMOL Molecular Graphics System software, DeLano Scientific, Palo Alto, California). Differently, PDBpaint (developed by us; see [50]) can be used to depict protein structures found in the Protein Data Bank and mark them with customized annotations and predictions from diverse sources, such as InterPro or Pfam.

1.6. Thesis outline

In this thesis, we exploit evolutionary information to predict the function of proteins. In the first place, we studied two evolutionary processes that are sources of variation for metazoan genomes.

The first one is the emergence of proteins with alpha-solenoid domains. We investigated the distribution of alpha-solenoid repeats across the tree of life. Their distribution is not uniform and shows that eukaryota have a much higher percentage of such motives than bacterial sequences. We show that prokaryotic sequences with alpha-solenoids are not related to the eukaryotic ones and have emerged independently. More interestingly, we find that alpha-solenoids are more enriched in two bacterial taxa, Cyanobacteria and Planctomycetes, which are morphologically more complex than other taxa, with sketches of organelles, and also have larger genomes. We conclude that alpha-solenoids are necessary for cellular functions associated to increased cellular complexity, but require evolved folding machineries.

The second evolutionary process is the evolution of the regulation of hypertension. A phylogenetic analysis was performed to study the emergence and evolution of the system's components. The major result is that anatomical features of the system started to appear in an ancestral chordate, while bony fishes already have all the major elements. This new physiological apparatus started to build 400 Million years ago. One of its key proteins, angiotensinogen, probably appeared in an ancestral cartilage fish around that time. Surprisingly, protein sequences of this system are also present in animals that do not display its anatomical features, meaning that these proteins had other functions and were later co-opted to create a new function. Our analysis shows that comparing sequences can be valuable to understand the embryological and evolutionary events that made the apparition of physiological systems possible.

In a third chapter, we study the consequences of mutations on protein stability, using structural and evolutionary information. We explain PDBpaint, a tool we developed to visualize annotation of different sources on 3D structures of proteins. Using alignments, and visualization and prediction tools, we study the impact of mutations within a human myosin, MYH6, and a huntingtin interacting partner, CRMP-1, and their potential impact on disease.

Publications. Chapter 2 and figure 2 are modified from a publication in PLoS One [51]. Chapter 3 and figure 1 are modified from a publication in Journal of Molecular Medicine, at Springer [16]. Chapter 4 contains parts modified from three publications: one in Bioinformatics, at Oxford Press [50], one in PLoS One [52] and a third one submitted [53].

2. Emergence of proteins with alpha-solenoids

2.1. Introduction

Living systems show an ability to evolve and develop new phenotypes. This ability comes from their capacity to intrinsically evolve upon changes in the environment and in their social context, and this process has been termed “evolvability” [22]. In this context of evolvability of living systems, we study here a protein feature, the alpha-solenoid domains, which are constituted of repeats of about 39 residues and are an important feature that we find associated to compartmentalization of metazoans. Alpha-solenoids might have played a role in the evolution of simple life forms into more complex ones, notably because of their capacity to promote protein-protein interactions and also to manage intracellular trafficking of proteins and nucleic acids, as this type of feature is often found in proteins that transport cellular material from one part of a cell to another.

To study the properties of alpha-solenoids, we first developed a tool to detect alpha-solenoids inside of protein sequences. We later studied their functions and diversity of shape, and we finally describe their distribution across the tree of life. From our observations, we conclude that this type of repeats very likely emerged independently in prokarya and eukarya; moreover, they are clearly in higher amount in taxa of higher cellular complexity, in bacteria as well as in metazoans, hinting their potential participation in the formation of compartments in living systems. These results are explained in sections 2.2 to 2.5 of this chapter, which are inspired for the most part by a publication from us [51].

Aside from functional and evolutionary analyses, we also provide an example of the importance of alpha-solenoids in the context of genetic diseases, by studying the sequence of huntingtin, a large protein comprising several alpha-solenoids that is involved in Huntington’s disease. By doing this we demonstrate how our methodology can be used to propose functional and structural predictions for a protein of medical interest. This study of huntingtin is exposed in section 2.6.

2.1.1. Functional and genomic analyses of alpha-solenoid proteins

Alpha-solenoids are long protein domains made of structurally similar pairs of anti-parallel alpha-helices. They form a structure shaped as a hollow rod or solenoid, from which they got their name [54]. Three classes of repeats are known to belong to the alpha-solenoid family ([55]; sometimes referred as Armadillo family by InterPro [56] or other tools): HEAT repeats, famous to be present in protein huntingtin ([36,37]; figure 2A), Armadillo repeats ([57,58]; figure 2B), and HAT repeats [59]. In some cases alpha-solenoid protein sequences have very little sequence homology with already known alpha-solenoids, hinting that similar structural elements might have appeared by convergent evolution [51].

Regarding physical properties, alpha-solenoids are highly flexible and therefore exhibit an extraordinary capability to stretch [29,33], which allows them to display a great compliance of shapes [60], and render them suitable for protein-protein interactions [37]. For instance, the regulatory subunit B of protein phosphatase 2A (coded by gene PPP2R5C) displays elastic changes of shape that impact the opening and closing of the binding site in the catalytic subunit C of PP2A so target proteins can bind or not [29]. Alpha-solenoids are also used for protein transportation, mostly via importins, which bring cargo proteins from cytoplasm to nucleus [61]. The dynamics of importins have been studied; these proteins show a tendency to self-wrap, creating a super-helix which is itself stretchable, adding more potential for regulation [60].

Detection of alpha-solenoids from sequence alone when no associated structure is available is especially relevant to get hints about the tertiary structure of the given protein potentially with a good precision. It is also relevant to understand the biology of proteins of special focus in medicine, when no structure is yet available. This is the case of the protein huntingtin, whose mutation causes neurodegenerative disease [62], and also of mTOR, which is a key protein in pathways related to cancer [63]. These two proteins are particularly large, which render the resolution of their structure experimentally tricky or impossible at the current state of knowledge. As a consequence, computational analyses are often the only tool available to accurately study the relationship of the structure of these molecules to disease [55,64].

Most methods to detect protein domains are based on homology, and users primarily rely on profile-based methods, using information stored in databases such as Pfam [65] or SMART [66]. In such databases, the profiles of sequences of alpha-solenoids are mostly based on canonical HEAT and Armadillo repeats; upon query one may miss many diverse alpha-solenoids as many of them are highly divergent or have no homology at all between each other, which can in certain cases limit detection by sequence similarity searches.

As a consequence, methods of detection that are not based on homology should be considered. Among such methods, algorithms based on artificial neural networks use very little prior knowledge for pattern identification and could be potentially more efficient at detecting relevant information in unknown sequences. This type of approach has been successfully applied to detection of structural motives such as secondary structure or transmembrane helices [67]. Following the same concept, an algorithm based on an artificial neural network (ARD, for Alpha-rod Repeat Detector) has been applied to the detection of alpha-solenoid repeats [55].

In the recent years, new structures of alpha-solenoids have been published, some of which have no similarity to known alpha-solenoids, so we hypothesized that the available neural network might benefit from inclusion of new structures to its training set. We further improved the algorithm of the program by allowing the detection of repeats with more variability of structure than before, in an attempt to enlarge the spectrum of possible detectable alpha-solenoids. We optimized the tool by testing it on known alpha-solenoids from the Protein Data Bank (PDB), and proved that our new tool, named ARD2, has an increased coverage.

We finally applied our algorithm to all available sequences from the TrEMBL database, and we explored the distribution of alpha-solenoids across the tree of life. Our analysis strongly suggests that alpha-solenoids have emerged independently several times, as several bacterial and eukaryotic groups are apparently not related by sequence homology, which supports events of convergent evolution. Moreover, we show that the distribution of alpha-solenoids is not uniform; they are highly represented in eukaryotic organisms; in bacterial species they are comparatively rarer than in eukaryota, and are interestingly enriched in a few taxa that display pseudo-compartments within their cells. These groups are cyanobacteria [68,69] and planctomycetes [70], among others. As a result, we conclude that alpha-solenoids are associated to events of high cellular or physiological compartmentalization (in bacteria and metazoans, respectively), which clearly demonstrates the functional importance of putative protein transport alpha-solenoid proteins in carrying proteins from one compartment of cells to another and therefore in participating to the potential of living systems to evolve more complex functions.

2.1.2. Function of huntingtin alpha-solenoid region and prediction of consequence of mutations for its structure

Moreover, we apply the ARD2 algorithm to the modeling of the alpha-solenoids of huntingtin, the protein whose mutation is responsible for Huntington's disease. We show that the first alpha-solenoid of huntingtin displays a functional signal: it specifically contacts proteins that are part of ribosome subunits, pointing a potential role for this domain in translation of mRNA into proteins. We later produce a model of this alpha-solenoid based on a threading method confirmed by ARD2 predictions of repeats positions. We then use this model to predict residues potentially involved in protein-protein interactions. In the end, in the context of PPI networks, these could be especially relevant to understand how this protein functions by interacting with other proteins.

2.2. Detection of alpha-solenoids

The main databases for protein domain annotation use *a priori* determined protein alignments, called profiles, of several homologous sequences that can be used to detect domains in sequences provided by the user as query. Examples of these databases are Pfam [65] and SMART [66]. Domains made of repeats are no exception and can be detected using the same type of homology-based methods. The first tools that were developed to detect specifically protein repeats were based on profiles of aligned sequences [71,72,73]. More recently, profile-profile comparison methods have been put into place, such as the HHrepID server, which detects repeats *ab initio* by creating a new profile for the query sequence from homologous sequences [74].

Nevertheless, protein repeats usually show extreme divergence, and this property can render some very divergent repeats difficult to detect with homology-based methods [31]. Other possible methods have been proposed as alternatives (see [31] for a review), such as Fourier transformation [75,76,77], short string extension algorithms [78,79], and methods comparing portions of a protein sequence to other regions of the same protein that look for internal

similarity of sequence within a protein [80,81]. One last possibility is to combine evidence coming from several of these tools to identify repeats more accurately [31]. The commonality between these different methods is that they are more naïve than profile-based methods and need no or very little preliminary knowledge about protein sequence to be able to perform their detection.

Recently, a neural network based method has been successfully applied to the detection of alpha-solenoid repeats in an attempt to detect as many alpha-solenoid structures as possible with the smallest possible amount of prior knowledge on sequences (ARD, for Alpha-rod Repeat Detection [55]). The network is trained with a reduced set of sequences comprising alpha-solenoid repeats, and is used to assign a score to any query sequence. Here we present the development and application of an improved version of this method.

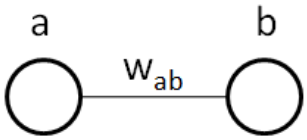
Section 2.2.1 is an introduction to artificial neural networks. In section 2.2.2, we describe the general features of the ARD algorithm and in section 2.2.3, we describe the improvements that we have implemented. We apply the new tool in later sections of the current chapter to do a phylogenetic and functional analysis of alpha-solenoids, and we finally apply the method to modeling the structure of huntingtin.

2.2.1. Introduction to artificial neural networks

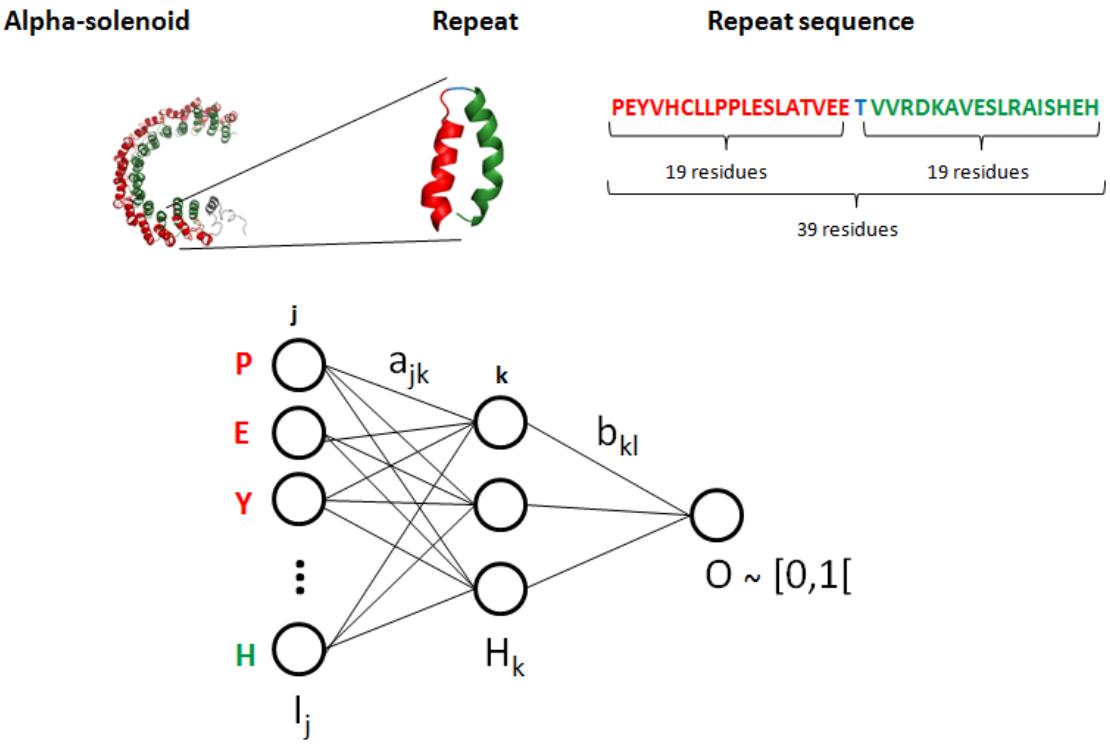
Such as real neuron networks, artificial neural networks react to a stimulus or a combination of stimuli and give an appropriate answer in response to it. The answer depends on the property of the stimulus and on the wiring of the neurons within the network. In order for a network to give an appropriate answer to a given stimulus, it needs to be properly wired, i.e. to display an efficient interconnection of its neurons. A random wired network that has not learnt anything yet, can be trained to recognize specific patterns (named ‘inputs”) and deliver an appropriate answer, such as a yes/no answer (or “output”). Artificial neural networks have been used in several areas, such as pattern recognition (for face identification for instance) and decision-making programs (such as chess simulators) [82]. Programming an artificial neural network requires two phases. The first one is the training of the network. In this phase, a series of inputs (forming a training set) are given to a randomly wired network, which is modified if it does not give the answers it is supposed to give in response to the data present in the training set.

The most basic neural network imaginable has two connected neurons (figure 3A). The left neuron receives a certain stimulus; say for instance a certain concentration X of neurotransmitter. This stimulation may eventually depolarize the membrane of the neuron and be transmitted along the axon of neuron A that points toward neuron B. Nevertheless, the axon has a certain conductance and might not completely deliver the message to neuron B. Therefore, to know the output value that will be acquired by neuron B after processing by neuron A, the value X of input will have to be multiplied by a factor or weight W that represents the fraction of signal transmitted from A to B. In the end, the value is transformed by a function F, which converts a

A


a
b
w_{ab}

B


Alpha-solenoid
Repeat
Repeat sequence
PEYVHCLLPPLESLATVEE TVVRDKAVESLRAISHEH
19 residues
19 residues
39 residues
j
P
E
Y
H
a_{jk}
k
b_{kl}
O ~ [0,1[
H_k
I_j

C

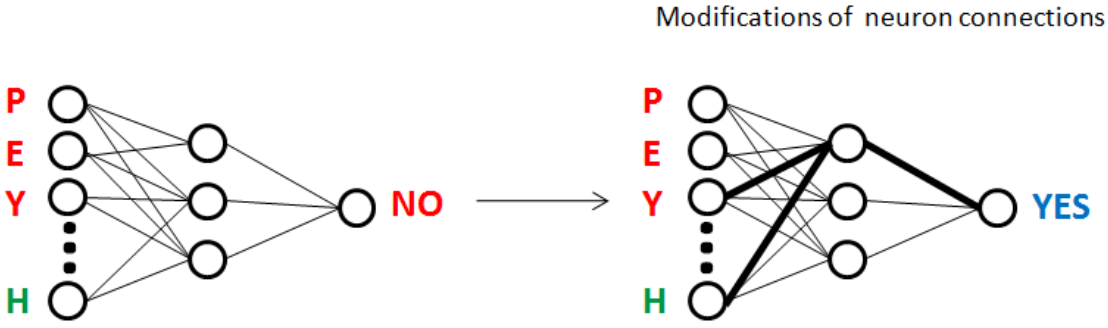

Modifications of neuron connections
P
E
Y
H
NO
P
E
Y
H
YES

Figure 3. An artificial neural network for the detection of alpha-solenoid repeats. A. Simple case of a neural network comprising two neurons. Neuron A is the input neuron, or transmitter and neuron B is the output neuron, which processes the message arriving from A and gives an answer in response to it. W_{ab} is the weight associated to the connection and can be seen as a conductance. The lower the weight, the weaker the transmission will be. **B.** View of a three layer neural network. The first layer, I_j, is made of input neurons, each one acquiring a part of the input message. The message is layer processed to a second layer of neurons, H_k, to be processed again, to be finally transmitted to the output neuron layer O, which delivers the final answers of the network. In the context of the detection of alpha-solenoid repeats, the input is a sequence of 39 amino-acids (two alpha-helices in red and green separated by a hinge, here in blue). **C.** Training of the network. Initially, the network is naïve and does not give the answer requested by the training set. As training progresses, the weights that bind neurons with each other are modified using a training algorithm, so the network gives a proper output. All weights are changed at the same time to facilitate the computation.

positive value that follows, for example, a non-linear sigmoid function in the range [0,1[. For a given input a, the corresponding output value b can therefore be calculated as follows:

$$b=F(w_{ab} \cdot a)$$

$$\text{with } F(x)=1/(1+ e^{-x})$$

This formula implies that a connection can be turned off by assigning to the weight a value of 0.

More complex networks comprise several neurons, which display contacts with their neighbors (figure 3B). Three types of neurons are usually intervening. Some are receiving the input of the network, which can be either a numeric value or a discrete value converted into a number. Some neurons are outputs and play the role of neuron B in figure 3A. Finally, several neurons usually play the role of intermediate and expand the possible connectivity of the network.

Firstly, we explain how information is processed in a network, to get the output from a given input, whether the network is "clever" (i.e. trained or correctly wired) or "naive" (i.e. not trained yet). At first, the input neurons acquire the input values (I); then transmit them to the intermediary neurons. In this case, the value of output of a neuron in the hidden layer (H) (see figure 3B) will be calculated as follows [82]:

$$H_k=F(\sum_j(a_{j,k}I_j + \Theta_j)) \quad \text{with } j=1, \dots , 39\times20 \text{ and } k=1, 2, 3.$$

$$\text{With } F(x) =1/(1+ e^{-x})$$

For each neuron, the value received from a given neuron from the input layer is multiplied by the weight a_{jk}. The product is then added to a value called bias Θ, always constant. Then transmission happens between the hidden layer and the output layer. In this case, the value of output of the network will be calculated like this:

$$O=F(\sum_{k}(b_{k,1}H_k + \Theta_k)) \quad \text{with } k=1, 2, 3$$

$$\text{With } F(x) =1/(1+ e^{-x})$$

This formula gives the unique value of output for a given input and a given network. Again, the output value of each neuron from the hidden layer is added to the bias Θ. The function of the sum of the outputs finally determines the value O of the unique neuron of the output layer.

Secondly, we explain how to train a network. Let's assume that the network is naïve, another word to say that it has close-to-random values of weights, and is therefore not capable to deliver an appropriate output to a given input. To train this network, we will give it different inputs sequentially and we will rewire the connections between neurons so the network delivers an appropriate answer. There are several algorithms to perform this rewiring, the most used being back-propagation [83]. This is the one we use here.

The training set is a list of two-dimensional vectors $\{(x_1, y_1), (x_2, y_2), \ldots (x_n, y_n)\}$. It consists of n inputs x_i associated to their associated theoretical output y_i, that have to be taught to the network. The output y_i are expected to be given by the network given input x_i.

For each couple (x_i, y_i), an associated deviation can be computed to give an idea of the difference between theoretical values y_i and observed value y'_i. This deviation is called quadratic error (E) and can be calculated using the following formula:

$$E=1/2 \mid y'_i - y_i \mid ^2$$

Errors E for all couples (x_i, y_i) of the training set are calculated and are then added up to get a total error E_T. The value of E_T has to be minimal if the neural network is properly wired and delivers the outputs it is expected to deliver.

The next step is the back-propagation itself. In consists in calculating for each weight the value $\delta E/\delta b_{kl}$, where E is the quadratic error of the output neuron and b_{kl} the weight of edge kl between output and hidden layer, and change the weights in order to minimize the quadratic error of the network in response to the training set. To modify the weights, i.e. rewire the network, the following formula is applied at each node to compute the weight increment Δb_{kl}:

$$\Delta b_{kl} = - \gamma (\delta E/\delta b_{kl})$$

γ being a learning constant. As the function that helps to pass signal from one neuron to another neuron of a downstream layer is b·x (product of the weight b to the input neuron value x), the back-propagation of the error in the other direction (from the single neuron of the output layer to a neuron of the hidden layer) can be done by the simple product of the error by b, b being the primitive of function b·y (b is the weight, and so is a constant and y is the value of the output neuron, a variable). By this simple product, the error can be transmitted to the output neuron to neurons of the hidden layer. The weights are then modified according to the value requested by

calculation of error deviation. In a second step, back-propagation happens between neurons from the hidden layer to neurons from the input layer, using the same process. Weights a_{jk} are modified in the same way as were weights b_{kl}.

After these two steps (back-propagation of error from output to hidden layer, and back-propagation of error from hidden to input layer), the network is given a new sequence x associated to its theoretical output y for training. The total quadratic error is computed and the weights are modified again. After the whole training set has been used to teach the network, and E_T has converged to an acceptable value, the neural network is validated. On the contrary, a new round of weight increments is performed, until convergence is eventually observed. After a certain number of iterations, if no convergence is visible, then the neural network is considered to be incapable to learn using this combination of training set, learning constant and function F; consequently some of these parameters should be modified to make learning possible.

2.2.2. Presentation of the neural network of ARD

ARD is an artificial neural network coded in Fortran and is similar to the one presented in figure 3B. This network is made of three layers of neurons, a first one of 39×20 neurons, a second layer of 3 neurons, and a third made of one neuron. All neurons of a layer n are related to all the neurons of layer n+1 by weights and these weights are stored in a matrix. The first layer takes sequences of 39 residues as input (portion of a protein sequence). For each position there are 20 neurons, one for each of the 20 amino acids. The intermediary layer is purely computational and adds complexity to the system. The third and last layer is the output layer and gives values between 0 and 1, depending on the input given in the first layer and the connections between the different layers. If the network is appropriately designed, given an input sequence of 39 residues, it should give a high score if an alpha-solenoid repeat is detected, a low score otherwise. This allowed the repeats to be detected accurately at precise positions. Sequences of 39 residues within a sequence resulting in high scores identify an alpha-solenoid repeat, the linker being at position 20 surrounded by two sequences of 19 residues. Typically, if a query protein sequence has to be tested, the network will start by testing the residues 1 to 39, then the residues 2 to 40, and so on until all sequences of 39 consecutive amino acids of the protein are tested. For each sequence of 39 amino acids presented to the network as input, an output equal to “1” was expected if the position was in the middle of an alpha-solenoid repeat, and value 0 was expected at other positions. Consequently, most of positions of an alpha-solenoid protein are expected to be 0, and these positions constituted a negative set for the network. To optimize the algorithm, we considered a sequence to be an alpha-solenoid if several repeats in a row were identified with a high score (see later).

The network was trained using a back-propagation learning algorithm (figure 3C, see [83]) in the way exposed in previous section 2.2.1.

2.2.3. Improvements of ARD

Firstly, we attempted to provide a better performing training set than in ARD. We decided to use a set of 27 proteins with alpha-solenoids determined by high quality alignments that were already used for ARD [55,84]. We later tried to add more sequences from our set of alpha-solenoids with known structures to this basic training set (table 1; see section 2.2.4 for details). Nevertheless, the algorithm was very sensitive to addition of sequences. Only one sequence, an ankyrin protein (PDB ID: 2AJA [85]), could improve the results. The final training set, which comprises 28 protein sequences, is presented on table 2.

One could argue that the sensitivity of our network to modification of the training set could be due to overfitting, an event happening when a network learns from a training set, but later is incapable to detect something outside of the sub-universe of the training set. In case of training sets made of protein sequences, overfitting could be due to the homology between the different sequences. Nevertheless, sequences of our training set, as for many alpha-solenoids, display a very weak homology [37]. As a consequence, as mentioned later in sections 2.3, 2.4 and 2.5, we were capable of detecting several types of repeats that were not homologous to sequences of the training set. One example of these non-homologous alpha-solenoid sequences are PBS lyases (see section 2.5 on evolution of alpha-solenoids for more details), which proves that our algorithm has potential to detect new sequences, especially the ones with no homology to the ones of the training set, ruling out that our network is affected by overfitting.

We then attempted to improve the performance of the neural network by modifying the values of the parameters. The fundamental structure of the network remains identical to the one of the first version of ARD; it displays three layers of neurons and uses a non-linear sigmoid activation function to adjust output values. Differently though, four elements are under variation: the score threshold, the window shift, the distance between repeats, and finally the number of repeats. A comparison of the performance of the network associated to different combinations of parameters, with the new training set, is presented in table 2. The way this performance was tested is explained in the next section. Here we first detail the parameters that were optimized.

Score threshold. This is a real value between 0 and 1. As 0.8 was used as threshold in ARD and we now potentially expand the number of structures but also the false positives, we thought that being more restrictive and setting a higher threshold would be necessary to reduce the rate of false positives.

Window shift. The algorithm of ARD assumes that the linker between two alpha-helices of an alpha-solenoid repeat has the same length in all proteins. This is factually not true, as hinges of different lengths, are visible in alpha-solenoid proteins, sometimes with very long insertions and this apparently without disturbing the stability of the correct protein folding [86]. Therefore, we allowed the central linker of a repeat to have a length greater than 1. For each position tested as

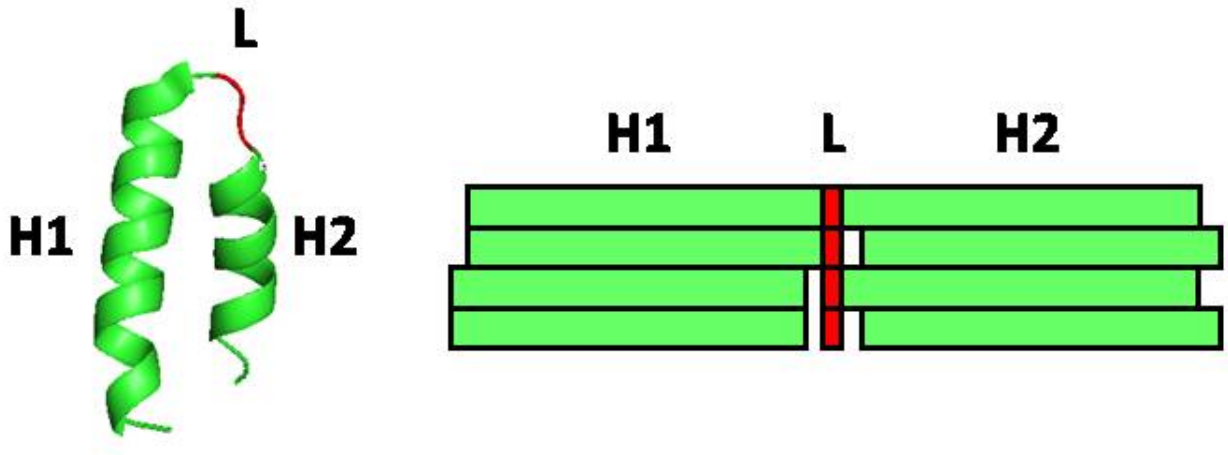


Figure 4. Diagram showing the window shift for repeat detection. A repeat is made of two helices (H1 and H2) separated by a linker sequence (L). Two detection windows of 19 amino acids are considered, one for each helix. During detection, different window shifts are tested by sliding the input windows H1 or H2 one residue apart from the middle-residue (red box), as indicated by the gaps between red and green boxes.

central residue of a repeat, we not only tested the immediate 19 neighboring residues on both sides of the central residue but also examined as alternatives the 19 residues neighbor to position -1 and +1 from the central residue. This window shifts now allow the linker to be 1 to 3 amino acids long (figure 4). Therefore, for each central position, four combinations of windows displacements were tested and the maximum value and corresponding windows displacements were reported.

Distance between repeats. The distance between positions with scores above threshold was also subjected to variation. We tested several values of distance between repeats and found that setting distance in the range 30 to 135 gave the best performance (data not shown).

Number of repeats. To be identified as an alpha-solenoid, a given sequence had to display a minimal number of repeats. We started to test a minimal number of 3 repeats, and then increased the value to see how many were needed to optimize the algorithm. We found that 3 repeats were an acceptable minimal number that shows a high precision in combination to appropriate values of other parameters. Minimal numbers of 1 and 2 repeats have previously been tested for ARD and proven to be associated to a high number of false positives [55]. This is also following the trend that alpha-solenoids contain in the range of 3 to 50 repeats, though some could be made of only 2 repeats (PDB ID: 1ZQ1; [87]).

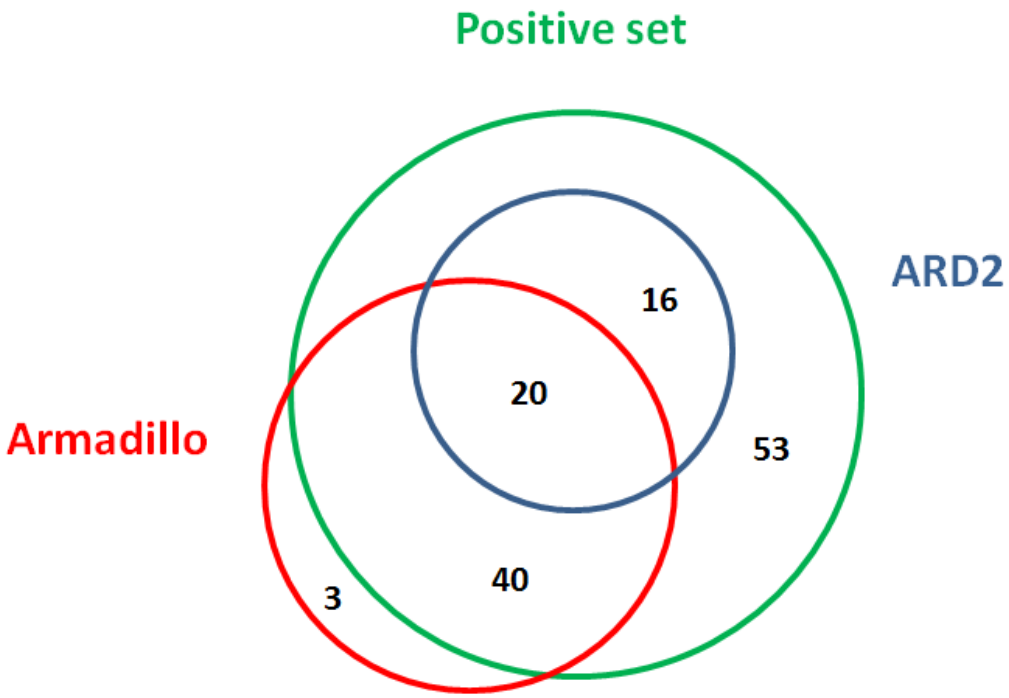


Figure 5. Comparison of structures recalled from the positive set by the Armadillo profile from InterPro and ARD2. The positive set is presented on table 1. Proteins detected outside of the positive set circle (Green) are consequently false positives (See table 3 for a detailed list of the proteins detected by each of the three methods).

2.2.4. Evaluation of ARD2 performance

In order to determine if ARD2 discriminates correctly alpha-solenoid from non alpha-solenoid proteins, we tested it on a set of 19,769 structures from the Protein Data Bank (PDB, [85]). This set was obtained by reducing the redundancy of proteins available in PDB. Initially, 174,488 sequences were extracted from this database and were classified into 23,710 clusters using a conservative algorithm [88]. Short sequences of less than 20 residues were then removed, as well as structures whose quality was not acceptable according to the NCBI standard (which is defined as a file at URL ftp://ftp.ncbi.nih.gov/mmdb/nrtable/nrpdb.latest). In the end, 19,769 sequences remained. For each cluster, one PDB structure was chosen according to the following parameters (in decreasing order of importance): best resolution of solved structure, lowest percentage of unknown residues, lowest percentage of missing residues, and longest sequence. We tested ARD2 on this set of proteins and the structures positively identified to contain alpha-solenoid repeats were manually checked using the protein visualizing tool PDBpaint (developed by us; see section 4.2 and associated publication [50]). From the prediction of ARD2, and with the help of additional manual literature analysis, we could identify a total of 129 alpha-solenoid proteins among the 19,769 non-redundant protein sequences from PDB. The PDB identifiers are listed in table 1. As the remaining 19,640 proteins from PDB, with the exception of 3 of them, are not

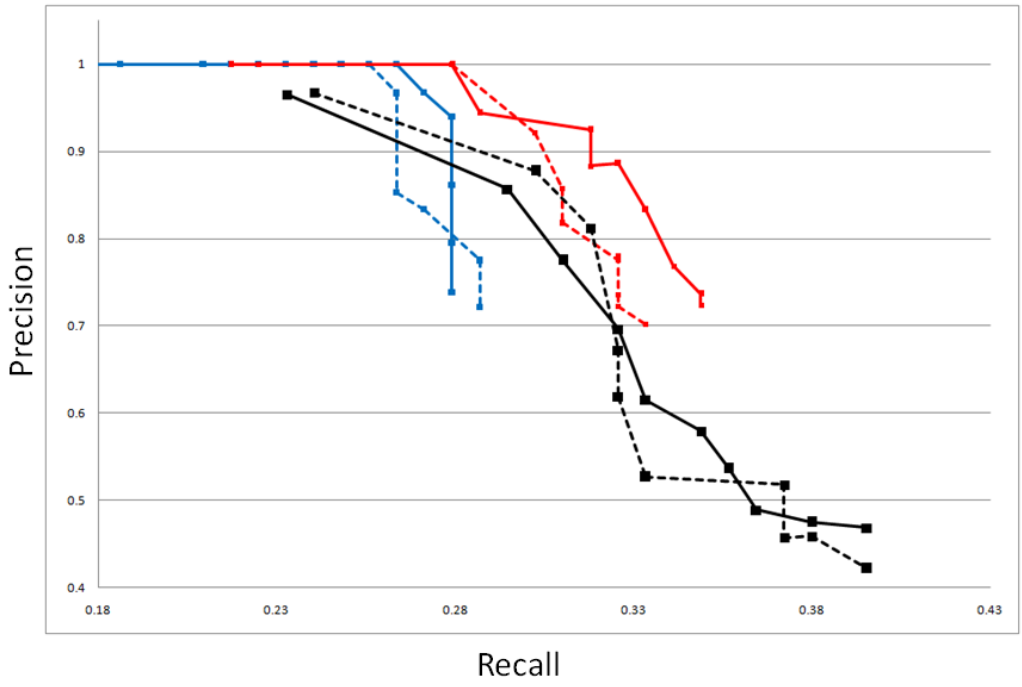


Figure 6. Precision-recall curves comparing the performance of ARD2 in identifying alpha-solenoids in our PDB set using different sets of parameters. The main parameter is the value of the window shift, which represents the possible length of the hinge between the two helices of an alpha-solenoid repeat (blue curves for absence of window shift, red curves for a window shift of 1 and for recollection we show result for a window shift of 2, with black curves). For each of the three values of the shift (0, 1 or 2), two training set are tested, the initial one from ARD (discontinuous lines) and the expanded one of ARD2 (continuous lines; see table 2 for a list of the proteins in the expanded training set). Finally, for each combination of window shift and training set used, different threshold of the network are tested, from 0.8 to 0.9. For all values, minimal number of repeats required to detect an alpha-solenoid within a protein was 3. For each combination, we calculated the number of proteins detected among the set of positives from the Protein Data Bank (table 1), checked if they were true, and then inferred precision and recall for each value. The best recall for a 100% precision is obtained when using the window shift and a score threshold of 0.87 (precision: 1.00, recall: 0.28). The ARD2 training set produced generally better results than the ARD training set, and resulted in the best value of precision x recall for a threshold score of 0.86 (precision = 0.93, recall = 0.32). Window shift of 2 did not perform as well as shift of 1. Though the recall can be higher with low thresholds (0.80 to 0.84), the precision rapidly drops, and is all cases lower than 100%. The best performing value in both precision and recall (defined as value precision x value) was obtained with the expanded training set, a window shift of 1 and a threshold of 0.84.

predicted as alpha-solenoids by InterPro (figure 5, only 3 proteins out of the set are predicted to be alpha-solenoid, which were found to be false after manual checking), these proteins constituted our negative set for the performance evaluation of ARD2 and other tools.

After this preliminary determination of a positive set of alpha-solenoids (with a putatively suboptimal combination of parameters of ARD2), we tested ARD2 performance by determining which combination of parameters had the best recall for a precision of 100% (table 3 & figure 6). The highest recall for a 100% precision had a value of 0.28. Parameters were set as follows: 3 repeats minimum, a distance range between repeats of [30,135], and a score threshold of 0.87. The method was able to identify sequences as alpha-solenoids that had no significant sequence similarity to any of the 28 sequences used in the training set. For example, the E-values of sequence similarity (according to BLAST) to the best match to the sequences in the training dataset were above 0.01 for human rotatin (UniProt ID: Q86VV8) (E-value = 0.071) and for predicted proteins UniProt ID: Q7ULY0 (from *Rhodopirellula baltica*, E-value = 0.16) and UniProt ID: A8JFV2 (from *Chlamydomonas reinhardtii*, E-value = 0.047).

We kept the different parameters (minimum of 3 repeats, distance range in [30,135] and threshold of 0.87) for the new version of the neural network, which we call ARD2. We later tested more relaxed parameters for ARD2 and found that we could identify more proteins but with an important proportion of false positives. A threshold of 0.5 for instance improves the recall to 0.46 but at the same time reduces precision to 0.22, with identification of 206 false positives.

We then compared the structures annotated as alpha-solenoids by ARD2 to the ones stored in InterPro [56], a database that gathers protein structure predictions from various sources, including profile-based tools Pfam [65], PROSITE [89], and SMART [66]. We compared these sets to the alpha-solenoids detected among the available PDB structures. Coverage of InterPro and ARD2 of the alpha-solenoids from the PDB were different, with some overlap (figure 5), but InterPro annotated falsely three proteins as alpha-solenoids, outside of the positive set (see table 3 for more details). These results suggest that ARD2 can complement profile-based tools and point out new types of alpha-solenoids with no homology to known alpha-solenoid structures, and this with a good precision.

We decided to provide the algorithm of ARD2 as a web interface where users can give a protein sequence as query to search for alpha-solenoid repeats. As we found that these predictions with sub-optimal scores were potentially useful information for the user, we decided to provide accessibility to all predictions from ARD2, associated to optimal scores or not. Nevertheless, caution should be taken regarding the predictions with sub-optimal scores. Predictions should be confirmed by additional information from secondary structure prediction, and other tools such as de novo structure determination tools. ARD2 is available at http://cbdm.mdc-berlin.de/~ard2/. Improvements of the neural network were coded in Fortran 90, as was the initial program [55]; the Fortran code was compiled with gFortran 4.4 on a Linux platform. The web service was

coded in Perl 5.10.1 and the server runs under an Apache 2.0 web server. An option to annotate PDB structures according to ARD2 predictions was added to the visualization tool PDBpaint [50], presented later in section 4.2, to allow seeing the 3D structure of a protein and the associated ARD2 prediction in the same view.

2.3. Structure of alpha-solenoids

In section 2.2.2, we have presented a set of curated positive structures (table 1). We have studied the entire set and review here their different structural properties; some of them seem to challenge the canonical description of alpha-solenoids (figure 2A and 2B; [90]).

We found that the set of structures of alpha-solenoids available in PDB is dominated by the presence of karyopherins (alpha-importins alpha, beta-importins and transportins), which represent 26 proteins out of 129 (20%), though this percentage is very likely biased toward choices of experimentalists, and may not give a complete picture of the diversity of functions of alpha-solenoids. Other functions of alpha-solenoids include activation of transcription factors, regulation of translation, vesicle trafficking, DNA repair and RNA processing, either tRNA, pre-mRNA or pre-miRNA. We give a review of our different findings that challenge the former knowledge about alpha-solenoids. The diversity of structures found in the PDB is displayed in table 4.

2.3.1. Some types of alpha-helical repeats are newly classified as alpha-solenoids

Alpha-solenoids comprise Armadillo, HEAT, and HAT repeats [55], among others, which do not have significant sequence homology although adopt similar structures. Our method proved to expand the definition of alpha-solenoids by pointing to structures that were not classified as such before. These new types of alpha-solenoids show no significant sequence similarity to Armadillo and HEAT repeats, which hints that our neural network can detect proteins of similar structure with no sequence homology.

We firstly identified a protein with TPR (Tetratricopeptide) repeats to be an alpha-solenoid (figure 7A). It is an important virulence factor from *Bacillus thuringiensis* involved in quorum sensing of bacteria (PDB ID: 2QFC [91]). Moreover, we identify Pumilio repeats to be alpha-solenoids (PDB ID: 3K62). These repeats are specialized in binding RNA [92] but are not classified as alpha-solenoids by SMART or Pfam.

We also identified the ankyrin repeats of protein of UniProt ID Q5ZSV0 as being an alpha-solenoid. Though ankyrin repeats show a similar structure to alpha-solenoids, their middle coils are usually longer and twisted [93]. The structure identified (PDB ID: 2AJA) has a more classical alpha-solenoid look than typical ankyrin repeat structures and can be considered a very good match (figure 7B). As explained before, this ankyrin protein was already detected in preliminary attempts to explore alpha-solenoid function and structure using the sole improvement of window shift but with no training set expansion of ARD; this protein was later

successfully included in the training set of ARD2, and demonstrate its capability to improve alpha-solenoid detection using non-homologous proteins (see section 2.2.2).

Finally, we identified a very irregular alpha-solenoid in a bacterial glutamyl-tRNA synthetase (PDB ID: 3AL0; see [94]), which comprises repeats that show a much higher twist (about 90º) than canonical alpha-solenoid structures ([60]; see figure 7C). We also point that some proteins with leucine-rich repeats could also be alpha-solenoids: protein of UniProt ID Q44534 is such a case, with repeats having a clear alpha-solenoid shape [95], though their length (average of 24.4 residues) is shorter than the typical value of 39 residues of the average alpha-solenoid repeat.

2.3.2. Alpha-solenoids can interact with nucleic acids and lipids

Alpha-solenoids are not only having direct contact to proteins. They can also physically interact with DNA, RNA and lipids.

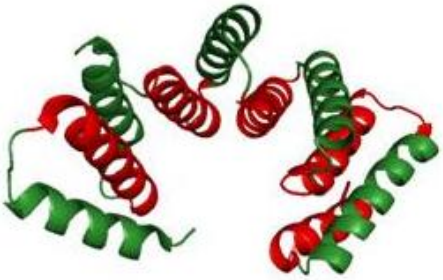
ARD2 identified an alkylpurine DNA glycosilase as alpha-solenoid; it has been previously described as being made of HEAT-repeats directly contacting DNA molecules via their sugar-phosphate backbone ([96]; PDB ID: 3JXY). Moreover, a bacterial DNA-binding protein involved in DNA repair mechanisms is also detected to display an alpha-solenoid ([97]; PDB ID:1XG7).

Alpha-solenoids are also capable of binding tRNA. ARD2 detects Xpot, a karyopherin involved in the export of tRNA from the nucleus to the cytoplasm ([98]; PDB ID: 3IBV/3ICQ). Xpot binds a tRNA by both 5'and 3' ends of the nucleotide sequence and the phospho-ribose backbone of the RNA molecule. The binding is therefore not specific and Xpot can bind all types of tRNA.

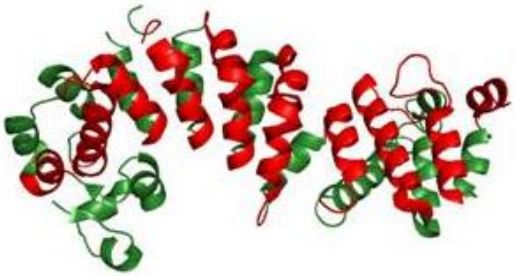
Alpha-solenoids can also contact micro-RNA. Exportin-5 is a large alpha-solenoid that transports pre-microRNA from the nucleus to the cytoplasm, and at the same time protects them from degradation by nucleases ([99]; PDB ID: 3A6P). For interaction to happen, a series of HEAT-repeats forming a very flexible ribbon wrap around the immature pre-microRNA in complex with RanGTP (see figure 7D). Similarly to the DNA glycosilase mentioned above, the solenoid only binds the backbone of the RNA molecule, which means that this probably happens independently of the RNA sequence.

