Universitat Autònoma de Barcelona

Development of Bioinformatic Tools for the Study of Membrane Proteins

Luis Eduardo Mayol Escuer

Doctoral thesis
PhD program in Biochemistry, Molecular Biology and Biomedicine.

Development of Bioinformatic Tools for the Study of Membrane Proteins

Eduardo Mayol Escuer

Directors: Arnau Cordomí and Mireia Olivella

Tutor: Mireia Duñach

Medicine Faculty, Biostatistics Department. Laboratory of Computational Medicine.
Universitat Autònoma de Barcelona (UAB)

2019

UAB
Universitat Autònoma de Barcelona

Doctoral thesis
PhD program in Biochemistry, Molecular Biology and Biomedicine.

Development of Bioinformatic Tools for the Study of Membrane Proteins

This work has been carried out by Luis Eduardo Mayol Escuer under the supervision of Dr. Arnau Cordomí Montoya and Dr. Mireia Olivella García to obtain the degree of Doctor in Biochemistry, Molecular Biology and Biomedicine.

Luis Eduardo Mayol Escuer
Candidate.

Dr. Arnau Cordomí Montoya
Researcher, UAB.

Dr. Mireia Olivella García
Associate Professor, International School of Studies, ESCI-UPF.

Agradecimientos

En esta tesis ha colaborado mucha gente dentro y fuera del laboratorio.

Primero tengo que agradecer a Arnau y Mireia, por su dirección a lo largo de estos años, su paciencia y su actitud siempre positiva. Muchas gracias, he aprendido mucho de vosotros. También un especial agradecimiento a todos los coautores de los artículos publicados y los que están en proceso.

A Leonardo, gracias por darme la oportunidad de desarrollar esta etapa tan importante de mi formación. Ha sido un placer trabajar estos años en el laboratorio. Merche, muchas gracias por tu asesoramiento estadístico y también al resto de post-docs por crear un buen ambiente de trabajo.

Quiero agradecer también a todos mis compañeros de doctorado, con todos los que empecé, compartimos muy buenos momentos durante esos primeros años, en la facultad, los workshops y congresos, gracias a vosotros tendré un mejor recuerdo de esta etapa. También a los nuevos compañeros de doctorado, que han traído una nueva ola de ilusión y se agradece, ¡ánimo que se puede!

Igualmente, tengo que agradecer a Peter Hildebrand, por acogerme durante unos meses en su grupo en Berlín, permitirme ver como se trabaja en un gran centro internacional y darme la oportunidad de conocer muchos excelentes científicos. Un abrazo grande a Peter, Johanna, Heiko y Johen. Danke! Me alegro mucho de haber podido coincidir allí con Ramón y reencontrarnos después para seguir compartiendo buenos ratos, y los no tan buenos, pero al menos con unas risas. Gracias por el asesoramiento científico, las conversaciones productivas y todas las demás.

Tengo que agradecer al COST (action CM1207) por financiar mi estancia en Berlín y la asistencia a parte de los congresos internacionales por toda Europa a los que he podido asistir y en los que he aprendido mucho y conocido a muchos compañeros.

Fuera de la universidad también he recibido mucho apoyo para poder llevar a cabo este largo proceso, muchas gracias a mis amigos Raul y Sergi, por estar ahí desde el principio del camino, a todos los que me he cruzado y a los que ya no están. Toni, muchas gracias por todo, dentro y fuera de las clases. Pablo, gracias por estar ahí siempre con una sonrisa. No me olvido de vosotras, Esther, Cris, Irene, Sara, Marta, aunque no nos veamos mucho, también sois parte de esto. Quiero agradecer a todos los compañeros y amigos de la universidad, Albert, Vendrell, Carod, Nuria, Clara y demás por todos los buenos ratos y los que quedan. Y especialmente a Laia, por acompañarme todos estos años y darme todo el apoyo que me hacía falta, y por soportarme, que a veces no es fácil, muchas gracias de corazón.

Por último, pero no menos importante, a mi familia, mis padres, Montse y Ramón, por haberme permitido desarrollar profesionalmente facilitándome mucho el camino. Muchas gracias, sin vosotros no habría llegado hasta aquí. Y mis hermanos, Noemí y Carlos, muchas gracias por vuestro apoyo incondicional y todos los ánimos, os aseguro que me han llegado. También al resto de mi gran familia, sobrino/as, abuelo/as, tío/as, primo/as. Gracias, entre todos habéis hecho que pueda crecer en lo personal y en lo profesional. ¡Mil gracias a todos!

Abstract

Membrane proteins are fundamental elements for every known cell, accounting for a quarter of genes in the Human genome, they play essential roles in cell biology. About 50% of currently marketed drugs have a membrane protein as target, and around a third of them target G-protein-coupled receptors (GPCRs). The current difficulties and limitations in the experimental work necessary for microscopic studies of the membrane as well as membrane proteins urged the use of computational methods. The scope of this thesis is to develop new bioinformatic tools for the study of membrane proteins and also for GPCRs in particular that help to characterize their structural features and understand their function. In regard to membrane proteins, a cornerstone of this thesis has been the creation of two databases for the main classes of membrane proteins: one for α-helical proteins (TMalphaDB) and another for β-barrel proteins (TMbetaDB). These databases are used by a newly developed tool to find structural distortions induced by specific amino acid sequence motifs (http://lmc.uab.cat/tmalphadb and http://lmc.uab.cat/tmbetadb) and in the characterization of inter-residue interactions that occur in the transmembrane region of membrane proteins aimed to understand the complexity and differential features of these proteins. Interactions involving Phe and Leu residues were found to be the main responsible for the stabilization of the transmembrane region. Moreover, the energetic contribution of interactions between sulfur-containing amino acids (Met and Cys) and aliphatic or aromatic residues were analyzed. These interactions are often not considered despite they can form stronger interactions than aromatic-aromatic or aromatic-aliphatic interactions. Additionally, G-protein coupled receptor family, the most important family of membrane proteins, have been the focus of two web applications tools dedicated to the analysis of conservation of amino acids or sequence motifs and pair correlation (GPCR-SAS, http://lmc.uab.cat/gpcrsas) and to allocate internal water molecules in receptor structures (HomolWat, http://lmc.uab.cat/HW). These web applications are pilot studies that can be extended to other membrane proteins families in future projects. All these tools and analysis may help in the development of better structural models and contribute to the understanding of membrane proteins.

Resumen

Las proteínas de membrana son elementos fundamentales de todas las células conocidas, que representan una cuarta parte de los genes del genoma humano, y desempeñan funciones esenciales en la biología celular. Alrededor del 50% de los medicamentos comercializados actualmente tienen una proteína de membrana como objetivo, y alrededor de un tercio de todos ellos se dirigen a los receptores acoplados a proteína G (GPCR). Las dificultades y limitaciones en el trabajo experimental necesario para los estudios microscópicos de la membrana, así como las proteínas de membrana, impulsaron el uso de métodos computacionales. El alcance de esta tesis es desarrollar nuevas herramientas bioinformáticas para el estudio de las proteínas de membrana y en particular para GPCRs que ayudan a caracterizar sus rasgos estructurales y ayudar a la comprensión de su función. Con respecto a las proteínas de membrana, una piedra angular de esta tesis ha sido la creación de dos bases de datos para las principales clases de proteínas de membrana: una para helices-α (TMalphaDB) y otra para proteínas barriles-β (TMbetaDB). Estas bases de datos son empleadas por una herramienta recientemente desarrollada para encontrar distorsiones estructurales inducidas por motivos específicos de secuencias de aminoácidos (http://lmc.uab.cat/tmalphadb y http://lmc.uab.cat/tmbetadb). También se usaron en la caracterización de las interacciones entre residuos que se producen en la región transmembrana de estas proteínas con el objetivo de favorecer la comprensión de la complejidad y las características diferenciales de las proteínas de membrana. Se encontró que las interacciones que involucran los residuos de Phe y Leu son las principales responsables de la estabilización de la región transmembrana. Además, se analizó la contribución energética de las interacciones entre los aminoácidos que contienen azufre (Met y Cys) y los residuos alifáticos o aromáticos. Estas interacciones normalmente no se tienen en gran consideración a pesar de que pueden formar interacciones más fuertes que las interacciones aromático-aromático o aromático-alifático. Asimismo, la familia de GPCRs, la más importante de proteínas de membrana, ha sido el foco de dos aplicaciones web dedicadas al análisis de conservación de aminoácidos o motivos de secuencia y correlación de pares (GPCR-SAS, http://lmc.uab.cat/gpcrsas) y para incorporar moléculas de agua internas en estructuras de estos receptores (HomolWat, http://lmc.uab.cat/HW). Estas aplicaciones web son estudios piloto que pueden extenderse a otras familias de proteínas de membrana en proyectos futuros. Todas estas herramientas y análisis pueden ayudar en el desarrollo de mejores modelos estructurales y contribuir a la comprensión de las proteínas de membrana.

“The good thing about science is that it’s true
whether or not you believe in it”

Neil deGrasse Tyson

Table of contents

1. Introduction

1. INTRODUCTION

1.1. Proteins: from sequence to structure and function

Proteins are large biomolecules that constitute one of the main building blocks of cells, they represent most of the cell's dry mass (Alberts *et al.*, 2014, Milo & Phillips 2015), and are responsible for cell assembly. The human genome contains approximately 20,000 protein-coding genes (Uhlén *et al.*, 2015). Proteins execute many functions in the cell: catalysing reactions, giving structural support, storing molecules, transducing signals, and many others. In order to carry out this wide diversity of functions, proteins possess specific sequence and structural features. Thus, amino acids sequences of proteins have been evolutively selected for its capacity to fold in certain structure in order to develop specific functions. The biological activity of a protein depends on its three-dimensional shape, and its structure is inherent to its protein sequence. Understanding the relation between protein sequence and its structure therefore is crucial in order to get insight on protein function. Protein structure is organized at many levels of complexity, going from primary to quaternary structure.

1.1.1. Primary Structure

The **primary structure** of proteins is its amino acid sequence, that is determined by the translation of DNA in its coding regions. Proteins are assembled from a set of 20 different building blocks (20 amino acids), that consist of a common backbone of a carboxylic acid group and an amino group linked by an α-carbon atom. Amino acids differ in the side-chain that branches from the α-carbon atom and provides specific physicochemical properties. The structure of the 20 amino acids commonly found in proteins are shown in Figure 1.1 together with the one-letter and the three-letter code abbreviations. These abbreviations are commonly used to simplify the written sequence of a peptide or protein. The sum and combination of all side-chains contained in a protein provides a variety of chemical properties that determines the 3D structure of the protein and its function.

Amino acids are bound covalently through peptide bonds forming a polypeptide chain. Peptide bonds are formed by condensation reaction of the amino group of one residue to the carboxyl group of the next residue in the sequence. The synthesis of peptide bonds is an enzymatically controlled process that takes place in the ribosomes directed by mRNA templates. Although the resulting amide bonds are very stable in water at neutral pH, their hydrolysis in cells is also enzymatically controlled (Petsko & Ringer, 2004). Amino acid sequences in proteins can be determined directly through chemical analysis or indirectly by determining the base sequence in the parent DNA. The average of length of human proteins is about 375 amino acids (Brocchieri *et al.*, 2005). However, we can find small peptides such as Insulin, with just 51 amino acids, or huge protein complexes, such as Titin, a muscle protein with ~30,000 amino acids. The conformation of proteins is not just maintained by the peptidic bond. Non-covalent interactions are also responsible for stabilization of the protein conformation (see section 1.1.4).

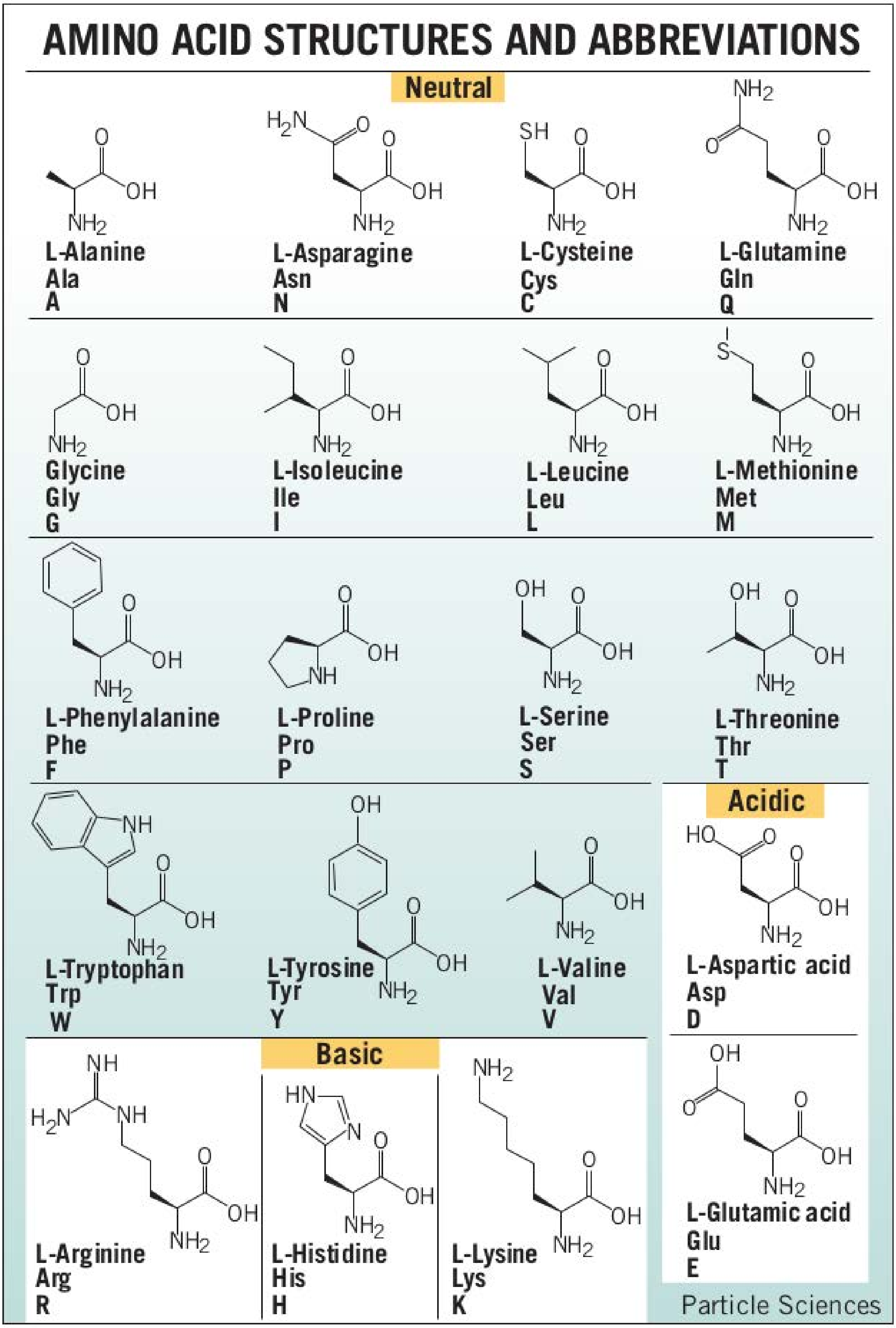


Figure 1.1. Chemical structure for the 20 amino acid and their three-letter and one-letter codes. Adapted from Technical Brief 2009 Volume 8 Particle Sciences Inc.

1.1.2. Secondary Structure

Secondary structure refers to local folded conformations that forms within a polypeptide due to interactions between atoms of the backbone. Our knowledge about secondary structure have its origin in the work of Linus Pauling in the first half of the XXth century. In the 1930s he started his studies with X-ray diffraction in amino acids and small peptides in order to get their three-dimensional structure. After many experiments, in 1951, Linus Pauling and Robert Corey proposed the two secondary structures of proteins: α-helices and β-sheets (Pauling & Corey, 1951). Both structures are held in shape by hydrogen bonds formed between the carbonyl group of one residue and the amino group of another, although they differ in the hydrogen bond pattern.

α-helices are the most common secondary structure, their backbone atoms are arranged in a right-handed helical structure where the carbonyl oxygen atom of each residue (i) accepts a hydrogen bond from the amide nitrogen four residues further in the sequence (i+4), completing a turn every 3.6 amino acids. Other type of biological helices is the so-called 3_{10}, much less common than α-helices, they constitute about 10-15% of all helices. Lastly, there is a much less represented helix named π-helix, it is known that around 15% of helices contain a π-helical segment (Cooley *et al.* 2010). Any complete π-helix have been found in nature so far. Some of the parameters and a representation of these helices and their hydrogen bond pattern are summarized in Table 1.1 and Figure 1.2.

Helix type	Hydrogen bond pattern	Φ / Ψ	Unit twist
α-helix	i → i+4	-60º/-50º	100º
3_{10}-helix	i → i+3	-49º/-26º	120º
π-helix	i → i+5	-57º/-70º	87º

Table 1.1. Values for properties of canonical helices, hydrogen bond pattern, typical Phi (Φ) and Psi (Ψ) torsion angles and the unit twist, that is, considering the 360º of a complete turn, how many degrees rotate a residue from the previous one. Thinner helices present a higher unit twist whereas wider helices have lower unit twist than α-helices.

β-sheets are patches of 3 to 10 consecutive amino acids (named strands) with backbone in an extended conformation connected laterally with each other in a parallel or anti-parallel manner. The hydrogen bonding in a β-sheet is between strands (inter-strand) rather than within strands (intra-strand) as in α-helices, complicating the identification of this type of secondary structure from the sequence of amino acids. The sheet conformation consists of pairs of strands lying side-by-side. The carbonyl oxygens in one strand hydrogen bond with the amino hydrogens of the adjacent strand. The two strands can be either parallel or anti-parallel depending on whether the strand directions (N-terminus to C-terminus) are the same or opposite. Anti-parallel β-sheets are more stable due to a better matching of the hydrogen bond patterns (Figure1.3).

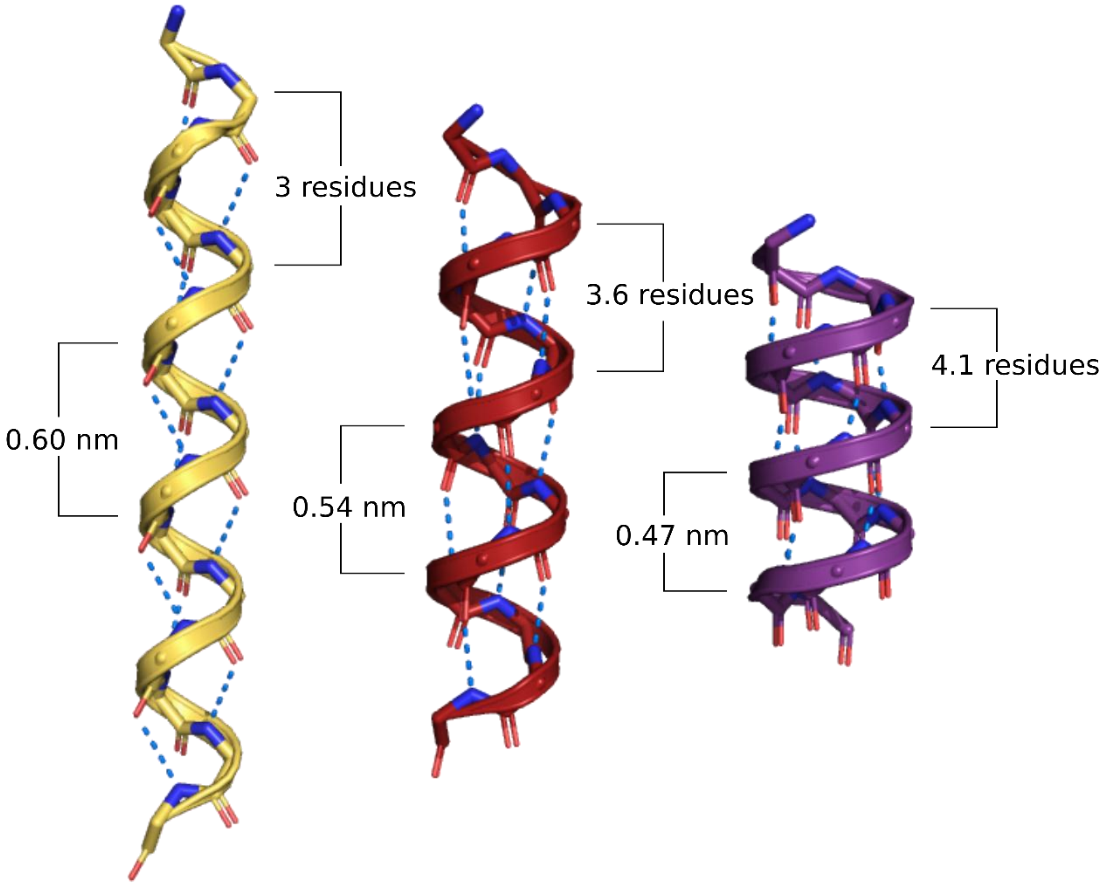


Figure 1.2. Canonical structure and hydrogen bond pattern for 3_{10}-helix (yellow), α-helix (red), and π-helix (purple). All of them contain the same number of residues (16). Blue dashed lines represent hydrogen bond patterns. Features of each complete turn for each helix type.

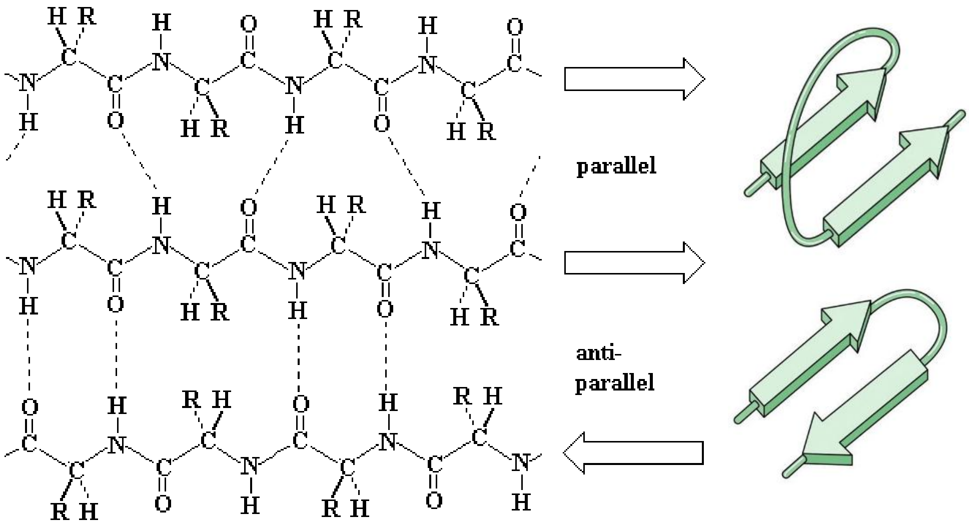


Figure 1.3. *Hydrogen bond pattern for parallel and anti-parallel β-sheets. From http://www.chembio.uoguelph.ca/educmat/phy456/456lec01.htm*

Thus, secondary structure is the three-dimensional conformation defined by patterns of hydrogen bonds between the main chain peptide groups and the allowed torsion angles. The peptide bond is important for the stability of proteins. Due to resonance, the delocalization of electrons increases the polarity of the peptide bond and confer a partial double-bond character, which results in a coplanarity of O=C-N-Cα atoms, with C=O and N-H groups positioning in opposite directions of the plane, hence the rotation about the peptidic bond is limited (Figure 1.4). That is the reason why there are only two backbone bonds that are free to rotate allowing some flexibility to peptidic chains. The torsion angles between these bonds are the commonly named Phi (Φ, between C and Cα of the first amino acid) and Psi (Ψ, between N and Cα of the second amino acid). In the mid-1960s, the group of G. N. Ramachandran first represented the confrontation of the torsion angles in a plot that now is part of every biochemistry textbook (Ramakrishnan & Ramachandran, 1965). They developed a way to visualize energetically allowed regions of backbone dihedral angles. Typical values for Φ and Ψ angle pairs are shown in the Ramachandran plot. Figure 1.5 shows that α-helices and β-sheets appear at different regions of the Φ and Ψ space.

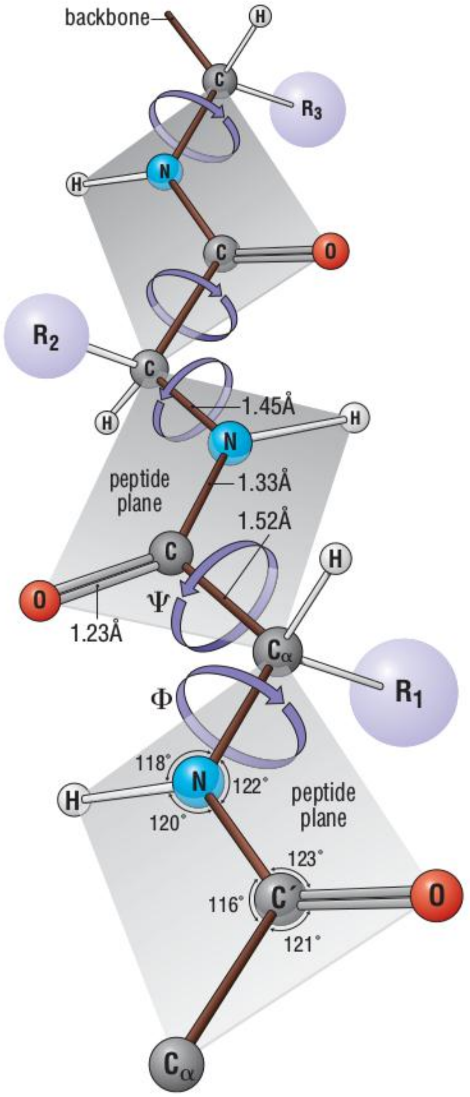


Figure 1.4. Extended polypeptide chain showing the typical backbone lengths and angles. Adapted from Petsko & Ringer 2004.

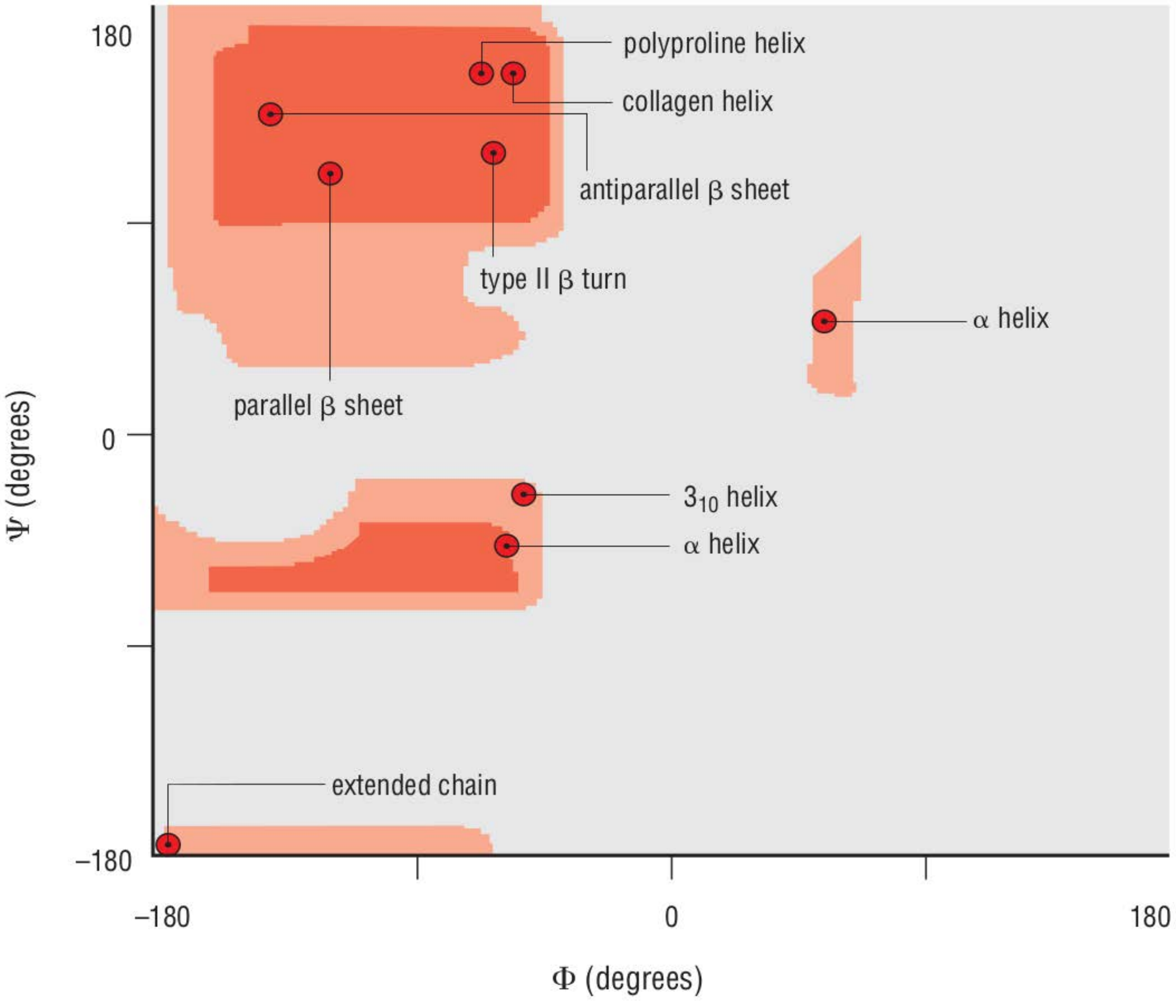


Figure 1.5. Schematic representation of a Ramachandran Plot, shown in red are those combinations of psi and phi backbone torsion angles allowed (without steric clashes) Adapted from Petsko & Ringe, 2004.

1.1.3. Supersecondary structure

Many proteins show recurrent patterns of interactions between helices and sheets. The combination of such secondary structure patterns is the supersecondary structure. These supersecondary structures are local structures formed by residues in a contiguous segment of the sequence. The main motifs found in proteins include: i) the **alpha-helix hairpin** (two alpha helices joint by a loop), ii) **Beta Hairpin** (two beta strands joined by a loop) and iii) **Beta-alpha-beta** (two β-strands and an α-helix joint by loops). The interactions between secondary structure elements include van der Waals packing and hydrogen bonds. The combination of different supersecondary structures or motifs can form a wide diversity of 3D structure.

1.1.4. Domain and Tertiary structure

A **domain** is a conserved part of the protein that can evolve, function and exist independently of the rest of the protein chain. Based on their secondary structural elements, protein domains can be classified into five broad classes, based on the predominant secondary structure:

(A) **Alpha domains**, comprised entirely of alpha-helices.
(B) **Beta domains**, contain only beta-sheet.
(C) **Alpha/beta** domains, contain beta strands with connecting helical segments.
(D) **Alpha+beta**, domains that contain separated beta sheet and helical regions.
(E) **Cross-linked domains**, with an undefined secondary structure but stabilized by several disulfide bridges or metal ions.

Different arrangements in each class are possible, translating in a wide variety of domains, some of them with capacity to bind DNA such as the helix-turn-helix or the Leucine Zipper.

The overall three-dimensional shape of an entire protein molecule is the **tertiary structure**. The tertiary structure emerges from the combination of both interactions between proximal consecutive amino acids in the primary sequence (mainly via hydrogen bonds that stabilize alpha-helices or beta sheets) and between amino acids from distant parts of the primary sequence that intermingle via disulfide, charge-charge, hydrophobic, or other non-covalent interactions. Such weak noncovalent interactions with a much lesser energetic contribution than covalent bonds are the main contributors to protein folding (Petsko & Ringer, 2004). Weakly polar interactions rarely commit even one-tenth of the enthalpy contributed by a single covalent bond, however, in any folded protein structure there may be hundreds to thousands of them, adding up a very large contribution (Petsko & Ringer, 2004). The two more important are the van der Waals interaction and the hydrogen bond. A list of these non-covalent interactions is shown in Table 1.2.

	Type of interaction	**Factors responsible for the interaction**	**Energy range (kcal/mol)**	**Distance dependency on energy**	**Example**
Coulomb	Ion-ion	Ion charge	20-40	1/r	–NH3+ ···–OOC–
	Ion-dipole (H-bond)	Ion charge, dipole magnitude	10-25	$1/r^2$	–NH3+ ··· O=C<
	Dipole-dipole (H-bond)	Dipole magnitude, electronegativity	2-7	$1/r^3$	>C=O ··· HN<
van de Waals	Dipole-induced dipole	Dipole magnitude, polarizability	0.5-20.5	$1/r^4$	–OH ··· –CH3
	Dispersion	Polarizability	0.1-3	$1/r^6$	–CH3 ··· –CH3

Table 1.2. Main types and features of non-covalent interactions. Adapted from Cordomí *et al.*, 2013.

1.1.5. Quaternary structure

Many proteins require the assembly of several polypeptide subunits to be functional. The **quaternary structure** is the result of the interaction of two or more folded polypeptides to form a complex. If the final protein is made of two subunits, the protein is said to be a dimer. If three subunits must come together, the protein is said to be a trimer, four subunits make up a tetramer, etc. If the subunits are identical, the prefix “homo” is used, as in “homodimer.” If the subunits are different, we use “hetero,” as in “heterodimer”. (Figure 1.6).

Overall, throughout a given protein family quaternary structure is less conserved than tertiary structure, i.e. while the fold of a polypeptide chain remains structurally similar the number of subunits forming the biologically relevant quaternary structure can vary significantly (Aloy *et al.*, 2003, Levy *et al.*, 2008). However, if a specific interaction between two protein chains plays a structural or functional role, it is reasonable to expect that residues at the corresponding interface are less free to vary hence increasing evolutionary conservation in these region (Elcock & McCammon, 2001, Capra & Singh, 2007).

The subunits in a quaternary structure must be specifically arranged for the entire protein to function properly. Any alteration in the structure of the subunits or how they are associated causes marked changes in biological activity (Ouellette & Rawn, 2015).

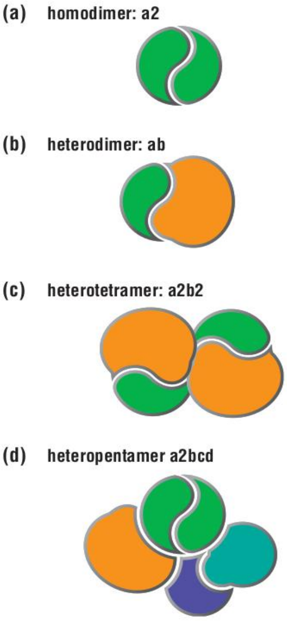


Figure 1.6. Schematic representation of different oligomers. Adapted from Petsko and Ringer 2004

1.1.6. Post-translational modifications

A final element that also modulates structure are post-translational modifications (PTMs) such as phosphorylation, glycosylations or palmitoylations. These occur in almost all proteins and modulate their structure and dynamics. A complete description of all PTMs and tools for their prediction can be found in the recent work of Audagnotto and Dal Peraro (Audagnotto & Dal Peraro, 2017). Phosphorylation occurs in amino acids with hydroxylic side chain, such as Ser, Thr and Tyr, and allows the recognition of other proteins or can induce conformational changes in proteins. Palmitoylations, are attachments of the fatty acid palmitoyl typically to Cys residues and help the protein to anchor the membrane (Blaskovic *et al.*, 2013).

1.1.7. Protein Functions

Proteins are the most versatile macromolecules in living systems and serve crucial functions in essentially all biological processes. They function as catalysts, they can transport and store other molecules such as oxygen, provide mechanical support and immune protection, generate movement, transmit nerve impulses, and control growth and differentiation, among others (see Table 1.3). All in all, protein functions are as diverse as protein structures.

Type	Function
Enzymes	Catalyze covalent bond breakage or formation
Structural proteins	Provides mechanical support to cells and tissues
Transport proteins	Carry small molecules or ions
Motor proteins	Generate movement in cells and tissues
Storage proteins	Storage of amino acids or ions
Signal proteins	Carry extracellular signals from cell to cell
Receptor proteins	Detect signals and transmit them to the cell's response machinery
Gene regulatory proteins	Bind to DNA to switch genes on or off

Table 1.3. Type of proteins and their main function. Adapted from Alberts *et al.*, 2014.

1.1.8. Evolution of protein structure and function

Evolution of proteins is the exploration of the space of amino acids sequences in search for selectively advantageous variants, a process that has been taking place for millions of years. The change in DNA sequence is usually translated in a change in protein sequence that can affect protein structure and protein function. Many domain families are found in all forms of life and repeatedly found in diverse proteins. Domains are used by nature to generate new proteins. The majority of proteins are multidomain proteins that have been created as a result of gene duplication events followed by speciation events (Siltberg-Liberles *et al.*, 2011). Evolution for instance, can generate a new enzyme from one that is “close”, that is, shares elements of mechanism or machinery from which the new activity can be built. Nature co-opts old machinery to do new jobs. And sometimes the ability to do the new job is already there, at least at a low level (Arnold 2018).

Many of the current domains once existed as independent proteins. The evolutionary classification of protein domains has been based on sequence and structural homologies that make use of phylogenetic tools and advanced bioinformatic methods (Marsden & Orengo, 2008). Protein families group together sequences that share a common ancestry, but generally do so with a low hierarchical inequality; the reliability of comparative methods break down when reaching the so-called ‘twilight zone’ of < 30 % sequence identity. (Caetano-Anollés *et al.*, 2009).

1.1.9. Classification of proteins

Various classification systems are commonly used for proteins based on their sequence, structure or function. One of the most used classification schemes divides proteins based on their structure into **fibrous**, **globular** and **membrane** proteins. Fibrous proteins perform mainly structural functions, globular proteins folds into a spherical shape with an irregular surface and can perform many different functions (see section 1.1.7), and finally membrane proteins, located at the membranes of cells (see section 1.2). Based on their folding, proteins can also be classified into

families where each member has an amino acid sequence and a three-dimensional conformation that closely resemble those of the other family members. Evolutionary families appear because once a protein has reached a stable conformation with the required properties, its structure can be modified over time to acquire new functions.

There are two main endeavours to classify proteins. **SCOP**, for Structural Classification Of Proteins, and **CATH**, for Class, Architecture, Topology and Homologous superfamily. And a third, **PFAM** that sorts out every protein into families (commented above). These schemes sort out all possible protein folds. Presumably few or any new fold is going to be discovered since no new folds were identified after 2008 and 2012 respectively for SCOP and CATH. Both schemes are based in the evolutionary relationships.

- **SCOP** (Andreeva *et al.*, 2007, 2014). Almost all proteins have structural similarities with other proteins and, for some cases, share a common evolutionary origin. Knowing these relationships is crucial for a complete understanding of the evolution of proteins and of development. Proteins are classified to reflect both structural and evolutionary relatedness. There are many levels in the hierarchy, but the main ones are:

  - **Family**: proteins with related sequence but typically with distinct function.
  - **Superfamily**: bridge together protein families with common functional and structural features inferring probable common ancestors.
  - **Fold**: similar structural elements.
  - **Class**: folds with similar structure.

The exact position of boundaries between these levels are difficult to determine. The evolutionary classification is generally conservative (doubts about relatedness translates into new divisions at the family and superfamily levels). A new level is incorporated to SCOP2, **Hyperfamily** domain, a common region shared by different superfamilies. Proteins are defined to share a fold if they have same major secondary structures in same arrangement and with the same topological connections. Different proteins with the same fold often have peripheral elements of secondary structure and turn regions that differ in size and conformation. In some cases, these differing peripheral regions may comprise half the structure. Proteins placed together in the same fold category may not have a common evolutionary origin: the structural similarities could arise just from the physics and chemistry of proteins favouring certain packing arrangements and chain topologies.

- **CATH** (Dawson *et al.*, 2016) is also a hierarchical classification of protein domain structures, which clusters proteins at four major levels, class, architecture, topology and homologous superfamily. Annotation of domains is both manual and automatic. Class, similar to class from SCOP, is defined by the secondary structure content (All alpha, all beta, alpha/beta etc.). Architecture is the clustering of structurally similar arrangement of secondary elements, independent of their connectivity. Topology, or fold family, is the structural grouping depending on both overall 3D shape and connectivity. And finally, homologous superfamilies are groups of protein domains with a common ancestor.

- **PFAM** (Finn *et al.*, 2015) provide a complete and accurate classification of protein families and domains. Each Pfam entry is represented by a set of aligned sequences with their probabilistic representation - called a profile hidden Markov model (HMM). The profile HMM is trained on a small representative set of aligned sequences that are known to belong to the family (the 'seed' alignment). This model is then used to search exhaustively against a large sequence database (e.g. UniProtKB) to find all homologous sequences. Those sequences that are significantly similar to the model are aligned to the profile HMM in order to provide the full alignment. Related Pfam entries may be grouped into sets, labelled as 'Clans'. These are typically large and divergent superfamilies, where a single profile HMM is insufficient to capture all members of a sequence.

1.2. Membrane Proteins

Membrane proteins (MPs) are a fundamental component of every cell as they mediate the information between both sides of the membrane. About 15% to 30% of all proteins in currently sequenced genomes correspond to MP (Arinaminpathy *et al.*, 2009, Almén *et al.,* 2009, Rawlings, 2018). The Human Genome Project estimated that 20% of human genes encode for MPs (International Human Genome Sequencing Consortium, 2001). More recent studies increased this percentage to 26-35% (Faberger *et al.,* 2010). Currently, the **Human Protein Atlas** (https://www.proteinatlas.org), contains 5480 genes encoding predicted MPs, about a 28% of all human genes (version 18.1, accessed on March 2019, search protein_class:Predicted membrane proteins).

The ancestral origin of MP, as for life on Earth, is not completely clear. Because of their relative simplicity, membranes were likely to be the first biologically relevant structures that appeared on the early Earth (Pohorille & Deamer, 2009). Probably, the initial function of membrane proteins was to act as ion channels that regulated cell volume through equilibrating the osmotic pressure between the interior of ancient cells and the environment. Acquirement of novel protein functions often starts with gene-duplication or gene-fusion events (Shimizu *et al.*, 2004). Subsequent evolutionary changes in sequence, and consequently in structure, permitted to acquire new functions. After millions of years of evolution, this mechanism has led to the outstanding variety of MPs.

MPs are interesting targets for the pharmaceutical industry -about 50% to 60% of drug targets are MP- (Almeida *et al.*, 2017, Overington *et al.*, 2006). Most drugs bind specifically to G protein coupled receptors (GPCRs) or ion channels (Hauser *et al.*, 2017, Santos *et al.*, 2016). However, due to the key role of lipids in cell membranes, lipids are suggested to constitute a novel target for drugs with effect in the properties of membranes, and thus modulating the activity of MPs (Lucio *et al.*, 2010). Moreover, now it is feasible to use small hydrophobic drugs that could directly target protein-protein interactions, hence modulating MP oligomerization (Yin & Flynn, 2016).

1.2.1. Biological membranes

The plasma membrane of modern eukaryotic cells is a ~30 Å semi-permeable bilayer of glycerophospholipid molecules that surrounds every cell. Some organelles (e.g. mitochondria, nucleus, Golgi apparatus and chloroplasts in plant cells), bacteria and also, some virus, are also surrounded by a membrane, with different thickness depending on the composition. Membranes also protect from oxidation and maintains electrochemical gradient (Müller *et al.*, 2008). Beyond encapsulating cell components, even the simplest cell contains hundreds of different proteins that mediate (in a highly regulated manner) the import and export of metabolites and/or polymers among other functions (Mulkidjanian & Galperin, 2010).

The model of biological membrane used nowadays is based in the fluid-mosaic membrane model (F-MMM), first proposed in 1972 by Seymour J. Singer and Garth L. Nicolson (Singer & Nicolson, 1972) and recently reviewed by Nicolson 40 years after (Nicolson, 2014). F-MMM features of the first model assumes that the membrane is composed by a bilayer of amphipathic lipids that possess a hydrophobic and a hydrophilic moiety, as originally advanced by Gorter and Grendel (Gorter & Grendel, 1925). Due to this amphipathic character, in aqueous solution they can rearrange themselves to organize the hydrophobic part facing each other, and the polar part oriented to the aqueous space. This model also describes proteins associated to the membrane, peripheral proteins attached to the polar head-groups of one layer, or integral proteins, those completely embedded in the membrane, proposed for the first time in this study. Another feature of the F-MMM is that both proteins and lipids are in constant movement, which contributes the fluidic name of the model. Finally, the model describes membranes as asymmetric, due to the movement of the different leaflets and/or their different composition, along with the movement of proteins. However, since this model was proposed, a huge amount of experimental data has appeared and urged for an adequation of the model to this information. For instance, the influence of cytoskeletal component to the mobility of integral proteins or the possibility of less mobile lipid-lipid domains. Additionally, some thermodynamic considerations regarding membrane deformation, curvature, compression and expansion were not possible to study at that moment without more recent developed techniques. A complete review of these considerations can be found in the work of Goñi (Goñi, 2014).

The lipid and protein composition of biological membranes varies among organisms, tissues, cells, and intracellular organelles (Ernst 2016). For instance, in plasma membrane the protein/phospholipid ratio is ~1:1 by weight, the same as in mitochondrial outer membrane, whereas the inner membrane has a higher protein/phospholipid ratio is >3:1 by weight and it is rich in cardiolipin (Müller *et al.*, 2008). The variability also extents to regions within the same membrane, the two membrane leaflets and also has a dynamic component. Eukaryotic cells and their organelles synthesize hundreds to thousands of lipid molecules differing in their molecular structures, physicochemical properties, and molar abundances. This stunning diversity derives from the combinatorial complexity of the lipid 'building blocks' (van Meer *et al.*, 2008). The LIPID MAPS Structure Database accounts for a classification of biological relevant lipids, so far it contains more than 43.000 unique lipid structures (http://www.lipidmaps.org/data/structure) (Fahy *et al.*, 2009), however, not all of them take part in membranes. There is growing evidence that lipids can modulate the function of proteins. As an example, it has recently been described how cardiolipin modulates the activity of respiratory complex I (Jussupow *et al.*, 2019).

The main lipidic components of membranes are glycerophospholipids (GPLs), sphingolipids and sterols (mainly cholesterol in mammals). The predominant GPLs (Figure 1.7) are phosphatidylcholine (PC), phosphatidylethanolamine (PE), phosphatidylserine (PS), phosphatidylinositol (PI) and phosphatidic acid (PA). All of them share the hydrophobic portion as diacylglycerol (DAG), with diversity in the hydrophilic portion, containing saturated or unsaturated fatty acyl chains of varying chains. In most of eukaryotic cells PC account for more than 50% of the total amount of phospholipid, while PE and PS account for 10% and 15% respectively, 10% corresponds to sphingolipids, 10% more to cholesterol and about a 1% to PI (Müller *et al.*, 2008). PC self-assemble spontaneously in a planar bilayer, with the lipidic tail pointing towards the centre of the bilayer and the polar head-groups interfacing with the water solvent (van Meer *et al.*, 2008) A nice review of lipid composition and its variability in membranes was recently published by Harayama and Riezman (Harayama & Riezman, 2018). An important aspect of cell membrane is their asymmetry. Sphingolipids are usually located on the outer membrane leaflet, while negatively charged lipids like PS and PI, are found in the inner leaflet (Honigmann & Pralle, 2016). Biomembranes typically consist of some hundred different lipid species that often are distributed asymmetrically between both leaflets of the bilayer (Alberts *et al.*, 2014).

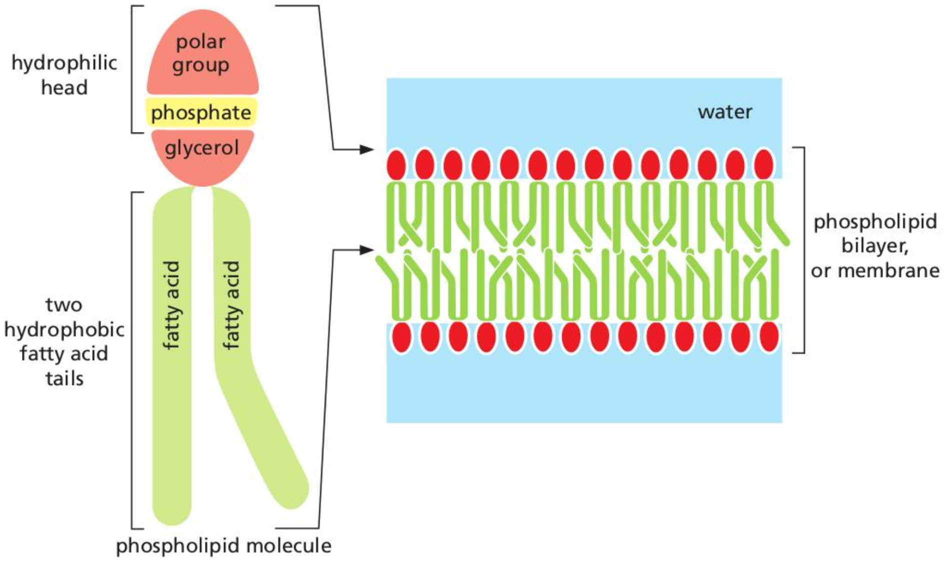


Figure 1.7. Schematic representation of a GPL and its disposition in a bilayer. Adapted from Alberts *et al.*, 2014.

1.2.2. Classification and Distribution of MP

MPs are divided in two major classes, **integral,** when they are completely embedded in the membrane, and **peripheral**, when they are temporarily attached either to the lipid bilayer or to integral proteins. Integral MPs can be classified in two broad categories: **monotopic proteins**, that are attached to the membrane from one side, only one helix spans the membrane and serves as an anchor for the protein (i.e. σ1R, UniProtID: Q99720). Monotopic proteins can also serve as a connector between an extracellular domain and a cytosolic domain (i.e. HER2, UniProtID: P04626). The other type of integral proteins is **bitopic** or **polytopic proteins**. Both types are also known as transmembrane proteins and have two or more helices (or sheets) spanning the membrane. Examples of such types are receptors or transporters (i.e. CNR1, UniProtID: P21554 for α-helices or OMR, UniProtID: Q9K0U9 for a β-barrel). A representation of these structures is shown in Figure 1.8.

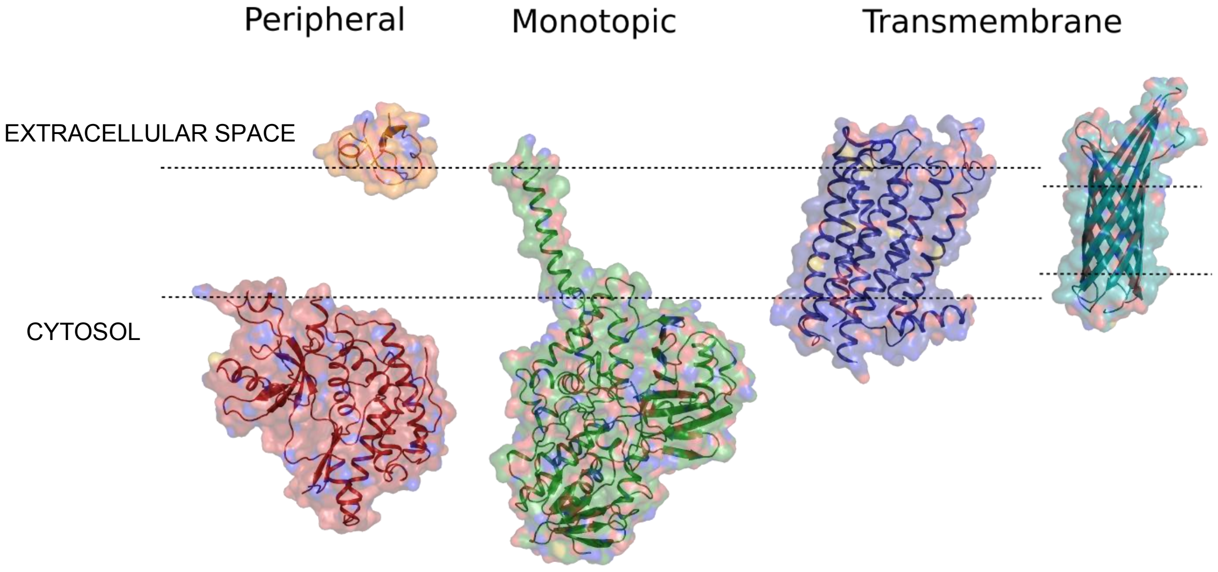


Figure 1.8. Representation of different membrane proteins. PDBID 6CS9 (orange) and PDBID 4PLA (red) are peripheral proteins. PDBID 2Z5X (green) corresponds to a monotopic protein, PDBID 6BQG (blue) is a transmembrane α-helical protein and PDBID 3QRA (cyan) represents a transmembrane β-barrel. Coordinates from OPM, image generated with PyMol.

1.2.3. Folding of Membrane Proteins

The folding and membrane insertion of MPs is still not completely understood. One problem is the lack of suitable approaches that allow investigating the process by which polypeptides insert and fold into membranes (Serdiuk *et al.*, 2017). The formation of most polytopic membrane proteins depends critically on co-translational insertion and folding in the cytoplasmic membrane, which is facilitated by ribosome-translocon complexes (Reid & Nicchitta, 2012)

The current model suggests membrane protein folding is achieved in two steps: i) insertion of the independent helices into the membrane bilayer and ii) the inserted helical segments fold into a functional TM structure by lateral interactions (Popot *et al.*, 1987 and 1990, White *et al.*, 2001). This model was further refined by the addition of the third step, in which binding of prosthetic groups, folding of long inter-helical loops, entry of other regions of polypeptide chain into the TM region or oligomerization of subunits occurs (Engelman *et al.*, 2003). Thus, the complex of protein polypeptide with bound lipids perhaps is critical for achieving a mature conformation without membrane disruption during initial insertion or later during fusion of small vesicles with the (plasma) membranes (Müller *et al.*,2008)

The length of the lipid-exposed hydrophobic segments is approximately equal to the hydrophobic membrane thickness, in order to avoid unfavourable exposure of hydrophobic surfaces to a hydrophilic environment. Yet, proteins that are encountered in a membrane can have different lengths of their hydrophobic parts compared to the thickness of the membrane that embeds them. This leads to what it is known as hydrophobic mismatch (Killian, 1998). Hydrophobic mismatch is energetically unfavourable and thus tends to be reduced by two main mechanisms: either i) the membrane thickens or ii) the protein (normally with a single TM helix) tilts or bends (Bowie, 2005).

1.2.4. Structure of Membrane Proteins

The structure and stability of proteins are the consequence of a delicate balance between protein–protein and protein–solvent interactions (White *et al.*, 1999). For membrane proteins, the surface-exposed residues are dominantly nonpolar so that they make favourable contacts with the hydrocarbon core of lipid bilayers (Adamian *et al.*, 2005). MPs are built from TM segments than span the membrane. A fundamental aspect of the structure of transmembrane proteins is the TM region topology, that is, the number of transmembrane segments, their position in the protein sequence and their orientation in the membrane. The segments can either be in the form of hydrophobic α-helices or β-sheets, which create a barrel. While α-helical MPs exist in all super-kingdoms, β-barrel MPs are only found in bacterial porins and (because of their supposed endosymbiotic origin (Margulis, 1967)) also in the inner membrane of mitochondria and chloroplasts of eukaryotic cells. TM segments contain a high proportion of non-polar amino acids (aliphatic and aromatic) as it could be expected for a hydrophobic environment (von Heijne, 1992, Li & Deber, 1992, Deber & Goto, 1996). To facilitate the anchorage of TM segments, on either side of the hydrophobic transmembrane domain are often found tyrosine and tryptophan residues (Landolt-Marticorena *et al.*, 1993, Arkin & Brunger,1998, Ulmschneider & Sansom, 2001). In the interface region, the polar groups of these amino acids can interact with the phosphate groups, while the hydrophobic rings can interact with the lipid-chains. It has even been observed that there is a preference for Tyrosine and Tryptophan to point toward the phosphlipid headgroups, that is, residues that are located outside the interface region point the polar groups inward, while the ones located outside point the polar groups outward (Hedin *et al.*, 2011, Yeagle, 2016). Besides, charged amino acids are scarce in the transmembrane domain, however, positively charged residues tend to concentrate in the cytosolic part of this kind of proteins, what it is known as the 'positive inside rule' (von Heijne, 1989 and 1992). This amino acid distribution differs from globular proteins, as globular proteins are surrounded by a hydrophilic environment, with polar and charged residues expected at the surface of the protein.

The differences of the environment between globular and membrane proteins, are also responsible for other specific features of membrane proteins that differ from globular proteins.

1.2.5. Functions of Membrane Proteins

Membrane proteins serve many functions. Some transport particular nutrients, metabolites, and ions across the lipid bilayer. Others anchor macromolecules to the membrane on either side. Still others function as receptors that detect chemical signals in the cell's environment and relay them into the cell interior, or work as enzymes to catalyse specific reactions nearby the membrane. Each type of cell membrane contains a different set of proteins, reflecting the specialized functions of the particular membrane some of them represented in figure 1.9. including:

- **Junction**, serve to connect and join cells together.
- **Enzymatic catalysis**, some steps of the metabolic pathways occur in, or near, the membrane.
- **Transport**, responsible for facilitated diffusion and active transport.
- **Recognition**, MPs may function as markers for cellular identification, proteins which allow cells to attach to other cells to enable cell communication.
- **Anchorage**, attachment point for cytoskeleton and extracellular matrix.
- **Transduction**, functioning as receptors and signal transducers. Signal transduction is the process by which an extracellular signaling molecule activates a membrane receptor that in turn alters intracellular molecules creating a response
- **Regulation**, MPs also participate in the regulation of membrane composition and stability.

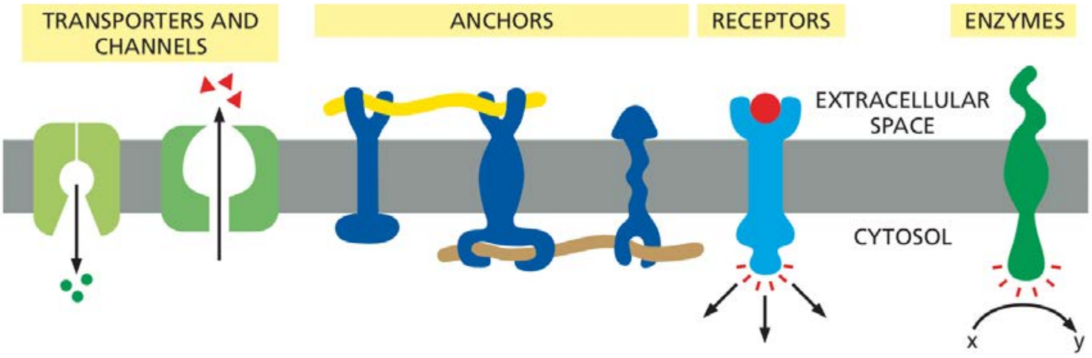


Figure 1.9. Schematic representation of some functions of MPs. Adapted from Alberts *et al.*, 2014

1.2.6. Determination of Membrane Proteins Structure

Knowing the 3D structure of proteins is fundamental to understand their function. Nevertheless, this is a complicated task, especially for MPs due to their instability in the conditions that are usually required for structure determination. First the difficulty in over expression of eukaryotic MPs in prokaryotic cells, mainly *Escherichia coli*, where the membrane lipid composition differs from an eukaryotic native membrane environment. They are affected by the membrane and various specific factors, such as cholesterol content or the hydrophobic thickness of the lipid bilayer, but also influence the membrane structure itself. Second, the purification process, in order to obtain sufficient amounts of protein that usually involves some detergent, so the dissociation from their native coupled lipids during this process can render non-functional proteins (Martinac & Vandenberg, 2015). Finally, in the case of X-ray structures, the crystallization of these insoluble proteins has to be done in a hydrophilic environment in order to maintain the native 3D structure of these kind of proteins (Almeida *et al.*, 2017).

All these aspects above contribute to the technical experimental difficulties in the structural characterization of MPs, which explains their relatively low number in the Protein Data Bank (PDB) (Berman *et al.*, 2000), despite their high proportion in the human proteome. Different strategies have been followed in order to avoid these difficulties, such as thermostabilizing mutations, the construction of chimeras that stabilize specific conformations or fusing tags which overcome the insolubility of the MP in aqueous solution (Rawlings, 2018). 3D structures of various MPs have been obtained in the recent years by several experimental methods, mostly X-ray crystallography, cryo-electron microscopy (cryo-EM) and nuclear magnetic resonance (NMR). Traditionally X-ray crystallography has been the most used technique and about 90% of deposited structures in the PDB have been solved using it (PDB statistics). X-ray crystallography requires proteins to be packed together into a stable organized crystals that later are bombed with X-rays to produce a diffraction pattern. However, many proteins are too floppy to line up into a crystal. When it happens, for instance for big protein complexes, it is still possible to get a structure at high resolution. Cryo-EM, a technique that has recently gained a lot of popularity among structural biologists (Almeida *et al.*, 2017), has the advantage that the protein is kept in a very thin layer and then freezed. The protein does not need to be crystallized, a great advantage compared to X-ray crystallography. After freezing, an electron beam pass through the sample and a specially designed high-tech camera captures the electrons to form an image. This camera can also record a movie instead of a steady picture allowing to see different states or conformations within a single sample. Finally, a computer algorithm sorts the images, and reconstructs a 3D model from the collection of 2D images. Until recently, its main drawback was the lower resolution obtained for membrane compared to X-ray crystallography (Almeida *et al.*, 2017). However, after the first MP structure solved by cryo-EM without crystallization there has been an explosion of new structures solved by this technique (Liao et al 2013) and methodological improvements have allowed to reach X-ray resolution. A recent example is the cannabinoid receptor 1 in complex with a G protein at 3 Å (PDBID 6N4B, Kumar *et al.*, 2019).

By April of 2019, the PDB contained >5300 structures of membrane proteins, just a scarce 3.5% of all determined structures, despite MPs represent 25% of the human proteome. This number include multiple submissions of the same protein, in different conditions or with different bound ligands. The group of Stephen White handles a database of MPs where they monitor the number of unique MPs with available structure (http://blanco.biomol.uci.edu/mpstruc/), which currently is 884, barely a 20% of them corresponding to Human MPs (Figure 1.10). This low number is in contrast with the high amount of sequence of MPs, more than 77.400 of the reviewed entries found in UniProt (The UniProt Consortium, 2019). Considering just Human reviewed sequences in Uniprot, about 20.000, a quarter of them correspond to MPs.

The complete understanding of the structure of MPs require the study of the lipids surrounding the protein. It is well known that lipids can modulate the structure and function of MPs (Hedger & Sansom, 2016, Simons, 2016, Stangl & Schneider,2015). Indeed, many crystallographic structures reveal specific binding sites for lipids (Hedger & Sansom, 2016). Thus, lipid-MP interactions have caught the attention of many researchers, as they are central to the function of cellular membranes (Battle *et al.*, 2015). To do so, Mass Spectrometry (MS) has become a very useful technique for the study of these interactions (Barrera *et al.*, 2013, Zhou & Robinson, 2014, Calabrese & Radford, 2018), as it can reveal which lipids are attached to MPs, association between proteins and even elucidate different binding modes.

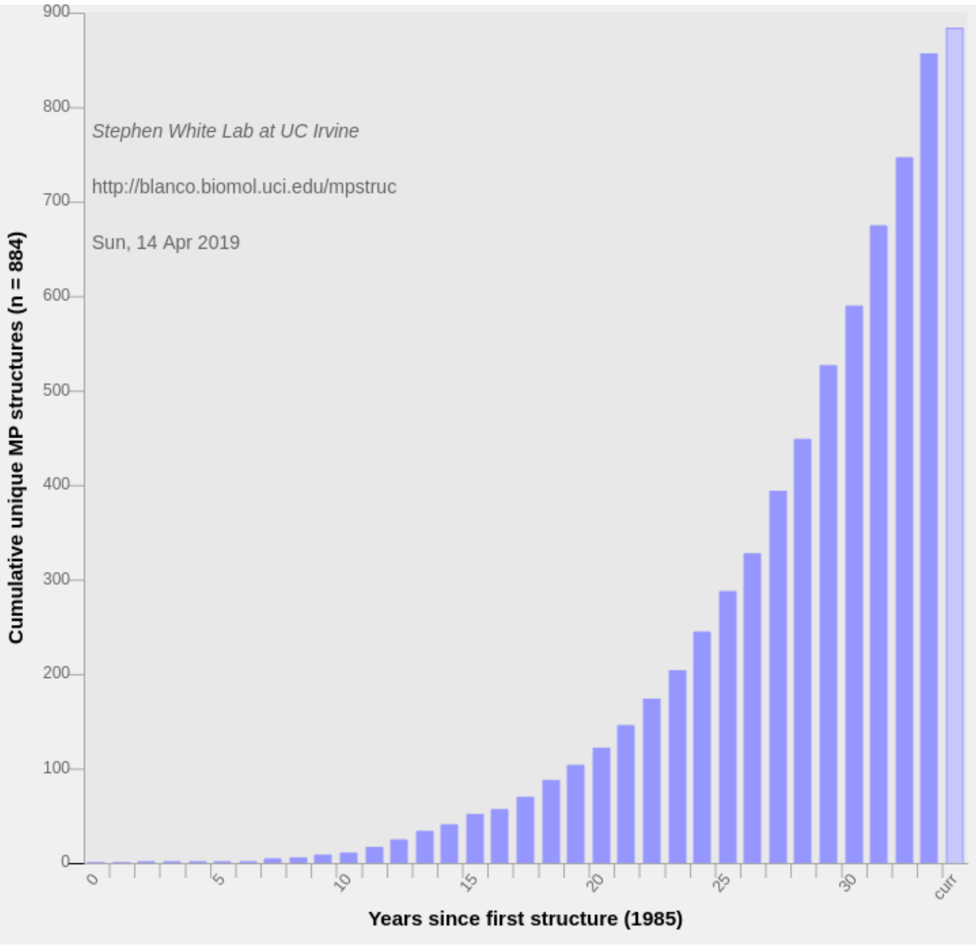


Figure 1.10. Cumulative number of unique MP structures deposited in the PDB. From http://blanco.biomol.uci.edu/mpstruc

1.3. G Protein-Coupled Receptors

1.3.1. Importance of GPCRs and classification

G Protein-Coupled Receptors (GPCRs) comprise the largest family of cell-surface receptors encoded in the human genome. The GPCR family is integrated by almost 800 different receptors, half of them olfactory receptors. The first phylogenetic classification of GPCRs was presented by Kolakowski in 1994, denoted as the A-F classification (Kolakowski, 1994). Almost ten years later, Fredriksson and colleagues (Fredicksson *et al.*, 2003) showed that human GPCRs can be divided in five main families, namely Rhodopsin (class A), Secretin (class B1), Adhesion (class B2), Glutamate (class C) and Frizzled/Taste2 (class F). Among these 5 main classes, class A is by far the largest and includes important receptors like the representative GPCR rhodopsin, olfactory receptors, opioid, adrenergic, chemokine, angiotensin, histamine, dopamine or adenosine receptors (Figure 1.12). Despite their high diversity in sequence, all GPCRs share one common structural feature: seven TM α-helices joined by a set of three extracellular and three intracellular loops (Katritch *et al.*, 2013). Main structural differences among classes are located in the extracellular part (Figure 1.11).

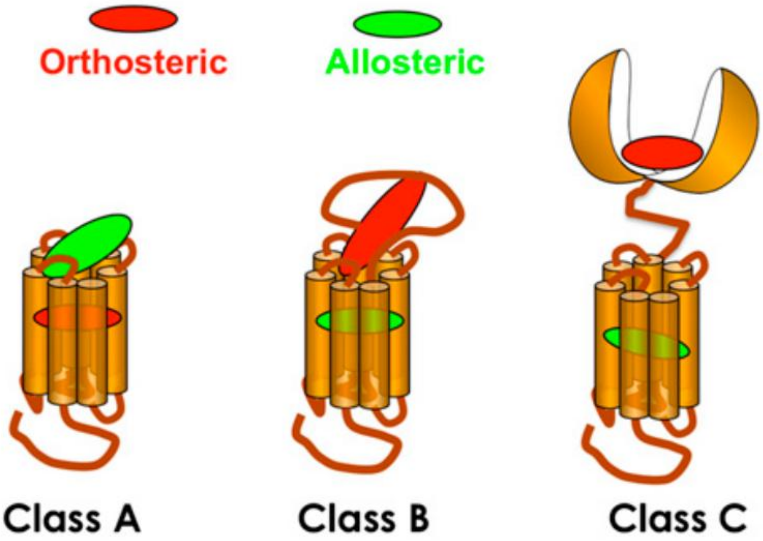


Figure 1.11. Schematic representation of Class A, B and C of GPCRs, showing the similar 7 TM disposition and differences in the extracellular part, as in the common binding site for orthosteric and allosteric ligands. Adapted from van der Westhuizen *et al.*, 2015.

The differences among class A GPCRs arise in the length and conformation of these loops and the extracellular N-terminal and intracellular C-terminal, moreover, the binding site location is also different between classes (Shonberg *et al.*, 2015, van der Westhuizen *et al.*, 2015) (Figure 1.11). Even though GPCRs share high structural similarity, their ligands are diverse in nature and range from a photon to odorants and pheromones, endogenous signals such as neurotransmitters, proteases, or even peptides. This diversity makes these receptors to be involved in many physiological processes, such as visual sense, taste, smell and regulation of behaviour and mood or immune system among others. Therefore, GPCRs are considered the largest family of targets for approved drugs, around one third of drugs in the market (Tehan *et al.*, 2014, Hauser *et al.*, 2017), hence, the study of GPCRs is of utmost importance for pharmaceutical companies in order to obtain new drugs and for academia to completely understand their varied functionality. In fact, in 2012 Brian Kobilka and Robert Lefkowitz were awarded with the Nobel Prize in Chemistry for the discoveries that reveal the working of such proteins.

The first crystal structure of a GPCR, bovine rhodopsin (PDBID 1F88), was resolved in the year 2000 (Palczewski *et al.*, 2000). It would take 7 years to obtain the structure of a second type of receptor of this family, the human β_2-adrenergic receptor (PDBID 2RH1) (Cherezov *et al.*, 2007). Both (and subsequently solved) structures corresponded members of the Class A family. It was not

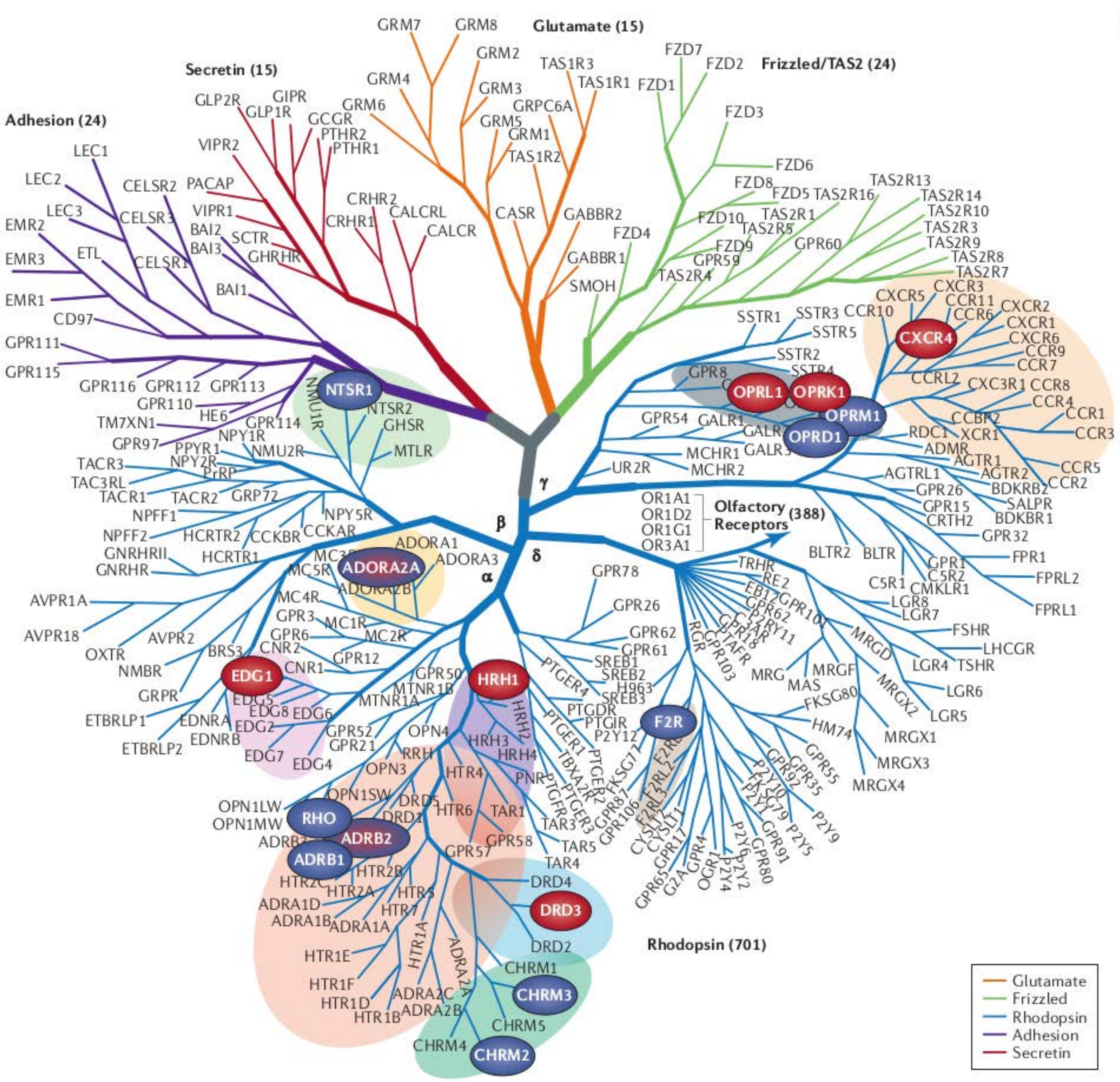


Figure 1.12. The resulting phylogenetic tree from Fredriksson classification. Adapted from Stevens *et al.*, 2013.

until 2013 that the first structures of non-Class A GPCR were solved, first the Class F smoothened receptor (PDBID 4JKV) (Wang *et al.*, 2013) and second the Class B corticotropin-releasing factor receptor 1 (PDBID 4K5Y) (Hollenstein *et al.*, 2013). 2014 brought the first Class C structure, the metabotropic glutamate receptor 1 (PDBID 4OR2) (Wu *et al.*, 2014). Since the first crystal structure, and specially along the last few years, hundreds of new structures have appeared. By April of 2019, just for class A, 283 complete GPCR structures are available, corresponding to 52 unique receptors. The total number of GPCR structures solved by year and grouped by families is shown in Figure 1.13. Classes B1, C and F, are also represented with 16, 8 and 12 structures respectively, however just 6 unique receptors for class B1, 2 for class C and 2 for Class F. No structures for class B2 are available so far, or at least not openly available. Altogether it sums 319 different crystal structures from 62 different receptors for all GPCRs, this does not even represent a 10% of all different proteins in this superfamily.

Multiple sequence alignment has shown that, despite the low sequence identity, there are highly conserved position and motifs in each of the 7 TMs. This feature was considered by Ballesteros and Weinstein to define a unified numbering scheme based on the most conserved positions in the TMs of class A family (Ballesteros 1995). According to these numbering system, each residue is defined by two numbers: the first corresponds to the TM number and the second indicates its position relative to the most conserved residue, arbitrarily assigned to 50 (numbers <50 correspond to residues towards the N-terminus of the TM helix relative to position 50, whereas numbers >50 correspond to residues towards the C-terminus). As an example, the most conserved residue in the third TM helix is R3.50, the preceding amino acid in sequence is E3.49 and the following is R3.51. The Ballesteros-Wenstein notation (Ballesteros & Weinstein, 1995) allows a straightforward comparison between residues of different receptors. In order to expand this nomenclature to Class B, C and F, a new numbering scheme has been proposed (Isberg *et al.*, 2015), the so-called generic numbering.

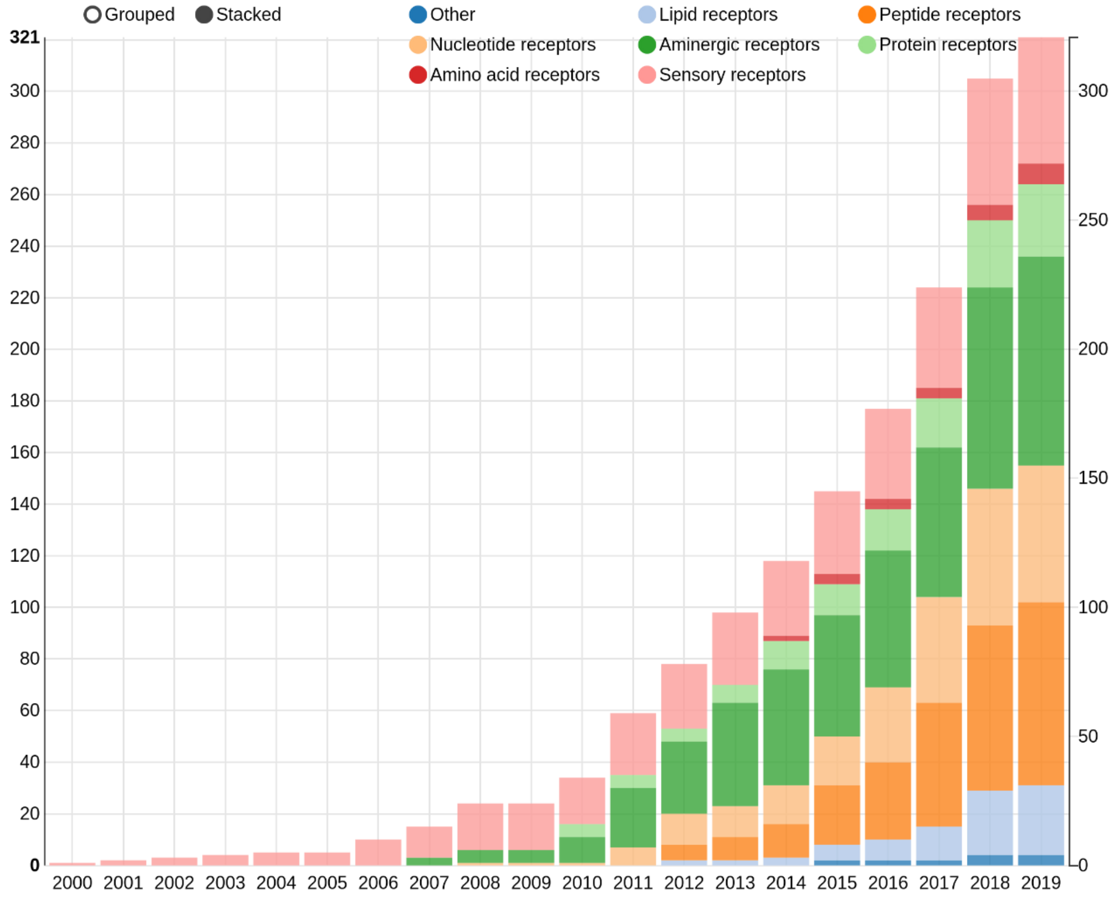


Figure 1.13. Cumulative number of crystallized receptors per year for class A GPCRs. From gpcrdb.org

1.3.2. GPCR Signaling and Activation

GPCRs are proteins that, upon agonist activation, trigger intracellular signal transduction after a conformational rearrangement by recruiting cytosolic proteins. There are two main pathways of signal transduction carried out by these receptors. The canonical signalling occurs via coupling to a heterotrimeric G-protein (Gα, Gβ and Gγ), followed by the activation of second messengers, such as cyclic AMP (cAMP), calcium or inositol phosphate, which in turn may initiate a broad spectrum of downstream signaling pathways. There is a second mechanism via a G-protein independent pathway that involves the phosphorylation of the C-terminus by specific GPCR Kinases (GRK) and a recruitment of multifunctional proteins, arrestins, which are believed to play a central role in receptor desensitization (Dwivedi *et al.*, 2018, Hilger *et al.*, 2018). Figure 1.14 displays a scheme of these mechanism.

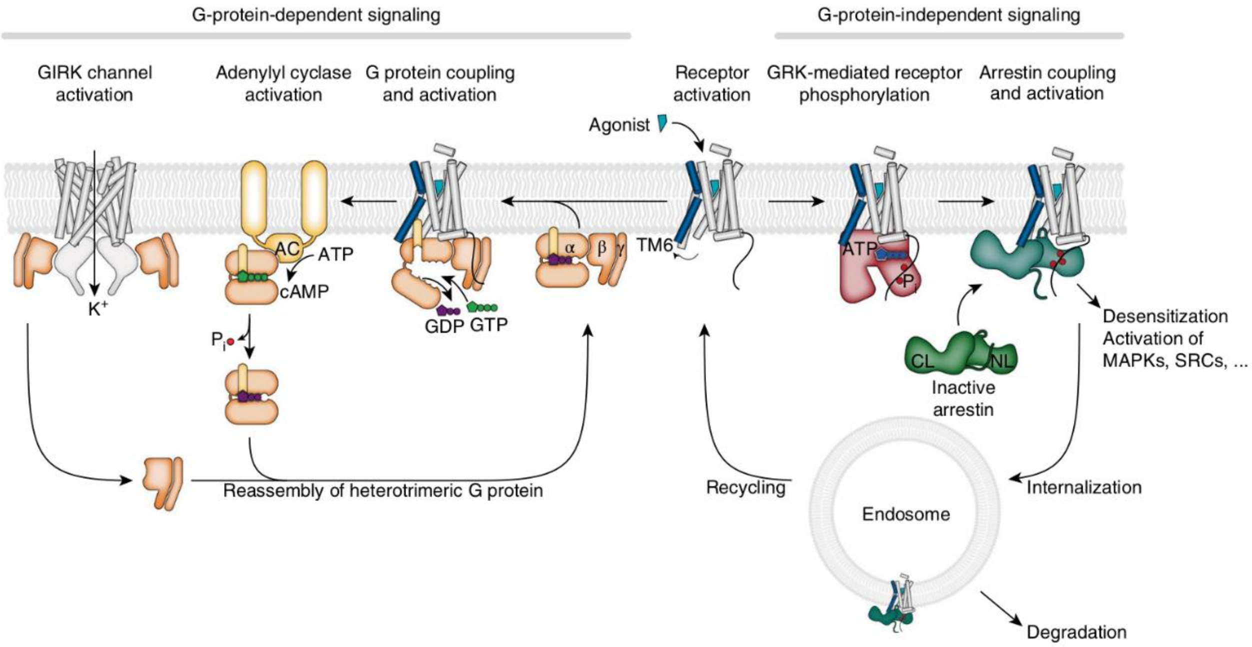

G-protein-dependent signaling
G-protein-independent signaling
GIRK channel activation
Adenylyl cyclase activation
G protein coupling and activation
Receptor activation
GRK-mediated receptor phosphorylation
Arrestin coupling and activation
Agonist
AC
ATP
cAMP
K+
Pi
GDP GTP
α
β
γ
TM6
ATP
P
CL
NL
Inactive arrestin
Desensitization
Activation of MAPKs, SRCs, ...
Reassembly of heterotrimeric G protein
Recycling
Endosome
Internalization
Degradation

Figure 1.14. GPCR signal transduction mechanism. Adapted from Hilger *et al.*, 2018.

Some ligands can trigger one of these pathways but not the other for a specific receptor, what it is known as bias signaling. The complete activation process is not fully understood, however, what we do know is that it involves the transmission switch, an inward movement of W6.48, the Tyr tooggle, corresponding to the outward movement of Tyr in the conserved motif NPxxY in TM7 and a disruption of the ‘ionic lock’, an ionic interaction between the conserved DRY motif in TM3 and E6.30 in TM6, producing the opening of TM6 in the cytosolic region in order to couple transducer proteins and movements in TM3, TM5 and TM7 (Park, 2012, Trzaskowski *et al.*, 2012). It has been recently postulated that different water networks are formed in the active or inactive states that help in the signal transduction (Venkatakrishnan *et al.*, 2019).

The large amount of different crystal structures available, in different conformations, with different ligands (agonists, antagonist or inverse agonist, and even allosteric modulators and/or in complex with signaling proteins such as G proteins or arrestins), have contributed extensively to the understanding of the activation process of this kind of receptors. However, the complete comprehension of the process requires further investigation of the conformational dynamics of receptor-transducer complexes, since crystal structures show a static picture of the receptor and cannot unveil the complex activation mechanism.

1.3.3. Importance of Functional Water Molecules in GPCRs

With the emergence of high-resolution structures, it become clear that some water molecules were located inside the protein core and that they are implicated in many processes such as ligand binding, catalytic reactions or even forming part of the functional structure of proteins. Hence, water molecules in proteins play a crucial role in their functionality. Thanks to their physical and chemical properties water can act as a solvent but also can be incorporated to the secondary structure of proteins, mediating in protein-protein contacts. Moreover, they may act as prosthetic groups for a proper protein function and participates in the folding of the three-dimensional structure. Some of these water molecules that act as a part of the protein are also called structural waters (Zhong *et al.*, 2011). For all these reasons water has been the subject of many studies for decades, since Perutz, Kendrew and others confirmed the presence of water inside proteins in the 1950s (Perutz *et al.*, 1960, Kendrew *et al.*,1958).

Water, with its tetrahedral coordination of oxygen atoms, can perform hydrogen bonds with up to 4 neighbours but many experiments indicate that, on average and at room temperature, this number is about 3.5 (Ball, 2008). Due to its small size, the dipolar nature cause by the different charge distribution between atoms and the capacity to act as a hydrogen bond donor and acceptor, water is involved in a wide diversity of cellular functions and can be considered one of the most important molecules in living systems.

In an apolar solvent like the center of a membrane bilayer that has a low dielectric constant and no competitive hydrogen bonding potential, hydrogen bond contributions could be even stronger than for globular proteins (Bowie, 2011).

The fact that secondary structure persists in denatured membrane proteins (Haltia & Freire, 1995, Lau & Bowie, 1997, Riley *et al.*, 1997, Engelman *et al.*, 2003) suggests that backbone hydrogen bonds in the membrane are more stable than in water solution where isolated helices are usually unstable (Bowie, 2011). Moreover, backbone hydrogen bonds in TM helices are shorter and more regular on average than backbone hydrogen bonds in water soluble protein helices (Hildebrand *et al.*, 2004, Joh *et al.*, 2008, Page *et al.*, 2008) Moreover, hydrogen bonds can be strong if there is evolutionary pressure driving their optimization. It could be thought that the most potent evolutionary pressure for optimizing hydrogen bonds would only occur for functional reasons rather than to stabilize the protein (Bowie, 2011).

Many works have been focused on the role of water in proteins, based on analysis of high-resolution proteins deposited in PDB, classifying them in stable/unstable, happy/unhappy, cold and hot or of relevance (Ahmed *et al.*, 2011). Hot or ‘unhappy’ waters are those that are mobile or with a significant internal energy, unstable in certain parts of the protein such as in contact with hydrophobic regions. In this situation is difficult for water molecules to maximize the number of hydrogen bonds, therefore are easily moved towards the bulk water. However, they can be found in high-resolution X-ray structures and the may play a role in protein macromolecular structure or in their functionality (Spyrakis *et al.*, 2017). On the other hand, cold waters are those that form part of the protein structure or protein machinery, with less tendency to displacement to the bulk. In general, cold waters have a lower B-factor and higher residence time than hot waters, among other features reviewed by Spyrakis and collaborators (Spyrakis *et al.*, 2017).

Due to the need of high-resolution structures to solve internal water positions, an increasing number of tools and methods are available to predict water placement, following different approaches, from geometrical scoring to Molecular Dynamics simulations, a recent review of them, including comparison and classification can be found in the work of Nittinger and collaborators (Nittinger *et al.*, 2018) However, none of them can be accessed from a web server and they try to predict position for a wide variety of proteins.

For GPCRs, water plays and important role mediating in GPCR activation. Since the first crystal of a GPCR appeared (rhodopsin, PDBID 1F88) showing internal waters, hundreds of structures are currently available for different receptors, with different ligands and also in different conformations. Many of them solved with high resolution, enough to capture internal water molecules. In the recent years many works have been devoted to understanding the role of these waters (Pardo *et al.*, 2007, Angel *et al.*, 2009 (A and B), Yuan *et al.*, 2014, Yuan *et al.*, 2013, Sun *et al.*, 2014) using statistical analysis and Molecular Dynamics. More recently, the group of Kobilka has analysed water networks inside GPCRs, differentiating among active and inactive conformations of the same receptor (Venkatakrishnan *et al.*, 2019). All these works share evidences to the understanding of water contribution to the correct functionality of these receptors.

1.4. Bioinformatics for the analysis of Membrane proteins

Bioinformatics is an interdisciplinary field that combines biology, computer science, information engineering, mathematics and statistics to analyse and interpret biological data.

Despite the generalized thinking, it was protein analysis and not DNA analysis that originated the emergence of bioinformatics field. Currently, a major part of this field focuses in genetics since the Human Genome Project published the sequence of the human genome in 2001, after more than ten years of work and more than 2 billion $ expended. The advances in next generation sequencing (NGS) a few years later boosted the amount of data and, therefore, the growth of fields such as molecular biology, evolutionary biology, population genomics and many others. Nowadays sequencing a complete human genome costs around 1000$ and takes just a few days (https://genome.gov National Human Genome Research Institute).

What follows is a brief recompilation of relevant facts for the development of bioinformatics as a discipline. For a more detailed view see (Gauthier *et al.*, 2018, Clément *et al.*, 2018, Hogeweg, 2011). Since the publication of the first sequence of a protein, bovine insulin (Sanger 1951), the number of available sequences has not stop growing. The accumulation of sequences urged to use computers in molecular biology already in the decade of 1950s. On the other hand, one of the major achievements for structural biology in the XXth century was the determination of the first crystal structure of a protein, carried out by John Kendrew It was the structure of myoglobin (Kendrew, 1958). Solving structures from an X-ray diffraction pattern required developing computational applications.
It was Margaret Dayhoff who first used a computer to analyse the protein sequences obtained (Dayhoff 1962). It was then, that the first known bioinformatics software to solve the issue of sort and classify this protein sequences appeared. Dayhoff along with Robert Ledley, developed COMPROTEIN (Dayhoff & Ledley, 1962), a computer program written in FORTRAN on punch-cards designed to determine the sequence of Edman peptides, this program was the first one of what we currently call *de novo* sequence assembler, implemented with the three-letter code to represent amino acids. In an effort to reduce the size of the data files needed to describe the sequence of amino acids in a protein Dayhoff developed the one-letter code in her work with Eck in 1965, the *Atlas of Protein Sequence and Structure* (Dayhoff, 1965), the first biological sequence database. They hypothesized that protein sequences would reflect the evolutionary history of species. During those years, the incipient field of molecular evolution grew fast. Computers allowed to generate phylogenetic trees (Doolitle & Blombäck 1964, Fitch & Margoliash, 1967). The concept of orthology was defined in 1970 by Walter M. Fitch to describe homology that resulted from speciation event (Fitch, 1970). He found that the sequence differences between orthologs of various species was proportional to the evolutionary differences between those species.
Dayhoff, Schwartz and Orcutt achieved another breakthrough in bioinformatics. They created the first probabilistic model of amino acid substitution (Dayhoff *et al.*, 1978) based on the observation of more than 1500 point accepted mutations (PAMs) in the phylogenetic trees of 71 families of proteins sharing at least 85% of identity. As a result, they got a 20x20 asymmetric substitution matrix, with the probability of each amino acid to change in a given evolutionary interval. The PAM matrix introduced the substitution rate as a measure of evolutionary change.

However, until the end of 1970s, ‘minicomputers’ were the size of a common refrigerator, its size complicated their acquisition for individuals or small group. It was in 1974 when the first desktop computer emerged. Rapidly, more desktop computers were available and with them the start of an extensive use, accessible to the general public. Biologist did not stay behind and started to develop specific software for these machines. In 1984 the first software collection for sequence analysis appeared (Devereux *et al.*, 1984). It was composed by many tools to manipulate RNA, DNA and protein sequences. All and all, desktop computers brought and explosion in the amount of software and data that has not stop yet.
The arrival of the Internet in the early 1990s supposed a new revolution in science and technology. Over all, to the accessibility of biological information. In 1993 the EMBL made available on the web the first nucleotide sequence database, created 10 years before distributed in books and CDs so

far, their Nucleotide Sequence Data Library. The next year appeared NCBI's website, including BLAST tool, which allows to perform pairwise alignments. With the expansion of the Web, bioinformatics tools and resources broadened, many through web server and graphical user interfaces. Even currently, with an increasing number of available tools, this amount is insignificant compared with the volume of biological data deposited in public databases.

All and all, to gather, sort and classify the huge amount of biological information and facilitate the extraction of knowledge regarding their relationships is of paramount importance for bioinformaticians. However, many problems could arise in the automatization of this organization. The different data types and diversity in nature of the information, along with the rapidness in the growth of the volume of data, that will continue in the coming years, requires the development of big data analytics techniques in order to solve this issue, such as machine learning (Kashyap *et al.*, 2016, Min *et al.*, 2017, Angermueller *et al.*, 2016). Machine learning has been recently used in many different areas of bioinformatics such as prediction and/or refinement of secondary structure (Wang *et al.*, 2016, Spencer *et al.*, 2015, Cao *et al.*, 2017, Yang *et al.*, 2017), in the prediction of protein-ligand interactions, many of these studies have been recently reviewed by Colwell (Colwell, 2018), or in drug discovery (Stephenson *et al.*, 2018, Lo *et al.*, 2018, Vamathevan *et al.*, 2019).

One of the many areas of bioinformatics is **structural bioinformatics**, the branch related to the analysis and prediction of the three-dimensional structure of biological macromolecules such as proteins, RNA, and DNA. It deals with generalizations about macromolecular 3D structure such as comparisons of overall folds and local motifs, principles of molecular folding, evolution, and binding interactions, and structure/function relationships, working both from experimentally solved structures and from computational models.

This thesis has focused on the analysis of the **structure of membrane proteins**, and the development of tools for this purpose, making use of the enormous quantity of available data in public databases.

1.5. References

Adamian, L., Nanda, V., DeGrado, W. F., & Liang, J. (2005). Empirical lipid propensities of amino acid residues in multispan alpha helical membrane proteins. *PROTEINS: Structure, Function, and Bioinformatics*, *59*(3), 496-509.

Ahmed, M. H., Spyrakis, F., Cozzini, P., Tripathi, P. K., Mozzarelli, A., Scarsdale, J. N., Safo, S. A. & Kellogg, G. E. (2011). Bound water at protein-protein interfaces: partners, roles and hydrophobic bubbles as a conserved motif. *PloS one*, *6*(9), e24712.

Alberts, B., Bray, D., Hopkin, K., Johnson, A. D., Lewis, J., Raff, M., Roberts, K. & Walter, P. (2014, Fourth Edition). *Essential cell biology*. Garland Science.

Almeida, J. G., Preto, A. J., Koukos, P. I., Bonvin, A. M., & Moreira, I. S. (2017). Membrane proteins structures: A review on computational modeling tools. *Biochimica et Biophysica Acta (BBA)-Biomembranes*, *1859*(10), 2021-2039.

Almén, M. S., Nordström, K. J., Fredriksson, R., & Schiöth, H. B. (2009). Mapping the human membrane proteome: a majority of the human membrane proteins can be classified according to function and evolutionary origin. *BMC biology*, *7*(1), 50.

Aloy, P., Ceulemans, H., Stark, A. & Russell, R. B. (2003). The relationship between sequence and interaction divergence in proteins. *J Mol Biol* 332, 989–998

Andreeva, A., Howorth, D., Chandonia, J. M., Brenner, S. E., Hubbard, T. J., Chothia, C., & Murzin, A. G. (2007). Data growth and its impact on the SCOP database: new developments. *Nucleic acids research*, *36*(suppl_1), D419-D425.

Andreeva, A., Howorth, D., Chothia, C., Kulesha, E., & Muzin, A. G. (2014). SCOP2 prototype: a new approach to protein structure mining (vol 42, pg D310, 2014). *Nucleic Acids Research*, *42*(18), 11847-11847.

Angel, T. E., Chance, M. R., & Palczewski, K. (2009). Conserved waters mediate structural and functional activation of family A (rhodopsin-like) G protein-coupled receptors. *Proceedings of the National Academy of Sciences*, *106*(21), 8555-8560. (A)

Angel, T. E., Gupta, S., Jastrzebska, B., Palczewski, K., & Chance, M. R. (2009). Structural waters define a functional channel mediating activation of the GPCR, rhodopsin. *Proceedings of the National Academy of Sciences*, *106*(34), 14367-14372.

Angermueller, C., Pärnamaa, T., Parts, L., & Stegle, O. (2016). Deep learning for computational biology. *Molecular systems biology*, *12*(7), 878.

Arinaminpathy Y, Khurana E, Engelman DM, Gerstein MB (2009) Computational analysis of membrane proteins: the largest class of drug targets. Drug Discov Today 14: 1130-1135.

Arkin, I. T., & Brunger, A. T. (1998). Statistical analysis of predicted transmembrane α-helices. *Biochimica et Biophysica Acta (BBA)-Protein Structure and Molecular Enzymology*, *1429*(1), 113-128.

Arnold, F. H. (2018). Directed evolution: bringing new chemistry to life. *Angewandte Chemie International Edition*, *57*(16), 4143-4148.

Audagnotto, M., & Dal Peraro, M. (2017). Protein post-translational modifications: In silico prediction tools and molecular modeling. *Computational and structural biotechnology journal*, *15*, 307-319.

Ball, P. (2008). Water as an active constituent in cell biology. *Chemical reviews*, *108*(1), 74-108.

Ballesteros JA, Weinstein H (1995) Integrated Methods for Modeling G-Protein Coupled Receptors. Methods Neurosci 25: 366-428.

Barrera, N. P., Zhou, M., & Robinson, C. V. (2013). The role of lipids in defining membrane protein interactions: insights from mass spectrometry. *Trends in cell biology*, *23*(1), 1-8.

Battle, A. R., Ridone, P., Bavi, N., Nakayama, Y., Nikolaev, Y. A., & Martinac, B. (2015). Lipid–protein interactions: Lessons learned from stress. *Biochimica et Biophysica Acta (BBA)-Biomembranes*, *1848*(9), 1744-1756.

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, et al. (2000) The Protein Data Bank. Nucleic Acids Res 28: 235-242.

Blaskovic, S., Blanc, M., & van der Goot, F. G. (2013). What does S-palmitoylation do to membrane proteins?. *The FEBS journal*, *280*(12), 2766-2774.

Bowie, J. U. (2011). Membrane protein folding: how important are hydrogen bonds?. *Current opinion in structural biology*, *21*(1), 42-49.

Bowie, J. U. (2005). Solving the membrane protein folding problem. *Nature*, *438*(7068), 581.

Brocchieri, L., & Karlin, S. (2005). Protein length in eukaryotic and prokaryotic proteomes. *Nucleic acids research*, *33*(10), 3390-3400.

Caetano-Anollés, G., Wang, M., Caetano-Anollés, D., & Mittenthal, J. E. (2009). The origin, evolution and structure of the protein world. *Biochemical Journal*, *417*(3), 621-637.

Calabrese, A. N., & Radford, S. E. (2018). Mass spectrometry-enabled structural biology of membrane proteins. *Methods*, *147*, 187-205.

Cao, R., Adhikari, B., Bhattacharya, D., Sun, M., Hou, J., & Cheng, J. (2017). QAcon: single model quality assessment using protein structural and contact information with machine learning techniques. *Bioinformatics*, *33*(4), 586-588.

Capra, J. A. & Singh, M. (2007). Predicting functionally important residues from sequence conservation. *Bioinformatics* 23, 1875–1882

Cherezov, V., Rosenbaum, D. M., Hanson, M. A., Rasmussen, S. G., Thian, F. S., Kobilka, T. S., Choi, H. J., Kuhn, P., Weis, W. I. Kobilka, B. K. & Stevens, R. C. (2007). High-resolution crystal structure of an engineered human β2-adrenergic G protein–coupled receptor. *science*, *318*(5854), 1258-1265.

Clément, L., Emeric, D., Laurent, M., David, L., Eivind, H., & Kristian, V. (2018). A data-supported history of bioinformatics tools. *arXiv preprint arXiv:1807.06808*.

Colwell, L. J. (2018). Statistical and machine learning approaches to predicting protein–ligand interactions. *Current opinion in structural biology*, *49*, 123-128.

Cooley, R. B., Arp, D. J., & Karplus, P. A. (2010). Evolutionary origin of a secondary structure: π-helices as cryptic but widespread insertional variations of α-helices that enhance protein functionality. *Journal of molecular biology*, *404*(2), 232-246.

Cordomi, A*., et al.* (2013) Sulfur-containing amino acids in 7TMRs: molecular gears for pharmacology and function, *Trends Pharmacol. Sci.*, 34, 320-331.

Dawson, N. L., Lewis, T. E., Das, S., Lees, J. G., Lee, D., Ashford, P., Orengo, C. A. & Sillitoe, I. (2016). CATH: an expanded resource to predict protein function through structure and sequence. *Nucleic acids research*, *45*(D1), D289-D295.

Dayhoff, M. O., & Ledley, R. S. (1962, December). Comprotein: a computer program to aid primary protein structure determination. In *Proceedings of the December 4-6, 1962, fall joint computer conference* (pp. 262-274). ACM.

Dayhoff, M. O. (1965). *Atlas of Protein Sequence and Structure, 1965*. National biomedical research Foundation.

Dayhoff, M. O., Schwartz, R. M., & Orcutt, B. C. (1978). 22 a model of evolutionary change in proteins. In *Atlas of protein sequence and structure* (Vol. 5, pp. 345-352). National Biomedical Research Foundation Silver Spring.

Deber, C. M., & Goto, N. K. (1996). Folding proteins into membranes. *Nature structural biology*, *3*(10), 815.

Devereux, J., Haeberli, P., & Smithies, O. (1984). A comprehensive set of sequence analysis programs for the VAX. *Nucleic acids research*, *12*(1Part1), 387-395.

Doolittle, R. F., & Blombäck, B. (1964). Amino-acid sequence investigations of fibrinopeptides from various mammals: evolutionary implications. *Nature*, *202*(4928), 147.

Dwivedi, H., Baidya, M., & Shukla, A. K. (2018). GPCR Signaling: The Interplay of Gαi and β-arrestin. *Current Biology*, *28*(7), R324-R327.

Elcock, A. H. & McCammon, J. A. (2001). Identification of protein oligomerization states by analysis of interface conservation. *Proc Natl Acad Sci USA* 98, 2990–2994

Engelman, D. M., Chen, Y., Chin, C. N., Curran, A. R., Dixon, A. M., Dupuy, A. D., Lee, A. S. Lehnert, U., Matthews E. E., Reshetnyak Y. K., Senes, A. & Popot, J. L. (2003). Membrane protein folding: beyond the two stage model. *FEBS letters*, *555*(1), 122-125.

Ernst, R., Ejsing, C. S., & Antonny, B. (2016). Homeoviscous adaptation and the regulation of membrane lipids. *Journal of molecular biology*, *428*(24), 4776-4791.

Fagerberg L, Jonasson K, von Heijne G, Uhlen M, Berglund L (2010) Prediction of the human membrane proteome. Proteomics 10: 1141-1149.

Fahy, E., Subramaniam, S., Murphy, R. C., Nishijima, M., Raetz, C. R., Shimizu, T., Spener, F., van Meer, G., Wakelam, M. J. O. & Dennis, E. A. (2009). Update of the LIPID MAPS comprehensive classification system for lipids. *Journal of lipid research*, *50*(Supplement), S9-S14.

Finn, R. D., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Mistry, J., Mitchell, A. L., Potter, S. C., Punta, M., Qureshi, M., Sangrador-Vegas, A., Salazar, G. A. Tate, J. & Bateman, A. (2015). The Pfam protein families database: towards a more sustainable future. *Nucleic acids research*, *44*(D1), D279-D285.

Fitch, W. M., & Margoliash, E. (1967). Construction of phylogenetic trees. *Science*, *155*(3760), 279-284.

Fitch, W. M. (1970). Distinguishing homologous from analogous proteins. *Systematic zoology*, *19*(2), 99-113.

Fredriksson R, Lagerstrom MC, Lundin LG, Schioth HB (2003) The G-protein-coupled receptors in the human genome form five main families. Phylogenetic analysis, paralogon groups, and fingerprints. Mol Pharmacol 63: 1256-1272.

Gauthier, J., Vincent, A. T., Charette, S. J., & Derome, N. (2018). A brief history of bioinformatics. *Brief Bioinform*, 1-16.

Goñi, F. M. (2014). The basic structure and dynamics of cell membranes: An update of the Singer–Nicolson model. *Biochimica et Biophysica Acta (BBA)-Biomembranes*, *1838*(6), 1467-1476.

Gorter, E., & Grendel, F. J. E. M. (1925). On bimolecular layers of lipoids on the chromocytes of the blood. *Journal of experimental medicine*, *41*(4), 439-443.

Haltia, T., & Freire, E. (1995). Forces and factors that contribute to the structural stability of membrane proteins. *Biochimica et Biophysica Acta (BBA)-Reviews on Biomembranes*, *1241*(2), 295-322.

Harayama, T., & Riezman, H. (2018). Understanding the diversity of membrane lipid composition. *Nature reviews Molecular cell biology*.

Hauser AS, Attwood MM, Rask-Andersen M, Schioth HB, Gloriam DE (2017) Trends in GPCR drug discovery: new agents, targets and indications. Nat Rev Drug Discov 16: 829-842.

Hedger, G., & Sansom, M. S. (2016). Lipid interaction sites on channels, transporters and receptors: recent insights from molecular dynamics simulations. *Biochimica et Biophysica Acta (BBA)-Biomembranes*, *1858*(10), 2390-2400.

Hedin, L. E., Illergård, K., & Elofsson, A. (2011). An introduction to membrane proteins. *Journal of proteome research*, *10*(8), 3324-3331.

Hildebrand, P. W., Preissner, R., & Frömmel, C. (2004). Structural features of transmembrane helices. *FEBS letters*, *559*(1-3), 145-151.

Hilger, D., Masureel, M., & Kobilka, B. K. (2018). Structure and dynamics of GPCR signaling complexes. *Nature structural & molecular biology*, *25*(1), 4.

Hogeweg, P. (2011). The roots of bioinformatics in theoretical biology. *PLoS computational biology*, *7*(3), e1002021.

Hollenstein, K., Kean, J., Bortolato, A., Cheng, R. K., Doré, A. S., Jazayeri, A., Cooke, R. M., Weir, M. & Marshall, F. H. (2013). Structure of class B GPCR corticotropin-releasing factor receptor 1. *Nature*, *499*(7459), 438.

Honigmann, A., & Pralle, A. (2016). Compartmentalization of the cell membrane. *Journal of molecular biology*, *428*(24), 4739-4748.

International Human Genome Sequencing Consortium. (2001). Initial sequencing and analysis of the human genome. *Nature*, *409*(6822), 860.

Isberg V, de Graaf C, Bortolato A, Cherezov V, Katritch V, et al. (2015) Generic GPCR residue numbers - aligning topology maps while minding the gaps. Trends in Pharmacological Sciences 36: 22-31.

Joh, N. H., Min, A., Faham, S., Whitelegge, J. P., Yang, D., Woods, V. L., & Bowie, J. U. (2008). Modest stabilization by most hydrogen-bonded side-chain interactions in membrane proteins. *Nature*, *453*(7199), 1266.

Jussupow, A., Di Luca, A., & Kaila, V. R. (2019). How cardiolipin modulates the dynamics of respiratory complex I. *Science advances*, *5*(3), eaav1850.

Kashyap, H., Ahmed, H. A., Hoque, N., Roy, S., & Bhattacharyya, D. K. (2015). Big data analytics in bioinformatics: architectures, techniques, tools and issues. *Network Modeling Analysis in Health Informatics and Bioinformatics*, *5*(1), 28.

Katritch V, Cherezov V, Stevens RC (2013) Structure-function of the G protein-coupled receptor superfamily. Annu Rev Pharmacol Toxicol 53: 531-556.

Kendrew, J. C., Bodo, G., Dintzis, H. M., Parrish, R. G., Wyckoff, H., & Phillips, D. C. (1958). A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. *Nature*, *181*(4610), 662-666.

Kendrew, J. C., Dickerson, R. E., Strandberg, B. E., Hart, R. G., Davies, D. R., Phillips, D. C., & Shore, V. C. (1960). Structure of myoglobin: A three-dimensional Fourier synthesis at 2 Å. resolution. *Nature*, *185*(4711), 422.

Killian, J. A. (1998). Hydrophobic mismatch between proteins and lipids in membranes. *Biochimica et Biophysica Acta (BBA)-Reviews on Biomembranes*, *1376*(3), 401-416.

Kolakowski LF, Jr. (1994) GCRDb: a G-protein-coupled receptor database. Receptors Channels 2: 1-7.

Kumar, K. K., Shalev-Benami, M., Robertson, M. J., Hu, H., Banister, S. D., Hollingsworth, S. A., Latorraca, N. R., Kato, H. E., Hilger, D., Maeda, S., Weis, W. I., Farrens, D. L., Dror, R. O., Malhotra, S. V., Kobilka, B., K. &Skiniotis, G. (2019). Structure of a Signaling Cannabinoid Receptor 1-G protein complex. *Cell*, *176*(3), 448-458.

Landolt-Marticorena, C., Williams, K. A., Deber, C. M., & Reithmeier, R. A. (1993). Non-random distribution of amino acids in the transmembrane segments of human type I single span membrane proteins.

Lau, F. W., & Bowie, J. U. (1997). A method for assessing the stability of a membrane protein. *Biochemistry*, *36*(19), 5884-5892.

Levy, E. D., Boeri Erba, E., Robinson, C. V. & Teichmann, S. A. (2008). Assembly reflects evolution of protein complexes. *Nature* 453, 1262–1265

Li, S. C., & Deber, C. M. (1992). Glycine and β-branched residues support and modulate peptide helicity in membrane environments. *FEBS letters*, *311*(3), 217-220.

Liao, M., Cao, E., Julius, D., & Cheng, Y. (2013). Structure of the TRPV1 ion channel determined by electron cryo-microscopy. *Nature*, *504*(7478), 107.

Lo, Y. C., Rensi, S. E., Torng, W., & Altman, R. B. (2018). Machine learning in chemoinformatics and drug discovery. *Drug discovery today*.

Lucio, M., Lima, J. L. F. C., & Reis, S. (2010). Drug-membrane interactions: significance for medicinal chemistry. *Current medicinal chemistry*, *17*(17), 1795-1809.

Margulis, L. (1967). On the origin of mitosing cells. *J Theor Bio*, *14*(3), 255-274.

Marsden, R. L., & Orengo, C. A. (2008). Target selection for structural genomics: an overview. In *Structural Proteomics* (pp. 3-25). Humana Press.

Martinac, B., & Vandenberg, J. (2015). 'Shooting gallery'for membrane proteins provides new insights into complexities of their function and structural dynamics. *The Journal of physiology*, *593*(2), 353-354.

Milo, R., & Phillips, R. (2015). *Cell biology by the numbers*. Garland Science.

Min, S., Lee, B., & Yoon, S. (2017). Deep learning in bioinformatics. *Briefings in bioinformatics*, *18*(5), 851-869.

Mulkidjanian, A. Y., & Galperin, M. Y. (2010). Evolutionary origins of membrane proteins. In *Structural Bioinformatics of Membrane Proteins* (pp. 1-28). Springer, Vienna.

Müller, D.J., Wu, N. and Palczewski, K. (2008) Vertebrate membrane proteins: structure, function, and insights from biophysical approaches, *Pharmacol. Rev.*, 60, 43-78.

Nicolson, G. L. (2014). The Fluid—Mosaic Model of Membrane Structure: Still relevant to understanding the structure, function and dynamics of biological membranes after more than 40 years. *Biochimica et Biophysica Acta (BBA)-Biomembranes*, *1838*(6), 1451-1466.

Nittinger, E., Flachsenberg, F., Bietz, S., Lange, G., Klein, R., & Rarey, M. (2018). Placement of Water Molecules in Protein Structures: From Large-Scale Evaluations to Single-Case Examples. *Journal of chemical information and modeling*, *58*(8), 1625-1637.

Olivella M, Gonzalez A, Pardo L, Deupi X (2013) Relation between sequence and structure in membrane proteins. Bioinformatics 29: 1589-1592.

Ouellette, R. J., & Rawn, J. D. (2015). *Principles of organic chemistry*. Academic Press.

Overington JP, Al-Lazikani B, Hopkins AL (2006) How many drug targets are there? Nat Rev Drug Discov 5: 993-996.

Page, R. C., Kim, S., & Cross, T. A. (2008). Transmembrane helix uniformity examined by spectral mapping of torsion angles. *Structure*, *16*(5), 787-797.

Palczewski, K., Kumasaka, T., Hori, T., Behnke, C. A., Motoshima, H., Fox, B. A., Le Trong, I., Okada, T., Stenkamp, R. E., Yamamoto, M. & Miyano M. (2000). Crystal structure of rhodopsin: AG protein-coupled receptor. *science*, *289*(5480), 739-745.

Pardo, L., Deupi, X., Dölker, N., López-Rodríguez, M. L., & Campillo, M. (2007). The role of internal water molecules in the structure and function of the rhodopsin family of G protein-coupled receptors. *ChemBioChem*, *8*(1), 19-24.

Park, P. S.-H. (2012). Ensemble of G protein-coupled receptor active states. *Current medicinal chemistry*, *19*(8), 1146-1154.

Pauling, L., & Corey, R. B. (1951). Configurations of polypeptide chains with favored orientations around single bonds: two new pleated sheets. *Proceedings of the National Academy of Sciences of the United States of America*, *37*(11), 729.

Perutz, M. F., Rossmann, M. G., Cullis, A. F., Muirhead, H., Will, G., & North, A. C. T. (1960). Structure of haemoglobin: a three-dimensional Fourier synthesis at 5.5-Å. resolution, obtained by X-ray analysis. *Nature*, *185*(4711), 416.

Petsko, G. A., & Ringe, D. (2004). *Protein structure and function*. New Science Press.

Pohorille, A., & Deamer, D. (2009). Self-assembly and function of primitive cell membranes. *Research in microbiology*, *160*(7), 449-456.

Popot, J. L., Gerchman, S. E., & Engelman, D. M. (1987). Refolding of bacteriorhodopsin in lipid bilayers: a thermodynamically controlled two-stage process. *Journal of molecular biology*, *198*(4), 655-676.

Popot, J. L., & Engelman, D. M. (1990). Membrane protein folding and oligomerization: the two-stage model. *Biochemistry*, *29*(17), 4031-4037.

Ramakrishnan, C., & Ramachandran, G. N. (1965). Stereochemical criteria for polypeptide and protein chain conformations: II. Allowed conformations for a pair of peptide units. *Biophysical journal*, *5*(6), 909-933.