ARD2 identified as alpha-solenoid a protein reported to bind messenger RNA. Differently from the previous examples, this alpha-solenoid, formed by Pumilio repeats, binds the nucleotides rather than the backbone of the RNA molecule ([100]; PDB ID: 3K62), and the binding itself is proved to depend on the presence of certain conserved nucleotides within the RNA molecule. The specificity of this protein to RNA interaction on the RNA sequence probably means that it has a regulatory rather than a pure transport function, this latter function being associated to the other types of alpha-solenoid that bind nucleic acids.

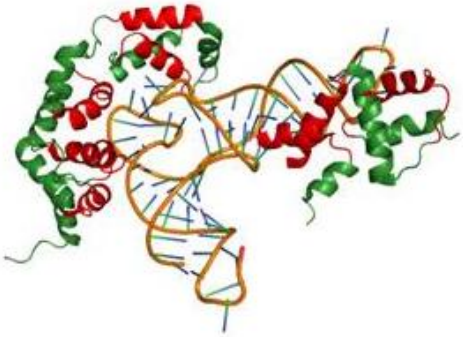
A



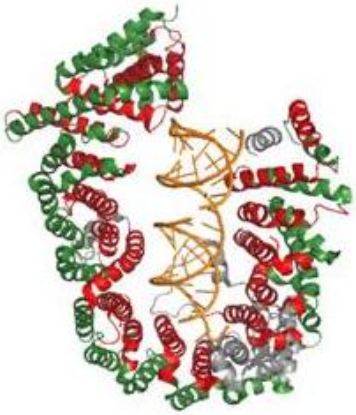
B



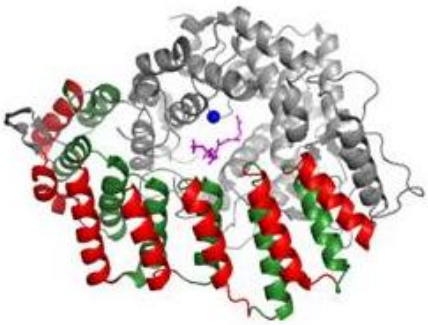
C



D



E



F

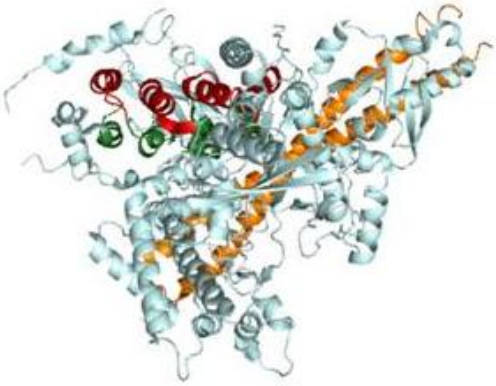


Figure 7. Examples of detected alpha-solenoid structures. Each repeat consists of two alpha-helices, depicted here in red and green. (A) TPR repeats protein, virulence regulator from *Bacillus thuringiensis* (PDB ID 2QFC [91]). (B) Ankyrin repeats protein Q5ZSV0 from *Legionella pneumophila* (PDB ID 2AJA [85]). (C) Irregular alpha-solenoid, glutamyl-tRNA synthetase from *Thermotoga maritima* (PDB ID 3AL0 [101]). (D) Alpha-solenoid binding RNA in exportin5 (PDB ID 3A6P [99]). (E) Lipid-binding protein. Isoprenoid lipid directly binding the HEAT repeats is colored in magenta, zinc atom in blue (PDB ID 3DRA [102]). (F) HEAT repeats buried in the core of the PI3KC catalytic subunit p110alpha (cyan), in complex with p85alpha (orange) (PDB ID 3HHM [103]).

Aside from binding to nucleic acids, we also found two alpha-solenoid structures that contact lipids. Lipovitellin is a lipoprotein that intervenes in the storage of lipids within the eggs of birds. Its shape is an open cone inside of which lipid is bound ([104]; PDB ID: 1LSH). This interface is detected by ARD2 to be partially made of alpha-solenoid repeats at positions described to be involved in lipid interaction. Another protein that we identified to bind lipids is the GGTase-I (geranylgeranyltransferase-I), which is known to catalyze the fusion of lipids on proteins ([102]; PDB ID: 3DRA; see figure 7E). Manual verification shows that the alpha-solenoid repeats bind the lipid. To our knowledge, these are the first structures showing direct contact of alpha-solenoid repeats to lipids.

2.3.3. Alpha-solenoids can be located outside as well as inside of proteins

As they are mostly involved in protein-protein interactions, alpha-solenoids are often found to be on the outer part of proteins; moreover, several alpha-solenoids are shaping entire proteins, including importins, beta-catenins and transportins [30,90,105]. In contrast, we found proteins with buried alpha-solenoid repeats. ARD2 detected the helical domain of p110alpha ([103]; PDB ID: 3HHM) as being an alpha-solenoid, though it is located inside the protein (see figure 7F, red and green part, the non alpha-solenoid region is in light blue). The domain is involved in the docking of p85alpha, one of several proteins that help PI3Kalpha to properly scaffold. To our knowledge, this is the only known case of a structure that presents an inner alpha-solenoid, along with a murine homologue (PDB ID: 2WXF).

2.4. Functions of alpha-solenoids

Similarly to sequences from the Protein Data Bank, we applied ARD2 to the sequences of the whole human proteome that are available in the Swiss-Prot database, in order to review the different functions they are involved in (human proteins comprise 20,328 sequences in the version 15.6 of Swiss-Prot (release date 28/07/2009). The prediction revealed a total of 99 alpha-solenoids, which represents 0.49% of the total proteome (see table 5 for a presentation of each sequence). Compared to ARD, this prediction increased the count of alpha-solenoids from 89 to 99 (see supplementary table 3 in [51], submitted). From data associated to the sequences, we

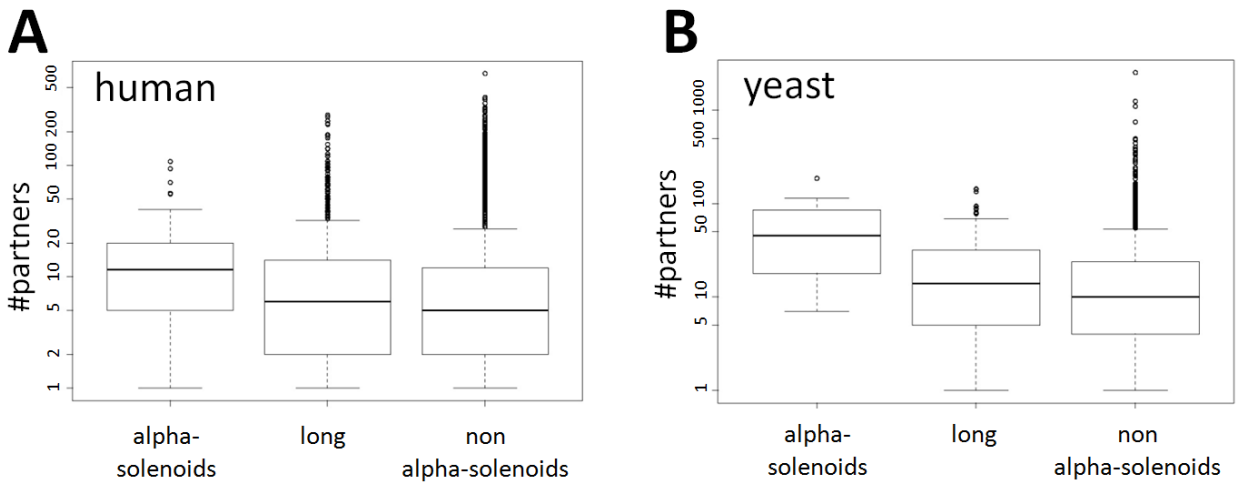


Figure 8. Distribution of number of interactions in alpha-solenoid and non alpha-solenoid proteins. The different box-plots display the distribution of the number of interactions for proteins with alpha-solenoids, non alpha-solenoid proteins with average length longer than alpha-solenoid proteins, and all non alpha-solenoid proteins. Boxes represent the values between the first and third quartile of the distributions. The horizontal line inside of the boxes indicates the median value. Whiskers indicate the standard deviation of the distribution. Circles indicate the outliers. (A) Human proteins. (B) Yeast proteins (*Saccharomyces cerevisiae*).

extracted information to better understand the functional characteristics of alpha-solenoids, and we briefly report this information in the sections 2.4.1 to 2.4.3.

2.4.1. Alpha-solenoid proteins are promiscuous

Alpha-solenoids are known for long to be involved in protein-protein interactions [34,54], and several structures of alpha-solenoids bound to other proteins have been published in the PDB in the last fifteen years. For instance, the alpha-solenoid of beta-catenin has been solved in complex with the Ran protein (PDB ID: 1IBR; see [106]); as well as the alpha-importin complex with the NLS (nuclear localization signal) of protein c-Myc (PDB ID: 1EE4; see [107]); Cand1 (TATA-binding-protein interacting protein) in complex with Cul1 (involved in the ubiquitination of proteins of the cell cycle) (PDB ID: 1U6G; see [108]); and the exportin CSE1P with karyopherin KAP60P and RanGTP (PDB ID: 1WA5; see [105]).

We decided to investigate whether human alpha-solenoids are actually contacting more proteins on average than other proteins. In the recent years, lots of data about protein-protein interactions have been accumulated, both from proteomic [109,110] and small-scale studies (for instance, the alpha-solenoid complexes cited in the previous paragraph). This information can be extracted from databases (for instance PDB), as well as from available literature as referenced in Pubmed. A database named HIPPIE has been recently developed by our group in order to collect and

evaluate such information on human PPIs [111,112]. We searched the HIPPIE database for protein-protein interactions involving the 99 human alpha-solenoid proteins detected by ARD2. We found that alpha-solenoid proteins have a significantly higher number of protein partners than average human proteins (Wilcoxon–Mann–Whitney test; p-value = 1.3E-6, see figure 8A). Nevertheless, alpha-solenoid proteins are on average longer than other human proteins (1,086 residues versus 553 residues for the human proteome), and their longer length could explain why they contact more proteins rather than the sole property of having an alpha-solenoid domain. Therefore, we also compared the number of interactions of alpha-solenoid proteins to that of non alpha-solenoid proteins with a length longer than the average of alpha-solenoids. The alpha-solenoid proteins had significantly more PPIs than this set of long proteins (p-value < 8E-4), which confirms that the alpha-solenoid domain is associated to an increased potential of proteins to contact partners. We also found similar results using a PPI network for the yeast *Saccharomyces cerevisiae* (retrieved from BioGrid; see [113]). In this case, we also found that alpha-solenoids have significantly more interactions than average proteins, and also more interactions than long proteins (p-values of 7E-9 and 5E-6, respectively; see figure 8B).

2.4.2. Alpha-solenoid proteins are primarily involved in intracellular trafficking

Though our findings prove that alpha-solenoid proteins display more protein partners than other proteins of the same length, this property is rather general and does not inform about the mechanisms in which these domains are involved. In order to get more information about the precise functions of alpha-solenoid proteins, we analyzed whether they are significantly associated with specific functions, described as gene ontology (GO) terms [114] (see summary in table 6). To perform such GO analyses, we used the DAVID tool [115].

We found the most significant functional enrichment for keyword “intracellular protein transport” (Benjamini-corrected p-value of 4.4E-23). This is due to the high frequency of importins (18%) and adaptins (13%) in the alpha-solenoid set; this result is similar to the trend observed in the PDB, with 20% of alpha-solenoids being involved in protein transport and somehow suggests that the trend is not just due to an experimental bias. Moreover, we know that transport characteristics of alpha-solenoids are not confined to proteins, but can also involve nucleic acids (tRNA or pre-miRNA), as we have shown in section 2.3.2. Now, one can ask, how old is the transport function of alpha-solenoid proteins? The involvement of alpha-solenoids in protein transport might date from ancestral unicellular organisms that reproduced by division, and needed to bring cellular material from the parent to the forming cell; accordingly, we have identified an alpha-solenoid protein in yeast (named She4, PDB ID: 3OPB) supposed to be involved in transport of proteins from the mother to the daughter cell [116].

Alpha-solenoids are also associated to the keyword “mitotic cell cycle” (corrected p-value of 1.5E-7), due to the presence of condensins and protein of the proteasome; to “Golgi vesicle transport” (8.8E-3), due to presence of adaptins and mTOR; and to “regulation of defense response to virus” (5.5E-2).

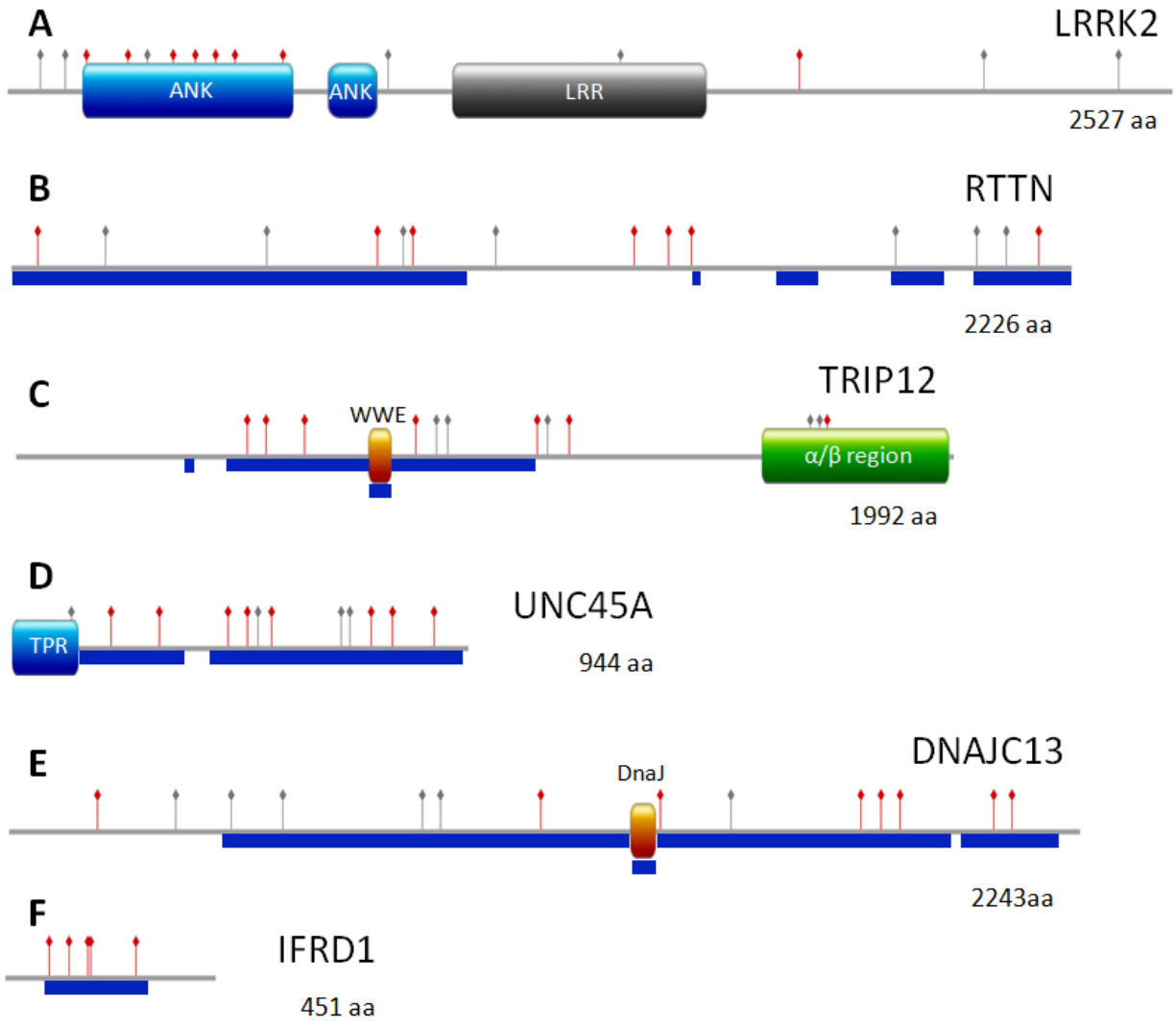


Figure 9. Domain organization of six predicted alpha-solenoid proteins. Alpha-solenoid repeat units predicted by ARD2 are displayed with red needles (score equal or greater than 0.87). Other scores above threshold 0.30 are represented with grey needles. For comparison, Armadillo regions predicted by InterPro are displayed as blue boxes. Other predicted domains are displayed with labels. (A) LRRK2, (B) RTTN (rotatin), (C) TRIP12, (D) UNC45A, (E) DNAJC13, and (F) IFRD1.

Analysis of keywords linked to cellular localization did not identify alpha-solenoids as specifically associated to the cytoplasm and nucleus. Differently, they are significantly enriched in keyword “nuclear pore” (1.2E-17), because transport proteins such as importins and exportins are present in both compartments and need to go through and bind the nuclear pore to transit between the two regions of the cell. “Coated membrane” was also found, due to presence of adaptins (1.3E-12), as well as “Golgi apparatus”.

Finally, we find that alpha-solenoids are more subject to alternative splicing than the average protein (corrected p-value: 7.9E-2).

2.4.3. Some proteins are newly detected as containing alpha-solenoids

We present here six examples of human proteins identified as alpha-solenoids for the first time: LRRK2, RTTN, TRIP12, UNC45, DNAJC13, and IFRD1 (table 7). All of these proteins follow the same trend as most alpha-solenoids: they display homologs exclusively in eukaryota, and are especially well-conserved within chordata. Our predictions are confirmed by InterPro, which annotates these sequences as ARM-like, though the positions predicted by InterPro do not always agree with predictions of ARD2 and overlap sometimes with domains known to adopt non-alpha-solenoid structures.

LRRK2 (figure 9A) is a serine/threonine protein kinase of 2527 residues, some mutations of which have been proved to be linked to parkinsonism [117]. ARD2 predicts region 360-494 to be an alpha-solenoid. The positions of the different repeats are consistent with the prediction of secondary structure for this region (obtained with Jpred3 [118], a neural network that predicts the secondary structure of a query protein sequence); ARD2 predicts repeats exactly in between alpha-helices (data not shown). More information has recently been added to the sequence [119]: a domain of ankyrin repeats (newly classified as alpha-solenoids, see section 2.3.1), a WD40 domain (beta-sheet repeats), and a leucine rich repeat (LRR) domain (solenoid made of alternating alpha and beta repeats). In total, this analysis predicted protein LRRK2 to contain 13 repeats between residues 13 to 657. ARD2 could predict 10 of these repeats when we included predictions with low scores (equal or higher than 0.10; see figure 9A). This shows that sub-optimal ARD2 predictions should always be examined; they might reveal important information but should be validated by other sources of information such as secondary structure prediction tools to verify their accuracy.

RTTN or rotatin (figure 9B) is a protein of 2226 residues involved in axial rotation and left-right specification of the body [120]. It interacts with the centrosome during mitosis and is also involved in cilia function [121]. ARD2 predicts significant hits in the region 1300-1450 (outputs of the network equal or higher than 0.87); but other hits were predicted outside of that zone, though mostly with suboptimal scores (positions 56, 198, 767, 842, 1853, 2023 and 2154 are predicted as repeats with scores equal or higher than 0.80). Moreover, InterPro predicts repeats in regions 1-954, 1602-1691, 1846-1956 and 2017-2225. These findings from ARD2 and InterPro combined suggest that the protein might be a giant alpha-solenoid, potentially forming a super helix, as proposed elsewhere for a different alpha-solenoid protein [122]. To support our hypothesis, we aligned the human sequence of rotatin to various homologues, including sequences from distant species such as the green algae *Chlamydomonas reinhardtii*, and found that the positions predicted as repeats by ARD2 are mostly highly conserved (see figure 10), adding more evidence of the reality of their presence. Residue conservation is usually associated to functional and structural signals [123].

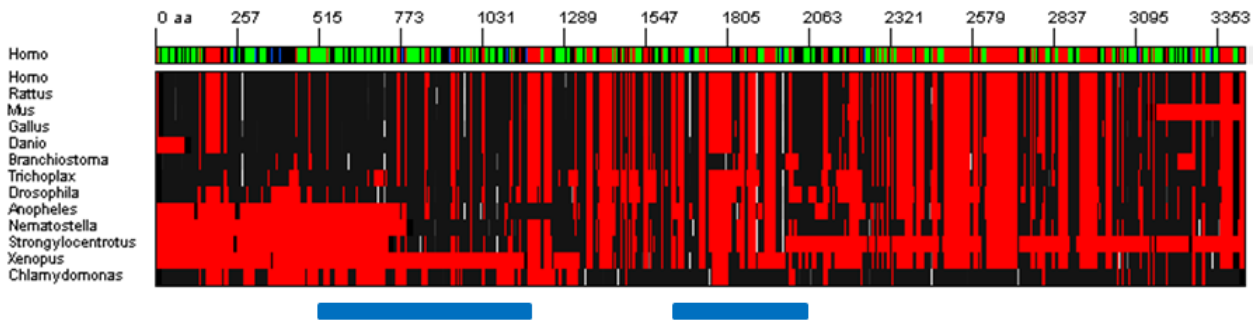


Figure 10. Alignment of rotatin homologs. A multiple sequence alignment of human rotatin and homologs in other species was produced and represented using BiasViz [124]. Top lane: Jpred3 2D prediction for human rotatin (red: gaps, green: alpha-helix, blue: beta-strand). Bottom part: multiple sequence alignment (red: gaps, black to white: score of ARD2 prediction from 0 to 1). Most of the secondary structure prediction is alpha-helical. Clusters of periodic alpha-solenoid hits can be seen at the positions indicated by the blue bars. Other scattered hits are distributed through the entire alignment.

TRIP12 (figure 9C) is a large E3 ubiquitin-protein ligase of 1992 residues involved in the regulation of the cellular response to DNA damage [125]. A former work presenting the analysis of its sequence proposed that the molecule is made of two HEAT-repeat domains interspersed by an ADP-ribose binding module termed WWE at residues 749 to 798 [126]. Hits with outputs higher or equal than threshold 0.87 were predicted by ARD2 on both sides, but not within the WWE domain, whose secondary structure is mostly beta-strand [127]. On the opposite, prediction from InterPro found a continuous domain of HEAT-repeats, failing to predict the WWE region. This result shows that ARD2 can be very relevant to define the precise limit of an alpha-solenoid domain.

UNC45 (figure 9D) is a co-chaperon of Hsp90 involved in the correct folding of myosin during development [128]; its ortholog in *Drosophila* is a key protein for the development and function of the heart [129]. Predictions of ARD2 and InterPro show two alpha-solenoid regions (residues 138-350 and 403-932 with 2 and 6 significant scores predicted by ARD2, respectively); moreover, InterPro predicts that the N-terminal region of the protein displays tetratricopeptide repeats (TPR) (residues 1-135; PDB:2DBA; unpublished, RIKEN structural genomics initiative). We also studied the paralog UNC45B, of close sequence, which gave a similar result (data not shown).

DNAJC13 (or RME8, figure 9E) is a co-chaperon of Hsc70 of 2243 residues. It is involved in receptor-mediated endocytosis [130]. Similarly as for TRIP12, the protein displays a DnaJ or J domain (non alpha-solenoid) at residues 1301 to 1354, surrounded by two alpha-solenoids, which cover most of the protein. ARD2 appropriately predicts optimal or suboptimal alpha-solenoid repeats all over the protein, except in the DnaJ domain, while InterPro falsely predicts

the two alpha-solenoids as one single alpha-solenoid domain and misses the information from the DnaJ domain. This shows again the precision of ARD2 at correctly showing the boundaries of alpha-solenoids, versus tools such as InterPro, which can potentially detect a wide range of different alpha-solenoids, but with less accuracy on the precise location of repeats.

IFRD1 (figure 9F) is an interferon-related protein that is possibly involved in regulation of gene transcription in pathways involving nerve growth factor. It is found to be involved in muscle development and in cystic fibrosis lung disease [131]. ARD2 prediction gives five optimal hits in the central region of the protein, confirmed by InterPro prediction. Prediction of the secondary structures of the N- and C-terminal regions of the protein by Jpred3 confirms this finding, with the first alpha-helix appearing at residue 76 (the middle point of the first repeat of the alpha-solenoid is predicted to be located at position 84 by ARD2); another repeat though might have been missed both by ARD2 and InterPro around position 393; Jpred3 predicts this position to be surrounded by two alpha-helices that could together form another repeat. We finally performed the same analyses on a human paralog of IFRD1, IFRD2, and found similar results, confirming the central position of the alpha-solenoid in this protein.

2.5. Distribution of alpha-solenoid proteins across the tree of life

For long, alpha-solenoids have been shown to have little conservation between each other, and despite this low conservation, we wondered if this type of structure could be detected across the tree of life using ARD2. To achieve this, a survey on all the proteins available in UniProt (about 22 millions of sequences; release 2012_05) was performed. 18,910 proteins were detected to contain alpha-solenoid repeats. We calculated the percentages of alpha-solenoids present in total protein sets of species belonging to 31 taxonomic divisions (figure 11).

Firstly, one can observe that there is a broad range of values between the major taxa. The lowest value is found for viruses with a frequency of 1.16E-5 (for only 16 alpha-solenoids), bacteria come next with a frequency of 2.72E-4, then archaea (8.17E-4), and the highest value is associated to eukaryota (2.57E-3). Eukaryota appears to be the most homogenous group, with values for its different taxa above 2.0E-3. Conversely, bacteria show a much important diversity of frequencies, with groups of very high frequencies while others are quite low. Cyanobacteria and planctomycetes were found to show the greatest values of all bacteria, 1.97E-3 and 2.27E-3 respectively, which is similar to the average value of eukaryota.

We then analysed the origin of alpha-solenoids in these two taxa. UniProt annotations of alpha-solenoid proteins found in cyanobacteria were found to be annotated as HEAT PBS domain. Phycobilisome (or PBS) is a complex of molecules involved in the harvest of light. This domain has been previously annotated as alpha-solenoid (identified as HEAT PBS in the Pfam database), and ARD2 could detect it, though no HEAT PBS was included in the training set (see table 2). In the cyanobacteria species *Nostoc punctiforme*, for instance, 15 out of 17 alpha-solenoids detected by ARD2 are annotated as HEAT PBS in UniProt. We could find HEAT PBS

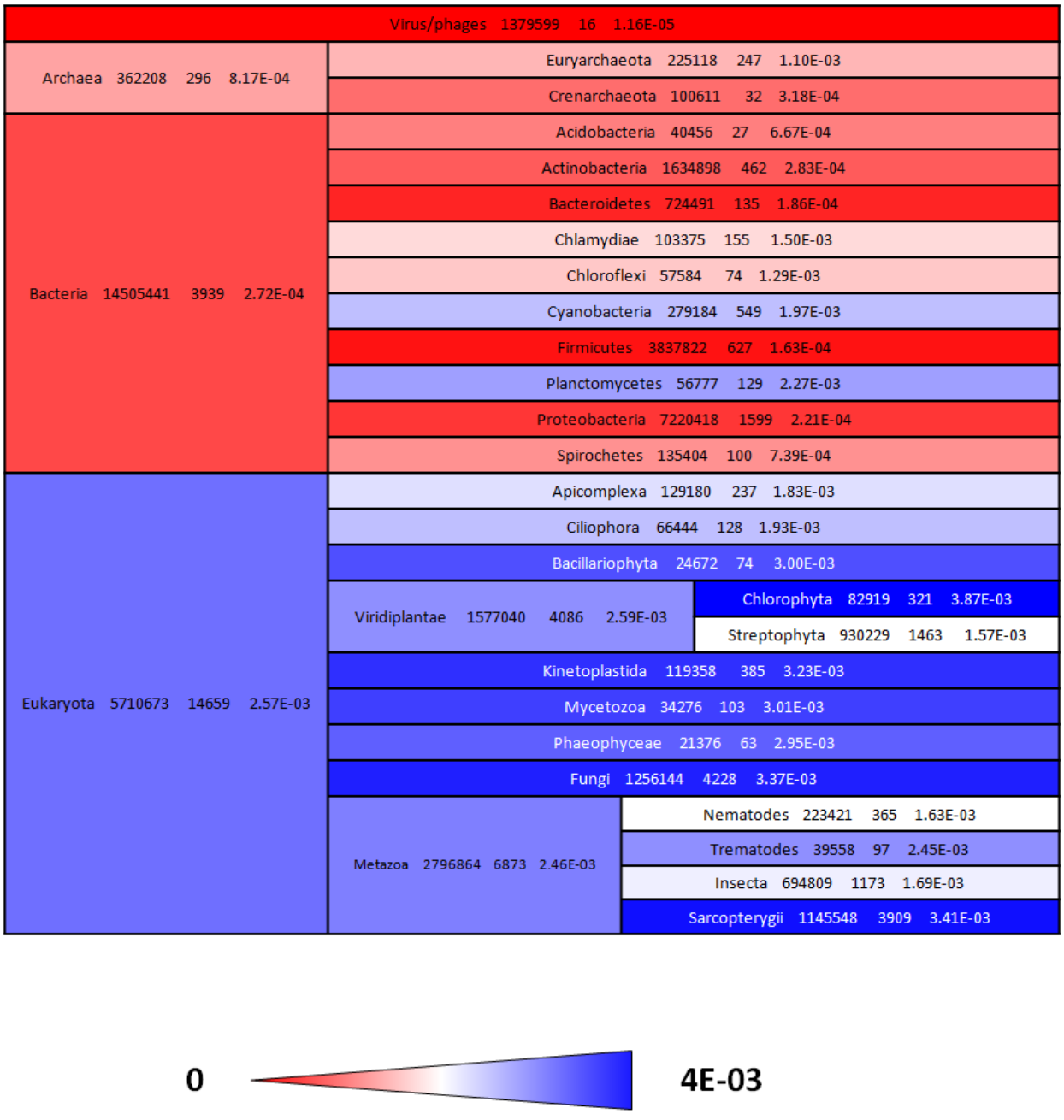


Figure 11. Alpha-solenoids in complete genomes. Fraction of alpha-solenoids in proteins from 31 taxonomic divisions. The blue/red scale indicates the percentage of alpha-solenoids present in the different taxa. Groups depicted in blue scale were the most enriched in alpha-solenoids, while the ones in red were poor in alpha-solenoids.

sequences in the tree major cyanobacterian groups (chroococcales, nostocales and stigonematales), which means that the domain was likely present in the common ancestor of the species of these three groups. Emergence of this domain might therefore have occurred 3.5 billion years ago, the probable time of appearance of this group of bacteria. Cyanobacterian sequences were checked for ontology enrichment using the DAVID tool [115] and no keyword was found apart from "Energy production and conversion" (p-value corrected: 6.00E-14) and HEAT PBS (2.60E-35), which are two close keywords.

Differently, planctomycetes alpha-solenoids showed more diversity of protein functions than cyanobacteria. To be able to understand the origin of these proteins, we classified the sequences into different families. A family was defined as a group of proteins whose members are all at least homologous to another member of the group with a p-value associated equal or lower than 1E-8. BLAST was used to check for homology. The program Pisces [132] was later used to check for consistency of the BLAST results. We found 21 families of proteins in *Blastopirellula marina* DSM 3645. The largest of these families comprised 9 sequences, but its function is uncharacterized. In another species, *Rhodopirellula baltica*, we observed HEAT PBS proteins (though planctomycetes do not harvest light, contrary to cyanobacteria) in 2 of the 13 alpha-solenoid proteins detected. To get information about the function of these alpha-solenoids, we retrieved information about them in UniProt. Among the 129 alpha-solenoid sequences in planctomycetes, 28 (21.7%) showed the keyword "dehydrogenase", 13 (10%) the keyword "heme", 15 (11.6%) the keyword "PBS lyase", and 6 (5.6%), the keyword "glucose" or "cytochrome c". It is likely that the proteins supposedly involved in light harvesting originate from gene duplication, and were later used for a different function in the planctomycetes lineage. Following the presence of PBS lyases in taxa as diverse as cyanobacteria, which harvest light, and planctomycetes, which harvest sulfate, we speculate that alpha-solenoid proteins found in these two groups are possibly related to pathways involved in energy production. This involvement in energy production has long been proven to be one of the functions of alpha-solenoids, notably in dinoflagellates, a group of protists. In this group, alpha solenoid proteins are subunits of the chlorophyll proteins [133].

We found that the majority of sequences identified in archaea are homologous to bacterial sequences, including the proteins containing a HEAT PBS domain. We found only 90 (30%) specific archaean families (research performed using BLAST on each archaean sequence against all bacterial sequences available in UniProt). For instance, we found that *Methanoculleus marisnigri*, a species found in anaerobic digestors and aquatic sediments, presents 10 alpha-solenoids in its proteome, 9 of which are annotated as PBS domains. In *Halobacterium salinarum*, a halophilic marine eukaryota, protein OE2401F, displaying a PBS domain, is found to be associated to flagella [134]. We found its closest homologue in prokaryota to belong to cyanobacteria (the two homologous sequences match by their entire sequences and have 28% identical amino-acids), which suggest that some HEAT PBS proteins of cyanobacteria might be involved in motility.

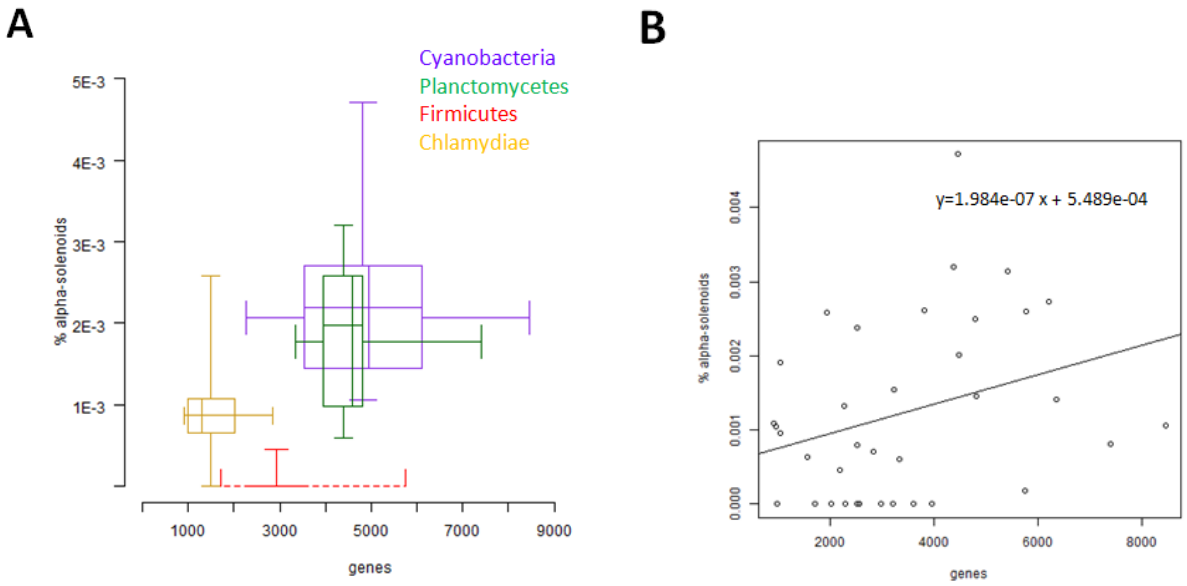


Figure 12. Percentage of alpha-solenoids versus number of genes. A. Two-dimensional box plot of percentage of alpha-solenoids against genome size averaged for several representative species with completely sequenced genomes from four bacterial groups: cyanobacteria, planctomycetes, firmicutes and chlamydiae. Each box shows the distribution of one of these four groups and summarizes two distributions: the percentage of alpha-solenoids associated to the genome of species of that group in the vertical direction, and the size of the genomes of the species of that group in horizontal direction. In each direction, the box is limited by first and third quartile of the distributions. The middle line (horizontal or vertical) inside of the boxes indicates the median value and the whiskers indicate standard deviation. B. Correlation between the number of genes and the percentage of alpha-solenoids of genomes of the same four groups of bacteria. Non-parametric Spearman coefficient ρ equals to 0.3634681.

Cyanobacteria and planctomycetes possess a higher morphological complexity [135] and their own cell compartments [68,136,137]. These two taxa have been shown to display a fairly large genome size compared to other bacterial groups [138]. We plotted the distribution of frequency of alpha-solenoid proteins in genomes against the distribution of the size of genomes for four different bacterial groups, including cyanobacteria and planctomycetes (figure 12A). We found that the genome sizes of these two groups are larger than the ones of firmicutes and chlamydiae, and that their larger genome sizes are associated to more frequent occurrences of alpha-solenoid proteins. Pooling data from the species of these four groups altogether, we found a tendency showing that the frequency of alpha-solenoids increases with the size of the genome (non-parametric Spearman coefficient ρ=0.36; figure 12B).

Finally, we could find only 16 sequences from virus with alpha-solenoids. We detected homologues (at least 60% of homology on 75% of their length) to their hosts in 7 of them. These hosts were either *Chlorella* or *Streptococcus*. Among the 9 other proteins, some belong to human virus of hepatitis C. These sequences might be as well the result of a horizontal gene transfer

from their hosts. For instance, viral alpha-solenoid protein G8DER4 from *Phaeocystis globosa* virus 14T is found to have sequence similarity to translation elongation factor 3.

2.6. Modeling of an alpha-solenoid region of protein huntingtin

Here, we apply ARD2 to the sequence of protein huntingtin in order to build a model of its first HEAT repeats region, which we hypothesize to have potential specialized functionality. We later use this model in order to predict the outcome of mutations on huntingtin's structure and PPIs. These results could later be used for testing and improve the understanding of the structural mechanisms that underlie huntingtin's biology and binding to its protein partners.

2.6.1. Introduction

Huntington's disease is the 12th most common genetic disease in human, with a prevalence of 5-10 cases per 100,000 [139]. The pathology originates from a mutation on a protein called huntingtin. In patients affected by the disease, huntingtin shows an expansion of a polyglutamine (polyQ) repeated motif located in the N-terminal region. The pathology is the consequence of the aggregation of expanded polyQ huntingtin fragments in the cytoplasm of certain types of neuron located in the caudate nucleus, a region of the brain involved in coordination of information coming from the cortex to other parts of the brain [62]. In the long term, this aggregation causes neurons necrosis. As these neurons are involved in coordination of thoughts and movements, people bearing the disease show motor impairment and degradation of their mental faculties [62].

Though a lot of information has been published on the behavior of mutated huntingtin, some suggest that focus should be rather put on investigating its wild-type function rather than solely focusing on understanding the mechanisms of aggregation [62,140]. What do we currently know about the function of huntingtin? Numerous studies have been published, where the protein was shown to be involved in transcription [141], vesicular transport [142], and, in a more recent study, in the formation of mitotic spindles [143].

As no clear pathway is associated to huntingtin, and the protein is big, probably causing it to be involved in many protein-protein interactions involved in various cross-talking pathways, studies have attempted to study the biology of huntingtin using "systems biology" approaches. In this case, proteins are not only considered as individually participating to function, but rather working synergistically in a PPI network [144]. Such studies led to the conclusion that huntingtin interacts with a large group of protein partners [109,145].

Moreover, huntingtin is well-known for containing 3 HEAT repeats regions [36,55], domains that belong to the family of alpha-solenoids [31]. These alpha-solenoid regions are important because they help the protein to fold correctly, by making intra-molecular interactions with the other alpha-solenoid regions of the same molecule [55]. Moreover we have shown in our work (in section 2.4.1, see also figure 8) that alpha-solenoids have a tendency to be promiscuous (see

also [51]), so we hypothesize that the alpha-solenoids of huntingtin are potentially good candidate regions for protein-protein interaction.

As a consequence, we believe that putting effort in understanding the biology of these alpha-solenoid regions could contribute to understand the function of huntingtin and be useful in the description of the mechanism of Huntington's disease. In order to achieve this goal, we used three approaches.

- From Y2H data (Wanker group, MDC Berlin-Buch, unpublished), we find that the region of huntingtin comprising its first alpha-solenoid shows functional signal and is associated with partners participating in protein biosynthesis; this confirms our hypothesis that alpha-solenoids of huntingtin have biological importance.

- Secondly, we modeled the structure of this alpha-solenoid region. Huntingtin's major characteristic is its giant size (3144 residues), which partly explains the instability of its structure. Huntingtin comprises several unfolded regions, which confer high flexibility; as a result, huntingtin shows large variations of shape [146]. As the protein is huge, it is hard to crystallize and therefore no high-resolution structure is available so far apart form a small fragment encoded by exon1 of the gene (N-terminal amino acids 1 to 71). Here we propose a model for huntingtin, designed using a threading algorithm and then confirmed by ARD2 prediction of alpha-solenoid repeats positions.

- Finally, we tried to locate the functional residues on our alpha-solenoid model, in other terms, the residues that are likely to be involved in protein-protein interactions. These residues display certain properties that differ from those of internal amino-acids, which are mostly involved in the correct shaping and cohesion of the protein [147]. Experimental verification of the involvement of these residues in protein interaction would be done by mutating them and observing a disruption of a protein interaction. To ensure that this is not due to a general alteration of the protein's structure, we will seek exposed residues whose mutation is not predicted to disrupt the structure of the protein.

2.6.2. Methods

Y2H data analyses

Y2H (Yeast-two-hybrid) is a technique used to test protein-protein interactions on a large scale, and has been used in the determination of a PPI network for huntingtin [109]. To test if two proteins interact (one being called "bait" and the other "prey"), each of them is expressed in yeast associated to one half of a sequence of the same transcription factor. If interaction between the two proteins happens within the yeast cell, the two portions of the transcription factor are brought into vicinity and transcription of the reporter gene under control of this factor can potentially happen and be detected; this reporter gene can be lacZ for β-galactosidase assays. In the case of huntingtin, one of the two proteins is a fragment of huntingtin itself and the other one is the protein to test. As huntingtin is a huge protein, one usually only expresses an N-terminal fragment of 500 residues as bait [109] (Figure 13). Nevertheless, different fragments of huntingtin have also been utilized for testing against each other in order to show intra-molecular interactions happening between different regions of the protein [55].

Y2H data from the group of Erich Wanker (Max-Delbrück Center, Berlin) (data not shown, unpublished) were studied. Fragments used as baits included exon1 of huntingtin with a polyQ of 23 residues (wild-type exon1), exon1 of huntingtin with expanded polyQ, fragment of wild-type huntingtin comprising residues 1 to 506 (Q23), fragment 1-548 with no polyQ tract (or Q0), longer fragments with expanded polyQ (1-506-Q80, 1-548-Q80, 1-513-Q49 &1-513-Q68), a fragment comprising residues 2563 to 3144 and one comprising residues 2721 to 3144. The whole human proteome constituted the set of preys. Experiments to test interaction in the yeast system were performed by the group of Erich Wanker as described elsewhere [110]. Gene ontology (GO) enrichment of genes interacting with the different regions of huntingtin was tested using DAVID web tool [115]. Annotations given by DAVID coming from various sources were scrutinized; among others, we checked for significant annotations given by GO biological process, GO molecular function, PIR superfamily name, Pfam, InterPro, and SMART. A keyword was selected as significant if the p-value with Benjamini correction was equal or lower than 5e-2.

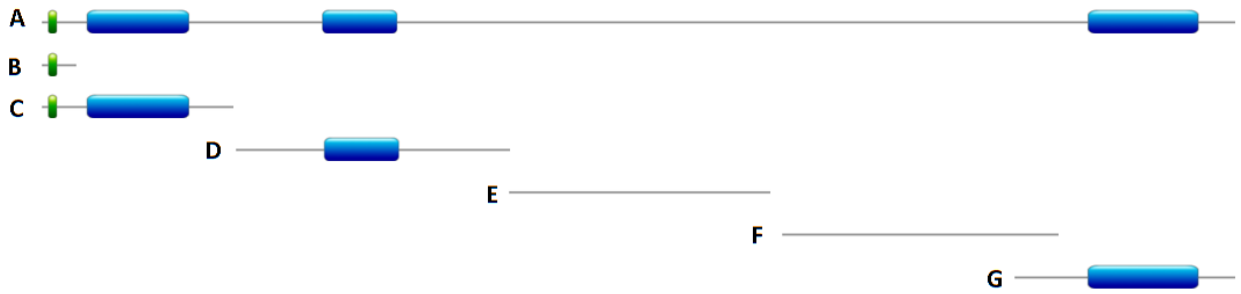


Figure 13. Diagram representing huntingtin and several fragments of the protein used in Y2H experiments. A. Full-length wild-type huntingtin. The polyQ tract is depicted in green (residues 18 to 40), the three alpha-solenoid regions in blue, as predicted by ARD2 (119-387, 740-937 & 2755-3046). **B.** Wild-type exon1 (first 91 residues of huntingtin, with a polyQ of Q23). **C.** Wild-type fragment comprising residues 1 to 506. **D.** Fragment comprising residues 507 to 1230. **E.** Fragment comprising residues 1223 to 1941. **F.** Fragment comprising residues 1934 to 2666. **G.** Fragment comprising residues 2563 to 3144.