Rawlings, A. E. (2018). Membrane protein engineering to the rescue. *Biochemical Society Transactions*, *46*(6), 1541-1549.

Reid, D. W., & Nicchitta, C. V. (2012). Primary role for endoplasmic reticulum-bound ribosomes in cellular translation identified by ribosome profiling. *Journal of Biological Chemistry*, *287*(8), 5518-5527.

Riley, M. L., Wallace, B. A., Flitsch, S. L., & Booth, P. J. (1997). Slow α helix formation during folding of a membrane protein. *Biochemistry*, *36*(1), 192-196.

Sanger, F., & Tuppy, H. (1951). The amino-acid sequence in the phenylalanyl chain of insulin. 1. The identification of lower peptides from partial hydrolysates. *Biochemical Journal*, *49*(4), 463.

Santos R, Ursu O, Gaulton A, Bento AP, Donadi RS, et al. (2017) A comprehensive map of molecular drug targets. Nat Rev Drug Discov 16: 19-34.

Serdiuk, T., Mari, S. A., & Müller, D. J. (2017). Pull-and-paste of single transmembrane proteins. *Nano letters*, *17*(7), 4478-4488.

Shimizu, T., Mitsuke, H., Noto, K., & Arai, M. (2004). Internal gene duplication in the evolution of prokaryotic transmembrane proteins. *Journal of molecular biology*, *339*(1), 1-15.

Shonberg, J., Kling, R. C., Gmeiner, P., & Löber, S. (2015). GPCR crystal structures: medicinal chemistry in the pocket. *Bioorganic & medicinal chemistry*, *23*(14), 3880-3906.

Siltberg-Liberles, J., Grahnen, J. A., & Liberles, D. A. (2011). The evolution of protein structures and structural ensembles under functional constraint. *Genes*, *2*(4), 748-762.

Simons, K. (2016). Cell membranes: A subjective perspective. *Biochimica et Biophysica Acta (BBA)-Biomembranes*, *1858*(10), 2569-2572.

Singer, S. J., & Nicolson, G. L. (1972). The fluid mosaic model of the structure of cell membranes. *Science*, *175*(4023), 720-731.

Spencer, M., Eickholt, J., & Cheng, J. (2015). A deep learning network approach to ab initio protein secondary structure prediction. *IEEE/ACM transactions on computational biology and bioinformatics (TCBB)*, *12*(1), 103-112.

Spyrakis, F., Ahmed, M. H., Bayden, A. S., Cozzini, P., Mozzarelli, A., & Kellogg, G. E. (2017). The roles of water in the protein matrix: a largely untapped resource for drug discovery. *Journal of medicinal chemistry*, *60*(16), 6781-6827.

Stangl, M., & Schneider, D. (2015). Functional competition within a membrane: Lipid recognition vs. transmembrane helix oligomerization. *Biochimica et Biophysica Acta (BBA)-Biomembranes*, *1848*(9), 1886-1896.

Stephenson, N., Shane, E., Chase, J., Rowland, J., Ries, D., Justice, N., Zhang, J., Chan, L. & Cao, R. (2018). Survey of machine learning techniques in drug discovery. *Current drug metabolism*.

Stevens, R. C., Cherezov, V., Katritch, V., Abagyan, R., Kuhn, P., Rosen, H., & Wüthrich, K. (2013). The GPCR Network: a large-scale collaboration to determine human GPCR structure and function. *Nature reviews Drug discovery*, *12*(1), 25.

Sun, X., Ågren, H., & Tu, Y. (2014). Functional water molecules in rhodopsin activation. *The journal of physical chemistry B*, *118*(37), 10863-10873.

Tanford, C. (1978). The hydrophobic effect and the organization of living matter. *Science*, *200*(4345), 1012-1018.Technical Brief 2009 Volume 8 Particle Sciences Inc

Tehan, B. G., Bortolato, A., Blaney, F. E., Weir, M. P., & Mason, J. S. (2014). Unifying family A GPCR theories of activation. *Pharmacology & therapeutics*, *143*(1), 51-60.

The UniProt Consortium. (2018). UniProt: a worldwide hub of protein knowledge. *Nucleic acids research*, *47*(D1), D506-D515.

Trzaskowski B, Latek D, Yuan S, Ghoshdastider U, Debinski A, et al. (2012) Action of Molecular Switches in GPCRs - Theoretical and Experimental Studies. Current Medicinal Chemistry 19: 1090-1109.

Uhlén, M., Fagerberg, L., Hallström, B. M., Lindskog, C., Oksvold, P., Mardinoglu, A., Sivertsson, A., Kampf, C., Sjöstedt, E., Asplund, A., Olsson, I., Edlund, E., Lunberg, E., Navani, S., Al-Khalili Szigyarto, C., Odeberg, J., Djureinovic, D., Takanen, J. O., Hober, S., Alm, T., Edqvist, P. H., Berling, H., Tegel, H., Mulder, J., Rockberg, J., Nilsson, P., Schwenk, J. M., Hamsten, M., von Feilitzen, K., Forsberg, M., Persson, L., Johansson, F., Zwahlen, M., von Heijne, G., Nielsen, J. & Pontén, F. (2015). Tissue-based map of the human proteome. *Science*, *347*(6220), 1260419.

Ulmschneider, M.B. and Sansom, M.S. (2001) Amino acid distributions in integral membrane protein structures, *Biochim. Biophys. Acta*, 1512, 1-14.

Vamathevan, J., Clark, D., Czodrowski, P., Dunham, I., Ferran, E., Lee, G., Li, B., Madabhushi, A., Shah, P., Spitzer, M. & Zhao, S. (2019). Applications of machine learning in drug discovery and development. *Nature Reviews Drug Discovery*, 1.

van der Westhuizen, E. T., Valant, C., Sexton, P. M., & Christopoulos, A. (2015). Endogenous allosteric modulators of G protein–coupled receptors. *Journal of Pharmacology and Experimental Therapeutics*, *353*(2), 246-260.

van Meer, G., Voelker, D. R., & Feigenson, G. W. (2008). Membrane lipids: where they are and how they behave. *Nature reviews Molecular cell biology*, *9*(2), 112.

Venkatakrishnan, A. J., Ma, A. K., Fonseca, R., Latorraca, N. R., Kelly, B., Betz, R. M., Asawa, C., Kobilka, B. K. & Dror, R. O. (2019). Diverse GPCRs exhibit conserved water networks for stabilization and activation. *Proceedings of the National Academy of Sciences*, *116*(8), 3288-3293.

von Heijne, G. (1989). Control of topology and mode of assembly of a polytopic membrane protein by positively charged residues. *Nature*, *341*(6241), 456.

von Heijne, G. (1992) Membrane protein structure prediction. Hydrophobicity analysis and the positive-inside rule, *J. Mol. Biol.*, 225, 487-494

Wang, C., Wu, H., Katritch, V., Han, G. W., Huang, X. P., Liu, W., Siu, F. Y., Roth, B. L., Cherezov, V. & Stevens, R. C. (2013). Structure of the human smoothened receptor bound to an antitumour agent. *Nature*, *497*(7449), 338.

Wang, S., Peng, J., Ma, J., & Xu, J. (2016). Protein secondary structure prediction using deep convolutional neural fields. *Scientific reports*, *6*, 18962.

White, S. H., & Wimley, W. C. (1999). Membrane protein folding and stability: physical principles. *Annual review of biophysics and biomolecular structure*, *28*(1), 319-365.

White, S. H., Ladokhin, A. S., Jayasinghe, S., & Hristova, K. (2001). How membranes shape protein structure. *Journal of Biological Chemistry*, *276*(35), 32395-32398.

Wu, H., Wang, C., Gregory, K. J., Han, G. W., Cho, H. P., Xia, Y., NIswender, C. M., Katritch, V., Meiler, J., Cherezov, V., Conn, P. J., Stevens R. C. (2014). Structure of a class C GPCR metabotropic glutamate receptor 1 bound to an allosteric modulator. *Science*, *344*(6179), 58-64.

Yang, Y., Heffernan, R., Paliwal, K., Lyons, J., Dehzangi, A., Sharma, A., Wang, J., Sattar, A. & Zhou, Y. (2017). Spider2: A package to predict secondary structure, accessible surface area, and main-chain torsional angles by deep neural networks. In *Prediction of protein secondary structure* (pp. 55-63). Humana Press, New York, NY.

Yeagle, P. L. (2016). *The membranes of cells*. Academic Press.

Yin, H., & Flynn, A. D. (2016). Drugging membrane protein interactions. *Annual review of biomedical engineering*, *18*, 51-76.

Yuan, S., Vogel, H., & Filipek, S. (2013). The Role of Water and Sodium Ions in the Activation of the µ-Opioid Receptor. *Angewandte Chemie International Edition*, *52*(38), 10112-10115.

Yuan, S., Filipek, S., Palczewski, K., & Vogel, H. (2014). Activation of G-protein-coupled receptors correlates with the formation of a continuous internal water pathway. *Nature communications*, *5*, 4733.

Zhong, D., Pal, S. K., & Zewail, A. H. (2011). Biological water: A critique. *Chemical Physics Letters*, *503*(1-3), 1-11.

Zhou, M., & Robinson, C. V. (2014). Flexible membrane proteins: functional dynamics captured by mass spectrometry. *Current opinion in structural biology*, *28*, 122-130.

2. Methods

2. METHODS

The work described in this thesis has been performed using structural bioinformatics methods and public data from biological databases with experimental information. The following sections describe the methods employed, among others: sequence and structural alignments, visualization of 3D structures, statistical analysis of protein features using Python scripts and development of tools in a web server format. Besides, the specific details for methods used in each project will be described within each study.

2.1. Protein Databases used for the analysis of Membrane Proteins

What follows is a (far from complete) list of databases related to proteins, membrane proteins and their sequences and structures that have been relevant for this thesis. Some of them contain structural data, others protein and genomic sequences or information about protein families and their mutations. They can be divided them in three categories: i) generalists, ii) dedicated to membrane proteins and iii) dedicated to GPCRs. A recent comparative study of some of the databases of membrane protein structures was carried out by Shimizu and colleagues (Shimizu *et al.*, 2018). They compare and analyse the number of structures and evaluate the relationships between the databases and the overlapping in the datasets.

In the category of generalistic databases we can find:

Protein Data Bank (PDB): https://www.rcsb.org
A database for three-dimensional structural data of biological molecules, such as proteins, nucleic acids and complexes of both. Data is obtained mainly by X-ray crystallography, NMR spectroscopy and more recently by cryo-electron microscopy. Since the first structure was released in 1976 (PDBID 1MBN, Watson, 1969), the number of available structures has not stopped growing, the current number of deposited structures in PDB is almost 150.000. The ratio of releases per year has also increased substantially in the last years to achieve the outstanding number of more than 10.000 structures released in 2018.

Protein families (Pfam): https://pfam.xfam.org
A project that gather all protein families, represented by multiple sequence alignments and hidden Markov models. Pfam narrows the number of transmembrane protein families to 578 in their last update from a total of 17.929 families. (version 32).

Universal Protein Resource (UniProt): https://www.uniprot.org
A comprehensive resource for protein sequence and annotation data for over 120 million proteins across all branches of life. Moreover, there are a considerable number of Reference Proteomes. It contains a large amount of information about the biological function of proteins derived from the research literature.

The following databases related to Membrane Protein have been used:

Orientation of Proteins in Membranes (OPM): https://opm.phar.umich.edu
Provides spatial arrangement of MPs with respect to the lipid bilayer. OPM includes all unique experimental structures of transmembrane proteins, some peripheral proteins and membrane-bound peptides from PDB with their calculated membrane boundaries. Entries are also divided in different classifications and assemblies

Protein Data Bank of Transmembrane Proteins (PDBTM): http://pdbtm.enzim.hu
This database contains MPs from PDB, its interest resides on the use of TMDET algorithm to locate the transmembrane region of each protein. This information can be easily downloaded in xml format.

Membrane Proteins of Known Structure (mpstruc): http://blanco.biomol.uci.edu/mpstruc
It contains all known MPs solved so far, updated weekly. Classified as monotopic, α-helical and β-barrels. By April 14th of 2019 it contained the coordinates of >2700 structures representing 884 unique proteins.

Finally, a database focused on GPCRs:

GPRCDB: https://gpcrdb.org
Contains reference data of GPCRs including sequence alignments for all species in UniProt, a large collection of crystal structures, 3D models in the active or inactive state, and receptor mutants. Moreover, recently has incorporated many interactive web browser tools and diagrams such as phylogenetic trees, sequence motif search and receptor sequence snake plots.

2.2. Sequence Alignments for the analysis Membrane Proteins

A milestone in the field of structural bioinformatics was achieved when Emile Zuckerkandl and Linus Pauling hypothesized that orthologues evolved from divergence from a common ancestor (Zuckerkandl & Pauling, 1965). By comparing the sequence of hemoglobin of different species, they could predict the ‘ancestral sequence’ of this protein and follow its evolutionary history up to current forms. At that time, however, automatic comparison and alignment of protein sequences was yet not available. An initial solution was provided by Needleman and Wunch, who developed the first dynamic programming algorithm for pairwise protein sequence alignments (Needleman & Wunsch, 1970). However, it was not until the early 1980s that the first multiple sequence alignment (MSA) algorithm, a generalization of the Needleman-Wunch algorithm was developed appeared, incorporating a scoring matrix whose dimensionality was the same as the length of the sequences (Murata *et al.*, 1985). This method was highly computational demanding, hence, not very useful for large sequences. In 1987 emerged the first truly practical MSA software, developed by Feng and Doolittle, called ‘progressive sequence alignment’ (Feng & Doolittle, 1987). The next year, based in a simplification of Feng-Doolittle algorithm, CLUSTAL software was developed (Higgins & Sharp, 1988). This software is still maintained and commonly used nowadays (Sievers & Higgins, 2014).

2.2.1. Sequence Alignment (Sequence-based)

A sequence alignment is a way of arranging DNA, RNA, or protein sequences to identify regions of similarity that may be a consequence of functional, structural, or evolutionary relationships between the sequences. Aligned sequences are typically represented as rows within a matrix with gaps inserted between residues so that identical or similar characters are aligned in successive columns. Methods for multiple-sequence alignment have become almost everyday tools for computational biologists (e.g., Psi-BLAST or ClustalW).

Although short sequences can be aligned by hand, this is unfeasible for common problems, which require using alignment algorithms. These are of two main types:

- **Global alignments**: attempt to align every residue in every sequence, are most useful when the sequences in the query set are similar and of roughly equal size. A general global alignment technique is the Needleman–Wunsch algorithm (Needleman & Wunsch, 1970), which is based on dynamic programming.

- **Local alignments**: are more useful for dissimilar sequences that are suspected to contain regions of similarity or similar sequence motifs within their larger sequence context. The Smith–Waterman algorithm (Smith & Waterman, 1981) is a general local alignment method based on the same dynamic programming scheme but with additional choices to start and end at any place.

The difference between a local and a global alignment is that in a local alignment, the query is intended to match with a portion of the target sequence, whereas in a global alignment, an end to end alignment is performed which may produce many gaps if the length of both sequences is dissimilar. Global alignments are usually done for comparing homologous sequences whereas local alignment can be used to find homologous domains in otherwise non-homologous sequences.

Thus, sequence alignment is the procedure of comparing two (pairwise alignment) or more (multiple sequence alignment) sequences by searching for a series of individual characters or character patterns that are in the same order in the sequences.

Pairwise sequence alignment methods are used to find the best-matching piecewise (local or global) alignments of two query sequences. Pairwise alignments are efficient to calculate and are often used for methods that do not require extreme precision (such as searching a database for sequences with high similarity to a query). The three primary methods of producing pairwise alignments are dot-matrix methods, dynamic programming, and word methods. Although each method has its individual strengths and weaknesses, all three pairwise methods have difficulty with highly repetitive sequences of low information content - especially where the number of repetitions differ in the two sequences to be aligned.

Multiple sequence alignment is an extension of pairwise alignment to incorporate more than two sequences at a time. Multiple alignment methods try to align all the sequences in a given query set. Multiple alignments are often used in identifying conserved sequence regions across a group of sequences hypothesized to be evolutionarily related. Such conserved sequence motifs can be used in conjunction with structural and mechanistic information to locate the catalytic active sites of enzymes. Alignments are also used to aid in establishing evolutionary relationships by constructing phylogenetic trees.

In order to evaluate sequence alignments **substitution matrices** are commonly used. Substitution matrices are two-dimensional matrices with score values describing the probability of one amino acid or nucleotide being replaced by another during sequence evolution. For proteins, PAM (Dayhoff *et al.*, 1978) and BLOSUM (Henikoff & Henikoff, 1992) methods are the most used to estimate the target frequencies and thus the log-odds scores of a substitution matrix in sequence alignments. These matrices contain positive and negative values, reflecting the likelihood of each amino acid substitution in related proteins. Using these tables, an alignment of a sequential set of amino acid pairs with no gaps receives an overall score that is the sum of the positive and negative log-odds scores for each individual amino acid pair in the alignment. The higher this score, the more significant is the alignment, or the more it resembles alignments in related proteins. The score given for gaps in aligned sequences is negative, because such misaligned regions should be uncommon in sequences of related proteins (Mount, 2001).

2.2.2. Structural Alignment

Structural alignment refers to the alignment, in three dimensions, between two or more molecular models. In the case of proteins, this is usually performed without reference to the sequences of the proteins. When the models align well, it suggests evolutionary and functional relationships that may not be discernible from sequence comparisons. The structural differences between two optimally aligned models are usually measured as the Root Mean Square Deviation (RMSD) between the aligned α-carbon positions.

$$RMSD = \sqrt{\frac{1}{N}\sum_{i=1}^{N} d_i^2}$$

where d_i is the distance between atom *i* and either a reference structure or the mean position of the *N* equivalent atoms.

Structural alignment of proteins can in principle provide more reliable alignments than sequence based (Shatsky *et al.*, 2004). Many online resources and tools are available for high-quality pairwise alignments. However, much more information can be derived from multiple-structure alignment. Recognition of a structural core common to a set of protein structures has many applications such as in the studies of protein evolution and classification, analysis of similar functional binding sites, and homology modeling. (Akutsu & Sim, 1999, Goldsmith-Fischman & Honig, 2003).

In this thesis, structural pairwise alignment is the basis of Homolwat (see chapter 4.7). We have used the following methods from PyMol software:

- **align**: performs a sequence alignment followed by a structural superposition, and then carries out zero or more cycles of refinement in order to reject structural outliers found during the fit. Suited for proteins with decent sequence similarity (identity >30%).
- **super**: does a **sequence-independent** structure-based dynamic programming alignment followed by a series of refinement cycles intended to improve the fit by eliminating pairing with high relative variability. **Super is more robust than align** for proteins with low sequence similarity.

2.3. Protein Visualization

The easiest way to represent a protein structure is to build a model of the 3D coordinates of its atoms. Visualization programs allow user to view and rotate these models and change the model structure to find the desirable information from it. Common way to visualize proteins are building a wire framework of the amino acid chain and to present alpha helices and beta sheets with ribbons to make them stand out. Many different programs exist, the most widely used are PyMol, Chimera and VMD, between them, PyMol is the one that has been more extensively used along this thesis.

PyMol (Version 2.0.5 Schrödinger, LLC) is a molecular graphics system with an embedded Python interpreter designed for real-time visualization and rapid generation of high-quality molecular graphics images and animations.

Chimera (Pettersen *et al.*, 2004) is a highly extensible program for interactive visualization and analysis of molecular structures and related data, including density maps, supramolecular assemblies, sequence alignments, docking results, trajectories, and conformational ensembles. High-quality images and animations can be generated.

Visual Molecular Dynamics, VMD (Humphrey *et al.*, 1996) is a molecular visualization program for displaying, animating, and analysing large biomolecular systems using 3-D graphics and built-in scripting. VMD is designed for modeling, visualization, and analysis of biological systems such as proteins, nucleic acids, lipid bilayer assemblies, etc.

2.4. Analysis of Sequence Conservation

When the sequences of a given protein are compared between similar species, using MSA, differences between sequences most often represent mutations that were allowed (by evolution) to persist because they were harmless. Where the sequences are identical, it is said that sequence is **conserved** (Martz & Hodis, 2013). If two or more protein sequences are highly similar it is likely that they have similar structure and function.

Many different methods have been used to quantify conservation in sequences, some of them are available in GPCRSAS (section 4.5) and the interactions analysis (section 4.3), they are explained below.

2.4.1. Conservation analysis

Given a multiple sequence alignment, the most trivial way of analyzing the conservation of a certain position (or range of positions) in a family is to monitor **residue frequencies**.

$$f_i = \frac{n_i}{N}$$

where f_i is the frequency of the residue *i* in a sequence of length *N*, and n_i the number of times this residue appears in the sequence.

Another interesting metric for a given position in the alignment is the **entropy**, which uses information theory. **Information entropy** is the average rate at which information is produced by a stochastic source of data. In the case of a protein multiple sequence alignment, for each position or range of positions *i*, the entropy of the information contained *H(i)* is defined according to Shannon's theorem (Shannon, 1948) as:

$$H(i) = -\sum_x p_x(i) \log_b p_x(i)$$

where $p_x(i)$ is the probability mass function for the amino acid(s) at position (or group of positions) *i*. The logarithm base *b* serves to scale the entropy in the range [0, 1] for one or more positions. Consequently, *b* is *20n*, with *n* being the number of positions used for the calculation. A position or group of positions with low variability (high conservation) has an entropy *H(i)* close to 0, while high variability (low conservation) gives an entropy close to 1.

2.4.2. Logo: a representation of Sequence Conservation

A sequence logo is a graphical representation of the conservation of nucleotides or amino acids in a sequence. A sequence logo is created from a collection of aligned sequences and depicts the consensus sequence and diversity of the sequences. The consensus logo depicts the degree of conservation of each position using the height of the consensus character at that position. However, it can be used to display frequencies of amino acids at specific positions as has been used in the inter-residue interactions study (section 4.3)

2.4.3. Covariance and correlation between two positions

Although conserved regions among orthologues and paralogues are considered core structures required for basic biological function (Yang & Honig, 2000), co-variant amino acids have attracted attention recently for their roles in differentiation and phenotype, which reflect a subtle correlation between the polymorphism and stability of proteins (Yeang & Haussler, 2007, Atchley *et al.*, 2000). Several algorithmic approaches have been used to explore the coevolution of protein residues, such as explicit likelihood (Dekker *et al.*, 2004), maximum likelihood (Pollock *et al.*, 1999), joint probability estimation (Jeong & Kim, 2012), and correlation coefficient tests (Fares & Travers, 2006).

To analyse the covariance of two positions, GPCR-SAS (section 4.5) uses the Observed Minus Expected Squared (OMES) (Fodor & Aldrich, 2004), that is based on a χ^2 test, and a corrected mutual information method (MIp) (Dunn *et al.*, 2008). Both methods have previously been employed by Pele and collaborators to identify evolutionary hubs between pairs of residues in GPCRs (Pele *et al.*, 2014).
OMES calculates the difference between the observed and expected frequencies of each possible pair of amino acids or groups of amino acids *(x, y)*, at positions *i* and *j* of the alignment:

$$OMES(i,j) = \frac{1}{N(i,j)} \sum_{x,y} (N_{x,y}^{obs}(i,j) - N_{x,y}^{\exp}(i,j))^2$$

being the number of sequences in the alignment with non-gapped residues, the observed frequency and the expected frequency, respectively, at positions (or list of positions) *i* and *j*.
The Mutual Information content *MI(i,j)* between two positions (or lists of positions) *i* and *j* on an alignment is based on the probability of joint occurrence of events and is defined as:

$$MI(i,j) = \sum_{x,y} p_{x,y}(i,j) \ln \frac{p_{x,y}(i,j)}{p_x(i)\, p_y(j)}$$

where *px(i)*, *py(j)* and *px,y(i,j)* are respectively the frequencies of amino acid *x* at position *i*, amino acid *y* at position *j* and the amino acid pair *(x, y)* at positions *i*and *j*.
The corrected MIp version is defined as:

$$MIp(i,j) = MI(i,j) - \frac{\frac{1}{n-1}\sum_{j \neq i} MI(i,j) \frac{1}{n-1} \sum_{i \neq j} MI(i,j)}{\frac{2}{n(n-1)} \sum_{i,j} MI(i,j)}$$

with *n* being the number of columns in the alignment.
To evaluate the statistical significance for the computed OMES and MIp values, GPCR-SAS provides the Z-scores and the associated p-values, which are computed by comparing with the mean value for all combinations of two positions.
To determine the correlation between two sequence positions, the occurrence of the amino acid or motif at the first position or range of positions is associated with the occurrence of the amino acid or motif at the second position or range of positions. The occurrences are used to compute an odds ratio (and the associated 95% confidence interval) that estimates how strongly the presence/absence of one of the first amino acids or motif is correlated with the presence/absence of the second amino acid or motif. To facilitate the comparison with the other categories at the chosen level of classification and in the subcategories, the output of a correlation analysis also returns the same analysis for these groups.

2.5. Homology Modelling

One of the main drawbacks of the study of proteins, and particularly MP, is the lack of experimental structural information, hence there is a need of constructing 3D atomic models of not resolved proteins. Normally these models are based on a known structure of a similar protein. Typically, the most accurate models of protein structure are achieved through Homology Modelling.
Homology Modelling is a structure prediction method that is adopted to construct a three-dimensional model of a target protein from a template structure of a related protein with known structure, and a sequence alignment between both proteins (target and template). The method relies on the fact that during evolution structure is more conserved than amino acid sequence. The quality of a model is linked with the sequence identity (SI, number of amino acids that match at a given position of the alignment) between template and target sequences (Chothia & Lesk, 1986; Rost, 1999). Usually the worst parts of

a model are the loop regions because they tend to be less conserved. This idea serves as a hypothesis for the placement of internal water in GPCR models (section 4.6), where many residues and interactions are conserved, we hypothesized that water molecules can also be conserved and thus, can be placed as homologous waters.

The past years have shown a steady increase in the use of homology modelling techniques, especially for MPs. Contributing factors are the increase of determined MP structures and therefore the availability of suitable templates (Koehler Leman *et al.*, 2015). Homology modelling requires a template structure with high sequence similarity to the target sequence: typically, the higher the sequence similarity, the higher the accuracy of the resulting model. Sequence similarities of ~70% can yield models with an RMSD of 1-2 Å whereas highest-quality models with sequence similarities of 25% to the template typically have RMSDs of 3-4 Å (Baker & Sali, 2001), consequently, the quality of the sequence alignment also influences the accuracy of the resulting model. After a high-quality alignment has been obtained, the alignment and the template can be submitted to a homology modelling tool. For some tools, the alignment step is already included in the calculation and the target sequence and template structure are sufficient for modelling. Several tools are available, the most commonly used are RosettaMembrane (Yarov-Yarovoy *et al.*, 2006, Barth *et al.*, 2007) and MODELLER (Eswar *et al.*, 2003) and designed for MPs, MEDELLER (Kelm *et al.*, 2010).
However, as the structure of membrane proteins is more conserved than the structure of globular proteins (Olivella *et al.*, 2013); i.e. less sequence identity is required to maintain the overall fold, and thus, the use of structure-based methods can achieve better results.

2.6. Analysis of Membrane Proteins Structures

2.6.1. Location of transmembrane regions in Membrane Proteins

One of the main problems of MP crystal structures is that the position of the lipid bilayer cannot be directly determined. Even when there is proof for the existence of transmembrane domains, it is difficult to determine their **boundaries.** In this context, several methods have been developed these last years, such as Phobius, TMPred or TMHMM. Information regarding membrane location for TMalphaDB and TMbetaDB was obtained from PDBTM and OPM (see section 2.1). For a complete review of experimental measurements and techniques to determine the orientation and localizations of proteins bound to the membrane check the work of Hohlweg and colleagues (Hohlweg *et al.*, 2012)

2.6.2. Structural parameters to characterize residues in Membrane Proteins

In order to characterize TM segments, many conformational parameters can be taken into account. TMalphaDB and TMbetaDB (section 4.2) allow to calculate the dihedral angles Φ and Ψ and the first rotamer χ_1 of each residue of a selected TM segment, download and plot this information. Likewise, for unit Twist and Bend angles, two relevant parameters to measure local distortions of TM helices. The unit twist angles are calculated, for each set of four contiguous Cα atoms along the helix, to analyse helical uniformity. An ideal α-helix, with approximately 3.6 residues per turn, has a unit twist of approximately 100° (360°/3.6). A closed helical segment, with <3.6 residues per turn, possesses a unit twist >100°, whereas an open helical segment, with >3.6 residues per turn, possesses a unit twist <100°. A variation greater than 20° in the unit twist angle will result in a change in the orientation of the amino acid side chain. Local bend angles are calculated as the angle between the axis of the cylinders formed by the Cα atoms of the residues preceding (i-3, i) and following (i, i + 3) a given amino acid i. Unit Twist and Bend angles are calculated employing the program HELANAL (Bansal et al., 2000).

2.6.3. Insideness of Residues or water molecules

In order to determine if residues or water molecules are buried to the interior of the protein or exposed to the surface, two different approaches have been used. For the study of inter-residue interactions (section 4.3) **Circular Variance** (CV) (Mezei, 2003) has been used for a-helical proteins. CV provides a quantitative measure of “insideness” of residues or atoms in proteins (it ranges between 0 and 1, where 0 means completely exposed and 1 completely buried). Residues with CV ≤ 0.7 are classified as OUT and those with CV > 0.7 are classified as IN. This method has been also implemented in HomolWat algorithm (section 4.6) for the classification of water molecules in proteins, following the same criteria for IN and OUT. On the other hand, for β-barrels another method has been used. The **direction of the vector connecting Cα and Cβ** allows to classify residues as IN (vector pointing towards the center of the barrel) or OUT (pointing in the opposite direction).

2.6.4. Characterization of interactions

Residue-residue Interactions

The quantification of inter-residue interactions is the basis of the studies of interactions in MPs and sulfur-containing residues (sections 4.3 and 4.4). The criteria used there is the following: two residues were considered to interact if the distance between any two heavy atoms (including both side-chain and backbone) is ≤ 4.5 Å. This distance cutoff was chosen in accordance with a previous analysis (Yuan *et al.*, 2012). The analysis of such interactions in different data sets is detailed in section 4.3. Specific interactions of sulfur-containing amino acids with aliphatic residues is reported in section 4.4.

Quantum Mechanical Calculations for inter-residue interactions

The interaction energy between residues can be calculated using *ab initio* methods. These are methods of computational chemistry, that attempt to solve the electronic Schrödinger equation given the position of the nuclei and the number of electrons in order to yield useful information such as electron densities, energies and other properties of the system. The ability to obtain a ‘good enough’ solution to this equation for systems containing tens or hundreds of atoms has revolutionized the ability of theoretical chemistry to address many problems in a wide range of disciplines (Friesner & Guallar, 2005). The development of this methods deserved a Nobel Prize to John Pople and Walter Kohn in 1998.

We have used ab initio methods to calculate interaction energies on small-molecule model systems mimicking sidechain- sidechain interactions for interactions involving Met and Cys (see section 4.4). Residues were mimicked by dimethyl sulfide (DMS) for Met, bymethanethiol (MT) for Cys, by propane (PRP) for Leu, and Phe by benzene (BNZ). All chosen model structures were optimized at the MP2/6-31+G(d,p) level of theory, which has been shown to provide reasonably good geometries (Riley *et al.*, 2012, Hobza *et al.*, 1996). Single point energy calculations for interacting pairs were performed at the CCSD(T)/6-311+G(3df,2p) level. In order to minimize the basis set superposition error, counterpoise method by Boys and Bernardi (Boys & Bernardi, 1970) was utilized. These calculations were performed using GAUSSIAN09 program (Gaussian 09, Frisch *et al.*, 2009)

2.7. Programming as a tool to automatise Bioinformatics pipelines

The main tool used in this thesis has been programming. In this thesis I have developed Python, Bash and (to a lower extend) R scripts for data collection, analysis and visualization.

The Python scripts heavily relied on packages:
- NumPy (Oliphant, 2006): mathematical and numerical calculations
- SciPy (Jones *et al.*, 2001): math, science and engineering.

- Pandas (McKinney, 2010): data manipulation and analysis
These three packages have been used to structure the data and perform measurement and statistical analyisis.
- Biopython (Chapman & Chang, 2000), its goal is to make as easy as possible the use of Python for bioinformatics by creating high-quality, reusable modules and classes. Biopython is able to parse many different formats such as FASTA, Clustalw or Genbank, and have lots of functionalities, for instance perform common sequence operations (translation, transcription, etc.), deal with alignments, or perform classification of data using k Nearest Neighbors, Naïve Bayes or Support Vector Machines, and many other functionalities.
- Matplotlib (Hunter, 2007): for plotting, in Python scripts, the Python shell, Jupyter notebooks, web application servers, and other graphical user interface toolkits.
- PyMol has been used as a Python package in some scripts to automatize structural superpositions and extract structural features such as angles.

2.8. Web Applications

It is common that the outcome of bioinformatics works are web applications, that is, computer programs that run in a web browser. This thesis includes work related to three different web applications:

- TMalphaDB and TMbetaDB: use a php backend
- GPCRSAS: use the Django (https://www.djangoproject.com) Python web framework. Django's primary goal is to ease the creation of complex, database-driven websites. On the other hand, for HomolWat,has been used.
- Homolwat: use the Flask (http://flask.pocoo.org) web framework. Flask is a lightweight WSGI web application framework, designed to make web applications straightforward in Python, based on Werkzeug and Jinja.

2.9. References

Akutsu, T., & Sim, K. L. (1999). Protein threading based on multiple protein structure alignment. *Genome Informatics*, *10*, 23-29.

Atchley, W. R., Wollenberg, K. R., Fitch, W. M., Terhalle, W., & Dress, A. W. (2000). Correlations among amino acid sites in bHLH protein domains: an information theoretic analysis. *Molecular biology and evolution*, *17*(1), 164-178.

Baker, D., & Sali, A. (2001). Protein structure prediction and structural genomics. *Science*, *294*(5540), 93-96.

Bansal, M., Kumart, S., & Velavan, R. (2000). HELANAL: a program to characterize helix geometry in proteins. *Journal of Biomolecular Structure and Dynamics*, *17*(5), 811-819.

Barth, P., Schonbrun, J., & Baker, D. (2007). Toward high-resolution prediction and design of transmembrane helical protein structures. *Proceedings of the National Academy of Sciences*, *104*(40), 15682-15687.

Boys, S. F.; Bernardi, F. The calculation of small molecular interactions by the differences of separate total energies. Some procedures with reduced errors. Mol Phys 1970, 19(4), 553-566.

Chapman, B., & Chang, J. (2000). Biopython: Python tools for computational biology. *ACM Sigbio Newsletter*, *20*(2), 15-19.

Chothia, C., & Lesk, A. M. (1986). The relation between the divergence of sequence and structure in proteins. *The EMBO journal*, *5*(4), 823-826.

Dayhoff, M. O., Schwartz, R. M., & Orcutt, B. C. (1978). 22 a model of evolutionary change in proteins. In *Atlas of protein sequence and structure* (Vol. 5, pp. 345-352). National Biomedical Research Foundation Silver Spring.

Dekker, J. P., Fodor, A., Aldrich, R. W., & Yellen, G. (2004). A perturbation-based method for calculating explicit likelihood of evolutionary co-variance in multiple sequence alignments. *Bioinformatics*, *20*(10), 1565-1572.

Dunn S. D., Wahl L. M., Gloor G. B. (2008) Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. Bioinformatics 24: 333-340.

Eswar, N., Webb, B., Marti-Renom, M. A., Madhusudhan, M. S., Eramian, D., Shen, M. Y., Pieper, U. & Sali, A. (2007). Comparative protein structure modeling using MODELLER. *Current protocols in protein science*, *50*(1), 2-9.

Fares, M. A., & Travers, S. A. (2006). A novel method for detecting intramolecular coevolution: adding a further dimension to selective constraints analyses. *Genetics*, *173*(1), 9-23.

Feng, D. F., & Doolittle, R. F. (1987). Progressive sequence alignment as a prerequisitetto correct phylogenetic trees. *Journal of molecular evolution*, *25*(4), 351-360.

Fodor A. A., Aldrich R. W. (2004) Influence of conservation on calculations of amino acid covariance in multiple sequence alignments. Proteins 56: 211-221.

Friesner, R. A., & Guallar, V. (2005). Ab initio quantum chemical and mixed quantum mechanics/molecular mechanics (QM/MM) methods for studying enzymatic catalysis. *Annu. Rev. Phys. Chem.*, *56*, 389-427.

Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, J. A.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, J. M.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J.: Wallingford CT, 2009.

Goldsmith-Fischman, S., & Honig, B. (2003). Structural genomics: computational methods for structure analysis. *Protein Science*, *12*(9), 1813-1821.

Henikoff, S., & Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences*, *89*(22), 10915-10919.

Higgins, D. G., & Sharp, P. M. (1988). CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. *Gene*, *73*(1), 237-244.

Hobza, P.; Selzle, H. L.; Schlag, E. W. Potential energy surface for the benzene dimer. Results of ab initio CCSD(T) calculations show two nearly isoenergetic structures: T-shaped and parallel-displaced. J Phys Chem 1996, 100(48), 18790-18794.

Hohlweg, W., Kosol, S., & Zangger, K. (2012). Determining the orientation and localization of membrane-bound peptides. *Current Protein and Peptide Science*, *13*(3), 267-279.

Humphrey, W., Dalke, A. and Schulten, K., "VMD - Visual Molecular Dynamics", J. Molec. Graphics, 1996, vol. 14, pp. 33-38.

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in science & engineering*, *9*(3), 90.

Jeong, C. S., & Kim, D. (2012). Reliable and robust detection of coevolving protein residues. *Protein Engineering, Design & Selection*, *25*(11), 705-713.

Jones, E., Oliphant, T., & Peterson, P. (2014). {SciPy}: Open source scientific tools for {Python}.

Kelm, S., Shi, J., & Deane, C. M. (2010). MEDELLER: homology-based coordinate generation for membrane proteins. *Bioinformatics*, *26*(22), 2833-2840.

Koehler Leman, J., Ulmschneider, M. B., & Gray, J. J. (2015). Computational modeling of membrane proteins. *Proteins: Structure, Function, and Bioinformatics*, *83*(1), 1-24.

Martz E, Hodis E, 2013, "Conservation, Evolutionary", *Proteopedia*, DOI: https://dx.doi.org/10.14576/108030.1766168

McKinney, W. (2010, June). Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference* (Vol. 445, pp. 51-56).

Mezei, M. (2003) A new method for mapping macromolecular topography, *J. Mol. Graph. Model.*, 21, 463-472.

Mount, D. W. (2001). *Bioinformatics: sequence and genome analysis* (Vol. 2). New York:: Cold spring harbor laboratory press.

Murata, M., Richardson, J. S., & Sussman, J. L. (1985). Simultaneous comparison of three protein sequences. *Proceedings of the National Academy of Sciences*, *82*(10), 3073-3077.

Needleman, S. B., & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, *48*(3), 443-453.

Oliphant, T. E. (2006). *A guide to NumPy* (Vol. 1, p. 85). USA: Trelgol Publishing.

Olivella M, Gonzalez A, Pardo L, Deupi X (2013) Relation between sequence and structure in membrane proteins. Bioinformatics 29: 1589-1592.

Pele J, Moreau M, Abdi H, Rodien P, Castel H, et al. (2014) Comparative analysis of sequence covariation methods to mine evolutionary hubs: examples from selected GPCR families. Proteins 82: 2141-2156.

Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., & Ferrin, T. E. (2004). UCSF Chimera—a visualization system for exploratory research and analysis. *Journal of computational chemistry*, *25*(13), 1605-1612.

Pollock, D. D., Taylor, W. R., & Goldman, N. (1999). Coevolving protein residues: maximum likelihood identification and relationship to structure. *Journal of molecular biology*, *287*(1), 187-198.

Riley, K. E.; Platts, J. A.; Rezac, J.; Hobza, P.; Hill, J. G. Assessment of the performance of MP2 and MP2 variants for the treatment of noncovalent interactions. J Phys Chem A 2012, 116(16), 4159-4169.

Rost, B. (1999). Twilight zone of protein sequence alignments. *Protein engineering*, *12*(2), 85-94.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell system technical journal*, *27*(3), 379-423.

Shatsky M, Nussinov R, Wolfson HJ (2004) A method for simultaneous alignment of multiple protein structures. Proteins 56: 143-156.

Shimizu, K., Cao, W., Saad, G., Shoji, M., & Terada, T. (2018). Comparative analysis of membrane protein structure databases. *Biochimica et Biophysica Acta (BBA)-Biomembranes*, *1860*(5), 1077-1091.Sievers, F., & Higgins, D. G. (2014). Clustal Omega, accurate alignment of very large numbers of sequences. In *Multiple sequence alignment methods* (pp. 105-116). Humana Press, Totowa, NJ.

Smith, T. F., & Waterman, M. S. (1981). Identification of common molecular subsequences. *Journal of molecular biology*, *147*(1), 195-197.

The PyMOL Molecular Graphics System, Version 2.0.5 Schrödinger, LLC.

Watson, H. C. (1969). The stereochemistry of the protein myoglobin. *Prog. Stereochem*, *4*(299), 5.

Yang, A. S., & Honig, B. (2000). An integrated approach to the analysis and modeling of protein sequences and structures. I. Protein structural alignment and a quantitative measure for protein structural distance. *Journal of molecular biology*, *301*(3), 665-678.

Yarov-Yarovoy, V., Schonbrun, J., & Baker, D. (2006). Multipass membrane protein structure prediction using Rosetta. *Proteins: Structure, Function, and Bioinformatics*, *62*(4), 1010-1025.

Yeang, C. H., & Haussler, D. (2007). Detecting coevolution in and among protein domains. *PLoS computational biology*, *3*(11), e211.

Yuan, C., Chen, H. and Kihara, D. (2012) Effective inter-residue contact definitions for accurate protein fold recognition, *BMC Bioinformatics*, 13, 292.

Zuckerkandl, E., & Pauling, L. (1965). Evolutionary divergence and convergence in proteins. In *Evolving genes and proteins* (pp. 97-166). Academic Press.

3. Objectives

3. OBJECTIVES

The aim of this thesis is to develop bioinformatics tools to assist in the study of membrane proteins. This includes a thorough characterization of membrane proteins using computational techniques and public biological data. In order to facilitate access to this information, various tools will be created in the format of web-services. These aims are detailed in:

1. Analysis of Membrane Proteins

1. Construct a non-redundant database for alpha and beta Membrane Proteins
2. Tool to analyse structural distortions induced by Residues and Sequence Motifs in the Structure of Membrane Proteins
3. Analyse inter-residue interactions in membrane Proteins

2. Computational Tools to analyse GPCRs

4. Develop a tool to analyse the Sequence Conservation in GPCRs
5. Develop a tool to place internal water molecules in GPCR structures.

4. Results

4. RESULTS

4.1. Overview

The results section comprises fives studies that were carried out during the thesis.

- ***STUDY 1****: TMalphaDB and TMbetaDB: web servers to study the structural role of sequence motifs in α-helix and β-barrel domains of membrane proteins*.
  Publication: Perea, M., Lugtenburg, I., Mayol, E., Cordomí, A., Deupí, X., Pardo, L., & Olivella, M. (2015). *TMalphaDB and TMbetaDB: web servers to study the structural role of sequence motifs in α-helix and β-barrel domains of membrane proteins*. *BMC Bioinformatics*, *16*(1), 266.
  Contribution: Construction and curation of the database, i.e., the selection of the best structure for each Uniprot code. This mostly involved writing Python scripts.

- ***STUDY 2****: Inter-residue interactions in alpha-helical transmembrane proteins*
  Publication*:* Mayol, E., Campillo, M., Cordomí, A., & Olivella, M. (2018). *Inter-residue interactions in alpha-helical transmembrane proteins*. *Bioinformatics*. bty978, https://doi.org/10.1093/bioinformatics/bty978
  Contribution: Doing the analysis and representations of all inter-residue interactions. his mostly involved writing Python scripts. I also wrote the Manuscript.

- ***STUDY 3****: Analysis of the interactions of sulfur-containing amino acids in membrane proteins*
  Publication*:* Gómez-Tamayo, J. C., Cordomí, A., Olivella, M., Mayol, E., Fourmy, D., & Pardo, L. (2016). *Analysis of the interactions of sulfur-containing amino acids in membrane proteins*. *Protein Science*, *25*(8), 1517-1524.
  Contribution: Developments of Python scripts to search for sulfur containing amino acids inter-residue interactions and the analysis of these interactions. I also performed the quantum-mechanics calculations for the different interaction pairs.

- ***STUDY 4****: GPCR-SAS: A web application for statistical analyses on G protein-coupled receptors sequences.*
  Publication: Gómez-Tamayo, J. C., Olivella, M., Ríos, S., Hoogstraat, M., Gonzalez, A., Mayol, E., Deupí, X., Campillo, M., Cordomí, A. (2018). *GPCR-SAS: A web application for statistical analyses on G protein-coupled receptors sequences*. *PloS one, 13(7), e0199843.*

  Contribution: Part of the data curation and validation of the software.

- ***STUDY 5****: Homolwat: a web server to incorporate internal water molecules into the structure of G-protein coupled receptors*
  Publication: Mayol, E., Garcia-Recio, A., Guixa-González R., Hildebrand, P., Pardo, L., Olivella, M., Cordomí, A. (2019). *Homolwat: a web server to incorporate internal water molecules into the structure of G-protein coupled receptors.* In preparation.

  Contribution: Design and validation of the software (writing of all the Python and Bash scripts for the study), curation of the database and writing of the manuscript. Part of this project was carried out during my stay in Berlin at the group of Dr. Hildebrand in the Charité - Universitätsmedizin.

Since the first structure of a MP was solved in 1985, the amount of available structures has raised substantially, especially during the last decade. By April 14th of 2019, the PDB contained to more than 5300 structures of MPs, from 884 different MPs (http://blanco.biomol.uci.edu/mpstruc). Despite the large number of membrane proteins structures available, data is redundant (some structures correspond to the same complex or to the same or similar protein) and annotations regarding the TM part of the membrane are scarce. Additionally, it is necessary to use programming tools to integrate and analyse this big amount of data in order to give insight into the comprehension of sequence and structure in membrane proteins.

We created two databases of MP database: α-helical proteins (**TMalphaDB**) and β-barrels (**TMbetaDB**). These DBs have two unique features that differentiate them from the existing databases such as **TMPDB**, **OPM** or **mpstruct** (see section 2.1): i) they contain only one (representative) structure for each different protein with available structure, and ii) they contain only the TM region. The web application on top of TMalphaDB and TMbetaDB (available at http://lmc.uab.cat/TMalphaDB/ and http://lmc.uab.cat/TMbetaDB) allow to search for specific sequence motifs on the determined TM segments and quantify structural parameters such as dihedral angles Ψ and Φ, side chain torsion angle χ_1, unit bend and unit twist (see section 2.6.2). The databases and the tools to search and analyse residues and sequence motifs in membrane proteins are part of **STUDY 1**.

Taking advantage **TMalphaDB** we performed a systematic analysis of the amino acid composition and inter-residue interactions in the TM region of alpha membrane proteins. We also identified the location of these residues within proteins and along the membrane axis. To complete the characterization, we performed the same analysis on beta membrane proteins (TMbetaDB) and alpha globular proteins (using a specially developed dataset). This permitted us to identify specific features of alpha membrane proteins (**STUDY 2**).

We characterized more deeply the interaction between Cys and Met residues with aromatic and hydrophobic residues and evaluated their strength in small-molecule model systems using *ab-initio* methods. We became interested in these interactions because they have received little interest despite several evidence that their interaction energies are strong (**STUDY 3**).

Unlike the first three studies that focus on all membrane proteins, the last two studies (**STUDIES 4 and 5**) focus on the development of tools for the study of one specific MP family: the family of GPCRs. We chose GPCRs because they are the largest family of membrane proteins and one of the most relevant drug targets.

First, we took advantage of the sequence/structural similarity among GPCR TM regions to develop a tool (**GPCR-SAS:** G-protein coupled receptor sequence analysis and statistics) that facilitates the statistical analysis of sequences including conservation, co-variance and correlation of residues or motifs within TM helices of all GPCR classes present in humans (**STUDY 4**). GPCR-SAS is a web-server application that is available at http://lmc.uab.cat/gpcrsas/.

Second, we developed a tool (**HomolWat**) that permits the incorporation of internal water molecules to GPCR structural models based on the homology to related receptors. We collected the increasing repertoire of GPCR crystal structures and gathered all internal water molecules and ions to develop a new method for placement internal water molecules in structures or models of GPCRs that contain few or do not contain internal water molecules (**STUDY 5**). HomolWat is a web application available at http://lmc.uab.cat/HW.

4.2. TMalphaDB and TMbetaDB

TMalphaDB and TMbetaDB: web servers to study the structural role of sequence motifs in α-helix and β-barrel domains of membrane proteins.

Abstract

Background
Membrane proteins represent over 25%-30% of human protein genes and account for more than 60% of drug targets due to their accessibility from the extracellular environment. The increasing number of available crystal structures of these proteins in the Protein Data Bank permits an initial estimation of their structural properties.

Description
We have developed two web servers—TMalphaDB for α-helix bundles and TMbetaDB for β-barrels—to analyse the growing repertoire of available crystal structures of membrane proteins. TMalphaDB and TMbetaDB permit to search for these specific sequence motifs in a non-redundant structure database of transmembrane segments and quantify structural parameters such as Ψ and Φ backbone dihedral angles, χ_1 side chain torsion angle, unit bend and unit twist.

Conclusions
The structural information offered by TMalphaDB and TMbetaDB permits to quantify structural distortions induced by specific sequence motifs, and to elucidate their role in the 3D structure. This specific structural information has direct implications in homology modelling of the growing sequences of membrane proteins lacking experimental structure. TMalphaDB and TMbetaDB are freely available at http://lmc.uab.cat/TMalphaDB and http://lmc.uab.cat/TMbetaDB.

Background

Membrane proteins represent over 25% of all proteins in sequenced genomes and mediate the interaction of the cell with its surroundings, including selective molecular transport, signaling, respiration and motility (Faberger *et al.*, 2010). Because of their accessibility from the extracellular environment, membrane proteins are targets of over 60% of currently marketed drugs (Overington *et al.*, 2006, Arinampathy *et al.*, 2009, Bakheet *et al.*, 2009). Due to the difficulty in over-expressing, purifying and crystallizing membrane proteins (Bill *et al.*, 2011), only 2% of the structures deposited in Protein Data Bank are membrane proteins (Berman *et al.*, 2000, Kozma *et al.*, 2013). Membrane proteins display specific features that differ from those of water-soluble ones, due to their different environment (Olivella *et al.*, 2013). For instance, the number of folds that membrane proteins can adopt is limited to α-helix bundles and β-barrels due to the physical constraints imposed by the lipid bilayer. The lipid bilayer, where the transmembrane (TM) regions are located, is predominantly lipophilic, lacks hydrogen-bonding potential, and provides little screening of electrostatic interactions. Thus, α-helix and β-sheets secondary structure elements maximize the hydrogen bond interactions among backbone atoms, whereas hydrophobic side chains are preferentially oriented toward the membrane lipids. This results in significant differences in amino acid composition (Donelly *et al.*, 1993) and in the probabilities of amino acid substitutions during evolution (Jones 1994, Li & Deber 1994) relative to globular proteins.
Biological function of membrane proteins involves conformational rearrangement of the TM regions. For example, activation of the G protein-coupled receptor family requires the binding of the C-terminal α-helix of the G protein to the intracellular cavity that is opened by the conformational rearrangement of TM6 (Rasmussen *et al.*, 2011). Similarly, multidrug transporters are flexible proteins that switch from outward-open to inward-open conformations, facilitating the release of the substrate (Masurel *et al.*, 2014). Such conformational changes require local flexibility or distortions in the TM regions, which can be provided by specific structural motifs. For instance, our laboratory has shown that serine or threonine,

either alone (Deupí *et al.*, 2010) or in combination with proline (Deupí *et al.*, 2004) induces distinctive TM distortions to accommodate the structural needs of specific protein functions (Sansuk *et al.*, 2011, Boiteux *et al.*, 2014). To address this issue, we have developed two non-redundant databases of 3D structures of TM segments consisting on α-helix bundles and β-barrels that are accessible through the TMalphaDB and TMbetaDB web servers, respectively. The main advantage of these servers is their ability to systematically survey sequences of TM regions and provide to the users main structural parameters, such as backbone Ψ and Φ dihedral angles and side chain χ_1 angle, as well as helix bend and twist angles. This structural information allows to quantify distortions induced by residues or motifs and to elucidate their role in the structure and function of membrane proteins.

Construction and Content

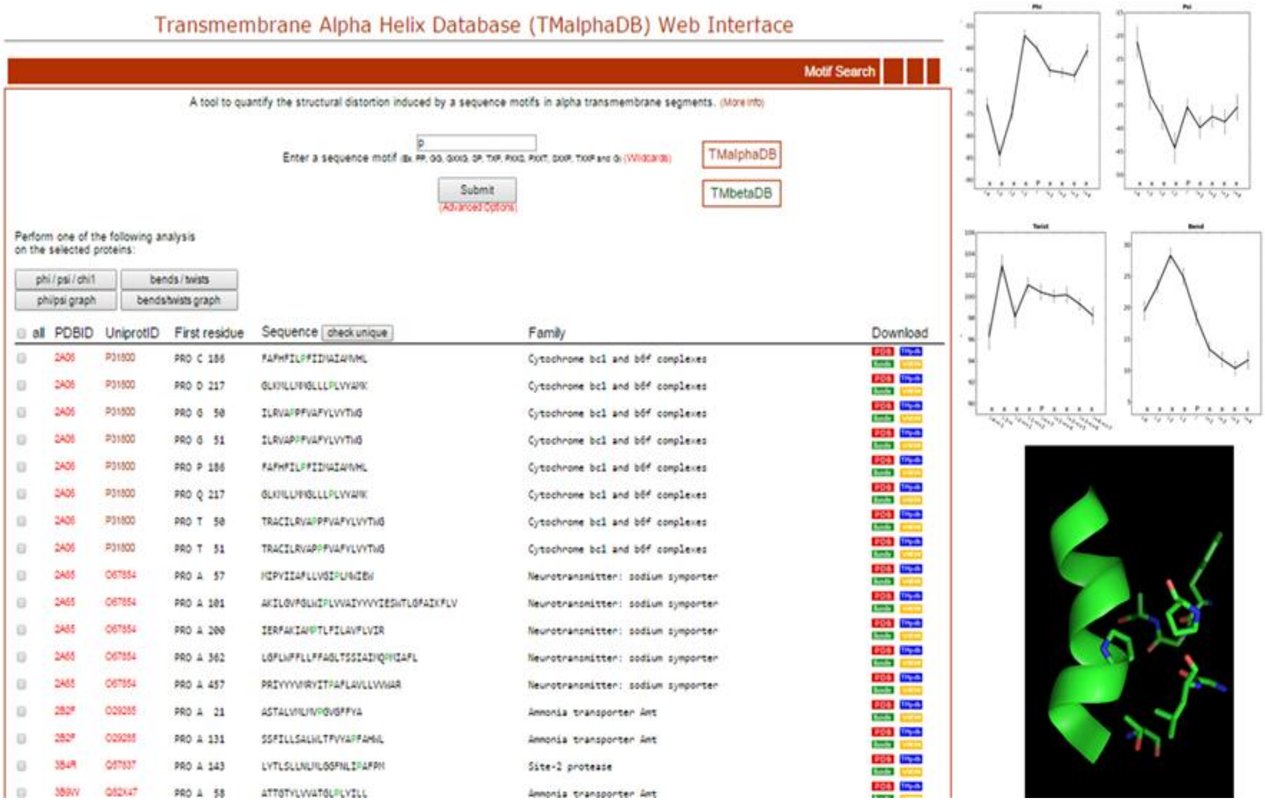
TMalphaDB and TMbetaDB are web-based servers that combine a MySQL database management system and Python programs with a dynamic web interface based on PHP.

Non-redundant databases of transmembrane segments structures of alpha and beta membrane proteins

TMalphaDB and TmbetaDB contain 330 structures of α-helix bundles and 107 structures of β-barrels, respectively, with a resolution lower than 3.5 Å. To avoid redundancy, only one structure for each protein is selected (i.e. one structure per UniProt accession code). Among different structures with the same UniProt accession code, the one with best resolution and resemblance to the native state (i.e. without mutations, native pH) is selected. Additionally, for multimeric proteins, only one subunit is extracted. The complete list of structures, together with the unique subunit database, can be downloaded at http://lmc.uab.cat/TmalphaDB/info.php and http://lmc.uab.cat/TmbetaDB/info.php. These databases are regularly and automatically updated, in order to include new solved proteins. Each structure is characterized by the Protein Data Bank identification code (PDBID) (Bernstein *et al.*, 1977), protein name, Uniprot accession code (UniProt 2014), family name according to Orientations of Proteins in Membranes (Lomize *et al.*, 2012) and organism. Moreover, because the hydrophobic nature of the lipid bilayer conditions the structure and features of the membrane-embedded regions relative to the water-exposed ones (Olivella *et al.*, 2013, Li & Deber 1994) we used PDBTM (Kozma *et al.*, 2013, Tusnady *et al.*, 2005) to download only the coordinates of the domain of the protein that is inserted in the lipid bilayer.

Tools to analyse sequence and structure of membrane proteins

The importance of TMalphaDB and TMbetaDB is their ability to search and analyse specific residues or sequence motifs in TM segments of membrane proteins. The search is performed using the single-letter amino acid code or/and in combination with wildcard characters such as '+' (positively charged, K/R), '-' (negatively charged, D/E), '*' (charged, K/R/H/D/E), '@' (aromatic, F/W/Y/H), '~' (hydrophobic, I/L/V/M/F/A/P), '^' (polar, D/E/N/Q/K/R/H/S/T/C/W/Y), '%' (aliphatic, L/V/I/M), '#' (distorting, P/G), '?' (hydroxylic, S/T/Y), '$' (sulphur-containing, C/M), '.' (tiny, G/A), '!' (aromatic amphipathic, W/Y/H), or 'x' (any amino acid). Moreover, the search (advanced options) can be filtered by the proximity of the sequence motif to the beginning/end of the TM domain (as the structural parameters can be highly influenced by the loops) or the presence of certain amino acids within the sequence motif (as, for instance, Pro or/and Gly can distort the secondary structure conformation). The output consists of a list of proteins, identified by the PDBID, Uniprot accession code, the name and identifier of the first residue in the motif, the sequence of the TM segment with the requested motif highlighted and the family name of the protein. The coordinates of the entire protein, the TM segments and/or a unique subunit of the protein can be downloaded for each entry. The user can select all TM segments, unique TM segments (i.e. only one TM segments is select for repeated subunits), or a manually selection can be performed. Average backbone Ψ and Φ angles and side chain χ_1 angle for the selected TM segments can be downloaded or/and displayed in a plot. When all the analysed sequences feature the same type of residue (according to the wildcards previously defined), for a specific position, the plot uses this representation. In TMalphaDB, bend and twist angles, two relevant parameters to measure local distortions of TM helices, are also calculated and plotted for the identified/selected TM segments using HELANAL (Bansal *et al.*, 2000) Local bend angles are calculated as the angle between the axis of the cylinders formed by the Cα atoms of the residues preceding (*i-3*, *i*) and following (*i*, *i+3*) a given amino acid *i*. Unit twist angles are calculated for sets of four consecutive Cα atoms, i.e. one turn, to analyze helical uniformity. An ideal α-helix, with approximately 3.6 residues per turn, has a unit twist of approximately 100º (360º/3.6). A closed helical segment, with <3.6 residues per turn, possesses a unit twist >100º, whereas an open helical segment, with >3.6 residues per turn, possesses a unit twist <100º. A variation greater than 20º in the unit twist angle will result in a change in the orientation of the amino acid side chain. Finally, JSMoL (Hansom *et al.*, 2013) sessions containing the coordinates of the requested motif, and all residues and ligands in its environment can also be displayed.



Transmembrane Alpha Helix Database (TMalphaDB) Web Interface

A tool to quantify the structural distortion induced by a sequence motifs in alpha transmembrane segments. (More info)

Enter a sequence motif (Ex. PP, GG, GXXG, SP, TXP, PXXG, PXXT, SXXP, TXXP and G) (Wildcards): P

Perform one of the following analysis on the selected proteins:

□ all	PDBID	UniprotID	First residue	Sequence	Family	Download
□	2A06	P31800	PRO C 186	FAFHFILPFIIDAIAVMHL	Cytochrome bc1 and b6f complexes	PDB
□	2A06	P31800	PRO D 217	GLKMLLMMGLLLPLVYAMK	Cytochrome bc1 and b6f complexes	PDB
□	2A06	P31800	PRO G 50	ILRVAPPFVAFYLVVTWG	Cytochrome bc1 and b6f complexes	PDB
□	2A06	P31800	PRO G 51	ILRVAPPFVAFYLVVTWG	Cytochrome bc1 and b6f complexes	PDB
□	2A06	P31800	PRO P 186	FAFHFILPFIIDAIAVMHL	Cytochrome bc1 and b6f complexes	PDB
□	2A06	P31800	PRO Q 217	GLKMLLMMGLLLPLVYAMK	Cytochrome bc1 and b6f complexes	PDB
□	2A06	P31800	PRO T 50	TRACILRVAPPFVAFYLVVTWG	Cytochrome bc1 and b6f complexes	PDB
□	2A06	P31800	PRO T 51	TRACILRVAPPFVAFYLVVTWG	Cytochrome bc1 and b6f complexes	PDB
□	2A65	O67854	PRO A 57	MIPVIIAFLLVGIPLMWIEW	Neurotransmitter: sodium symporter	PDB
□	2A65	O67854	PRO A 101	AKILGVFGLWIPLVVAIYVVYIESWTLGFAIKFLV	Neurotransmitter: sodium symporter	PDB
□	2A65	O67854	PRO A 200	IERFAKIAMPTLFILAVFLVIR	Neurotransmitter: sodium symporter	PDB
□	2A65	O67854	PRO A 362	LGFLWPFLLFFAGLTSSIAIMQPMIAFL	Neurotransmitter: sodium symporter	PDB
□	2A65	O67854	PRO A 457	PRIVYYVMRYITPAFLAVLLVVWAR	Neurotransmitter: sodium symporter	PDB
□	2B2F	O29285	PRO A 21	ASTALVMLMVPGVGFFYA	Ammonia transporter Amt	PDB
□	2B2F	O29285	PRO A 131	SSFILLSALWLTFVYAPFAHWL	Ammonia transporter Amt	PDB
□	3B4R	O57837	PRO A 143	LYTLSLLNLILGGFNLIPAFPM	Site-2 protease	PDB
□	3B9W	Q82X47	PRO A 58	ATTGTYLVVATGLPLVYILL	Ammonia transporter Amt	PDB

Figure 4.2-1**. Snapshots of the TMalphaDB output.** The output consists in a list of proteins containing the “P” motif (left panel), average backbone Φ and ψ angles (top right panel), average bend and twist angles (central right panel), and a JSMol session displaying all residues and ligands at a distance cutoff of 5 Å from the “P” motif (bottom right panel).

Utility and Discussion

Membrane proteins incorporate in the sequence of their TMs specific residues like Pro and Gly, introducing a flexible point and assisting in helix movements or stabilizing local regions of structural relevance (González *et al.*, 2012). In order to illustrate the use of TMalphaDB and TMbetaDB, we have surveyed and quantified structural distortions induced by P and PP motifs in TM α-helices and P and G residues in TM β-strands.

P and PP motifs in TM α-helices

Although Pro presents the smallest helix-forming tendency among naturally occurring amino acids (O'Neil & DeGrado,1990), Pro residues are often observed in TM helices (Senes *et al.*, 2000) where they induce a significant distortion. This is produced to avoid a steric clash between the pyrrolidine ring of Pro and the carbonyl oxygen of the residue in the preceding turn, leading to a bending of the helical structure (Rey *et al.*, 2010). Moreover, two consecutive Pro residues are also observed in sequences of membrane proteins. In order to study the distortion induced by the PP motif, relative to P, we scanned TMalphaDB. The search resulted in 349 unique TM helices containing P (search="P") and 8 TM helices containing PP (search="PP"). Figure 1 shows a snapshot of the obtained output for the "P" search, plots for phi and psi dihedrals and unit twist/bend and a PyMol session showing the residues located near "P". The obtained average bend angle plots for P and PP are shown to compare the structural distortion induced by each sequence motif (Figure 2A). Clearly, the distortion in bend induced by PP is lower than for P and is not the sum of individual Pro distortions, suggesting a modulating effect.

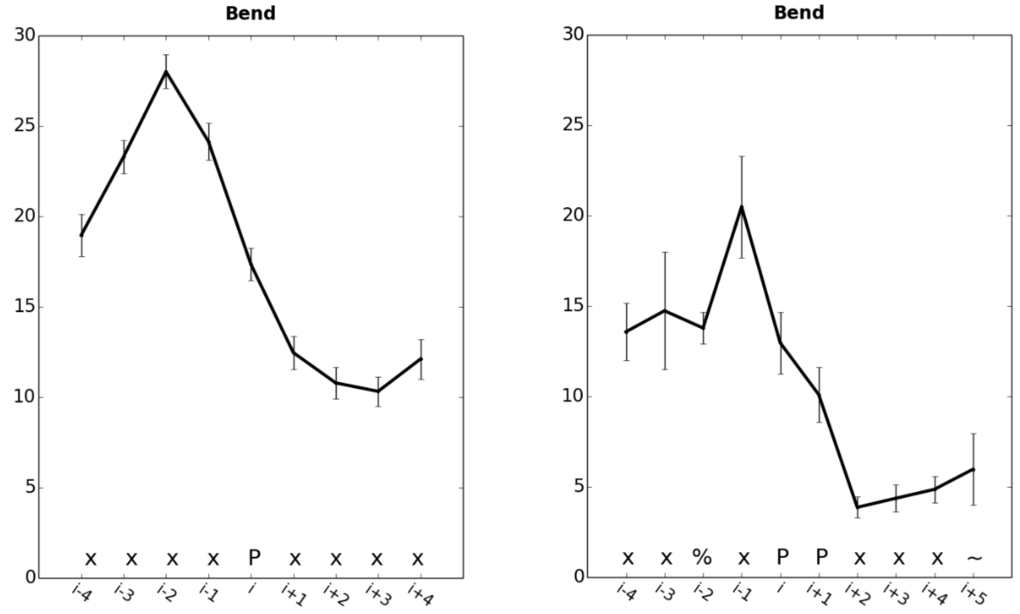


Figure 4.2-2. ***Bend angle of TM α-helices.*** Average bend angle of TM helices containing P (left panel;n=349, "P" search) and PP (right panel;n=8, "PP" search). Motifs located 4 positions from either the beginning/end of the TM domain or/and containing Pro/Gly within 4 residues of the motif were excluded.

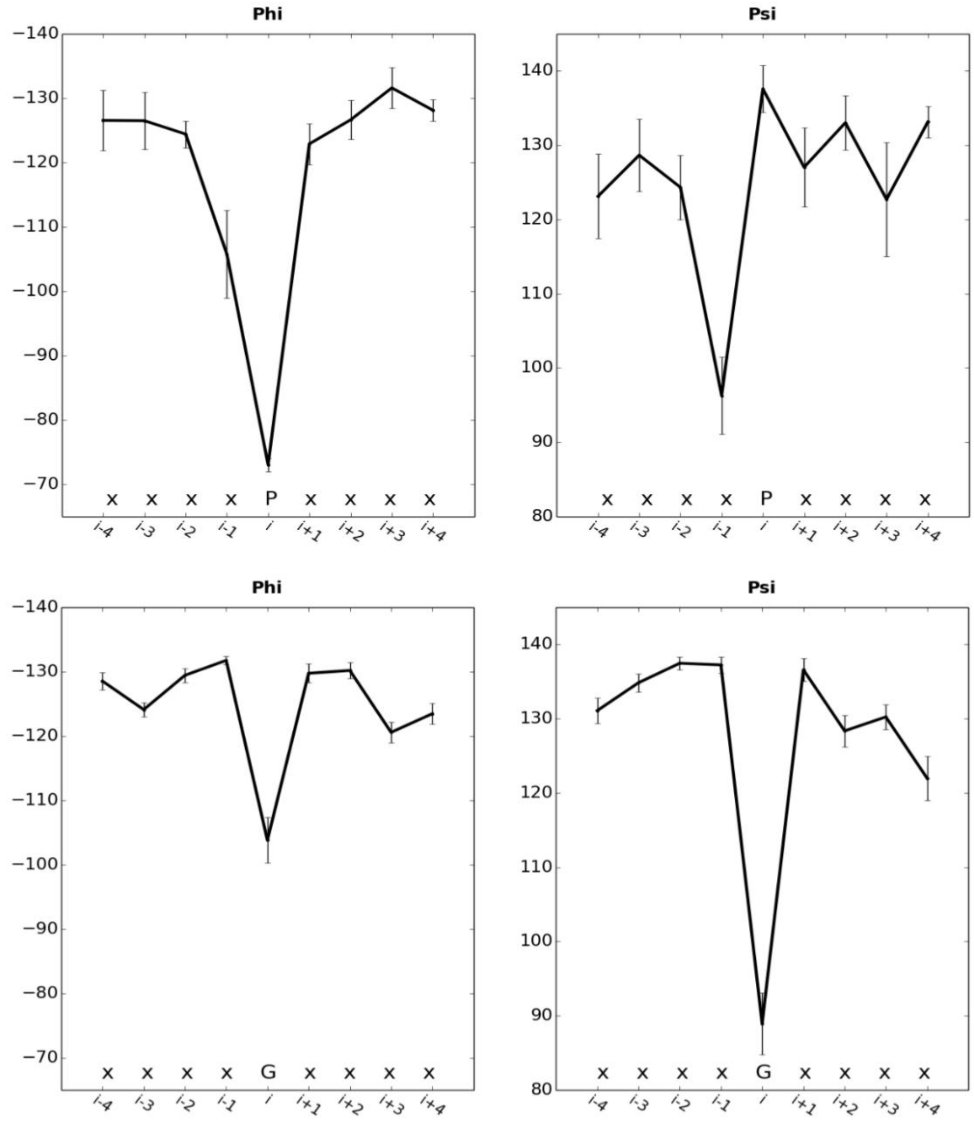


Figure 4.2-3. **Φ and ψ dihedral angles of TM β-strands.** Average Φ and ψ dihedral angles in TM β-strands containing P (n=172 search="P") and G (n=1031, search="G").

P and G in TM β-barrels

The cyclic structure of the side chain of Pro locks the Φ dihedral angle at approximately -60º, which is incompatible with Φ values near -130º observed in β-strands. We scanned TMbetaDB in order to calculate the backbone Φ and Ψ dihedral angles of Pro when located in β-barrel domains of membrane proteins. The TMbetaDB search resulted in 172 TM segments containing P whose average Φ and Ψ dihedral angles are plotted in Figure 3. Relative to the energetically preferred Φ and Ψ dihedral angles near -130º and 130º of β-strands, Pro increases Φ and triggers a decrease in Ψ at position *i-1*. In contrast to Pro, the absence of a side chain in Gly allows high flexibility in the polypeptide chain as well as dihedral angles. In order to calculate the observed Φ and Ψ dihedral angles of Gly in β-barrel domains TMbetaDB was scanned. The search resulted in 1031 TM segments containing Gly. Figure 3 shows that, on average, Gly increases Φ and decreases Ψ dihedral angles. These results indicate that both Pro and Gly induce a distortion in the conformation of main polypeptide chain in TM β-strands.

Conclusions

The structural data provided by TMalphaDB and TMbetaDB quantify structural distortions induced by specific amino acids or motifs and elucidate their role in the structure of membrane proteins. This specific structural information can be, for instance, incorporated in the homology modelling of membrane proteins lacking experimental structure. Thus, these servers emerge as valuable tools to fill the growing gap between the pool of known sequences of membrane proteins and the number of experimentally determined structures.

Availability and requirements

TMalphaDB and TMbetaDB are freely available at http://lmc.uab.cat/TMalphaDB and http://lmc.uab.cat/TMbetaDB.

References

Arinaminpathy Y, Khurana E, Engelman DM, Gerstein MB (2009) Computational analysis of membrane proteins: the largest class of drug targets. Drug Discov Today 14: 1130-1135.

Bakheet TM, Doig AJ (2009) Properties and identification of human protein drug targets. Bioinformatics 25: 451-457.

Bansal M, Kumar S, Velavan R (2000) HELANAL: a program to characterize helix geometry in proteins. J Biomol Struct Dyn 17: 811-819.

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, *et al.*, (2000) The Protein Data Bank. Nucleic Acids Res 28: 235-242.

Bernstein FC, Koetzle TF, Williams GJ, Meyer EF, Jr., Brice MD, *et al.*, (1977) The Protein Data Bank: a computer-based archival file for macromolecular structures. J Mol Biol 112: 535-542.

Bill RM, Henderson PJ, Iwata S, Kunji ER, Michel H, *et al.*, (2011) Overcoming barriers to membrane protein structure determination. Nat Biotechnol 29: 335-340.

Boiteux C, Vorobyov I, Allen TW (2014) Ion conduction and conformational flexibility of a bacterial voltage-gated sodium channel. Proc Natl Acad Sci U S A 111: 3454-3459.

Deupi X, Olivella M, Govaerts C, Ballesteros JA, Campillo M, *et al.*, (2004) Ser and Thr residues modulate the conformation of pro-kinked transmembrane alpha-helices. Biophys J 86: 105-115

Deupi X, Olivella M, Sanz A, Dolker N, Campillo M, *et al.*, (2010) Influence of the g- conformation of Ser and Thr on the structure of transmembrane helices. J Struct Biol 169: 116-123.

Donnelly D, Overington JP, Ruffle SV, Nugent JH, Blundell TL (1993) Modeling alpha-helical transmembrane domains: the calculation and use of substitution tables for lipid-facing residues. Protein Sci 2: 55-70.

Fagerberg L, Jonasson K, von Heijne G, Uhlen M, Berglund L (2010) Prediction of the human membrane proteome. Proteomics 10: 1141-1149.

Gonzalez A, Cordomí A, Caltabiano G, Campillo M, Pardo L (2012) Impact of helix irregularities on sequence alignment and homology modelling of G protein-coupled receptors. Chembiochem 13: 1393-1399.

Hanson RM, Prilusky J, Renjian Z, Nakane T, Sussman JL (2013) JSmol and the Next-Generation Web-Based Representation of 3D Molecular Structure as Applied to Proteopedia. Isr J Chem 53: 207-216.

Jones DT (1994) De novo protein design using pairwise potentials and a genetic algorithm. Protein Sci 3: 567-574.

Kozma D, Simon I, Tusnady GE (2013) PDBTM: Protein Data Bank of transmembrane proteins after 8 years. Nucleic Acids Res 41: D524-529.

Li SC, Deber CM (1994) A measure of helical propensity for amino acids in membrane environments. Nat Struct Biol 1: 368-373.

Lomize MA, Pogozheva ID, Joo H, Mosberg HI, Lomize AL (2012) OPM database and PPM web server: resources for positioning of proteins in membranes. Nucleic Acids Res 40: D370-376.

Masureel M, Martens C, Stein RA, Mishra S, Ruysschaert JM, *et al.*, (2014) Protonation drives the conformational switch in the multidrug transporter LmrP. Nat Chem Biol 10: 149-155.

Olivella M, Gonzalez A, Pardo L, Deupi X (2013) Relation between sequence and structure in membrane proteins. Bioinformatics 29: 1589-1592.

O'Neil KT, DeGrado WF (1990) A thermodynamic scale for the helix-forming tendencies of the commonly occurring amino acids. Science 250: 646-651.

Overington JP, Al-Lazikani B, Hopkins AL (2006) How many drug targets are there? Nat Rev Drug Discov 5: 993-996.

Rasmussen SG, DeVree BT, Zou Y, Kruse AC, Chung KY, *et al.*, (2011) Crystal structure of the beta2 adrenergic receptor-Gs protein complex. Nature 477: 549-555.

Rey J, Deville J, Chabbert M (2010) Structural determinants stabilizing helical distortions related to proline. J Struct Biol 171: 266-276.

Sansuk K, Deupi X, Torrecillas IR, Jongejan A, Nijmeijer S, *et al.*, (2011) A Structural Insight into the Reorientation of Transmembrane Domains 3 and 5 during Family A G Protein-Coupled Receptor Activation. Mol Pharmacol 79: 262-269.

Senes A, Gerstein M, Engelman DM (2000) Statistical Analysis of Amino Acid Patterns in Transmembrane Helices: The GxxxG Motif Occurs Frequently and in Association with b-branched Residues at Neighboring Positions. J Mol Biol 296: 921-936.

Tusnady GE, Dosztanyi Z, Simon I (2005) PDB_TM: selection and membrane localization of transmembrane proteins in the protein data bank. Nucleic Acids Res 33: D275-278.

UniProt (2014) Activities at the Universal Protein Resource (UniProt). Nucleic Acids Res 42: D191-198.

4.3. Inter-residue interactions in alpha-helical MPs

Introduction

Membrane proteins (MP) mediate the interaction between the cell interior and its surrounding and are involved in many cellular processes. They comprise receptors, ion channels, transporters and enzymes which account for approximately a quarter of the human proteome (Fagerberg, *et al.*, 2010). MPs are the targets for about two-thirds of marketed drugs (Overington, *et al.*, 2006; Santos, *et al.*, 2017). The environment and function of MPs are different from those of globular proteins and these lead to distinct properties. The TM region of MPs is predominantly lipophilic and consequently less prone to form hydrogen-bonds and with little screening of electrostatic interactions. At the sequence level, this originates differences in amino acid composition (Wallin, *et al.*, 1997), favoring residues with hydrophobic sidechains, especially at the lipid exposed surface (Eyre, *et al.*, 2004; Rees, *et al.*, 1989). Also, it is known that amino acids in membrane proteins exhibit different probabilities of amino acid substitutions during evolution compared to globular proteins (Donnelly, *et al.*, 1994), different secondary-structure propensities (Blondelle, *et al.*, 1997; Li and Deber, 1994; Monne, *et al.*, 1999; Ulmschneider and Sansom, 2001; Ulmschneider, *et al.*, 2005) and even different average backbone dihedrals for the α-helices (Olivella, *et al.*, 2002). Furthermore, the membrane restricts the possible relative orientations of helices, a fact that limits the structural diversity within MP bundles. In fact, its minimal sequence identity to preserve structure is lower than in globular proteins (Olivella, *et al.*, 2013). Knowing the 3D structures of MPs is crucial to understand their functioning at the molecular level and to develop novel pharmacological agents (Katritch, *et al.*, 2012). Although 65% of the protein families defined in PFAM (Finn, *et al.*, 2016) have a determined 3D structure (Ovchinnikov, *et al.*, 2017) the structural coverage of membrane proteins drops to only 16% (Khafizov, *et al.*, 2014), or 7% (Tsirigos, *et al.*, 2017) when we consider the coverage of TM domains only. This is due to difficulties in over-expression, purification and crystallization of membrane proteins (Bill, *et al.*, 2011). Consequently, characterization of common patterns in determined structures is an important research topic in structural biology.

An early attempt to characterize the pattern of inter-residue interactions in MPs was aimed to determine if contacts between residues (based on Cα-Cα distances) were proximal or distant in the sequence (Gromiha and Selvaraj, 2001; Gromiha and Selvaraj, 2004). This was done on a small set of crystal structures and using the limited tools then available for determining membrane spanning regions. Similar analyses of inter-residue interactions have been used to understand protein stability and the mechanisms of protein folding in globular proteins (Drablos, 1999; Gromiha and Selvaraj, 1999; Meysman, *et al.*, 2015; Miyazawa and Jernigan, 1996; Punta and Rost, 2005; Reva, *et al.*, 1997; Seno, *et al.*, 1998; Zhang and Skolnick, 1998). The strategy of mapping residue contacts using Cα-Cα distances between residues overestimates the number of contacts compared to methods that consider sidechains (Faure, *et al.*, 2008). More recently, Baeza-Delgado and co-workers addressed residue composition -but not interactions- based on a much larger dataset of membrane protein structures and obtained quantitative measurements of the distribution of the residues along the TM helices (Baeza-Delgado, *et al.*, 2013).

In this report, we present the first study covering sidechain inter-residue interactions in α-helical membrane proteins (ATM). To clear up the role of each residue in protein folding and stability, the results of inter-residue interactions are related to the location of the residues exposed either to the lipid bilayer or to the core of the protein and to the distribution of residues along the TM segment. The results are also compared to a data set of β-barrels (BTM) and α-helical globular proteins (GLOB). The elucidated common patterns of inter-residue interaction in membrane proteins give fundamental insight into membrane protein stability and can aid in the prediction of unknown structures.

Methods

TM bundles and globular proteins databases

The coordinates of the TM bundles for α-helical and β-barrel MPs were taken from TMalphaDB (http://lmc.uab.cat/TMalphaDB/) and TMbetaDB (http://lmc.uab.cat/TMbetaDB/) (Perea, *et al.*, 2015). These are two non-redundant (i.e. one structure per Uniprot accession code and one single subunit in the case of homo-multimeric proteins) databases of α-helical and β-barrel MPs with resolution ≤ 3.5 Å. The main attribute of these databases is that they capture the specific features of the TM region, as they only contain the membrane embedded portions and discard water exposed portions. ATM and BTM currently consist of 430 and 129 bundles, respectively (see Supplementary Table 1). A non-redundant dataset of alpha globular proteins (GLOB) was specially generated for this study with the same criteria as for ATM and BTM sets. Proteins classified as “alpha” by SCOP (http://scop.mrc-lmb.cam.ac.uk/scop/) (Murzin, *et al.*, 1995) and CATH (http://www.cathdb.info/) (Sillitoe, *et al.*, 2015) and not tagged as “membrane protein” were extracted from the Protein Data Bank (Berman, *et al.*, 2000). For the GLOB dataset, we only kept helices containing ten or more residues. This resulted in 1231 bundles and 19861 helices (see Supplementary Table 4.3-1). GLOB dataset is available at http://lmc.uab.cat/TMalphaDB/info.php.

Inter-residue interactions

Two residues were considered to interact if the distance between any two heavy atoms (including both sidechain and backbone) was ≤ 4.5 Å. This distance cutoff was chosen in accordance with a previous analysis (Yuan, *et al.*, 2012). Interactions were analyzed separately for sidechain-sidechain, sidechain-backbone and backbone-backbone and for inter-helical and intra-helical types. To classify each residue as buried in the protein core (‘IN’) or exposed to the environment (‘OUT’), we employed two different criteria. For ATM and GLOB the method used was the circular variance (CV) (Mezei, 2003), which provides a quantitative measure of “insideness” of residues or atoms in proteins (it ranges between 0 and 1, where 0 means completely exposed and 1 completely buried). Residues with CV ≤ 0.7 were classified as ‘OUT’ and those with CV > 0.7 were classified as ‘IN’. Interactions were considered as ‘OUT’ when both interacting residues are labeled as ‘OUT’ or as ‘IN’ otherwise. For β-barrels the criteria used was the direction of the vector connecting Cα and Cβ to classify residues as ‘IN’ (vector pointing towards the center of the barrel) or ‘OUT’ (pointing in the opposite direction). For Gly we used the vector connecting C and Cα instead. Residues classified as ‘IN’ either face the protein core (ATM and GLOB) or the central pore (BTM), whereas residues classified as ‘OUT’ either face the lipid bilayer (ATM and BTM) or the hydrophilic environment (GLOB).

Statistics

Pairwise comparisons for: (i) percentages of inter-residue participation for all residues and groups of residues in ATM to BTM and GLOB, (ii) frequencies IN and OUT for each residue in ATM, BTM and GLOB and (iii) percentages of residues and groups of residues in ATM to BTM and GLOB, were evaluated with a χ^2 of goodness-of-fit test, 0.05 significance level, with Bonferroni corrections for multiple comparisons.

Results

Inter-residue interactions in ATM. Comparison to BTM and GLOB.

We analyzed inter-residue interactions in the TM region of 3462 helices from 430 proteins (ATM dataset; see Methods). Absolute counts for the interactions are available in Supplementary Table 4.3-1. The results are presented as heatmaps that display: (i) all the interactions (Figure 4.3-1A), (ii) those in the protein interior (Figure 4.3-1B) or in the lipid-exposed protein surface (Figure 1C), (iii) those of inter-helical (Figure 1D) or intra-helical (Figure 1E) types, and (iv) interactions that involve either sidechain-sidechain, sidechain-backbone or backbone-backbone contacts (Supplementary Figures 4.3-1-2). The participation of each residue to the total number of inter-residue interactions are shown in Table 4.3-1. The distribution of the relative frequencies of inter-residue interactions in ATM presents a heterogenic profile with a noteworthy participation of aliphatic residues, followed by aromatic and polar residues (mainly Ser and Thr), and very few interactions participated by charged residues. The most frequent residue-residue pairs involve all combinations of aliphatic residues and Phe, mainly Leu-Phe, Leu-Ile and Leu-Val. Clearly, interactions involving branched aliphatic residues appear mainly at the protein surface, whereas interactions involving Ala mostly happen in the protein interior (Figure 4.3-1B-C). There are also many interactions of aliphatic residues or Phe with Ser or Thr. By contrast there are few polar-polar, charged-charged or polar-charged interactions, despite its well-known role in protein functioning (Muller, *et al.*, 2008) (Figure 4.3-1H). These interactions occur mainly in the protein core. Inter-helical interactions triplicate intra-helical interactions and are mainly performed by sidechain-sidechain interactions of hydrophobic and Phe residues and interactions between the sidechain of hydrophobic and Phe residues and the backbone of Gly and Ala residues (see Figure 4.3-1D and Supplementary Figure 4.3-2). Intra-helical interactions are mainly performed by Pro, Ser and Thr residues that interact through its sidechain with the backbone of residues in the preceding turn and also a minor contribution of hydrophobic and Phe sidechain-sidechain interactions (see Figure 4.3-1E and Supplementary Figure 4.3-2).

For comparison purposes, we analyzed a set of 129 β-barrel TM proteins (BTM) and a set of 1231 bundles (19861 helices) of α-helical globular proteins (GLOB) (see Methods; Supplementary Table 4.3-2 and Supplementary Figure 4.3-1 and 4.3-3. The distribution of interactions observed for the ATM set is remarkably different to that observed for the BTM set (Figure 4.3-1F and Supplementary Tables 4.3-3 and 4.3-4). The latter shows a more homogeneous distribution of the interactions, with less aliphatic-aliphatic and aliphatic-polar interactions and more interactions participated by aromatic. Tyr accounts for the highest participation in inter-residue interactions. Because the protein core of β-barrels is hydrophilic, the frequencies of polar-polar, polar-charged and charged-charged interactions are larger. The distribution of interactions in the GLOB set is notably similar to the ATM set (see Figure 4.3-1G and Supplementary Tables 4.3-3 and 4.3-4). The GLOB set features a similar frequency of polar interactions compared to the ATM set, more charged-charged interactions and less aliphatic interactions involving residues other than Leu.

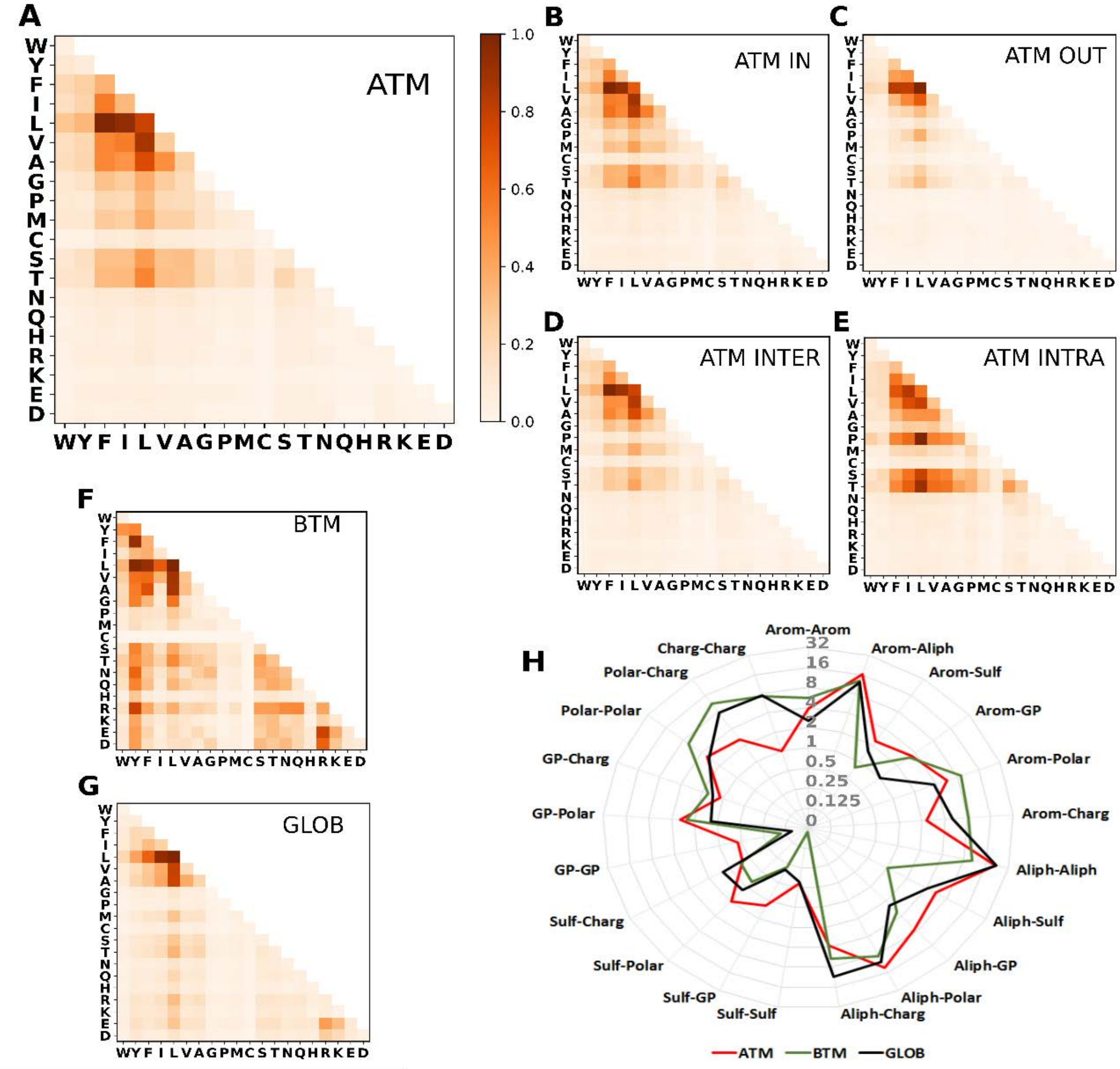


Figure 4.3-1. **The distribution of interactions in the ATM set compared to BTM and GLOB sets. (A-E)** Heatmaps of the normalized frequency of inter-residue interactions by residue in (A) α-helical membrane proteins (ATM set), (B) the interior (ATM IN) and (C) surface (ATM OUT) of α-helical membrane proteins, (D) inter-helical (INTER) and (E) intra-helical (INTRA) in α-helical membrane proteins (F) β-barrel membrane proteins (BTM set) and (G) α-helical globular proteins (GLOB set). (H) Net-plot representing the percentage of inter-residue interactions grouped by residue types: aromatic (Trp,Tyr, Phe), aliphatic (Ile, Leu, Val, Ala), Gly-Pro, sulfur containing (Met and Cys), polar (Ser, Thr, Asn, Gln), and charged (His, Arg, Lys, Glu, Asp) residues. Backbone-backbone interactions were not considered in these plots.

Table 4.3-1. Number of residues and participation in inter-residue interactions for each residue in α-helical membrane proteins.

	Number of residues	Participation in inter-residue interactions			
		Number of interacting residues		*Number of interactions*	
Trp	1846 (2.6%)	6801 (3.6%)	**Aromatic** 35234 (18.8%)	6643 (7.1%)	**Aromatic** 31650 (33.8%)
Tyr	2414 (3.4%)	8745 (4.7%)		8492 (9.1%)	
Phe	6071 (8.6%)	19688 (10.5%)		18446 (19.7%)	
Ile	7653 (10.8%)	18541 (9.9%)	**Aliphatic** 81930 (43.8%)	17389 (18.6%)	**Aliphatic** 62300 (66.6%)
Leu	11694 (16.5%)	29210 (15.6%)		26498 (28.3%)	
Val	7634 (10.8%)	17641 (9.4%)		16734 (17.9%)	
Ala	8141 (11.5%)	16538 (8.8%)		15699 (16.8%)	
Gly	6241 (8.8%)	9133 (4.9%)	**Gly-Pro** 16124 (8.6%)	9133 (9.8%)	**Gly-Pro** 15597 (16.7%)
Pro	1806 (2.5%)	6991 (3.7%)		6919 (7.4%)	
Met	2612 (3.7%)	8689 (4.6%)	**Sulfur containing** 11594 (6.2%)	8467 (9.0%)	**Sulfur containing** 11190 (12.0%)
Cys	1039 (1.5%)	2905 (1.6%)		2872 (3.1%)	
Ser	3755 (5.3%)	11384 (6.1%)	**Polar** 30659 (16.4%)	11013 (11.8%)	**Polar** 27685 (29.6%)
Thr	3728 (5.3%)	11927 (6.4%)		11541 (12.3%)	
Asn	1302 (1.8%)	4433 (2.4%)		4328 (4.6%)	
Gln	902 (1.3%)	2915 (1.6%)		2863 (3.1%)	
His	819 (1.2%)	2597 (1.4%)	**Charged** 11645 (6.2%)	2496 (2.7%)	**Charged** 10702 (11.4%)
Arg	1016 (1.4%)	2961 (1.6%)		2906 (3.1%)	
Lys	718 (1.0%)	1559 (0.8%)		1546 (1.7%)	
Glu	799 (1.1%)	2440 (1.3%)		2410 (2.6%)	
Asp	668 (0.9%)	2088 (1.1%)		2060 (2.2%)	

Absolute counts and percentages (in parentheses) of each amino acid on the composition and number of interacting residues and interactions in the α-helical membrane proteins (ATM set). Percentages of the number of interacting residues and number of interactions are calculated for each residue (or group of residues) dividing by the total number of interacting residues (187186) and by the total number of interactions (93593), respectively. Backbone-backbone interactions are not considered in this table.

The role of residues in ATM. Comparison to BTM and GLOB.

To understand the role of each residue in the above described inter-residue interactions in membrane proteins (Figure 4.3-1 and Supplementary Figure 4.3-1 and 4.3-2) it is important to know the preference of each residue (or groups of residues) for specific localizations within the bundle: for the protein interior or for the surface, and for the deep membrane regions or for interfacial regions. With this purpose we analyzed the frequency and distribution of each amino acid considering the overall protein (ALL), and residues located either in the protein core (IN) or in the lipid-exposed protein surface (OUT) (see Figure 4.3-2A and Supplementary Table 4.3-5). We also looked at the residue distribution along a membrane profile (see Figure 4.3-2D and Supplementary Figure 4.3-4). These data show that many residues do exhibit marked preference for being located at the protein interior or facing the lipid bilayer, and for lying at the center of the TM segments or at one or both (extracellular and cytoplasmic) ends. The frequencies of each residue and groups of residues and the participation to the overall number of interactions is shown in Table 4.3-1. χ^2 goodness-of-fit tests (see Methods) were performed to determine if inter-residue interactions observed in the ATM set (Figure 4.3-1) are statistically over-represented or under-represented (Supplementary Figure 4.3-5) compared to what would be expected assuming random interactions based on the observed composition (amino acid frequencies in Table 4.3-1 and Supplementary Table 4.3-5). The same analysis was carried out, for comparative purposes, in the BTM and GLOB sets (Figure 4.3-2F-G, Supplementary Table 4.3-5 and Supplementary Figure 4.3-6).

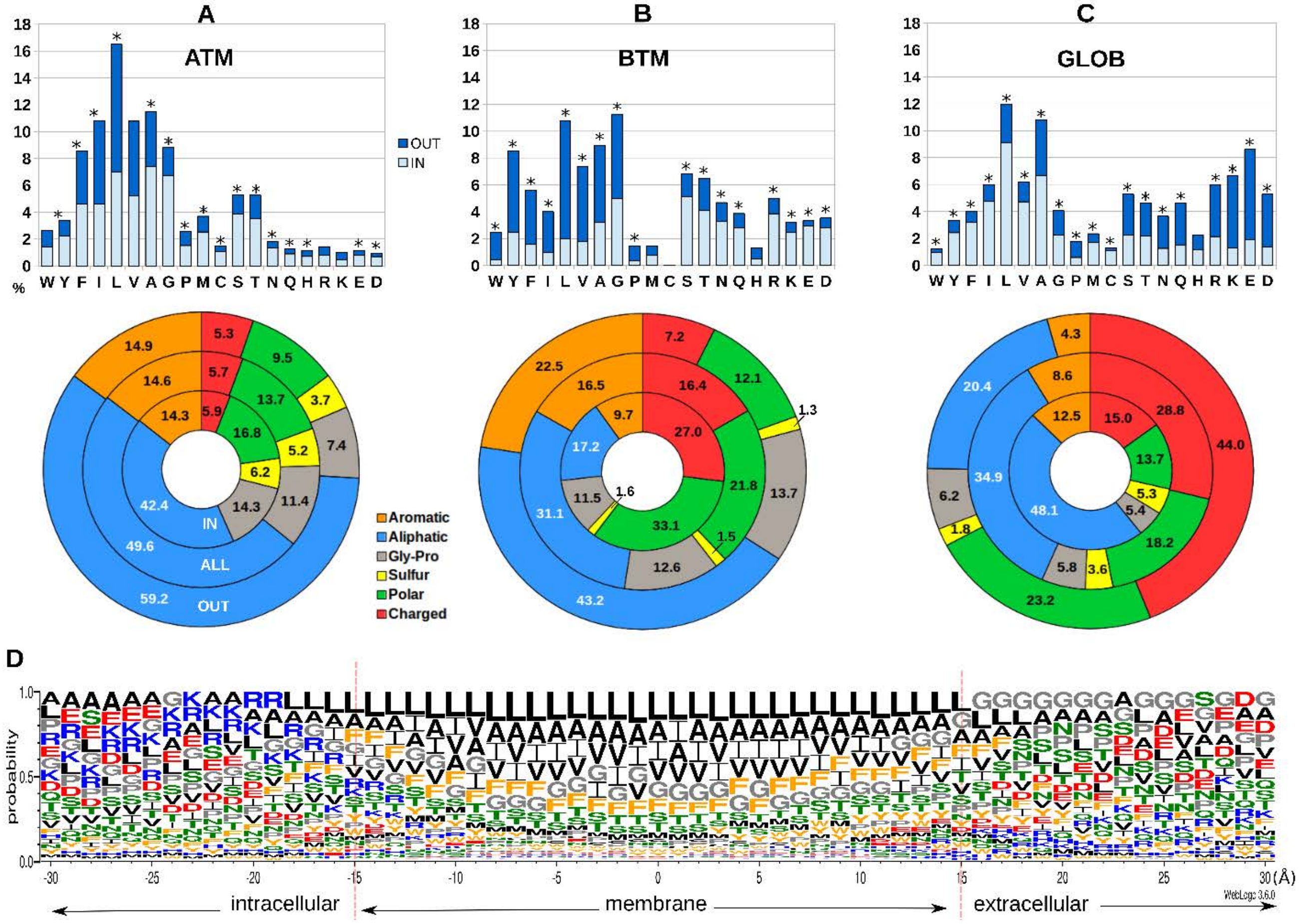


Figure 4.3-2. Composition, distribution and preferred localization of each amino acid. (A-C). The distribution of amino acid frequencies (top) and groups of residues (bottom; see Figure 1 for the definition of residue groups) in the ATM (A), BTM (B) and GLOB (C) datasets. The contribution from residues in the protein interior (IN) or in the protein surface (OUT) are also shown. * indicates statistical differences ($p<0.05$) between IN and OUT frequencies in a χ2 goodness-of-fit test with Bonferroni corrections. (D) Sequence logos for the residue composition along the TM helices of the ATM dataset. Image generated with WebLogo (Crooks, *et al.*, 2004). For each helix, the membrane center (set at 0 Å) is based on Orientations of Proteins in Membranes (Lomize, *et al.*, 2012). The membrane goes approximately between -15 (intracellular) and + 15 Å (extracellular).

Aliphatic residues

Aliphatic residues are the most common in TM helices. They account for half of the total number of residues and participate in two thirds of the inter-residue interactions in TM helices.

They are ubiquitously distributed along the hydrophobic membrane segment, though with a maximum at the bilayer center. Leu is by far the star residue and has the largest frequency (16%) and the largest participation in inter-residue interactions (28%). At the protein surface Leu, Ile, and Val are the most frequent residues, indicating that the membrane prefers to interact with branched residues. These residues interact with other Leu, Ile, Val (notably in Leu-Val, Leu-Leu and Leu-Ile interactions), as well as Phe and Ala residues located at the protein surface. These interactions are over-represented according to the individual residue frequencies. In the protein core, Ala is the most frequent aliphatic residue, followed by Leu, and consequently they are involved in most of the inter-residue interactions, mainly with other aliphatic (Leu, Ala, Val and Ile) residues and Phe through inter-helical sidechain-sidechain interactions. Ala also presents an important role in inter-helical interactions as its small side chain can allocate hydrophobic and polar side chains close to its backbone. Thus, aliphatic residues present an important role in helix packing.

In BTM, aliphatic residues are exposed to the lipid bilayer, participating in 40% of all interactions, through aliphatic-aliphatic and through aliphatic-aromatic interactions. The presence of aliphatic residues in the core of BTM is rare as water molecules can access the protein core. In GLOB, aliphatic residues participate in 60% of the interactions, mainly through aliphatic-aliphatic interactions. The core of GLOB proteins is highly rich in aliphatic residues, especially in Leu and Ala residues. Although Leu and Ala have similar frequencies, Leu participates in one third of all interactions, while Ala contributes little.

Aromatic residues

Aromatic residues account for 15% of all residues in TM helices and participate in nearly 36% of the interactions. Despite aromatic-aromatic interactions are known to contribute to a significant portion of the stabilizing forces in proteins (Goyal, *et al.*, 2017), the number of interactions of this type is scarce (about 4%). By contrast, aromatic residues interact mainly with aliphatic residues. Phe is the second most frequent residue om TM helices and by far, the most frequent aromatic residue. Phe residues are equally present in the surface or in the protein core. They exhibit a relatively flat distribution along the TM segment, though with a peak at the extracellular end. Phe mainly interacts with Leu in the surface and with aliphatic residues in general in the protein core. In fact, Phe-Leu is the most frequent inter-residue interaction in TM helices, where it has an important role in packing helices through sidechain-sidechain interhelical interactions. Tyr and Trp residues have small frequencies and have marked preference for the TM ends, in accordance with their role in interacting with the lipid head-groups (Baker, *et al.*, 2017; Muller, *et al.*, 2008; von Heijne, 1992) and with polar residues in the extracellular and intracellular interface. Whereas Tyr residues prefer the protein core, Trp residues appear equally buried in the protein interior or at the surface. Tyr and Trp participate in few inter-residue interactions, mainly with Leu. Yet, these interactions are over-represented compared to the interactions expected from the individual amino acid frequencies.

Aromatic residues account for 40% of the total interactions in BTM, mainly through Tyr and Trp, facing the lipid bilayer and performing aromatic-aromatic and aliphatic-aromatic interactions. Tyr performs a significant number of interactions, mainly with polar and charged residues in the protein core and with aliphatic and other aromatic residues in the surface of the protein. The few aromatic residues found in GLOB are mainly located at the core of the protein

Polar residues

Polar residues represent 14% of the total in TM segments and participate in 30% of the interactions, mainly involving Ser and Thr. They have clear preference for the protein core and are ubiquitously spread along the TM helix. Most of the interactions occur between the sidechain of Ser and Thr and the backbone carbonyl of the residue in the preceding turn inducing distortions (Ballesteros, *et al.*, 2000). Thus, Ser and Thr have an important role in intra-helical interactions. Few inter-helical interactions are also formed between the methylene group of Thr sidechain and hydrophobic residues or through C-H···O-H weak hydrogen bonds between the hydroxyl groups and hydrophobic residues (Desiraju, 2005). Although polar-polar interactions have an important role in the function proteins (Muller, *et al.*, 2008), there are relatively few such interaction in TM helices. Gln and Asn exhibit small frequencies and thus they participate in a marginal number of inter-residue interactions. They prefer the exterior of the protein, close to the lipid head-groups.

BTM presents the highest percentage of polar residues (around 20%) compared to ATM and GLOB. Polar residues interact with all residue types and are mainly located at the protein interior where they can interact with water molecules. Although polar residues are frequent in GLOB, they contribute little to interactions, as these residues are located at the protein surface and interact mainly with water molecules.

Gly and Pro

Gly and Pro are well known helix distorters or breakers. Gly prefers to be in the protein interior -in fact it is the second most frequent residue in the protein core- and in the central part of the TM helices. The lack of sidechain allows to allocate the sidechains of bulky neighbouring hydrophobic residues close to its backbone (Eilers, *et al.*, 2002). The neighbouring sidechains form weak C-H···O=C interactions with the backbone of Gly. Thus, Gly has an important role in backbone-sidechain interhelical interactions. Pro is present along the TM helix, although with preference for the extracellular region. Pro mainly interacts with Leu, either in the protein core or in the protein surface. Pro-aliphatic interactions are over-represented considering the frequencies of residues in TM helices. Pro has an important role in intra-helical interactions, as its sidechain interacts with the backbone of the preceding turn, as Ser and Thr residues. Thus, besides its role as helix breaker or as helix distortion inducer (Cordes, *et al.*, 2002), Pro contributes to stabilize TM bundles through intrahelical interactions.

In BTM and GLOB, Pro contributes very few interactions, whereas Gly participates in a significant number of interactions with hydrophobic and aromatic residues through its backbone. In GLOB, both Pro and Gly have small frequencies and thus little participation in interactions.

Cys and Met

The sulfur-containing residues Cys and Met show low frequencies (specially Cys) in TM helices and prefer the protein core and the center of the TM helices (see Figure 2). Cys residues interact mainly with aliphatic residues, Met and Phe. This is compatible with the fact that sulfur-containing residues form strong interactions with aliphatic and aromatic amino acids (Cordomi, *et al.*, 2013; Gómez-Tamayo, *et al.*, 2016). Most of Cys-Cys interaction distances are compatible with being non-bonded interactions rather than disulfide bonds. Met shows an important contribution to the overall inter-residue interactions despite its low frequency. In fact, interactions involving Met are over-represented compared to the interactions expected from the individual amino acid frequencies.

Cys and Met are even scarcer in BTM and GLOB and contribute few interactions. They have no preference for being IN or OUT.

Charged residues

The group of putatively charged residues (His, Lys, Arg, Glu, Asp) have the smallest frequencies in TM segments -they sum only 6%- and participate in little number of inter-residue interactions (around 11%). They are located close to the ends of the TM segments. Arg and Lys predominantly face the cytoplasmic part, following the “positive inside rule” (Baker, *et al.*, 2017; Muller, *et al.*, 2008; von Heijne, 1992). Despite the low participation in inter-residue interactions, the few charged-charged interactions are over-represented compared to the interactions expected from the individual amino acid frequencies consistent with their predicted role in protein function (Muller, *et al.*, 2008).

Charged residues are few in BTM, they are mainly found in the protein core where they perform interactions with charged and polar residues. By contrast they account for 40% of residues in the surface of globular proteins, where these residues can interact with the hydrophilic environment and with other charged residues.

Conclusions

The available membrane protein structures now cover a wide diversity of folds. At the same time, the methods to determine the TM portion of helices have become reliable. This has enabled a detailed study of inter-residue interactions in α-helical membrane proteins. The study sheds light into the contribution of each residue to the overall protein fold, and the differences relative to β-barrels and α-helical globular proteins. Our results show that most of the interactions involve aliphatic residues and Phe, with Leu participating to near one third of the overall interactions (Leu-Phe, Leu-Ile and Leu-Val are the most frequent pairs). A remarkable number of interactions implies hydrogen bonds between Thr or Ser side-chains and the backbone carbonyl of aliphatic and Phe residues. Clearly, there is a lack of polar-polar, polar-charged and charged-charged interactions.

References

Baeza-Delgado, C., Marti-Renom, M.A. and Mingarro, I. (2013) Structure-based statistical analysis of transmembrane helices, *Eur. Biophys. J.*, 42, 199-207.

Baker, J.A.*, et al.* (2017) Charged residues next to transmembrane regions revisited: "Positive-inside rule" is complemented by the "negative inside depletion/outside enrichment rule", *BMC biology*, 15, 66.

Ballesteros, J.A.*, et al.* (2000) Serine and threonine residues bend alpha-helices in the chi(1) = g(-) conformation, *Biophys. J.*, 79, 2754-2760.

Berman, H.M.*, et al.* (2000) The Protein Data Bank, *Nucleic Acids Res.*, 28, 235-242.

Bill, R.M.*, et al.* (2011) Overcoming barriers to membrane protein structure determination, *Nat. Biotechnol.*, 29, 335-340.

Blondelle, S.E.*, et al.* (1997) Secondary structure induction in aqueous vs membrane-like environments, *Biopolymers*, 42, 489-498.

Cordes, F.S., Bright, J.N. and Sansom, M.S. (2002) Proline-induced distortions of transmembrane helices, *J. Mol. Biol.*, 323, 951-960.

Cordomi, A.*, et al.* (2013) Sulfur-containing amino acids in 7TMRs: molecular gears for pharmacology and function, *Trends Pharmacol. Sci.*, 34, 320-331.

Crooks, G.E.*, et al.* (2004) WebLogo: a sequence logo generator, *Genome Res.*, 14, 1188-1190.

Desiraju, G.R. (2005) C-H...O and other weak hydrogen bonds. From crystal engineering to virtual screening, *Chemical communications*, 2995-3001.

Donnelly, D., Overington, J.P. and Blundell, T.L. (1994) The prediction and orientation of alpha-helices from sequence alignments: the combined use of environment-dependent substitution tables, Fourier transform methods and helix capping rules, *Protein Eng.*, 7, 645-653.

Drablos, F. (1999) Clustering of non-polar contacts in proteins, *Bioinformatics*, 15, 501-509.

Eilers, M.*, et al.* (2002) Comparison of helix interactions in membrane and soluble alpha-bundle proteins, *Biophys. J.*, 82, 2720-2736.

Eyre, T.A., Partridge, L. and Thornton, J.M. (2004) Computational analysis of alpha-helical membrane protein structure: implications for the prediction of 3D structural models, *Protein Eng. Des. Sel.*, 17, 613-624.

Fagerberg, L.*, et al.* (2010) Prediction of the human membrane proteome, *Proteomics*, 10, 1141-1149.

Faure, G., Bornot, A. and de Brevern, A.G. (2008) Protein contacts, inter-residue interactions and side-chain modelling, *Biochimie*, 90, 626-639.

Finn, R.D.*, et al.* (2016) The Pfam protein families database: towards a more sustainable future, *Nucleic Acids Res.*, 44, D279-285.

Gómez-Tamayo, C.J.*, et al.* (2016) Analysis of the interactions of sulfur-containing amino acids in membrane proteins, *Protein science: a publication of the Protein Society*, 25, 1517-1524.

Goyal, S.*, et al.* (2017) Role of Urea-Aromatic Stacking Interactions in Stabilizing the Aromatic Residues of the Protein in Urea-Induced Denatured State, *J. Am. Chem. Soc.*, 139, 14931-14946.

Gromiha, M.M. and Selvaraj, S. (1999) Importance of long-range interactions in protein folding, *Biophys. Chem.*, 77, 49-68.

Gromiha, M.M. and Selvaraj, S. (2001) Role of medium--and long-range interactions in discriminating globular and membrane proteins, *Int. J. Biol. Macromol.*, 29, 25-34.

Gromiha, M.M. and Selvaraj, S. (2004) Inter-residue interactions in protein folding and stability, *Prog. Biophys. Mol. Biol.*, 86, 235-277.

Katritch, V., Rueda, M. and Abagyan, R. (2012) Ligand-guided receptor optimization, *Methods Mol. Biol.*, 857, 189-205.

Khafizov, K.*, et al.* (2014) Trends in structural coverage of the protein universe and the impact of the Protein Structure Initiative, *Proc. Natl. Acad. Sci. U. S. A.*, 111, 3733-3738.

Li, S.C. and Deber, C.M. (1994) A measure of helical propensity for amino acids in membrane environments, *Nat. Struct. Biol.*, 1, 558.

Lomize, M.A.*, et al.* (2012) OPM database and PPM web server: resources for positioning of proteins in membranes, *Nucleic Acids Res.*, 40, D370-376.

Meysman, P., *et al.* (2015) Mining the entire Protein DataBank for frequent spatially cohesive amino acid patterns, *BioData mining*, 8, 4.

Mezei, M. (2003) A new method for mapping macromolecular topography, *J. Mol. Graph. Model.*, 21, 463-472.

Miyazawa, S. and Jernigan, R.L. (1996) Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading, *J. Mol. Biol.*, 256, 623-644.

Monne, M., Hermansson, M. and von Heijne, G. (1999) A turn propensity scale for transmembrane helices, *J. Mol. Biol.*, 288, 141-145.

Muller, D.J., Wu, N. and Palczewski, K. (2008) Vertebrate membrane proteins: structure, function, and insights from biophysical approaches, *Pharmacol. Rev.*, 60, 43-78.

Murzin, A.G., *et al.* (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures, *J. Mol. Biol.*, 247, 536-540.

Olivella, M., *et al.* (2002) Influence of the environment in the conformation of alpha-helices studied by protein database search and molecular dynamics simulations, *Biophys. J.*, 82, 3207-3213.

Olivella, M., *et al.* (2013) Relation between sequence and structure in membrane proteins, *Bioinformatics*, 29, 1589-1592.

Ovchinnikov, S., *et al.* (2017) Protein structure determination using metagenome sequence data, *Science*, 355, 294-298.

Overington, J.P., Al-Lazikani, B. and Hopkins, A.L. (2006) How many drug targets are there?, *Nature reviews. Drug discovery*, 5, 993-996.

Perea, M., *et al.* (2015) TMalphaDB and TMbetaDB: web servers to study the structural role of sequence motifs in alpha-helix and beta-barrel domains of membrane proteins, *BMC Bioinformatics*, 16, 266.

Punta, M. and Rost, B. (2005) Protein folding rates estimated from contact predictions, *J. Mol. Biol.*, 348, 507-512.

Rees, D.C., DeAntonio, L. and Eisenberg, D. (1989) Hydrophobic organization of membrane proteins, *Science*, 245, 510-513.

Reva, B.A., *et al.* (1997) Residue-residue mean-force potentials for protein structure recognition, *Protein Eng.*, 10, 865-876.