Protein structure modeling

The 3D model of huntingtin was obtained using I-TASSER, a prediction tool that uses threading to predict structure from protein sequences [148]. Jpred3 [118] was used to confirm the reliability of secondary structure predictions of I-TASSER. ARD2 [51] was used to check the accuracy of the positions of the alpha-solenoid repeats predicted by I-TASSER, with a precision of one amino-acid.

Prediction of interacting regions

We used a tool called Promate [149] to locate functional regions on the predicted model. This tool predicts which regions of a protein are more likely to be involved in protein-protein interactions according to the stability of the atoms and surrounding water molecules. Stability of atoms is associated to a value named b-factor, which measures the movements of atoms compared to their average position. The more stable the atoms are, the more likely they are involved in protein-protein interactions [150]. As a result, binding sites show a higher b-factor, and this is already true while the protein is in unbound state [149]. Secondly, water molecules are found to be present in higher quantity in the surroundings of putative binding residues [149]. Measurements of these two parameters, the b-factor and the number of molecules of water, for a given residue gives a score reflecting its potential to be involved in protein-protein interactions. Here, we considered as interacting sites residues with a score higher or equal to 70 [149].

Prediction of mutation outcome

To predict the outcome of protein mutations, various tools were used. MutationAssessor is based on multiple sequence alignments (MSA) of homologous proteins, and assumes that the more conserved a residue is, the less likely it would induce a change in the protein structure [5]. Polyphen-2 uses the same method, but rather than using alignments of sequences from different species, its MSA uses alignments of sequences from human patients, which only differ from one another by the eventual presence of SNPs [48]. Differently, SDM assumes that the mutations more likely to be disruptive are the ones which change the thermodynamic stability of the whole protein [49].

Protein visualization

In order to check the agreement between modeling of the first alpha-solenoid region and the annotations by ARD2, we used PDBpaint, a tool to display protein models with annotations from several sources, including ARD2 (see section 4.2 for a detailed description of PDBpaint). PyMOL (PyMOL Molecular Graphics System software, DeLano Scientific, Palo Alto, California) was used to display our model and annotate it according to the b-factor. This enabled us to locate functional regions potentially involved in protein-protein interactions.

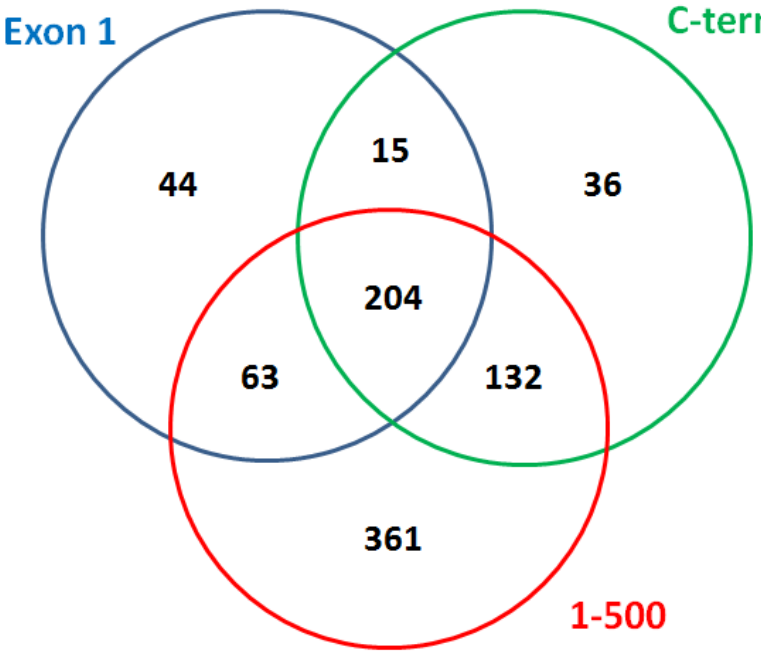
2.6.3. Results and discussion

The first alpha-solenoid region of huntingtin has functional importance

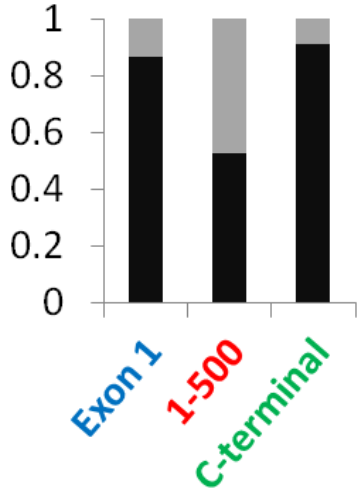
Here, we examined Y2H data (Wanker group, Max Delbrueck Center, Berlin, unpublished) of protein interactions involving diverse huntingtin fragments as baits and target proteins as preys (see section 2.6.2 for the list of huntingtin fragments studied). We extracted different kinds of information that are presented here.

The first alpha-solenoid of huntingtin shows a higher number of partners than exon1 and other regions; it also displays the highest number of unique gene partners. We summed up the results of the Y2H experiments in a Venn diagram showing the number of gene partners associated to fragments comprising exon1, fragments comprising residues 1 to 500 approximately, and the C-terminal fragments (figure 14A). We also displayed the percentage of gene partners for each of these groups (figure 14B). The region with the most unique protein partners is the one comprising residues 1 to 500 with a percentage of 47.5%; as these genes are not found to bind any exon1 fragment, we hypothesized that these interactors might be specific to the alpha-solenoid region. This high percentage of interactions compared to the one of exon1 (13.5% only) cannot be explained only by the fact that 1-500 fragments are longer than exon1 (which only displays 91 residues); for example, the C-terminal fragments have lengths of about 700 residues (Figure 13 D-G) and have much fewer unique partners (9.3%).

A


Exon 1
C-terminal
1-500
44
15
36
204
63
132
361

B


1
0.8
0.6
0.4
0.2
0
Exon 1
1-500
C-terminal

C

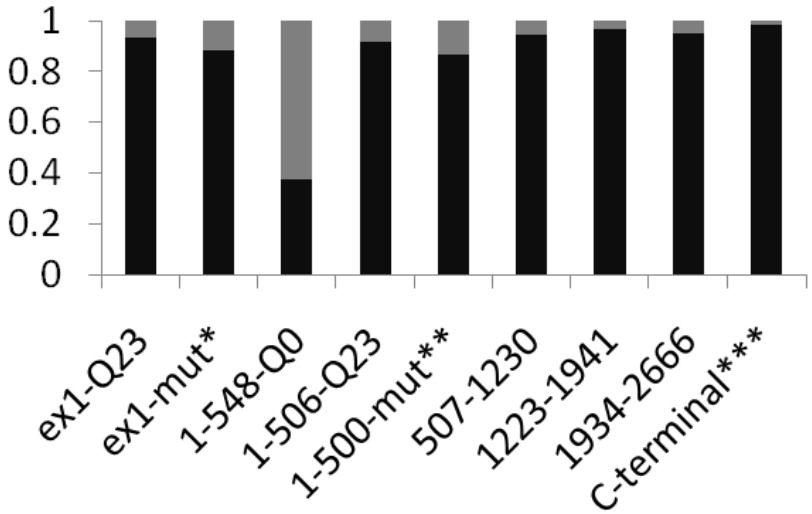

1
0.8
0.6
0.4
0.2
0
ex1-Q23
ex1-mut*
1-548-Q0
1-506-Q23
1-500-mut**
507-1230
1223-1941
1934-2666
C-terminal***

Figure 14. Distribution of huntingtin interactors in the different regions. A. Number of interactions per region. Proteins in the exon1 (blue circle) interact with huntingtin fragments comprising portions of exon1 only. Proteins in the 1-500 circle (red) interact with long huntingtin fragments comprising portions of huntingtin between residues 1 to 500. Proteins in the C-terminal circle (green) interact with huntingtin fragments located in the C-terminal part (fragments D to G on figure 14). Overlapping regions (e.g. the 63 sequences in common between Exon1 and 1-500 fragments) indicate proteins that bound two different types of fragments (in our example, the 63 proteins bind both at least one fragment of huntingtin in exon1 and one fragment comprising region 1-500). **B.** Percentage of non-specific interactors for each of the 3 groups of huntingtin fragments displayed in A, namely fragments comprising the exon1 only, fragments comprising the first 500 residues and the fragments in the C-terminal region of the protein. **C.** Percentage of non-specific interactors per fragments as displayed in figure 13. For each fragment of huntingtin tested in the two-hybrid screening (horizontal scale), the relative quantity of unique interactors (only interacting with that fragment, in red) and non-specific interactors (in blue) is displayed in vertical bars. * mut for mutated (i.e. elongation of the polyQ sequence); ** pooling of 4 fragments of mutated huntingtin (i.e. elongated poly Q): 506-Q80, 548-Q80, 513-Q49 & 513-Q68; *** pooling of two fragments: 2563-3144 & 2721-3144.

Moreover, we explored in more detail the different fragments tested one by one (see section 2.6.2 above on methods for more details), and found that the highest percentage of unique gene partners (62.7%) is associated to a fragment comprising the first alpha-solenoid of huntingtin, without the polyQ tract (fragment 1-548-Q0, see figure 14C). Other fragments display an average of 6.2% of unique interactors; the fragment with the greatest amount of unique interactors after fragment 1-548-Q0 shows a percentage of 13.8% only (1-500 fragments with expanded polyQ). Surprisingly, only 8.45% of interactors of fragment 1-506-Q23 (wild-type) are found to be unique, a number that we would have expected to be much higher, closer to the one found for the Q0 fragment. Given that they differ in the polyQ, this suggests that the polyQ has great importance as a modulator of PPIs in huntingtin. We demonstrate that there is a functional signal associated to the first alpha-solenoid of huntingtin lacking the polyQ in the next paragraph.

Finally, we hypothesized that analyzing the part of the huntingtin PPI network that interact with the Q0 fragment would help to learn which functions the first alpha-solenoid region might be involved in. To do so, we analyzed the Gene Ontology terms associated to its unique interactors, using the DAVID tool (see section 2.6.2 for details). Significant enrichment was found for keywords “ribonucleoprotein” (corrected p-value of 9.7E-3) and “nucleotide-binding” (corrected p-value of 4.7E-2), between others (table 8). Most of the enriched terms are related to the translational process. Proteins associated to the keyword “ribonucleoprotein” almost all belong to the subunits 40S or 60S of the ribosome or are involved in the pre-mRNA processing. Conversely, we found no significant GO enrichment in the unique interactors of the Q23 1-500 fragment (data not shown), but this could be due to the fact that much fewer interactors contact this fragment.

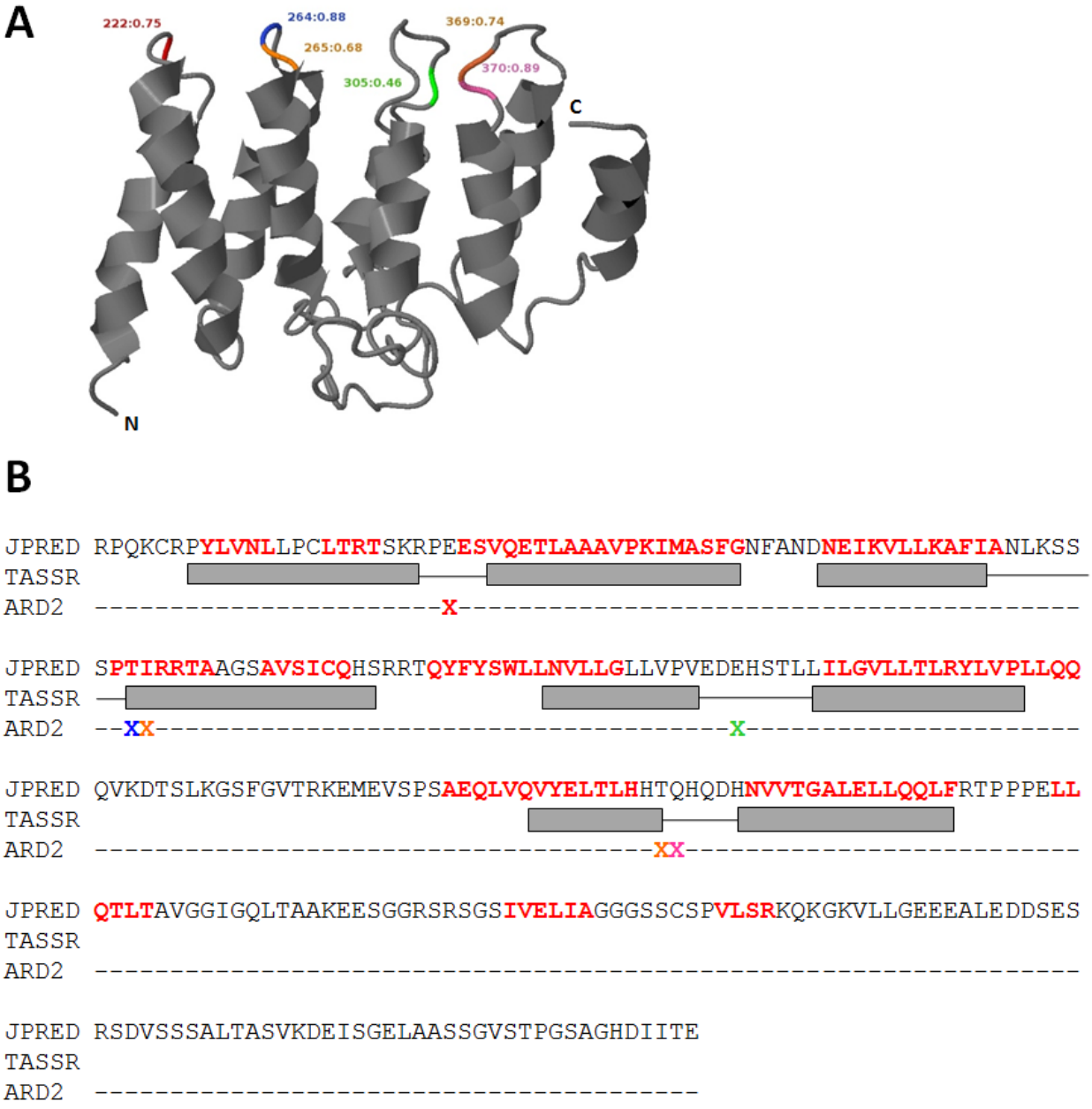


Figure 15. Huntingtin model as predicted by I-TASSER server and confirmed by ARD2 annotations. A. 3D representation of the first alpha-solenoid of huntingtin (residues 200 to 500), displayed using PDBpaint (see section 4.2 for more details about this tool). Annotations from ARD2 are depicted in colors with the residue position and the score of the neural network (on a scale from 0 to 1). **B.** Sequence of huntingtin annotated with Jpred3, I-TASSER and ARD2. In red, part of the sequence predicted to be alpha-helical according to Jpred3. Grey boxes are the alpha-helices of the alpha-solenoid predicted by I-TASSER, linked by lines figuring the coils in the middle of each repeat. Crosses, coloured as in the part A of the figure, show ARD2 predictions of repeats.

No significant GO enrichment was detected for specific interactors of other huntingtin fragments, except for those interacting with exon1 with expanded polyQ, where the term "coiled coil" was associated to 13 genes among 25 (corrected p-value of 5.1E-3). This finding agrees to the proposal that polyQ act as a modulator of the interaction between coiled coil proteins [140].

We conclude that the function of the first alpha-solenoid of huntingtin is probably specific to this region in particular, as no significant functional enrichment was observed for the interactors of fragments 507-1230 and 2563-3144 of huntingtin, which also comprise alpha-solenoids, both predicted by ARD2 (at positions 740 to 937 and 2755 to 3046, respectively) in accordance with Jpred3 prediction (data not shown).

A model of the first alpha-solenoid region of huntingtin

Aware of the potential importance of the region, we have elaborated a 3D model of the first 500 residues of the protein based on the prediction by I-TASSER server and ARD2. To verify the accuracy of the prediction by I-TASSER we used PDBpaint [50] to superimpose ARD2 predictions on the PDB structure predicted by the threading, and found that the predictions of the two methods match (figure 15A). Information from homology-based secondary prediction tool Jpred3, from I-TASSER and from ARD2 converge to the exact same four positions in the 100-430 region (222, 264/265, 305, 369/370; see figure 15B). Conservation of huntingtin sequences from human to *C. elegans* reveals that three of these alpha-solenoid repeats are very conserved, including ARD2 hits found in between motives of two alpha-helices predicted by Jpred3 (figure 16), adding more evidence to validate our model.

Specific residues of the alpha-solenoid model are predicted to be involved in protein-protein interactions

In order to locate functional residues on our huntingtin model, we identified regions with a high probability of interaction and amenable to experimental verification by mutation. This was achieved in two steps: we firstly selected candidate residues for interaction, from computational prediction and manual curation; in a second step we kept only the residues whose mutation does not disturb the protein structure, as it would be expected that disruption of the interaction would be tested by mutating those residues and therefore it is advantageous that such mutation do not disrupt the structure of the protein. Our procedure is summarized in figure 17.

- Prediction of interacting residues. We used the Promate tool [149] on our model of huntingtin in order to rank amino-acids according to their propensity to form protein interactions (see section 2.6.2 on methods for details). We selected the top residues (whose Promate score is higher than 70) and found a first set of 19 residues. Then we added residues specifically predicted by I-TASSER as involved in protein interactions, as well as some candidates that we selected manually, which had a Promate score below 70. These latter were selected as follows. We selected residues located in the outer part of the molecule, in the vicinity of the ones predicted to have the Promate top-scores. If a residue, though pointing outside, was also found to

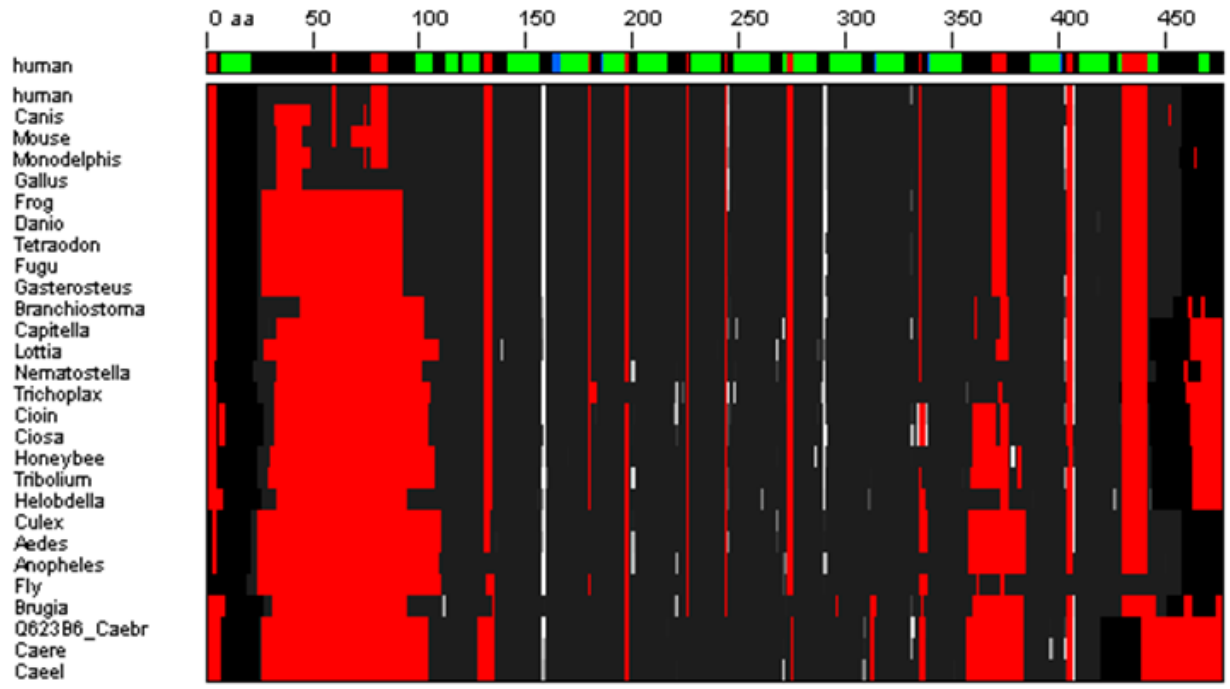


Figure 16. Visualization of an alignment of huntingtin sequences. Top lane: Jpred3 2D prediction for human huntingtin (red: gaps, green: alpha-helix, blue: beta-strand). Bottom part: multiple sequence alignment (red: gaps, black to white: score of ARD2 prediction from 0 to 1). Most of the secondary structure prediction for region 100-350 is alpha-helical. Periodicity of alpha-solenoid repeats matches the predictions of alpha-helices, conserved repeats around positions 160, 240 and 280 (which are positions of the alignment, not of the human sequence) being found between two alpha-helices. Visual output was generated using web tool BiasViz 2 [124].

interact with another one, it was discarded from the list. We added finally some interacting residues that were found by modeling the interaction of huntingtin with the HAP1 protein using the RosettaDock server ([151], data not shown). HAP1 was chosen as it is the interactor of huntingtin with the best score that could be found in the HIPPIE database ([112], see also section 2.4.1. for more explanation about this method). In the end, a list of 45 potentially good candidates for mutation was established in the first alpha-solenoid of huntingtin.

- Elimination of residues important for the stability of the whole protein. In a second step, locations of interest were studied for sensibility to mutations. We used different sources to predict the outcome of mutations of each of these 45 residues in terms of modification of the stability of the protein. To do so, we used three algorithms: Polyphen-2, SDM and MutationAssessor (see section 2.6.2 on methods for more details). Polyphen-2 predicts the outcome of mutations based on existing SNPs in the human population, assuming that varying residues are less likely to have functional importance. Nevertheless, it was found to be overly pessimistic, predicting almost all mutations tested to be harmful, probably due to the fact that

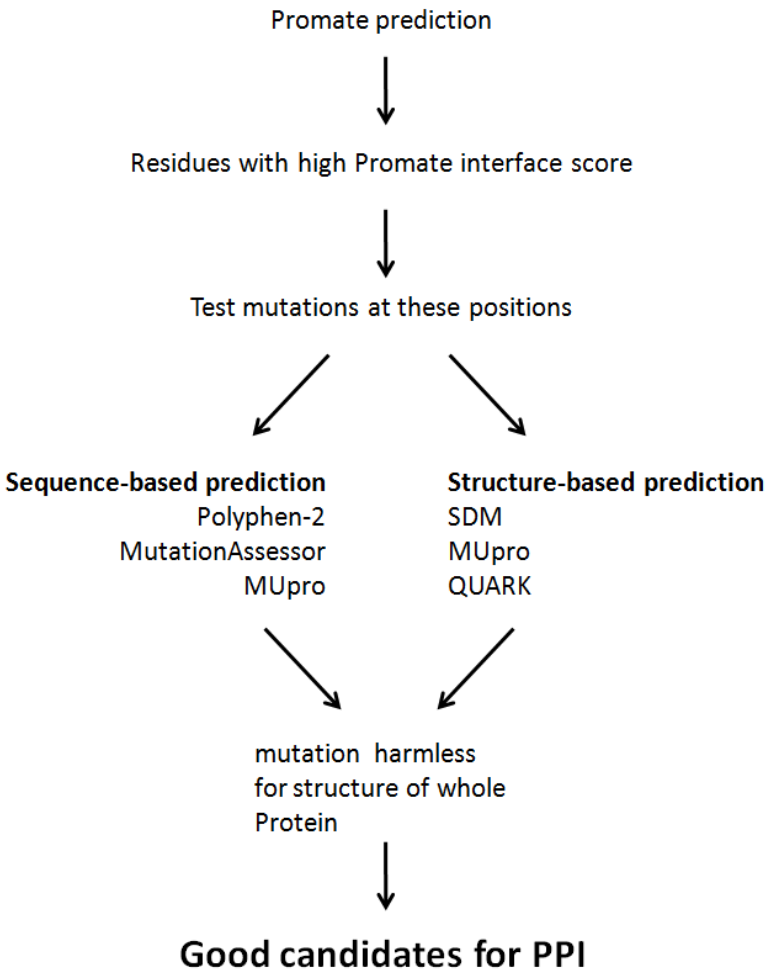


Figure 17. Flowchart of the procedure used to identify residues of putative functional importance. See section 2.6.2 on methods for details.

most of the residues mutated are conserved in homologous sequences of huntingtin in other species. We therefore relied only on SDM and MutationAssessor predictions. SDM uses structural evidence to predict if a given mutation might have an impact on the stability of a protein, by associating the mutation to a change of its deltaG (free enthalpy). MutationAssessor gives a score according to residue conservation, assuming that conserved residues display functional importance. From these two types of information, we kept mutations whose absolute value of deltaG was lower or equal to 0.7 for which MutationAssessor score was lower than 1.9 (values ranked of neutral or low consequence for the protein structure). In the end, we established a list of mutants that might specifically disrupt protein docking to huntingtin with no harm for its structure (table 9). Mutations were then mapped to the model of huntingtin HEAT region interacting with protein HAP1 (see figure 18).

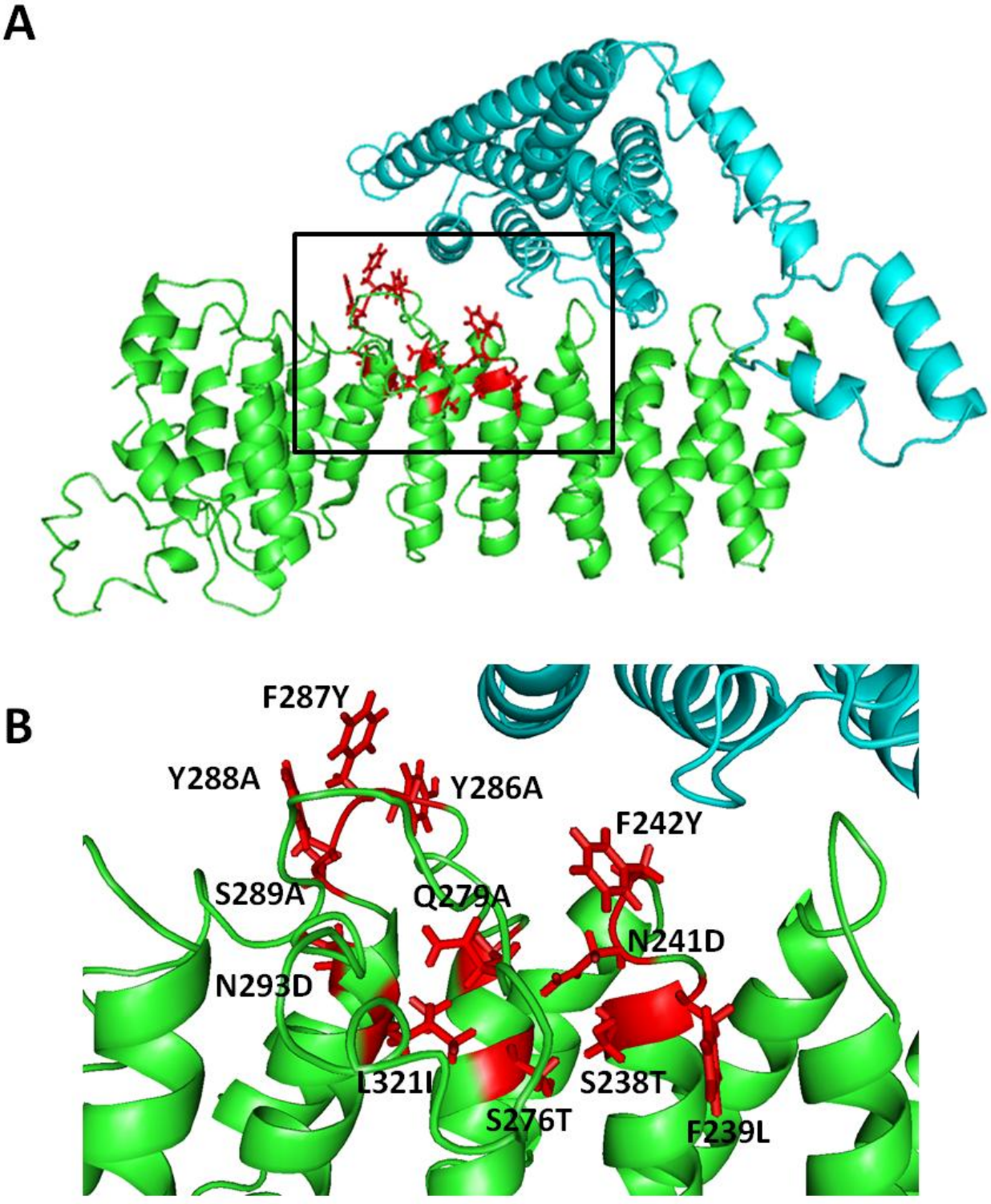

A
B
F287Y
Y288A
Y286A
F242Y
S289A
Q279A
N241D
N293D
L321I
S238T
S276T
F239L

Figure 18. Localization of sites proposed for mutations on a model of huntingtin first HEAT region interacting with protein HAP1. Green: protein huntingtin, residues 77 to 520. Blue: protein HAP1. Red: mutation sites listed in table 9. **A.** View of the two proteins with mutation sites displayed in red. **B.** Magnification of the region where the mutation sites are located. Mutations from table 9 are indicated close to their corresponding residue.

2.6.4. Conclusion

This study of functional enrichment study of the first alpha-solenoid of huntingtin is based on two assumptions. The first one is that the partners of the Q0 fragment only interact with the alpha-solenoid. We are aware though that other structural features are present in the Q0 fragment, including a polyP close to the N-terminal end and several disordered regions, that are also candidate regions for interaction with the different partners found to be uniquely binding the Q0 fragment. But we hypothesize that the alpha-solenoid is more likely to be the region involved in most of the interactions found by Y2H for Q0, as it is a stable domain and was more likely evolved for a clear functional purpose versus polyP or disordered regions, which appeared later in evolution [140].

The second assumption we made here is that these interactions are also happening *in vivo*. This can be put into question because the specific partners of the Q0 fragment did not bind the 1-500 fragments with wild-type polyQ, which should have happened under assumption that these partners specifically bind the alpha-solenoid region. An explanation is that the polyQ prevents these proteins to bind to the alpha-solenoid. But the Q0 fragment only exists in experimental conditions. In normal cellular conditions though, the interactions specifically found to bind Q0 might actually happen but be very transient by being under the control of the polyQ region, which might act as a switch to allow or hinder binding. In conclusion, here we only suggest a putative biological function of the alpha-solenoid; further evidence should be provided to prove this function and eventually describe a mechanism. Our findings agree with a recent report that relates huntingtin to proteins involved in translation and the ribosome using tandem mass-spectrometry [152]. In any case, we recommend taking more attention to this domain, which could be an important agent of wild-type function of huntingtin.

We have also generated a model of the alpha-solenoid region of huntingtin and tried to characterize functional regions on it using computational predictions. This model could be used by biochemists willing to evaluate the role of particular regions of this protein in interaction with other proteins. Nevertheless, our predictions should be used with caution because of putative high instability of the domain, which is likely to display variable conformations. The structural and interaction predicted information could be completed by the selection of huntingtin interacting partners to reveal their precise binding location using protein cross-linking followed by mass spectrometry [153].

Our findings show the potential of focusing on individual regions of huntingtin to understand its biology and pathology. In future experiments, our predicted mutations could be tested and the consequences for the huntingtin interaction network could be studied by Y2H. This could eventually lead to the identification of the function of huntingtin.

2.7. General conclusion of chapter 2

We have improved a neural network used to detect alpha-solenoids and we have applied it to study alpha-solenoids present in databases in order to get insight into their function, structure and evolutionary history. We further applied the algorithm to model huntingtin.

Firstly, we expanded the definition of alpha-solenoids in terms of function, demonstrating that they are more involved in protein-protein interaction than average proteins, but we also show that their function extends beyond proteins to interactions with DNA, RNA and lipids. In the case of nucleic acids, alpha-solenoids show a capacity to bind to specific nucleotidic sequences or be unspecific about the sequence by binding the backbone of the nucleic acid molecule. This wide spectrum of putative interactions of alpha-solenoids can stem from their capacity to show a variety of shapes via their exceptional stretching capabilities [33,60].

Regarding their ability to contact proteins and nucleic acids, it is no surprise that we found alpha-solenoids as being mostly involved in intracellular trafficking; nevertheless, they can have a wide range of other specialized functions (see table 4 for a summary). Protein synthesis is potentially one of them, and we showed in this chapter that the first alpha-solenoid region of huntingtin could be involved in such a function.

In terms of localization, we showed that alpha-solenoids are not only located outside of proteins (which is a requirement to perform protein-protein interactions), but some are found to be localized inside of proteins, such as the one found in PI3Kalpha.

In terms of morphology, we showed that alpha-solenoids are more diverse than previously expected, not only comprising Armadillo, HEAT repeats and HAT, but also possibly TPR, and some types of ankyrin repeats, Pumilio repeats and potentially proteins with leucine-rich repeats. Table 4 shows the diversity of structures that we found. The fact that we were able to detect these motifs does not mean that these new types of alpha-solenoids have any homology to Armadillo or HEAT repeats, or that their profiles are incorrect, but rather that ARD2 is capable to detect sequences that have structural homology that was achieved by convergent evolution using different evolutionary routes. Very different types of alpha-solenoid forming repeats may have other types of similarity. Hinting at this, experiments showed that ankyrin repeats [154] have similar mechanical behavior to HEAT repeats [31].

Continuing the trend defined by our work, we expect that the future discovery of new structures of alpha-solenoids will be valuable to improve even more the detection of alpha-solenoid repeats. Even if the neural network we used in this work was very sensitive to the addition of new

structures to the training set, the improved accuracy of our method was rendered possible by trial on a large set of structures, which helped to detect novel alpha-solenoid proteins in human sequences and proteins of the entire tree of life.

As the function of alpha-solenoids remained general and was mostly restricted to PPI formation and intracellular trafficking, we thought that studying the evolutionary history was important to understand in which biological context alpha-solenoid proteins appeared in evolutionary times. We presented here a distribution of alpha-solenoid repeats across the tree of life. We found that this feature is more present in eukaryotic taxa than in bacterial ones. They represent 1 in 400 proteins in the first group, a magnitude order lower in the second, and are found to be present in 1 in 1200 in the archaea kingdom. Though alpha-solenoids in archaea are the result of horizontal transfer, we speculate that several bacterial and eukaryotic families of alpha-solenoids have emerged independently. While absence of significant sequence similarity between these alpha-solenoids does not necessarily imply that these bacterial and the eukaryotic groups of alpha-solenoids have emerged independently, the specialization of functions found in bacterial sequences (as exemplified in section with analysis of cyanobacteria and planctomycetes groups) suggests that several events of emergence of alpha-solenoids, probably several inside of each kingdom, have happened.

Eukaryotic cells show a higher degree of complexity than bacterial cells have, notably with the presence of a nucleus and organelles, and in the case of metazoans, a more complex environment, with differentiated of cells within tissues and organs. Eukaryotic cells therefore show evolved protein transport machineries to manage the migration of proteic material from one compartment to another, or from one region of the cell to another. Alpha-solenoids show physical properties that render them natural scaffolds for protein interaction, easing transport and regulation of other proteins. They might promote cellular complexity by being available for evolution of new functions while continuing to perform old ones. On the other hand, alpha-solenoids are often long (more than 10 repeats) and show flexibility, which makes them difficult to fold properly upon mRNA translation, therefore requiring complex folding machineries involving chaperones. We have evidence toward this hypothesis as cyanobacteria and planctomycetes have been shown to display chaperones [155,156], which aims at showing that species from these bacterial groups have the molecular machinery to make the evolution of alpha-solenoid proteins possible. This might explain why alpha-solenoids are more likely emerging with eukaryota. Pfam and SMART find 98% of all Armadillo repeats in this group. Several other repeats forming solenoids follow the same tendency: ankyrin repeats (SMART: 87%; Pfam: 75%), HAT (SMART: 97%; Pfam: 100%), and leucine-rich repeats (SMART: 93%). Conversely, TPR repeats are present at 55% in bacteria according to SMART, but this could be due to the fact that TPR repeats are usually only 3 to 4 repeats long and might be easier to shape by the bacterial folding machinery.

An explanation for the success of alpha-solenoids, is that in evolutionary terms, they are cheap to evolve and expand from an initial repeat to expand later by duplication [34]. The initial repeat

could have dimerized with the equivalent repeat of another copy of the same protein, creating the first alpha-solenoid with two repeats on two different copies of the same protein (see [157] for the same explanation about proteins with beta-trefoil).

To conclude, the property of alpha-solenoid proteins to contact other proteins, their high flexibility and compliance of shape might render them potentially involved in multiple pathways, and likely to be easily switched from one to another when evolutionary forces require new protein material to evolve new functions. Huntingtin or UNC45 have shown the potential of alpha-solenoids to be versatile, the first one by being able potentially to contact a high number of partners, and the second one by being involved in several pathways, including cytokines, RNA transport, endocytosis and muscle development. These properties have important implications for the study of alpha-solenoids; for example, the need to study these proteins in a systems-biology context that takes into account PPI information, often a key component, along with gene expression and other genome-wide analyses, to understanding the function of biocellular pathways.

3. Emergence and evolution of the renin-angiotensin-aldosterone system

3.1. Introduction

For the past twenty years, evolutionary medicine has been a topic of special attention as it provides interesting insights in medicine by attempting to explain diseases and disorders of the human body as maladaptations to the environment. Let's take on example. The way the body reacts to a reduction in the cardiac output, for instance upon a heart hypertrophy, can be considered as not making sense regarding its efficiency. The diminution of cardiac output, meaning the quantity of blood pumped by the heart at each heart beat activates baroreceptors, which activates beta-receptors of the sympathetic nervous system. This results in increasing heart rate to compensate the low output, but the problem is that in the long-term the increase of the heart rate tires the muscular fibers of the heart, causing myocardial damages, which further diminishes the cardiac output. Why would the heart react to a decrease of cardiac output by decreasing the output even further? Why should our body respond to severe conditions by accelerating their processes? Part of the explanation is that built as it is, our body cannot cope with heart hypertrophy. In prehistoric times, a diminution of the heart output, for instance upon a state of shock or upon a loss of blood volume, was an acute condition. A decrease of cardiac output induced an increase of heart rate by the sympathetic system in order to keep the perfusion of organs normal. This regulation was happening in a matter of seconds and probably remained for a short period of time. At the end, if the individual had survived, cardiac output was coming back to normal and so the heart rate. Differently, heart hypertrophy is a chronic disease which can only appear in a society where heart suffers enough damage on a regular basis so it is not capable anymore to pump blood correctly to the systemic circulation. As the condition is chronic, the system of regulation is stimulating heart rate all the time in order to keep perfusion to organs the same. This abnormal feedback loop demands a lot to the heart. At some point, the heart cannot manage to sustain a high rhythm anymore, which results in myocardial damages; at such a stage, only a combination of change of habit and probably medication can help the patient to get better. The body does not have what it takes to respond to such harmful conditions typical of our modern environment. This evolutionary perspective on medical conditions can be very relevant to explain numbers of apparent flaws of our bodies and help to understand to which extent our modern environment is involved in their emergence.

The renin-angiotensin-aldosterone system, or RAAS, is the most important system involved in regulation of blood pressure in vertebrates. While its main anatomical, physiological and molecular features have been extensively studied in the past and are well-known for the most part, not so much is known about the molecular origin of the RAAS and how the evolutionary constraints on its emergence relate to important human diseases such as hypertension. We set to study the evolution of this system using information from protein evolution to understand the functional context in which the RAAS emerged in evolutionary times, potentially revealing new aspects of regulation of blood pressure.

In order to study the evolution of the system in molecular terms, we studied the phylogeny of the main genes related to RAAS, and found that some key actors emerged around a time corresponding roughly to the appearance of the first chordates and tunicates, some 500 million years ago. Most importantly, the main anatomical, physiological and molecular features of the system were all present together in the first bony fishes (-400 Myr), to the exception of the Mas receptor (which appeared after the divergence of bony fishes with tetrapods), hinting that this taxon is probably the oldest where the so-called regulation of blood pressure happened. Solid evidence show that angiotensinogen made its appearance in cartilage fishes.

Differently though, one of the proteins of the system, named the angiotensin-converting enzyme (referred to as ACE) appeared much earlier than the other ones in evolutionary times, and probably already existed in the first animals. It was obviously used at that time for a different purpose than it is used for today. Generally speaking, the presence of several RAAS genes in taxa which do not display the anatomical and physiological features associated to regulation of blood pressure means that those genes had a different function in a remote past.

In this introduction, we briefly present the main facts regarding regulation of blood pressure in the human body, with special focus on the renin-angiotensin-aldosterone system. We then give the main mechanisms suspected to be the cause of hypertension, a name for the abnormal elevation of blood pressure in the human body.

3.1.1. Introduction to regulation of blood pressure

Vascular circulation, a feature of chordates, is the key element in the delivery of oxygen and nutrients to the different tissues of the body. However, propulsion of blood by the heart in the different vessels of the body at high pressure also exposes the individual to life threatening situation upon wound. To prevent the body from bleeding completely, the process of scarring over has evolved, among other processes. While healing happens in response to liberation of factors in the local tissue following an injury, another process consists in detecting the modifications of blood pressure (i.e. pressure exerted by the blood flow on the walls of vessels) to keep the blood within the body by constricting the vessels. To achieve this regulation by vaso-constriction, the kidney plays a central role.

3.1.1.1. Definition of blood pressure

Blood pressure is the measure of tension exerted by blood on the walls of blood vessels. This measure depends on three components, the heart, which plays the role of a pump, vessels, which transport blood, and on which the pressure of blood is exerted, and finally blood itself, which is the fluid in motion. Blood pressure or mean arterial pressure (MAP) is the product of the cardiac output (CO) by the systemic vascular resistance (SVR):

$MAP = CO \times SVR$

Transport of blood to tissues and cells can be impaired by several events: blood loss through wound for instance, heart attack and dehydration, which can diminish the volume of water present in the body. Though behavior and local tissues can partially manage some of these problems, the cardio-vascular system importantly participates to manage its own deregulations. How is this achieved? Two types of components are involved in the regulation of blood pressure: sensors, which detect variations of blood pressure, and effectors, which modulate the blood pressure.

3.1.1.2. Sensors of blood pressure variation

Pressure can be detected by measuring two different parameters: either a change in blood volume, which appears upon hemorrhage for instance, or a change in the concentration of ions, osmolality, which can have similar destructive effects on tissues if not regulated.

Volume change can be detected by the so-called baroreceptors (or detectors of pressure). These sensory neurons located in various circulatory beds, such as carotid sinuses, aorta or kidney vessels, are excitable upon a stretch of blood vessels. There are also sensors and ion channels in the distal kidney that are also capable of detecting volume variation. At the level of an anatomical structure named juxtaglomerular apparatus, renal perfusion pressure can be detected by so-called granular cells, and leads to release of renin if the pressure is too low. Production of the enzyme renin is the entry point of the RAAS in the regulation of blood pressure.

Osmolality of blood and its variations can be detected by the hypothalamus. Upon hypertonic conditions (high osmotic pressure), receptor neurons in the hypothalamus become activated. Consequences are stimulation of thirst and subsequent drinking behavior and release of ADH (anti-diuretic hormone), which acts on the ion channels of the wall of the kidney to increase water retention.

3.1.1.3. Effectors of blood pressure

In response to variation of blood pressure, two kinds of actions can be taken by the organism. The human body can either act on volume or on osmolality to modify blood pressure.

- The different effectors that can act on volume are the glomerular filtration rate, the sympathetic nervous system, and the renin-angiotensin-aldosterone system.

The glomerular filtration rate (GFR) is the rate of filtration of plasma by the nephron, the anatomical unit of the kidney. Most of the fluid that enters the kidney is reabsorbed, meaning that it goes back to circulation, and only a small fraction of it is eliminated via urine, in other words, filtered. The value of GFR has to be kept constant in order for the kidney to function properly. When variation of plasma volume happens, the kidney regulates itself by modifying the contraction of the afferent arteriole (i.e. the entry vessel), by dilatation on constriction, to compensate the volume change and keep the GFR steady.