Santos, R., *et al.* (2017) A comprehensive map of molecular drug targets, *Nature reviews. Drug discovery*, 16, 19-34.

Seno, F., Maritan, A. and Banavar, J.R. (1998) Interaction potentials for protein folding, *Proteins*, 30, 244-248.

Sillitoe, I., *et al.* (2015) CATH: comprehensive structural and functional annotations for genome sequences, *Nucleic Acids Res.*, 43, D376-381.

Tsirigos, K.D., *et al.* (2017) Topology of membrane proteins-predictions, limitations and variations, *Curr. Opin. Struct. Biol.*, 50, 9-17.

Ulmschneider, M.B. and Sansom, M.S. (2001) Amino acid distributions in integral membrane protein structures, *Biochim. Biophys. Acta*, 1512, 1-14.

Ulmschneider, M.B., Sansom, M.S. and Di Nola, A. (2005) Properties of integral membrane protein structures: derivation of an implicit membrane potential, *Proteins*, 59, 252-265.

von Heijne, G. (1992) Membrane protein structure prediction. Hydrophobicity analysis and the positive-inside rule, *J. Mol. Biol.*, 225, 487-494.

Wallin, E., *et al.* (1997) Architecture of helix bundle membrane proteins: an analysis of cytochrome c oxidase from bovine mitochondria, *Protein Sci.*, 6, 808-815.

Yuan, C., Chen, H. and Kihara, D. (2012) Effective inter-residue contact definitions for accurate protein fold recognition, *BMC Bioinformatics*, 13, 292.

Zhang, L. and Skolnick, J. (1998) How do potentials derived from structural databases relate to "true" potentials?, *Protein Sci.*, 7, 112-122.

Supplementary Material

ATM

1BCC 1EYS 1FFT 1FX8 1GZM 1IJD 1J4N 1JB0 1KPL 1KQF 1LGH 1NKZ 1O5W 1OKC 1ORS 1P49 1PW4 1Q16 1Q90 1RC2 1RH5 1UAZ 1XIO 1XME 1XQF 1YEW 1Z98 1ZCD 1ZOY 1ZRT 2A06 2A65 2B2F 2BHW 2BL2 2BS2 2CFQ 2F2B 2GFP 2GSM 2H88 2HFE 2HVK 2HYD 2J58 2J7A 2J8C 2JAF 2NQ2 2NR9 2NTU 2OAR 2QI9 2QKS 2QTS 2R6G 2RH1 2V5Z 2WDQ 2WGM 2WIE 2WJM 2WQY 2WSC 2WSW 2X6C 2XND 2XOK 2XOV 2XQU 2YEV 2YIU 2Z5Y 2Z73 2ZUQ 2ZXE 3AM6 3AQP 3AYF 3B4R 3B9W 3BEH 3C02 3CX5 3D9S 3DDL 3DH4 3E86 3GD8 3H90 3HB3 3HD6 3HD7 3HFX 3I4D 3JYC 3K3F 3KCU 3KLY 3L1L 3LLQ 3M73 3MP7 3N5K 3NCY 3NE2 3NE5 3O7Q 3ODJ 3ODU 3ORG 3P4P 3PBL 3PCV 3Q7K 3QAP 3QBG 3QE7 3QNQ 3RGB 3RKO 3RLB 3RLF 3RQW 3RZE 3S8F 3SYA 3T9N 3TDO 3TLM 3TUI 3UG9 3UKM 3UON 3UX4 3V2Y 3V3C 3VMA 3VMT 3VOU 3VVN 3VW7 3W9I 3WDO 3WFD 3WGU 3WME 3WMM 3WO6 3WQJ 3WVF 3WXV 3WXW 3ZE3 3ZK1 3ZOJ 3ZUX 4A01 4A2N 4AL0 4AV3 4AYT 4B4A 4BBJ 4BEM 4BUO 4BVN 4BW5 4BWZ 4C7R 4C9G 4C9Q 4CAD 4CBK 4COF 4CSK 4CZA 4D1A 4D2E 4DAJ 4DJH 4DJK 4DKL 4DVE 4DW1 4DX5 4DXW 4ENE 4EV6 4EZC 4F35 4F4C 4F4S 4FBZ 4FC4 4FG6 4G1U 4G7V 4G7Y 4GC0 4GD3 4GX1 4H13 4H33 4H44 4HEA 4HFI 4HKR 4HUM 4HUQ 4HYG 4HYJ 4HYO 4I0U 4I7Z 4IAQ 4IB4 4IKV 4IL3 4IU9 4J05 4J7C 4JKV 4JQ6 4JR8 4JR9 4JTA 4K0J 4K1C 4K5Y 4KI0 4KJS 4KNF 4KPP 4KT0 4KX6 4LCZ 4LDS 4LP8 4LXJ 4M5B 4M64 4MBS 4MND 4MRS 4MYC 4N6H 4N7W 4NEF 4NH2 4NTF 4NV5 4O6M 4O6Y 4O93 4OAA 4OD4 4OGQ 4OR2 4P02 4P19 4P6V 4P79 4PD6 4PHU 4PHZ 4PIR 4PL0 4POP 4PX7 4PXK 4PXZ 4Q2G 4Q4A 4Q65 4QI1 4QNC 4QND 4QO2 4QTN 4QUV 4R0C 4RDQ 4RFS 4RI2 4RNG 4RP9 4RYO 4RYQ 4TKQ 4TKR 4TNW 4TPH 4TQ3 4TQU 4TSY 4TWD 4TWK 4U2P 4U4T 4U9N 4UMV 4US3 4UVM 4V1G 4W6V 4WAB 4WAV 4WD8 4WFF 4WGV 4WIS 4WMZ 4WOL 4X5M 4X89 4XK8 4XNV 4XP4 4XTL 4XU4 4Y7K 4Y9H 4YB9 4YBQ 4YMK 4YMU 4YSX 4YZF 4YZI 4Z35 4Z3N 4Z7F 4ZJ8 4ZP0 4ZR1 4ZUD 4ZW9 4ZWJ 4ZYO 5A1S 5A43 5AEX 5AEZ 5AJI 5AWW 5AX0 5AYN 5AZB 5AZD 5B0W 5B1A 5B2G 5B57 5B66 5BZ3 5C6N 5C6P 5C78 5C8J 5CBG 5CGD 5CKR 5CTG 5CXV 5D0Y 5DA0 5DHG 5DIR 5DJQ 5DOQ 5DQQ 5DSG 5DUO 5DWY 5EDL 5EE7 5EGI 5EH6 5EIK 5EKE 5EQG 5ER7 5EZM 5F1C 5FL7 5G28 5GLI 5GMY 5GPJ 5H35 5H36 5HK1 5HK7 5HYA 5I20 5I31 5I32 5IU4 5IWK 5IWS 5J4I 5JLC 5JNQ 5JSZ 5JWY 5KBN 5KBW 5KKZ 5KMD 5KO2 5KSD 5KUK 5LEV 5LRI 5LWE 5M87 5M94 5MKK 5NKQ 5SV0 5SVK 5SYT 5T1A 5T77 5TCX 5TIN 5TIS 5U09 5U1X

BTM

1A0S 1AF6 1EK9 1FEP 1I78 1K24 1KMO 1OSM 1P4T 1PHO 1PRN 1QD6 1QFG 1QJ8 1QJP 1T16 1TLY 1UYN 1WP1 1XKW 1YC9 2ERV 2F1V 2FGQ 2GR7 2GUF 2HDI 2J1N 2MPR 2O4V 2POR 2QOM 2VDF 2W16 2WJR 2X27 2X55 2X9K 2XE1 2XE3 2Y0L 2YNK 3A2S 3AEH 3ANZ 3BRY 3BS0 3CSL 3DWO 3DZM 3EFM 3EMN 3FHH 3FID 3FIP 3GP6 3JTY 3KVN 3NSG 3O0E 3O44 3QLB 3QQ2 3QRA 3RFZ 3SY7 3SY9 3SYB 3SYS 3SZD 3SZV 3T0S 3T24 3UPG 3V8X 3W9T 3WI5 3X2R 4AFK 4AIP 4AIQ 4B7O 4BUM 4C00 4CU4 4D5B 4D64 4D65 4E1S 4E1T 4EPA 4FQE 4FRT 4FRX 4FSO 4FSP 4FT6 4FUV 4GCS 4GEY 4K3B 4K3C 4MEE 4MT0 4MT4 4N4R 4N74 4N75 4Q35 4QL0 4RDR 4RHB 4RJW 4RL8 4RL9 4RLB 4RLC 4Y25 5DL5 5DL6 5DL7 5DL8 5FOK 5FP1 5FR8 5IV8 5IVA 5IXM 7AHL

GLOB

1A04 1A0P 1A17 1A26 1A32 1A4P 1A59 1A5T 1A62 1A6D 1A6M 1A6Q 1A76 1A7W 1A8O 1A9X 1AE7 1AEP 1AF7 1AH7 1AIL 1AIP 1AJ8 1AK0 1ALU 1ALV 1AOA 1AOR 1ARU 1ASH 1AU1 1AUA 1AUI 1AVC 1AVO 1AVS 1AW9 1AX8 1AXD 1AXN 1AZS 1B06 1B0B 1B0N 1B0X 1B1U 1B25 1B3Q 1B3U 1B43 1B48 1B4A 1B4F 1B4P 1B5L 1B67 1B6A 1B79 1B7V 1B89 1B8D 1B9M 1BAZ 1BBH 1BEA 1BG6 1BG7 1BG8 1BGC 1BGE 1BGF 1BGP 1BH9 1BHD 1BIA 1BJA 1BJF 1BJJ 1BK9 1BKR 1BM9 1BMO 1BOU 1BPO 1BR1 1BRW 1BS2 1BSM 1BT3 1BU3 1BUC 1BUN 1BVP 1BVS 1BWO 1C1K 1C3C 1C52 1C6R 1CB8 1CC5 1CCD 1CCR 1CDP 1CG5 1CH4 1CHU 1CI4 1CLC 1CMC 1CNO 1CNT 1CO6 1COT 1CPC 1CPQ 1CPT 1CQX 1CSH 1CTJ 1CUK 1CXC 1CY5 1CYJ 1D2T 1D2Z 1D4D 1D9C 1DBH 1DD4 1DGJ 1DGS 1DI1 1DJX 1DK5 1DK8 1DKZ 1DL2 1DLJ 1DLW 1DLY 1DM5 1DNP 1DNU 1DOF 1DOV 1DOW 1DP5 1DQE 1DTL 1DUG 1DVK 1DVO 1DW0 1DWK 1DXS 1E29 1E3P 1E6B 1E6I 1E6V 1E6Y 1E7L 1E8Y 1EAK 1ECA 1ECM 1ED1 1EE4 1EE8 1EEM 1EER 1EF1 1EFU 1EG3 1EGD 1EI7 1EIA 1ELK 1ELR 1ELW 1EM9 1EMU 1EMY 1ENH 1EQF 1ETE 1EUM 1EVS 1EVY 1EWR 1EXR 1EYH 1EYS 1EYV 1EYX 1EZ3 1EZF 1F0Y 1F1C 1F1E 1F1F 1F1M 1F1S 1F2E 1F3A 1F45 1F5N 1F6F 1F7C 1F80 1F99 1F9N 1FBV 1FC3 1FCD 1FCH 1FCY 1FE5 1FEW 1FFV 1FGJ 1FHE 1FHF 1FHJ 1FIO 1FIP 1FK5 1FNN 1FP1 1FP2 1FP3 1FPO 1FPS 1FQI 1FS0 1FS1 1FSE 1FSL 1FT5 1FT9 1FTS 1FUR 1FW1 1FW4 1FX7 1FYH 1G2N 1G2X 1G5Z 1G73 1G87 1G8I 1G8Q 1G9G 1GAI 1GAK 1GCV 1GDV 1GGQ 1GGZ 1GH0 1GJY 1GKM 1GKZ 1GLQ 1GMZ 1GNW 1GO3 1GP7 1GPJ 1GQA 1GRJ 1GS0 1GS9 1GSU 1GTE 1GU2 1GVD 1GVH 1GVJ 1GVN 1GWC 1GWI 1GWU 1GXM 1GXQ 1H12 1H1O 1H32 1H3L 1H3N 1H3O 1H4R 1H54 1H6G 1H6K 1H6P 1H7C 1H97 1H99 1HB6 1HBK 1HBN 1HC1 1HCU 1HDS 1HE1 1HF8 1HG4 1HH8 1HKQ 1HLB 1HLM 1HM6 1HN0 1HO8 1HQV 1HRO 1HSJ 1HSS 1HST 1HU3 1HUL 1HUS 1HUW 1HW1 1HW5 1HX8 1HXI 1HY5 1HYP 1HZ4 1HZF 1I0A 1I0H 1I1G 1I1R 1I27 1I2T 1I36 1I3D 1I4Y 1I5N 1I8O 1IA6 1IBR 1IDS 1IHG 1IJ5 1IJL 1IJX 1IJY 1ILE 1ILK 1IN4 1IO7 1IOK 1IOM 1IQ0 1IQC 1IQP 1IQV 1IRQ 1IRX 1IS9 1IT2 1ITH 1ITK 1IVH 1IXC 1IXS 1IYN 1IZO 1J02 1J09 1J1J 1J30 1J3U 1J55 1J5Y 1J77 1J8M 1JBG 1JDH 1JDL 1JEB 1JEQ 1JFB 1JFZ 1JGC 1JGS 1JHF 1JHG 1JI4 1JI5 1JI7 1JIA 1JIG 1JK0 1JLT 1JLV 1JMS 1JMW 1JNR 1JOG 1JR8 1JRO 1JS8 1JSW 1JUO 1JUQ 1K04 1K0D 1K0M 1K3Y 1K40 1K62 1K6K 1K8K 1K8U 1K94 1K9U 1KA8 1KB0 1KEA 1KF6 1KGS 1KHD 1KHO 1KHY 1KKC 1KL9 1KLX 1KMH 1KN1 1KNC 1KNY 1KO9 1KP8 1KRQ 1KS8 1KS9 1KT1 1KU1 1KU3 1KU5 1KU9 1KV9 1KVO 1KW4 1KXP 1KXU 1KYZ 1L0O 1L1Y 1L3P 1L6J 1L8Q 1L8S 1LB3 1LBD 1LBU 1LDD 1LE6 1LF6 1LFB 1LFK 1LHT 1LI5 1LJ8 1LJ9 1LKI 1LKP 1LLA 1LLP 1LNL

1LP1 1LRI 1LRV 1LRZ 1LS1 1LS9 1LSH 1LV2 1LVA 1LVF 1LWB 1M0U 1M15 1M1R 1M45 1M48 1M6Y 1M70 1M8T 1M98 1MA1
1MBA 1MBS 1MG6 1MGT 1MHQ 1MHY 1MI1 1MID 1MIJ 1MIX 1MKM 1MN8 1MNG 1MPG 1MQV 1MSD 1MTY 1MUN 1MXR
1MZ4 1MZB 1N00 1N1B 1N1C 1N1F 1N1J 1N1Q 1N2A 1N40 1N45 1N4K 1N5U 1N62 1N69 1N7O 1N81 1N83 1N8V 1N93 1N97
1NC5 1NF1 1NFV 1NG6 1NGN 1NHY 1NI2 1NIG 1NIR 1NKD 1NKT 1NLX 1NML 1NO1 1NOG 1NP7 1NP8 1NQ7 1NR6 1NT2 1NTY
1NVM 1NXC 1NXH 1NZN 1O0W 1O17 1O3U 1O3X 1O57 1O7D 1O7F 1O7X 1O82 1O9I 1O9R 1OAH 1OAI 1ODO 1OE8 1OFC 1OIP
1OJH 1OKR 1OKS 1OLP 1OMR 1ON2 1OOH 1OPC 1OQ9 1OQC 1OR4 1OR7 1ORJ 1OTK 1OU5 1OUT 1OUV 1OW4 1OWC 1OWL
1OXJ 1OYJ 1OYW 1OYZ 1OZ6 1P22 1P2F 1P2X 1P3Q 1P4X 1P5Q 1P7O 1PA2 1PAL 1PAQ 1PBW 1PD2 1PD3 1PDU 1PGJ 1PI1
1PK3 1PMT 1PN9 1PO5 1POC 1PP2 1PPR 1PS1 1PSR 1PU6 1PVA 1PVH 1PX5 1PXY 1PZN 1Q06 1Q08 1Q0G 1Q0Q 1Q1F 1Q1H
1Q3Q 1Q4G 1Q5D 1Q5N 1Q79 1Q8C 1QAZ 1QB2 1QC7 1QDB 1QGJ 1QGR 1QH4 1QHD 1QI9 1QK1 1QKR 1QKS 1QL3 1QLS
1QMG 1QN2 1QNT 1QO8 1QPA 1QPW 1QQE 1QQF 1QSA 1QSD 1QUU 1QV1 1QVR 1QX2 1QXP 1QZZ 1R03 1R0D 1R1T 1R1U
1R2F 1R2J 1R4V 1R5A 1R5Q 1R69 1R6T 1R76 1R7J 1R8I 1R8J 1R8S 1R9O 1RCD 1RCW 1RE5 1RFZ 1RHG 1RKT 1RM6 1RQG 1RR7
1RRH 1RRO 1RSS 1RT8 1RTR 1RTW 1RTY 1RWH 1RWY 1RX0 1RXQ 1RZ4 1RZL 1S1E 1S1F 1S29 1S35 1S3J 1S3Q 1S4K 1S69 1S6C
1S7O 1S8I 1S94 1S9U 1SCF 1SCH 1SD4 1SDI 1SED 1SES 1SFE 1SFX 1SGM 1SH5 1SIG 1SIQ 1SJ7 1SJP 1SK7 1SKV 1SKY 1SO2 1SPG
1SQG 1SR7 1STZ 1SUM 1SV0 1SXJ 1SYY 1SZ9 1SZP 1T0F 1T11 1T33 1T3Q 1T3W 1T6O 1T6S 1T6U 1T72 1T77 1T7R 1T7S 1T8K
1T98 1TAD 1TAF 1TAZ 1TBF 1TBX 1TC8 1TD6 1TF4 1TFR 1TJ7 1TJO 1TJV 1TLQ 1TU7 1TU9 1TUK 1TW3 1TW9 1TWF 1TX4 1TX9
1TXD 1TXG 1TZV 1TZY 1U00 1U2K 1U2W 1U3D 1U4J 1U5P 1U61 1U6R 1U7K 1U84 1U8V 1U9L 1UB2 1UB9 1UCR 1UDD 1UED
1UFB 1UG3 1UHK 1UHN 1UJ8 1UKW 1ULV 1ULY 1UNK 1UOU 1UP8 1UPK 1UPT 1UPV 1URU 1UT9 1UTG 1UUR 1UW4 1UX8
1UZR 1V2A 1V2Z 1V4E 1V4X 1VAP 1VCT 1VDK 1VF6 1VI0 1VIP 1VJX 1VKE 1VKU 1VLB 1VLG 1VMA 1VMG 1VP7 1VPD 1VQR
1VRP 1VYD 1W07 1W1W 1W27 1W3B 1W53 1W5T 1W6K 1W7J 1W9C 1WA5 1WB8 1WDC 1WE1 1WER 1WG8 1WLZ 1WMG
1WMU 1WOV 1WPB 1WRK 1WU3 1WZD 1X04 1X1I 1X90 1X9D 1XA6 1XAP 1XB2 1XB4 1XD7 1XF6 1XG7 1XGS 1XK4 1XLY 1XNF
1XNX 1XO1 1XO5 1XPC 1XQO 1XSM 1XSV 1XSZ 1XW6 1XWR 1XX7 1XZP 1Y0P 1Y0U 1Y2K 1Y2O 1Y88 1Y9B 1Y9I 1Y9Q 1YAR
1YBZ 1YCC 1YCQ 1YFM 1YG2 1YIO 1YLF 1YNB 1YNR 1YOZ 1YU5 1YYD 1YYV 1Z05 1Z0P 1Z0X 1Z3E 1Z67 1Z72 1Z7U 1Z8O 1ZAR
1ZCA 1ZCB 1ZK8 1ZKE 1ZL7 1ZWP 1ZWW 1ZX3 1ZYM 256B 2A06 2A0B 2A2R 2A5Y 2A61 2A6H 2ABK 2AHR 2AIB 2AO9 2AP3
2APL 2ASR 2AU5 2AUW 2AXI 2B50 2B6C 2BGC 2BJ7 2BO3 2BS2 2BUO 2BWB 2C08 2C0G 2C4J 2C8S 2CCA 2CCY 2CFX 2CG4
2CIW 2CJ4 2CPG 2CRO 2CVD 2CVZ 2CYY 2D29 2D5B 2D5X 2DDH 2DT5 2DTR 2DW4 2E2A 2E2R 2ELC 2END 2ERL 2ESH 2ESN
2ETD 2F2E 2F6S 2FB1 2FBA 2FBH 2FBI 2FBQ 2FD5 2FEF 2FHE 2FI0 2FJC 2FKZ 2FML 2FNA 2FOK 2FQ4 2FSW 2FUP 2FXA 2G03
2G3B 2G7G 2G7S 2GA1 2GAU 2GBO 2GC7 2GDM 2GEN 2GF4 2GFN 2GM8 2GMF 2GMY 2GNO 2GRR 2GS4 2GSQ 2GSR 2GST
2GYQ 2GZ4 2H5N 2H6F 2HBG 2HEK 2HH6 2HHP 2HKJ 2HKU 2HKV 2HMZ 2HOE 2HR2 2HR3 2HS5 2HTS 2HUE 2HUJ 2HYJ 2HZA
2HZT 2I10 2I15 2I1Q 2I5U 2I6H 2I76 2I9C 2IAZ 2IB0 2ICT 2ICW 2ID3 2ID6 2IDG 2IE7 2II0 2IJ2 2IJQ 2IM8 2INC 2IPQ 2ITB 2IZY
2J07 2J5Y 2J8W 2JDI 2LHB 2LIS 2LJR 2MHR 2NN4 2NNJ 2NP3 2NP5 2NR5 2NS0 2O02 2O35 2O38 2O3F 2O3L 2O4D 2O4J 2O4T
2O6K 2O7T 2O8I 2O8R 2O8S 2OBP 2OC5 2OEB 2OEE 2OEQ 2OFY 2OH3 2OI8 2OKU 2OO2 2OOC 2OPO 2OQM 2OTA 2OU6
2OUW 2OUX 2OX6 2OY9 2OYO 2OZ6 2P06 2P0M 2P0T 2P1A 2P1T 2P54 2P61 2P8T 2PBE 2PFX 2PG4 2PGD 2PIH 2PJQ 2PMR
2PNR 2PPX 2PQ7 2PRR 2PV4 2PV7 2PYQ 2Q0T 2Q0Z 2Q4C 2Q4X 2QGS 2QJZ 2QKW 2QSB 2QSS 2QZG 2R25 2R7G 2R9G 2RDZ
2RI9 2RLD 2SAS 2SCP 2SPC 2SQC 2TPT 2V1F 2VAN 2VHB 2VHD 2VKV 2VLQ 2YW6 2ZFD 2ZPY 3B57 3B7S 3BEJ 3BES 3BGE 3BJ1
3BLH 3BQO 3BRJ 3BRO 3BU8 3BUL 3BUX 3BVU 3BWG 3BZ6 3BZK 3C07 3C1V 3C2C 3CI0 3CJT 3CKC 3CTA 3CTD 3CZH 3D1K
3D7C 3DEU 3DHZ 3DJB 3DLQ 3DSS 3DTO 3DYN 3FAP 3LYN 3MDD 451C 4GTU 5CSM 5CYT 5PAL

Supplementary Table 4.3-1: List of Protein Data Bank (Bernstein *et al.*, 1977) IDs for the proteins that compose the ATM, BTM and GLOB datasets.

ATM

W	158																			
Y	386	253																		
F	688	857	1242																	
I	567	775	1940	1152																
L	977	1179	3459	3236	2712															
V	556	707	1810	1918	3025	907														
A	599	731	1789	1620	2540	1681	839													
G	358	510	1045	882	1290	903	646	0												
P	339	293	690	642	1126	652	613	455	72											
M	358	378	928	853	1306	830	826	484	318	222										
C	91	112	325	262	453	282	253	181	144	149	33									
S	359	563	1063	1064	1534	1033	1080	685	418	501	172	371								
T	450	491	1141	1162	1819	1094	1053	713	447	529	176	742	386							
N	180	264	326	308	431	330	361	225	162	153	87	399	315	105						
Q	106	177	262	191	276	197	245	162	116	162	42	247	214	143	52					
H	126	142	222	155	273	207	182	202	59	106	34	187	201	78	54	101				
R	103	219	235	216	299	193	210	140	99	124	19	186	210	122	78	45	55			
K	76	110	118	131	154	92	118	53	63	80	18	100	91	74	30	28	44	13		
E	78	178	192	189	237	184	158	134	137	92	17	151	161	95	71	61	137	65	30	
D	88	167	114	126	172	133	155	65	74	68	22	158	146	170	38	33	172	88	43	28
	W	Y	F	I	L	V	A	G	P	M	C	S	T	N	Q	H	R	K	E	D

BTM

W	21																			
Y	186	203																		
F	117	361	141																	
I	54	129	143	46																
L	137	382	362	264	392															
V	99	234	241	138	347	116														
A	74	218	279	51	342	122	39													
G	77	213	136	34	231	50	26	0												
P	31	64	58	51	107	66	27	13	14											
M	32	66	45	34	76	34	35	40	11	13										
C	0	1	1	1	1	0	0	0	0	0	0									
S	52	209	133	61	152	97	77	77	32	33	0	77								
T	96	192	140	93	202	121	108	88	37	47	1	166	115							
N	76	251	108	52	109	53	99	126	18	28	0	149	142	117						
Q	61	208	122	34	137	94	68	69	17	32	1	122	150	134	45					
H	19	110	33	14	36	20	24	34	5	8	0	39	48	28	18	8				
R	61	307	157	85	143	79	92	76	17	59	0	171	172	199	202	32	159			
K	28	155	61	36	62	48	30	18	7	16	1	110	123	72	86	12	91	43		
E	45	172	88	25	72	44	73	35	5	20	0	94	125	98	79	37	307	150	40	
D	39	169	65	33	61	64	39	87	18	21	0	128	106	130	84	37	256	111	40	30
	W	Y	F	I	L	V	A	G	P	M	C	S	T	N	Q	H	R	K	E	D

(**Supplementary Table 4.3-2; continues)**

GLOB

W	169																			
Y	452	510																		
F	680	1403	1245																	
I	670	1414	2272	1660																
L	1427	2928	4625	6719	7114															
V	616	1199	1965	2697	5633	1250														
A	634	1536	2016	2807	5548	2722	1544													
G	213	399	515	493	930	471	447	0												
P	96	284	255	338	658	336	455	141	42											
M	292	629	970	1014	2085	967	993	213	133	292										
C	114	284	446	425	897	344	446	116	74	240	182									
S	336	727	875	1093	1993	1028	1254	323	302	411	179	391								
T	387	747	951	1327	2480	1221	1526	392	270	468	194	807	455							
N	187	493	462	512	950	539	643	223	177	241	125	591	535	264						
Q	255	566	527	742	1431	669	771	207	230	325	103	590	756	512	315					
H	228	530	476	566	1020	485	515	153	133	218	103	360	365	218	309	233				
R	389	943	794	1093	2234	960	1143	338	338	423	189	911	915	723	854	435	530			
K	258	869	671	856	1673	706	785	197	239	307	158	699	723	586	634	246	552	220		
E	384	1089	823	1078	2027	989	1014	262	426	376	141	1032	1086	717	860	735	3094	2231	612	
D	200	735	430	471	920	496	623	210	179	210	78	712	697	458	537	462	1787	1280	628	239
	W	**Y**	**F**	**I**	**L**	**V**	**A**	**G**	**P**	**M**	**C**	**S**	**T**	**N**	**Q**	**H**	**R**	**K**	**E**	**D**

ATM IN

W	137																			
Y	350	234																		
F	607	776	1015																	
I	461	684	1582	763																
L	816	1035	2785	2571	1889															
V	472	650	1540	1491	2469	711														
A	559	689	1653	1445	2278	1519	788													
G	349	487	983	834	1174	836	634	0												
P	303	264	599	471	827	526	534	416	66											
M	329	356	843	776	1189	765	788	471	290	207										
C	79	107	289	235	401	271	242	179	137	146	32									
S	324	532	983	948	1334	936	1033	664	375	480	169	354								
T	398	458	1021	988	1555	954	948	668	415	487	169	687	352							
N	164	246	305	295	406	317	351	220	151	131	87	387	300	102						
Q	100	166	245	175	257	190	242	159	108	155	42	235	202	137	50					
H	109	133	204	135	238	189	162	188	53	100	34	175	180	75	50	97				
R	82	187	202	195	262	169	196	133	75	113	19	179	194	108	75	40	42			
K	66	94	96	108	124	82	110	51	49	73	18	83	73	62	25	28	37	9		
E	71	172	182	176	217	177	152	128	122	89	17	143	151	94	69	59	116	57	28	
D	87	159	108	121	166	128	153	61	68	65	20	154	141	166	37	31	164	81	42	25
	W	**Y**	**F**	**I**	**L**	**V**	**A**	**G**	**P**	**M**	**C**	**S**	**T**	**N**	**Q**	**H**	**R**	**K**	**E**	**D**

(**Supplementary Table 4.3-2; continues)**

ATM OUT

W	21																			
Y	36	19																		
F	81	82	227																	
I	106	91	359	389																
L	161	144	676	665	823															
V	84	57	271	427	556	196														
A	40	42	135	174	262	162	51													
G	9	23	62	48	116	67	12	0												
P	36	29	91	171	299	126	79	39	6											
M	29	22	87	79	117	65	38	13	28	15										
C	12	5	36	27	52	11	11	2	7	3	1									
S	34	31	82	117	202	98	47	21	43	21	3	17								
T	52	33	120	174	264	140	105	45	32	42	7	55	34							
N	16	18	21	13	25	13	10	5	11	22	0	13	15	3						
Q	6	11	17	16	19	7	3	3	8	7	0	12	12	6	2					
H	17	9	18	20	35	18	20	14	6	6	0	12	21	3	4	4				
R	21	32	33	21	37	24	14	7	24	11	0	7	16	14	3	5	13			
K	10	16	22	23	30	10	8	2	14	7	0	17	18	12	5	0	7	4		
E	7	6	10	13	20	7	6	6	15	3	0	8	10	1	2	2	21	8	2	
D	1	8	6	5	6	5	2	4	6	3	2	5	5	4	1	2	8	7	1	3
	W	Y	F	I	L	V	A	G	P	M	C	S	T	N	Q	H	R	K	E	D

ATM INTER

W	120																			
Y	337	212																		
F	591	738	1044																	
I	467	657	1568	914																
L	869	1066	3010	2700	2337															
V	476	600	1487	1450	2517	718														
A	516	653	1577	1323	2229	1374	728													
G	310	460	923	716	1131	723	558	0												
P	178	187	352	244	483	302	288	197	29											
M	316	340	804	714	1154	718	720	422	185	198										
C	82	94	279	206	390	225	215	157	64	135	31									
S	274	431	799	689	1028	670	743	487	198	369	124	252								
T	340	393	813	740	1234	704	678	439	216	391	125	451	215							
N	136	212	244	221	349	246	295	171	100	134	73	286	224	77						
Q	84	142	210	133	233	153	205	144	78	130	39	187	130	106	41					
H	101	123	180	116	220	156	136	177	22	92	30	159	157	60	50	93				
R	88	183	191	177	250	157	170	118	48	97	15	161	160	98	64	29	34			
K	58	97	90	104	118	74	95	38	35	69	14	67	66	62	22	21	36	11		
E	60	158	159	155	196	146	126	111	94	79	15	106	96	81	58	46	112	51	26	
D	80	145	87	93	125	95	126	46	44	54	15	118	97	144	29	25	139	67	38	23
	W	Y	F	I	L	V	A	G	P	M	C	S	T	N	Q	H	R	K	E	D

(**Supplementary Table 4.3-2; continues)**

ATM INTRA

W	38																			
Y	49	41																		
F	97	119	198																	
I	100	118	372	238																
L	108	113	449	536	375															
V	80	107	323	468	508	189														
A	83	78	212	297	311	307	111													
G	48	50	122	166	159	180	88	0												
P	161	106	338	398	643	350	325	258	43											
M	42	38	124	139	152	112	106	62	133	24										
C	9	18	46	56	63	57	38	24	80	14	2									
S	85	132	264	375	506	363	337	198	220	132	48	119								
T	110	98	328	422	585	390	375	274	231	138	51	291	171							
N	44	52	82	87	82	84	66	54	62	19	14	113	91	28						
Q	22	35	52	58	43	44	40	18	38	32	3	60	84	37	11					
H	25	19	42	39	53	51	46	25	37	14	4	28	44	18	4	8				
R	15	36	44	39	49	36	40	22	51	27	4	25	50	24	14	16	21			
K	18	13	28	27	36	18	23	15	28	11	4	33	25	12	8	7	8	2		
E	18	20	33	34	41	38	32	23	43	13	2	45	65	14	13	15	25	14	4	
D	8	22	27	33	47	38	29	19	30	14	7	40	49	26	9	8	33	21	5	5
	W	**Y**	**F**	**I**	**L**	**V**	**A**	**G**	**P**	**M**	**C**	**S**	**T**	**N**	**Q**	**H**	**R**	**K**	**E**	**D**

Supplementary Table 4.3-2. Absolute frequencies of all inter-residue interactions in the ATM, (A) BTM (B) and GLOB (C) sets and, for the ATM set only, separately for interactions in the protein interior (ATM IN, D), at the lipid-exposed protein surface (ATM OUT, E), inter-helical (ATM INER, F) and intra-helical (ATM INTRA, G).

residue	ATM		BTM			GLOB		
Trp	6801	(3.6%)	1326	(3.6%)		8156	(2.3%)	*
Tyr	8745	(4.7%)	4033	(11.0%)	*	18247	(5.2%)	*
Phe	19688	(10.5%)	2932	(8.0%)	*	23646	(6.8%)	*
Ile	18541	(9.9%)	1424	(3.9%)	*	29907	(8.6%)	*
Leu	29210	(15.6%)	4007	(10.9%)	*	60406	(17.3%)	*
Val	17641	(9.4%)	2183	(6.0%)	*	26543	(7.6%)	*
Ala	16538	(8.8%)	1862	(5.1%)	*	28966	(8.3%)	*
Gly	9133	(4.9%)	1430	(3.9%)	*	6243	(1.8%)	*
Pro	6991	(3.7%)	612	(1.7%)	*	5148	(1.5%)	*
Met	8689	(4.6%)	663	(1.8%)	*	11099	(3.2%)	*
Cys	2905	(1.6%)	7	(0.0%)	*	5020	(1.4%)	*
Ser	11384	(6.1%)	2056	(5.6%)	*	15005	(4.3%)	*
Thr	11927	(6.4%)	2387	(6.5%)		16757	(4.8%)	*
Asn	4433	(2.4%)	2106	(5.7%)	*	9420	(2.7%)	*
Gln	2915	(1.6%)	1808	(4.9%)	*	11508	(3.3%)	*
His	2597	(1.4%)	570	(1.6%)		8023	(2.3%)	*
Arg	2961	(1.6%)	2824	(7.7%)	*	19175	(5.5%)	*
Lys	1559	(0.8%)	1303	(3.6%)	*	14110	(4.0%)	*
Glu	2440	(1.3%)	1589	(4.3%)	*	20216	(5.8%)	*
Asp	2088	(1.1%)	1548	(4.2%)	*	11591	(3.3%)	*

group	ATM		BTM			GLOB		
Aromatic	81930	(43.8%)	9476	(25.8%)	*	145822	(41.8%)	*
Aliphatic	35234	(18.8%)	8291	(22.6%)	*	50049	(14.3%)	*
Gly-Pro	16124	(8.6%)	2042	(5.6%)	*	11391	(3.3%)	*
Sulfur	11594	(6.2%)	670	(1.8%)	*	16119	(4.6%)	*
Polar	30659	(16.4%)	8357	(22.8%)	*	52690	(15.1%)	*
Charged	11645	(6.2%)	7834	(21.4%)	*	73115	(20.9%)	*

Supplementary Table 4.3-3. Absolute counts and percentages (in parentheses) of each residue (top) and groups of residues (bottom) of the number of interacting residues in the ATM, BTM and GLOB sets. Percentages are calculated dividing by the total number of interacting residues. * indicates statistical differences ($p<0.05$) between ATM and BTM or ATM and GLOB in a χ^2 goodness-of-fit test with Bonferroni corrections. Residue groups are defined in Figure 1. Backbone-backbone interactions are not considered.

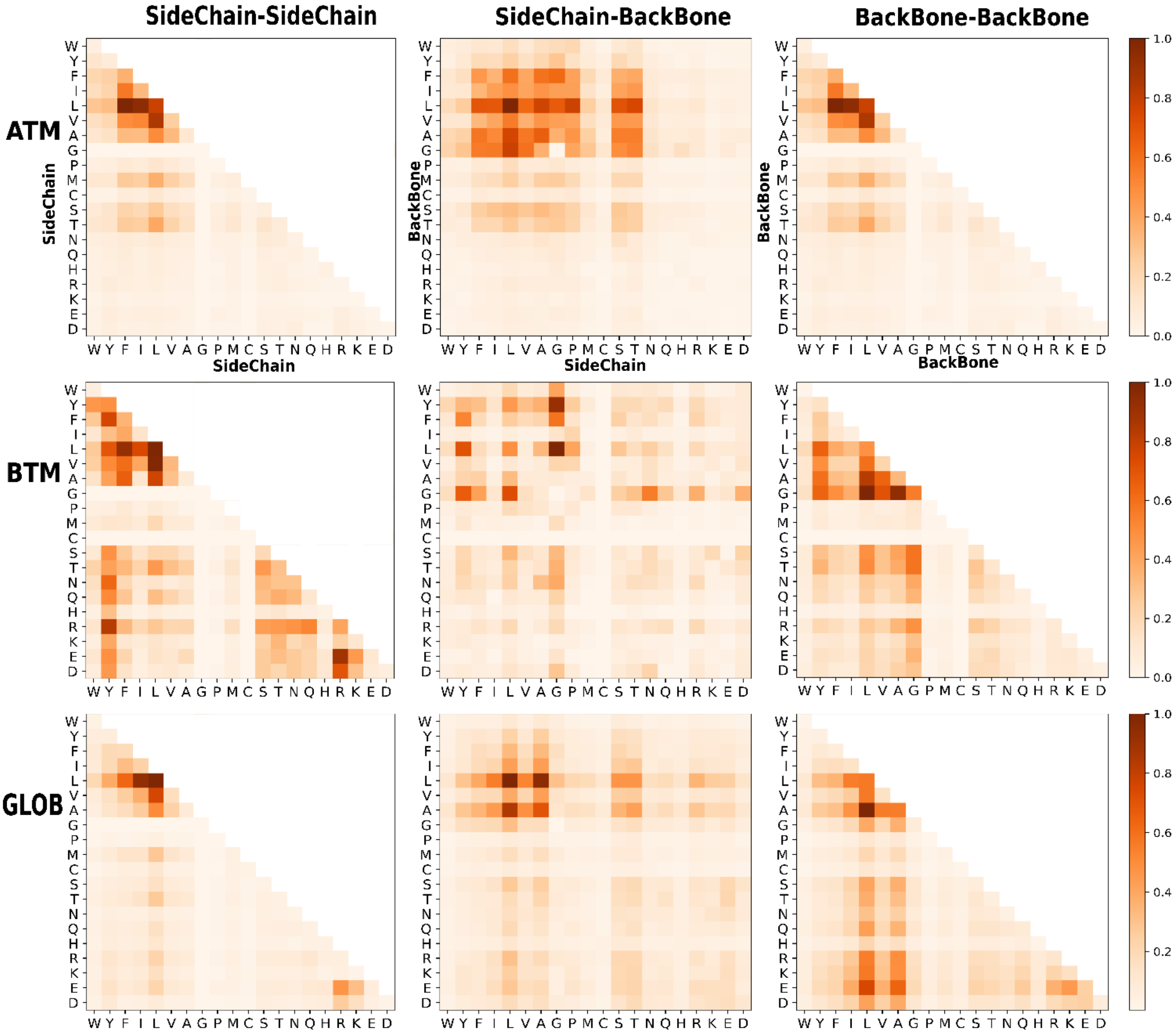
residue	ATM		BTM			GLOB		
Trp	6643	(7.1%)	1305	(7.1%)	*	7987	(4.6%)	
Tyr	8492	(9.1%)	3830	(20.9%)	*	17737	(10.2%)	*
Phe	18446	(19.7%)	2791	(15.2%)	*	22401	(12.8%)	*
Ile	17389	(18.6%)	1378	(7.5%)		28247	(16.2%)	*
Leu	26498	(28.3%)	3615	(19.7%)		53292	(30.5%)	*
Val	16734	(17.9%)	2067	(11.3%)	*	25293	(14.5%)	*
Ala	15699	(16.8%)	1823	(9.9%)	*	27422	(15.7%)	*
Gly	9133	(9.8%)	1430	(7.8%)	*	6243	(3.6%)	*
Pro	6919	(7.4%)	598	(3.3%)	*	5106	(2.9%)	*
Met	8467	(9.0%)	650	(3.5%)	*	10807	(6.2%)	*
Cys	2872	(3.1%)	7	(0.0%)	*	4838	(2.8%)	*
Ser	11013	(11.8%)	1979	(10.8%)	*	14614	(8.4%)	*
Thr	11541	(12.3%)	2272	(12.4%)		16302	(9.3%)	*
Asn	4328	(4.6%)	1989	(10.8%)	*	9156	(5.2%)	*
Gln	2863	(3.1%)	1763	(9.6%)	*	11193	(6.4%)	*
His	2496	(2.7%)	562	(3.1%)	*	7790	(4.5%)	*
Arg	2906	(3.1%)	2665	(14.5%)	*	18645	(10.7%)	*
Lys	1546	(1.7%)	1260	(6.9%)	*	13890	(8.0%)	*
Glu	2410	(2.6%)	1549	(8.4%)	*	19604	(11.2%)	*
Asp	2060	(2.2%)	1518	(8.3%)	*	11352	(6.5%)	*

group	ATM		BTM			GLOB		
Aromatic	31650	(33.8%)	7260	(39.6%)	*	45590	(26.1%)	*
Aliphatic	62300	(66.6%)	7617	(41.6%)	*	108128	(61.9%)	*
Gly-Pro	15597	(16.7%)	2015	(11.0%)	*	11208	(6.4%)	*
Sulfur	11190	(12.0%)	657	(3.5%)	*	15405	(8.8%)	*
Polar	27685	(29.6%)	7138	(38.9%)	*	47474	(27.2%)	*
Charged	10702	(11.4%)	6480	(35.4%)	*	59831	(34.3%)	*

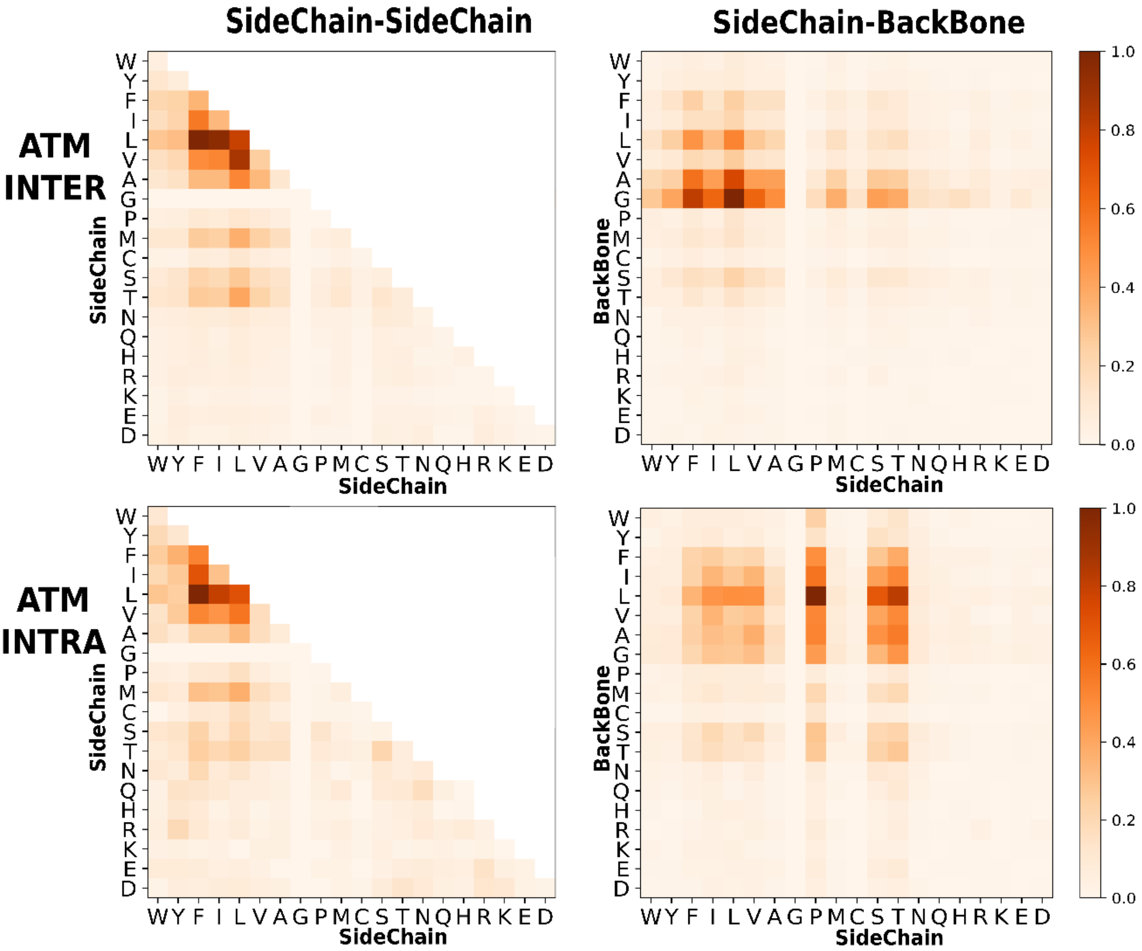
Supplementary Table 4.3-4. Absolute counts and percentages (in parentheses) of each residue (top) and groups of residues (bottom) of the number of interactions in the ATM, BTM and GLOB sets. Percentages are calculated dividing by the total number of interactions. * indicates statistical differences ($p<0.05$) between ATM and BTM or ATM and GLOB in a χ^2 goodness-of-fit test with Bonferroni corrections. Residue groups are defined in Figure 1. Backbone-backbone interactions are not considered.

	ATM			BTM			GLOB		
	TOTAL	IN	OUT	TOTAL	IN	OUT	TOTAL	IN	OUT
Trp	1846 (2.6%)	999 (2.5%)	847 (2.8%)	430 (2.4%)	77 (0.9%) *	353 (3.7%) *	2108 (1.2%) *	1611 (1.8%) *	497 (0.6%) *
Tyr	2414 (3.4%)	1576 (3.9%)	838 (2.8%)	1495 (8.5%) *	438 (5.4%) *	1057 (11.2%) *	5673 (3.3%)	4090 (4.6%) *	1583 (2.0%) *
Phe	6071 (8.6%)	3246 (8.0%)	2825 (9.3%)	985 (5.6%) *	275 (3.4%) *	710 (7.5%) *	6774 (4.0%) *	5377 (6.1%) *	1397 (1.7%) *
Ile	7653 (10.8%)	3273 (8.1%)	4380 (14.5%)	705 (4.0%) *	169 (2.1%) *	536 (5.7%) *	10087 (6.0%) *	8011 (9.0%) *	2076 (2.6%) *
Leu	11694 (16.5%)	4943 (12.2%)	6750 (22.3%)	1895 (10.8%) *	350 (4.3%) *	1545 (16.4%) *	20279 (12.0%) *	15404 (17.4%) *	4875 (6.0%) *
Val	7634 (10.8%)	3708 (9.1%)	3926 (13.0%)	1297 (7.4%) *	318 (3.9%) *	979 (10.4%) *	10478 (6.2%) *	7993 (9.0%)	2485 (3.1%) *
Ala	8141 (11.5%)	5258 (13.0%)	2883 (9.5%)	1569 (8.9%) *	564 (6.9%) *	1005 (10.7%) *	18321 (10.8%) *	11250 (12.7%)	7071 (8.8%) *
Gly	6241 (8.8%)	4748 (11.7%)	1493 (4.9%)	1974 (11.2%) *	878 (10.8%)	1096 (11.6%) *	6853 (4.0%) *	3808 (4.3%) *	3045 (3.8%) *
Pro	1806 (2.5%)	1071 (2.6%)	735 (2.4%)	250 (1.4%) *	60 (0.7%) *	190 (2.0%)	2973 (1.8%) *	979 (1.1%) *	1994 (2.5%)
Met	2612 (3.7%)	1791 (4.4%)	821 (2.7%)	254 (1.4%) *	130 (1.6%) *	124 (1.3%) *	3964 (2.3%) *	2858 (3.2%) *	1106 (1.4%) *
Cys	1039 (1.5%)	740 (1.8%)	299 (1.0%)	5 (0.0%) *	2 (0.0%) *	3 (0.0%) *	2207 (1.3%) *	1830 (2.1%) *	377 (0.5%) *
Ser	3755 (5.3%)	2752 (6.8%)	1003 (3.3%)	1198 (6.8%) *	903 (11.1%) *	295 (3.1%)	8933 (5.3%)	3822 (4.3%) *	5111 (6.3%) *
Thr	3728 (5.3%)	2477 (6.1%)	1251 (4.1%)	1138 (6.5%) *	723 (8.9%) *	415 (4.4%)	7859 (4.6%) *	3682 (4.2%) *	4177 (5.2%) *
Asn	1302 (1.8%)	945 (2.3%)	357 (1.2%)	822 (4.7%) *	575 (7.0%) *	247 (2.6%) *	6181 (3.6%) *	2125 (2.4%)	4056 (5.0%) *
Gln	902 (1.3%)	635 (1.6%)	267 (0.9%)	679 (3.9%) *	498 (6.1%) *	181 (1.9%) *	7858 (4.6%) *	2486 (2.8%) *	5372 (6.7%) *
His	819 (1.2%)	495 (1.2%)	324 (1.1%)	235 (1.3%)	89 (1.1%)	146 (1.6%) *	3831 (2.3%) *	1973 (2.2%) *	1858 (2.3%) *
Arg	1016 (1.4%)	555 (1.4%)	461 (1.5%)	881 (5.0%) *	678 (8.3%) *	203 (2.2%) *	10152 (6.0%) *	3614 (4.1%) *	6538 (8.1%) *
Lys	718 (1.0%)	320 (0.8%)	398 (1.3%)	565 (3.2%) *	430 (5.3%) *	135 (1.4%)	11308 (6.7%) *	2199 (2.5%) *	9109 (11.3%) *
Glu	799 (1.1%)	549 (1.4%)	250 (0.8%)	586 (3.3%) *	517 (6.3%) *	69 (0.7%)	14589 (8.6%) *	3246 (3.7%) *	11343 (14.0%) *
Asp	668 (0.9%)	488 (1.2%)	180 (0.6%)	621 (3.5%) *	492 (6.0%) *	129 (1.4%) *	8948 (5.3%) *	2274 (2.6%) *	6674 (8.3%) *

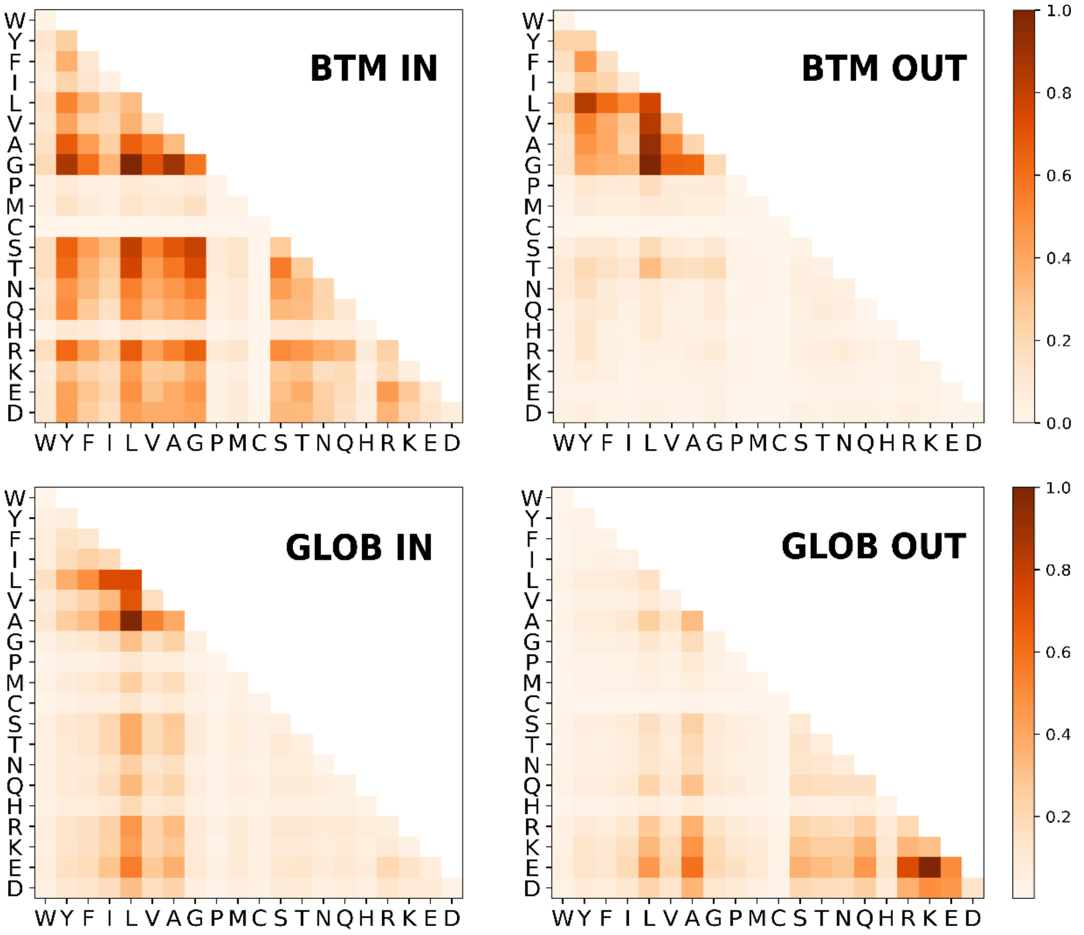
Supplementary Table 4.3-5. Absolute frequencies and percentages on total contributions (in parenthesis) of residues in the ATM, BTM and GLOB datasets for all residues (TOTAL), for those in the protein core (IN) or those facing the lipid-exposed protein surface (OUT). * indicates statistical differences in residue percentages of BTM and GLOB compared to ATM in a χ^2 goodness-of-fit test with Bonferroni corrections.



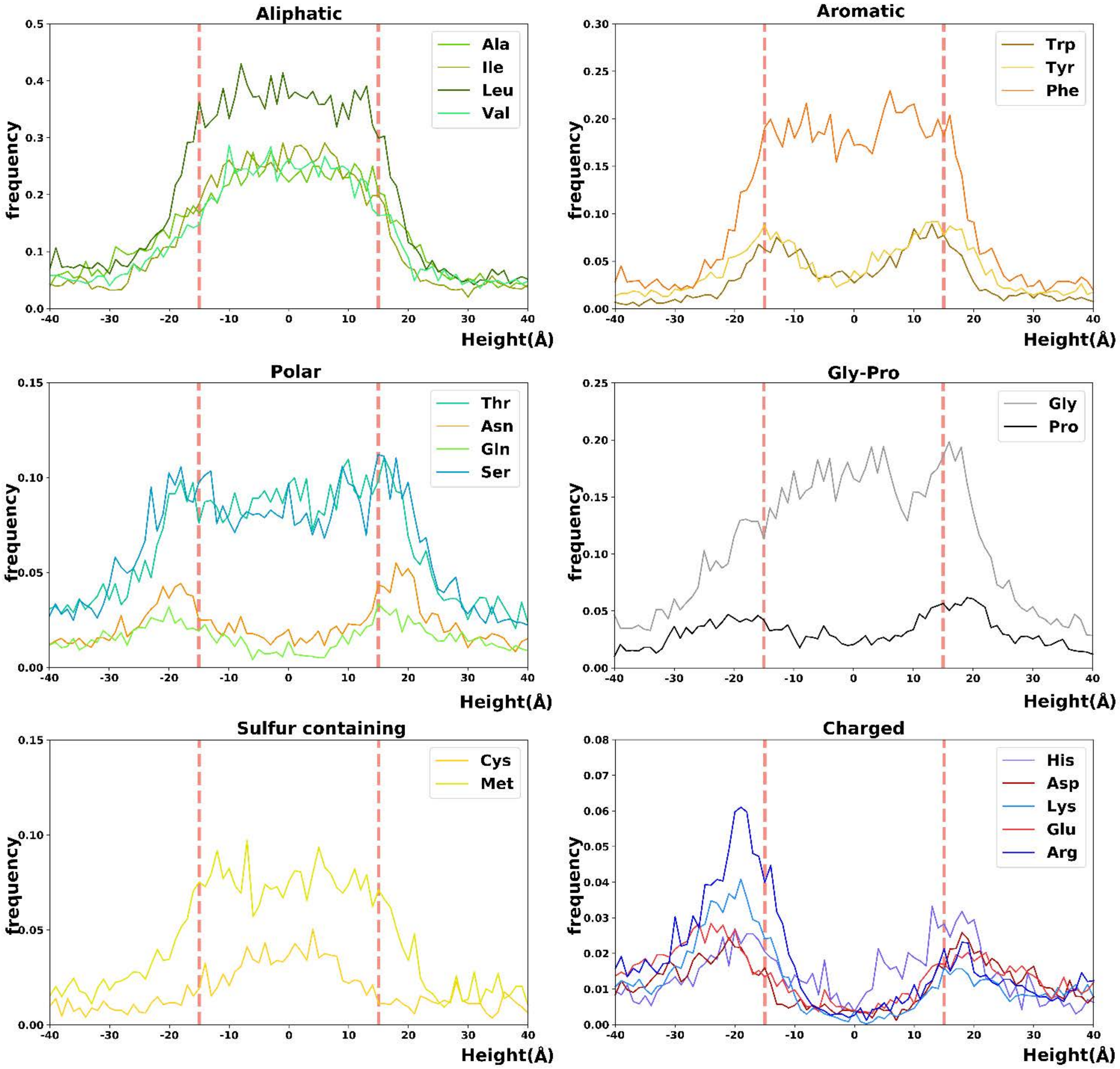
Supplementary Figure 4.3-1. Heatmaps of the inter-residue interaction frequencies of amino acids in the ATM, BTM and GLOB datasets separated by sidechain-sidechain, sidechain-backbone and backbone-backbone contributions normalized by dividing by the total number of interactions.