The sympathetic nervous system can also have an effect on the volume of the body. When baro-receptors detect a drop of blood pressure, they stimulate the sympathetic nervous system. This activation causes constriction of the afferent arteriole of the kidney, and thus the GFR, which increases blood pressure and causes indirectly water retention in the body. The modification of blood pressure involving baro-receptors happens very rapidly and can be named a short-term response.

The RAAS, finally, is considered to be the main volume regulator of the human body. It presents a fine mechanism, which can both respond to variation of blood pressure by short-term or long-term response. The RAAS has a key role for instance in maintaining high blood pressure upon hemorrhage. As a consequence, the study of this system is key in order to understand the mechanisms of hypertension [158]. The action of this volume effector is detailed in section 3.1.2.

We also briefly mention effectors that can act on osmolality. These effectors of osmolality concentrate the urine so the general solute concentration of the body diminishes, diminishing the pressure. They include stimulation of thirst, release of vasopressin, an anti-diuretic hormone, and inclusion of new aquaporin water channels in the renal collecting duct.

Thirst occurs when the brain detects an increase of plasma osmolality, for instance upon sweating, when water is evacuated but not the solute. Thirst, by definition, stimulates the motivation to drink, an action that will diminish the osmolality and increase the volume of fluid in the circulation.

Vasopressin or ADH (for anti-diuretic hormone) is produced in the neurohypophysis upon increase of plasma osmolality. The hormone impairs water excretion by stimulating the insertion of water channels in the distal tubule of the nephron, which promotes reabsorption of water into the general circulation, avoiding it to be excreted as urine [159]. Vasopressin also stimulates concentration of solute in urine by mobilizing urea transporters on the membrane of cells from the collecting duct upon water loss, in order to concentrate urea in the urine [160].

Combination of volume and osmolality regulation helps the body to respond appropriately to variation of blood pressure. In the rest of this chapter, we only focus on regulation of volume by RAAS.

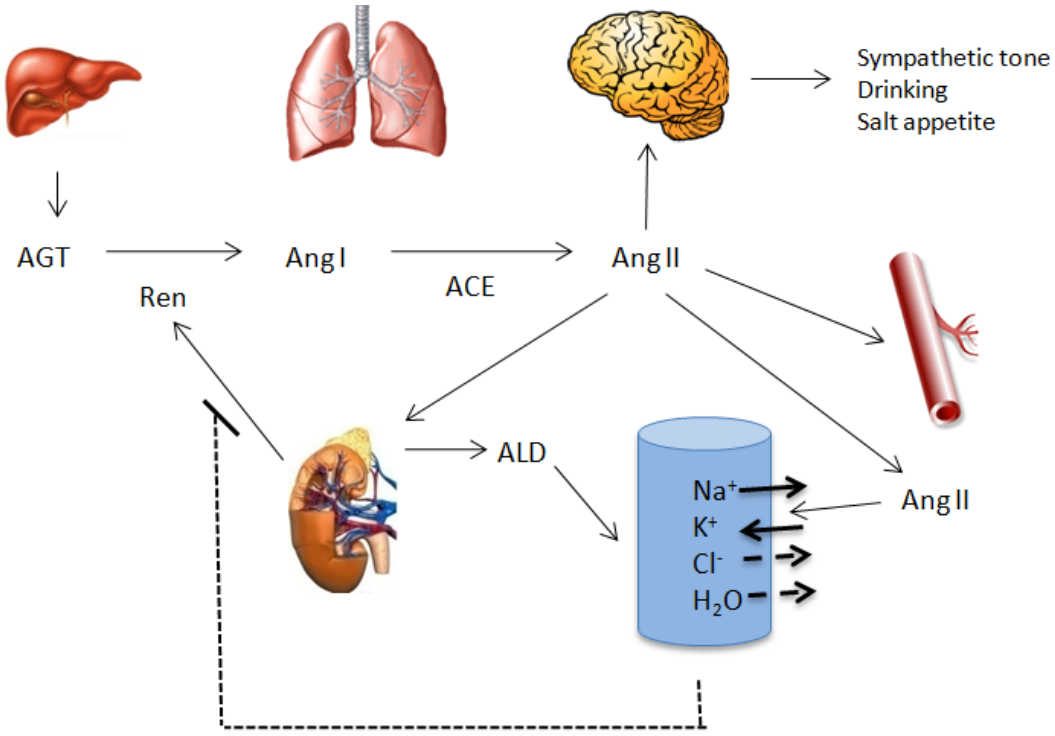


Figure 19. Scheme of the different components of the RAAS. Liver-produced angiotensin (AGT) is cleaved by renin from the kidney to the decapeptide angiotensin I (Ang I), which in turn is converted to angiotensin II (largely in the lung). The effector angiotensin II directs the adrenal gland to release aldosterone (ALD), which directs the brain to increase sympathetic tone, drinking, and salt appetite and also increases vasomotor tone. ALD, sympathetic tone, and angiotensin II act independently to affect NaCl reabsorption in the kidney. A reverse feedback mechanism exists.

3.1.2. Presentation of the renin-angiotensin-aldosterone system

3.1.2.1. Anatomical and physiological features

The different molecules (hormones, proteins and peptides) involved in the proper function of the RAAS are produced in various organs of the human body, with the kidney playing a central role. The main anatomical feature of the system is a zone of the kidney called the juxtaglomerular apparatus (JGA), situated between the region of the glomerulus and the cells of the wall of the renal tubule coming into its vicinity. Cells from the JGA secrete an aspartyl protease called renin, notably when perfusion of the kidney diminishes (volume loss) or if the concentration of salt is too low in the distal tube of the kidney. This low concentration, which can be detected for instance after a drop of the glomerular filtration rate, means that the rate of progression in the renal duct is low. A low rate of progression gives more time during the travel in the renal duct for electrolytes from the primitive urine to be reabsorbed. Their concentration will further decrease and be detected by the JGA as a hint for low volume input.

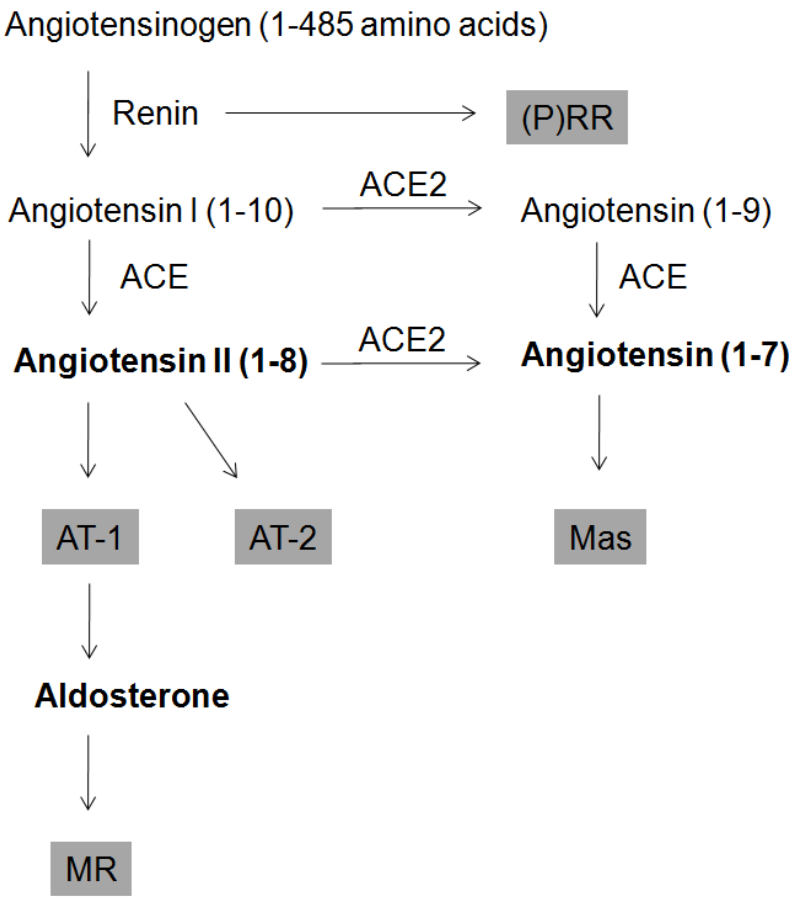


Figure 20. Molecular features of the RAAS. The features of the RAAS are a series of proteins or peptides: (P)RR prorenin receptor, ACE and ACE2 (angiotensin-converting enzymes 1 and 2), AT_1 and AT_2 (angiotensin receptors 1 and 2), Mas for the Mas receptor, MR as the mineralocorticoid receptor. In addition to the canonical RAAS, which is hormonal and circulates around the body, there are also local RAAS in tissues of the brain, the adrenal gland and the heart. Receptors are represented as grey boxes. Effectors with pharmacological effect are in bold.

3.1.2.2. Molecular features

Angiotensinogen plays a central role in the RAAS, as it is the precursor molecule of the peptide angiotensin II, a strong vaso-constrictor that is the effector of the system. Angiotensinogen is mainly produced in liver (figure 19). This protein is cleaved to a decapeptide called angiotensin (angiotensin I or Ang I), by protein renin produced in the JGA. Angiotensin I is later cleaved by another enzyme, the angiotensin-converting enzyme (or conversion enzyme, or ACE), into a smaller peptide of 8 residues, angiotensin II (Ang II), a highly active molecule with strong vaso-

constriction property. ACE is classified as a matrix metalloproteinase produced mainly in pulmonary endothelial cells, and whose measurement in medicine is a hint to diagnoses for several diseases. ACE has also been shown to have other functions in different tissues [161]. At the same time, other enzymes display ACE-like activities, one being protein chymase, which is mainly produced in the mast cells of the heart [162].

Angiotensin II primarily acts on blood pressure by exerting a strong vaso-constriction on vessel walls (figure 19). It has another action on the adrenal cortex in order to stimulate the release of aldosterone. Aldosterone has an action on cells of the collecting duct of the kidney to increase the reabsorption of Na^+ and Cl^-. Angiotensin II also exerts its own influence on the Na^+ reabsorption function of the collecting duct. It also stimulates thirst and salt appetite in the brain and increases sympathetic tone. These different actions tend to strongly increase blood pressure [163].

Renin is the limiting step in the production of angiotensin II. The system works with a negative feedback phenomenon that inhibits renin release, which indirectly impairs angiotensin II levels, inactivates baroreflex sensors, and inhibits sympathetic tone. The RAAS can be seen as a network of different molecules, either proteins or peptides (figure 20). Renin, whose primary function is to cleave angiotensinogen, also has a receptor called prorenin receptor (P)RR, whose function is not clearly known. This receptor binds both renin and renin's precursor prorenin [164]. It can induce MAP-kinase cascades in the intracellular compartment of the cells it is attached to; it apparently serves to activate prorenin. Whatever its function, (P)RR is not implied in volume regulation. Angiotensin II displays two target receptors, AT_1 and AT_2 of different functions. AT_1 displays two different isoforms (AT_{1A} and AT_{1B}) whose existence has been shown in mice [165]. Aside from the canonical axis angiotensinogen - angiotensin I - angiotensin II, other peptides can be produced and potentially help to regulate the system. Angiotensin II can be cleaved by a different ACE that was discovered in 2000 [166], ACE2, into a peptide of 7 residues called angiotensin (1-7). Angiotensin (1-7) can bind to the Mas receptor, a protein encoded by the Mas oncogene. Angiotensin I can be cleaved by the same ACE2 into a peptide of 9 residues, angiotensin (1-9), which can later be cleaved into angiotensin (1-7) by ACE. Moreover, aldosterone, whose production is stimulated by angiotensin II, binds the mineralocorticoid receptor (MR) to exert its effects on salt reabsorption. This salt reabsorption actually comes in two steps: first aldosterone stimulates reabsorption via the mineralocorticoid receptor, and then, salt reabsorption, along with vasoconstriction, is strongly stimulated by angiotensin II binding to the AT_1 receptor. Differently, AT_2 helps to modulate the response. Angiotensin (1-7) exerts an action that is globally the opposite of the one of angiotensin II, by binding the Mas receptor. The RAAS exerts a complex response in cells that has been reviewed extensively [163,167].

3.1.3. Putative mechanisms leading to hypertension

Hypertension is a complex phenomenon whose mechanisms, still not clearly known and under debate, can be described as an abnormal sustained increase of blood pressure. Hypertension is the result of an increase of peripheral resistance, in other terms it is an increase of the action of the vessels to oppose to the flow of blood, mostly by narrowing their diameter.

One explanation for long-term hypertension in the modern world is a consequence of a sustained high-salt diet. Hypertension manifests itself by the persistence of high renin levels, in circumstances that do not require high blood pressure. The permanent injection of salt in the circulation causes the system to produce renin in high amount in permanence. As a result, a major strategy to cure hypertension is to inhibit components of the RAAS. The understanding of the physiological and molecular features of the system is very important to imagine future cures for hypertension. High-salt diet as a cause for hypertension has been for long time under debate but is today generally accepted, though a clear connection between the two phenomenon remains unclear [168]. Other explanations include genetics background as a major component [169].

3.2. Evolution of anatomical and physiological features of the renin-angiotensin-aldosterone system (RAAS)

In section 3.1, we have presented the main facts regarding the regulation of blood pressure and its eventual deregulation in the human body. Here, we review what is known about the evolution of anatomical and physiological features of the RAAS. In section 3.3, we will present the evolution of the system, from a molecular perspective.

The study of the evolution of the RAAS can help to understand the biology of its components and molecules, and address specific questions that are still incompletely addressed today. Is the list of components of the system complete? For instance, local RAAS have been proved to exist in brain and splanchnic territory [170,171]. Are they all relevant and vital for the function of the respective tissues? Do they have local functions apart from helping regulate hypertension at the level of the organism? In which species is the RAAS conserved on all of these three biological levels? When exactly did the RAAS emerge and which anatomical, physiological and molecular constraints helped to evolve it? These questions are important to understand a system that has to do with numerous cardiovascular diseases, and to know how targeting the system might help to provide efficient treatments [167,172].

3.3. Analysis of DNA sequences of proteins of the renin-angiotensin-aldosterone system

In order to get a broader perspective on the evolution of RAAS, we conducted an analysis of the information available in public databases (mostly from the NCBI [173]) regarding proteins and genes of the system. The study of complete genomes can be very informative because it allows

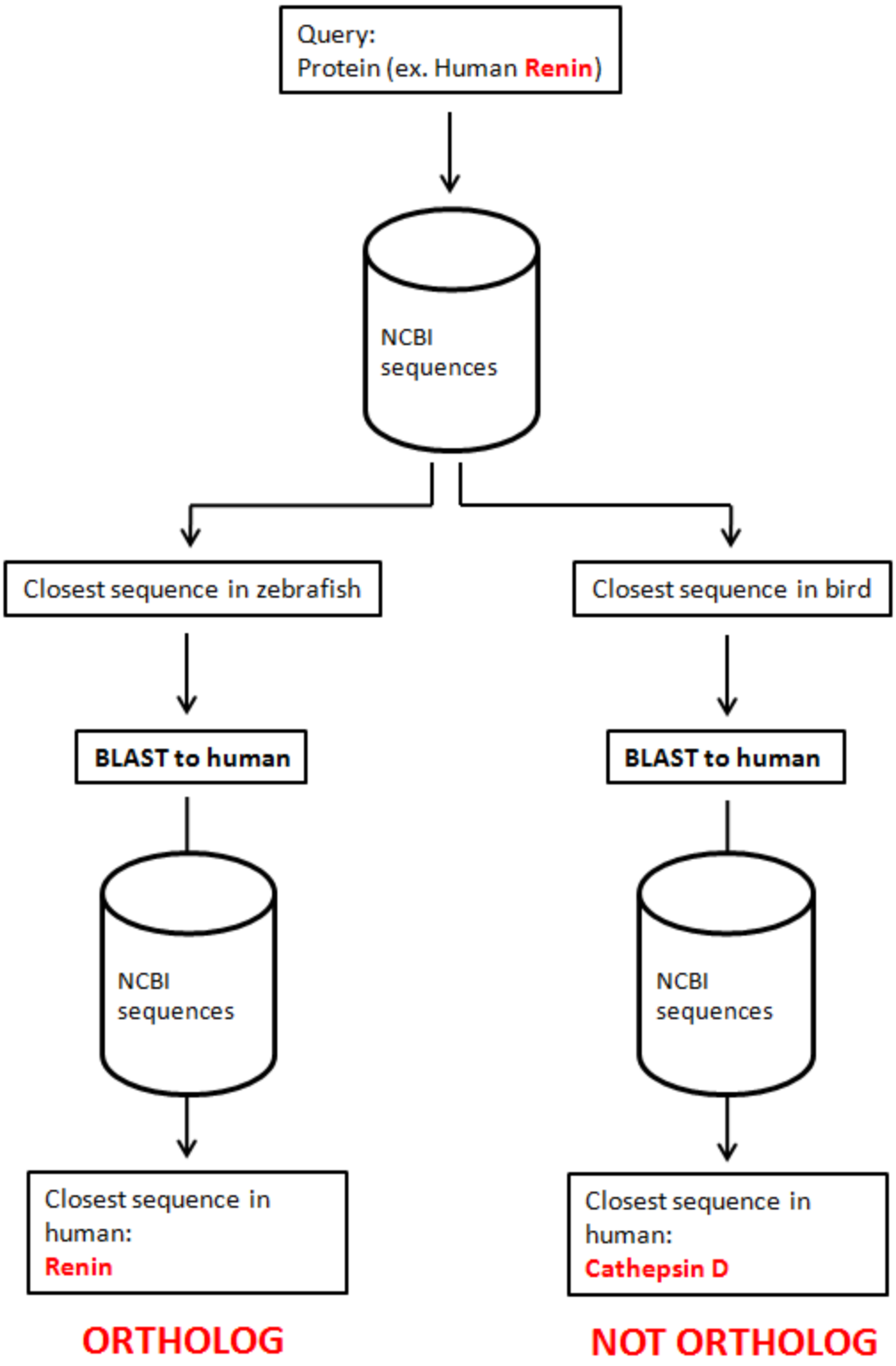

Query:
Protein (ex. Human Renin)
NCBI sequences
Closest sequence in zebrafish
Closest sequence in bird
BLAST to human
BLAST to human
NCBI sequences
NCBI sequences
Closest sequence in human:
Renin
Closest sequence in human:
Cathepsin D
ORTHOLOG
NOT ORTHOLOG

Figure 21. Reciprocal searches to demonstrate the orthology of two genes. Firstly, the BLAST tool is used to search for sequences close to the human renin in the NCBI database, against zebrafish sequences in the left example, and against bird sequences on the right example. This procedure aims at finding putative homologs of the query sequence, in zebrafish and bird, respectively. In both cases, if the best match has an E-value below threshold (for instance 10^{-10}), a BLAST of the best match is performed against the human genome. If we find the initial sequence on top of the results, i.e. human renin, the putative zebrafish or bird sequence found as the best match is a true ortholog. If this is not the case, it is closer to another human sequence that is not renin. In the example displayed, BLAST to zebrafish sequences identifies a renin ortholog, while the BLAST to birds identifies the closest gene as a cathepsin D, which is closer to human cathepsin than to human renin, and so it is not the ortholog of the query renin sequence, although it is homologous to it.

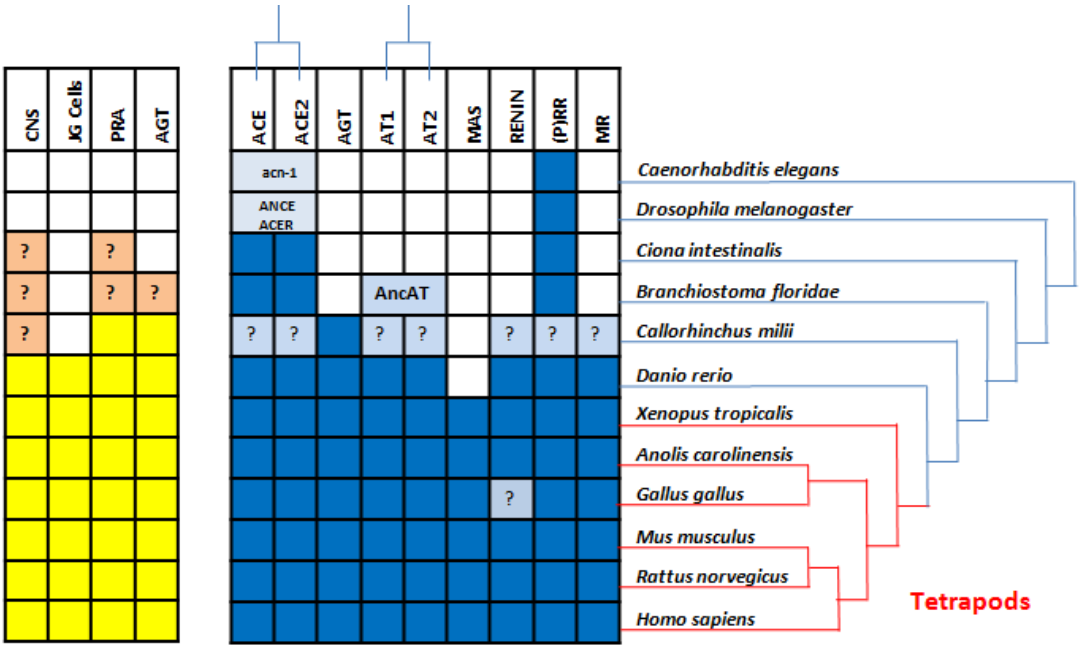
evaluating the presence or absence of proteins across species. Global studies of gene conservation have been especially revealing regarding the evolution of vertebrates [174]. Here, we focus the analysis on sequences homologous to genes of the RAAS present in 12 representative model organisms (figure 1 and table 10). The genomes of all of these organisms have been published, to the exception of *Callorhinchus milii*, the elephant shark. In order to perform our analysis, we used the BLAST algorithm [175] to search the NCBI databases. In order to verify homology of two sequences, we used reciprocal searches (figure 21). A reciprocal search consists in searching the NCBI protein database for sequences similar to a query sequence. If the best match has an E-value of probability below an acceptable threshold (here 10^{-10}), then this matching sequence is used as a query for BLAST against the species of the initial query sequence. If the best match of this second query is the initial query, we considered that the two sequences were orthologs. For each gene, we then performed an alignment of all the orthologs found and built a phylogenetic tree to verify orthology. An important remark is that most of the proteins we found are only predicted from gene sequence and await experimental validation. A first observation regarding our result (figure 22A) is that presence of RAAS genes is concomitant to the appearance of the anatomical features of the juxtaglomerular apparatus. We note that the zebrafish (*Danio rerio*), the representative of bony fishes, displays the most primitive JGA [176], but already shows eight of the nine proteins from the RAAS.

We briefly comment the conservation of each of the nine genes selected here to finally use all the information to sketch an evolutionary scenario for the emergence of the renin-angiotensin-aldosterone system.

3.3.1. Angiotensinogen

Human angiotensinogen (UniProt ID: P01019) is a protein of the serpin family (figure 23A and structure in figure 23B). It is a central element of the RAAS as it is the precursor of angiotensin I and II. In respect to that, it could be used to date the time where the RAAS emerged in evolutionary time. An ortholog of angiotensinogen was found in *Callorhinchus milii*, the

A


CNS
JG Cells
PRA
AGT
ACE
ACE2
AGT
AT1
AT2
MAS
RENIN
(P)RR
MR
acn-1
ANCE
ACER
AncAT
Caenorhabditis elegans
Drosophila melanogaster
Ciona intestinalis
Branchiostoma floridae
Callorhinchus milii
Danio rerio
Xenopus tropicalis
Anolis carolinensis
Gallus gallus
Mus musculus
Rattus norvegicus
Homo sapiens
Tetrapods

B

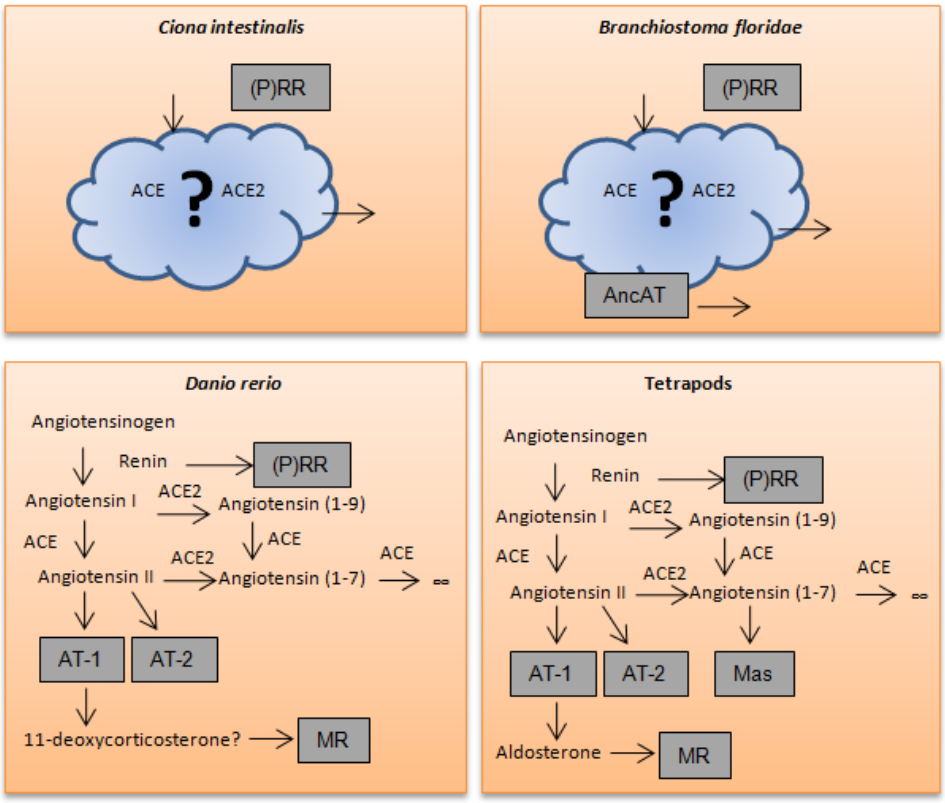

Ciona intestinalis
(P)RR
ACE
?
ACE2
Branchiostoma floridae
(P)RR
ACE
?
ACE2
AncAT
Danio rerio
Angiotensinogen
Renin
(P)RR
Angiotensin I
ACE2
Angiotensin (1-9)
ACE
ACE
ACE2
ACE
Angiotensin II
Angiotensin (1-7)
AT-1
AT-2
11-deoxycorticosterone?
MR
Tetrapods
Angiotensinogen
Renin
(P)RR
Angiotensin I
ACE2
Angiotensin (1-9)
ACE
ACE
ACE2
ACE
Angiotensin II
Angiotensin (1-7)
AT-1
AT-2
Mas
Aldosterone
MR

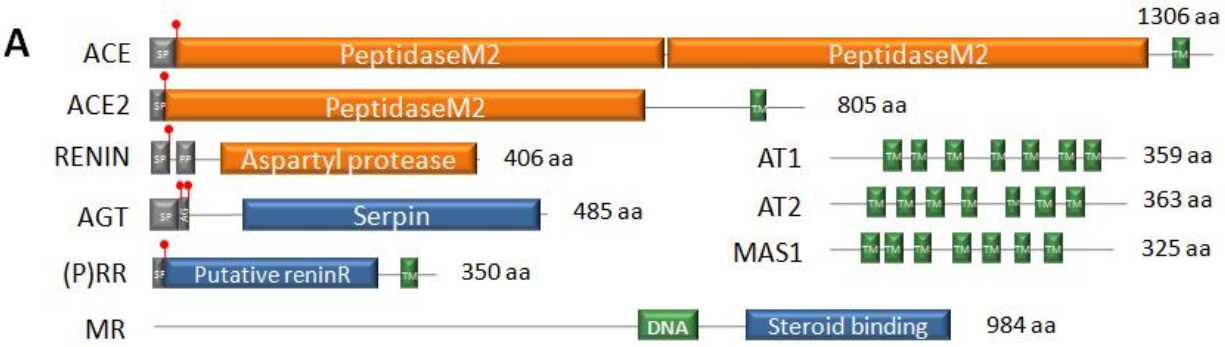
Figure 22. Comparison of the RAAS in multiple species. A. Left panel (yellow) includes data from physiological studies: presence known or supposed of RAAS in the central nervous system (CNS), juxtaglomerular cells (JGCells), plasma renin activity (PRA), and angiotensin or angiotensin-like activity known (AGT). Note that PRA does not measure renin, but rather the conversion of AGT to angiotensin I; renin is not the only enzyme with this capability. Right panel (blue) shows sequence data found by BLAST inquiry. Blanks indicate instances in which the property could not be found. Question marks denote instances of uncertain or contradictory data. **B.** Model of the stepwise emergence of the components of the RAAS based on their conservation across several taxonomic divisions. *Ciona intestinalis* contains both ACE and ACE2, as well as the prorenin receptor, but the many components missing show evidence that these three proteins have functions ancestral to the RAAS. *B. floridae* has an additional member, AncAT, an ancestral version of the angiotensin receptors. After a large gap, our next closer relatives whose complete genomes we know, the bony fishes, have a human-like system, with two notable differences: a possible use of a precursor of aldosterone and the absence of the Mas receptor. The tetrapods (mammals, reptiles, amphibians, birds) have the complete system, with the exception of renin, which could be missing in aves (see main text).

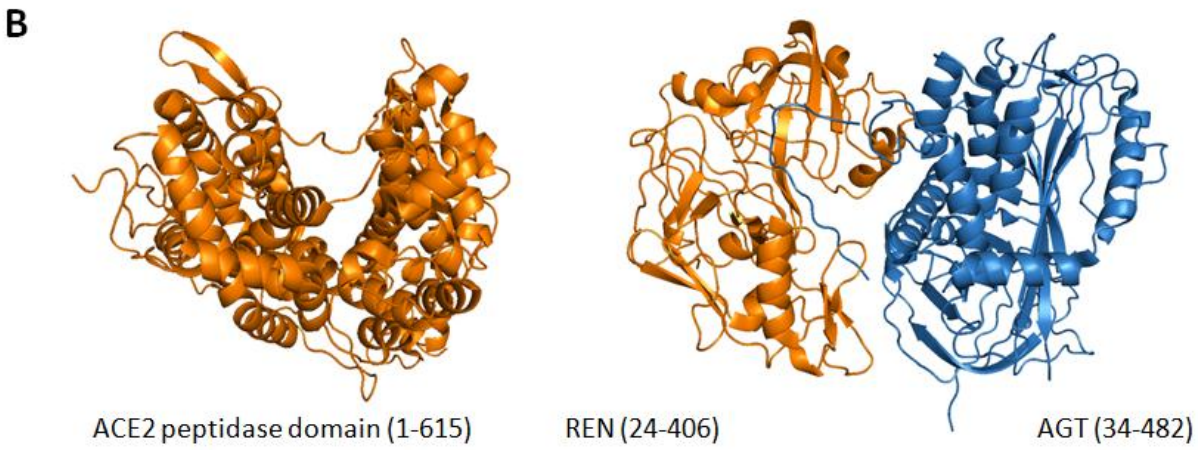
elephant shark (figure 24), and for all highest taxa, until human. Angiotensinogen is absent in invertebrates, including tunicates.

We then focused attention on lampreys, which are among the most primitive vertebrates. To study the presence of angiotensinogen peptides in this taxon, we performed a reciprocal sequence similarity search but did not find an ortholog of angiotensinogen. We could not determine if an active peptide equivalent to angiotensins is present in lampreys. In GenBank, there are sequences which are annotated as "putative angiotensinogens" (for instance FM954978 present in *Lampetra fluviatilis* [177]). Using reciprocal sequence similarity searches, we could find that these supposedly orthologs of angiotensinogen, are actually orthologs of SERPIND1 sequence in *Danio rerio*, and higher vertebrates. These findings confirm that the RAAS was probably established in an ancestor of the first jawed fishes.

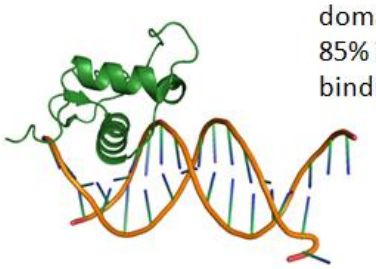
How could angiotensinogen have emerged in evolutionary times? This protein displays a serpin domain (figure 23A). These domains are involved in inhibiting proteases [178]. Ancestral sequences of angiotensinogen binding to proteases might have been a first step in the property of renin to cleave angiotensinogen into angiotensin I [179], the protein being stabilized while the peptide sequence could appear in its sequence.

There is clearly a need of more angiotensinogen sequences in a broader range of species at the basis of the vertebrate phylum in order to understand exactly when the RAAS peptides (angiotensin I and II) emerged.

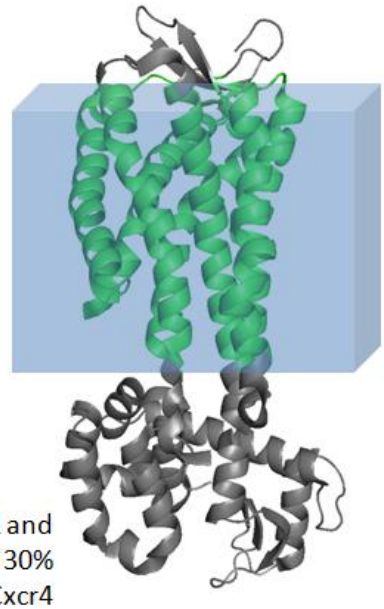




NRC32 steroid binding domain (712-984)



NRC32 DNA binding domain (600-686) is 85% id to NRC31 DNA binding domain



AGTR1 and AGTR2 are 30% id to Cxcr4

Figure 23. Structural features of nine human proteins relevant to the RAAS. A. Domain organization of ACE, ACE2, renin, AGT, (P)RR, and MR. Transmembrane alpha-helix (TM), signal peptide (SP), pro-peptide (PP). Red box on angiotensinogen diagram: angiotensin I sequence (AG). Red symbols indicate protein cleavage sites. **B.** Solved 3D structures of these proteins or homologs (when indicated). ACE2: peptidase domain (fragment 1-615, PDB:1R42); REN and AGT: complex of renin (blue) and AGT (orange). Note the N-terminal of AGT protruding into the renin molecule for processing (PDB:2X0B); NRC32: steroid binding domain (blue; PDB:2AA2) and DNA binding domain (green) with DNA (stick model) from 85% identical rat glucocorticoid receptor NRC31 (PDB:3G9P); AGTR1/AGTR2 are 30 % identical to the CXCR4 chemokine receptor whose structure is shown (TM helices in green; PDB:3OE0). All protein structures are represented using the PyMOL Molecular Graphics System software (DeLano Scientific, Palo Alto, California).

3.3.2. Angiotensin-converting enzymes

ACE (angiotensin-converting enzyme) is a well-known peptidase involved in the processing of angiotensinogen into angiotensin I and as such, has been used since the seventies as a target for drugs called ACE inhibitors, to treat hypertension.

In 2000, an isoform of ACE was discovered and named ACE2. ACE and ACE2 share a common domain, peptidase M2, and are the result of gene duplication. More interestingly, we found that ACE has two of these domains, a result of an additional internal duplication (figure 23A and 25, see also structure in figure 23B).

We then wanted to explore whether or not these sequences are present in other taxa. Starting with human sequences of ACE and ACE2, we were able to identify orthologous sequences in all representative species selected for this study, from *Caenorhabditis elegans* to vertebrates (figure 22A), though in *C.elegans* only one sequence was found, and presence in shark is only speculated as genome sequencing appears to be incomplete. Moreover, we found multiple instances of both ACE and ACE2 sequences in our two tunicate representatives (figure 25A). The different relations between the different sequences can be seen both in a phylogenetic tree of protein sequences (figure 25A) and on a sketch representing the different events of duplication along a tree (figure 25B).

Two of the sequences found from *Drosophila* are gathered in an ‘outgroup’. This means that the two drosophila sequences emerged after divergence of the common ancestor of insects and vertebrates. These sequences, ANCE (for angiotensin-converting enzyme gene, UniProt ID: Q10714) and ACER (for angiotensin converting enzyme related, UniProt ID: Q9VLJ6) belong to the same family of genes as the two human ACE (classified both in family 3.4.15.1 or ‘Dipeptidyl carboxypeptidase I’ in Swiss-Prot). They both have been proven to have an endopeptidase activity during the development of fly but do not show activity during adulthood [180]. It has been proven that ANCE hydrolyzes Angiotensin I [181], an activity that was also

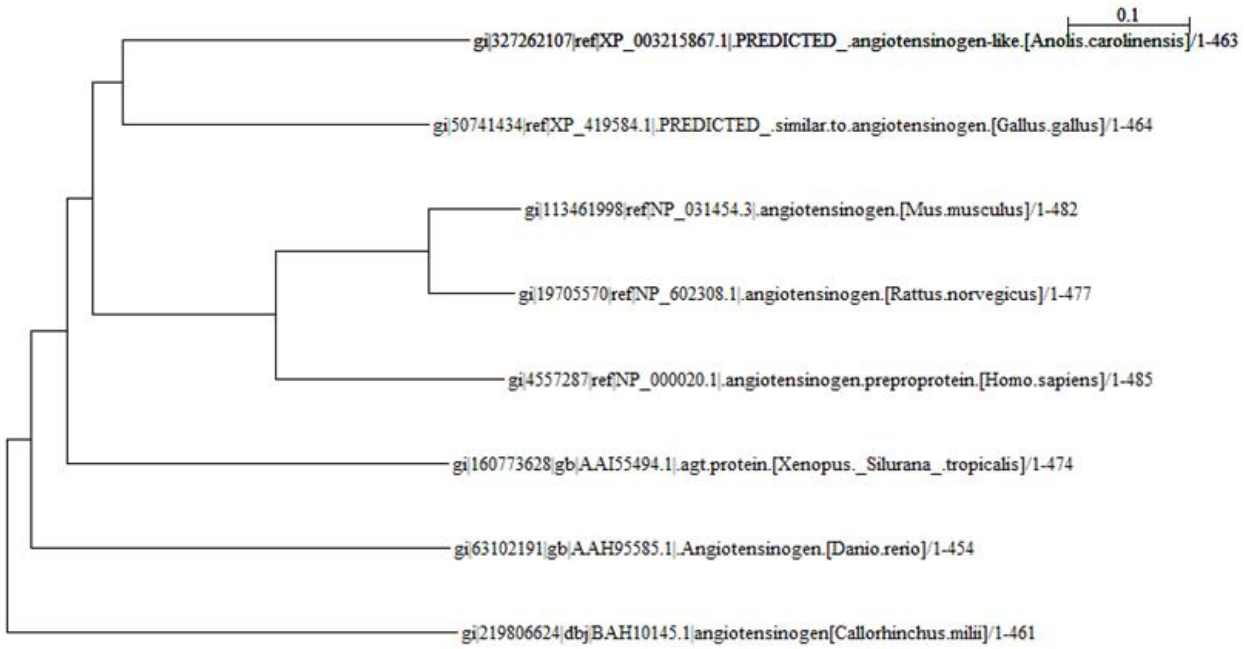
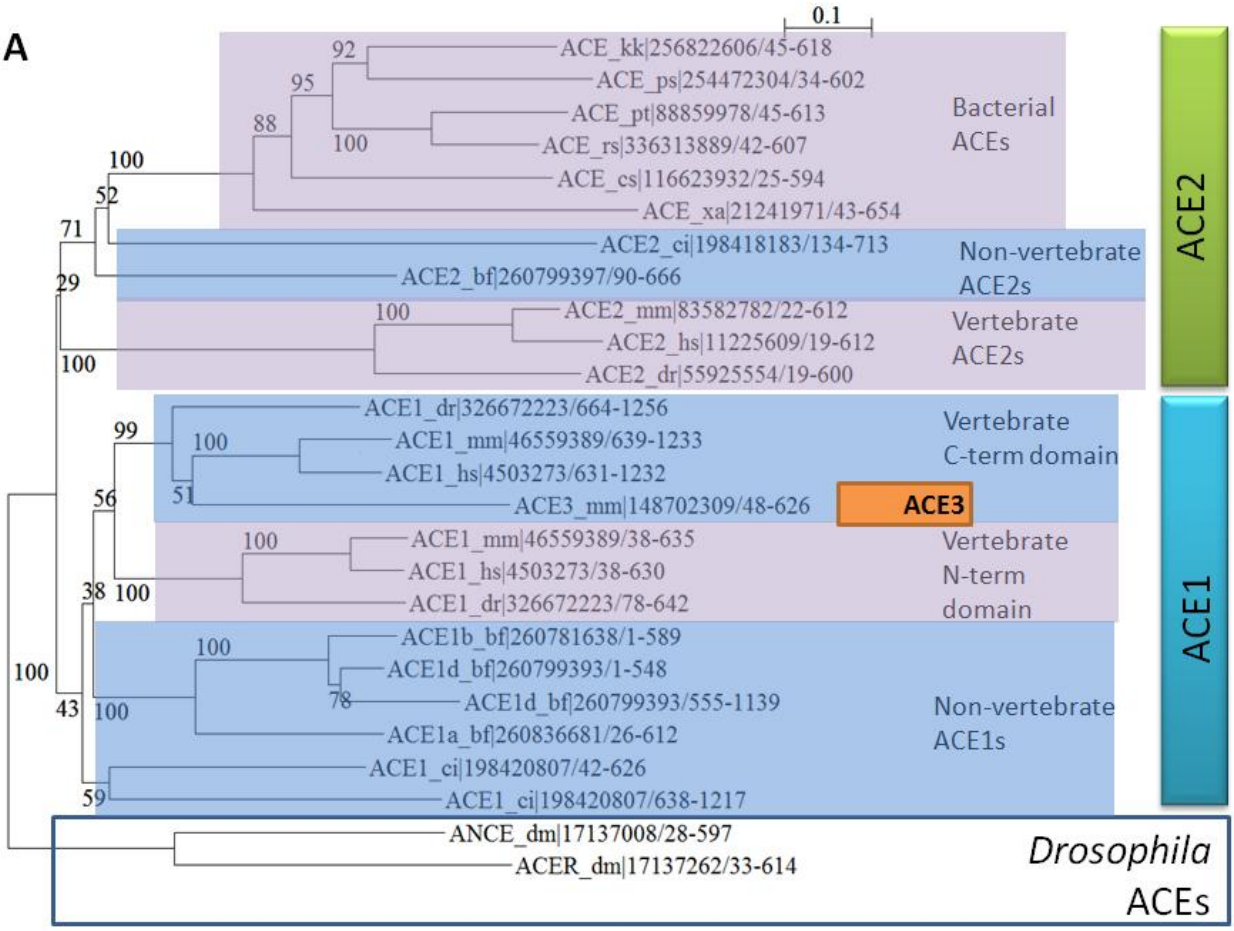


Figure 24. Evolution of angiotensinogen sequences. Multiple sequence alignment was produced using the MUSCLE method [182] as implemented at the EBI web server. The alignment was examined, and phylogenetic trees were generated using ClustalX Version 2.1 [183] excluding positions with gaps and correcting for multiple substitutions.

shown to exist in a bacterial ACE homolog, present in *Xanthomonas axonopodis pv. citri* [184]. This means that the potential of ACE sequences to act on angiotensins is much older than the existence of blood pressure regulation and presence of angiotensinogen itself. Angiotensinogen evolution might have been constrained by the potential of ACE to cleave certain sequences within proteins and peptides.

Interestingly, ACE activity is related to development in various species from *Drosophila* to mouse, and is highly expressed in testis (lung and thyroid being the other two locations of top-three most highly expressed ACE in adults according to BioGPS [185]). In the mosquito, ANCE is hypothesized to be involved as a "peptide-processing enzyme" in the seminal fluid [186]. ANCE has also been found to be potentially involved in embryogenesis regulation upon activation by blood metal [186]. In freshfly *Neobelliera bullata*, substrates and inhibitors of an identified ACE activity have been documented. They have been shown to play a role in development of ovaries, which adds more evidence for the role of an ACE activity in reproductive system in insects [187]. Moreover, an isoform of ACE was also found to be present in germ cells of male mice [188]. Male mice lacking this isoform show infertility [189]. This adds more evidence to the compliance of ACE sequences to switch from one function to another during evolutionary times.


A
0.1
ACE_kk|256822606/45-618
ACE_ps|254472304/34-602
ACE_pt|88859978/45-613
ACE_rs|336313889/42-607
ACE_cs|116623932/25-594
ACE_xa|21241971/43-654
Bacterial ACEs
ACE2_ci|198418183/134-713
ACE2_bf|260799397/90-666
Non-vertebrate ACE2s
ACE2_mm|83582782/22-612
ACE2_hs|11225609/19-612
ACE2_dr|55925554/19-600
Vertebrate ACE2s
ACE2
ACE1_dr|326672223/664-1256
ACE1_mm|46559389/639-1233
ACE1_hs|4503273/631-1232
ACE3_mm|148702309/48-626
Vertebrate C-term domain
ACE3
ACE1_mm|46559389/38-635
ACE1_hs|4503273/38-630
ACE1_dr|326672223/78-642
Vertebrate N-term domain
ACE1b_bf|260781638/1-589
ACE1d_bf|260799393/1-548
ACE1d_bf|260799393/555-1139
ACE1a_bf|260836681/26-612
ACE1_ci|198420807/42-626
ACE1_ci|198420807/638-1217
Non-vertebrate ACE1s
ACE1
ANCE_dm|17137008/28-597
ACER_dm|17137262/33-614
Drosophila ACEs

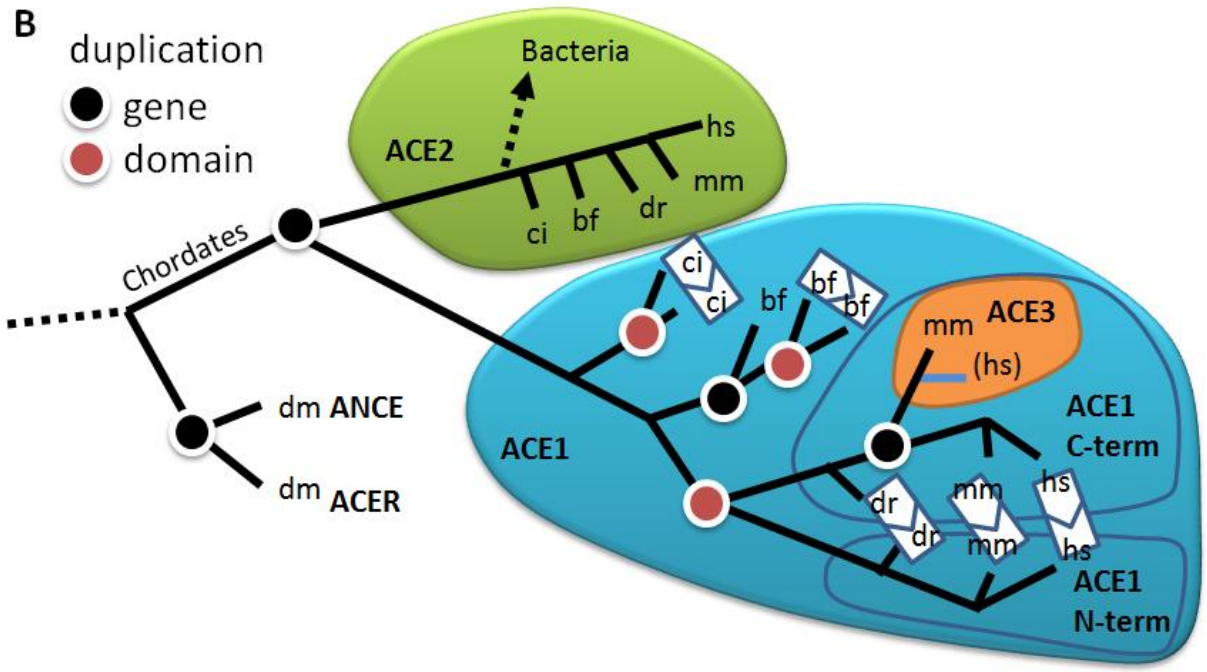

B
duplication
gene
domain
Bacteria
ACE2
hs
mm
ci
bf
dr
Chordates
dm ANCE
dm ACER
ACE1
ACE3
mm
(hs)
ACE1 C-term
ACE1 N-term

Figure 25. Evolution of the ACE family. A. Phylogenetic tree of the peptidase domains of selected eukaryotic and bacterial ACE homologs. The numbers at the branches indicate number of bootstrapping tests that resulted in the marked grouping: Values close to the total used (100) indicate reliable branches. The labels indicate the subfamily, a two letter abbreviation of the species name, GenPept identifier, and amino acid range. Species abbreviations of eukaryotic species are dm (*Drosophila melanogaster*), ci (Ciona intestinalis), bf (*B. floridae*), dr (*Danio rerio*), mm (*M. musculus*), and hs (*H. sapiens*). ACE_xa corresponds to the bacterial *Xanthomonas axonopodis* sequence; for the other bacterial species, please refer to the database records. Drosophila sequences contain a single domain (ANCE_dr, ; ACER_dm) and constitute an outgroup indicating that they are ancestral to chordate ACE1/ACE2. Multiple bacterial sequences (including the *X. axonopodis* sequence) contain a single protease domain that groups with single domain ACE2s and is not ancestral to both ACE1 and ACE2. This suggests that the bacterial sequences are a result of horizontal transfer from an ancestral chordate species. **B.** Interpretation of the phylogenetic tree. The ACE family originated before the divergence of chordates from arthropods. Gene duplications (black dots) have expanded this family, for example, leading to the existence of ACE1 and ACE2 in chordates. Multiple events of domain duplication (red dots) have happened in the ACE1 subfamily, an important one leading to the vertebrate ACE1, which contains an N-terminal and a C-terminal catalytic domain. ACE3 is a single domain ACE, which stems from duplication of the mammalian C-terminal domain of the ACE1. This sequence seems to have evolved into a pseudogene in humans (blue line). Orthologs of vertebrate ACE2 are present in many bacterial species. Their close homology to non-vertebrate ACE2s suggests that they are the result of a single event of horizontal transfer from an ancestral non-vertebrate species. The grouping in the phylogenetic tree of the bacterial sequences analyzed here suggests that this initial event was followed by further events of horizontal transfer between bacterial species, indicating that bacterial ACEs have acquired a function that confers an evolutionary advantage to the species bearing it. Methods to calculate and display the phylogeny are the same as in figure 24.

The most important finding of this phylogenetic analysis of ACE sequences is that, as we proposed for angiotensinogen, ACE and ACE2 seem to have an ancestral function besides the one they are currently known to have in higher taxa, i.e. regulation of hypertension, as there is no blood pressure in flies, or at least, not one comparable to the one of vertebrate, as flies got hemolymph. Both ACE and ACE2 have an ortholog in our tunicate representative, *Ciona intestinalis*, or sea squirt, and also in cephalochordates, whose representative in our study is the lancelet *Branchiostoma floridae*, or amphioxus. Nevertheless, at the same time, these two taxa do not display the morphological and molecular components of the RAAS.

Finally, we also show on our phylogenetic tree that some ACE sequences might be the result of horizontal gene transfer, potentially leading to new functions for these molecules. This is shown by the fact that the closest homologues to bacterial ACE-like sequences in eukaryota are chordate orthologs of human ACE2. One of them belongs to *Branchiostoma floridae* (GenPept ID: 260799397, e-value: 1E-168). Generally speaking, these bacterial sequences cluster with ACE2 sequences of *Ciona intestinalis* and *Branchiostoma floridae* (figure 25A). The event of gene transfer of these ACE2 sequences happened in an ancestor of *Ciona intestinalis* after

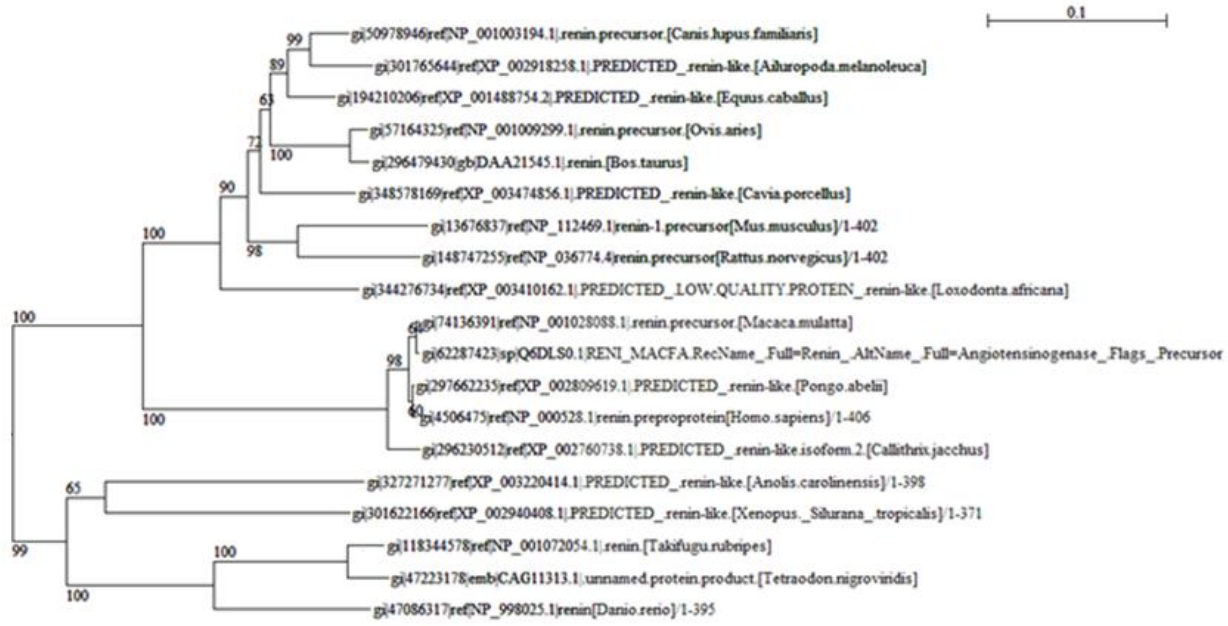


Figure 26. Evolution of renin. Methods to calculate and display the phylogeny are the same as in figure 24.

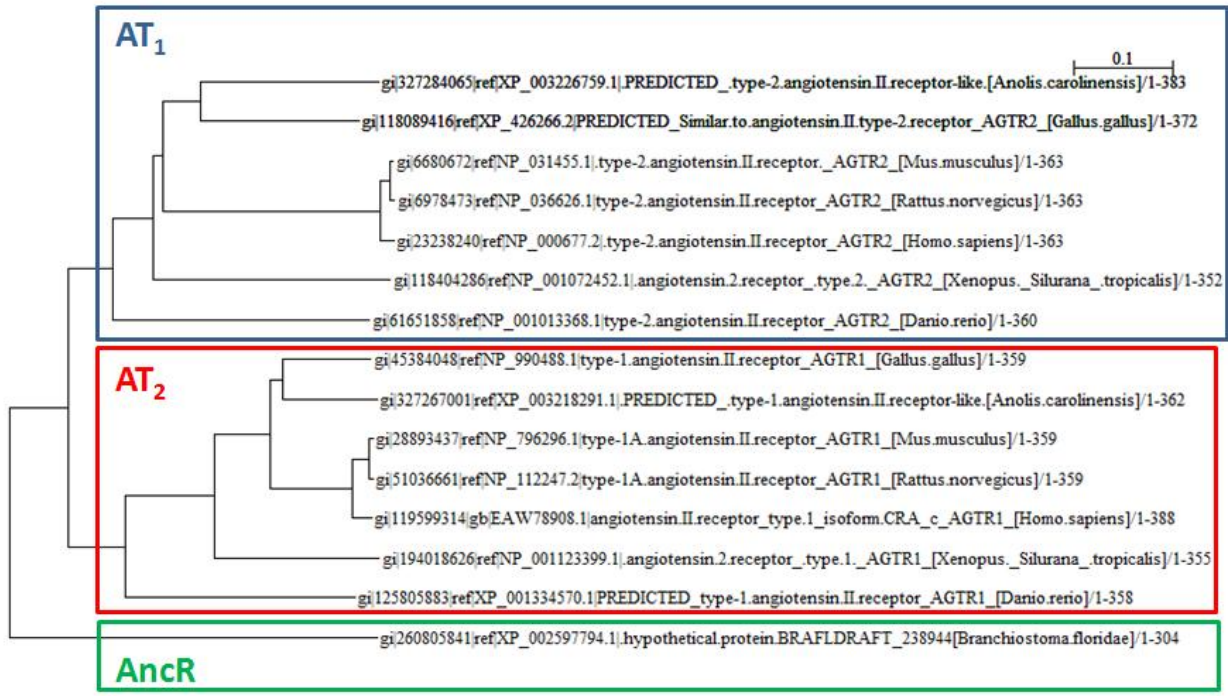
divergence of tunicates from the other chordates, in the direction of a bacteria ancestral to the ones that are today detected to bear ACE2 sequences. Our sketch summarizing the evolution of ACE sequences suggests further events of horizontal gene transfer (figure 25B).

Our study points that contrary to what one would think intuitively, ACE2 is not closer to ACER than ACE. Both ACE and ACE2 come from a unique common ancestor with ACER and ANCE. These latter are used exclusively during embryogenesis, while ACE and ACE2 have acquired specific functions that have proven to be essential in adulthood, though it is not clear until today whether they are mandatory for normal development. Genetic ablation of ACE for instance is not lethal, while ACE2 K.O. mice become adult though displaying severe defects of cardiac contractility [190].

3.3.3. Renin

Renin is the initiator of the long-term body response to volume change and, along with angiotensinogen, must have been present very early in the evolving hypertension regulation system. As a result, it has a detectable ortholog in the taxa where angiotensinogen is present (figure 22). Nevertheless, we did not find a renin protein in *Gallus gallus*, or chicken, which is our bird species in this study (figure 26). Looking at shotgun sequences from chromosome 26,

A



B

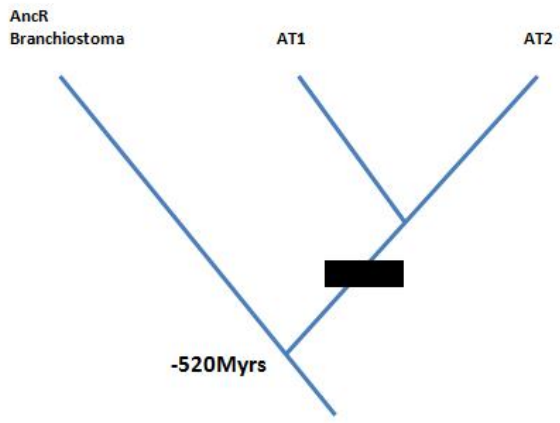


Figure 27. Evolution of AT receptors. A. Phylogeny of chordate AT receptors. **B.** Interpretation of the phylogenetic tree. AncR is the ancestral gene found in *Branchiostoma floridae*. A duplication event (represented by a black rectangle) led to emergence of two different receptors, AT_1 and AT_2. Methods to calculate and display the phylogeny are the same as in figure 24.

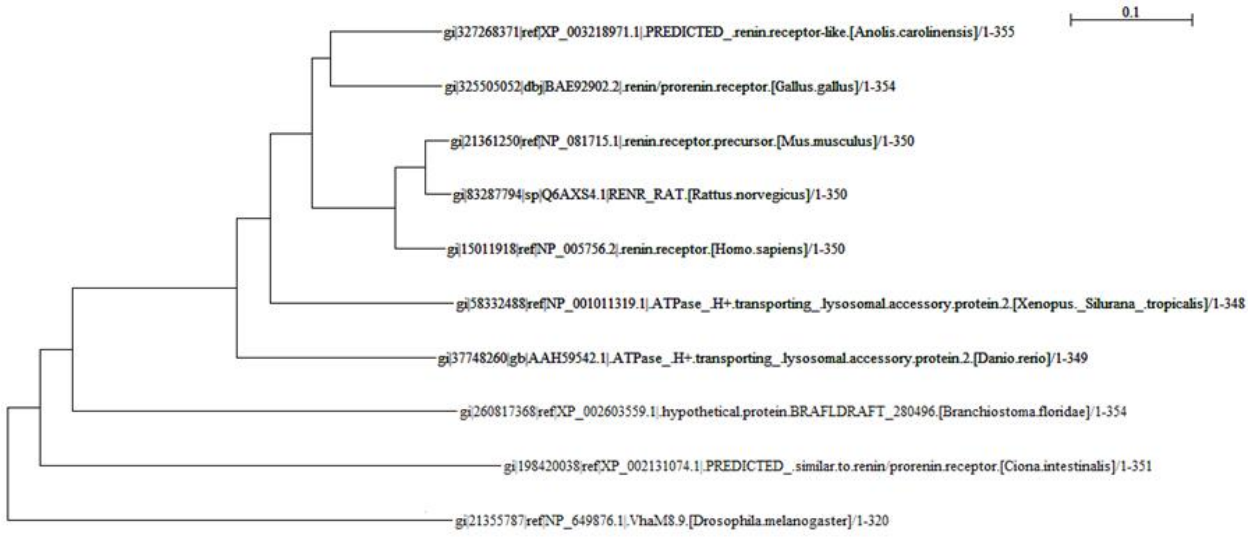


Figure 28. Evolution of (P)RR, the prorenin and renin receptor. Methods to calculate and display the phylogeny are the same as in figure 24.

though, we could find a sequence. These renin sequences probably have not been yet assembled to the *Gallus gallus* genome.

3.3.4. Evolution of RAAS targets

Renin has appeared relatively recently in evolutionary times compared to other components of RAAS, and especially its effectors.

3.3.4.1. AT_1 and AT_2

The two receptors target of angiotensin II, AT_1 and AT_2, are products of gene duplication as ACE are. We could identify an ancestral sequence of the two proteins in *Branchiostoma* (named AncAT, see figure 22A and figure 27). We could not find any homologue in *Ciona* though. The result suggests that the duplication event occurred after the divergence of tunicates from chordates. Here also, the actual binding of AncAT with ligand angiotensin II probably evolved later since angiotensinogen appeared later too (figure 22B). The ligands might not have been able to bind its modern receptor at first but evolved to do so later in evolutionary times. The angiotensin II receptor family probably has ancestral functions that will have to be explored in order to understand better the origins of the RAAS.

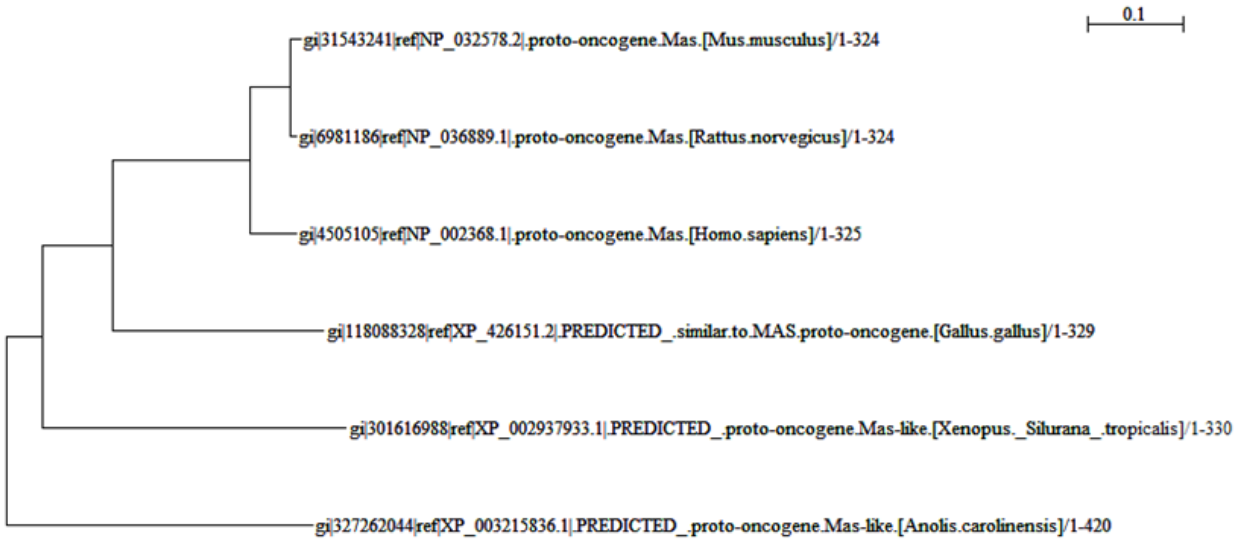


Figure 29. Evolution of the Mas receptor. Methods to calculate and display the phylogeny are the same as in figure 24.

3.3.4.2. (P)RR

(P)RR, also known as ATP6AP2, is one of the oldest proteins of the set, as it has orthologs in all organisms of our selection, including *Drosophila melanogaster* and *Caenorhabditis elegans* (figure 28). This indicates that the protein existed a long time before the emergence of the RAAS. We find the length of the protein to be very similar in all taxa, which suggests that the architecture of the gene might be very well conserved. While ATP6AP2 has a H+-ATPase function, in higher taxa, this protein evolved an additional function: to bind prorenin and renin [164]. There is still no evidence functionally relating the 'old' purpose of this protein, that is its H^+-ATPase function, to its 'new' function of binding renin and prorenin [191].

3.3.4.3. MAS

Among the different components of the system regulating hypertension, the Mas gene is the one that apparently appeared the latest. Contrary to all other genes studied, it is not present in fish and seems to appear for the first time in amphibians, here *Xenopus tropicalis* (figure 29). Mas is known to be bound by a degradation product of angiotensins, namely angiotensin (1-7) (figure 20). Contrary to angiotensin II, this peptide diminishes blood pressure and relieves cardio-vascular diseases such as thrombosis and atherosclerosis [192]. As a consequence, after the emergence in ancestral cartilage or bony fishes of the core components of the RAAS associated to a regulation of hypertension like we know today for terrestrial vertebrates, apparition of the

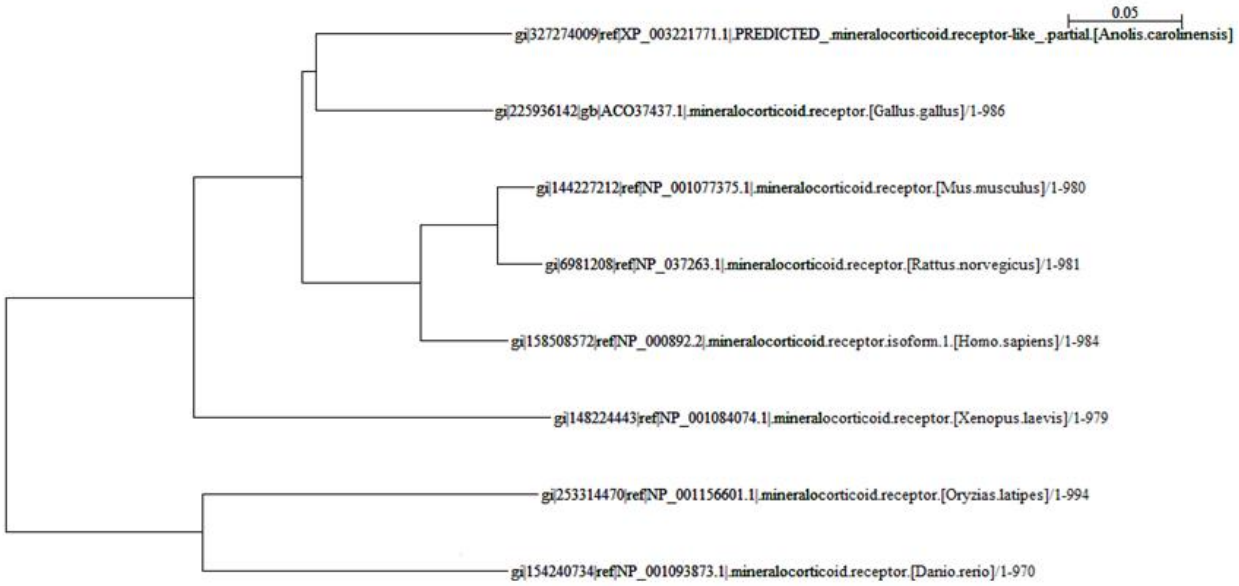


Figure 30. Evolution of the mineralocorticoid receptor. Methods to calculate and display the phylogeny are the same as in figure 24.

Mas gene increased the complexity of the system by creating a new potential way to modulate the effects of products of RAAS on vasoconstriction.

3.3.4.4. Mineralocorticoid receptor

The mineralocorticoid receptor was found to be present from fish to human (figure 22A and 30). We could not confirm its presence in *Callorhinchus milii* though it was identified in the skate, which is another type of cartilage fish [193]. Therefore, the MR probably predated angiotensinogen and renin. The ancestral gene of MR (able to bind a precursor of aldosterone, the 11-deoxycorticosterone) was duplicated in an ancestral organism (probably in an ancestral cartilage fish) before the bony fishes emerged [193,194]. One of the copies of the MR then evolved to become a glucocorticoid receptor (GR), which binds cortisol. In the fish, MR was still specialized in binding 11-deoxycortisone, and later evolved to bind aldosterone in a higher taxon. If the ancestral MR present in cartilage fish proves to have been involved in an early version of the RAAS, then it would mean that volume regulation is older than stress-related responses (stress in the sense of high vertebrates behavior). We are still left with the question of when the receptor appeared. Again, genome sequencing of more taxa such as lamprey or myxine would help solving this issue.

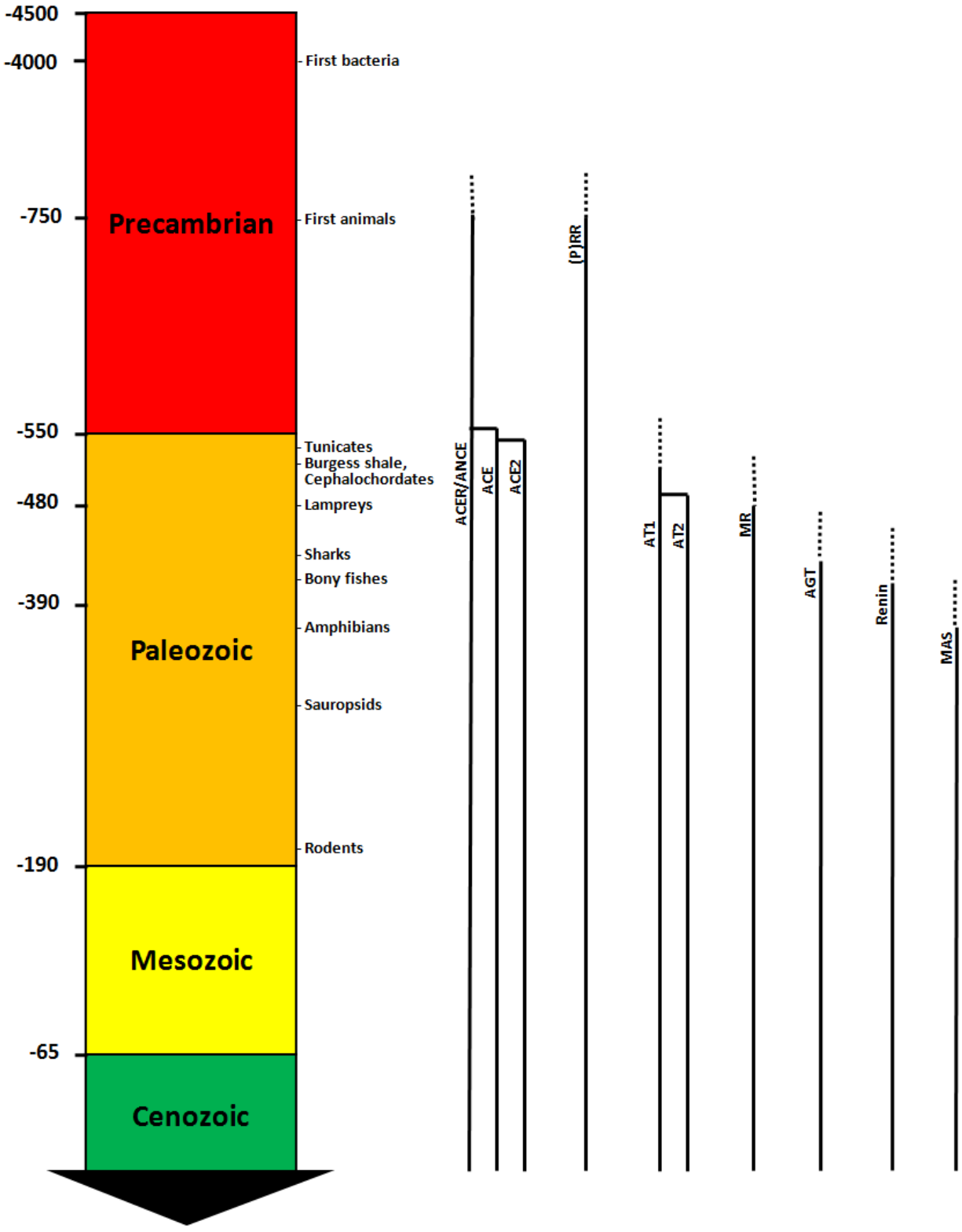

-4500
-4000
-750
-550
-480
-390
-190
-65
Precambrian
Paleozoic
Mesozoic
Cenozoic
First bacteria
First animals
Tunicates
Burgess shale, Cephalochordates
Lampreys
Sharks
Bony fishes
Amphibians
Sauropsids
Rodents
ACER/ANCE
ACE
ACE2
(P)RR
AT1
AT2
MR
AGT
Renin
MAS

Figure 31. Time-line of the emergence of the RAAS. Left geological eras and a time-line (scale in millions of years). While most genes appeared in the early Paleozoic, others might have emerged earlier in the Precambrian era and were adapted for their use as part of the RAAS. ACE is one such example and might have evolved from an initial developmental function to physiological actions on volume regulation in vertebrates.

3.4. Conclusion

The main finding of our study is that most of the components of the RAAS appeared at the time chordates and tunicates emerged. The main components were present in bony fishes, except the Mas oncogene, which is first seen in amphibians. Angiotensinogen, the precursor of the active peptides of the RAAS (angiotensin II and to a minor extent angiotensin (1-7)), appeared in cartilage fishes. Our analysis included the (P)RR, although it does not play a direct role in volume regulation. As the sequence of this receptor evolved from a vacuolar ATPase, it is probably much older than other RAAS-related proteins [164].

A lesson from our study is that the targets of a pathway appeared earlier than their substrates, hinting that evolution preferentially happens on circulating agents, rather than on receptors. The emergence of the RAAS is the result of the construction of a peptidase core and receptors that bound an unknown substrate in an ancestral organism (cloud in figure 22B). Peptidases with a broad spectrum of potential targets were present earlier to the emergence of the RAAS. This is the case for instance of ACE, which used to participate in embryo formation and was therefore expressed in various tissues of developing invertebrates and early vertebrates to later participate in the regulation of blood pressure. AT_1 and AT_2 were present in chordates earlier than angiotensinogen and as a result might also have played a different role in more ancient taxa. At some point, ACE and ACE2 would have become adapted to process angiotensinogen. This processing later included an earlier cleavage by renin, and finally new downstream receptor targets, such as Mas and the mineralocorticoid receptor. In fishes, the mineralocorticoid receptor first used probably 11-deoxycorticosterone, to later bind its derivate, aldosterone, in higher vertebrates [195].

Production of renin could have been first happening in the whole adrenal gland and late migrate to the glomerulus, as suggests the embryology of bony fishes [196]. In an ancestor of bony fishes, upon change of blood volume, renin produced in the adrenal glands could stimulate the local production of angiotensin II, which induced vasoconstriction. Production of renin later could have become confined to glomerulus because it is a strategic place to act on blood volume, by checking directly how the body is managing water in the kidney, the main place for volume extraction out of the body.

The different phylogenies presented in our study allow to suggest a time-line for the emergence and evolution of the RAAS components, from Precambrian to present time (figure 31). The

emergence of the RAAS could have taken about 150 million years in the Paleozoic, followed by a 400 million year period of relative stability. Our study confirms previous findings that date the modern RAAS at -400 million years, the time when cartilage and bony fishes diverged [197]. The presence of an ancestral RAAS in jawless fishes such as lamprey is not to be excluded, as renin-like activity and presence of angiotensin II has been shown in these taxa [198]. But more evidence at both molecular and physiological levels has to be shown before raising any hypothesis, especially, the precise location of cells with renin-like activity. It is possible that lamprey produces angiotensin II or a peptide with similar physiological activity using pathways different to those used in higher taxa.

What evidence shows beyond any doubt is the presence of the RAAS in bony fish, and its absence in tunicates, yet the zone in between the two is still to be investigated in depth. In the future, a study of all intermediate animals from tunicates to sharks, lampreys and myxines should be conducted to be sure if the elements of the RAAS are present, if they are expressed and produced and where exactly. The location in the tree of life of the emergence of the system regulating hypertension remains unclear. Aside from more genomic data, we ask that the activity of putatively present renin and ACE be measured in vitro for these key taxa; and that the sequencing of angiotensin peptides be performed on blood samples in each case. With such future experiments, we would aim at discovering what were the earliest mechanisms of hypertension, and from which previous item they emerged from.

4. Methods to study the impact of mutations on proteins related to disease using structural and evolutionary information

4.1. Introduction

Protein mutations are a major force of evolution of living beings. They are the subject of numerous studies in modern biology because they cause genetic diseases, and are thought to be the driving force of cancer, as cancer cells evolve and create a highly-replicable tissue that can cause severe damage to the human body. As a consequence, there is a need for methods to describe the consequences of residue mutations on protein structure and function. In the recent years, computational tools have been developed to predict the impact of mutations on protein fitness, for example PolyPhen-2 [48], SDM [49] or MutationAssessor [5]. Such predictions have been shown to be very important in understanding the mechanism of genetic diseases [46].

We have used graphical visualization tools, either PDBpaint, a visualization tool developed by us [50] or PyMOL (PyMOL Molecular Graphics System software, DeLano Scientific, Palo Alto, California), in combination with these mutations predictors, to study the impact of mutation in three different proteins related to disease: two proteins related to neurodegenerative diseases (CRMP-1 in section 4.3. of this chapter and Huntingtin in section 2.6.) and one related to heart septum defects (MYH6, in section 4.4.). Particularly, PyMOL has functions devoted to mutagenesis with customized parameters, which allowed us guessing if mutations are likely to disrupt the structure of the three proteins. PDBpaint helped us in the establishment of a huntingtin model that was later used to predict the outcome of mutations in this protein.

4.2. PDBpaint, a visualization tool to display proteins using functional annotations

4.2.1. Introduction

We have created PDBpaint, a visualization tool that displays protein structures using custom sequence annotations [50]. The development of this tool was motivated by two different observations.

- The function of proteins is constrained by their three dimensional structure. This structure is very crucial in particular to proteins such as enzymes, which have to recognize targets with a high precision using a key-to-lock mechanism. One single mutation in an interface region can lead to a dramatic change in local protein shape that could prevent the enzyme from interacting properly with its targets. Overall, all protein domains, including their post-translational modifications, have a shape that is crucial to understand the function of the protein. For this reason, protein databases provide abundant structural information, which is stored as sequence annotations. This information is eventually used to build protein classification. Web servers providing such structural information include PDB [85], SCOP [199] and Dali [200]; this information can be combined with domain predictions, for instance from TMHMM 2.0 [201] and Pfam [65]. These resources are very convenient because the information they store is easy to handle and analyze. Unfortunately, most proteins sequences are not associated with a solved structure [202]. In addition, information from 3D outputs in the PDB is limited to structure information and does not show protein annotations from protein databases (e.g. predicted transmembrane helices), such as those mentioned before. Therefore, we thought useful to design a web tool that could allow the users to call structures from PDB and have annotations from Pfam and other databases displayed directly on them.
- Secondly, in the last decade, prediction of protein features from sequence has been one of the most addressed topics in computational biology. Researchers often need to analyse sequences from proteins whose structure is known in order to verify that predictions are in accordance with the experimental structural data. Though useful, this method of checking structures can be very time-consuming if the number of proteins to check is large. We personally experienced this issue during the development of ARD2 (see chapter 2 to read about the tool). A large number of structures had to be checked in order to verify if the alpha-solenoids predicted by ARD2 were mapping to alpha-solenoid structures found in the PDB database. Therefore, we decided to create a tool to facilitate this procedure, a tool that can display multiple structures in a single web page with custom annotations from the user.

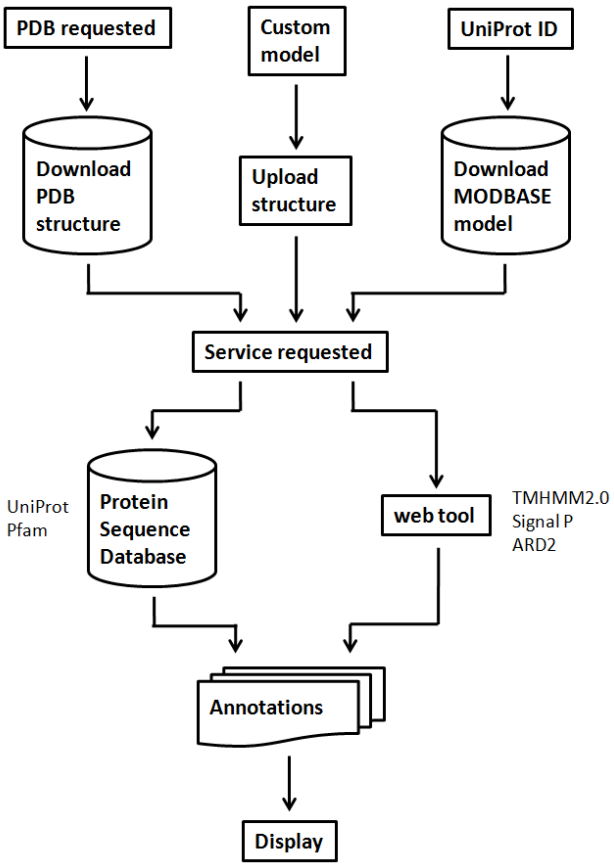


Figure 32. Flowchart of the PDBpaint webtool. The user can choose to request a sequence from PDB, upload a protein model from a local hard drive or give the UniProt ID of a protein sequence. In the first case the structure is downloaded from the PDB website. In the last case the MODBASE database is queried to find a structure for similar sequences to the query. In all cases the structure is saved on the local server of PDBpaint for a week before being automatically suppressed. In a next step, The user can also choose between 6 different webservices to annotate the protein structure. Finally the algorithm adds the custom annotations to display the protein structure with computed annotations associated. The process is repeated as many times as there are references in the query.

4.2.2. Functionalities of PDBpaint

During development, our agenda regarding specifications of PDBpaint was:

- To able the user to "paint" proteins with annotations from UniProt, Pfam and various webservices or, even more interestingly, to give custom annotations with custom colors. The format needed, very simple, had to relieve the user from learning how to write scripts in other tools such as PyMOL.

- To be able to provide either PDB identifiers (which means that no PDB structures will have to be downloaded on the user's hard drive) or import structures in PDB format, in case the user has a solved structure or a predicted model to study.
- To allow the user to display several structures at once on the same webpage, so the structures can be checked serially.

The PDBpaint algorithm follows three steps (see flow chart on figure 32). It first starts by loading the structure to study. If the user provides a PDB identifier, then the algorithm downloads directly the file from the PDB website. If the user has chosen a custom model, for example generated using a threading method such as I-TASSER [148,203], then the algorithm will attempt to upload the file selected by user and check that the format is correct. If the user provides a UniProt ID [204], the algorithm will send a request to the MODBASE database [205] and download protein models for the input sequence. The user can set manually the number of models to download, which cannot be more than 20.

In a second step, the program checks if the user has requested to use a webservice for annotation of the protein. Information can be acquired from two databases (Pfam protein domains [65] or UniProt sequence features) or predicted using web tools focused on specialized types of information (TMHMM2.0 to predict transmembrane domains [201], Signal P to predict signal peptides sequences [206] and ARD2 to predict alpha-solenoid proteins [51]).

In a next step, the custom annotations eventually requested by the user are stored. Finally, the output itself is built and showed on screen. The visual is displayed calling a graphical tool called Jmol (http://www.jmol.org; [207]). The entire procedure (figure 32) is repeated for each structure requested by the user, which are limited to 20 per query.

An example of a result page from PDBpaint is displayed on figure 33. In the first frame (figure 33.1), the user can set up the query. In the left dialog box, the user can provide one to 20 PDB IDs or UniProt IDs to display. In the second dialog box, the user can provide custom annotations. The format is very flexible and primarily consists in associating positions (such as 10 or 10-50) to colors in hexadecimal RGB format (for instance FF0000 for red). Then a menu allows to choose the webservice to use on the structures provided (for instance, Pfam). Finally, the user can set up the size of windows or the number of models to get from MODBASE for each of the UniProt IDs provided. In the second frame (figure 33.2), the user can manage custom models in PDB format. The first button allows choosing the file containing the model data. The next fields help to choose custom positions to tag and the webservice. Finally, if the user has already uploaded a model from the server, in which case a reference code has been provided, the structure can be recalled without uploading again by simply giving this reference code.

The output itself contains for each structure, a frame displaying the 3D visual of the protein, with eventual annotations showed in various colors (figure 33.3). General information about the protein (such as name, function, species) is provided (figure 33.4). Finally, the annotations from

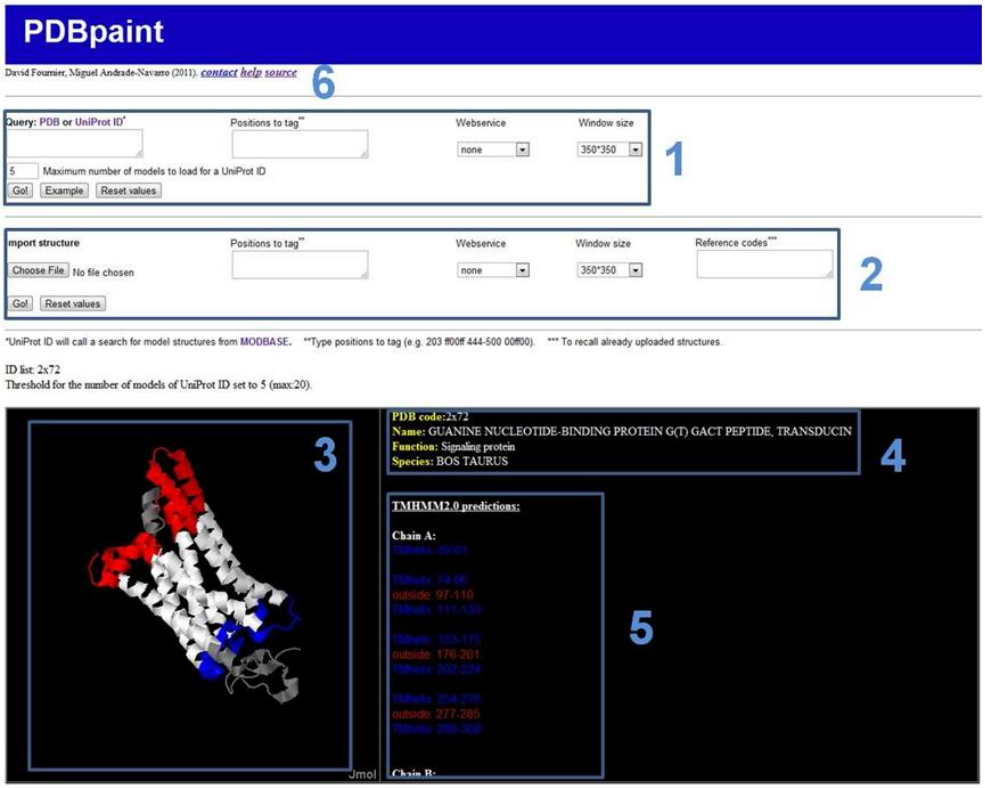


Figure 33. Example of a PDBpaint query. Bovine rhodopsin (PDB entry **2X72**) annotated according to prediction of TMHMM2.0, which detects transmembrane regions. 1-6: features of PDBpaint. 1: PDB/UniProt ID code input window, with different options (positions to tag, webservice and window size). 2: Custom structure upload window, with its options. 3: Output of the structure by Jmol. 4: Properties of the protein (from top to bottom, PDB code, name of protein, function of protein, species). 5: Legend of annotations performed by PDBpaint. In our example, webservice TMHMM2.0 for detection of transmembrane regions has been chosen. 6: Help page, to get some tips about PDBpaint.

the webservice chosen are displayed, with their associated colors, to make the visual output comprehensible (figure 33.5). Colors are chosen wisely for clarity of representation. Finally, on the top of the page (figure 33.6), a few links allow the user to send mail to the developers, to display a help page and to get the source code of PDBpaint available on the website of SourceForge, at http://sourceforge.net/projects/pdbpaint/files/.

To serve here as an example, we have used PDBpaint to study the localization of repeats in a region of huntingtin that contains an alpha-solenoid domain. As this region has no structure associated yet, we have used I-TASSER [148] to generate a model of the protein. Homology between proteins with alpha-solenoids is too weak to get an accurate model by similarity, so using threading-based prediction seemed more reliable than using predictions by MODBASE. We then uploaded our I-TASSER model in PDBpaint and used the ARD2 option in order to calculate and display positions with alpha-solenoid repeats on the 3D structure (figure 15). The

use of PDBpaint helped us to represent the agreement of the two predictions by ARD2 and I-TASSER (see section 2.6. for more details).