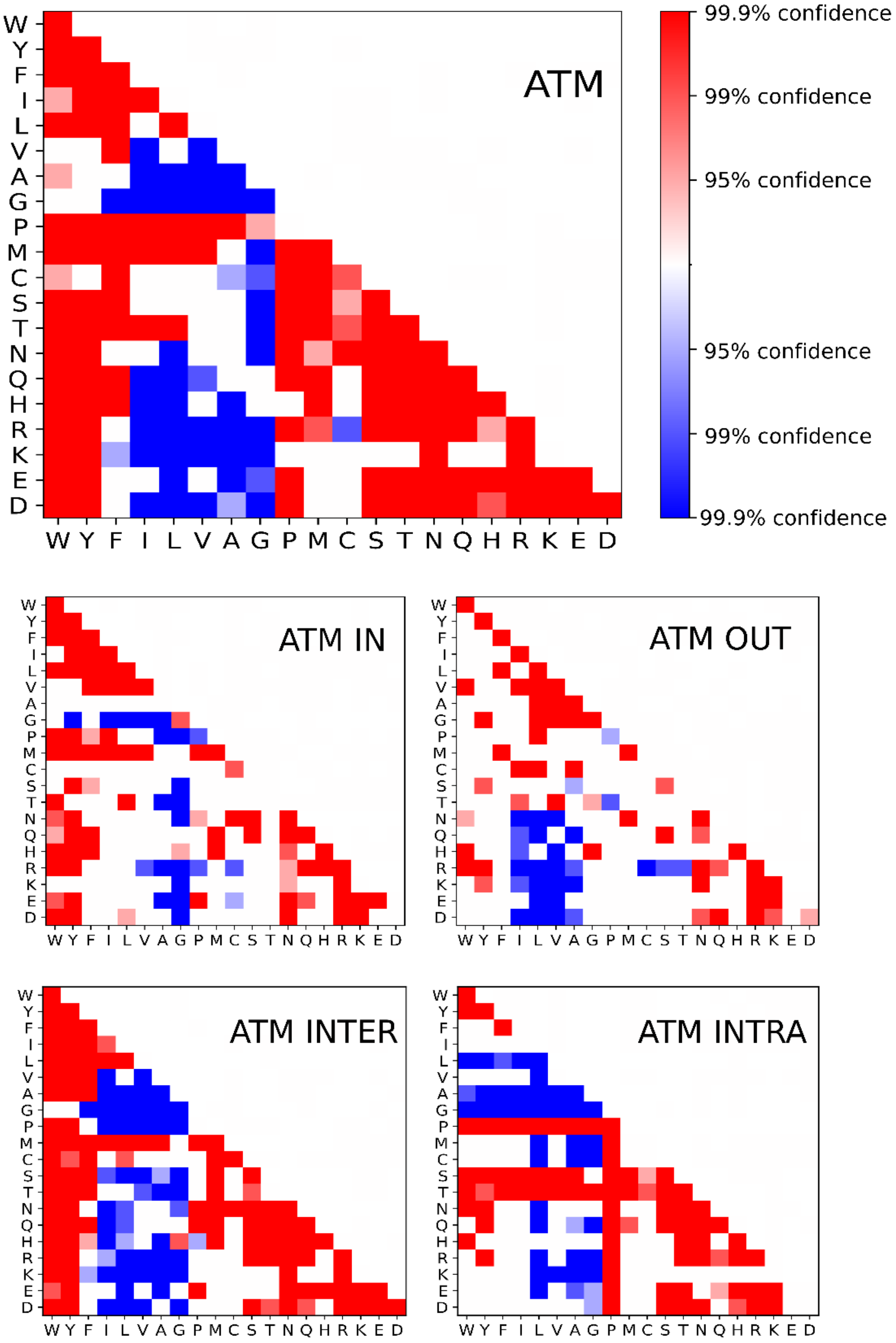
Supplementary Figure 4.3-2. Heatmaps of the inter-residue interaction frequencies of amino acids in the ATM inter-helical and ATM intra-helical datasets separated by sidechain-sidechain and sidechain-backbone normalized by dividing by the total number of interactions.



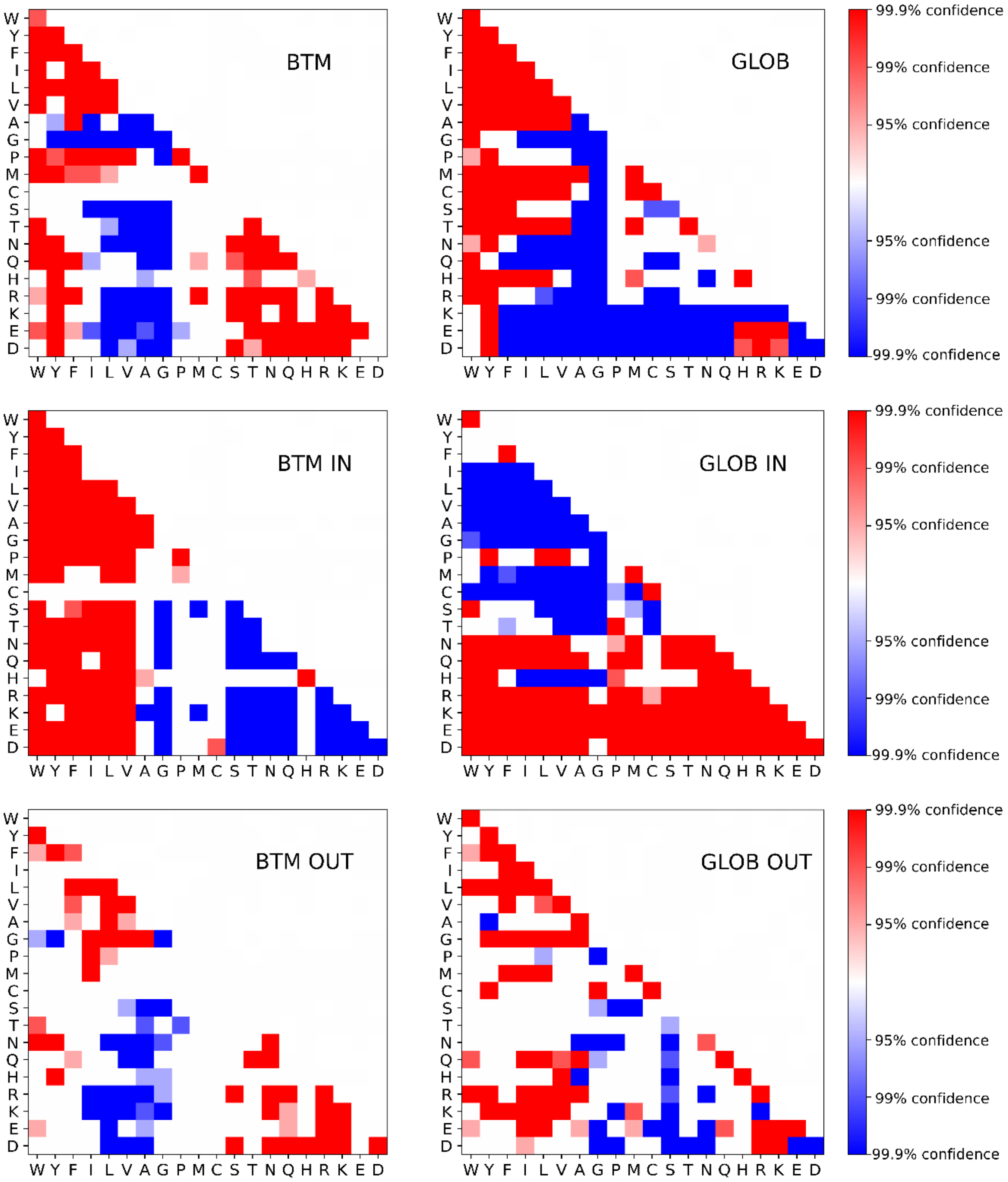
Supplementary Figure 4.3-3. Heatmaps of the inter-residue interaction frequencies of amino acids in the BTM and GLOB datasets separated by IN or OUT interactions normalized by dividing by the total number of interactions.



Supplementary Figure 4.3-4. Distribution of residues along the direction perpendicular to the membrane in the ATM dataset based on the localization predicted by Orientations of Proteins in Membranes (Lomize, *et al.*, 2012). Vertical red dashed lines indicate the approximate location of the lipid head-groups between -15 Å (intracellular) and +15 Å (extracellular).



Supplementary Figure 4.3-5. Heatmaps of the significance of inter-residue interactions comparing observed interactions to those expected from the residue frequencies and assuming random interactions for ATM (total set), ATM IN, ATM OUT, ATM Inter-helical and ATM intra-helical datasets. Under-represented and over-represented interactions are shown in blue and red, respectively, with the color intensity reflecting the significance level.



Supplementary Figure 4.3-6. Heatmaps of the significance of inter-residue interactions comparing observed interactions to those expected from the residue frequencies and assuming random interactions for BTM and GLOB (total set), BTM IN, BTM OUT, GLOB IN and GLOB OUT datasets. Under-represented and over-represented interactions are shown in blue and red, respectively, with the color intensity reflecting the significance level.

4.4. Interactions of sulfur containing amino acids

Abstract

The interactions of Met and Cys with other amino acid side chains have received little attention, in contrast to aromatic–aromatic, aromatic–aliphatic or/and aliphatic–aliphatic interactions. Precisely, these are the only amino acids that contain a sulfur atom, which is highly polarizable and, thus, likely to participate in strong Van der Waals interactions. Analysis of the interactions present in membrane protein crystal structures, together with the characterization of their strength in small-molecule model systems at the ab-initio level, predicts that Met–Met interactions are stronger than Met–Cys ≈ Met–Phe ≈ Cys–Phe interactions, stronger than Phe–Phe ≈ Phe–Leu interactions, stronger than the Met–Leu interaction, and stronger than Leu–Leu ≈ Cys–Leu interactions. These results show that sulfur-containing amino acids form stronger interactions than aromatic or aliphatic amino acids. Thus, these amino acids may provide additional driving forces for maintaining the 3D structure of membrane proteins and may provide functional specificity.

Introduction

Non-bonded interactions are crucial for protein stability, function and ligand binding. These comprise electrostatic (including hydrogen bonds) and van der Waals (dipole-dipole, dipole-induced dipole and induced dipole-induced dipole) interactions (Muller-Dethfel & Hobza 2000). All these types of interactions have been extensively characterized in terms of strength, directionality, and physicochemical properties (Meyer et al. 2003). However, their prevalence and importance vary depending on whether or not they occur in membrane or globular proteins due to their different environment. Both globular and membrane proteins position hydrophobic amino acid side chains toward the protein core and maximize hydrogen bond interactions among backbone atoms. However, in contrast to soluble proteins, the hydrophobic nature of the lipid bilayer imposes that residues pointing towards the membrane are also hydrophobic. Thus, dispersion forces (mainly aromatic-aromatic, aromatic-aliphatic or aliphatic-aliphatic) are involved in stabilizing the tertiary structure of the protein or in structural changes (Meyer et al. 2003, Nishio 2004, Brandl et al. 2001, Kim et al. 2011, Tsuzuki et al. 2006, Ringer et al. 2006).

As polarizabilities of the two interacting partners become larger, van der Waals forces become stronger. For example, the aromatic ring of aromatic amino acids has a quadrupole π system that is highly polarizable and provides strong aromatic-aromatic dispersion interactions. Thus, aromatic side chains importantly contribute to the folding and thermodynamic stability of proteins (Hong et al. 2007). Similarly, sulfur-containing amino acids are also highly polarizable, as sulfur has filled 3p and empty 3d orbitals and contain a permanent dipole (Groosfield & Wolf 2002). Surprisingly, non-bonded interactions (dipole-induced dipole or dispersion) involving sulfur-containing amino acids (Met and Cys) have received little attention (Valley et al. 2012, Pal & Chakrabarti 1998 and 2001) in contrast to interactions involving aromatic amino acids (Meyer et al. 2003). More than 30 years ago, Morgan and coworkers observed a high frequency of contacts between sulfur-containing residues and aromatic residues in proteins, and identified large stacked arrangements composed of aromatic and Met or Cys residues (Morgan et al. 1978). Further studies also demonstrated that Cys- and Met-aromatic interactions were fairly common in protein crystal structures (Pal & Chakrabarti 2001, Zauhar et al. 2000, Samanta et al. 2000).

In the present work we aim to evaluate the occurrence of interactions involving Met and Cys side-chains in crystal structures of membrane proteins and to characterize their strength in small-molecule model systems at the *ab-initio* level. The employed level of theory improves previous calculations in analogous systems (Pranata 1997, Cabaleiro-Lago et al. 2004, Duan et al. 2001, Ringer et al. 2007). Our results show that Met-Met, Met-Phe, Met-Leu and Cys-Phe interactions are stronger in magnitude than Phe-Phe interactions.

Material and Methods

Analysis of membrane protein structures

A non-redundant dataset of 327 α-helical transmembrane bundles were taken from TMalphaDB (Perea et al. 2015). This data set consists of crystallographic structures deposited in the Protein Data Bank (Berman et al. 2000) with resolution <3.5 Å.
Residues were classified, based on their circular variance (CV) (Mezei 2003) of vectors drawn from the Cα atom of a given residue to the Cα atoms of neighbor residues, as exposed (CV > 0.7) or buried (CV ≤ 0.7) to the membrane. Met-Met, Met-Phe, Met-Leu, Cys-Met, Cys-Phe, and Cys-Leu interactions were considered if the distance d between the two side-chains (measured as the distance between the atoms S_δ of Met, S_γ of Cys or C_γ of Leu or the centroid of the aromatic ring of Phe) was < 6Å. The relative orientation of the interacting side-chains was defined by the distance d, the angle P between side-chain planes (each plane defined by atoms C_γ, S_δ and C_ε of Met; C_α, C_β and S_γ of Cys; $C_{\delta1}$, C_γ and $C_{\delta2}$ of Leu; and the aromatic ring of Phe), and the angle θ between the plane defined by side-chain A and the vector connecting the central atoms of each side-chain (A and B). Definitions of P and θ angles were those used by Chakrabarti et al. to describe benzene dimer geometries (Chakrabarti & Bhattacharyya 2007). These interactions were clustered according to the conformational space defined by the distance d and the angles P and θ (see Supplementary Tables S1-S3 for a detailed description).

Quantum mechanical calculations

For the representative structure of each cluster, the energy of interaction between side chains was calculated using *ab initio* methods on small-molecule models systems: Met was mimicked by dimethyl sulfide (DMS), Cys by methanethiol (MT), Leu by propane (PRP), and Phe by benzene (BNZ). All chosen model structures were optimized at the MP2/6-31+G(d,p) level of theory, which has been shown to provide reasonably good geometries (Riley et al. 2012, Hobza et al. 1996). Next, single point energy calculations were performed at the CCSD(T)/6-311+G(3df,2p) level. In order to minimize the basis set superposition error, counterpoise method by Boys and Bernardi (Boys & Bernardi 1970) was utilized. Moreover, we also computed the interaction energies using the AMBER force field (Cornell et al. 1995) (see Supplementary Figures S4.4-1-S4.4-5) in order to evaluate the ability of protein force fields in reproducing these interactions. All calculations were performed using GAUSSIAN09 program (Gaussian 09, Frisch et al. 2009).

Results and Discussion

Structural bioinformatics analysis of the presence of Cys and Met in membrane proteins

Table 1 summarizes the occurrence of the most frequent amino acids, together with Cys and Met, in the transmembrane region (i.e. excluding water-soluble domains or loops) of α-helix bundles of membrane proteins with known crystal structure (see Methods). Amino acids such as Leu, Ile, Val and Phe are the most frequent at the membrane-embedded region, without preference for being localized in the protein core or in the membrane-exposed region. In contrast, sulfur-containing amino acids (i.e. Met and Cys) show lower frequencies and are mostly found buried in the core of the protein. This might indicate a functional role in stabilizing the 3D structure. Analysis of inter-residue interaction of Met and Cys reveals that a significant percentage of Met residues form interactions with aliphatic residues (Leu, Ala, Ile, Val) and Phe, while a smaller proportion interact with Met and Cys (Table 4.4-1). Thus, we aim to determine the orientation and strength of these interactions of sulfur-containing amino acids.

	Amino acid distribution			Side chain-side chain interactions						
	Total	**Buried**	**Exposed**	**Cys**	**Met**	**Phe**	**Val**	**Ile**	**Ala**	**Leu**
Leu	8,896 (17%)	3,787 (43%)	5,111 (57%)	474	1,129	2,415	3,427	2,49	3,764	2,319
Ala	6,198 (12%)	3,988 (64%)	2,209 (36%)	396	795	1,616	3,437	2,42	2,039	
Ile	5,801 (11%)	2,493 (43%)	3,309 (57%)	401	806	1,671	2,126	723		
Val	5,761 (11%)	2,826 (49%)	2,935 (51%)	357	700	1,276	1,363			
Phe	4,651 (9%)	2,464 (53%)	2,188 (47%)	276	707	859				
Met	1,933 (4%)	1,336 (69%)	597 (31%)	111	199					
Cys	768 (1.4%)	557 (73%)	211 (27%)	37						

Table 4.4-1. **Structural bioinformatics analysis of membrane proteins**. Amino acid type distribution (absolute frequencies and relative frequencies in percentage) observed in the survey of transmembrane domains of α-helix bundles (the five most frequent residues and Met and Cys) classified as buried or exposed to the membrane. The most significant side-side chain interactions of the five most frequent amino acids and Met and Cys.

The orientation and strength of Met-Phe interactions

The Met-Phe interactions identified in the crystal structures of membrane proteins (a total of 707 pairs, Table 1) were clustered based on relative distances and angles between the two amino acids (see Methods). Figure 4.4-1 shows the 2D histograms with the distribution of the Met-Phe interactions projected on the conformational space defined by P and θ angles (see Methods and Supplementary Table S4.4-1). In order to evaluate the magnitude of the energy of interaction between both side chains, we performed high level *ab initio* calculations (see Methods) in small-molecule model systems (Met and Phe were mimicked by dimethyl sulfide (DMS) and the benzene ring (BNZ), respectively, Supplementary Figure S4.4-1). Figure 4.4-2 shows that the interaction energy along the distance between DMS and BNZ exhibits a wide minimum located between 4 and 6 Å. Clusters **I** (containing 24% of the observed interactions) and **II** (37%) reproduce the most favorable arrangements of the side chains (-2.9 kcal/mol) as calculated in comparable model systems **1** and **2**. Arabic numbers depict optimized *ab initio* models whereas roman numbers represent clusters observed in crystal structures. The two planes defined by DMS and BNZ molecules are almost parallel in model **1** and perpendicular in **2**, but in both cases a methyl group of DMS is located on top of the negative charge density at the center of BNZ ring (π electrons) and the sulfur atom on top of the positive charged density at the exterior of BNZ ring (–CH groups). In Cluster **III** (12%) the CH atoms of Met are pointing to the aromatic ring of Phe and the sulfur atom is pointing toward opposite direction, which results in an interaction of -2.4 kcal/mol in model **3**. Finally, cluster **IV** (28%) accounts for interactions in which the sulfur atom of Met acts as hydrogen bond acceptor for a –CH group from the phenyl ring of Phe. The interaction energy was -2.0 kcal/mol in the comparable model **4**. Interestingly, these computed energies are of the same magnitude as the values experimentally obtained for peptides in water (Viguera & Serrano 1995).

In order to study the influence in the energy of interaction of the highly polarizable sulfur, compared to oxygen or the methylene group, we performed analogous *ab initio* calculations with model compounds that replace the sulfur atom (dimethyl sulfide, DMS, mimicking Met) by oxygen (dimethyl ether, DME) or –CH_2– (propane, PRP, mimicking aliphatic amino acids) (Supplementary Figure S4.4-1).

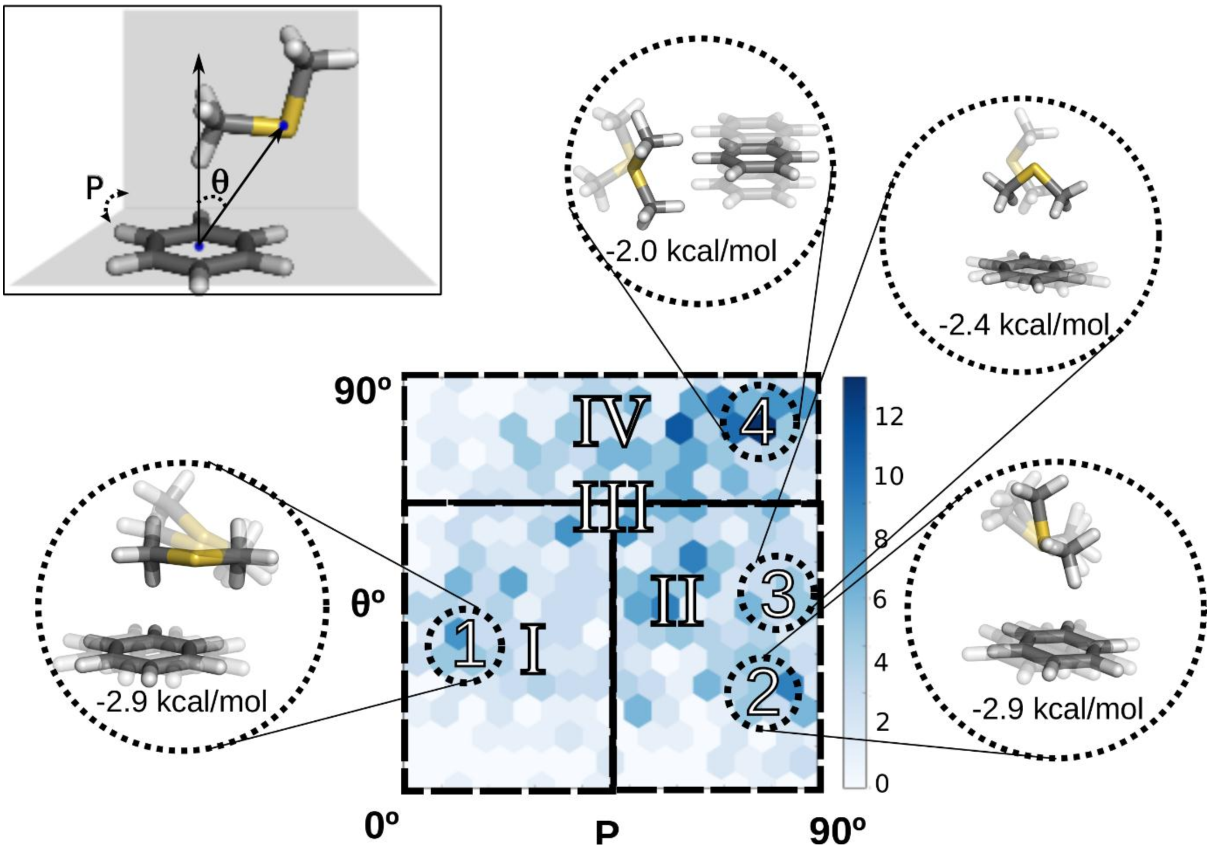


Figure 4.4-1. **The orientation and strength of Met-Phe interactions.** 2D histograms of the frequencies of occurrence of these interactions, clustered according to the conformational space defined by the distance d and the angles P and θ (see Supplementary Table S4.4-1 for a detailed description). Roman and arabic numbers indicate the position in the 2D histogram of the most representative structure in the cluster and the energy-minimized structure, respectively. Ab initio geometry optimization at the MP2/6-31+G(d,p) level and calculated energy of interaction at the CCSD(T)/6-311+G(3df,2p) level (see Methods) are shown inside dotted circles as solid sticks. Representative structures obtained in the cluster analysis are shown as transparent sticks.

Comparison of these energies of interactions of model compounds **1-4** in DMS-BNZ complexes with analogous conformations of DME-BNZ shows that in all cases the sulfur-containing molecule (DMS) interacts stronger with the aromatic ring (BNZ) than in the oxygen-containing one (DME) with the exception of model **4**. This suggests that the induced positive charge density on the methyl group, involved in the interaction with the π electrons of the ring in models **1-3**, is larger in the presence of the sulfur atom than in the presence of oxygen. Reasonably, because sulfur is a poorer hydrogen bond acceptor, in model **4** the sulfur atom forms weaker S···HC hydrogen bond interaction with the CH group of the ring than oxygen (O···HC). Importantly, the energies of interactions of model compounds **1-4** in DMS-BNZ complexes are always more stable than in PRP-BNZ complexes, indicating that the interaction of aromatic rings with sulfur-containing groups is always stronger than with aliphatic groups.

Because aromatic-aromatic interactions are considered key in the stability of membrane proteins (Hong et al. 2007), we next compared the energies of interaction in DMS-BNZ complexes with those in BNZ-BNZ complexes. Comparison with the well-characterized (Janowski & Puay 2007, DiStasio et al. 2007, Sinnokrot & Sherrill 2006, Tsuzuki et al. 2002) lowest energy configurations of BNZ-BNZ (Supplementary Figure S4.4-2), the T-shaped (-2.4 kcal/mol) and parallel displaced (-2.1 kcal/mol), indicates that Met forms more stable interactions with aromatics rings (DMS-BNZ) than aromatic-aromatic interactions (BNZ-BNZ).

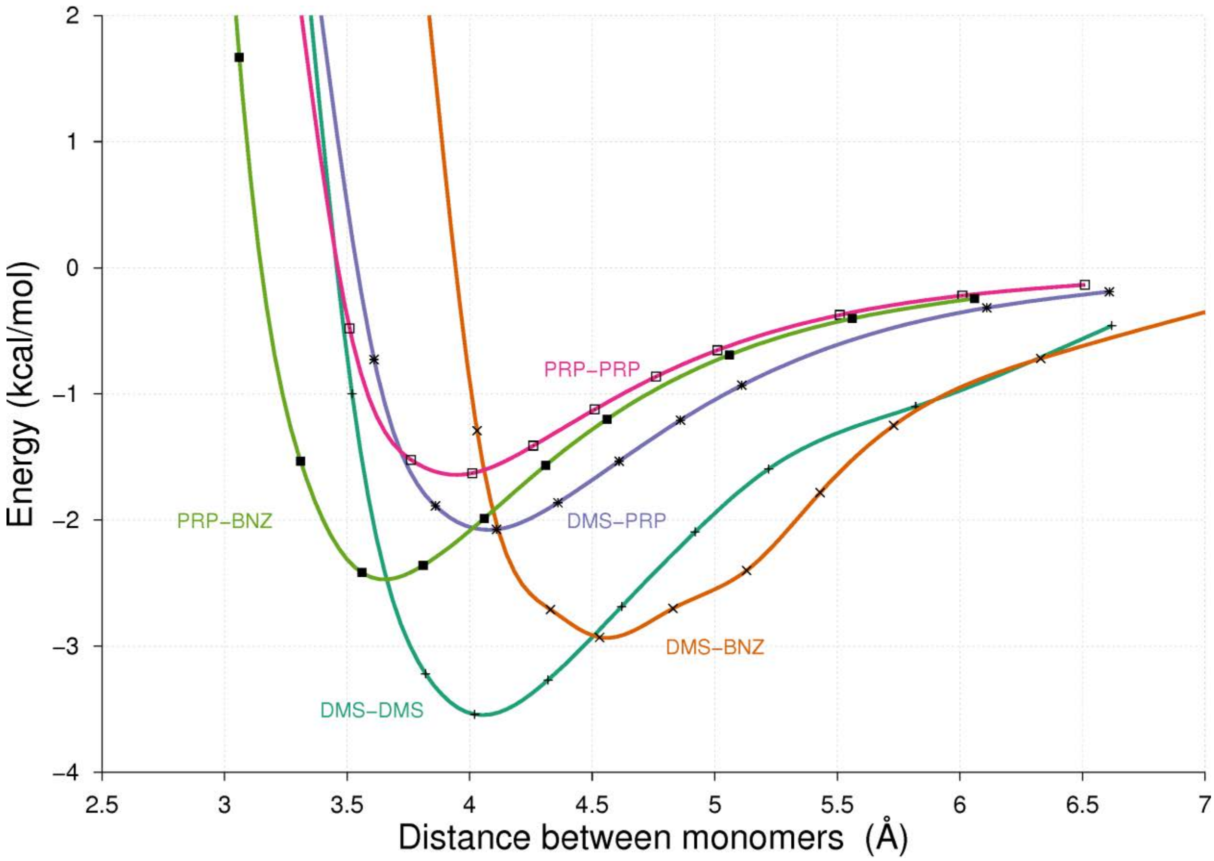


Figure 4.4-2. **Influence of the distance on the interaction energy.** Calculations were done on the DMS-DMS (Met-Met), DMS-BNZ (Met-Phe), DMS-PRP (Met-Leu), PRP-PRP (Leu-Leu), and PRP-BNZ (Leu-Phe) model systems with the lowest energy. d refers either to the sulfur-sulfur or the sulfur-benzene (centroid) distance.

The orientation and strength of Met-Met interactions

Clusters **I-V** in Figure 4.4-3, obtained from the 199 Met-Met interactions present in membrane proteins (Table 4.4-1), are calculated in a similar way to the clusters of Met-Phe (see above). Cluster **I**, containing 11% of the interactions, corresponds to an anti-parallel orientation, exhibiting the largest interaction energy (-3.5 kcal/mol) in the comparable model system **1**. The orientation of the Met side chains in cluster **II** (46%) is in the T-shaped configuration, being the interaction energy of -3.0 kcal/mol in the comparable model system **2**. In these configurations **1** and **2** each sulfur atom interacts respectively with four and three hydrogen atoms of the methyl groups (S···HC interactions) that have positive charge density. The Met side chains in cluster **III** (5%) are in a parallel-displaced orientation, in a head-to-head configuration with both sulfur atoms engaged in the interaction (-2.2 kcal/mol in model **3**). Clusters **IV** (15%) and **V** (25%) account for the least favored Met-Met interactions (-1.5 and -1.3 kcal/mol in models **4** and **5**, respectively). Models mimicking these clusters reproduce a structure with a single sulfur atom interacting with four and two CH hydrogen atoms (S···HC interactions), respectively. Cluster **IV** shows a parallel orientation in a head-to-tail configuration of the Met side-chains, while cluster **V** shows a T-shaped orientation in which the interactions occur through the methyl groups. The influence of the highly polarizable sulfur atom in these interactions was evaluated by performing analogous *ab initio* calculations with model compounds that replace the sulfur atom (Supplementary Figure S4.4-3) (dimethyl sulfide, DMS) by oxygen (dimethyl ether, DME). The anti-parallel orientation of model **1** (-3.5 vs. -2.7 kcal/mol) and the T-shaped configuration of model **2** (-3.0 vs. -2.8 kcal/mol) are more stable in the sulfur-containing DMS-DMS complex than in the oxygen-containing DME-DME complex. The opposite is observed for models **3** (-2.2 vs. -2.4 kcal/mol), **4** (-1.5 vs. -1.6 kcal/mol) and **5** (-1.3 vs. -1.5 kcal/mol).

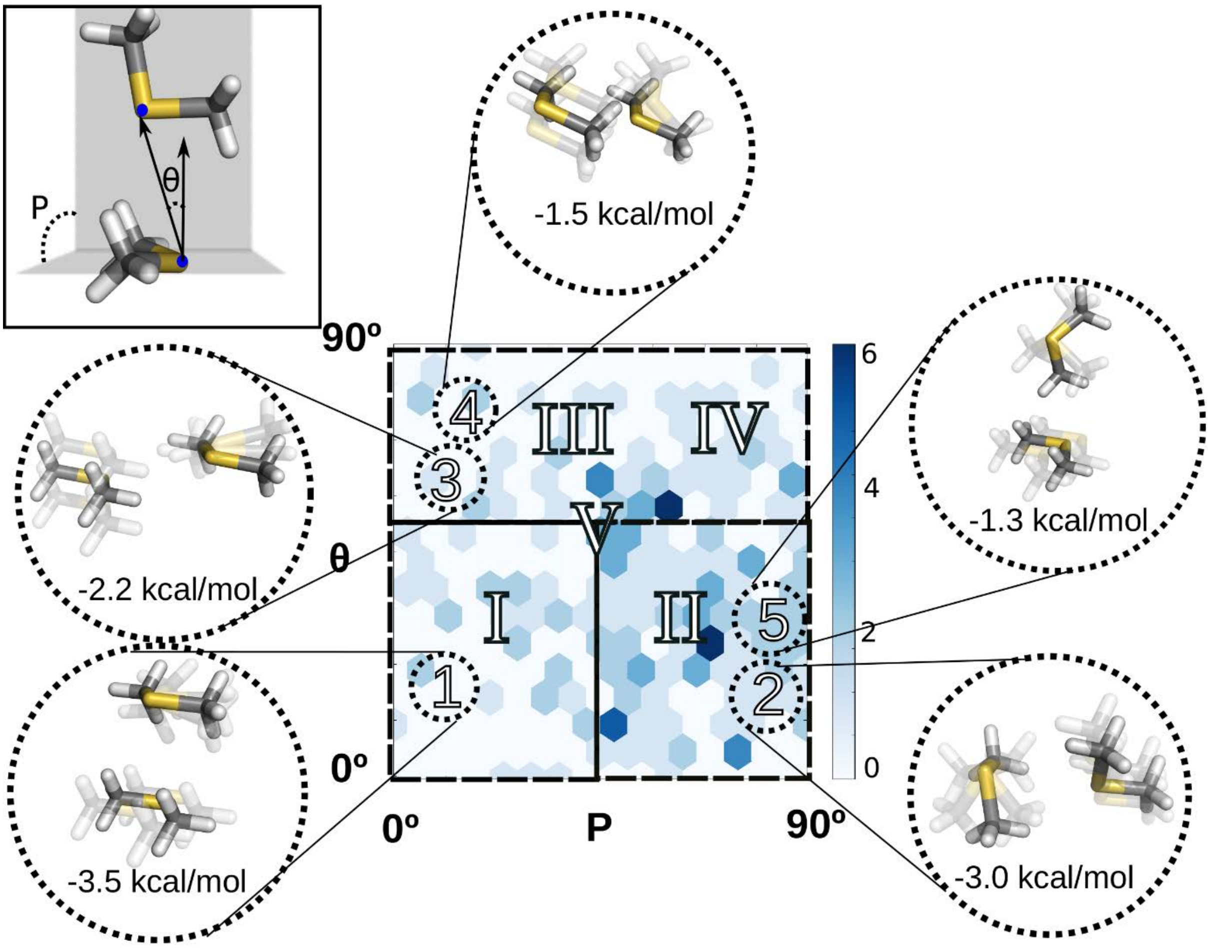


Figure 4.4-3. **The orientation and strength of Met-Met interactions.** 2D histograms of the frequencies of occurrence of these interactions, clustered according to the conformational space defined by the distance d and the angles P and θ (Supplementary Table S4.4-2). See Legend of Figure 1 for further details.

The orientation and strength of Met-Leu interactions

We have selected Leu as a representative residue to study the interactions of Met with aliphatic amino acids. The 1,129 Met-Leu interactions present in membrane proteins (Table 1) were clustered in a similar manner as in Met-Met interactions (see above). Because the C_γ, S_δ and C_ε atoms of Met are analogous to the $C_{\delta1}$, C_γ and $C_{\delta2}$ atoms of Leu, the relative orientation of Met-Leu residues in clusters **I-V** (Figure 4.4-4) were taken in analogy with clusters **I-V** of Met-Met (Figure 4.4-3). The computed interaction energies in comparable model systems **1-5** (Supplementary Figure S4.4-4) show that the anti-parallel orientation in model **1** exhibits the largest interaction energy (-2.1 kcal/mol). Comparison of the interaction energies in DMS-DMS (Met-Met clusters), DMS-PRP (Met-Leu clusters) and PRP-PRP (Leu-Leu clusters, not shown) models allow us to study the influence of the sulfur atom in the interaction energy. Clearly, the rank order of energies on interaction is DMS-DMS (2 sulfur atoms) < DMS-PRP (1 sulfur atom) < PRP-PRP (0 sulfur atom) in models **1** (-3.5 < -2.1 < -1.7 kcal/mol, respectively), **2** (-3.0 < -1.5 < -1.4 kcal/mol), **3** (-2.2 < -1.4 < -1.0 kcal/mol), **4** (-1.5 < -1.3 < -1.1 kcal/mol), and **5** (-1.3 < -1.2 < -1.0 kcal/mol).

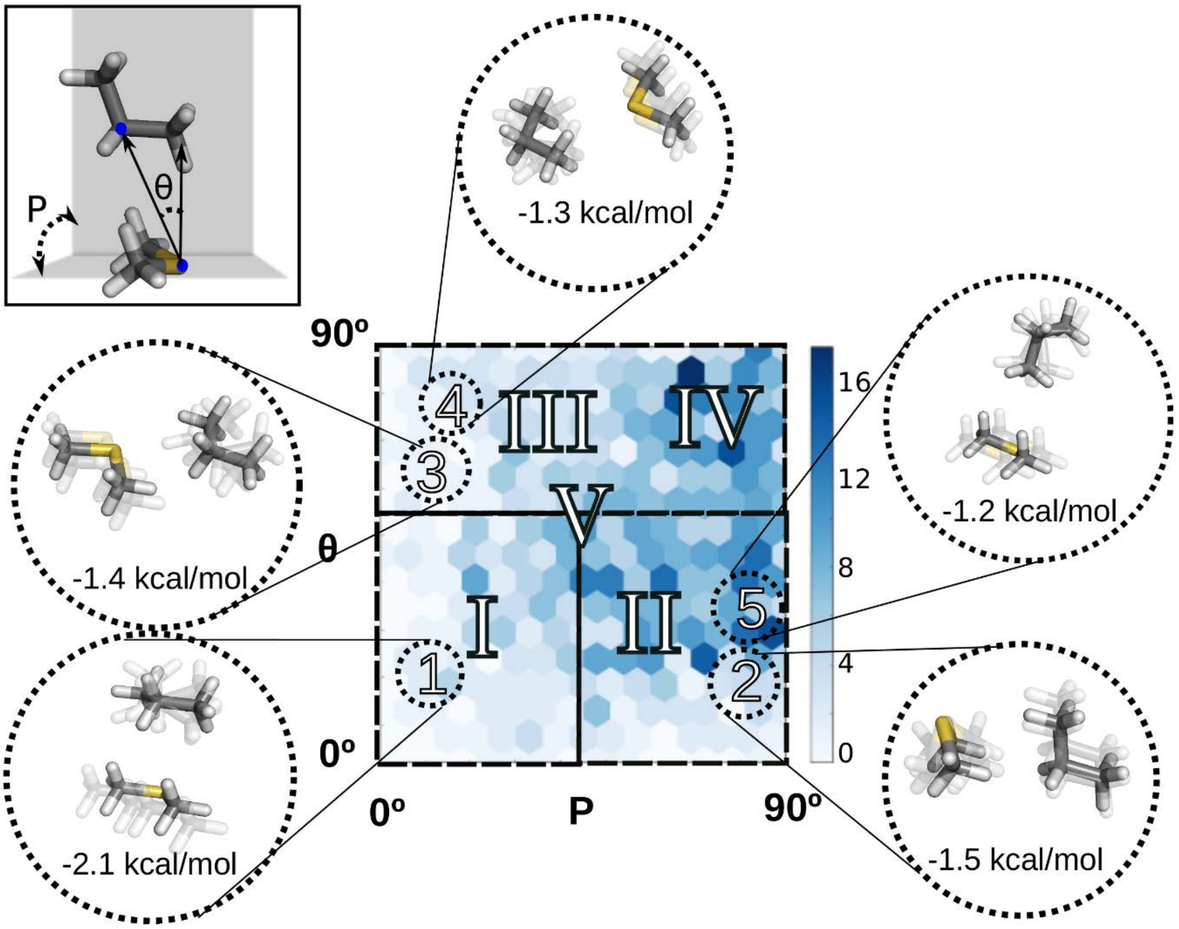


Figure 4.4-4. **The orientation and strength of Met-Leu interactions.** 2D histograms of the frequencies of occurrence of these interactions, clustered according to the conformational space defined by the distance d and the angles P and θ (Supplementary Table S4.4-3). See Legend of Figure 1 for further details.

The interactions of Cys

Cys interacts with Phe (a total of 276 pairs), Met (111 pairs) and hydrophobic amino acids such as Leu (474 pairs), Ala (396 pairs), Ile (401 pairs), and Val (357 pairs) (Table 4.4-1). In addition, Cys can interact with other Cys through a covalent disulfide bridge. Excluding disulfide bridges (S-S distances < 3Å) only 37 Cys-Cys pairs were observed in crystal structures (Table 4.4-1) in which Cys was acting as hydrogen bond donor and/or acceptor. These Cys-Cys interactions were not further analyzed, as they belong to the common hydrogen bond interaction. Analysis of the crystal structures of Cys-Phe interactions in membrane proteins revealed three main interaction modes (Figure 4.4-5). In cluster **I** (12%) the sulfur S_γ atom is located on top of the aromatic ring, in cluster **II** (49%) the C_β atom is located on top of the ring, and in cluster **III** (39%) the sulfur S_γ atom is coplanar to the phenyl ring. *Ab initio* energy optimizations of model compounds (Cys and Phe were mimicked by methanethiol (MT) and the benzene ring (BNZ), respectively, Supplementary Figure S4.4-5) positioned the sulfhydryl hydrogen absent in the crystal structures. In model **1** (-3.0 kcal/mol) MT forms a S-H··π hydrogen bond with the phenyl ring, whereas in model **2** (-2.9 kcal/mol) the S atom of MT is located on top of the positive charged density at the exterior of the BNZ ring and the methyl group on top of the negative charge density at the center of the ring. Importantly, *ab initio* energy minimization of model compound **3**, mimicking cluster **III**, led either to models **1** or **2**.

Clustering of the 111 Cys-Met and 474 Cys-Leu interactions (Table 4.4-1) was challenging due to the absence of the sulfhydryl hydrogen in the crystal structures. Thus, we performed a*b initio* energy optimizations of model compounds (MT-DMS or MT-PRP, Supplementary Figure S4.4-5) in which one of the methyl groups of DMS, in the reported Met-Met (DMS-DMS, Supplementary Figure S4.4-3) and Met-Leu (DMS-PRP, Supplementary Figure S4.4-4) interactions, was replaced by hydrogen. Comparison of these energies of interaction in MT-DMS complexes (-3.0, -2.9, -1.4, -1.7, -1.3 kcal/mol for models **1-5**, respectively) with analogous conformations of DMS-DMS (-3.5, -3.0, -2.2, -1.5, -1.3 kcal/mol for **1-5**, respectively) shows that DMS-DMS interactions are stronger with the exception of model **4**.

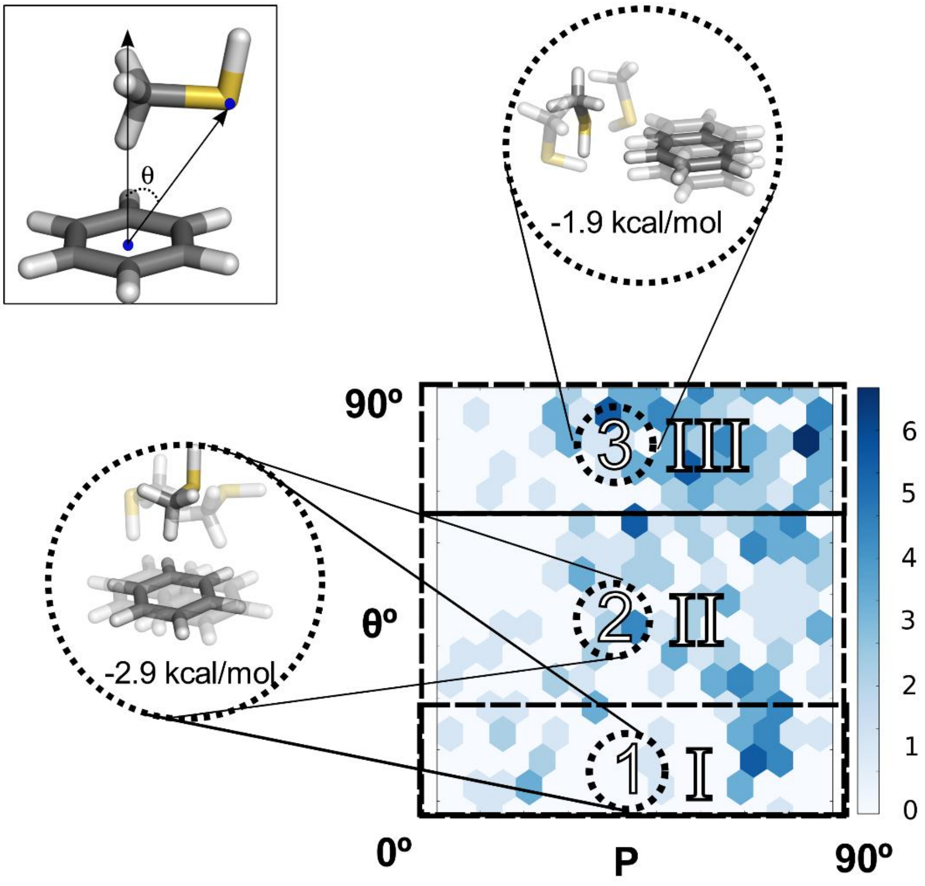


Figure 4.4-5. **The orientation and strength of Cys-Phe interactions.** 2D histograms of the frequencies of occurrence of these interactions, clustered according to the conformational space defined by the distance d and the θ angle. See Legend of Figure 1 for further details

Interaction energy comparison with the AMBER force field

We computed the interaction energies of the small-molecule model compounds using the AMBER force field (Cornell et al. 1995), with the aim of assessing the accuracy of this force field in reproducing the interactions of Met or Cys (see Methods). The results shown in Supplementary Figure S4.4-5 indicate a reasonable quantitative agreement in the interactions of Cys (mimicked by MT) with Met (DMS), Leu (PRP), and Phe (BNZ). The deviations are larger for Met (Supplementary Figures S4.4-1, S4.4-3, S4.4-4) with an average difference relative to CCSD(T) of ~0.5 kcal/mol. The larger deviations correspond to conformations in which sulphur atoms are in close proximity or when the sulphur atom is located on top of the benzene ring. Moreover, the rank order of Met-Phe interactions is not fully reproduced: CCSD(T) predicts 1 = 2 < 3 < 4 in DMS-BNZ models, while AMBER predicts 3 < 1 < 2 < 4 (Supplementary Figure S1). Similarly, CCSD(T) predicts 1 < 2 < 3 < 4 < 5 in DMS-DMS models, while AMBER predicts 1 < 2 < 4 < 3 = 5 (Supplementary Figure S4.4-3). In contrast, the interactions of Met with Leu are highly consistent, both in magnitude and rank order

(Supplementary Figure S4.4-4). Overall, these results are in line with a recent report on π-π, CH/π, and SH/π interactions (Sherrill et al. 2009).

Conclusions

Others and we have previously outlined the structural and functional role of Met-aromatic and Met-Met interactions in the family of G protein-coupled receptors (Cordomí et al. 2013, Nygaard et al. 2013, Magnan et al. 2013). In the present report we addressed a quantitative characterization of such interactions in membrane proteins. The analysis of the inter-residue interactions in crystal structures of membrane protein revealed that Met and Cys often interact with Leu, Ile, Val, Phe, and other Met or Cys. The characterization of their strength using *ab-initio* calculations in small-molecule model systems, predicted that Met-Met, Met-Phe, Cys-Phe, Met-aliphatic and Cys-aliphatic interactions are stronger in magnitude than aliphatic-aliphatic interactions. Remarkably, Met-Met, Met-Phe, and Cys-Phe interactions are stronger than aromatic-aromatic. Thus, we can conclude that these types of interactions, which have often been misled, need to be taken into account when considering the forces that stabilize the overall fold in membrane proteins. In addition to the stronger interactions of Met and Cys, their more flexible side-chains may provide extra versatility and adaptation to conformational changes. We believe that these interactions are also likely to be important in the interior of globular proteins or in the formation of protein-ligand or protein-protein complexes. However, further studies would be necessary in these regards.

References

Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. Nucleic Acids Res 2000, 28(1), 235-242.

Boys, S. F.; Bernardi, F. The calculation of small molecular interactions by the differences of separate total energies. Some procedures with reduced errors. Mol Phys 1970, 19(4), 553-566.

Brandl, M.; Weiss, M. S.; Jabs, A.; Suhnel, J.; Hilgenfeld, R. C-H...pi-interactions in proteins. J Mol Biol 2001, 307(1), 357-377.

Cabaleiro-Lago, E. M.; Hermida-Ramon, J. M.; Rodriguez-Otero, J. Computational study of the interaction in (CH3)(2)X dimer and trimer (X = O, S). J Phys Chem A 2004, 108(22), 4923-4929.

Chakrabarti, P.; Bhattacharyya, R. Geometry of nonbonded interactions involving planar groups in proteins. Prog Biophys Mol Biol 2007, 95(1-3), 83-137.

Cordomi, A.; Gomez-Tamayo, J. C.; Gigoux, V.; Fourmy, D. Sulfur-containing amino acids in 7TMRs: molecular gears for pharmacology and function. Trends Pharmacol Sci 2013, 34(6), 320-331.

Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. A 2nd Generation Force-Field for the Simulation of Proteins, Nucleic-Acids, and Organic-Molecules. J Am Chem Soc 1995, 117(19), 5179-5197.

DiStasio, R. A., Jr.; von Helden, G.; Steele, R. P.; Head-Gordon, M. On the T-shaped structures of the benzene dimer. Chem Phys Let 2007, 437(4-6), 277-283.

Duan, G. L.; Smith, V. H.; Weaver, D. F. Characterization of aromatic-thiol pi-type hydrogen bonding and phenylalanine-cysteine side chain interactions through ab initio calculations and protein database analyses. Mol Phys 2001, 99(19), 1689-1699.

Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, J. A.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, J. M.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J.: Wallingford CT, 2009.

Grossfield, A.; Woolf, T. B. Interaction of tryptophan analogs with POPC lipid bilayers investigated by molecular dynamics calculations. Langmuir 2002, 18(1), 198-210.

Hobza, P.; Selzle, H. L.; Schlag, E. W. Potential energy surface for the benzene dimer. Results of ab initio CCSD(T) calculations show two nearly isoenergetic structures: T-shaped and parallel-displaced. J Phys Chem 1996, 100(48), 18790-18794.

Hong, H.; Park, S.; Jimenez, R. H.; Rinehart, D.; Tamm, L. K. Role of aromatic side chains in the folding and thermodynamic stability of integral membrane proteins. J Am Chem Soc 2007, 129(26), 8320-8327.

Janowski, T.; Pulay, P. High accuracy benchmark calculations on the benzene dimer potential energy surface. Chem Phys Lett 2007, 447(1-3), 27-32.

Kim, K.; Karthikeyan, S.; Singh, J. How Different Are Aromatic π Interactions from Aliphatic π Interactions and Non-π Stacking Interactions? J Chem Theory Comput 2011, 7(11), 3471-3477

Magnan, R.; Escrieut, C.; Gigoux, V.; De, K.; Clerc, P.; Niu, F.; Azema, J.; Masri, B.; Cordomi, A.; Baltas, M.; Tikhonova, I. G.; Fourmy, D. Distinct CCK-2 receptor conformations associated with beta-arrestin-2 recruitment or phospholipase-C activation revealed by a biased antagonist. J Am Chem Soc 2013, 135(7), 2560-2573.

Meyer, E. A.; Castellano, R. K.; Diederich, F. Interactions with aromatic rings in chemical and biological recognition. Angewandte Chemie 2003, 42(11), 1210-1250.

Mezei, M. A new method for mapping macromolecular topography. J Mol Graph Model 2003, 21(5), 463-472.

Morgan, R. S.; Tatsch, C. E.; Gushard, R. H.; McAdon, J.; Warme, P. K. Chains of alternating sulfur and pi-bonded atoms in eight small proteins. Int J Pept Protein Res 1978, 11(3), 209-217.

Muller-Dethlefs, K.; Hobza, P. Noncovalent interactions: A challenge for experiment and theory. Chem Rev 2000, 100(1), 143-167.

Nishio, M. CH/pi hydrogen bonds in crystals. CrystEngComm 2004, 6, 130-158.

Nygaard, R.; Zou, Y.; Dror, R. O.; Mildorf, T. J.; Arlow, D. H.; Manglik, A.; Pan, A. C.; Liu, C. W.; Fung, J. J.; Bokoch, M. P.; Thian, F. S.; Kobilka, T. S.; Shaw, D. E.; Mueller, L.; Prosser, R. S.; Kobilka, B. K. The Dynamic Process of beta(2)-Adrenergic Receptor Activation. Cell 2013, 152(3), 532-542.

Pal, D.; Chakrabarti, P. Different types of interactions involving cysteine sulfhydryl group in proteins. J Biomol Struct Dyn 1998, 15(6), 1059-1072.

Pal, D.; Chakrabarti, P. Non-hydrogen bond interactions involving the methionine sulfur atom. J Biomol Struct Dyn 2001, 19(1), 115-128.

Perea, M.; Lugtenburg, I.; Mayol, E.; Cordomi, A.; Deupi, X.; Pardo, L.; Olivella, M. TMalphaDB and TMbetaDB: web servers to study the structural role of sequence motifs in alpha-helix and beta-barrel domains of membrane proteins. BMC Bioinformatics 2015, 16, 266.

Pranata, J. Sulfur aromatic interactions: A computational study of the dimethyl sulfide benzene complex. Bioorg Chem 1997, 25(4), 213-219.

Riley, K. E.; Platts, J. A.; Rezac, J.; Hobza, P.; Hill, J. G. Assessment of the performance of MP2 and MP2 variants for the treatment of noncovalent interactions. J Phys Chem A 2012, 116(16), 4159-4169.

Ringer, A. L.; Figgs, M. S.; Sinnokrot, M. O.; Sherrill, C. D. Aliphatic C-H/pi interactions: Methane-benzene, methane-phenol, and methane-indole complexes. J Phys Chem A 2006, 110(37), 10822-10828.

Ringer, A. L.; Senenko, A.; Sherrill, C. D. Models of S/pi interactions in protein structures: comparison of the H2S benzene complex with PDB data. Protein Sci 2007, 16(10), 2216-2223.

Samanta, U.; Pal, D.; Chakrabarti, P. Environment of tryptophan side chains in proteins. Proteins 2000, 38(3), 288-300.

Sherrill, C. D.; Sumpter, B. G.; Sinnokrot, M. O.; Marshall, M. S.; Hohenstein, E. G.; Walker, R. C.; Gould, I. R. Assessment of standard force field models against high-quality ab initio potential curves for prototypes of pi-pi, CH/pi, and SH/pi interactions. J Comput Chem 2009, 30(14), 2187-2193.

Sinnokrot, M. O.; Sherrill, C. D. High-accuracy quantum mechanical studies of pi-pi interactions in benzene dimers. J Phys Chem A 2006, 110(37), 10656-10668.

Tsuzuki, S.; Honda, K.; Uchimaru, T.; Mikami, M.; Tanabe, K. Origin of attraction and directionality of the pi/pi interaction: model chemistry calculations of benzene dimer interaction. J Am Chem Soc 2002, 124(1), 104-112.

Tsuzuki, S.; Honda, K.; Uchimaru, T.; Mikami, M. Estimated MP2 and CCSD(T) interaction energies of n-alkane dimers at the basis set limit: comparison of the methods of Helgaker et al. and Feller. J Chem Phys 2006, 124(11), 114304.

Valley, C. C.; Cembran, A.; Perlmutter, J. D.; Lewis, A. K.; Labello, N. P.; Gao, J.; Sachs, J. N. The methionine-aromatic motif plays a unique role in stabilizing protein structure. J Biol Chem 2012, 287(42), 34979-34991.

Viguera, A. R.; Serrano, L. Side-chain interactions between sulfur-containing amino acids and phenylalanine in alpha-helices. Biochemistry 1995, 34(27), 8771-8779.

Zauhar, R. J.; Colbert, C. L.; Morgan, R. S.; Welsh, W. J. Evidence for a strong sulfur-aromatic interaction derived from crystallographic data. Biopolymers 2000, 53(3), 233-248.