PDBpaint is available at http://cbdm.mdc-berlin.de/~pdbpaint with an example accessible via simply clicking on "Example".

4.2.3. Technical specifications of PDBpaint

The generation PDBpaint web pages was performed using a CGI file developed in Perl 5.10.1. The CGI script builds the webpage using HTML and elements of JavaScript. To display the graphical interface window of the web tool, which shows the 3D structure of the protein requested by the user, the Perl code calls Java (version 1.5) to run a program named Jmol (http://www.jmol.org; [207]). Jmol is a tool that displays protein structures and allows adding and representing annotations.

PDBpaint runs under an Apache 2.0 web server. Development of PDBpaint was performed on a Linux platform. Time to display a structure, including the annotations, varies depending on the size of the protein, but it is usually less than five seconds. For bigger proteins (1000 residues and more), it takes less than ten seconds.

4.2.4. Comparison with other tools

The problem of automating the mapping of features to structures is almost as old as the molecular graphic programs themselves. In 1994, Saqi and Sayle already presented a script to map protein motifs detected with regular expressions on PDB protein structure [208] using the RasMol viewer. The PDB website displays secondary structure annotations using Jmol. Standalone graphical molecular viewers like RasMol [209] or PyMOL (PyMOL Molecular Graphics System, Version 1.3, Schrödinger, LLC.) can also do this. Pfam [65] allows the display of Pfam domains of a protein in a dynamic protein 3D viewer (Jmol) when the protein's structure is available from the PDB. Aside from these well-known tools, several small web services allow to display a variety of features on 3D structures. Motif3D is an online tool that focuses on displaying protein motifs from the prints database on structures from the PDB database [210]. Amino acid conservation of a sequence using related homologs can be displayed on a 3D structure using ConSurf [211]. Suits of bioinformatics tools such as SRS 3D [212] or UTOPIA [213] can be used to display annotations on PDB structures like PDBpaint does; however these tools require local installation of software and have a significant learning curve.

4.2.5. Conclusion of section 4.2.

PDBpaint has been created to be an easy to use tool for the annotations of PDB structures according to features from sequence annotation collections or custom information provided by the user. The display of protein features for databases such as Pfam or UniProt directly on the 3D structure helps to better understand protein function. The user can also validate predictions from computational tools on multiple protein structures in a serial and convenient way, without needing to download any structures. Moreover, users that have just solved a new structure or built a model can see if their structure matches information stored in databases. We believe that PDBpaint can be very useful to researchers who want to design experiments, such as mutations, by being able to grasp all the structural and linear information at once in a comprehensive output.

An expansion of a poly-glutamine motif on N-terminal of protein huntingtin is one of the elements leading to diagnose Huntington's disease. This mutation is likely to cause accumulation of huntingtin fragments in neurons by aggregation of polyQ fragments, leading to the degeneration of these neurons, eventually causing

4.3. Study of deleterious mutations in huntingtin interacting protein CRMP-1

4.3.1. Introduction

An expansion of a poly-glutamine motif on N-terminal of protein huntingtin is one of the elements leading to diagnose Huntington's disease. This mutation is likely to cause accumulation of huntingtin fragments in neurons by aggregation of polyQ fragments, leading to the degeneration of these neurons, eventually causing the motor and cognitive impairments of patients baring the mutation [62]. As a consequence, finding drugs to suppress this aggregation has been thought to be a potential cure for the disease [214].

Huntingtin has been shown to contact more than a thousand of proteins [215], a vast majority of them waiting for experimental evidences in order to show their putative contribution to huntingtin wild-type or mutant phenotype. Collaborators from the group of Erich Wanker (Max Delbrueck Center for molecular medicine, Berlin Buch) have investigated 14 protein partners of huntingtin which they found deregulated in patients baring the disease ([53], data not shown). Four of the proteins had never been shown to be associated with Huntington's disease before, one of them being CRMP-1 (for collapsing response mediator protein-1), a protein expressed in brain which has an important function in neuronal development [216]. CRMP-1 was found by the laboratory of Erich Wanker to interact with kinase ROCK1, an enzyme that regulates actin and microtubules [217]. As a consequence, the down regulation of CRMP-1 in patients with Huntington's could lead to important perturbation of neurons. Unsurprisingly, CRMP-1 over-expression in cells over-expressing mutant huntingtin was found to impair polyQ aggregation (data not shown). More evidence has been provided, one of them exploiting the relation between structure and function of a protein. In this view, to gather more evidence for the relation between a protein and a function, mutations at specific sites can be performed to see the effect on the protein function.

Consequently, we designed a list of point mutations in various regions of the CRMP-1 protein, choosing locations that may have an impact on the structure. For each of these residues, we have predicted the consequence of the mutation for protein stability using evidence from different computational tools. To date, only one of these mutations, D408V, has been tested experimentally, and our collaborators have successfully proven to impair the function of aggregation reduction ([53], data not shown). This first result adds more evidence to other experiments that show that CRMP-1 is a potential suppressor of polyQ aggregation and that its down regulation in Huntington's disease participates in the development of the degeneration of neurons.

4.3.2. Methods

Impact of mutations for protein stability

Each of the positions selected were tested for mutation using tools Polyphen-2 [48], SDM [49], MUpro and MutationAssessor [5]. The structure used to run SDM and MUpro was found in the Protein Data Bank (PDB ID: 4B3Z, [218]). It is important to note that this crystal solved structure is in tetramer state and that two of the mutations were chosen in the interface between subunits, to see probable impact on protein function.

SDM and MUpro use data from the 3D structure to make predictions regarding protein stability. Polyphen-2, MUpro and MutationAssessor use information coming from sequence. Polyphen-2 is a tool that attributes two scores to a given mutation, HumDiv and HumVar. The HumDiv score is the probability of a mutation to cause disease according to records of mutations associated to Mendelian diseases stored in UniProt. HumVar calculates a similar score taking into account mutations causing disease and non-synonymous single-nucleotide polymorphisms that are not involved in disease and considered harmless. MutationAssessor is a tool that predicts the possible impact of mutation for the protein using conservation. More conserved residues are associated with increased functional importance, and their mutation is predicted to potentially cause disease with a higher probability. SDM predicts the thermodynamic effects of a mutation according to modification of the 3D structure and gives a prediction regarding the consequences for the stability of the protein. MUpro is a support-vector machine algorithm trained with 1615 site mutations from 42 proteins in order to predict the potential consequences of mutation for query proteins.

Conservation of CRMP-1 sequences

To study the conservation of residue 408, we aligned human CRMP-1 to various CRMP homologous sequences from vertebrates and invertebrates, as well as paralogous human CRMP sequences. CRMP sequences are very well conserved and we could observe 51% of identity (sequence similarity E-value = 1e-113) between CRMP-1 and *Drosophila* CRMP, while CRMP-1 and murine CRMP display 97% of identity (E-value = 0). Figure 34 represents an alignment of CRMP sequences for region 393-423, which contains site 408 whose mutation has been analyzed in more details by our collaborators (see section 4.3.3.2.). Alignment was performed using ClustalW [183].

4.3.3. Results and discussion

4.3.3.1. Design of CRMP-1 mutants

In order to study the relation between CRMP-1 structure and function, we have drawn a list of mutations to propose for experimentation (table 11). They were previously performed on mice mutants and some have been proven to impair capability of the murine CRMP to bind interacting

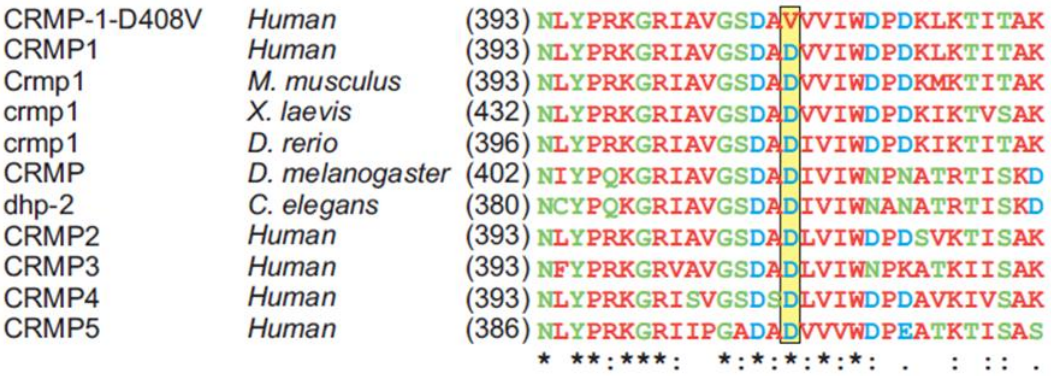


Figure 34. Multiple sequence alignment of the protein sequence of human CRMP-1 (corresponding to residues 393-423) in relation to paralogous human sequences and homologous sequences of other organisms. Conservation is displayed for each residue. A "*" is a position with 100% of conservation. A ":" is a conserved position with 2 different possible residues, while a "." is a conserved position with more than 2 different residues. Note that aspartic acid in position 408 of CRMP-1 highlighted in yellow is conserved in all sequences (a difference is only visible for the mutant). Alignment was performed using ClustalW.

partners ([216], mutants with an '*' in table 11). As a result, a hypothesis is that they could also impair human CRMP function and possibly have an influence on polyQ aggregation. Moreover, mutation D408V was added to the list. As a matter of fact, the residue at position 408 of the human CRMP-1 is homologous to a residue in the *Drosophila* CRMP whose mutation has been shown to alter the CRMP capacity to silence the effects of deleterious mutations in *Drosophila* mutants [219], so mutating this position in human CRMP-1 might possibly disrupt its ability to reducing polyQ aggregation. Figure 35 shows that most of our mutants are not located on the regions involved in CRMP-1 tetramerization (red residues and bottom face of the protein), to the exception of mutants 487 and 489, which are in the red interface region. All the mutations are present in the outer part of the protein, except D408V. Almost all the mutations previously shown to be harmful in mice or *Drosophila* were predicted to be harmful to the human CRMP by at least one tool, in some cases 2 (V51A, P52A). Positions 49 to 52 were predicted to be destabilizing by both SDM and MUpro. This could be explained by their patterns of conservation, which reflect the importance of the beta-sheet region in probable interactions of CRMP-1 with protein partners [216]. In any case, the positions that are predicted by far to be the most harmful are the four mutations on residues located in the region comprising positions 397 to 408 (residues depicted in figure 35B).

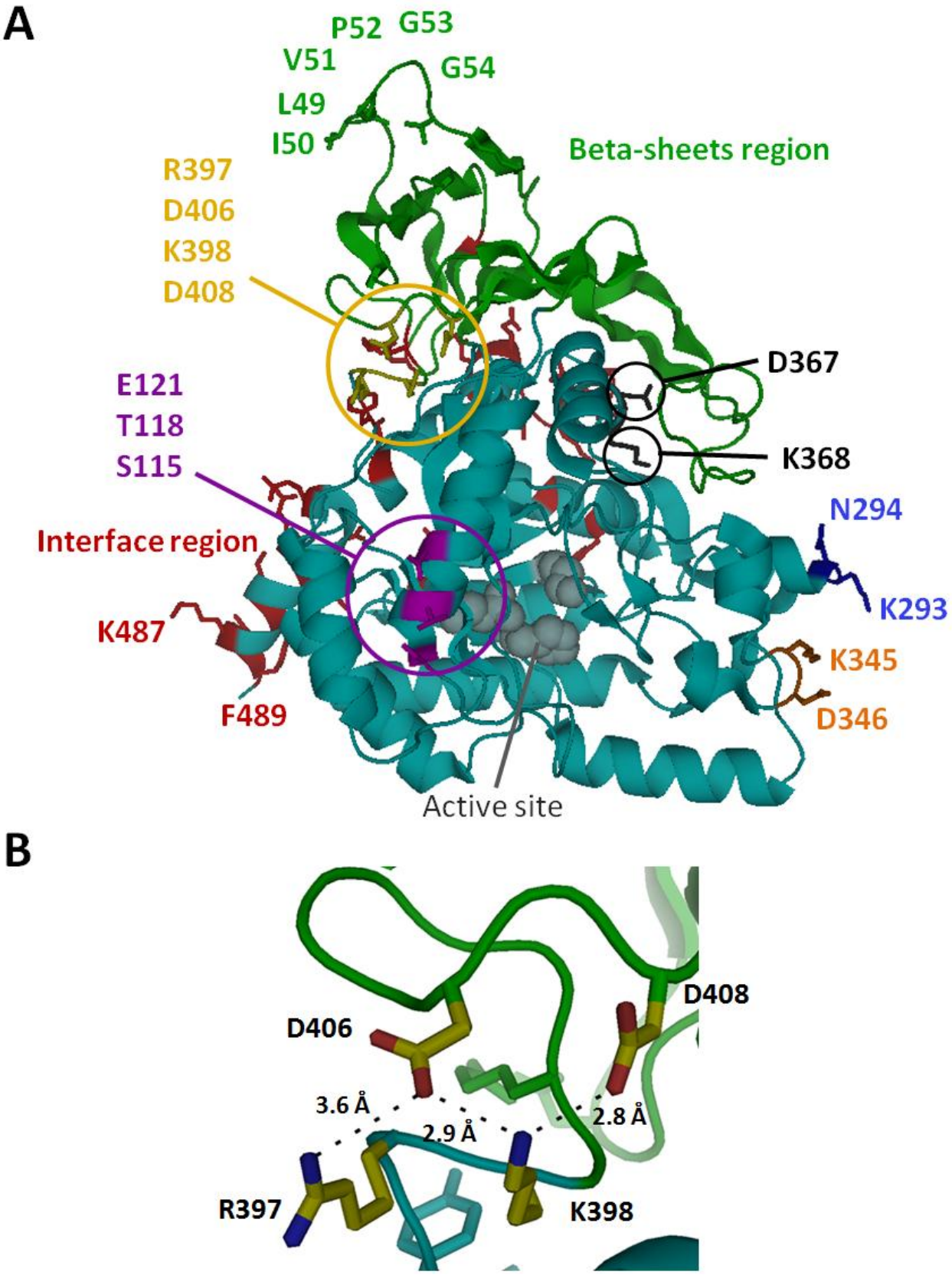

A
P52
G53
V51
G54
L49
I50
Beta-sheets region
R397
D406
K398
D408
D367
E121
T118
S115
K368
N294
Interface region
K293
K487
K345
D346
F489
Active site
B
D408
D406
3.6 Å
2.8 Å
2.9 Å
R397
K398

Figure 35. Location of different potential mutants for CRMP-1. A. List of mutants tested in table 11. The CRMP-1 protein is made of two lobes: the lower lobe, here in turquoise, and the upper lobe, in green. The lower lobe shows different regions: an interface region, depicted in red, that interact with a neighbor monomer, another interface region, which roughly corresponds to the bottom part of the molecule, and an active site represented here with grey spheres. Six positions are located in the beta-sheet region of the upper lobe (positions 49-54). Several positions were chosen on the outer part of the protein, outside interacting regions, and are potentially involved in interactions with other proteins (positions S115, T118, E121, K293, N294, D367, K368) Two mutations are associated with positions in one of the two interface regions (K487 and F489). Finally, four positions at the interface between upper and lower lobe are also shown (positions R397, K398, D406 and D408). **B.** Magnification of region of residue D408. Segment 397-408 presents three salt bridges that might be involved in the cohesion of upper and lower lobe. Residue D408 interacts with residue K397, which also interacts with D406. D406 is also interacting with R398. Mutations on all positions are predicted to be highly damaging as showed in table 11. The structure used to show the different mutations was found in the Protein Data Bank (PDB ID: 4B3Z, no publication associated). Software used for display is PyMOL (PyMOL Molecular Graphics System software, DeLano Scientific, Palo Alto, California).

4.3.3.2. Impact of mutation D408V on the function of CRMP-1

Convergence of evidences showing a probable impact of D408V on CRMP-1 structure

From all our candidates, the mutation predicted to be the most harmful of all is D408V (table 11). It was predicted to be deleterious in *Drosophila* with impact on the protective function of CRMP-1 against proteins with expanded polyQ [219]. Residue D408 is highly conserved from human to *Drosophila* (figure 34), which suggests that its mutation might influence the function of the human protein. D408 is located inside the protein (yellow coloration on figure 35A) and as such is probably not directly involved in protein-protein interactions. As shown in figure 35B, it is involved in a salt bridge with lysine of position 398 (distance: 2.8Å, a value close to the salt bridge optimal distance of 3Å [220]), which suggests that D408 is very likely to participate in the tight cohesion between the beta-sheet region (green color on figure 35A) and the rest of the

protein. As a result, mutation of this residue may cause the upper part of the protein to unfold and therefore to disrupt interaction of CRMP-1 with other proteins. We performed the same computational analysis on this mutation as the other listed on table 11. D408V was predicted to be highly damaging by all the different tools, HumVar score being 0.894 and classified as "possibly damaging", very close to threshold value 0.9 corresponding to "probably damaging". Prediction from MUpro and SVM shows impact for the stability of the protein, though they disagree on the outcome of the mutation (MUpro predicting it destabilizing and SVM stabilizing). High conservation of the residue predicted by MutationAssessor and strong association to disease both by HumDiv and HumVar confirms the potential harm of the mutation for the protein.

As it is predicted as a very good candidate for mutation, D408V functional impact was tested by collaborators from the group of Erich Wanker. The mutant was generated and its capacity to form stable homodimers in yeast-two-hybrid assays. Formation of homodimers is known to be a biological feature of the wild-type CRMP1 [221] and this was confirmed by correct binding when wild-type CRMP-1 is used both as bait and prey. Nevertheless, CRMP-1 mutant D408V was not able to form homodimers, while a weak binding was found between wild-type and mutant forms. These results mean that the protein is correctly produced but changes in the shape prevent it from correctly forming dimers.

Positions in the vicinity of D408 were investigated. Residues number 397, 398 and 406 form several salt bridges whose disruption might induce the same kind of protein misfolding that is caused by D408V, as they are all at the junction between the two lobes of CRMP-1. Residue K397 potentially forms a salt bridge with both D408 and D406 (distance of 2.9A, close to the theoretical value 3Å) and might be another critical position for protein stability. Nevertheless, mutation R397A is only predicted harmful by HumDiv score. MUpro and SDM predict the outcome of the mutation as being destabilizing. HumVar predict the mutation to be possibly damaging. Only MutationAssessor predicts no harmful consequence for the protein. Conservation of the position is maximal in vertebrates though it is a glutamine in invertebrates instead of an arginine (figure 34), which means that the salt bridges R398-D406 and R398-D408 did not exist at the time of emergence of the vertebrate lineage (more than 500 million years ago), and the residue might later have change to stabilize the structure. Residue D406 probably forms two salt bridges with neighbor residues R398 and K397 (distance of 3.6Å). Its mutation is predicted to be highly harmful by all the tools, with the exception of MUpro. It is by far the most harmful mutation predicted, along with D408V. Such a result can be explained by the potential involvement of the residue in two salt bridges. Finally, R397A mutation is predicted to be disease-related by HumDiv and HumVar scores, while SDM predicts it to be destabilizing. Despite a score of 0 (which refers to a harmless mutation) attributed by MutationAssessor, position D406 is perfectly conserved (figure 34) and may have functional importance. To conclude, it is very likely that disruption of any of the four residues 397, 398, 406 or 408 may cause important modifications of protein CRMP-1's structure and function.

Possible scenario explaining the impact of D408V on CRMP-1

Residue D408 is situated at the interface of upper and lower lobe of CRMP-1. The lower lobe (blue region on figure 35A and 35B) contains the two interface regions for interaction with the other subunits, as well as the active site. The upper lobe, or beta-sheet region (green region on figure 35A and 35B), has been proven to be involved in interaction with the COS7 protein [216]. We speculate that this region might be involved in important protein-protein interactions, some related to neuronal disorders. As a consequence, we predict that disruption of this region by mutation D408V or other mutations susceptible to disrupt the region might have dramatic consequences for the capacity of CRMP-1 to bind other proteins and function properly. Upon

mutation of D408V, the two lobes interaction could weaken and the upper lobe could unhook a little from its neighbor lobe. The disruption of the salt bridge between D408 and R397 by the D408V mutation may impair the stability of CRMP-1 oligomers, which are not capable anymore to suppress spontaneous mutant Htt misfolding and proteotoxicity. We speculate that such a mechanism could happen upon mutation of residues R397, K398, or D406 too, as they are neighbor sites of D408, also forming salt bridges at the interface of the two lobes. Structural and functional consequences of these mutations, along with mutations on sites 49 to 56, still have to be explored. For instance the impact of these mutations on the PPIs of CRMP-1 could be studied. Such experiments may further aid to understand the cellular dysfunctions leading to Huntington's disease and other neurological and neurodegenerative disorders.

4.3.4. Conclusion

We have predicted the possible outcome of mutations on the biological function of CRMP-1. One of these mutations has been tested and was proven by collaborators from the group of Erich Wanker to cause impairment of CRMP-1's capacity to reduce polyQ aggregation [53]. Other mutations are waiting for further exploration, especially located on the outer part of the beta-sheet region (green zone in figure 35). This region has been shown in mice to be important for the interaction of CRMP-1 to various proteins [216]. As impairment of huntingtin aggregation is believed to happen by direct interaction of CRMP-1 to polyQ fragments [53], we speculate that this region could be part of the interface zone and mutations of the positions 49 to 56 possible cause impairment of CRMP-1 ameliorative function. Further experimental validation is required in order to prove such a hypothesis.

4.4. Study of myosin mutations involved in cardiac septal defects.

4.4.1. Introduction

Secundum atrial septal defects (ASDII) account for 10% of cardiac malformation and have an incidence of 1 out of 1,500 people in the population. This type of malformation implies a communication between atria, which is normal in developing fetus, but is closed in the first three months after birth. Moreover, some familial and sporadic mutations have previously been shown to be factors of incidence for cardiac atrial septal defects. While some of these mutations involve transcription factors, some are present in the heavy chain of myosin VI (MYH6), a type of myosin found mainly in the atria [222] [223]. In collaboration with the group of Maximilian Posch of the Deutsches Herzzentrum Berlin and other teams, new mutations in MYH6 were recently identified in patients with ASDII in a study involving 31 patients with familial septal heart defects. MYH6 is an alpha-myosin only expressed in atria. A total of 13 sarcomeric genes were analyzed in these 31 patients. Among the mutations found by the sequencing of these genes, four had never been found before and were located in the coding sequence of myosin VI. Such mutations were neither present in 370 control patients nor in the database of the exome variant server (EVS) [224], which stores human mutations within coding sequences. In this collaborative work, our task was to explain the molecular and phenotypic effect of these mutations using structural and evolutionary information [52]. Mutations were R17H, C539R, K543R, and A1004S. The three first are located in the head of myosin heavy chain, known to be the place that pivots upon binding of actin, while A1004S is located in a region called the neck of the protein.

4.4.2. Methods

Analysis of conservation of human myosin heavy chains

To analyze the conservation of human myosin heavy chains, we ruled out using the sequences of orthologs from other species, as they are not divergent enough from human myosin. Identity ranged from 87% to the *Danio rerio* MYH6 sequence to 97% to the murine sequence. A multiple alignment of these sequences clearly indicated the high conservation of myosin VI across species (data not shown). In order to find more variability, we aligned myosin VI to 37 different human myosin chains found in the Entrez protein database. Alignment was performed using the MUSCLE algorithm [182] via the EBI server. Gene names are visible on supplementary figure S1 (see Appendix).

Representation of MYH6 mutants

To better study the impact of mutations for the structure of myosin VI, we analyzed a model of its 3D structure. As no structure is available for the human myosin VI, we searched the closest sequence of human myosin VI head with a solved structure associated in the NCBI sequence database. To achieve this, we used BLAST against PDB. The closest homologue sequence was

the myosin head domain of a chicken alpha-myosin heavy chain found in skeletal muscle (PDB ID: 2MYS, [225]). Identity between the two sequences is 81% for a coverage of 98% (E-value = 0). The region comprising residues 505 to 530 with the mutation sites C539 and K543 displayed 85% identity with the homologous sequence from chicken myosin for a coverage of 89% (E-value = 7e-23). In fine, as the similarity between human myosin VI and chicken skeletal myosin seemed acceptable, chicken myosin served as a model to represent myosin VI 3D structure. Mutation A1004 was not represented as it is outside of the sequence whose structure was determined. 3D structures were represented using PyMOL Molecular Graphics System software (DeLano Scientific, Palo Alto, California).

Computational analysis of outcome of mutations

The analysis is similar to the one which was performed to study CRMP-1. Please find details of the procedure in section 4.3.2.

4.4.3. Results and discussion

Study of mutant R17H

This mutation was found in three siblings who showed congenital heart defect, whether ASDII or AVSD (atrioventricular septal defect). Diverse evidence may explain why this mutation is associated to heart disease. Firstly, the residue is very well conserved across human myosins (see supplementary figure S2A in Appendix) in comparison with its neighbor residues, which aims at showing that R17 may have functional importance. Computational analysis of the outcome of R17H mutation reveals that it is the most harmful of the four mutations studied here, with convergence of prediction of our five scores (table 12). SDM predicts the mutation to be highly destabilizing, with a score of -2.47, which agrees with MUpro predicting the mutation to be destabilizing, with a score of -0.99. Polyphen-2 predicts the mutation as probably damaging and disease-causing, and MutationAssessor of high impact, probably due to high conservation of the position. Though structure of this region seems to be only partially resolved in our model (PDB ID: 2MYS, [225]), the residue is close to interacting residues D4, R24 and I25 known to contact the NH2 terminal side of actin [225]. Though R17 does not interact directly with these important residues, its mutation to a bulky residue such as histidine may weaken the interaction of myosin to actin (figure 36).

Study of mutants C539R and K543R

Mutant C539R was found in three generations of women from the same family, while K543R in five persons from the same family, two of which have a clear heart defect phenotype, the other being uncertain. Residue C539 is very well conserved across the different human myosins. It is conserved in myosin heavy chains class I, II, V and XV. In classes III, VII, IX and X, it is a serine, while it is an asparagine in class VI and a valine in class XVIII (supplementary figure S1

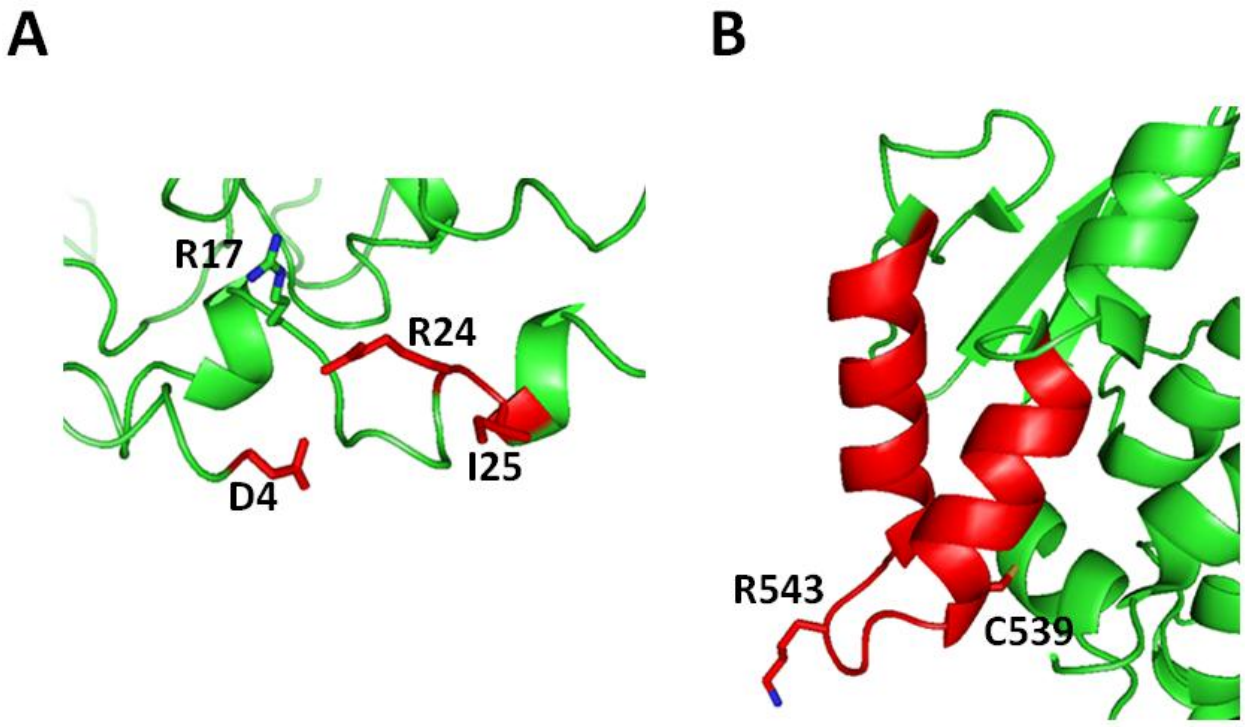


Figure 36. Localization of mutations on models for myosin VI heavy chain. A. Representation of residue R17, closed to interacting residues to NH2 terminal end of actin according to [225] (these residues are tagged in red: D4, R24 and I25). **B.** Representation of residues C539 and K543 in the local interface region to actin. The two residues are highlighted with side chains represented as sticks. Interface region is tagged in red and corresponds to residues 505 to 530 according to [225]. Space surrounding residue C539 is reduced and may imply harm for the protein upon mutation to a long amino acid such as it is the case of mutant C539R. Model used for both figures is the structure of a chicken alpha-myosin from skeletal muscle (PDB ID: 2MYS, [225]).

in Appendix). K543 is also well conserved and particularly across myosin heavy chains of class II. While C539R is predicted to be harmful by SDM, K543R is predicted to be harmless (table 12). This can be explained because K543 points outside and the structure of the model lacks information about interaction of this region with actin. We speculate that the same analysis performed on the structure of the actin/myosin complex, if it was available, may show probable harmful structural modifications of the protein upon mutation K543R. MutationAssessor predicts both mutations to have a high impact on function, probably due to good sequence conservation. Interestingly, polyphen-2 predicts C539R to be probably damaging with a score of 0.977 and K543R possibly damaging with a lower score of 0.644. 3D representation (figure 36B) helps to understand how C539 can be more damaging: K543 points toward the outside of the structure, while C539 is confined in a small space in the inside of the protein so mutation of this latter to a bulkier residue may impair protein stability. Nevertheless, both mutations are located in a helix-loop-helix region that is known to bind actin [225], which could explain why both are found

associated with cardiac defects. C539 is located in the loop of the helix-loop-helix region while K543 points outside. Mutation of the first one might disrupt myosin structure and therefore impair indirectly the correct binding of myosin and actin, while K543, probably in direct contact with actin surface residues, might impair strength of contact upon mutation.

Study of mutant A1004S

Mutant A1004S was also found to be associated with congenital heart defect in a girl and a boy who are sister and brother, while the father also bears the mutation, without a clear phenotype. Residue A1004 is well conserved across human myosin heavy chain sequences (supplementary figure S2B in Appendix). Polyphen-2 predicted the mutation to be benign, while MutationAssessor found it disruptive, but not as high as the other mutations. Contrary to the other 3 mutations, A1004S is located in the neck of myosin head protein, a region that is not involved in actin binding. The region surrounding the residue was predicted as alpha-helical by the tool Jpred3 [118]. The good conservation of residue A1004 is the only supportive piece of information to explain the possible link between the mutation and heart defect phenotype.

4.4.4. Conclusion

In this work we have analyzed four mutations associated with heart defect reported for the first time. All the mutations are located in a highly conserved region of the alpha-myosin motor domain, which is involved in myosin-actin interaction. This result shows that part of ASDII cases are related to mutations in sarcomere genes, such as MYH6. To link the genotype to the phenotype, and obtain a more mechanistic explanation for the underlying disease, we have shown that a combined analysis of conservation, modification of stability and structure is particularly useful.

4.5. Conclusion to chapter 4

In this chapter, we have shown that combining the analysis of conservation with structural information could be helpful in order to predict the outcome of mutations and the eventual consequence for protein functions and disease. Performing a simple alignment can be very informative about the relevant positions of a structure and help design mutants in order to test the importance of that site or domain for the function of the protein. Such a method is an example of the many possible connections between evolutionary concepts and medicine.

5. General conclusion

In the present work, we have conducted different studies to show the importance of taking into account facts from evolutionary biology in order to understand biological processes, particularly those related to human disease. We believe that the most complete way to know the function of a biological system is to consider the events that led to its appearance. Consequently, we have studied protein sequences in data from very different taxa to obtain mechanistic information on different biological systems relevant for human disease. In the case of alpha-solenoids, our nearly exhaustive analysis of available sequences helped to understand the potential link between this type of repeat and the increased demand for a protein transport system in complex taxa such as Eukaryota or Cyanobacteria. Such a statement could not have been supported by solely relying on data on human proteins. Differently, the study of ten carefully selected proteins in various taxa helped to build a scenario for the emergence of the system regulating hypertension. In this case, the evolutionary cues were helpful to show that the system was partially built up from ancestral proteins that necessarily had a different function before its emergence. Lastly, we used evolutionary information to obtain hints on the possible outcome of mutations of human disease related proteins by studying the conservation of particular amino acids in protein sequences in combination with protein structural information.

In the future, we expect that evolutionary concepts will be used as often as the common biology tools of today in order to explain biological processes. The diversification of experimental models of disease used in laboratories is one of the main elements that will eventually lead to a better understanding of how biological systems have emerged in evolutionary times. One of the various ways of including the evolutionary perspective in experimental work is to compare a system to its orthologous systems in different animal models. For instance, a biologist studying a given biological pathway in mammals, could do the same analysis in fish, and maybe in an invertebrate to try to see what is common and what is different. Such analyses could eventually shed light on some important facts that did not seem so important at first glance. To illustrate this, we cite a recent work from Zeron-Medina et al. who have identified potential cancer markers based on genes under positive selection in human populations more susceptible to certain types of cancer [226]. Such a result exemplifies that studying gene conservation can reveal functionality that is not visible with standard approaches. Reading such studies, we are confident that in a close future, evolutionary concepts will become major tools of biologists to understand processes happening in nature. The living world is not only a highly complex intricate network of processes, but before all it is the result of a story started several billions years ago, which has constrained for the most part the way living beings function today.

Summary

Evolution is a major actor of function of living beings. Studying biological processes with the perspective of an evolutionary biologist is important in order to have the most complete picture possible of the processes acting in nature. Following this idea, we have studied protein sequences to study two different biological systems. Such information was used to build evolutionary scenarios for our two questions.

We first studied alpha-solenoid repeat proteins. We have improved a method to detect such motives inside of protein sequences and applied this updated method to all sequences available in protein databases. The study of the distribution of such sequences in the tree of life shows that eukaryota are the taxons displaying the most this type of structure, as well as two groups of bacteria, cyanobacteria and planctomycetes. Importantly, the three groups of alpha-solenoid show limit similarity. We speculate that they appeared independently. Equally important, the three groups, eukaryota and the two bacteria taxons, are associated with increased cellular complexity versus classical bacteria groups. We hypothesized that the increased demand of protein important to protein transport and synthesis in living beings with compartmentalized cells induces a higher recruitment of alpha-solenoid proteins, as they require more complex protein machinery in order to be built. This high pressure could have occurred in the three groups independently, increased the recruitment of alpha-solenoid proteins.

The second evolutionary scenario we tried to put together is about the renin-angiotensin system which regulates hypertension in higher vertebrates. To perform this, we conducted a phylogenetic analysis of a dozen of proteins involved in the system, from higher vertebrates to invertebrates. We found that contrary to naïve thinking, some of the components of the system appeared before the set of the system, and had a complete different function, showing orthologue sequences in invertebrates. Some proteins, present in taxons with no regulation of hypertension such as *Drosophila*, were previously used for development a long time before being co-opted for homeostasis regulation in vertebrates. We could confirm the onset of the system around the appearance of cartilage fishes, around 400 million years ago.

Both analysis, of alpha-solenoid repeat proteins and protein sequences from the renin-angiotensin system, showed the importance of using evolutionary cues in order to better comprehend how living being work.

Aside from these evolutionary scenarios, we also used evolutionary along with structural information in order to study the impact of mutations for the structure of various proteins, and the relation of such mutations to disease and function. For these analyses, we have developed a tool called PDBpaint to visualize various annotations on the structure of proteins.

Zusammenfassung

Evolution prägt die Funktionsweise aller Lebewesen. Es ist notwendig biologische Prozesse aus der Perspektive eines Evolutionsbiologen zu betrachten, wenn man ein möglichst vollständiges Bild von den natürlichen Vorgängen erhalten möchte. In diesem Sinne haben wir Proteinsequenzen herangezogen um zwei verschiedene biologische Systeme zu untersuchen und um Evolutionsszenarios für unsere beiden Fragestellungen zu entwickeln.

Zunächst analysierten wir Proteine mit sich wiederholenden Alpha-Solenoidsequenzen. Wir verbesserten eine Methode zur Detektion solcher Motive innerhalb von Proteinsequenzen und wandten die verbesserte Methode auf alle Sequenzen an, die in Proteindatenbanken erhältlich waren. Die Untersuchung der Verteilung solcher Motive im phylogenetischen Stammbaum der Arten zeigte, dass besonders Eukaryonten und zwei Gruppen von Bakterien, die Cyanobakterien und Planctomyceten, diese Struktur tragen. Wichtig hierbei ist, dass die drei Gruppen von Alpha-Solenoidsequenzen nur begrenzte Ähnlichkeit aufweisen. Wir nehmen an, dass sie unabhängig voneinander entstanden sind. Genauso wichtig ist, dass die drei Gruppen, Eukaryonten und die zwei bakteriellen Taxa, ein erhöhtes Maß an zellulärer Komplexität aufweisen im Vergleich mit anderen bakteriellen Gruppen. Unsere Hypothese ist, dass Alpha-Solenoid Proteine durch den zunehmendem Bedarf an Proteinen für Transport und Synthese in Lebewesen mit kompartmentalisieren Zellenvermehrt entstehen, da sie eine komplexe Proteinmaschinerie benötigen, um gebildet zu werden. Dieser Druck könnte in allen drei Gruppen unabhängig bewirkt haben, dass Alpha-Solenoid Proteine in immer größerer Zahl hervorgebracht wurden.

Das zweite evolutionäre Szenario, das wir untersuchten, das Renin-Angiotensin System, reguliert Bluthochdruck in höheren Vertebraten. Wir unternahmen eine phylogenetische Analyse von zwölf Proteinen, die in das System involviert sind, von höheren Vertebraten bis hin zu Invertebraten. Wir fanden heraus, dass entgegen unseren Erwartungen einige Komponenten lange vor dem eigentlichen Regulationssystem auftraten, mit völlig anderen Funktionen, gezeigt anhand von orthologen Sequenzen in Invertebraten. Einige Proteine, die in Taxa ohne Regulation von Bluthuchdruck zu finden sind, so wie *Drosophila*, wurden schon lange vorher für die Regulation von Entwicklung verwendet. Wir konnten bestätigen, dass das System etwa gleichzeitig mit den Knorpelfischen entstand, ca. vor 400 Millionen Jahren.

Beide Analysen, sowohl die der Alpha-Solenoid Proteine als auch der Proteinsequenzen des Renin-Angiotensin Systems, haben gezeigt, wie wichtig es ist, Hinweise aus der Evolution in die Betrachtungen mit einzubeziehen, bei dem Versuch Lebewesen und ihre Funktionsweise, zu verstehen.

Neben diesen Evolutionsszenarios verwendeten wir evolutionäre gleichzeitig mit strukturellen Informationen um den Einfluss von Mutationen auf die Struktur verschiedener Proteine zu erforschen, sowie die Beziehung solcher Mutationen zu Krankheit und Funktion. Für solche Analysen haben wir PDBpaint entwickelt, ein Werkzeug zur Visualisierung von Annotationen in Proteinstrukturen.

Appendix

Figure S1. Multiple sequence alignment of human myosin heavy chains around C539 and K543 of myosin VI heavy chain. The yellow blocks represent alpha-helix in solved structures.

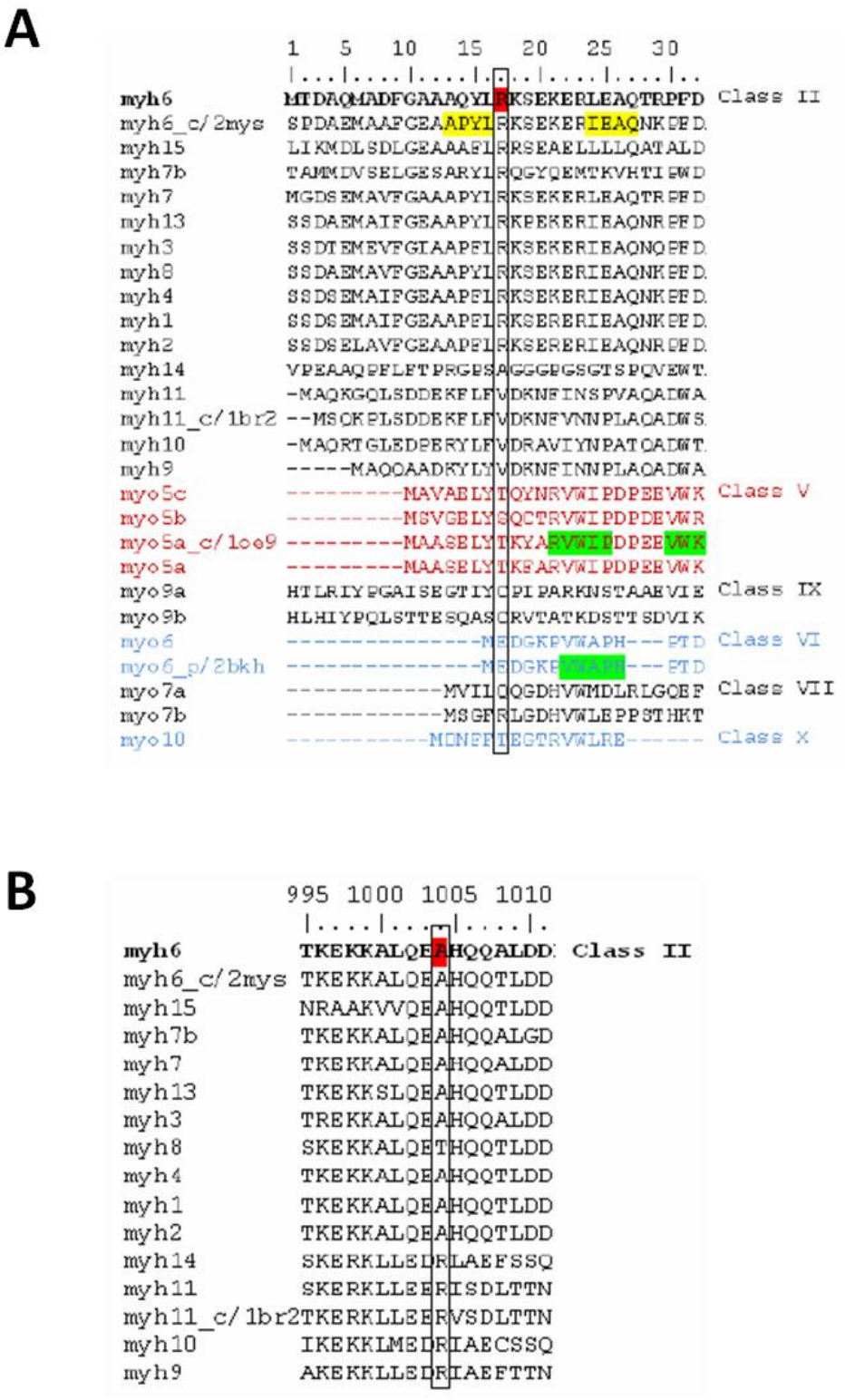


Figure S2. Multiple sequence alignment of human myosin heavy chains around R17 (A) and A1004 (B) of myosin VI heavy chain. The yellow and green blocks represent alpha-helix and beta-sheet in solved structures, respectively.

Table 1. Alpha-solenoid structures from PDB.