Supplementary Information

Table S4.4-1. Cluster analysis of Met-Phe interactions in crystal structures. Number (and percentage) of Met-Phe interactions and values of P (angle between the planes defined by C_γ, S_δ and C_ε atoms of Met and the aromatic ring of Phe) and θ (angle between the normal vector of the plane defined by the aromatic ring of Phe and the vector connecting the centroid R of the aromatic ring of Phe and S_δ of Met), and distance criteria (R accounts for the centroid of the aromatic ring of Phe, and Sδ, Cγ and Cε represent the atoms of the Met side-chain), in clusters I-IV (see Fig 4.4-1).

Cluster	Number	P	θ	Distance criteria
I	191 (24%)	0-45°	0-70°	d(R-Sδ) < d(R-Cγ) v d(R-Sδ) < d(R-Cε)
II	131 (37%)	45-90°	0-70°	d(R-Sδ) < d(R-Cγ) v d(R-Sδ) < d(R-Cε)
III	181 (12%)	0-90°	0-90°	d(R-Sδ) > d(R-Cγ) ^ d(R-Sδ) > d(R- Cε)
IV	204 (28%)	0-90°	70-90°	-

Table S4.4-2. Cluster analysis of Met-Met interactions in crystal structures. Number (and percentage) of Met-Met interactions and values of P (angle between the planes defined by C_γ, S_δ and C_ε atoms of Met) and θ (angle between the normal vector of the plane defined by the C_γ, S_δ and C_ε atoms of Met and the vector connecting the S_δ atoms of Met), and distance criteria (subindexes A and B refer to atoms in distinct side-chains) in clusters I-V (see Fig 4.4-2).

Cluster	Number	P	θ	Distance criteria
I	26 (11%)	0-45°	0-60°	[d($S\delta_A$-$S\delta_B$) < d($S\delta_A$-$C\gamma_B$) v d($S\delta_A$-$S\delta_B$) < d($S\delta_A$-$C\varepsilon_B$)] ^ [d($S\delta_A$-$S\delta_{B)}$ < d($S\delta_{B-}C\gamma_A$) v d($S\delta_A$-$S\delta_B$) < d($S\delta_B$-$C\varepsilon_A$)]
II	71 (46%)	45°- 90°	0-60°	[d($S\delta_A$-$S\delta_B$) < d($S\delta_A$-$C\gamma_B$) v d($S\delta_A$-$S\delta_B$) < d($S\delta_A$-$C\varepsilon_B$)] ^ [d($S\delta_A$-$S\delta_{B)}$ < d($S\delta_{B-}C\gamma_A$) v d($S\delta_A$-$S\delta_B$) < d($S\delta_B$-$C\varepsilon_A$)]
III	12 (5%)	0-90°	60-90°	[d($S\delta_A$-$S\delta_B$) < d($S\delta_A$-$C\gamma_B$) ^ d($S\delta_A$-$S\delta_B$) < d($S\delta_A$-$C\varepsilon_B$)] ^ [d($S\delta_A$-$S\delta_{B)}$ < d($S\delta_{B-}C\gamma_A$) ^ d($S\delta_A$-$S\delta_B$) < d($S\delta_B$-$C\varepsilon_A$)]
IV	43 (15%)	0-90°	60-90°	[d($S\delta_A$-$S\delta_B$) > d($S\delta_A$-$C\gamma_B$) v d($S\delta_A$-$S\delta_B$) > d($S\delta_A$-$C\varepsilon_B$)] ^ [d($S\delta_A$-$S\delta_{B)}$ < d($S\delta_{B-}C\gamma_A$) v d($S\delta_A$-$S\delta_B$) < d($S\delta_B$-$C\varepsilon_A$)]
V	47 (25%)	0-90°	0-90°	[d($S\delta_A$-$S\delta_B$) > d($S\delta_A$-$C\gamma_B$) ^ d($S\delta_A$-$S\delta_B$) > d($S\delta_A$-$C\varepsilon_B$)] ^ [d($S\delta_A$-$S\delta_{B)}$ > d($S\delta_{B-}C\gamma_A$) v d($S\delta_A$-$S\delta_B$) > d($S\delta_B$-$C\varepsilon_A$)]

Table S4.4-3. Cluster analysis of Met-Leu interactions in crystal structures. Number (and percentage) of Met-Leu interactions and values of P (angle between the planes defined by the C_γ, S_δ and C_ε atoms of Met and the $C_{\delta1}$, C_γ and $C_{\delta2}$ atoms of Leu), θ (angle between the normal vector of the plane defined by the C_γ, S_δ and C_ε atoms of Met and the vector connecting the S_δ atom of Met and the C_γ atom of Leu) and distance criteria (subindexes A and B refer to atoms in distinct side-chains) in clusters I-V (see Fig 4.4-3).

Cluster	Number	P	θ	Distance criteria
I	226 (20%)	0-45º	0-60º	$[d(S\delta_A - C\gamma_B) < d(S\delta_A - C\varepsilon1_B) \vee d(S\delta_A - C\gamma_B) < d(S\delta_A - C\varepsilon2_B)] \wedge [d(S\delta_A - C\gamma_B) < d(C\gamma_B - C\gamma_A) \vee d(S\delta_A - C\gamma_B) < d(C\gamma_B - C\varepsilon_A)]$
II	304 (27%)	45º-90º	0-60º	$[d(S\delta_A - C\gamma_B) < d(S\delta_A - C\varepsilon1_B) \vee d(S\delta_A - C\gamma_B) < d(S\delta_A - C\varepsilon2_B)] \wedge [d(S\delta_A - C\gamma_B) < d(C\gamma_B - C\gamma_A) \vee d(S\delta_A - C\gamma_B) < d(C\gamma_B - C\varepsilon_A)]$
III	34 (3%)	0-90º	60-90º	$[d(S\delta_A - C\gamma_B) < d(S\delta_A - C\varepsilon1_B) \wedge d(S\delta_A - C\gamma_B) < d(S\delta_A - C\varepsilon2_B)] \wedge [d(S\delta_A - C\gamma_B) < d(C\gamma_B - C\gamma_A) \vee d(S\delta_A - C\gamma_B) < d(C\gamma_B - C\varepsilon_A)]$
IV	143 (13%)	0-90º	60-90º	$[d(S\delta_A - C\gamma_B) > d(S\delta_A - C\varepsilon1_B) \vee d(S\delta_A - C\gamma_B) > d(S\delta_A - C\varepsilon2_B)] \wedge [d(S\delta_A - C\gamma_B) < d(C\gamma_B - C\gamma_A) \vee d(S\delta_A - C\gamma_B) < d(C\gamma_B - C\varepsilon_A)]$
V	422 (37%)	0-90º	0-90º	$[d(S\delta_A - C\gamma_B) > d(S\delta_A - C\varepsilon1_B) \wedge d(S\delta_A - C\gamma_B) > d(S\delta_A - C\varepsilon2_B)] \wedge [d(S\delta_A - C\gamma_B) > d(C\gamma_B - C\gamma_A) \vee d(S\delta_A - C\gamma_B) > d(C\gamma_B - C\varepsilon_A)]$

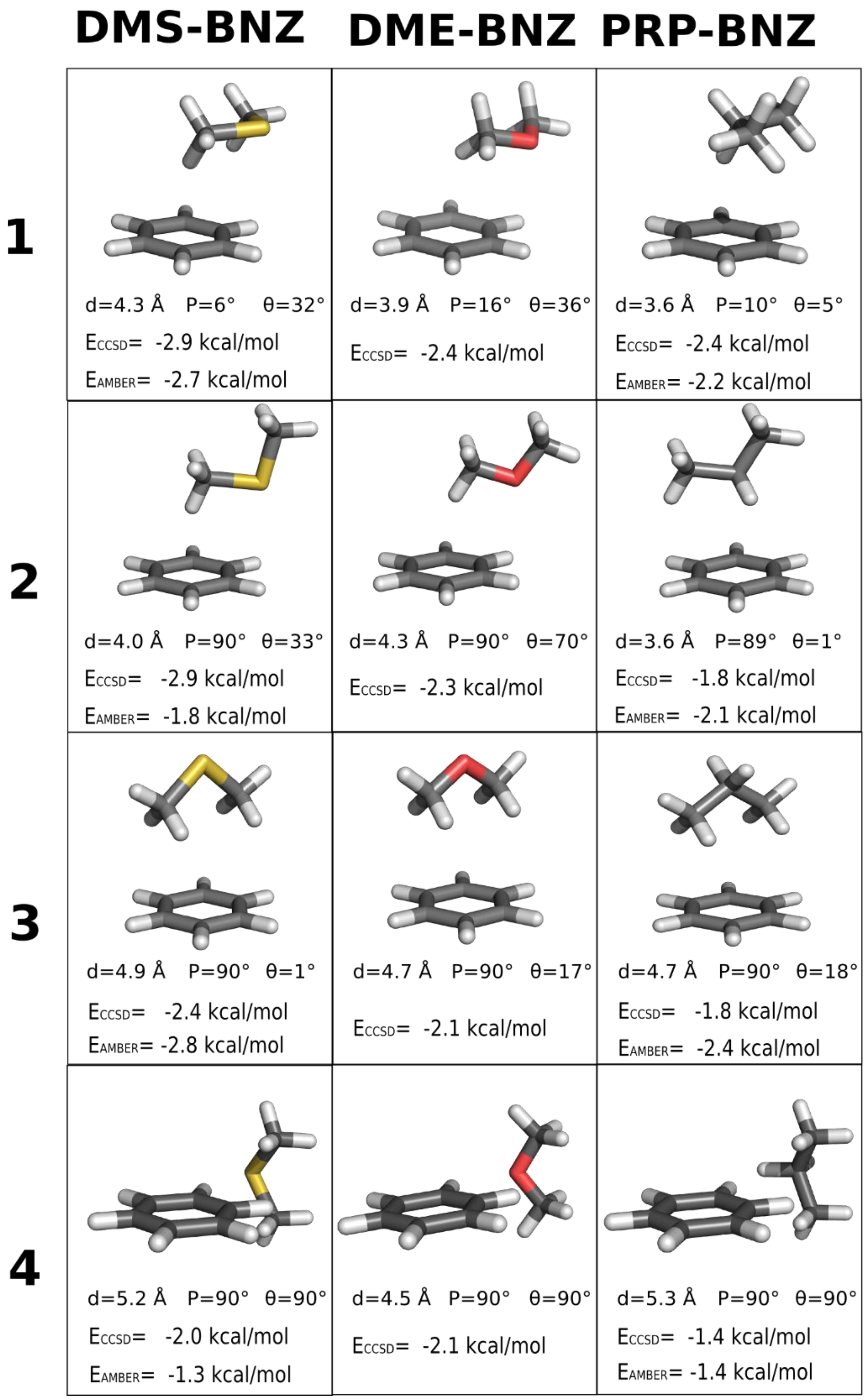


Figure S4.4-1. Small-molecule models systems mimicking Met-Phe interactions. Geometry optimized, at the *ab-initio* MP2/6-31+G(d,p) level of theory, of the interactions between benzene (BNZ, mimicking Phe) and dimethyl sulfide (DMS, mimicking Met), dimethyl ether (DME), and propane (PRP, mimicking Leu). Each optimized structure is designated by an arabic number that corresponds to a roman number of the obtained clusters in crystal structures (see Fig 1). The values of d, P, and θ (see Suppl. Table 1 and Fig 1 for definition), and single point energy calculations at the *ab-initio* CCSD(T)/6-311+G(3df,2p) level of theory (E_{CCSD}) and by molecular mechanics using the AMBER99 forcefield (E_{AMBER}) are shown.

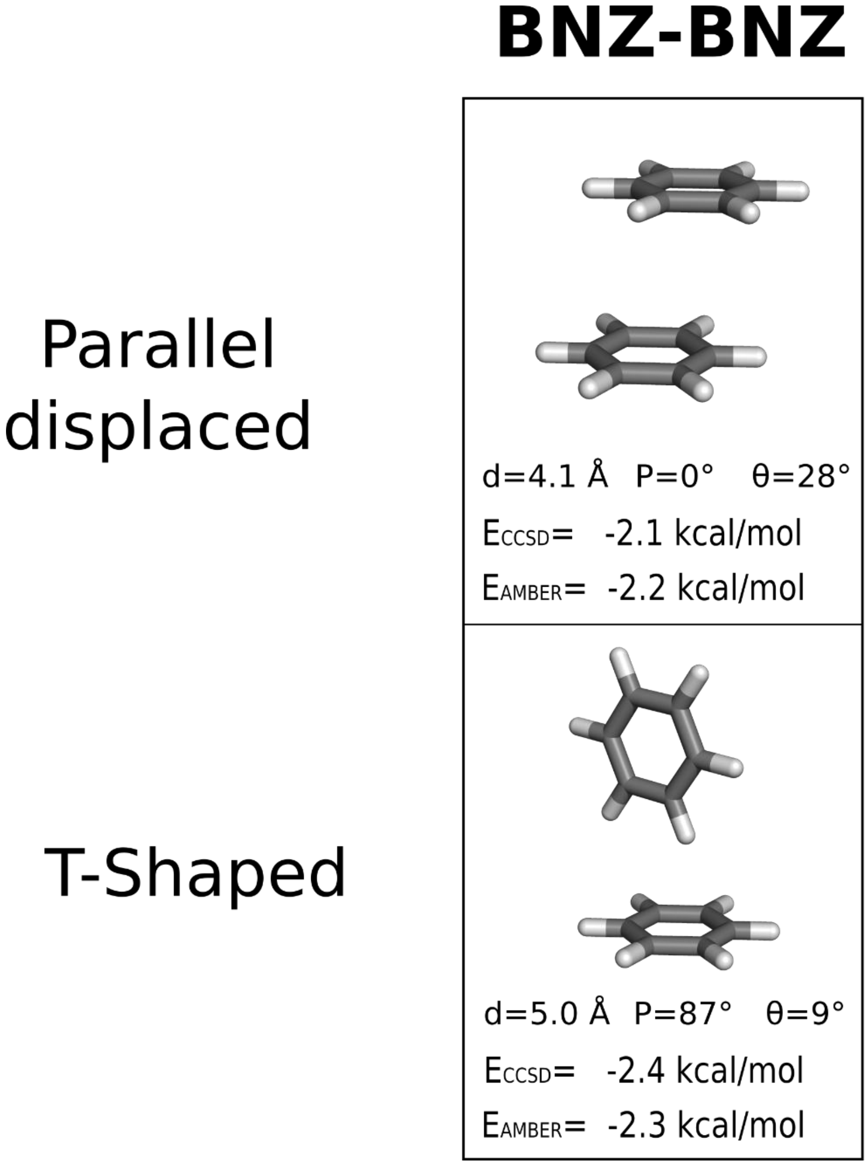


Figure S4.4-2. **Small-molecule models systems mimicking Phe-Phe interactions.** Geometry optimized models, at the *ab-initio* MP2/6-31+G(d,p) level of theory, of benzene-benzene (BNZ, mimicking Phe) interactions in the lowest parallel displaced and T-shaped energy configurations. The values of d (calculated as the distance between the centroid R of the aromatic ring of BNZ), P (calculated as the angle between the planes defined by the aromatic rings of BNZ), and θ (angle between the normal vector of the plane defined by the aromatic ring of Phe and the vector connecting the centroids R of the aromatic rings of BNZ), and single point energy calculations at the *ab-initio* CCSD(T)/6-311+G(3df,2p) level of theory (E_{CCSD}) and by molecular mechanics using the AMBER99 forcefield (E_{AMBER}).

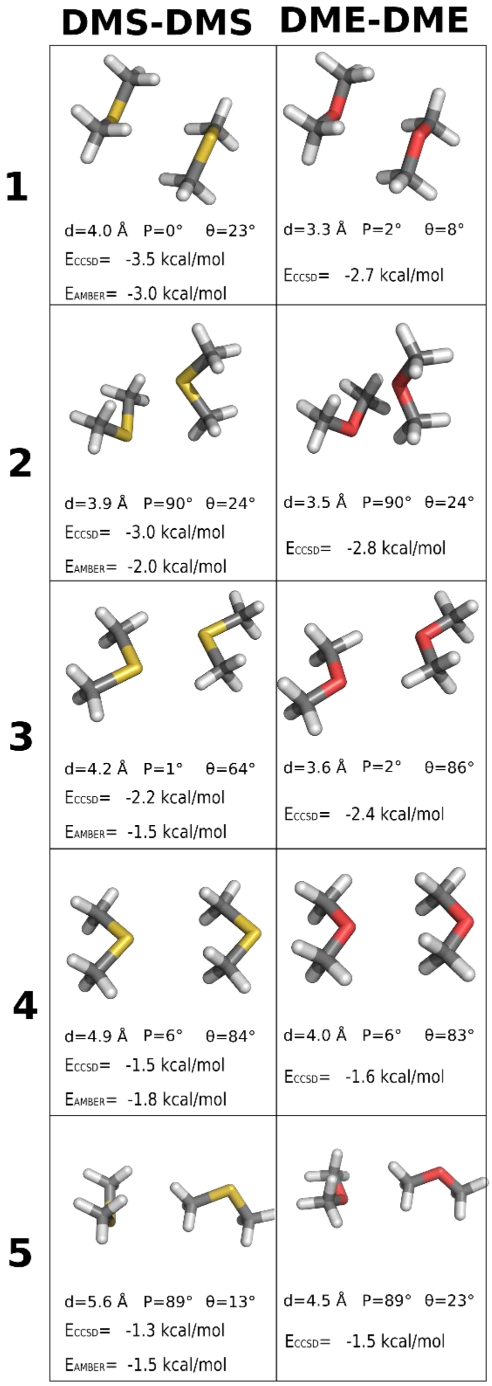

DMS-DMS
DME-DME
1
d=4.0 Å P=0° θ=23°
E_{CCSD}= -3.5 kcal/mol
E_{AMBER}= -3.0 kcal/mol
d=3.3 Å P=2° θ=8°
E_{CCSD}= -2.7 kcal/mol
2
d=3.9 Å P=90° θ=24°
E_{CCSD}= -3.0 kcal/mol
E_{AMBER}= -2.0 kcal/mol
d=3.5 Å P=90° θ=24°
E_{CCSD}= -2.8 kcal/mol
3
d=4.2 Å P=1° θ=64°
E_{CCSD}= -2.2 kcal/mol
E_{AMBER}= -1.5 kcal/mol
d=3.6 Å P=2° θ=86°
E_{CCSD}= -2.4 kcal/mol
4
d=4.9 Å P=6° θ=84°
E_{CCSD}= -1.5 kcal/mol
E_{AMBER}= -1.8 kcal/mol
d=4.0 Å P=6° θ=83°
E_{CCSD}= -1.6 kcal/mol
5
d=5.6 Å P=89° θ=13°
E_{CCSD}= -1.3 kcal/mol
E_{AMBER}= -1.5 kcal/mol
d=4.5 Å P=89° θ=23°
E_{CCSD}= -1.5 kcal/mol

Figure S4.4-3. Small-molecule models systems mimicking Met-Met interactions. Geometry optimized models, at the *ab-initio* MP2/6-31+G(d,p) level of theory, of dimethyl sulfide (DMS, mimicking Met)-DMS and dimethyl ether (DME)-DME interactions. Each energy-minimized structure is designated by an arabic number that corresponds to a roman number of the obtained clusters in crystal structures (see Fig 4.4-2). The values of d, P, and θ (see Suppl. Table 2 and Fig 2 for definition), and single point energy calculations at the *ab-initio* CCSD(T)/6-311+G(3df,2p) level of theory (E_{CCSD}) and by molecular mechanics using the AMBER99 forcefield (E_{AMBER}) are shown.

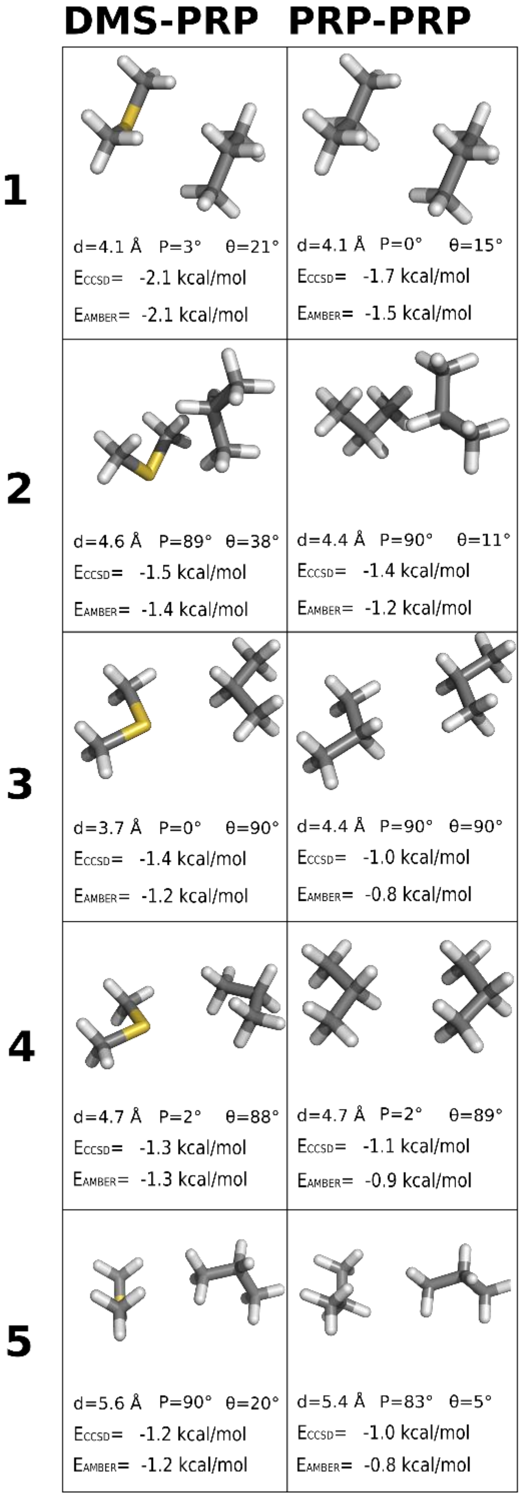

DMS-PRP
PRP-PRP
1
d=4.1 Å P=3° θ=21°
E_{CCSD}= -2.1 kcal/mol
E_{AMBER}= -2.1 kcal/mol
d=4.1 Å P=0° θ=15°
E_{CCSD}= -1.7 kcal/mol
E_{AMBER}= -1.5 kcal/mol
2
d=4.6 Å P=89° θ=38°
E_{CCSD}= -1.5 kcal/mol
E_{AMBER}= -1.4 kcal/mol
d=4.4 Å P=90° θ=11°
E_{CCSD}= -1.4 kcal/mol
E_{AMBER}= -1.2 kcal/mol
3
d=3.7 Å P=0° θ=90°
E_{CCSD}= -1.4 kcal/mol
E_{AMBER}= -1.2 kcal/mol
d=4.4 Å P=90° θ=90°
E_{CCSD}= -1.0 kcal/mol
E_{AMBER}= -0.8 kcal/mol
4
d=4.7 Å P=2° θ=88°
E_{CCSD}= -1.3 kcal/mol
E_{AMBER}= -1.3 kcal/mol
d=4.7 Å P=2° θ=89°
E_{CCSD}= -1.1 kcal/mol
E_{AMBER}= -0.9 kcal/mol
5
d=5.6 Å P=90° θ=20°
E_{CCSD}= -1.2 kcal/mol
E_{AMBER}= -1.2 kcal/mol
d=5.4 Å P=83° θ=5°
E_{CCSD}= -1.0 kcal/mol
E_{AMBER}= -0.8 kcal/mol

Figure S4.4-4. Small-molecule models systems mimicking Met-Leu interactions. Geometry optimized models, at the *ab-initio* MP2/6-31+G(d,p) level of theory, of dimethyl sulfide (DMS, mimicking Met) and propane (PRP, mimicking Leu) and PRP-PRP interactions. Each optimized structure is designated by an arabic number that corresponds to a roman number of the obtained clusters in crystal structures (see Fig 4.4-3). The values of d, P, and θ (see Suppl. Table 4.4-3 and Fig 4.4-3 for definition), and single point energy calculations at the *ab-initio* CCSD(T)/6-311+G(3df,2p) level of theory (E_{CCSD}) and by molecular mechanics using the AMBER99 forcefield (E_{AMBER}) are shown.

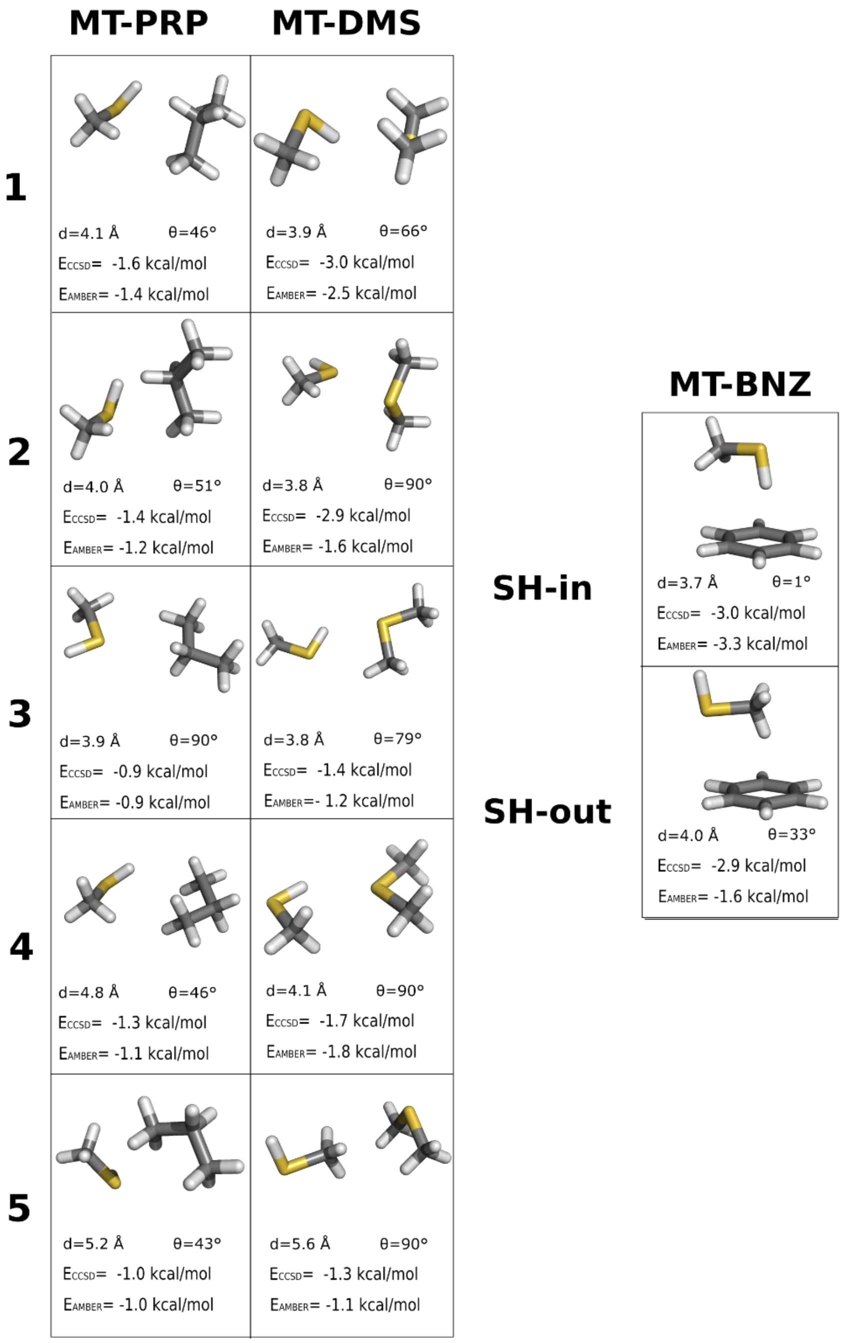


Figure S4.4-5. Small-molecule models systems mimicking Cys-Phe, Cys-Met and Cys-Leu interactions. Geometry optimized models, at the *ab-initio* MP2/6-31+G(d,p) level of theory, of the interactions between methanethiol (MT, mimicking Cys) and benzene (BNZ, mimicking Phe), dimethyl sulfide (DMS, mimicking Met) and propane (PRP, mimicking Leu). The values of d (calculated as the distance between S and Cγ in PRP or the centroid R of the aromatic ring in BNZ, respectively) and θ (angle between the normal vector of the plane defined by the aromatic ring of Phe or the plane defined by $C_β$, $S_γ$ and HS atoms and the vector connecting the centroid R of the aromatic ring of BNZ or the central atom of the side-chain to the center of the other side-chain), and single point energy calculations at the *ab-initio* CCSD(T)/6-311+G(3df,2p) level of theory (E_{CCSD}) and by molecular mechanics using the AMBER99 force field (E_{AMBER}).

4.5. GPCR-SAS

A web application for statistical analyses on G protein-coupled receptors sequences.

Abstract

G protein coupled receptors (GPCRs) are one of the largest protein families in mammals. They mediate signal transduction across cell membranes and are important targets for the pharmaceutical industry. The G Protein-Coupled Receptors - Sequence Analysis and Statistics (GPCR-SAS) web application provides a set of tools to perform comparative analysis of sequence positions between receptors, based on a curated structural-informed multiple sequence alignment. The analysis tools include: (i) frequency and entropy for one or more amino acids in a sequence position or range, (ii) covariance of two positions, (iii) correlation between two amino acids in two positions (or two sequence motifs in two range of positions), and (iv) snake-plot representation for a specific receptor or for the consensus sequence of a group of selected receptors. The analysis of conservation of residues and motifs across transmembrane (TM) segments is key in the study of ligand and G protein selectivity or family-specific mechanisms of activation. As an example, here we analyze the amino acids of the "transmission switch", that initiates receptor activation following ligand binding. The tool is freely accessible at http://lmc.uab.cat/gpcrsas/.

Introduction

G protein coupled receptors (GPCRs) are one of the most prevailing protein families in mammalian genomes (Overington *et al.,* 2006) and the major protein family as drug targets, with about one third of marketed compounds targeting them (Santos *et al.,* 2017, Hauser *et al.,* 2017). They are involved in most signal transduction processes across membranes, including the response to hormones and neurotransmitters and the senses of sight, smell and taste. GPCRs transduce extracellular signals across the cell membrane through G protein dependent but also through G protein independent processes (Sun *et al.,* 2007). GPCRs are classified into six families or classes named A to F according to sequence similarities (Kolakowski 1994, Altwood & Findlay 1994, Hom *et al.,* 2003). Only classes A, B, C and F are present in humans and class A or rhodopsin-like comprises, by far, the largest number of members (Fredriksson *et al.,* 2003). The available GPCRs crystal structures showed a conserved TM structure with a common fold formed by an extracellular N–terminus, seven transmembrane helices (TM1-7), connected by alternating intracellular (ICL1 to ICL3) and extracellular (ECL1 to ECL3) hydrophilic loops, and a cytoplasmic C–terminus (Liapakis *et al.,* 2012, Katrich *et al.,* 2013, Venkatakrishnan *et al.,* 2013) that starts with an α-helix (Hx8) parallel to the cell membrane in classes A, B and F. This structural similarity in the TM domain facilitates comparative analysis between members of the family. This was early recognized by Ballesteros and Weinstein when they developed the common residue numbering system (Ballesteros & Wenstein 1995) and have been exploited in the GPCRdb (Isberg *et al.,* 2016), a dedicated database for GPCRs, with sequence, structural and ligand information.

Here we present GPCR-SAS, a web application that permits easy comparison and statistical analysis of sequence positions or motifs within the TM helices and helix 8 across GPCRs of classes A, B, C and F. Our tool can be of help in identifying residues undergoing correlated evolution, and thus represents a useful instrument to rationalize ligand selectivity, G protein recognition or receptor activation, among others. As an example, here we illustrate the utility of GPCR-SAS by analyzing the amino acids that belong to the "transmission switch", that initiates receptor activation following ligand binding (Deupí *et al.,* 2012, Trzaskowski *et al.,* 2012, Sansuk et al, 2011).

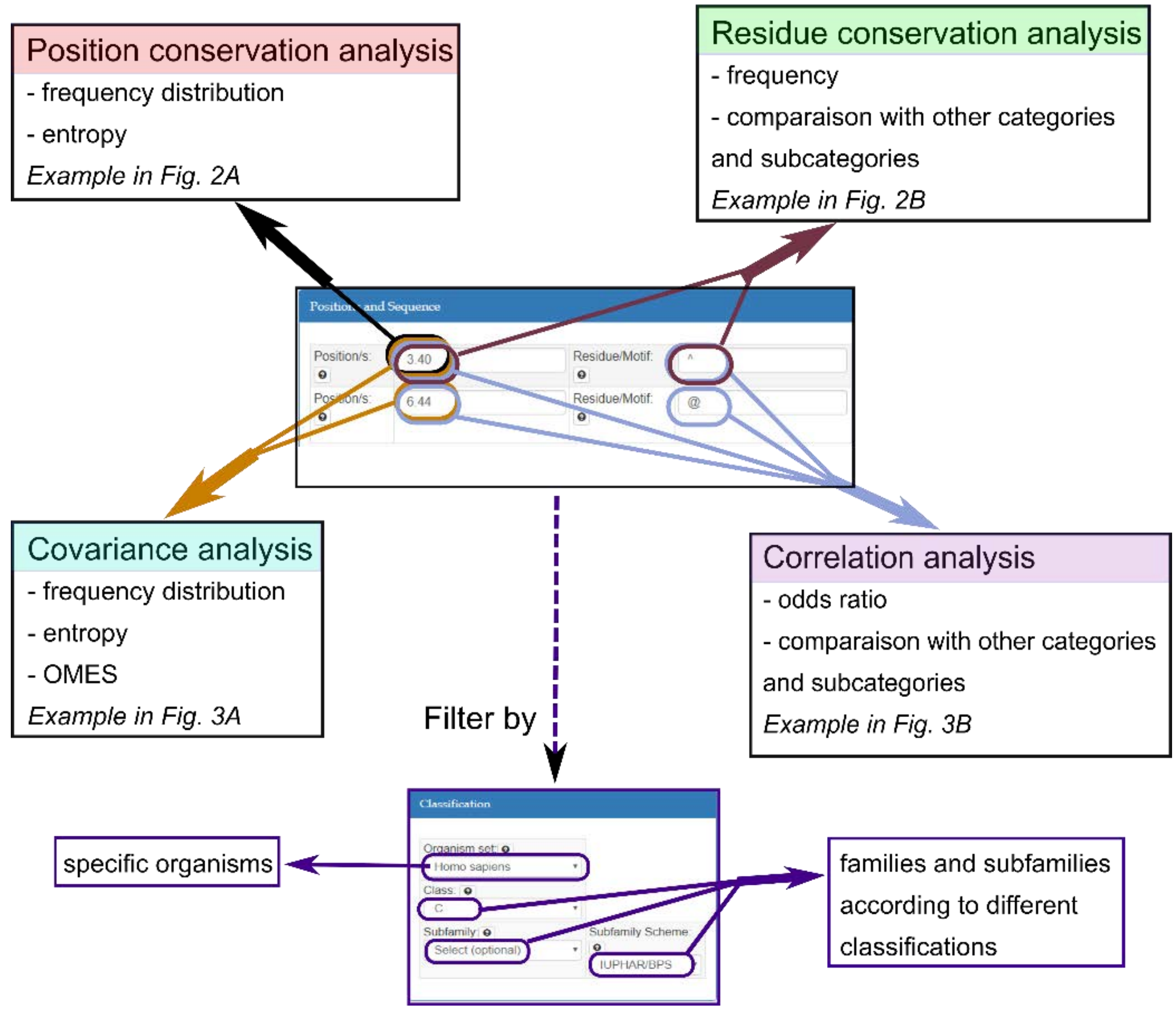


Figure 4.5-1. Schematic representation of the input forms of GPCR-SAS and the possible outcomes. The input of GPCR-SAS consists in two main sections: (i) Positions and Sequence: where the user can introduce a position/set of positions and a residue or sequence motif and (ii) Classification, that provides navigable multilevel hierarchical classification of GPCRs in families, branches and various levels of subfamilies according to different implemented schemes.

Material and methods

GPCR-SAS web application

GPCR-SAS is a web application freely accessible at http://lmc.uab.cat/gpcrsas/. The main tool is written in Python (version 2.7; available at http://www.python.org) and employs Django framework (version 1.5; available at https://djangoproject.com). The application relies on a MySQL database (version 5.1.73; available at https://www.mysql.com/) that contains the sequence alignments of all GPCRs sequences and previously reported classification schemes (Fredriksson *et al.,* 2003, Isberg *et al.,* 2016, Davies *et al.,* 2007, Surgand *et al.,* 2006, Deville *et al.,* 2009). Its design and implementation permit automatic incorporation of additional sequences as they are incorporated or edited in the UniProt (The UniProt Consortium, 2017). The input of GPCR-SAS consists in two main sections: (i) ***Positions and Sequence***: where the user can introduce a position/set of positions and a residue or sequence motif and (ii) ***Classification***, that provides navigable multilevel hierarchical classification of GPCRs in families, branches and various levels of subfamilies according to different implemented schemes. The output provides conservation, covariance or correlation analysis for different classification sub-levels depending on the input provided, as schematically shown in Figure 4.5-1.

Database of GPCRs sequence alignments

GPCRS-SAS database currently contains multiple sequence alignments of the transmembrane helices and helix 8 of 2982 GPCRs sequences of all species: 2377 class A, 206 class B, 111 class C and 297 class F. Amino acid sequences for all GPCRs belonging to classes A, B, C and F were retrieved from the UniProtKB/Swiss-Prot database (http://www.uniprot.org) on May the 15th 2018 (The UniProt Consortium, 2017). For each class, the sequences of receptors with an available crystal structure were aligned in a first step using MultiProt (Shatski *et al.,* 2004). This structural alignment was used to define consensus boundaries for each TM helix on every GPCR class and to construct an initial sequence profile aware of the TM segments. Because the main purpose of GPCR-SAS is the comparison between sequences, we chose to be conservative in terms of gap introduction. Otherwise the alignment cannot be used for unambiguous predictions (Cvicek *et al.,* 2016). In the class A, irregularities observed between different structure patterns were handled using gaps on TMs 2 and 5 as previously described (González *et al.,* 2012, Becu *et al.,* 2013, Isberg *et al.,* 2015). For each class, we aligned the sequences in three steps: first human sequences, next the rest of mammalian sequences and finally, the remaining vertebrate and invertebrate sequences. All multiple sequence alignments were performed with Clustal Omega (Sievers *et al.,* 2011). At each step the alignment was manually curated ensuring lack of gaps on TM regions other than those at TMs 2 and 5 in class A receptors (González *et al.,* 2012, Becu *et al.,* 2013) or those associated to trivial deletions. Finally, the alignments for each class were assembled based on the structural alignment between crystallized receptors from the different classes (Siu *et al.,* 2013, Spyridaki *et al.,* 2014). The final alignment is similar to the one presented by Cvicek and collaborators (Cvicek *et al.,* 2016) except for the gaps in TMs 2 and 5 (positions 2x551 and 5x461 according to the generic residue numbering (Isberg *et al.,* 2015)), whereas it does not contain other gaps that appear in the GPCRdb alignment (Isberg *et al.,* 2015).
The updated list of GPCRs sequences and the alignment are available at the help panel (http://lmc.uab.cat/gpcrsas/about/). By default, the class A set in the GPCRS-SAS database consists of the 1824 non-olfactory GPCRs, since olfactory receptors are excluded to avoid biasing of results towards this subfamily -it accounts for almost two thirds (561 receptors) of the human class A GPCRs. The sequence alignment for TM helices and the short helix 8 perpendicular to de membrane for a selected group of receptors is shown in S4.5-1 Fig. Regular updates are planned every 3 months. This step implies aligning new sequences to the previous alignment and assigning the different classification categories. We have scheduled a yearly full update where we will rebuild the complete alignment to account for possible changes in the structural alignment due to new structures.

Sequence numbering scheme

Each position in the TM segments is numbered according to the Ballesteros & Weinstein numbering scheme for class A GPCRs (Ballesteros & Wenstein 1995). In this numbering, the position of each residue is described by two numbers: the helix in which the residue is located and the position relative to a conserved residue in that helix, arbitrarily assigned to 50, separated by a dot. In class A, these amino acids are: N1.50 in TM1 (97.6% conserved in human class A excluding olfactory receptors; data from GPCR-SAS), D2.50 in TM2 (92.1%), R3.50 in TM3 (94.8%), W4.50 in TM4 (95.8%), P5.50 in TM5 (76.0%), P6.50 in TM6 (98.3%), and P7.50 in TM7 (93.7%). Although GPCRs of classes other than A most often do not have such conserved amino acids at these positions (Isberg *et al.,* 2015), extrapolation of the class A numbering-scheme is now possible thanks to structure-based sequence alignment between classes (Siu *et al.,* 2013, Spyridaki *et al.,* 2014). S4.5-1 Fig shows the sequence alignment for the TM helices of a selected group of receptors from different classes.

GPCR classifications

GPCR-SAS queries permit to filter receptor sequences based on previously reported classification systems including: (i) Fredriksson(Fredriksson *et al.,* 2003), based on a phylogenetic analysis of human GPCR sequences (branches and one subfamily level); (ii) GPCRdb (Vroling *et al.,* 2011), which uses a pharmacologic classification of the receptors (three subfamily levels); (iii) BIAS-PROF GDS (Davies *et al.,* 2007), based on the comparison of the protein sequence using the physicochemical properties of the amino acids (two subfamily levels); (iv) Rognan (Surgand *et al.,* 2006), that relies on the phylogenetic analysis of 30 positions putative involved on the ligand binding site (one subfamily level), (v) Chabbert (Deville *et al.,* 2009), which employs multidimensional scaling to cluster GPCRs (branches and one subfamily level), and (vi) GPCR SARfari (https://www.ebi.ac.uk/chembl/sarfari/gpcrsarfari), based on chemogenomic data (three subfamily levels). As a complementary filter option, we defined five sequence sets based on different levels of taxonomic classification: Human, Mammals, Vertebrates, Eukaryotes and All (sequences for all species).

Analysis Tools

The relevance of GPCR-SAS is its capability to analyze conservation of residues or sequence motifs across TM segments of GPCRs and to identify correlations between two positions for various classification schemes at different levels (class, branch and subfamilies) and five different taxonomic sets.

Conservation analysis for a position or a set/range of positions

The entropy can be given for a single position (i.e. 3.50) or for a range of consecutive (i.e. 3.50-3.54) or non-consecutive (3.50, 4.50, 5.50) positions. When the input is a single position, a graph with the amino acids counts is also displayed. For each position or range of positions *I*, the entropy of the information contained *H(i)* is computed according to Shannon's theorem (Shannon, 1948) as:

$$H(i) = -\sum_x p_x\,(i)\,log_b\,p_x\,(i) \qquad (1)$$

where $p_x(i)$ is the probability mass function for the amino acid(s) at position (or group of positions) *i*. The logarithm base *b* serves to scale the entropy in the range [0, 1] for one or more positions. Consequently, *b* is 20^n, with *n* being the number of positions used for the calculation. A position or group of positions with low variability (high conservation) has an entropy *H(i)* close to 0, while a position with high variability (low conservation) has an entropy close to 1.

The percentage of occurrence of an amino acid or motif in a specific position(s) (i.e. N 7.49, P 7.50) can also be computed. In this case, the output allows comparison to other categories and subcategories. To represent residues or motifs with specific physicochemical properties the user can utilize one-character wildcards as a residue or as part of a motif with the following correspondences: 'X' (any amino acid), '@' (aromatic, W/Y/F/H), '~' (apolar, I/L/V/A/F/P), '+' (positively charged, R/H/K), '-' (negatively charged, D/E), '*' (charged, R/H/K/D/E) and '^' (polar, D/E/N/Q/K/R/H/S/T/C/W/Y).

Covariance analysis

To analyze the covariance of two positions, GPCR-SAS uses the Observed Minus Expected Squared (OMES) (Fodor & Aldrich, 2004), that is based on a χ^2 test and a corrected mutual information method (MIp) (Dunn *et al.,* 2008). Both methods have previously been employed by Pele *et al.,* to identify evolutionary hubs between pairs of residues in GPCRs (Pele *et al.,* 2014).

OMES calculates the difference between the observed and expected frequencies of each possible pair of amino acids *(x, y)*, at positions *i* and *j* of the alignment:

$$OMES(i,j) = \frac{1}{N(i,j)}\sum_{x,y}(N_{x,y}^{obs}(i,j) - N_{x,y}^{exp}(i,j))^2 \qquad (2)$$

with $N(i,j)$, $N_{x,y}^{obs}$ and $N_{x,y}^{exp}$ being the number of sequences in the alignment with non-gapped

residues, the observed frequency and the expected frequency, respectively, at positions (or list of positions) *i* and *j*.
The *MI* content *MI(i,j)* between two positions (or lists of positions) *i* and *j* on an alignment is based on the probability of joint occurrence of events and is defined as:

$$MI(i,j) = \sum_{x,y} p_{x,y}\,(i,j)\, ln \frac{p_{x,y}(i,j)}{p_x(i)p_y(j)} \qquad (3)$$

where $p_x(i)$, $p_y(j)$ and $p_{x,y}(i,j)$ are respectively the frequencies of amino acid *x* at position *i*, amino acid *y* at position *j* and the amino acid pair *(x*, *y)* at positions *i* and *j*.
The corrected MIp version is defined as:

$$MIp(i,j) = MI(i,j) - \frac{\frac{1}{n-1}\sum_{j\neq i} MI(i,j)\frac{1}{n-1}\sum_{i\neq j} MI(i,j)}{\frac{2}{n(n-1)}\sum_{i,j} MI(i,j)} \qquad (4)$$

with *n* being the number of columns in the alignment.
To evaluate the statistical significance for the computed OMES and MIp values, GPCR-SAS provides the Z-scores and the associated p-values, which are computed by comparing with the mean value for all combinations of two positions.

Correlation analysis

To determine the correlation between two sequence positions, the occurrence of the amino acid or motif at the first position or range of positions is associated with the occurrence of the amino acid or motif at the second position or range of positions. The occurrences are used to compute an odds ratio (and the associated 95% confidence interval) that estimates how strongly the presence/absence of one of the first amino acids or motif is correlated with the presence/absence of the second amino acid or motif. To facilitate the comparison with the other categories at the chosen level of classification and in subcategories, the output of a correlation analysis also returns the same analysis for these groups.

Snake-plot representations

GPCR-SAS can also provide snake-plot representations for the sequence of a specific receptor or for the consensus sequence of a group of receptors (S4.5-2 Fig). Each residue is represented by a circle with a letter in gray-gradient (representing the frequency of the residue on the class) and an outline in blue-gradient (representing the frequency of the residue on the selected group of receptors). In single-receptor snake-plots, residues colored in green are those that do not match the most conserved residue for the selected group of receptors (i.e. the selected subfamily).

Results

To illustrate the use of GPCR-SAS, we analyzed conservation, covariance and correlation of the residues of the “transmission switch”, which is one of the initial steps of receptor activation following ligand binding in class A GPCRs (Deupí *et al.,* 2012, Trzaskowski *et al.,* 2012, Sansuk et al, 2011). The “transmission switch” involves positions 3.40, 5.50 and 6.44. Rearrangement of the packing between these residues following ligand binding at the extracellular side of the receptor weakens the interface between TM helices 5 and 6 and triggers local conformational changes that are transmitted towards the cytoplasmic side, where G proteins and β-arrestins bind (Rasmussen *et al.,* 2011, Kang *et al.,* 2015) (Fig 4.5-2A). By using GPCR-SAS we will determine if a “transmission switch” is likely to exist in receptors belonging to classes B and C of GPCRs.

The transmission switch in Class A GPCRs

In class A GPCRs, the comparison between crystal structures of inactive versus active states showed a rearrangement of residues 3.40, 5.50 and 6.44 of the “transmission switch” (Rasmussen *et al.,* 2011, Cherezov *et al.,* 2007). The changes observed in the β$_2$-adrenergic receptor (β$_2$-AR) are displayed in Fig 4.5-2B and 4.5-2C. The percentage of occurrence of each amino acid at position 3.40 for human non-olfactory class A GPCRs computed with GPCR-SAS is shown in Fig 4.5-3A. The information displayed includes also the entropy of the position (as a measure of variability, see Methods) and a histogram with the counts of each amino acid at this position grouped according to subfamily categories. The most prevalent residue in this position is Ile (39.2% conserved), followed by Val (24.1%) and Leu (11.3%), all of them sharing hydrophobic properties. The entropy value of 0.6 indicates a moderate variability (entropy values ranges between 0 and 1, see Methods). The histogram shows that most amine receptors, including the β2-AR, feature Ile at this position. Indeed, the frequency of Ile raises to 76.2% when restricting the query to the amine subfamily (according to Fredriksson's classification scheme (Fredriksson *et al.,* 2003)). The output also provides (“Click to show receptors” button) the list of UniProt entry names for the receptors that matched the query. Clicking on an amino acid allows easy comparison with the same position in the other categories at the same level of classification (Same-level button) and within the categories of the child subfamilies (Sub-level button). In this particular example, the categories at the same level mean the other receptor classes (B, C and F) and the child categories are Fredriksson's branches (Fredriksson *et al.,* 2003). The output of the query for Ile (shown in Fig 4.5-3D) indicates that this residue is rarely/never found at position 3.40 in classes B (0%), C (13.6%) and F (5.6%). Regarding the sublevels, GPCR-SAS tells that Ile is preferentially found at the α (48.0%) and δ branches (43.3%). To expand the initial search, we used the wildcard for apolar amino acids (see Methods) which also accounts for Leu, Val, Ala, Phe and Pro. The frequencies for apolar residues add to 86.9% in class A, 79.2% in class B, 50% in class C, 69.4% in class F GPCRs. Thus, this analysis shows that despite Ile3.40 is not conserved in GPCRs of classes B, C and F, these receptors have mostly kept hydrophobic amino acids. Similar queries for the content of positions 5.50 and 6.44 in class A GPCRs reveals that Pro (78.4%) and aromatic residues (Phe:80.4%, Tyr:6.9%), respectively, are the most prevalent amino acids. The β2-AR has both Pro5.50 and Phe6.44 (see Fig 4.5-2B and 4.5-2C). To identify the most common residue triad, the three positions for human non-olfactory class A GPCRs were used as query (coma-separated in the box position). GPCR-SAS returns that the most frequent “transmission switch” residues are Ile3.40-Pro5.50-Phe6.44 as in the β2-AR (32.3%), Val3.40-Pro5.50-Phe6.44 (16.5%), and other triads with a percentage of occurrence smaller than 5% each. Overall, the triad Ile/Val3.40-Pro5.50-Phe6.44 comprises nearly half of the class A GPCRs.

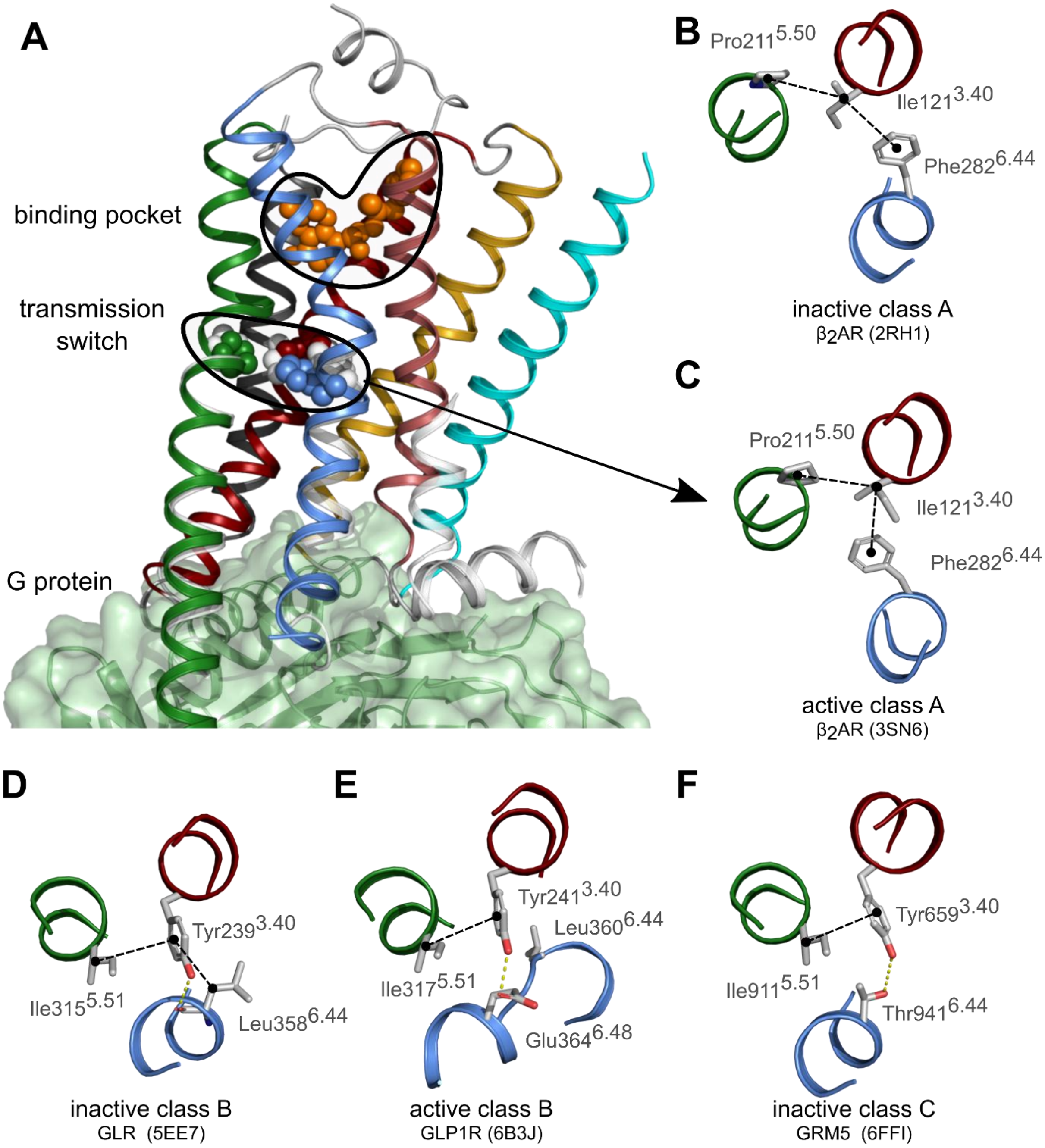
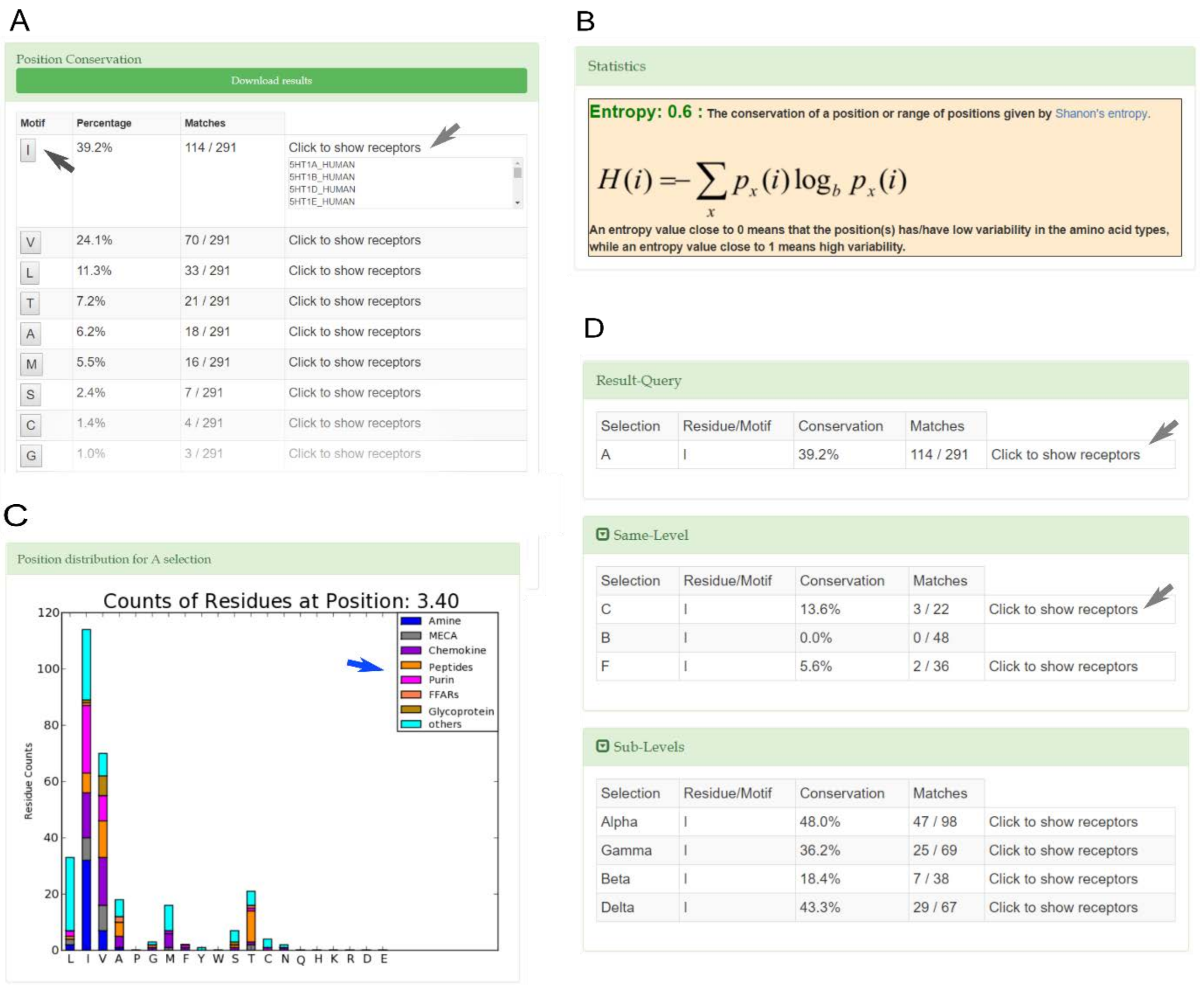


Fig 4.5-2. The transmission switch in the crystal structures of GPCRs. (A) Cartoon representation of the crystal structure of the active β2-adrenergic receptor (PDB id 3SN6) in complex with the G protein (green surface) illustrating the localization of the transmission switch just below the orthosteric binding pocket. The sidechains of residues of the transmission switch (3.40, 5.50 and 6.44) and the ligand (in orange) are shown as spheres. The color-code for the TM helices is 1:cyan, 2:gold, 3:red; 4:dark-gray, 5:green, 6_blue, 7:pale-red. Helix 8 and loops are shown in light-gray. Superposed (in white) the cytoplasmic ends of TMs 5, 6 and 7 and the residues of the transmission switch in the inactive structure (PDB id 2RH1). **(B-F)** Detail of the ªtransmission switchº in the crystal structures of inactive **(B)** and active **(C)** β2-adrenergic receptor (ADRB2), inactive glucagon receptor (GLR) **(D)**, active glucagon-like peptide 1 receptor (GLP1R) **(E)**, and inactive



Position Conservation

Motif	Percentage	Matches	
I	39.2%	114 / 291	Click to show receptors 5HT1A_HUMAN 5HT1B_HUMAN 5HT1D_HUMAN 5HT1E_HUMAN
V	24.1%	70 / 291	Click to show receptors
L	11.3%	33 / 291	Click to show receptors
T	7.2%	21 / 291	Click to show receptors
A	6.2%	18 / 291	Click to show receptors
M	5.5%	16 / 291	Click to show receptors
S	2.4%	7 / 291	Click to show receptors
C	1.4%	4 / 291	Click to show receptors
G	1.0%	3 / 291	Click to show receptors

Statistics

Entropy: 0.6 : The conservation of a position or range of positions given by Shanon's entropy.

$$H(i) =-\sum_{x} p_x(i)\log_b p_x(i)$$

An entropy value close to 0 means that the position(s) has/have low variability in the amino acid types, while an entropy value close to 1 means high variability.