PDB ID	Chain	Information	Species
1B3U	A	Protein phosphatase 2A	*Homo sapiens*
1DL2	A	Mannosidase	*Saccharomyces cerevisiae*
1DVP	A	VHS & FYVE domains of HRS	*Drosophila melanogaster*
1E7U	A	PI3-kinase	*Sus scrofa*
1EE4	B	Importin-alpha	*Saccharomyces cerevisiae*
1F1S	A	Hyaluronate 2 lyase	*Streptococcus agalactiae*
1G6I	A	Mannosidase	*Saccharomyces cerevisiae*
1H2T	C	Cap-binding-complex (CBC)	*Homo sapiens*
1H6K	C	Cap binding complex	*Homo sapiens*
1HO8	A	Subunit H of the V-type 2 ATPase	*Saccharomyces cerevisiae*
1HU3	A	Initiation factor 4GII	*Homo sapiens*
1IAL	A	Importin-alpha	*Mus musculus*
1IB2	A	Pumilio domain	*Homo sapiens*
1IBR	B	Importin-beta/Ran complex	*Homo sapiens*
1JDH	A	Beta-catenin	*Homo sapiens*
1LDJ	A	Ubiquitin 2 ligase complex	*Homo sapiens*
1LRV	A	Leucine-rich repeat variant	*Azotobacter vinelandii*
1LSH	A	Lipovitellin	*Ichthyomyzon unicuspus*
1N52	A	Cap binding complex	*Homo sapiens*
1NA0	A	TPR motif	*unidentified*
1NF1	A	Neurofibromin	*Homo sapiens*
1O6P	B	Importin-beta	*Homo sapiens*
1OXJ	A	Smaug RNA-binding domain	*Drosophila melanogaster*
1OYZ	A	YibA	*Escherichia coli*
1OZN	A	Reticulon-4 receptor	*Homo sapiens*
1P22	A	Beta-TRCP1/Beta-catenin	*Homo sapiens*
1PAQ	A	Translation initiation factor 2B epsilon	*Saccharomyces cerevisiae*
1RZ4	A	Translation initiation factor	*Homo sapiens*
1T08	A	Beta-catenin	*reptilian*
1TE4	A	No known function	*Methanothermobacter thermautotrophicus*
1U6G	C	TBP-interacting protein	*Homo sapiens*
1UPK	A	Calcium-binding protein	*Homo sapiens*
1UW4	B	Regulator of nonsense transcripts	*Homo sapiens*
1VSY	4	Proteasome activator complex	*Saccharomyces cerevisiae*
1W63	I	AP1 clathrin adaptor core	*Mus musculus*
1WA5	B	Importin-alpha	*Saccharomyces cerevisiae*
1WA5	C	Importin-alpha re-exporter	*Saccharomyces cerevisiae*
1X9D	A	Mannosidase	*Homo sapiens*
1XG7	A	DNA lyase	*Pyrococcus furiosus*
1XI5	A	Clathrin D6 coat	*Bos taurus*
1XM9	A	Plakophilin 1	*Homo sapiens*
1XQR	B	HSPBP1 core domain	*Homo sapiens*
1XQS	A	HSPBP1 core domain	*Homo sapiens*
1Y2A	C	Importin-alpha	*Mus musculus*
1ZQ1	C	Amidotransferase	*Pyrococcus abyssi*
2B39	B	Third component of complement	*Bos taurus*
2BPT	A	Importin-beta	*Saccharomyces cerevisiae*
2C1M	A	Importin-alpha	*Mus musculus*
2DB0	B	Uncharacterized protein	*Pyrococcus horikoshii OT3*
2DQ6	A	Aminopeptidase	*Escherichia coli*

2E9H	A	Translation initiation factor	*Homo sapiens*
2FNO	A	Glutathione s-transferase	*Agrobacterium fabrum str. C58*
2FUL	E	Eukaryotic translation initiation factor	*Saccharomyces cerevisiae*
2FV2	C	Part of a transcription complex	*Homo sapiens*
2GRR	B	Ran GTPase-activating	*Homo sapiens*
2I2O	A	Translation of histone mRNA	*Danio rerio*
2IW3	B	Elongation factor	*Saccharomyces cerevisiae*
2JDQ	B	Importin-alpha	*Homo sapiens*
2NSZ	A	Inhibition of translation initiation	*Mus musculus*
2OF3	A	Microtubules organization	*Caenorhabditis elegans*
2OT8	A	Importin-beta	*Homo sapiens*
2OVR	A	Ubiquitination complex	*Homo sapiens*
2PHG	A	Transcription initiation factor IIB	*Homo sapiens*
2PN5	A	Thioester-containing protein	*Anopheles gambiae*
2PZI	A	Protein kinase	*Mycobacterium tuberculosis*
2QFC	A	Transcription activator	*Bacillus thuringiensis*
2QK1	A	Microtubules organization	*Saccharomyces cerevisiae*
2QNA	A	Importin-beta	*Homo sapiens*
2R7R	A	RNA-dependent RNA polymerase	*Simian rotavirus*
2RHS	C	Phenylalanyl-tRNAsynthetase	*Staphylococcus haemolyticus*
2SQC	A	Squalene cyclase	*Alicyclobacillus acidocaldarius*
2UY1	A	mRNA stability	*Encephalitozoon cuniculi*
2VGL	A	AP2 clathrin adaptor core	*Rattus norvegicus*
2VGL	B	AP2 clathrin adaptor core	*Homo sapiens*
2VSO	E	Translation initiation complex	*Saccharomyces cerevisiae*
2W3C	A	Globular head of a vesicular transport factor	*Homo sapiens*
2WXF	A	PI3-kinase	*Mus musculus*
2X1G	F	Importin 13	*Drosophila melanogaster*
2XQ0	A	Leukotriene hydrolase	*Saccharomyces cerevisiae*
2XWU	B	Importin	*Homo sapiens*
2Z6H	A	Beta-catenin	*Homo sapiens*
2ZRK	A	Recombinase	*Mycobaterium smegmatis str. MC2 155*
3A6P	A	pre-miRNA-binding importin-alpha	*Homo sapiens*
3AA0	A	Actin regulator	*Gallus gallus*
3AL0	C	Glutamyl-tRNAsynthetase	*Thermotoga maritima MSB8*
3B34	A	Aminopeptidase	*Escherichia coli K-12*
3B7S	A	Leukotriene hydrolase	*Homo sapiens*
3BCT	A	Beta-catenin	*Mus musculus*
3BG1	B	Nucleoporin	*Homo sapiens*
3BWT	A	RNA binding domain	*Saccharomyces cerevisiae*
3C5W	A	Protein phosphorylase 2A	*Homo sapiens*
3CHT	A	Oxygenase	*Streptomyces thioluteus*
3D3M	A	Translation initiation factor	*Homo sapiens*
3DAD	A	Actin regulator	*Homo sapiens*
3DRA	A	Geranylgeranyltransferase-I 2	*Candida albicans*
3EBB	A	Ubiquitin regulation	*Homo sapiens*
3FGA	B	Protein phosphorylase 2A	*Mus musculus*
3GAE	B	Ubiquitin regulation	*Saccharomyces cerevisiae*
3GE3	B	Toluene oxygenase	*Pseudomonas mendocina*
3GRL	A	Globular head of a vesicular transport factor	*Bos taurus*
3GS3	A	pre-mRNA processing	*Drosophila melanogaster*
3H3D	Y	RNA binding domain	*Drosophila melanogaster*
3H7L	A	Endoglucanase	*Vibrio parahaemolyticus*

3HHM	A	PI3-kinase	*Homo sapiens*
3I4R	B	Nucleoporin	*Homo sapiens*
3IBV	A	tRNA binding exportin	*Schizosaccharomyces pombe*
3IHY	C	PI3-kinase	*Homo sapiens*
3IP4	B	Glutamyl-tRNAamidotransferase	*Staphylococcus aureus*
3JRO	A	Nucleoporin	*Saccharomyces cerevisiae*
3JUI	A	Translation initiation factor	*Homo sapiens*
3JXY	A	DNA glycosylase	*Bacillus cereus*
3K62	A	mRNA binding protein	*Caenorhabditis elegans*
3K8P	D	Protein transport protein	*Saccharomyces cerevisiae*
3KND	A	Importin-alpha	*Mus musculus*
3L22	A	Membrane protein	*Bacteroides fragilis*
3L6A	A	Translation initiation factor	*Homo sapiens*
3L6X	A	Delta-catenin	*Homo sapiens*
3L9T	A	Uncharacterized protein	*Streptococcus mutans*
3LTJ	A	Consensus engineered sequence	*synthetic*
3LTM	A	Consensus engineered sequence	*synthetic*
3LY8	A	Transcriptional activator	*Escherichia coli K-12*
3M1I	C	Exportin	*Saccharomyces cerevisiae*
3O2Q	A	pre-mRNA processing	*Homo sapiens*
3O2T	A	pre-mRNA processing	*Homo sapiens*
3O4Z	D	Telomere regulator	*Saccharomyces cerevisiae*
3OBV	D	Actin regulator	*Mus musculus*
3OC3	A	Helicase	*Encephalitozoon cuniculi*
3OPB	A	Protein transport	*Saccharomyces cerevisiae*
3TJZ	B	Coatomer subunit	*Bos taurus*

Table 2. Training set of ARD2.

PDB id	Sequence id	Structure	Description
no	147899589	other	Condensin complex subunit 1
no	19113121	HEAT	Condensin, non-SMC subunit Cnd1 [*S. pombe*]
no	6323302	HEAT	Ycs4p [*S. cervisiae*]
no	148234026	HEAT	Condensin complex subunit 3
no	19075707	HEAT	Condensin, non-SMC subunit Cnd3 [*S. pombe*]
no	2132501	HEAT	Probable membrane protein YDR325w – yeast [*S. cerevisiae*]
no	336268148	HEAT	PDS5/BimD/Spo76 protein [S. macrospora k-shell]
no	4559410	Other	Androgen-induced prostate proliferative shutoff associated protein [*Homo sapiens*]
no	168025	HEAT	bimD [*Emericella nidulans*]
no	33088246	HEAT	Adherin Nipped-B [*Drosophila melanogaster*]
no	1353390	HEAT	DNA repair and meiosis protein Rad9 [*Coprinopsis cinerea*]
no	6320386	HEAT	Sccp2 [*S. cerevisiae*]
no	347834059	HEAT	Adherin, cohesion loading factor Mis4 [*S. pombe*]
1U6G_C	CAND1_HUMAN	HEAT	Cand1 – ubiquitin ligase inhibitor
no	27477070	HEAT	TATA-binding protein-associated factor 172 [*Homo sapiens*] 20% homologous to 1Z6A:A
no	6325175	HEAT	Mot1p [*S. cerevisiae*] 20% homologous to 1Z6A:A
no	6323074	Other	Stup2 [*S. cerevisiae*] 27% identical to 2QK1
no	19075285	Other	Microtubule-associated protein Dis1 [*S. pombe*]
no	6633953	other	KIAA0097 protein [*Homo sapiens*] Cytoskeleton-associated protein
no	17538165	HEAT	ZYGote defective : embryonic lethal family member (zyg-9) 18% identical to 2OF3 :A
no	5915683	Other	Tubulin-folding cofactor D
no	27806383	HEAT	Tubulin-specific chaperone D [*Bos taurus*]
no	6321700	Other	Apl6p [*S. cerevisiae*]
no	3885988	Other	AP-3 complex subunit beta-3A
no	6320444	Other	Sec26p [*S. cerevisiae*]
no	6324042	Other	Sec21p [*S. cerevisiae*]
2P8Q_A	IMB1_HUMAN	HEAT	Importin-subunit beta1 [*Homo sapiens*]
2AJA_A	Q5ZSV0_LEGPH	Other	Ankyrin repeat family protein Q5ZSV0 [*Legionella pneumophila*]

Table 3. Comparison of performances for ARM profile and ARD2.

PDB	Positive set	ARM	ARD2
1B3U_A	1	1	11
1EE4_B	1	0	6
1IAL_A	1	1	4
1IBR_B	1	0	3
1O6P_B	1	1	3
1OYZ_A	1	1	3
1OZN_A	1	0	3
1U6G_C	1	1	13
1WA5_B	1	1	6
1XQR_B	1	1	3
1XQS_A	1	0	3
1Y2A_C	1	0	4
2BPT_A	1	1	8
2C1M_A	1	0	4
2DB0_B	1	0	0
2IW3_B	1	1	4
2JDQ_B	1	1	6
2OF3_A	1	1	3
2OT8_A	1	0	10
2QK1_A	1	1	4
2QNA_A	1	0	6
2VGL_A	1	1	4
2VGL_B	1	1	8
2W3C_A	1	1	3
2XWU_B	1	1	3
2ZRK_A	1	1	10
3C5W_A	1	0	4
3GRL_A	1	1	4
3GS3_A	1	0	3
3KND_A	1	0	4
3LTJ_A	1	0	5
3LTM_A	1	0	5
3O2Q_A	1	1	1
3O2T_A	1	0	3
3OC3_A	1	0	3
3TJZ_B	1	1	1
1DL2_A	1	0	0
1DVP_A	1	0	0
1E7U_A	1	1	0
1F1S_A	1	0	0
1G6I_A	1	0	0
1H2T_C	1	1	0
1H6K_C	1	0	0
1HO8_A	1	1	0
1HU3_A	1	0	0
1IB2_A	1	0	0
1JDH_A	1	0	0
1LDJ_A	1	0	0
1LRV_A	1	0	0
1LSH_A	1	0	0

1N52_A	1	0	0
1NA0_A	1	0	0
1NF1_A	1	0	0
1OXJ_A	1	1	0
1P22_A	1	0	0
1PAQ_A	1	1	0
1RZ4_A	1	1	0
1T08_A	1	1	0
1TE4_A	1	0	0
1UPK_A	1	1	0
1UW4_B	1	1	0
1VSY_4	1	1	0
1W63_I	1	1	0
1WA5_C	1	1	0
1X9D_A	1	0	0
1XG7_A	1	0	0
1XI5_A	1	0	0
1XM9_A	1	1	0
1ZQ1_C	1	0	0
2B39_B	1	0	0
2DQ6_A	1	0	0
2E9H_A	1	1	0
2FNO_A	1	0	0
2FUL_E	1	1	0
2FV2_C	1	1	0
2GRR_B	1	1	0
2I2O_A	1	1	0
2NSZ_A	1	1	0
2OVR_A	1	0	0
2PHG_A	1	0	0
2PN5_A	1	0	0
2PZI_A	1	0	0
2QFC_A	1	0	0
2R7R_A	1	0	0
2RHS_C	1	0	0
2SQC_A	1	0	0
2UY1_A	1	0	0
2VSO_E	1	1	0
2WXF_A	1	1	0
2X1G_F	1	0	0
2XQ0_A	1	1	0
2Z6H_A	1	0	0
3A6P_A	1	1	0
3AA0_A	1	0	0
3AL0_C	1	0	0
3B34_A	1	0	0
3B7S_A	1	1	0
3BCT_A	1	1	0
3BG1_B	1	0	0
3BWT_A	1	1	0
3CHT_A	1	0	0
3D3M_A	1	1	0
3DAD_A	1	1	0

3DRA_A	1	0	0
3EBB_A	1	0	0
3FGA_B	1	1	0
3GAE_B	1	0	0
3GE3_B	1	0	0
3H3D_Y	1	1	0
3H7L_A	1	0	0
3HHM_A	1	1	0
3I4R_B	1	0	0
3IBV_A	1	1	0
3IHY_C	1	1	0
3IP4_B	1	0	0
3JRO_A	1	0	0
3JUI_A	1	1	0
3JXY_A	1	0	0
3K62_A	1	1	0
3K8P_D	1	0	0
3L22_A	1	0	0
3L6A_A	1	1	0
3L6X_A	1	1	0
3L9T_A	1	0	0
3LY8_A	1	0	0
3M1I_C	1	1	0
3O4Z_D	1	0	0
3OBV_D	1	1	0
3OPB_A	1	1	0
1C9B_A	0	1	0
1NF1_A	0	1	0
2RHQ_A	0	1	0

The last three structures highlighted in red are falsely predicted to contain alpha-solenoid repeats by InterPro.

Table 4. Functions of proteins with alpha-solenoids.

Function	Interaction	PDB ID	Protein name	Type	Reference
Protein transport	P/P	2JDQ	Importin subunit alpha-1	ARM	[227]
	P/P	1IAL	Importin subunit alpha-2/pendulin	ARM	[228]
	P/P	1IBR	Importin subunit beta-1/importin 90	HEAT	[106]
	P/P	2OT8	Transportin-1 (Importin beta-2)	HEAT	[229]
	P/P	1WA5	Re-exporter of importin subunit alpha	HEAT	[105]
TF coactivators	P/P	2Z6H	Catenin beta-1	ARM	[30]
	P/P	3OC3	Helicase MOT1	HEAT	[230]
Protein biosynthesis	P/N?	2IW3	Elongation factor 3A	HEAT	[231]
	P/N	3AL0	Glutamyl-tRNAsynthetase	HEAT	[101]
Enzyme scaffolding	P/P	2IAE	Protein Phosphatase PP2A subunit A	HEAT	[232]
	P/P	2PZI	Protein kinase PknG	HEAT	[233]
	P/P	2DQ6	Aminopeptidase N	HEAT	[94]
	P/P	3HHM	PI3Kalpha	HEAT	[103]
Substrate catalysis	P/P	2IAE	Protein Phosphatase PP2A subunit B	HEAT	[232]
Vesicle trafficking	P/P	1W63	AP1 Clathrin adaptor core	HEAT	[234]
	P/P	2VGL	AP2 Clathrin adaptor core	HEAT/ARM	[235]
	P/P	3GRL	p115 tether globular head domain	HEAT	[236]
Cytoskeleton	P/P	3OPB	She4p	HEAT	[237]
	P/P	2QK1	Protein STU2	HEAT	[238]
Ubiquitination/proteasome	P/P	1U6G	Cand1	HEAT	[108]
	P/P	1XQS	Hsp70-binding protein 1	ARM	[239]
	P/P	1VSY	Proteasome activator BLM10	HEAT	[240]
	P/P	3GAE	Ubiquitin fusion degradation 3	ARM	[241]
DNA damage	P/N	3JXY	Alkylpurine DNA glycosulaseAlkD	HEAT	[96]
	P/N	1XG7	N-glycosylase/DNA lyase	HEAT	[97]
micro-RNA processing	P/N	3A6P	Exportin-5	HEAT	[99]
mRNA processing	P/P	3O2Q	Symplekin	HEAT/ARM	[242]
	P/N	3K62	mRNA binding protein	PUMILIO	[100]
	P/P	1N52	Cap-binding protein	HEAT	[243]
	P/P	3D3M	Death associated protein 5 (DAP5)	HEAT	[244]
tRNA processing	P/N	3IBV	tRNA export factor	HEAT	[98]
Lipid metabolism	-	3DRA	Geranylgeranyltransferase-I	HEAT	[102]
	P/L	1LSH	Lipovitellin	HEAT	[104]
Tumor suppressing	P/P	1UPK	Calcium-binding protein 39	ARM	[245]
Other functions	P/P	2DB0	Hypothetical protein	HEAT	[246]
	P/P	2AJA	Ankyrin repeat protein	ANK	[247]
	P/P	2QFC	Virulence regulator	TPR	[91]
	-	3LTJ	Artificial protein	HEAT	[248]
	-	1LRV	Leucine-rich repeat protein	L-rich	[95]

Each protein is displayed with its PDB ID and the type of interaction its repeats are involved in. Though most of structures dock to proteins, we here point out the involvement of alpha-solenoids in protein-protein (P/P), protein-lipid (P/L) and protein-nucleic acid (P/N), either DNA or RNA. The diversity of function is broader than previously known. Structures referenced as "PDB" were recorded in the Protein Data Bank but have no publication associated.

Table 5. Human protein sequences from Swiss-Prot predicted to contain alpha-solenoids by ARD2.

UniProt ID	Entrez gene ID	Gene Symbol	NCBI description
P52294	3836	KPNA1	Importin
P52292	3838	KPNA2	Importin
O00505	3839	KPNA3	Importin
O00629	3840	KPNA4	Importin
O15131	3841	KPNA5	Importin
O60684	23633	KPNA6	Importin
A9QM74	402569	KPNA7	Importin
Q14974	3837	KPNB1	Importin
Q9UI26	51194	IPO11	Importin
O94829	9670	IPO13	Importin
Q8TEX9	79711	IPO4	Importin
O00410	3843	IPO5	Importin
O15397	10526	IPO8	Importin
Q96P70	55705	IPO9	Importin
Q96T76	64210	MMS19	DNA excision repair
O60518	26953	RANBP6	RAN binding
Q92973	3842	TNPO1	Transportin
O14787	30000	TNPO2	Transportin
Q10567	162	AP1B1	Adaptin
O75843	8906	AP1G2	Adaptin
O95782	160	AP2A1	Adaptin
O94973	161	AP2A2	Adaptin
P63010	163	AP2B1	Adaptin
O00203	8546	AP3B1	Adaptin
Q13367	8120	AP3B2	Adaptin
O14617	8943	AP3D1	Adaptin
Q9Y6B7	10717	AP4B1	Adaptin
P53618	1315	COPB1	Adaptin
Q9Y678	22820	COPG	Adaptin
Q9UBF2	26958	COPG2	Adaptin
Q6AI08	63897	HEATR6	
Q15021	9918	NCAPD2	Condensin
Q9BPX3	64151	NCAPG	Condensin
P42695	23310	NCAPD3	Condensin
Q86XI2	54892	NCAPG2	Condensin
Q96LV5	285905	INTS4L1	Integrator complex
Q2T9F4	644619	INTS4L2	Integrator complex
Q96HW7	92105	INTS4	Integrator complex
Q9BYG7	83876	MRO	Maestro
Q9NTI5	23047	PDS5B	Cohesion maintenance
Q8N122	57521	RPTOR	Complex of MTOR
P42345	2475	MTOR	mTOR
O00750	5287	PIK3C2B	PIK3
Q99570	30849	PIK3R4	PIK3
P78527	5591	PRKDC	Protein kinase
Q5VYK3	23392	KIAA0368	KIAA0368
Q92616	10985	GCN1L1	Amino-acid synthesis
Q86XA9	25938	HEATR5A	

Q9P2D3	54497	HEATR5B	
Q14008	9793	CKAP5	
Q7Z460	23332	CLASP1	Cytoplasmic linker protein
O75122	23122	CLASP2	Cytoplasmic linker protein
Q6KC79	25836	NIPBL	Nipped-B
Q8WVM7	10274	STAG1	Cohesin complex 3
Q8N3U4	10735	STAG2	Cohesin complex 3
Q6ZUX3	165186	FAM179A	
Q9Y4F4	23116	FAM179B	
Q86VP6	55832	CAND1	Cullin-associated
O75155	23066	CAND2	Cullin-associated
P30153	5518	PPP2R1A	Protein phosphatase
P30154	5519	PPP2R1B	Protein phosphatase
Q99460	5707	PSMD1	Proteasome 26S subunit
Q16401	5711	PSMD5	Proteasome 26S subunit
Q8N2F6	83787	ARMC10	Armadillo repeats
Q5W041	219681	ARMC3	Armadillo repeats
Q5T2S8	55130	ARMC4	Armadillo repeats
Q8IUR7	25852	ARMC8	Armadillo repeats
O00192	421	ARVCF	Armadillo repeats
Q6PI77	80823	BHLHB9	Helix-loop-helix domain
O14981	9044	BTAF1	TFIID-associated
Q68CQ1	374977	C1orf175	
Q6PJG6	221927	C7orf27	ATM activator
A6NGR9	642475	C8orf73	
A2RTY3	256957	C17orf66	
O75165	23317	DNAJC13	DnaJ (Hsp40)
P52306	5910	RAP1GDS1	GTP-GDP dissociation
Q86Y56	54919	HEATR2	
Q86WZ0	399671	HEATR4	
Q9NZL4	23640	HSPBP1	Heat shock protein binding
Q8NDA8	727957	HEATR7A	HEAT repeat containing 7A
O00458	3475	IFRD1	Developmental regulator
Q5S007	120892	LRRK2	Leucine-rich repeat kinase
Q8NG97	284383	OR2Z1	Olfactory receptor
Q29RF7	23244	PDS5A	Cohesion maintenance
Q8TF05	9989	PPP4R1	Protein phosphatase
Q6NUP7	57718	PPP4R4	Protein phosphatase
Q5JTH9	23223	RRP12	rRNA processing
Q9UHP6	27156	RTDR1	Rhabdoid tumor deletion
Q9BZR6	65078	RTN4R	Reticulon receptor
Q86VV8	25914	RTTN	Rotatin
O75533	23451	SF3B1	Splicing factor
Q9NRP7	27148	STK36	Serine/threonine kinase
Q92797	8189	SYMPK	Symplekin
Q9BTW9	6904	TBCD	Tubulin folding cofactor
Q9C0B7	79613	TMCO7	Transmembrane protein
Q14669	9320	TRIP12	Hormone receptor interactor
Q9H3U1	55898	UNC45A	Unc-45
O60763	8615	USO1	Vesicle docking
Q8N398	90113	VWA5B2	Von Willebrand domain

Table 6. Gene ontology terms found to be significantly enriched in human alpha-solenoids (99 sequences). The total human proteome is used as background (20,328 sequences). All sequences come from the Swiss-Prot database. GO analysis was performed using the DAVID tool [115].

Functions		
GO term	**genes**	**p-value***
Intracellular protein transport	29	4.4E-23
Mitotic cell cycle	15	1.5E-7
Golgi vesicle transport	6	8.8E-3
Regulation of defense response to virus	3	5.5E-2
Alternative splicing**	50	7.9E-2
Cellular localization		
GO term	**genes**	**p-value***
Nuclear pore	16	1.2E-17
Coated membrane	12	1.3E-12
Golgi apparatus	17	1.2E-4

*p-value is the chance to find the same enrichment by picking randomly the same number of proteins among the proteins of the background. Here, we show Benjamini-corrected p-values. Benjamini is a corrected p-value to reduce the false discovery rate. Significance was validated if the Benjamini value was equal or lower to 5e-2. **This gene ontology means that the enriched proteins are alternative spliced, not that they participate to the mechanism of alternative splicing itself.

Table 7. Human proteins newly identified as alpha-solenoids.

Name (Swiss-Prot accession number)	Function	Conservation[1]	ARD2	InterPro ARM
LRRK2. Leucine-rich repeat serine/threonine-protein kinase (Q5S007)	serine/threonine kinase	Dm Bf Ci	360; 408; 452; 494	163-619
RTTN. Rotatin (Q86VV8)	Axial rotation, left-right specification of body	Dm Bf Ci	1305; 1377; 1425;	1-954, 1422-1445, 1602-1691, 1846-1956, 2017-2225
TRIP12. E3 ubiquitin-protein ligase (Q14669)	ubiquitination	At ScDm Bf Ci	491; 532; 613	357-379, 436-938
UNC45A (Q9H3U1)	Co-chaperone of Hsp90, cell proliferation, muscle cell development, cytoskeletal function	Dm Bf Ci	448; 488; 537	89-350, 403-932
DNAJC13. Required for receptor-mediated endocytosis 8 (O75165)	Co-chaperone of Hsc70, receptor mediator endocytosis	At Dm Bf Ci	1783; 1826; 1865	445-1968, 1988-2191
IFRD1. Interferon-related developmental regulator 1 (O00458)	Embryonic development, muscle development	At ScDm Bf Ci	93; 136; 176	84-326

[1]Orthologs were searched for in Sc: *Saccharomyces cerevisiae*, At: *Arabidopsis thaliana*, Dm: *Drosophila melanogaster*, Bf: *Branchiostoma floridae*, Ci: *Ciona intestinalis*.

Table 8. Gene Ontology terms found to be significantly enriched in the genes uniquely interacting with the first alpha-solenoid region of huntingtin.

GO term	genes	p-value
Ribonucleoprotein	19	9.70E-3
Nucleotide-binding	43	4.70E-2
Ribosomal protein	13	2.90E-2
Protein biosynthesis	16	2.80E-2
Gene expression	23	1.90E-2
Metabolism of proteins	20	1.10E-2
Influenza infection	16	1.10E-2
3'UTR-mediated translational regulation	14	4.70E-2

P-value is the chance to find the same enrichment by picking randomly the same number of proteins among the proteins of the background. Significance was validated if the Benjamini-corrected value was equal or lower than 5e-2.

Table 9. Mutations designed for studies of huntingtin PPI.

Position	b-score	Mutation	SDM	MutationAssessor	Polyphen-2
321	**85.04**	**L321I**	**-0.37**	**1.725**	**0.998**
238	**83.31**	**S238T**	**0.15**	**1.1**	**0.346**
289	**80.48**	**S289A**	**0.16**	**-0.145**	**0**
276	**79.23**	**S276T**	**0.54**	**1.525**	**0.986**
287	**78.82**	**F287Y**	**-0.12**	**1.525**	**0.994**
241	**77.47**	**N241D**	**-0.21**	**1.61**	**0.911**
293	**76.31**	**N293D**	**-0.63**	**1.725**	**0.732**
288	**75.85**	**Y288A**	**0**	**1.12**	**0.998**
286	**75.82**	**Y286A**	**0.03**	**1.825**	**0.998**
279	**74.38**	**Q279A**	**0.22**	**1.79**	**0.565**
239	**74.22**	**F239L**	**0.24**	**-0.17**	**0.002**
242	**70.22**	**F242Y**	**-0.27**	**1.295**	**0.994**
346	67.98	E346D	0.17	0.17	0.002
285	65.81	Q285E	0.19	1.555	0.229
192	64.39	E192Q	-0.06	1.81	0.998
347	55.22	V347I	0.39	0	0.004
197	54.92	V197I	-0.17	-0.045	0.028
350	47.07	S350T	0.01	1.405	0.064
348	46.49	S348T	0.58	0.605	0.004
344	34.19	E344Q	-0.09	1.12	0.998
342	32.2	R342K	-0.16	1.01	0.111
343	29.16	K343A	-0.15	1.825	0.998
161	28.5	S161A	0.16	1.9	0.986
156	21.64	K156R	0.24	1.32	0.998
304	15.78	D304N	0.3	0.55	0.22
335	14.68	K335A	-0.15	1.435	0.998
115	10.11	V115I	0.35	0.345	0.001

Each position predicted to be involved in protein-protein interactions is associated to its Promate b-score, a mutation predicted harmless for the structure of the protein, the value of the deltaG given by SDM, the MutationAssessor score, and finally the polyphen-2 score. Mutations listed here have a $|\text{deltaG}| \leq 0.7$ and a MutationAssessor score inferior to 1.9. Top b-scores as predicted by Promate are highlighted in bold. All mutations tested include the following amino-acid replacements: amino-acid to glycine, amino-acid to alanine, amide to acid, acid to amide, R to K, K to R, W to H, H to W, L to D, D to L, S to T, T to S, F to Y, Y to F. In the end, a total of 119 mutations were tested for outcome on protein structure.

Table 10. Homologous sequences of human sequences of proteins of the renin-angiotensin-aldosterone system. The identifiers are those of GenPept. Dark grey boxes indicate that a sequence has no ortholog for a given species. Light grey boxes indicate sequences that are the ancestor of several sequences.

	Angiotensin I converting enzyme 1	Angiotensin I converting enzyme 2	Angioten-sinogen	Angiotensin II receptor type 1	Angiotensin II receptor type 2	Renin receptor	MAS1 oncogene	Mineralo-corticoid receptor	Renin
	ACE1	ACE2	AGT	AGTR1	AGTR2	ATP6AP2	MAS1	NR3C2	REN
Homo sapiens	4503273	11225609	4557287	119599314	23238240	15011918	4505105	158508572	4506475
Rattus norvegicus	6978757	58865588	19705570	51036661	6978473	83287794	6981186	6981208	148747255
Mus musculus	46559389	83582782	113461998	28893437	6680672	21361250	31543241	144227212	13676837
Gallus gallus	268370291	118084115	50741434	45384048	118089416	325505052	118088328	225936142	
Anolis carolinensis	327275269	327268244	327262107	327267001	327284065	327268371	327262044	327274009	327271277
Xenopus tropicalis	183986763	301617730	160773628	194018626	118404286	58332488	301616988	148224443	301622166
Danio rerio	326672223	55925554	63102191	125805883	61651858	37748260		154240734	47086317
Branchiostoma floridae	260781638 260799393 260836681	260799397		260805841 (AGTR)		260817368			
Ciona intestinalis	198420807	198418183				198420038			
Drosophila melanogaster	17137008 (ANCE) 17137262 (ACER)					21355787			
Caenorhabditis elegans	71985287 (acn-1)					17569199			

Table 11. Prediction of the outcome of different mutations of human CRMP-1 using computational tools.

Mutation	SDM	Polyphen-2 HumDiv	Polyphen-2 HumVar	MutationAssessor	MUpro
L49A*	-1.64	0.758	0.273	-	**-1**
I50A*	-1.18	0	0	-	**-1**
V51A*	**-2.18**	0.002	0.005	0.615	**-1**
P52A*	-0.91	0.009	0.003	1.895	**-1**
G53A*	**2.23**	0.001	0.001	1.32	0.51
G54A*	1.82	0.039	0.017	1.28	-0.52
V55A*	**-2.73**	0.002	0.009	0.02	-0.11
K56A*	-0.45	0.510	0.055	-	0.78
S115A	0.60	0.083	0.203	-	**1**
T118A	**2.48**	0	0	-1.225	**-0.89**
E121A	0.07	0.080	0.074	1.625	**-1**
K293A	-0.15	0.730	0.207	-	-0.38
N294A	-0.82	0.177	0.091	-	-0.31
K297A	1.06	0.032	0.073	-	0.39
K345A	-0.06	0.277	0.142	-	-0.15
D346A	0.09	0.328	0.291	2.38	**-1**
D367A*	**2.97**	**0.833**	0.438	2.235	-0.04
K368A*	1.06	0.122	0.066	-	0.11
R397A	-1.43	**0.955**	0.518	-	-0.77
K398A	-0.64	**1**	**0.999**	-	-0.25
D406A	**2.86**	**0.997**	**0.963**	**4.123**	0.28
D408A	1.81	**0.975**	0.805	**4.1**	**-0.99**
D408V**	**2.18**	**0.952**	0.894	**4.1**	**-0.99**
K487A*	0.93	0.081	0.054	-	**-1**
V488A*	1.57	0.021	0.024	1.3	**-1**
F489A*	1.01	0	0	-	**-1**

Colors used here are the same that are used in figure 35 to represent the mutations in 3D space. SDM and MUpro values marked in red classify mutations as highly disruptive, either stabilizing (positive value) or destabilizing (negative value). Polyphen-2 scores (HumDiv and HumVar) marked in red are classified as probably damaging. MutationAssessor values marked in red are associated with high impact of mutation, according to conservation. * indicates mutants that were shown to impair functional activity of murine CRMP in [216]. ** indicates the mutant that was shown to impair function of fly CRMP in [219]. - indicates that MutationAssessor could not perform the prediction for an unknown reason.

Table 12. Prediction of the outcome of four mutations associated with cardiac defects using different computational methods.

Mutation	SDM	HumDiv	HumVar	MutationAssessor	MUpro
R17H	-2.47	0.997	0.906	3.21	-0.99
C539R	0.71	0.977	0.968	4.625	-1
K543R	-0.09	0.644	0.767	2.67	-0.05
A1004S	-	0.024	0.035	1.72	-0.89

Details regarding the different tools and associated scores used in this table are explained in section 4.3.2. (Methods). No SDM score could be calculated for A1004S as no model was available for this region of myosin VI.

Bibliography

1. Aristotle History of animals. Harvard University Press.
2. Cosans CE (1997) Galen's critique of rationalist and empiricist anatomy. J Hist Biol 30: 35-54.
3. Bernard C (1957) An introduction to the Study of Experimental Medicine: Dover edition.
4. Losos JB, Arnold SJ, Bejerano G, Brodie ED, 3rd, Hibbett D, Hoekstra HE, Mindell DP, Monteiro A, Moritz C, Orr HA, Petrov DA, Renner SS, Ricklefs RE, Soltis PS, Turner TL (2013) Evolutionary biology for the 21st century. PLoS Biol 11: e1001466.
5. Reva B, Antipin Y, Sander C (2011) Predicting the functional impact of protein mutations: application to cancer genomics. Nucleic Acids Res 39: e118.
6. Hopf TA, Colwell LJ, Sheridan R, Rost B, Sander C, Marks DS (2012) Three-dimensional structures of membrane proteins from genomic sequencing. Cell 149: 1607-1621.
7. Williams GC, Nesse RM (1991) The dawn of Darwinian medicine. Q Rev Biol 66: 1-22.
8. Palmer AC, Kishony R (2013) Understanding, predicting and manipulating the genotypic evolution of antibiotic resistance. Nat Rev Genet 14: 243-248.
9. Snitkin ES, Zelazny AM, Thomas PJ, Stock F, Henderson DK, Palmore TN, Segre JA (2012) Tracking a hospital outbreak of carbapenem-resistant Klebsiella pneumoniae with whole-genome sequencing. Sci Transl Med 4: 148ra116.
10. Moxon ER, Rainey PB, Nowak MA, Lenski RE (1994) Adaptive evolution of highly mutable loci in pathogenic bacteria. Curr Biol 4: 24-33.
11. Bush RM, Bender CA, Subbarao K, Cox NJ, Fitch WM (1999) Predicting the evolution of human influenza A. Science 286: 1921-1925.
12. Censorinus (2007) The Birthday Book (De die natali Liber) (1st complete English translation): Chicago: University of Chicago Press.
13. de Maupertuis P-L (1750) Essai de Cosmologie: Amsterdam.
14. de Lamarck J-B (1809) Philosophie zoologique: Edition Dentu.
15. Darwin CR, editor (1859) On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life. London: John Murray ed.
16. Fournier D, Luft FC, Bader M, Ganten D, Andrade-Navarro MA (2012) Emergence and evolution of the renin-angiotensin-aldosterone system. J Mol Med (Berl) 90: 495-508.
17. Futuyma DJ (2009) Evolution: Sinauer Associates, Sunderland, Massachusetts.
18. Gerhart J, Kirschner M (2007) The theory of facilitated variation. Proc Natl Acad Sci U S A 104: 8582-8589.
19. Gerhart J, Kirschner M (1997) Cells, Embryos, and Evolution: Blackwell, Malden, MA.
20. Jacob F (1977) Evolution and tinkering. Science 196: 1161-1166.
21. Abzhanov A, Protas M, Grant BR, Grant PR, Tabin CJ (2004) Bmp4 and morphological variation of beaks in Darwin's finches. Science 305: 1462-1465.
22. Kirschner M, Gerhart J (1998) Evolvability. Proc Natl Acad Sci U S A 95: 8420-8427.
23. Ohno S (1970) Evolution by Gene Duplication: Springer.
24. Lynch M, Conery JS (2000) The evolutionary fate and consequences of duplicate genes. Science 290: 1151-1155.
25. Zhang J (2003) Evolution by gene duplication: an update. Trends Ecol Evol 18: 292-298.
26. Putnam NH, Butts T, Ferrier DE, Furlong RF, Hellsten U, Kawashima T, Robinson-Rechavi M, Shoguchi E, Terry A, Yu JK, Benito-Gutierrez EL, Dubchak I, Garcia-Fernandez J, Gibson-Brown JJ, Grigoriev IV, Horton AC, de Jong PJ, Jurka J, Kapitonov VV, Kohara Y, Kuroki Y, Lindquist E, Lucas S, Osoegawa K, Pennacchio LA, Salamov AA, Satou Y, Sauka-Spengler T, Schmutz J, Shin IT, Toyoda

A, Bronner-Fraser M, Fujiyama A, Holland LZ, Holland PW, Satoh N, Rokhsar DS (2008) The amphioxus genome and the evolution of the chordate karyotype. Nature 453: 1064-1071.
27. Jaillon O, Aury JM, Brunet F, Petit JL, Stange-Thomann N, Mauceli E, Bouneau L, Fischer C, Ozouf-Costaz C, Bernot A, Nicaud S, Jaffe D, Fisher S, Lutfalla G, Dossat C, Segurens B, Dasilva C, Salanoubat M, Levy M, Boudet N, Castellano S, Anthouard V, Jubin C, Castelli V, Katinka M, Vacherie B, Biemont C, Skalli Z, Cattolico L, Poulain J, De Berardinis V, Cruaud C, Duprat S, Brottier P, Coutanceau JP, Gouzy J, Parra G, Lardier G, Chapple C, McKernan KJ, McEwan P, Bosak S, Kellis M, Volff JN, Guigo R, Zody MC, Mesirov J, Lindblad-Toh K, Birren B, Nusbaum C, Kahn D, Robinson-Rechavi M, Laudet V, Schachter V, Quetier F, Saurin W, Scarpelli C, Wincker P, Lander ES, Weissenbach J, Roest Crollius H (2004) Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype. Nature 431: 946-957.
28. Marcotte EM, Pellegrini M, Yeates TO, Eisenberg D (1999) A census of protein repeats. J Mol Biol 293: 151-160.
29. Grinthal A, Adamovic I, Weiner B, Karplus M, Kleckner N (2010) PR65, the HEAT-repeat scaffold of phosphatase PP2A, is an elastic connector that links force and catalysis. Proc Natl Acad Sci U S A 107: 2467-2472.
30. Xing Y, Takemaru K, Liu J, Berndt JD, Zheng JJ, Moon RT, Xu W (2008) Crystal structure of a full-length beta-catenin. Structure 16: 478-487.
31. Kajava AV (2012) Tandem repeats in proteins: from sequence to structure. J Struct Biol 179: 279-288.
32. Jorda J, Xue B, Uversky VN, Kajava AV (2010) Protein tandem repeats - the more perfect, the less structured. FEBS J 277: 2673-2682.
33. Kim M, Abdi K, Lee G, Rabbi M, Lee W, Yang M, Schofield CJ, Bennett V, Marszalek PE (2010) Fast and forceful refolding of stretched alpha-helical solenoid proteins. Biophys J 98: 3086-3092.
34. Andrade MA, Perez-Iratxeta C, Ponting CP (2001) Protein repeats: structures, functions, and evolution. J Struct Biol 134: 117-131.
35. Buljan M, Bateman A (2009) The evolution of protein domain families. Biochem Soc Trans 37: 751-755.
36. Andrade MA, Bork P (1995) HEAT repeats in the Huntington's disease protein. Nat Genet 11: 115-116.
37. Andrade MA, Petosa C, O'Donoghue SI, Muller CW, Bork P (2001) Comparison of ARM and HEAT protein repeats. J Mol Biol 309: 1-18.
38. Gould SJ (2002) The Structure of Evolutionary Theory: Belknap Press.
39. Kugler JM, Lasko P (2009) Localization, anchoring and translational control of oskar, gurken, bicoid and nanos mRNA during Drosophila oogenesis. Fly (Austin) 3: 15-28.
40. Kieny M, Mauger A, Sengel P (1972) Early regionalization of somitic mesoderm as studied by the development of axial skeleton of the chick embryo. Dev Biol 28: 142-161.
41. Zhang J, Wagh P, Guay D, Sanchez-Pulido L, Padhi BK, Korzh V, Andrade-Navarro MA, Akimenko MA (2010) Loss of fish actinotrichia proteins and the fin-to-limb transition. Nature 466: 234-237.
42. Silbernagl S, Despopoulos A (2003) Taschenatlas der Physiologie: Thieme.
43. Mayr E (1964) From Molecules to Organic Diversity. Fed Proc 23: 1231-1235.
44. Stuart GR, Oda Y, de Boer JG, Glickman BW (2000) Mutation frequency and specificity with age in liver, bladder and brain of lacI transgenic mice. Genetics 154: 1291-1300.
45. Waters LS, Minesinger BK, Wiltrout ME, D'Souza S, Woodruff RV, Walker GC (2009) Eukaryotic translesion polymerases and their roles and regulation in DNA damage tolerance. Microbiol Mol Biol Rev 73: 134-154.
46. Scriver CR (2007) The PAH gene, phenylketonuria, and a paradigm shift. Hum Mutat 28: 831-845.
47. Studer RA, Dessailly BH, Orengo CA (2013) Residue mutations and their impact on protein structure and function: detecting beneficial and pathogenic changes. Biochem J 449: 581-594.

48. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR (2010) A method and server for predicting damaging missense mutations. Nat Methods 7: 248-249.
49. Worth CL, Preissner R, Blundell TL (2011) SDM--a server for predicting effects of mutations on protein stability and malfunction. Nucleic Acids Res 39: W215-222.
50. Fournier D, Andrade-Navarro MA (2011) PDBpaint, a visualization webservice to tag protein structures with sequence annotations. Bioinformatics 27: 2605-2606.
51. Fournier D, Palidwor GA, Shcherbinin S, Szengel A, Schaefer MH, Perez-Iratxeta C, Andrade-Navarro MA (2013) Functional and genomic analyses of alpha-solenoid proteins. PLoS One 8: e79894.
52. Posch MG, Waldmuller S, Muller M, Scheffold T, Fournier D, Andrade-Navarro MA, De Geeter B, Guillaumont S, Dauphin C, Yousseff D, Schmitt KR, Perrot A, Berger F, Hetzer R, Bouvagnet P, Ozcelik C (2011) Cardiac Alpha-Myosin (MYH6) Is the Predominant Sarcomeric Disease Gene for Familial Atrial Septal Defects. PLoS One 6: e28872.
53. Stroedicke M, Bounab Y, Yigit S, Strempel N, Chaurasia G, Friedrich RP, Li S, Riechers S-P, Russ J, Nicoletti C, Schnoegl S, Fournier D, Graham RK, Hayden MR, Sigrist S, Bates G, Priller J, Andrade-Navarro MA, Futschik ME, Wanker EE Interaction network filtering identifies CRMP-1 as a modifier of mutant huntingtin misfolding and neurotoxicity. Submitted.
54. Kobe B, Kajava AV (2000) When protein folding is simplified to protein coiling: the continuum of solenoid protein structures. Trends Biochem Sci 25: 509-515.
55. Palidwor GA, Shcherbinin S, Huska MR, Rasko T, Stelzl U, Arumughan A, Foulle R, Porras P, Sanchez-Pulido L, Wanker EE, Andrade-Navarro MA (2009) Detection of alpha-rod protein repeats using a neural network and application to huntingtin. PLoS Comput Biol 5: e1000304.
56. Hunter S, Jones P, Mitchell A, Apweiler R, Attwood TK, Bateman A, Bernard T, Binns D, Bork P, Burge S, de Castro E, Coggill P, Corbett M, Das U, Daugherty L, Duquenne L, Finn RD, Fraser M, Gough J, Haft D, Hulo N, Kahn D, Kelly E, Letunic I, Lonsdale D, Lopez R, Madera M, Maslen J, McAnulla C, McDowall J, McMenamin C, Mi H, Mutowo-Muellenet P, Mulder N, Natale D, Orengo C, Pesseat S, Punta M, Quinn AF, Rivoire C, Sangrador-Vegas A, Selengut JD, Sigrist CJ, Scheremetjew M, Tate J, Thimmajanarthanan M, Thomas PD, Wu CH, Yeats C, Yong SY (2012) InterPro in 2011: new developments in the family and domain prediction database. Nucleic Acids Res 40: D306-312.
57. Peifer M, Berg S, Reynolds AB (1994) A repeating amino acid motif shared by proteins with diverse cellular roles. Cell 76: 789-791.
58. Hatzfeld M (1999) The armadillo family of structural proteins. Int Rev Cytol 186: 179-224.
59. Preker PJ, Keller W (1998) The HAT helix, a repetitive motif implicated in RNA processing. Trends Biochem Sci 23: 15-16.
60. Forwood JK, Lange A, Zachariae U, Marfori M, Preast C, Grubmuller H, Stewart M, Corbett AH, Kobe B (2010) Quantitative structural analysis of importin-beta flexibility: paradigm for solenoid protein structures. Structure 18: 1171-1183.
61. Cook AG, Conti E (2010) Nuclear export complexes in the frame. Curr Opin Struct Biol 20: 247-252.
62. Cattaneo E, Zuccato C, Tartari M (2005) Normal huntingtin function: an alternative approach to Huntington's disease. Nat Rev Neurosci 6: 919-930.
63. Zoncu R, Efeyan A, Sabatini DM (2011) mTOR: from growth signal integration to cancer, diabetes and ageing. Nat Rev Mol Cell Biol 12: 21-35.
64. Knutson BA (2010) Insights into the domain and repeat architecture of target of rapamycin. J Struct Biol 170: 354-363.
65. Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, Pang N, Forslund K, Ceric G, Clements J, Heger A, Holm L, Sonnhammer EL, Eddy SR, Bateman A, Finn RD (2012) The Pfam protein families database. Nucleic Acids Res 40: D290-301.

66. Letunic I, Doerks T, Bork P (2012) SMART 7: recent updates to the protein domain annotation resource. Nucleic Acids Res 40: D302-305.
67. Punta M, Rost B (2008) Neural networks predict protein structure and function. Methods Mol Biol 458: 203-230.
68. Liberton M, Howard Berg R, Heuser J, Roth R, Pakrasi HB (2006) Ultrastructure of the membrane systems in the unicellular cyanobacterium Synechocystis sp. strain PCC 6803. Protoplasma 227: 129-138.
69. Yeates TO, Kerfeld CA, Heinhorst S, Cannon GC, Shively JM (2008) Protein-based organelles in bacteria: carboxysomes and related microcompartments. Nat Rev Microbiol 6: 681-691.
70. Fuerst JA (2005) Intracellular compartmentation in planctomycetes. Annu Rev Microbiol 59: 299-328.
71. Gribskov M, McLachlan AD, Eisenberg D (1987) Profile analysis: detection of distantly related proteins. Proc Natl Acad Sci U S A 84: 4355-4358.
72. Andrade MA, Ponting CP, Gibson TJ, Bork P (2000) Homology-based method for identification of protein repeats using statistical significance estimates. J Mol Biol 298: 521-537.
73. Bucher P, Karplus K, Moeri N, Hofmann K (1996) A flexible motif search technique based on generalized profiles. Comput Chem 20: 3-23.
74. Biegert A, Soding J (2008) De novo identification of highly diverged protein repeats by probabilistic consistency. Bioinformatics 24: 807-814.
75. McLachlan AD, Stewart M (1976) The 14-fold periodicity in alpha-tropomyosin and the interaction with actin. J Mol Biol 103: 271-298.
76. Coward E, Drablos F (1998) Detecting periodic patterns in biological sequences. Bioinformatics 14: 498-507.
77. Gruber M, Soding J, Lupas AN (2005) REPPER--repeats and their periodicities in fibrous proteins. Nucleic Acids Res 33: W239-243.
78. Newman AM, Cooper JB (2007) XSTREAM: a practical algorithm for identification and architecture modeling of tandem repeats in protein sequences. BMC Bioinformatics 8: 382.
79. Jorda J, Kajava AV (2009) T-REKS: identification of Tandem REpeats in sequences with a K-meanS based algorithm. Bioinformatics 25: 2632-2638.
80. Heger A, Holm L (2000) Rapid automatic detection and alignment of repeats in protein sequences. Proteins 41: 224-237.
81. Szklarczyk R, Heringa J (2004) Tracking repeats using significance and transitivity. Bioinformatics 20 Suppl 1: i311-317.
82. Rojas R (1996) Neural networks: Springer-Verlag, Berlin.
83. Rumelhart DE, Hinton GE, Williams RJ (1986) Learning representations by back-propagating errors. Nature 323: 533-536.
84. Neuwald AF, Hirano T (2000) HEAT repeats associated with condensins, cohesins, and other complexes involved in chromosome-related functions. Genome Res 10: 1445-1452.
85. Rose PW, Beran B, Bi C, Bluhm WF, Dimitropoulos D, Goodsell DS, Prlic A, Quesada M, Quinn GB, Westbrook JD, Young J, Yukich B, Zardecki C, Berman HM, Bourne PE (2011) The RCSB Protein Data Bank: redesigned web site and web services. Nucleic Acids Res 39: D392-401.
86. Striegl H, Andrade-Navarro MA, Heinemann U (2010) Armadillo motifs involved in vesicular transport. PLoS One 5: e8991.
87. Schmitt E, Panvert M, Blanquet S, Mechulam Y (2005) Structural basis for tRNA-dependent amidotransferase function. Structure 13: 1421-1433.
88. Perez-Iratxeta C, Palidwor G, Andrade-Navarro MA (2007) Towards completion of the Earth's proteome. EMBO Rep 8: 1135-1141.

89. Sigrist CJ, Cerutti L, de Castro E, Langendijk-Genevaux PS, Bulliard V, Bairoch A, Hulo N (2010) PROSITE, a protein domain database for functional characterization and annotation. Nucleic Acids Res 38: D161-166.
90. Cingolani G, Petosa C, Weis K, Muller CW (1999) Structure of importin-beta bound to the IBB domain of importin-alpha. Nature 399: 221-229.
91. Declerck N, Bouillaut L, Chaix D, Rugani N, Slamti L, Hoh F, Lereclus D, Arold ST (2007) Structure of PlcR: Insights into virulence regulation and evolution of quorum sensing in Gram-positive bacteria. Proc Natl Acad Sci U S A 104: 18490-18495.
92. Zamore PD, Williamson JR, Lehmann R (1997) The Pumilio protein binds RNA through a conserved domain that defines a new class of RNA-binding proteins. RNA 3: 1421-1433.
93. Michaely P, Tomchick DR, Machius M, Anderson RG (2002) Crystal structure of a 12 ANK repeat stack from human ankyrinR. EMBO J 21: 6387-6396.
94. Ito K, Nakajima Y, Onohara Y, Takeo M, Nakashima K, Matsubara F, Ito T, Yoshimoto T (2006) Crystal structure of aminopeptidase N (proteobacteria alanyl aminopeptidase) from Escherichia coli and conformational change of methionine 260 involved in substrate recognition. J Biol Chem 281: 33664-33676.
95. Peters JW, Stowell MH, Rees DC (1996) A leucine-rich repeat variant with a novel repetitive protein structural motif. Nat Struct Biol 3: 991-994.
96. Rubinson EH, Gowda AS, Spratt TE, Gold B, Eichman BF (2010) An unprecedented nucleic acid capture mechanism for excision of DNA damage. Nature 468: 406-411.
97. Chang J, Zhao M, Horanyi P, Xu H, Yang H, Liu Z-J, Chen L, Zhou W, Habel J, Tempel W, Lee D, Lin D, Chang S-H, Eneh JC, Hopkins RC, Jenney Jr. FE, Lee H-S, Li T, Poole II FL, Shah C, Sugar FJ, Chen C-Y, Arendall III WB, Richardson JS, Richardson DC, Rose JP, Adams MWW, Wang B-C, Genomics SCfS (2004).
98. Cook AG, Fukuhara N, Jinek M, Conti E (2009) Structures of the tRNA export factor in the nuclear and cytosolic states. Nature 461: 60-65.
99. Okada C, Yamashita E, Lee SJ, Shibata S, Katahira J, Nakagawa A, Yoneda Y, Tsukihara T (2009) A high-resolution structure of the pre-microRNA nuclear export machinery. Science 326: 1275-1279.
100. Wang Y, Opperman L, Wickens M, Hall TM (2009) Structural basis for specific recognition of multiple mRNA targets by a PUF regulatory protein. Proc Natl Acad Sci U S A 106: 20186-20191.
101. Takai H, Xie Y, de Lange T, Pavletich NP (2010) Tel2 structure and function in the Hsp90-dependent maturation of mTOR and ATR complexes. Genes Dev 24: 2019-2030.
102. Hast MA, Beese LS (2008) Structure of protein geranylgeranyltransferase-I from the human pathogen Candida albicans complexed with a lipid substrate. J Biol Chem 283: 31933-31940.
103. Mandelker D, Gabelli SB, Schmidt-Kittler O, Zhu J, Cheong I, Huang CH, Kinzler KW, Vogelstein B, Amzel LM (2009) A frequent kinase domain mutation that changes the interaction between PI3Kalpha and the membrane. Proc Natl Acad Sci U S A 106: 16996-17001.
104. Thompson JR, Banaszak LJ (2002) Lipid-protein interactions in lipovitellin. Biochemistry 41: 9398-9409.
105. Matsuura Y, Stewart M (2004) Structural basis for the assembly of a nuclear export complex. Nature 432: 872-877.
106. Vetter IR, Arndt A, Kutay U, Gorlich D, Wittinghofer A (1999) Structural view of the Ran-Importin beta interaction at 2.3 A resolution. Cell 97: 635-646.
107. Conti E, Kuriyan J (2000) Crystallographic analysis of the specific yet versatile recognition of distinct nuclear localization signals by karyopherin alpha. Structure 8: 329-338.
108. Goldenberg SJ, Cascio TC, Shumway SD, Garbutt KC, Liu J, Xiong Y, Zheng N (2004) Structure of the Cand1-Cul1-Roc1 complex reveals regulatory mechanisms for the assembly of the multisubunit cullin-dependent ubiquitin ligases. Cell 119: 517-528.

109. Goehler H, Lalowski M, Stelzl U, Waelter S, Stroedicke M, Worm U, Droege A, Lindenberg KS, Knoblich M, Haenig C, Herbst M, Suopanki J, Scherzinger E, Abraham C, Bauer B, Hasenbank R, Fritzsche A, Ludewig AH, Bussow K, Coleman SH, Gutekunst CA, Landwehrmeyer BG, Lehrach H, Wanker EE (2004) A protein interaction network links GIT1, an enhancer of huntingtin aggregation, to Huntington's disease. Mol Cell 15: 853-865.
110. Stelzl U, Worm U, Lalowski M, Haenig C, Brembeck FH, Goehler H, Stroedicke M, Zenkner M, Schoenherr A, Koeppen S, Timm J, Mintzlaff S, Abraham C, Bock N, Kietzmann S, Goedde A, Toksoz E, Droege A, Krobitsch S, Korn B, Birchmeier W, Lehrach H, Wanker EE (2005) A human protein-protein interaction network: a resource for annotating the proteome. Cell 122: 957-968.
111. Schaefer MH, Lopes TJ, Mah N, Shoemaker JE, Matsuoka Y, Fontaine JF, Louis-Jeune C, Eisfeld AJ, Neumann G, Perez-Iratxeta C, Kawaoka Y, Kitano H, Andrade-Navarro MA (2013) Adding protein context to the human protein-protein interaction network to reveal meaningful interactions. PLoS Comput Biol 9: e1002860.
112. Schaefer MH, Fontaine JF, Vinayagam A, Porras P, Wanker EE, Andrade-Navarro MA (2012) HIPPIE: Integrating protein interaction networks with experiment based quality scores. PLoS One 7: e31826.
113. Chatr-Aryamontri A, Breitkreutz BJ, Heinicke S, Boucher L, Winter A, Stark C, Nixon J, Ramage L, Kolas N, O'Donnell L, Reguly T, Breitkreutz A, Sellam A, Chen D, Chang C, Rust J, Livstone M, Oughtred R, Dolinski K, Tyers M (2013) The BioGRID interaction database: 2013 update. Nucleic Acids Res 41: D816-823.
114. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25: 25-29.
115. Huang da W, Sherman BT, Lempicki RA (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 4: 44-57.
116. Jansen RP, Dowzer C, Michaelis C, Galova M, Nasmyth K (1996) Mother cell-specific HO expression in budding yeast depends on the unconventional myosin myo4p and other cytoplasmic proteins. Cell 84: 687-697.
117. Zimprich A, Biskup S, Leitner P, Lichtner P, Farrer M, Lincoln S, Kachergus J, Hulihan M, Uitti RJ, Calne DB, Stoessl AJ, Pfeiffer RF, Patenge N, Carbajal IC, Vieregge P, Asmus F, Muller-Myhsok B, Dickson DW, Meitinger T, Strom TM, Wszolek ZK, Gasser T (2004) Mutations in LRRK2 cause autosomal-dominant parkinsonism with pleomorphic pathology. Neuron 44: 601-607.
118. Cole C, Barber JD, Barton GJ (2008) The Jpred 3 secondary structure prediction server. Nucleic Acids Res 36: W197-201.
119. Mills RD, Mulhern TD, Cheng HC, Culvenor JG (2012) Analysis of LRRK2 accessory repeat domains: prediction of repeat length, number and sites of Parkinson's disease mutations. Biochem Soc Trans 40: 1086-1089.
120. Manca A, Capsoni S, Di Luzio A, Vignone D, Malerba F, Paoletti F, Brandi R, Arisi I, Cattaneo A, Levi-Montalcini R (2012) Nerve growth factor regulates axial rotation during early stages of chick embryo development. Proc Natl Acad Sci U S A 109: 2009-2014.
121. Stevens NR, Dobbelaere J, Wainman A, Gergely F, Raff JW (2009) Ana3 is a conserved protein required for the structural integrity of centrioles and basal bodies. J Cell Biol 187: 355-363.
122. Lee SJ, Imamoto N, Sakai H, Nakagawa A, Kose S, Koike M, Yamamoto M, Kumasaka T, Yoneda Y, Tsukihara T (2000) The adoption of a twisted structure of importin-beta is essential for the protein-protein interaction required for nuclear transport. J Mol Biol 302: 251-264.
123. Mount DM (2004) Bioinformatics: Sequence and Genome Analysis (2nd ed.): Cold Spring Harbor Laboratory Press: Cold Spring Harbor, NY.

124. Huska MR, Buschmann H, Andrade-Navarro MA (2007) BiasViz: visualization of amino acid biased regions in protein alignments. Bioinformatics 23: 3093-3094.
125. Park Y, Yoon SK, Yoon JB (2009) The HECT domain of TRIP12 ubiquitinates substrates of the ubiquitin fusion degradation pathway. J Biol Chem 284: 1540-1549.
126. Aravind L (2001) The WWE domain: a common interaction module in protein ubiquitination and ADP ribosylation. Trends Biochem Sci 26: 273-275.
127. He F, Tsuda K, Takahashi M, Kuwasako K, Terada T, Shirouzu M, Watanabe S, Kigawa T, Kobayashi N, Guntert P, Yokoyama S, Muto Y (2012) Structural insight into the interaction of ADP-ribose with the PARP WWE domains. FEBS Lett 586: 3858-3864.
128. Hutagalung AH, Landsverk ML, Price MG, Epstein HF (2002) The UCS family of myosin chaperones. J Cell Sci 115: 3983-3990.
129. Melkani GC, Bodmer R, Ocorr K, Bernstein SI (2011) The UNC-45 chaperone is critical for establishing myosin-based myofibrillar organization and cardiac contractility in the Drosophila heart model. PLoS One 6: e22579.
130. Girard M, Poupon V, Blondeau F, McPherson PS (2005) The DnaJ-domain protein RME-8 functions in endosomal trafficking. J Biol Chem 280: 40135-40143.
131. Gu Y, Harley IT, Henderson LB, Aronow BJ, Vietor I, Huber LA, Harley JB, Kilpatrick JR, Langefeld CD, Williams AH, Jegga AG, Chen J, Wills-Karp M, Arshad SH, Ewart SL, Thio CL, Flick LM, Filippi MD, Grimes HL, Drumm ML, Cutting GR, Knowles MR, Karp CL (2009) Identification of IFRD1 as a modifier gene for cystic fibrosis lung disease. Nature 458: 1039-1042.
132. Wang G, Dunbrack RL, Jr. (2005) PISCES: recent improvements to a PDB sequence culling server. Nucleic Acids Res 33: W94-98.
133. Hofmann E, Wrench PM, Sharples FP, Hiller RG, Welte W, Diederichs K (1996) Structural basis of light harvesting by carotenoids: peridinin-chlorophyll-protein from Amphidinium carterae. Science 272: 1788-1791.
134. Schlesner M, Miller A, Streif S, Staudinger WF, Muller J, Scheffer B, Siedler F, Oesterhelt D (2009) Identification of Archaea-specific chemotaxis proteins which interact with the flagellar apparatus. BMC Microbiol 9: 56.
135. Lane N, Martin W (2010) The energetics of genome complexity. Nature 467: 929-934.
136. Studholme DJ, Fuerst JA, Bateman A (2004) Novel protein domains and motifs in the marine planctomycete Rhodopirellula baltica. FEMS Microbiol Lett 236: 333-340.
137. Fuerst JA, Sagulenko E (2012) Keys to eukaryality: planctomycetes and ancestral evolution of cellular complexity. Front Microbiol 3: 167.
138. Fogel GB, Collins CR, Li J, Brunk CF (1999) Prokaryotic Genome Size and SSU rDNA Copy Number: Estimation of Microbial Relative Abundance from a Mixed Population. Microb Ecol 38: 93-113.
139. Driver-Dunckley E, Caviness JN (2007) Neurology and Clinical Neuroscience: Elsevier.
140. Schaefer MH, Wanker EE, Andrade-Navarro MA (2012) Evolution and function of CAG/polyglutamine repeats in protein-protein interaction networks. Nucleic Acids Res 40: 4273-4287.
141. Steffan JS, Kazantsev A, Spasic-Boskovic O, Greenwald M, Zhu YZ, Gohler H, Wanker EE, Bates GP, Housman DE, Thompson LM (2000) The Huntington's disease protein interacts with p53 and CREB-binding protein and represses transcription. Proc Natl Acad Sci U S A 97: 6763-6768.
142. Qin ZH, Wang Y, Sapp E, Cuiffo B, Wanker E, Hayden MR, Kegel KB, Aronin N, DiFiglia M (2004) Huntingtin bodies sequester vesicle-associated proteins by a polyproline-dependent interaction. J Neurosci 24: 269-281.
143. Godin JD, Colombo K, Molina-Calavita M, Keryer G, Zala D, Charrin BC, Dietrich P, Volvert ML, Guillemot F, Dragatsis I, Bellaiche Y, Saudou F, Nguyen L, Humbert S (2010) Huntingtin is required for mitotic spindle orientation and mammalian neurogenesis. Neuron 67: 392-406.

144. Kitano H (2002) Systems biology: a brief overview. Science 295: 1662-1664.
145. Wanker EE, Rovira C, Scherzinger E, Hasenbank R, Walter S, Tait D, Colicelli J, Lehrach H (1997) HIP-I: a huntingtin interacting protein isolated by the yeast two-hybrid system. Hum Mol Genet 6: 487-495.
146. Seong IS, Woda JM, Song JJ, Lloret A, Abeyrathne PD, Woo CJ, Gregory G, Lee JM, Wheeler VC, Walz T, Kingston RE, Gusella JF, Conlon RA, MacDonald ME (2010) Huntingtin facilitates polycomb repressive complex 2. Hum Mol Genet 19: 573-583.
147. Chothia C, Janin J (1975) Principles of protein-protein recognition. Nature 256: 705-708.
148. Roy A, Kucukural A, Zhang Y (2010) I-TASSER: a unified platform for automated protein structure and function prediction. Nat Protoc 5: 725-738.
149. Neuvirth H, Raz R, Schreiber G (2004) ProMate: a structure based prediction program to identify the location of protein-protein binding sites. J Mol Biol 338: 181-199.
150. Yuan Z, Bailey TL, Teasdale RD (2005) Prediction of protein B-factor profiles. Proteins 58: 905-912.
151. Lyskov S, Gray JJ (2008) The RosettaDock server for local protein-protein docking. Nucleic Acids Res 36: W233-238.
152. Culver BP, Savas JN, Park SK, Choi JH, Zheng S, Zeitlin SO, Yates JR, 3rd, Tanese N (2012) Proteomic analysis of wild-type and mutant huntingtin-associated proteins in mouse brains identifies unique interactions and involvement in protein synthesis. J Biol Chem 287: 21599-21614.
153. Walzthoeni T, Leitner A, Stengel F, Aebersold R (2013) Mass spectrometry supported determination of protein complex structure. Curr Opin Struct Biol 23: 252-260.
154. Lee G, Abdi K, Jiang Y, Michaely P, Bennett V, Marszalek PE (2006) Nanospring behaviour of ankyrin repeats. Nature 440: 246-249.
155. Chitnis PR, Nelson N (1991) Molecular cloning of the genes encoding two chaperone proteins of the cyanobacterium Synechocystis sp. PCC 6803. J Biol Chem 266: 58-65.
156. Wecker P, Klockow C, Ellrott A, Quast C, Langhammer P, Harder J, Glockner FO (2009) Transcriptional response of the model planctomycete Rhodopirellula baltica SH1(T) to changing environmental conditions. BMC Genomics 10: 410.
157. Ponting CP, Russell RB (2000) Identification of distant homologues of fibroblast growth factors suggests a common ancestor for all beta-trefoil proteins. J Mol Biol 302: 1041-1047.
158. Cushman DW, Ondetti MA (1999) Design of angiotensin converting enzyme inhibitors. Nat Med 5: 1110-1113.
159. Wilson JL, Miranda CA, Knepper MA (2013) Vasopressin and the regulation of aquaporin-2. Clin Exp Nephrol.
160. Sands JM, Blount MA, Klein JD (2011) Regulation of renal urea transport by vasopressin. Trans Am Clin Climatol Assoc 122: 82-92.
161. Shen XZ, Xiao HD, Li P, Lin CX, Billet S, Okwan-Duodu D, Adams JW, Bernstein EA, Xu Y, Fuchs S, Bernstein KE (2008) New insights into the role of angiotensin-converting enzyme obtained from the analysis of genetically modified mice. J Mol Med (Berl) 86: 679-684.
162. Urata H, Boehm KD, Philip A, Kinoshita A, Gabrovsek J, Bumpus FM, Husain A (1993) Cellular localization and regional distribution of an angiotensin II-forming chymase in the heart. J Clin Invest 91: 1269-1281.
163. Bader M, Ganten D (2008) Update on tissue renin-angiotensin systems. J Mol Med (Berl) 86: 615-621.
164. Nguyen G, Muller DN (2010) The biology of the (pro)renin receptor. J Am Soc Nephrol 21: 18-23.
165. Tsuchida S, Matsusaka T, Chen X, Okubo S, Niimura F, Nishimura H, Fogo A, Utsunomiya H, Inagami T, Ichikawa I (1998) Murine double nullizygotes of the angiotensin type 1A and 1B receptor genes duplicate severe abnormal phenotypes of angiotensinogen nullizygotes. J Clin Invest 101: 755-760.

166. Donoghue M, Hsieh F, Baronas E, Godbout K, Gosselin M, Stagliano N, Donovan M, Woolf B, Robison K, Jeyaseelan R, Breitbart RE, Acton S (2000) A novel angiotensin-converting enzyme-related carboxypeptidase (ACE2) converts angiotensin I to angiotensin 1-9. Circ Res 87: E1-9.
167. Bader M (2010) Tissue renin-angiotensin-aldosterone systems: Targets for pharmacological therapy. Annu Rev Pharmacol Toxicol 50: 439-465.
168. Reusch HP, Luft FC (1991) [The role of chlorides in sodium-induced "salt-sensitive" hypertension]. Klin Wochenschr 69 Suppl 25: 90-96.
169. Luft FC (2004) Present status of genetic mechanisms in hypertension. Med Clin North Am 88: 1-18, vii.
170. Ganten D, Hayduk K, Brecht HM, Boucher R, Genest J (1970) Evidence of renin release or production in splanchnic territory. Nature 226: 551-552.
171. Ganten D, Minnich JL, Granger P, Hayduk K, Brecht HM, Barbeau A, Boucher R, Genest J (1971) Angiotensin-forming enzyme in brain tissue. Science 173: 64-65.
172. Steckelings UM, Paulis L, Unger T, Bader M (2011) Emerging drugs which target the renin-angiotensin-aldosterone system. Expert Opin Emerg Drugs 16: 619-630.
173. Pruitt KD, Tatusova T, Klimke W, Maglott DR (2009) NCBI Reference Sequences: current status, policy and new initiatives. Nucleic Acids Res 37: D32-36.
174. Blomme T, Vandepoele K, De Bodt S, Simillion C, Maere S, Van de Peer Y (2006) The gain and loss of genes during 600 million years of vertebrate evolution. Genome Biol 7: R43.
175. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25: 3389-3402.
176. Sokabe H, Ogawa M (1974) Comparative studies of the juxtaglomerular apparatus. Int Rev Cytol 37: 271-327.
177. Ragg H, Kumar A, Koster K, Bentele C, Wang Y, Frese MA, Prib N, Kruger O (2009) Multiple gains of spliceosomal introns in a superfamily of vertebrate protease inhibitor genes. BMC Evol Biol 9: 208.
178. Potempa J, Korzus E, Travis J (1994) The serpin superfamily of proteinase inhibitors: structure, function, and regulation. J Biol Chem 269: 15957-15960.
179. Wang Y, Ragg H (2011) An unexpected link between angiotensinogen and thrombin. FEBS Lett 585: 2395-2399.
180. Houard X, Williams TA, Michaud A, Dani P, Isaac RE, Shirras AD, Coates D, Corvol P (1998) The Drosophila melanogaster-related angiotensin-I-converting enzymes Acer and Ance--distinct enzymic characteristics and alternative expression during pupal development. Eur J Biochem 257: 599-606.
181. Williams TA, Michaud A, Houard X, Chauvet MT, Soubrier F, Corvol P (1996) Drosophila melanogaster angiotensin I-converting enzyme expressed in Pichia pastoris resembles the C domain of the mammalian homologue and does not require glycosylation for secretion and enzymic activity. Biochem J 318 (Pt 1): 125-131.
182. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32: 1792-1797.
183. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG (2007) Clustal W and Clustal X version 2.0. Bioinformatics 23: 2947-2948.
184. Riviere G, Michaud A, Corradi HR, Sturrock ED, Ravi Acharya K, Cogez V, Bohin JP, Vieau D, Corvol P (2007) Characterization of the first angiotensin-converting like enzyme in bacteria: Ancestor ACE is already active. Gene 399: 81-90.

185. Wu C, Macleod I, Su AI (2013) BioGPS and MyGene.info: organizing online, gene-centric information. Nucleic Acids Res 41: D561-565.
186. Ekbote U, Coates D, Isaac RE (1999) A mosquito (Anopheles stephensi) angiotensin I-converting enzyme (ACE) is induced by a blood meal and accumulates in the developing ovary. FEBS Lett 455: 219-222.
187. Vandingenen A, Hens K, Macours N, Schoofs L, De Loof A, Huybrechts R (2002) Presence of angiotensin converting enzyme (ACE) interactive factors in ovaries of the grey fleshfly Neobellieria bullata. Comp Biochem Physiol B Biochem Mol Biol 132: 27-35.
188. Corvol P, Eyries M, Soubrier F (2004) Peptidyl-dipeptidase A/angiotensin I-converting enzyme. In: Barrett AJ, Rawlings ND, Woessner JF. (eds) Handbook of proteolytic enzymes: Elsevier Academic Press edn, London.
189. Hagaman JR, Moyer JS, Bachman ES, Sibony M, Magyar PL, Welch JE, Smithies O, Krege JH, O'Brien DA (1998) Angiotensin-converting enzyme and male fertility. Proc Natl Acad Sci U S A 95: 2552-2557.
190. Crackower MA, Sarao R, Oudit GY, Yagil C, Kozieradzki I, Scanga SE, Oliveira-dos-Santos AJ, da Costa J, Zhang L, Pei Y, Scholey J, Ferrario CM, Manoukian AS, Chappell MC, Backx PH, Yagil Y, Penninger JM (2002) Angiotensin-converting enzyme 2 is an essential regulator of heart function. Nature 417: 822-828.
191. Sihn G, Rousselle A, Vilianovitch L, Burckle C, Bader M (2010) Physiology of the (pro)renin receptor: Wnt of change? Kidney Int 78: 246-256.
192. Jiang T, Gao L, Lu J, Zhang YD (2013) ACE2-Ang-(1-7)-Mas Axis in Brain: A Potential Target for Prevention and Treatment of Ischemic Stroke. Curr Neuropharmacol 11: 209-217.
193. Bridgham JT, Carroll SM, Thornton JW (2006) Evolution of hormone-receptor complexity by molecular exploitation. Science 312: 97-101.
194. Carroll SM, Ortlund EA, Thornton JW (2011) Mechanisms for the evolution of a derived function in the ancestral glucocorticoid receptor. PLoS Genet 7: e1002117.
195. Bury NR, Sturm A (2007) Evolution of the corticosteroid receptor signalling pathway in fish. Gen Comp Endocrinol 153: 47-56.
196. Jones CA, Sigmund CD, McGowan RA, Kane-Haas CM, Gross KW (1990) Expression of murine renin genes during fetal development. Mol Endocrinol 4: 375-383.
197. Taylor AA (1977) Comparative physiology of the renin-angiotensin system. Fed Proc 36: 1776-1780.
198. Brown JA, Cobb CS, Frankling SC, Rankin JC (2005) Activation of the newly discovered cyclostome renin-angiotensin system in the river lamprey Lampetra fluviatilis. J Exp Biol 208: 223-232.
199. Murzin AG, Brenner SE, Hubbard T, Chothia C (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 247: 536-540.
200. Holm L, Kaariainen S, Rosenstrom P, Schenkel A (2008) Searching protein structure databases with DaliLite v.3. Bioinformatics 24: 2780-2781.
201. Krogh A, Larsson B, von Heijne G, Sonnhammer EL (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol 305: 567-580.
202. Liu ZP, Wu LY, Wang Y, Zhang XS, Chen L (2008) Bridging protein local structures and protein functions. Amino Acids 35: 627-650.
203. Zhang Y (2008) I-TASSER server for protein 3D structure prediction. BMC Bioinformatics 9: 40.
204. Magrane M, Consortium U (2011) UniProt Knowledgebase: a hub of integrated protein data. Database (Oxford) 2011: bar009.
205. Pieper U, Webb BM, Barkan DT, Schneidman-Duhovny D, Schlessinger A, Braberg H, Yang Z, Meng EC, Pettersen EF, Huang CC, Datta RS, Sampathkumar P, Madhusudhan MS, Sjolander K, Ferrin

TE, Burley SK, Sali A (2011) ModBase, a database of annotated comparative protein structure models, and associated resources. Nucleic Acids Res 39: D465-474.
206. Emanuelsson O, Brunak S, von Heijne G, Nielsen H (2007) Locating proteins in the cell using TargetP, SignalP and related tools. Nat Protoc 2: 953-971.
207. Jmol (2011) Jmol: an open-source Java viewer for chemical structures in 3D.
208. Saqi MA, Sayle R (1994) PdbMotif--a tool for the automatic identification and display of motifs in protein structures. Comput Appl Biosci 10: 545-546.
209. Sayle RA, Milner-White EJ (1995) RASMOL: biomolecular graphics for all. Trends Biochem Sci 20: 374.
210. Gaulton A, Attwood TK (2003) Motif3D: Relating protein sequence motifs to 3D structure. Nucleic Acids Res 31: 3333-3336.
211. Ashkenazy H, Erez E, Martz E, Pupko T, Ben-Tal N (2010) ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. Nucleic Acids Res 38: W529-533.
212. O'Donoghue SI, Meyer JE, Schafferhans A, Fries K (2004) The SRS 3D module: integrating structures, sequences and features. Bioinformatics 20: 2476-2478.
213. Pettifer SR, Sinnott JR, Attwood TK (2004) UTOPIA-User-Friendly Tools for Operating Informatics Applications. Comp Funct Genomics 5: 56-60.
214. Wanker EE (2000) Protein aggregation in Huntington's and Parkinson's disease: implications for therapy. Mol Med Today 6: 387-391.
215. Chaurasia G, Malhotra S, Russ J, Schnoegl S, Hanig C, Wanker EE, Futschik ME (2009) UniHI 4: new tools for query, analysis and visualization of the human protein-protein interactome. Nucleic Acids Res 37: D657-660.
216. Deo RC, Schmidt EF, Elhabazi A, Togashi H, Burley SK, Strittmatter SM (2004) Structural bases for CRMP function in plexin-dependent semaphorin3A signaling. EMBO J 23: 9-22.
217. Da Silva JS, Medina M, Zuliani C, Di Nardo A, Witke W, Dotti CG (2003) RhoA/ROCK regulation of neuritogenesis via profilin IIa-mediated control of actin stability. J Cell Biol 162: 1267-1279.
218. Liu SH, Lin YH, Huang SF, Niou YK, Huang LL, Chen YJ (2013).
219. Rawls JM, Jr. (2006) Analysis of pyrimidine catabolism in Drosophila melanogaster using epistatic interactions with mutations of pyrimidine biosynthesis and beta-alanine metabolism. Genetics 172: 1665-1674.
220. Karshifoff A, Jelesarov I (2008) Salt bridges and conformational flexibility: effect on protein stability. Biotechnol and biotechnol eq: 606-611.
221. Wang LH, Strittmatter SM (1997) Brain CRMP forms heterotetramers similar to liver dihydropyrimidinase. J Neurochem 69: 2261-2269.
222. Ching YH, Ghosh TK, Cross SJ, Packham EA, Honeyman L, Loughna S, Robinson TE, Dearlove AM, Ribas G, Bonser AJ, Thomas NR, Scotter AJ, Caves LS, Tyrrell GP, Newbury-Ecob RA, Munnich A, Bonnet D, Brook JD (2005) Mutation in myosin heavy chain 6 causes atrial septal defect. Nat Genet 37: 423-428.
223. Granados-Riveron JT, Ghosh TK, Pope M, Bu'Lock F, Thornborough C, Eason J, Kirk EP, Fatkin D, Feneley MP, Harvey RP, Armour JA, David Brook J (2010) Alpha-cardiac myosin heavy chain (MYH6) mutations affecting myofibril formation are associated with congenital heart defects. Hum Mol Genet 19: 4007-4016.
224. Exome Variant Server NGESPE, Seattle, WA (URL: http://evs.gs.washington.edu/EVS/) [2011].
225. Rayment I, Holden HM, Whittaker M, Yohn CB, Lorenz M, Holmes KC, Milligan RA (1993) Structure of the actin-myosin complex and its implications for muscle contraction. Science 261: 58-65.
226. Zeron-Medina J, Wang X, Repapi E, Campbell MR, Su D, Castro-Giner F, Davies B, Peterse EF, Sacilotto N, Walker GJ, Terzian T, Tomlinson IP, Box NF, Meinshausen N, De Val S, Bell DA, Bond

GL (2013) A Polymorphic p53 Response Element in KIT Ligand Influences Cancer Risk and Has Undergone Natural Selection. Cell 155: 410-422.
227. Tarendeau F, Boudet J, Guilligay D, Mas PJ, Bougault CM, Boulo S, Baudin F, Ruigrok RW, Daigle N, Ellenberg J, Cusack S, Simorre JP, Hart DJ (2007) Structure and nuclear import function of the C-terminal domain of influenza virus polymerase PB2 subunit. Nat Struct Mol Biol 14: 229-233.
228. Kobe B (1999) Autoinhibition by an internal nuclear localization signal revealed by the crystal structure of mammalian importin alpha. Nat Struct Biol 6: 388-397.
229. Cansizoglu AE, Lee BJ, Zhang ZC, Fontoura BM, Chook YM (2007) Structure-based design of a pathway-specific nuclear import inhibitor. Nat Struct Mol Biol 14: 452-454.
230. Wollmann P, Cui S, Viswanathan R, Berninghausen O, Wells MN, Moldt M, Witte G, Butryn A, Wendler P, Beckmann R, Auble DT, Hopfner KP (2011) Structure and mechanism of the Swi2/Snf2 remodeller Mot1 in complex with its substrate TBP. Nature 475: 403-407.
231. Andersen CB, Becker T, Blau M, Anand M, Halic M, Balar B, Mielke T, Boesen T, Pedersen JS, Spahn CM, Kinzy TG, Andersen GR, Beckmann R (2006) Structure of eEF3 and the mechanism of transfer RNA release from the E-site. Nature 443: 663-668.
232. Cho US, Xu W (2007) Crystal structure of a protein phosphatase 2A heterotrimeric holoenzyme. Nature 445: 53-57.
233. Scherr N, Honnappa S, Kunz G, Mueller P, Jayachandran R, Winkler F, Pieters J, Steinmetz MO (2007) Structural basis for the specific inhibition of protein kinase G, a virulence factor of Mycobacterium tuberculosis. Proc Natl Acad Sci U S A 104: 12151-12156.
234. Heldwein EE, Macia E, Wang J, Yin HL, Kirchhausen T, Harrison SC (2004) Crystal structure of the clathrin adaptor protein 1 core. Proc Natl Acad Sci U S A 101: 14108-14113.
235. Collins BM, McCoy AJ, Kent HM, Evans PR, Owen DJ (2002) Molecular architecture and functional model of the endocytic AP2 complex. Cell 109: 523-535.
236. An Y, Chen CY, Moyer B, Rotkiewicz P, Elsliger MA, Godzik A, Wilson IA, Balch WE (2009) Structural and functional analysis of the globular head domain of p115 provides insight into membrane tethering. J Mol Biol 391: 26-41.
237. Shi H, Blobel G (2010) UNC-45/CRO1/She4p (UCS) protein forms elongated dimer and joins two myosin heads near their actin binding region. Proc Natl Acad Sci U S A 107: 21382-21387.
238. Slep KC, Vale RD (2007) Structural basis of microtubule plus end tracking by XMAP215, CLIP-170, and EB1. Mol Cell 27: 976-991.
239. Shomura Y, Dragovic Z, Chang HC, Tzvetkov N, Young JC, Brodsky JL, Guerriero V, Hartl FU, Bracher A (2005) Regulation of Hsp70 function by HspBP1: structural analysis reveals an alternate mechanism for Hsp70 nucleotide exchange. Mol Cell 17: 367-379.
240. Sadre-Bazzaz K, Whitby FG, Robinson H, Formosa T, Hill CP (2010) Structure of a Blm10 complex reveals common mechanisms for proteasome binding and gate opening. Mol Cell 37: 728-735.
241. Zhao G, Li G, Schindelin H, Lennarz WJ (2009) An Armadillo motif in Ufd3 interacts with Cdc48 and is involved in ubiquitin homeostasis and protein degradation. Proc Natl Acad Sci U S A 106: 16197-16202.
242. Xiang K, Nagaike T, Xiang S, Kilic T, Beh MM, Manley JL, Tong L (2010) Crystal structure of the human symplekin-Ssu72-CTD phosphopeptide complex. Nature 467: 729-733.
243. Calero G, Wilson KF, Ly T, Rios-Steiner JL, Clardy JC, Cerione RA (2002) Structural basis of m7GpppG binding to the nuclear cap-binding protein complex. Nat Struct Biol 9: 912-917.
244. Liberman N, Dym O, Unger T, Albeck S, Peleg Y, Jacobovitch Y, Branzburg A, Eisenstein M, Marash L, Kimchi A (2008) The crystal structure of the C-terminal DAP5/p97 domain sheds light on the molecular basis for its processing by caspase cleavage. J Mol Biol 383: 539-548.

245. Milburn CC, Boudeau J, Deak M, Alessi DR, van Aalten DM (2004) Crystal structure of MO25 alpha in complex with the C terminus of the pseudo kinase STE20-related adaptor. Nat Struct Mol Biol 11: 193-200.
246. Handa N, Nishino A, Kishishita S, Murayama K, Shirouzu M, Yokoyama S (2006).
247. Kuzin AP, Chen Y, Acton T, Xiao R, Conover K, Ma C, Kellie R, Montelione GT, Tong L, Hunt JF (2005).
248. Urvoas A, Guellouz A, Valerio-Lepiniec M, Graille M, Durand D, Desravines DC, van Tilbeurgh H, Desmadril M, Minard P (2010) Design, production and molecular structure of a new family of artificial alpha-helicoidal repeat proteins (alphaRep) based on thermostable HEAT-like repeats. J Mol Biol 404: 307-327.

List of publications

Stroedicke M, Bounab Y, Strempel N, Chaurasia G, Friedrich RP, Li S, Riechers SP, Plaßmann S, Russ J, Nicoletti C, Schnoegl S, **Fournier D**, Graham RK, Hayden MR, Sigrist S, Bates G, Priller J, Andrade-Navarro MA, Futschik M and Wanker EE. Interaction network filtering identifies CRMP-1 as a potent modifier of mutant huntingtin misfolding and neurotoxicity. Submitted.

Fournier D, Palidwor GA, Shcherbinin S, Szengel A, Schaefer MH, Andrade-Navarro M. Functional and genomic analyses of alpha-solenoid proteins. *PLoS One*. 2013;8(11):e79894. Epub 2013 Nov 21.

Guéguen L, Gaillard S, Boussau B, Gouy M, Groussin M, Rochette NC, Bigot T, **Fournier D**, Pouyet F, Cahais V, Bernard A, Scornavacca C, Nabholz B, Haudry A, Dachary L, Galtier N, Belkhir K, Dutheil JY. Bio++: efficient, extensible libraries and tools for computational molecular evolution. *Molecular Biology and Evolution*. Epub 2013 May 21.

Fournier D, Luft FC, Bader M, Ganten D and Andrade-Navarro MA. 2012. Emergence and evolution of the renin-angiotensin-aldosterone system. *Journal of Molecular Medicine*. Epub 2012 Apr 14.

Posch MG, Waldmuller S, Müller M, Scheffold T, **Fournier D**, Andrade-Navarro MA, De Geeter B, Guillaumont S, Dauphin C, Yousseff D, Schmitt KR, Perrot A, Berger F, Hetzer R, Bouvagnet P, Özcelik C. Cardiac alpha-myosin (MYH6) is the predominant sarcomeric disease gene for familial atrial septal defects. *PLoS One*. 2011;6(12):e28872. Epub 2011 Dec 14.

Fournier D, Andrade-Navarro MA. PDBpaint, a visualization webservice to tag protein structures with sequence annotations. *Bioinformatics*. 2011 Sep 15;27(18):2605-6. Epub 2011 Jul 14.