Position distribution for A selection

Result-Query

Selection	Residue/Motif	Conservation	Matches	
A	I	39.2%	114 / 291	Click to show receptors

Same-Level

Selection	Residue/Motif	Conservation	Matches	
C	I	13.6%	3 / 22	Click to show receptors
B	I	0.0%	0 / 48	
F	I	5.6%	2 / 36	Click to show receptors

Sub-Levels

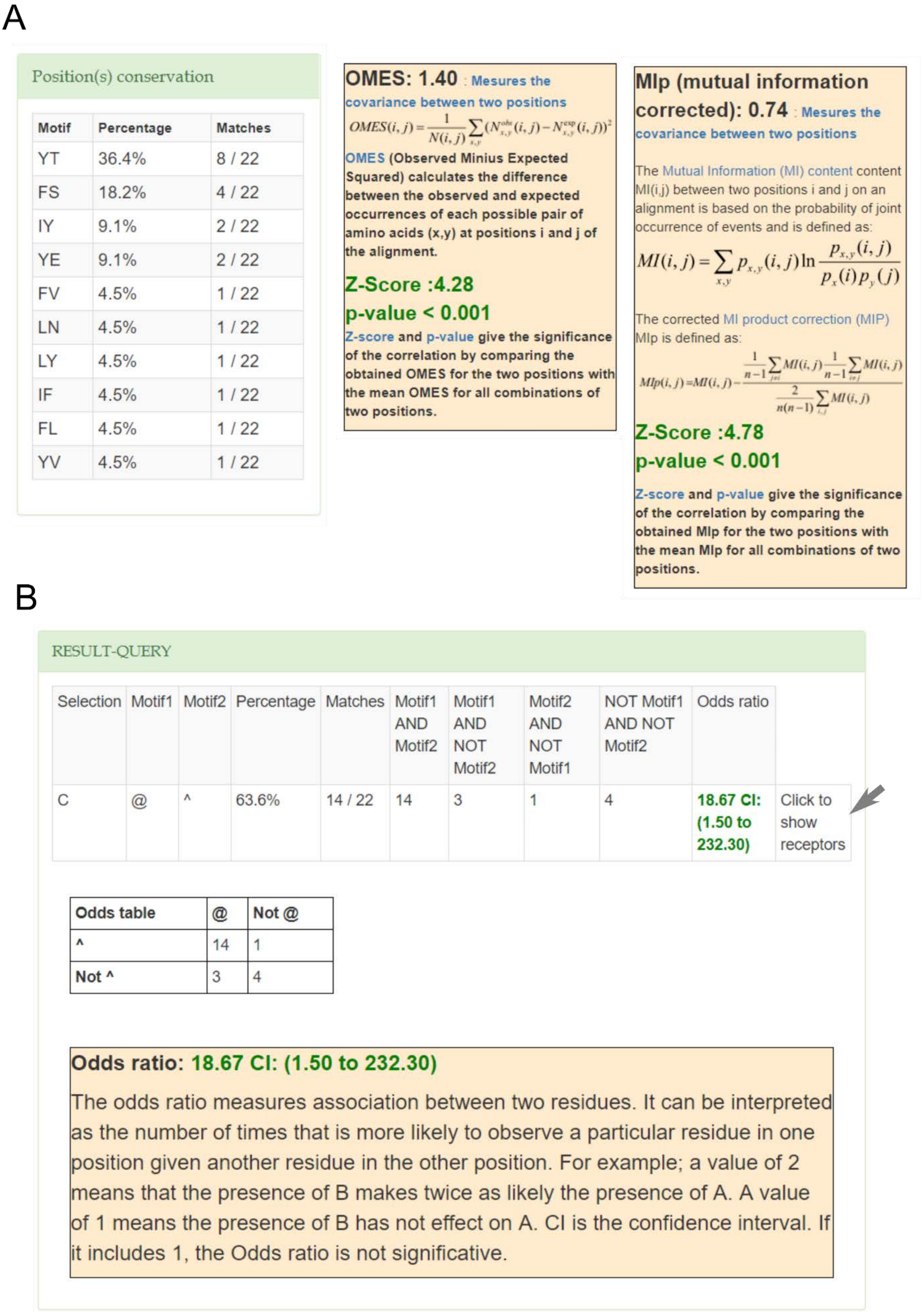
Selection	Residue/Motif	Conservation	Matches	
Alpha	I	48.0%	47 / 98	Click to show receptors
Gamma	I	36.2%	25 / 69	Click to show receptors
Beta	I	18.4%	7 / 38	Click to show receptors
Delta	I	43.3%	29 / 67	Click to show receptors

Fig 4.5-3. Conservation analysis queries for position 3.40 (A-C) and for Ile at position 3.40 (D) in human non-olfactory class A GPCRs. **(A)** the frequencies of amino acids at position 3.40; (**B**) the histogram of the amino acid frequencies at position 3.40 for the most the major subcategories (blue arrow); **(C)** the entropy of position 3.40; (**D**) the frequencies of Ile at position 3.40 for class A GPCRs compared the other classes (B, C and F; Same Level panel) and for the different class A branches (Sub-Level panel). Gray arrows indicate that it is possible to get the list of receptors that contain a certain residue or motif; the black arrow indicates that a click on the residue type in (**A**) provides the output displayed in (**D**).

Transmission switch in classes B and C GPCRs

Next, we analyzed positions 3.40, 5.50 and 6.44 in GPCRs of classes B and C. GPCR-SAS shows that human class B GPCRs mostly feature aromatic residues at position 3.40 (Tyr:66.7% and Phe:20.8%), aliphatic residues at position 5.50 (39.6% Ile, 20.8% Val, 14.6% Ala) and Leu (64.6%) or Phe (20.8%) at position 6.44. Compared to class A GPCRs, class B receptors have switched residue types at positions 3.40 and 6.44 (from aliphatic-aromatic in class A to aromatic-aliphatic in class B), but still exhibit conserved residue types in these positions in this class. The results are compatible with our recent proposal, based on mutagenesis studies, that residues Phe3.40 and Leu6.44 form the “transmission switch” in the corticotropin-releasing factor 1 receptor (Spyridaki *et al.,* 2014). The crystal structures of class B GPCRs show that the side chain of Tyr/Phe3.40 interacts with Leu/Phe6.44 in the inactive state but that this interaction is lost in the active state (Fig 4.5-2D and 4.5-2E). In addition, in the glucagon family of receptors (which contain Tyr3.40) the hydroxyl group of Tyr forms a hydrogen bond with the backbone of residue 6.44 in the inactive state and with the side-chain of Glu6.48 in the active state. GPCR-SAS tells that Glu6.48 is exclusive of the glucagon family of receptors. The most common pair for positions 3.40-6.44 is Phe/Leu. Class B receptors also lack the characteristic Pro at position 5.50 present in class A GPCRs.

For human class C receptors, GPCR-SAS shows that position 3.40 mostly contains an aromatic residue (Tyr:50%, Phe:27.3%; entropy 0.4), but position 6.44 exhibits more variability (Thr:36.4%, Ser:18.2%, Tyr:13.6%, Glu:9.1%, Val:9.1%; entropy of 0.6). The most common pairs of residues at these positions in the class C receptors are aromatic-polar (Tyr-Thr:36.4%, Phe-Ser:18.2%, Tyr-Glu:9.1%) and aliphatic-aromatic (Ile-Tyr: 9.1%). The subfamily of metabotropic glutamate receptors (mGluRs) all contain Tyr3.40 and Thr6.44 (see mGluR5 in Fig 2F). Like class B receptors, class C receptors lack Pro5.50. We next analyzed the covariance of positions 3.40 and 6.44, that is, if changes in position 3.40 occur together with changes in position 6.44. Both GPCR-SAS revealed statistically significant covariance using OMES and MIp analyses (p-values <0.001 in both cases) with Z-scores of 4.28 and 4.78 respectively (Fig 4.5-4A). Furthermore, A GPCRs-SAS correlation analysis for aromatic residues at position 3.40 and polar residues at position 6.44 provided an odds ratio of 18.67 (with a 95% confidence interval of 1.50 to 232.3). For better statistics, we increased the number of sequences by using the “All organism” set. In this case we got a stronger association (odds ratio of 40.89) and a narrower confidence interval (13.48 to 123.99), (Fig 4.5-4B). Similarly, (with the “All organism set”) we obtain an odds ratio of 28.27 for an apolar residue at position 3.40 and an aromatic residue at position 6.44 (with a 95% confidence interval between 3.67 and 217.83), supporting a coordinated role for both amino acids. This covariance of positions 3.40 and 6.44, together with correlation of specific amino acids at these positions clearly suggest that both residues have a functional role in class C receptors as part of the same molecular switch. In fact, this is in accordance with our recent proposal, based on mutagenesis experiments and molecular modelling, that Tyr3.40 and Thr6.44 are part of the “transmission switch” in the GRM20 (Pérez-Benito *et al.,* 2017).



Position(s) conservation

Motif	Percentage	Matches
YT	36.4%	8 / 22
FS	18.2%	4 / 22
IY	9.1%	2 / 22
YE	9.1%	2 / 22
FV	4.5%	1 / 22
LN	4.5%	1 / 22
LY	4.5%	1 / 22
IF	4.5%	1 / 22
FL	4.5%	1 / 22
YV	4.5%	1 / 22

OMES: 1.40 : Mesures the covariance between two positions

$$OMES(i,j)=\frac{1}{N(i,j)}\sum_{x,y}(N^{obs}_{x,y}(i,j)-N^{exp}_{x,y}(i,j))^2$$

OMES (Observed Minius Expected Squared) calculates the difference between the observed and expected occurrences of each possible pair of amino acids (x,y) at positions i and j of the alignment.

Z-Score :4.28

p-value < 0.001

Z-score and p-value give the significance of the correlation by comparing the obtained OMES for the two positions with the mean OMES for all combinations of two positions.

MIp (mutual information corrected): 0.74 : Mesures the covariance between two positions

The Mutual Information (MI) content content MI(i,j) between two positions i and j on an alignment is based on the probability of joint occurrence of events and is defined as:

$$MI(i,j)=\sum_{x,y}p_{x,y}(i,j)\ln\frac{p_{x,y}(i,j)}{p_x(i)p_y(j)}$$

The corrected MI product correction (MIP) MIp is defined as:

$$MIp(i,j)=MI(i,j)-\frac{\frac{1}{n-1}\sum_{j\neq i}MI(i,j)\frac{1}{n-1}\sum_{i\neq j}MI(i,j)}{\frac{2}{n(n-1)}\sum_{i,j}MI(i,j)}$$

Z-Score :4.78

p-value < 0.001

Z-score and p-value give the significance of the correlation by comparing the obtained MIp for the two positions with the mean MIp for all combinations of two positions.

RESULT-QUERY

Selection	Motif1	Motif2	Percentage	Matches	Motif1 AND Motif2	Motif1 AND NOT Motif2	Motif2 AND NOT Motif1	NOT Motif1 AND NOT Motif2	Odds ratio	
C	@	^	63.6%	14 / 22	14	3	1	4	18.67 CI: (1.50 to 232.30)	Click to show receptors

Odds table	@	Not @
^	14	1
Not ^	3	4

Odds ratio: 18.67 CI: (1.50 to 232.30)

The odds ratio measures association between two residues. It can be interpreted as the number of times that is more likely to observe a particular residue in one position given another residue in the other position. For example; a value of 2 means that the presence of B makes twice as likely the presence of A. A value of 1 means the presence of B has not effect on A. CI is the confidence interval. If it includes 1, the Odds ratio is not significative.

Fig 4.5-4. Covariance (A) and correlation (B) analysis for positions 3.40 and 6.44 in class C GPCRs (A). In (**A**) the left panel shows a list of all motifs and its frequencies and the right panel shows the statistical tests: the OMES and MIp analyses indicates statistical significant covariance between both positions. In (**B**) the odds ratio for an aromatic residue at position 3.40 and a polar residue at position 6.44, which indicates statistical significant correlation. The gray arrow indicates that it is possible to get the list of receptors that contain a certain residue or motif.

Conclusions

Here we present GPCR-SAS, a web application that allows to perform frequency, covariance and correlation analyses for sequence positions or motifs in GPCRs. The tool takes advantage of the structural similarity in the TM domain of GPCRs and allows performing comparisons and statistical analyses of sequence positions or motifs within the TM helices and helix 8 for receptors of classes A, B, C and F. As an example of use, we here show its utility in the successful extrapolation of the “transmission switch” of class A GPCRs to classes B and C. GPCR-SAS tells that nearly half the class A GPCRs have as triads for positions 3.40-5.50-6.44 either Ile-Pro-Phe (32.3%) or Val-Pro-Phe (16.5%). Class B and Class C receptors lack Pro5.50 and contain alternative, but also conserved, pairs at positions 3.40-6.44. The preferred pairs in class B receptors are Phe-Leu (45.8%) and Tyr-Leu (12.5%). This implies a switch between residue types compared to the same positions in class A. The preferred pairs in class C receptors are Tyr-Thr (36.4%) and Phe-Ser (18.2.1%). Overall, as supported by previous mutagenesis data (Spyridaki *et al.,* 2014, Pérez-Benito *et al.,* 2017), this analysis suggests that different “transmission switches” involving the same positions but different residue types exist in classes A, B and C, and are likely to represent common activation pieces within the whole superfamily. The type of statistical analyses that GPCR-SAS performs can be used to find functionally important residues or groups of residues undergoing correlated evolution, and thus represents a useful instrument to rationalize ligand selectivity, G protein recognition or receptor activation, among others.

References

Attwood TK, Findlay JB (1994) Fingerprinting G-protein-coupled receptors. Protein Eng 7: 195-203.

Ballesteros JA, Weinstein H (1995) Integrated Methods for Modeling G-Protein Coupled Receptors. Methods Neurosci 25: 366-428.

Becu JM, Pele J, Rodien P, Abdi H, Chabbert M (2013) Structural evolution of G-protein-coupled receptors: a sequence space approach. Methods Enzymol 520: 49-66.

Cherezov V, Rosenbaum DM, Hanson MA, Rasmussen SG, Thian FS, *et al.,* (2007) High-resolution crystal structure of an engineered human beta2-adrenergic G protein-coupled receptor. Science 318: 1258-1265.

Cvicek V, Goddard WA, 3rd, Abrol R (2016) Structure-Based Sequence Alignment of the Transmembrane Domains of All Human GPCRs: Phylogenetic, Structural and Functional Implications. PLoS Comput Biol 12: e1004805.

Davies MN, Secker A, Freitas AA, Mendao M, Timmis J, *et al.,* (2007) On the hierarchical classification of G protein-coupled receptors. Bioinformatics 23: 3113-3118.

Deupi X, Standfuss J, Schertler G (2012) Conserved activation pathways in G-protein-coupled receptors. Biochem Soc Trans 40: 383-388.

Deville J, Rey J, Chabbert M (2009) An indel in transmembrane helix 2 helps to trace the molecular evolution of class A G-protein-coupled receptors. J Mol Evol 68: 475-489.

Dunn SD, Wahl LM, Gloor GB (2008) Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. Bioinformatics 24: 333-340.

Fodor AA, Aldrich RW (2004) Influence of conservation on calculations of amino acid covariance in multiple sequence alignments. Proteins 56: 211-221.

Fredriksson R, Lagerstrom MC, Lundin LG, Schioth HB (2003) The G-protein-coupled receptors in the human genome form five main families. Phylogenetic analysis, paralogon groups, and fingerprints. Mol Pharmacol 63: 1256-1272.

Gonzalez A, Cordomi A, Caltabiano G, Pardo L (2012) Impact of helix irregularities on sequence alignment and homology modeling of G protein-coupled receptors. Chembiochem 13: 1393-1399.

Hauser AS, Attwood MM, Rask-Andersen M, Schioth HB, Gloriam DE (2017) Trends in GPCR drug discovery: new agents, targets and indications. Nat Rev Drug Discov 16: 829-842.

Horn F, Bettler E, Oliveira L, Campagne F, Cohen FE, *et al.,* (2003) GPCRDB information system for G protein-coupled receptors. Nucleic Acids Res 31: 294-297.

Isberg V, de Graaf C, Bortolato A, Cherezov V, Katritch V, *et al.,* (2015) Generic GPCR residue numbers - aligning topology maps while minding the gaps. Trends in Pharmacological Sciences 36: 22-31.

Isberg V, Mordalski S, Munk C, Rataj K, Harpsoe K, *et al.,* (2016) GPCRdb: an information system for G protein-coupled receptors. Nucleic Acids Research 44: D356-D364.

Kang Y, Zhou XE, Gao X, He Y, Liu W, *et al.,* (2015) Crystal structure of rhodopsin bound to arrestin by femtosecond X-ray laser. Nature 523: 561-567.

Katritch V, Cherezov V, Stevens RC (2013) Structure-function of the G protein-coupled receptor superfamily. Annu Rev Pharmacol Toxicol 53: 531-556.

Kolakowski LF, Jr. (1994) GCRDb: a G-protein-coupled receptor database. Receptors Channels 2: 1-7.

Liapakis G, Cordomi A, Pardo L (2012) The G-protein coupled receptor family: actors with many faces. Curr Pharm Des 18: 175-185.

Overington JP, Al-Lazikani B, Hopkins AL (2006) How many drug targets are there? Nat Rev Drug Discov 5: 993-996.

Pele J, Moreau M, Abdi H, Rodien P, Castel H, *et al.,* (2014) Comparative analysis of sequence covariation methods to mine evolutionary hubs: examples from selected GPCR families. Proteins 82: 2141-2156.

Pérez-Benito L, Doornbos MLJ, Cordomí A, Peeters L, Lavreysen H, *et al.,* (2017) Molecular Switches of Allosteric Modulation of the Metabotropic Glutamate 2 Receptor. Structure 25: 1153-1162.

Rasmussen SG, DeVree BT, Zou Y, Kruse AC, Chung KY, *et al.,* (2011) Crystal structure of the beta2 adrenergic receptor-Gs protein complex. Nature 477: 549-555.

Santos R, Ursu O, Gaulton A, Bento AP, Donadi RS, *et al.,* (2017) A comprehensive map of molecular drug targets. Nat Rev Drug Discov 16: 19-34.

Sansuk K, Deupi X, Torrecillas IR, Jongejan A, Nijmeijer S, *et al.,* (2011) A structural insight into the reorientation of transmembrane domains 3 and 5 during family A G protein-coupled receptor activation. Mol Pharmacol 79: 262-269.

Shannon CE (1948) A Mathematical Theory of Communication. Bell System Technical Journal 27: 623-656.

Shatsky M, Nussinov R, Wolfson HJ (2004) A method for simultaneous alignment of multiple protein structures. Proteins 56: 143-156.

Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, *et al.,* (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol 7: 539.

Siu FY, He M, de Graaf C, Han GW, Yang D, *et al.,* (2013) Structure of the human glucagon class B G-protein-coupled receptor. Nature 499: 444-449.

Spyridaki K, Matsoukas MT, Cordomi A, Gkountelias K, Papadokostaki M, *et al.,* (2014) Structural-functional analysis of the third transmembrane domain of the corticotropin-releasing factor type 1 receptor: role in activation and allosteric antagonism. J Biol Chem 289: 18966-18977.

Sun Y, McGarrigle D, Huang XY (2007) When a G protein-coupled receptor does not couple to a G protein. Mol Biosyst 3: 849-854.

Surgand JS, Rodrigo J, Kellenberger E, Rognan D (2006) A chemogenomic analysis of the transmembrane binding cavity of human G-protein-coupled receptors. Proteins 62: 509-538.

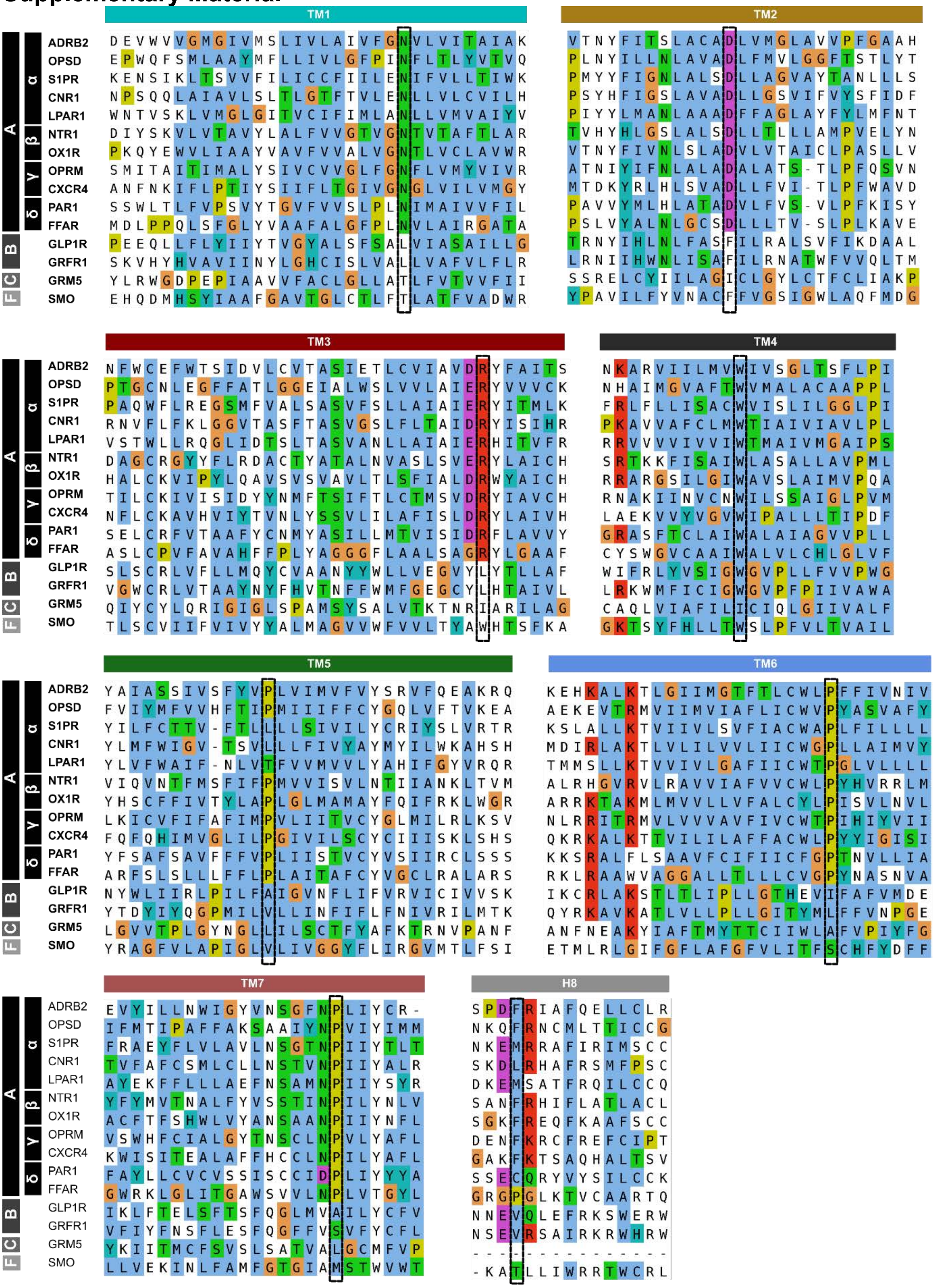
The UniProt C (2017) UniProt: the universal protein knowledgebase. Nucleic Acids Res 45: D158-D169.

Trzaskowski B, Latek D, Yuan S, Ghoshdastider U, Debinski A, *et al.,* (2012) Action of Molecular Switches in GPCRs - Theoretical and Experimental Studies. Current Medicinal Chemistry 19: 1090-1109.

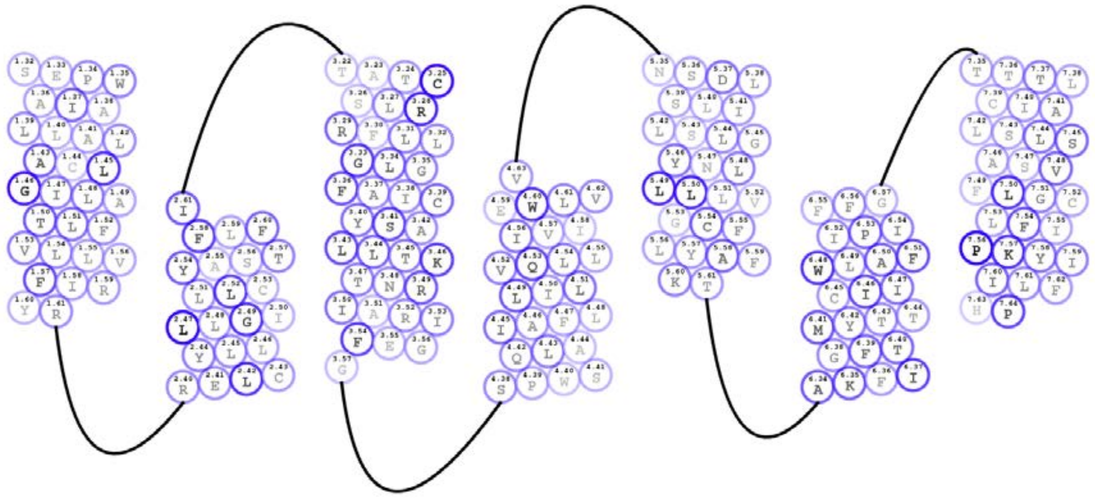
Venkatakrishnan AJ, Deupi X, Lebon G, Tate CG, Schertler GF, *et al.,* (2013) Molecular signatures of G-protein-coupled receptors. Nature 494: 185-194.

Vroling B, Sanders M, Baakman C, Borrmann A, Verhoeven S, *et al.,* (2011) GPCRDB: information system for G protein-coupled receptors. Nucleic Acids Res 39: D309-319.

Supplementary Material



S4.5-1 Fig. The sequence alignment of the TM segments and helix 8 of selected GPCRs from classes A, B, C and F. The alignment is colored using Clustal scheme. Receptor abbreviations are as in UniProt. Greek letters α, β, γ and δ represent the branches described by Fredriksson and collaborators for the class A. Receptors of the class C do not have helix 8 according to the presently available structures



S4.5-2 Fig. Snake-plot representation of the human Class C GPCRs. Each residue is represented with circle with a blue outline the one-letter amino acid code and the Ballesteros-and-Weinstein numbering scheme. The intensity of the blue outline represents the conservation of the position and the intensity of the grey-black letter represents the conservation of the specific amino acid relative to the consensus sequence.

4.6 HomolWat

A web server to incorporate internal water molecules into the structure of G-protein coupled receptors.

Abstract

Internal water molecules play important roles in stabilizing the structure of proteins, recognizing ligands and enabling conformational changes associated to protein function. Here we present a web application (HomolWat, available at http://lmc.uab.cat/HW), that implements a protocol for incorporating internal water molecules to GPCR structures based on the "homology" with similar structures. The application suits for both experimental structures lacking internal water molecules or computational models. HomolWat systematically allocates water molecules to a target structure from a database of crystallographic internal water molecules of high-resolution structures of homologous receptors. In our test cases, HomolWat provided similar recovery to presently available methods for hydrating protein structures such as Dowser. However, HomolWat has the advantage that waters introduced are internal water molecules whose existence has been proven experimentally. Furthermore, we show that as more experimental structures become available, HomolWat will be superior to Dowser+/++.

Introduction

G protein-coupled receptors (GPCRs) are the largest family of membrane proteins. Although GPCRs can recognize a wide diversity of extracellular signals, like hormones, neurotransmitters or entire proteins, the over 800 human GPCRs share a common 7 α-helical transmembrane (TM) architecture and similar conformation changes upon activation (Tehan *et al.*, 2014, Venkatakrishnan *et al.,* 2016). GPCRs are the target of more than 30% of drugs in the market (Hauser *et al.*, 2017), as they participate in many different cellular pathways. Since the first crystal structure of a GPCR appeared (PDBID 1F88, Palczewski *et al, 2000*), hundreds of structures have been released for different receptors, with different ligands, and also in different conformations (Cong *et al.*, 2017, Cvicek *et al.*, 2016). These structures have shed light into the transition from inactive to active conformation, still the mechanism is not completely understood. Anyhow, each conformation presents a different internal water network (Tehan *et al.,* 2014, Yuan *et al.,* 2014, Venkatakrishnan *et al.,* 2019)
Many studies have been devoted to understanding the role of these waters in the structure and function of GPCRs (Pardo *et al.,* 2007, Angel *et al.,* 2009 (A and B), Yuan *et al.,* 2013, Sun *et al.,* 2014). Recently, a conserved water-mediated polar interactions network inside GPCRs have been identified (Venkatakrishnan *et al.,* 2019). This water network plays a central role in receptor activation, differentiating among water network that is maintained across the inactive and the active states and water networks that rearranges upon activation. Thus, similarly to a conserved networks of amino acids contacts and ionic interactions in a common mechanism of activation among GPCRs, a network of water molecules is also conserved, especially in the lower half close to the G protein binding site.
The importance of water molecules in the structure and function of proteins, have given rise to an increasing number of new tools and methods to predict water placement, following different approaches, from geometrical scoring to Molecular Dynamics simulations (see Nittinger *et al.*, 2018 for a review), We present HomolWat (HW) a web server that incorporates internal water molecules and sodium ions to GPCR models. The incorporation of these molecules relies in a database of crystallographic GPCR structures with solved internal water molecules. HW takes advantage of the increasing number of high resolution X-ray crystal structures of GPCRs deposited in the Brookhaven Protein Data Bank (rcsb.org, Berman *et al.*, 2000) as the algorithms thrives on those solved water molecules in the core of this receptors. HomolWat places internal water molecules in the user's model according to homology with GPCRs in the database containing all GPCR structures released so far. A

comparison to Dowser+ (Morozenko *et al.,* 2014) and Dowser++ (Morozenko & Stuchebrukhov, 2016), two popular software for automatically hydrating protein structures, shows improved results of water molecule recovery. HW server can be accessed at: http://lmc.uab.cat/HW/

Methods

HW core is written in Python (v2.7). The HW web application interface has been developed using the Flask Web Framework (v1.02, http://flask.pocoo.org) with Bootstrap CSS public libraries (https://getbootstrap.com/). HW requires an input PDB file that the user needs to provide. Next, the user can select various options (incorporate sodium ion, prioritize active structures or use Dowser+ along with HW). The output is a PDB file that contains the input structure plus all the identified internal water molecules which can be visualized and downloaded. The web server relies on a regularly updated database of PDB crystallized GPCR structures.

Database of internal water molecules in GPCR (GPCRwatDB)

The coordinates of all GPCRs from Protein Data Bank were downloaded, resulting in 326 available structures from which around 280 crystals belong to class A structures, corresponding to 52 different receptors. From these, only those structures containing internal water molecules were kept. The database currently contains 162 chains from 122 GPCRs X-ray crystal structures containing internal water molecules and representing 31 unique receptors (see http://lmc.uab.cat/HW/gpcr_table).
To determine if the water molecules present in the downloaded structure were internal or not Circular Variance and B-factor was used. Circular variance (CV) (Mezei, 2003) is a measure of vectors from the oxygen atom of a water molecule to the surrounding atoms up to 10 Å. CV values range from 0 (completely exposed) to 1 (completely buried). B-factor or temperature factor is related to temperature-dependent atomic vibrations of the atoms and it reflects the fluctuation of an atom around its average position; large B-factor indicates high mobility of individual atoms and side chains while low B-factors indicates more ordered and stable parts of the structure (Yuan *et al.*, 2005). Internal water molecules included in the database were those with CV > 0.6 and B-factor < 45 Å^2. Molecules other than the orthosteric ligand, filtered waters and sodiums were cleared from the structures, along with nanobodies or other complementary proteins.

Incorporation of internal water molecules to the query model

HW protocol of water placement consist of three main steps: i) Identification of GPCR class A subfamily, ii) Superimposition of all GPCR crystals with internal waters to the target structure and iii) water molecule incorporation. The initial query model is provided by the user, as a PDB file with the coordinates of the GPCR, if water molecules, ions or ligands are present, they are maintained and prioritized.
The query sequence is aligned to the sequence of all GPCRs in the Database using NCBI-BLAST v2.6.0+ (Camacho *et al.*, 2009) in order to obtain a hierarchical list of structurally determined GPCRs sorted according to sequence identity. By default, all GPCR are considered for water placement, although it is also possible to restrict the incorporated waters to receptors with a minimum sequence identity with the target. Next, all selected GPCRs structures containing internal water molecules from the BLAST-derived hierarchical list, are structurally superimposed to the query model using PyMol (v2.0.5). Structural superimpositions are performed using sequence and structure-based modules from PyMol and only considering Cα atoms. The structural superimposition (sequence-based or structure-based) that produces the lowest root mean square deviation, is kept. By default, the protocol incorporates water molecules from inactive receptors before the active conformations, due to a lower representation of the formers in the database. Nonetheless, if the query conformation belongs to an active state, the active over the inactive structures can be prioritized just marking the ‘Active’ checkbox.

Afterwards, internal water molecules from superimposed structures are incorporated one by one in descending order from the hierarchical list, from highest to lowest resolution when more than one structure is available for the same receptor, and from lowest to highest B factor, when the water would not clash with atoms from query model or to previous incorporated water molecules. We used a cutoff distance of 2.4 Å (van Beusekom *et al.*, 2017). The hydrogen positions of the water molecules are then computed according to pdb2pqr (Dolinsky *et al.*, 2004) using the CHARMM forcefield (Vanommeslaeghe *et al.*, 2010). Only those water molecules able to perform hydrogen bond with the query structures are kept. If sodium checkbox was selected in the query, sodium is incorporated from the structure with the highest sequence identity and best resolution, with a cutoff distance to the sodium ion of 2.35Å (Mancinelli *et al.*, 2007). If Dowser+ option was checked, besides the addition of waters from crystals, Dowser+ is applied to the model. The list of waters provided from Dowser+ is compared with those obtained by HomolWat. Water molecules incorporated by Homolwat and Dowser+ are labelled as coincident if the two waters are within 2Å. Non coincident waters molecules are all added to the query model. For coincident water molecules, the one (from the crystal structures or from Dowse+) which is able to perform more hydrogen bonds is selected. In case of identical number of hydrogen bonds, the water molecule from HomolWat is prioritized.

The final model is shown using a NGL (Rose & Hildebrand, 2015) embedded viewer and/or can be downloaded. The Uniprot name and the PDB id of the structures that provided each incorporated water molecule are written in the final PDB, in columns 80-89.

HomolWat Validation

To validate the ability of the program to predict the position of structural water molecules, we evaluated the percentage of recovery as a ratio of the number of water molecules predicted and the total number of internal waters for 18 GPCR class A. The selected structures had a resolution <2.8Å and contained a minimum of 5 internal water molecules. Allocated molecules were considered as recovered if the distance to an experimental resolved water was <2Å. This is the same criteria used in previous predictions (Morozenko & Stuchebrukhov, 2016, Lai et al., 2017). The number of internal waters in the initial crystal structure was used to compute the percentage of recovery for each method. Initial water molecules, ions and other co-crystallized molecules were deleted from the initial crystal structures, although the orthosteric ligand was kept. Next, either HW, Dowser+ or Dowser++ were used to incorporate internal waters without considering water molecules from the tested crystal structure.

Table 4.6-1 and Figure 4.6-1 show the recovery of water molecules for the tested 18 query structures. Comparison to Dowser+ and Dowser++ is also shown. The median for the percentage of recovery for HW in the 18 structures is over 75%, very similar to Dowser+ and slightly better than Dowser++. The receptor for which HW performs better is Adenosine receptor A_{2A} (AA2AR), the receptor with the highest amount of internal waters to recover. This can be explained by the fact that many structures of AA2A receptors have been resolved at high-resolution and they largely contribute to the total number of internal water molecules in the database (see table S4.6-1 for a complete list of the number of crystals per receptor, chains and the number of water molecules with which they contribute). The structures with the least recovery rate (~45%) are EDNRB and CCR5. For the ENDRB the most probable reason for the low recovery is the lack of close phylogenetic relatives in the database for these receptors. Likewise, for CCR5 the reason is that many of the internal water molecules present in the structure with PDB id 5UIW are in the interface between CCR5 and the chemokine ligand (CCL5 variant) and 5UIW is the only structure in complex with a chemokine ligand that contributes water molecules Dowser+ and Dowser++ perform much better in both cases. Regarding the amount of internal water molecules incorporated to each model, HW allocates a median of 80 molecules, similarly to Dowser programs after the CV filtering, as these methods also incorporate water molecules in the surface exposed to lipids. These values agree with recent experimental data that determined that rhodopsin activation is coupled to a bulk influx of 80-100 water molecules into the protein core (Fried *et al.*, 2019).

crystal			HomolWat			Dowser+			Dowser++		
receptor	pdb	internal waters	waters	recov.	(%)	waters	recov.	(%)	waters	recov.	(%)
OPSD	1GZM	13	68	11	84.6	72	12	92.3	61	11	84.6
CXCR4	3ODU	17	91	13	76.5	79	13	76.5	93	12	70.6
PAR1	3VW7	21	62	13	61.9	68	15	71.4	62	14	66.7
ADRB1	4BVN	20	91	17	85.0	80	16	80.0	92	13	65.0
OPRM	4DKL	5	88	3	60.0	76	3	60.0	68	2	40.0
OPRD	4N6H	28	84	23	82.1	67	20	71.4	64	17	60.7
P2Y12	4PXZ	7	70	4	57.1	87	6	85.7	79	5	71.4
AA2AR	5IU4	53	101	53	100.0	65	38	71.7	63	39	73.6
FFAR1	5TZR	10	62	10	100.0	50	9	90.0	58	9	90.0
CNR1	5U09	6	66	4	66.7	75	5	83.3	76	2	33.3
CCR5	5UIW	23	66	10	43.5	83	19	82.6	72	14	60.9
DRD4	5WIU	16	86	13	81.3	73	8	50.0	70	11	68.8
OX2R	5WQC	32	97	26	81.3	85	22	68.8	70	25	78.1
ACM2	5YC8	15	87	12	80.0	74	14	93.3	78	13	86.7
C5AR1	6C1R	32	94	24	75.0	73	23	71.9	82	27	84.4
CCR2	6GPX	10	75	6	60.0	50	6	60.0	49	4	40.0
NK1R	6HLP	7	65	6	85.7	59	5	71.4	59	6	85.7
EDNRB	6IGK	31	83	14	45.2	101	28	90.3	139	27	87.1
	median	17	84	13	**75.8**	74	14	**74.2**	70	13	**71.0**

Table 4.6-1: Recovery of water molecules with different methods (HW, Dowser+ and Dowser++) in the test set. For each method the table lists the total number of waters (waters), the number of recovered water molecules considering a cutoff distance of 2.0 Å (recov.) and its percentage. The bottom line corresponds to the median of each column.

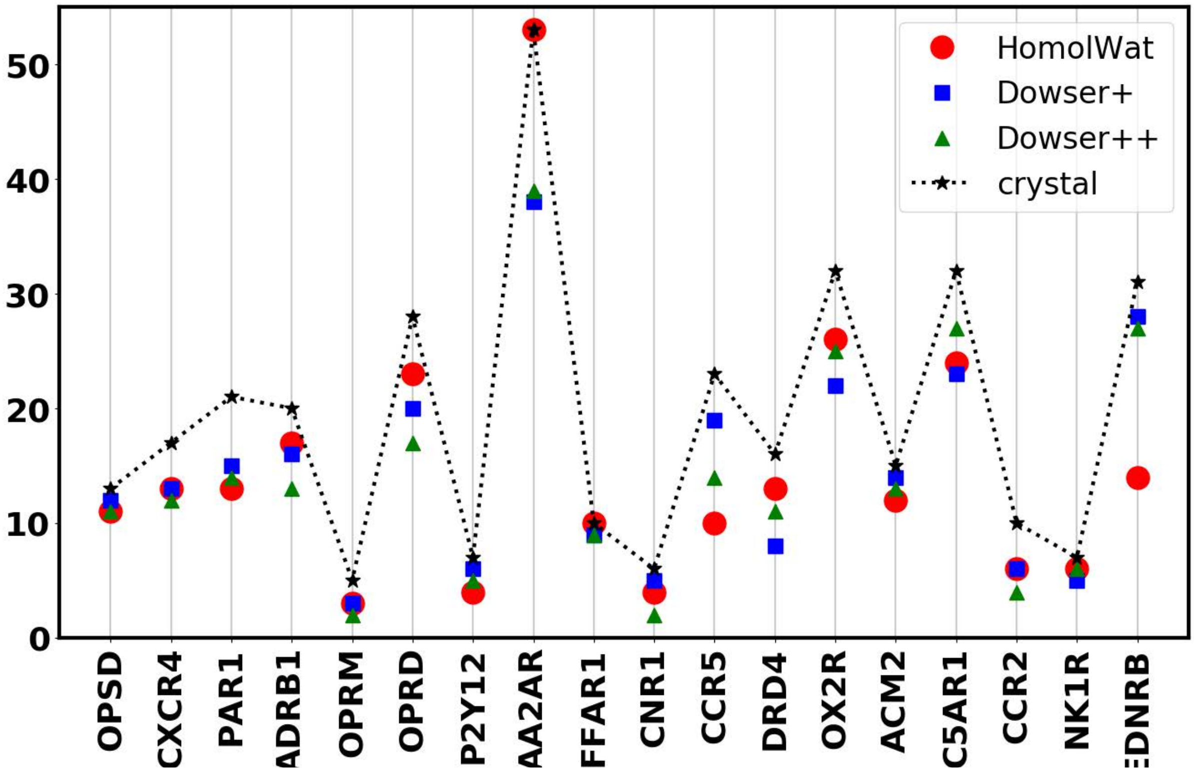


Figure 4.6-1: Number of internal water molecules recovered by each method. The number of internal waters in the original crystal (black star, see Table 4.6-1 for the associated PDB ids), and waters recovered by HomolWat (red dot), Dowser+ (blue square) and Dowser++ (green triangle).

Figure 4.6-2 shows the evolution of HW performance, using the accumulated available waters for each year. Dark blue bars display the amount of internal waters obtained each year, whereas light blue bars show the accumulated number of waters along these years. Starting in 2007, when only Rhodopsin receptors with internal waters were available and finishing with last update of the database corresponding to February of 2019. This rising evolution is based on the increasing number of high-resolution GPCRs deposited in the PDB. Although, since 2016, with about 300 new internal water molecules per year, the recovery ratio has seen its rise reduced. This reduction may be explained by the location of these new water molecules. The internal cavities of GPCRs have a limited space and the available waters already occupy it (see Figure S4.6-1). Thus, future improvement of HW algorithm rely on the release of structures of receptors not crystalized so far along with high-resolution structures of current crystallized receptors with resolved internal water molecules not present in current structures.

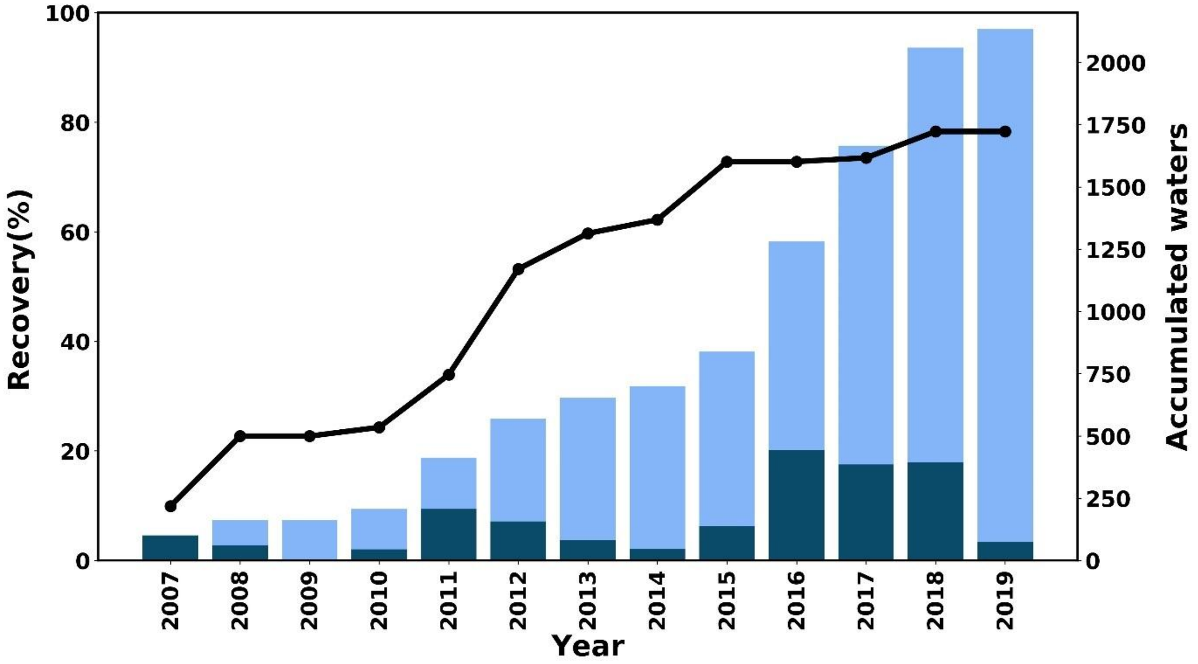


Figure 4.6-2: Simulated evolution of the percentage of recovery (line) and the number of internal waters resolved each year (blue bars) and the accumulated number of waters over the years (grey bars). Recovery of HW method over time for the median of 18 tested query structures at a cutoff distance of 2.0. As the number of determined X-ray crystal structures increase, hence the number of internal waters, the percentage of recovery also increases.

Example of use

To illustrate the use of HomolWat, we cover two different cases: i) the structure of the cannabinoid receptor 2 (CNR2) with PDB id 5ZTY (Li *et al.*, 2019), and ii) the structure of the adenosine A2A receptor (AA2AR) with PDB id 5NM4 (Weinert *et al.*, 2017). The CNR2 structure with PDB id 5ZTY has recently been obtained and does not contain any internal water. This is the only CNR2 structure available so far, and there is a single structure of the related CNR1 receptor contributing 6 internal water molecules only. The AA2AR structure with PDB id 5NM4 has a relatively 'low' content of internal waters. However, there are >30 structures available of the AA2AR, many of them with many internal water molecules (see Table S4.6-1). Figure 4.6-3 shows the structures of the CNR2 and AA2AR prior and after adding the molecules with HW. If the model used has any internal water, HW incorporates water molecules filling all the protein interior, respecting the position occupies by the ligand. Moreover, HW offers the possibility of incorporating internal sodium conserved in many inactive GPCR structures. However, if water molecules or ions are present in the query structure, HW maintains these ions and molecules and incorporates more water molecules until any more water molecule of the database can be incorporated to the model.

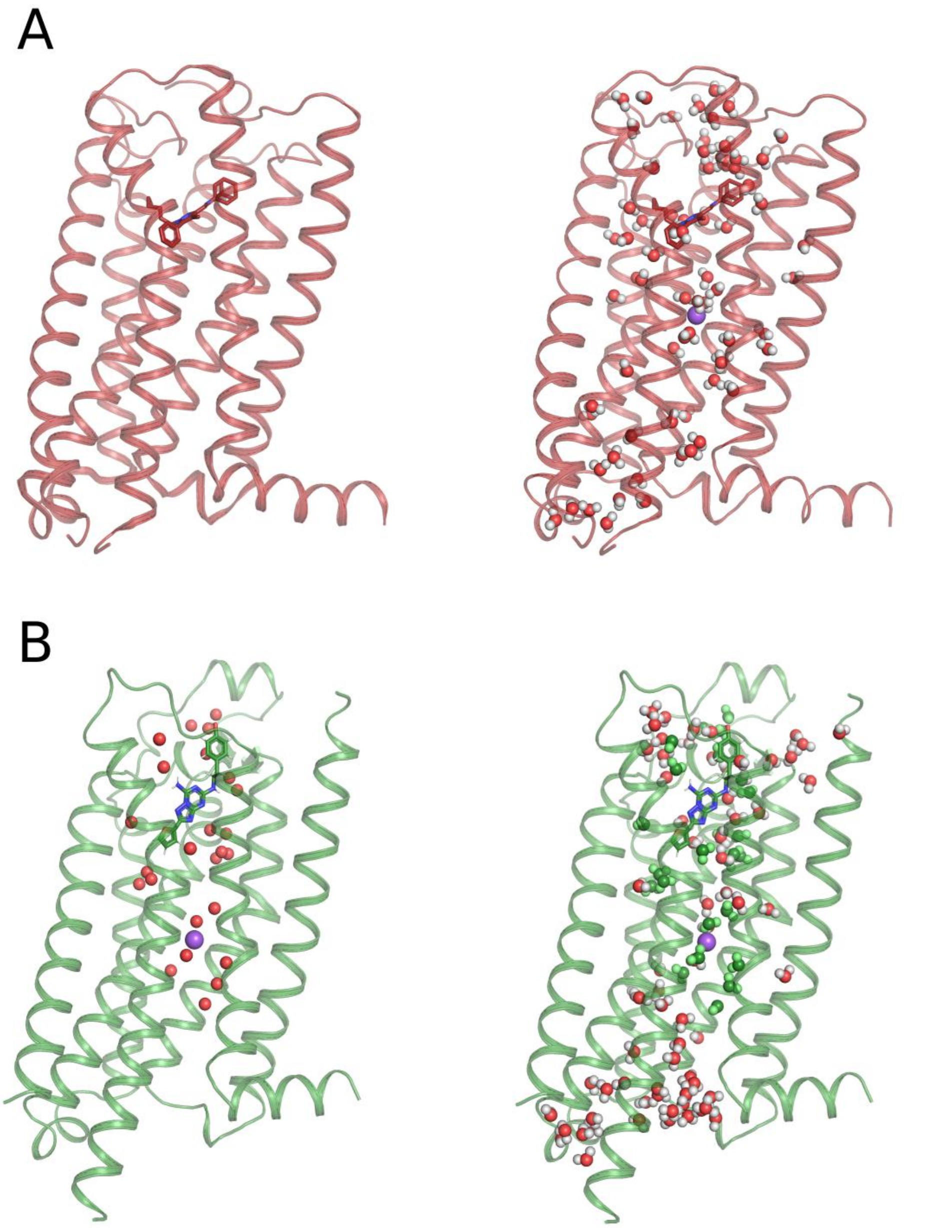


Figure 4.6-3. GPCR models before (left) and after (right) applying HW protocol for CNR2 (PDBID 5ZTY) (A) and for AA2AR (PDBID 5NM4) (B). Water molecules are represented as ball-and-stick, with the original oxygen waters coloured in green in the HW models (right).

Conclusions

We have developed a tool (HomolWat) that is able to introduce internal water molecules in given structures using the resolved water molecules present in experimental structures of members of the same family deposited in the PDB. The tool is available via a web application that uses an up-to-date, curated database of GPCR structures with internal water molecules to allocate them in GPCR models or experimental structures with few or none internal water molecules. The algorithm of water placement has been validated and compared to other methods such as Dowser+ and Dowser++, with HomolWat achieving slightly superior recovery ratios. The foreseeable increase in the number of resolved high-quality GPCR structures predicts that HomolWat will perform even better in the future. Better GPCR models that explicitly introduce internal water molecules may for instance improve docking calculations or MD simulations.

References

Angel, T. E., Chance, M. R., & Palczewski, K. (2009). Conserved waters mediate structural and functional activation of family A (rhodopsin-like) G protein-coupled receptors. *Proceedings of the National Academy of Sciences*, *106*(21), 8555-8560. (A)

Angel, T. E., Gupta, S., Jastrzebska, B., Palczewski, K., & Chance, M. R. (2009). Structural waters define a functional channel mediating activation of the GPCR, rhodopsin. *Proceedings of the National Academy of Sciences*, *106*(34), 14367-14372.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). The protein data bank. *Nucleic acids research*, *28*(1), 235-242.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: architecture and applications. *BMC bioinformatics*, *10*(1), 421.

Cong, X., Topin, J., & Golebiowski, J. (2017). Class A GPCRs: structure, function, modeling and structure-based ligand design. *Current pharmaceutical design*, *23*(29), 4390-4409.

Cvicek V, Goddard WA, 3rd, Abrol R (2016) Structure-Based Sequence Alignment of the Transmembrane Domains of All Human GPCRs: Phylogenetic, Structural and Functional Implications. PLoS Comput Biol 12: e1004805.

Dolinsky, T. J., Nielsen, J. E., McCammon, J. A., & Baker, N. A. (2004). PDB2PQR: an automated pipeline for the setup of Poisson–Boltzmann electrostatics calculations. *Nucleic acids research*, *32*(suppl_2), W665-W667.

Hauser AS, Attwood MM, Rask-Andersen M, Schioth HB, Gloriam DE (2017) Trends in GPCR drug discovery: new agents, targets and indications. Nat Rev Drug Discov 16: 829-842.

Fried, S. D., Eitel, A. R., Weerasinghe, N., Norris, C. E., Vos, M. R., Somers, J. D., Fitzwater, G. I., Pitman, M. C., Struts, A. V., Perera, S. M. D. C. & Brown, M. F. (2019). Hydration Modulates G-Protein-Coupled Receptor Signaling. *The FASEB Journal*, *33*(1_supplement), 462-1.

Lai, J. K., Ambia, J., Wang, Y., & Barth, P. (2017). Enhancing structure prediction and design of soluble and membrane proteins with explicit solvent-protein interactions. *Structure*, *25*(11), 1758-1770.

Li, X., Hua, T., Vemuri, K., Ho, J. H., Wu, Y., Wu, L., Popov, P., Benchama, O., Zvonok, N., Locke, K., Qu, L., Han, G. W., Iyer, M. R., Cinar, R., Coffey, N. J., Wang, J., Wu, M., Katritch, V., Zhao, S., Kunos, G., Bohn, L. M., Makriyannis, A., Stevens, R. C. & Liu, Z. (2019). Crystal Structure of the Human Cannabinoid Receptor CB2. *Cell*, *176*(3), 459-467.

Mancinelli, R., Botti, A., Bruni, F., Ricci, M. A., & Soper, A. K. (2007). Hydration of sodium, potassium, and chloride ions in solution and the concept of structure maker/breaker. *The Journal of Physical Chemistry B*, *111*(48), 13570-13577.

Mezei, M. A new method for mapping macromolecular topography. J Mol Graph Model 2003, 21(5), 463-472.

Morozenko, A., Leontyev, I. V., & Stuchebrukhov, A. A. (2014). Dipole moment and binding energy of water in proteins from crystallographic analysis. *Journal of chemical theory and computation*, *10*(10), 4618-4623.

Morozenko, A., & Stuchebrukhov, A. A. (2016). Dowser++, a new method of hydrating protein structures. *Proteins: Structure, Function, and Bioinformatics*, *84*(10), 1347-1357.

Nittinger, E., Flachsenberg, F., Bietz, S., Lange, G., Klein, R., & Rarey, M. (2018). Placement of Water Molecules in Protein Structures: From Large-Scale Evaluations to Single-Case Examples. *Journal of chemical information and modeling*, *58*(8), 1625-1637.Morozenko & Stuchebrukhov 2016),

Palczewski, K., Kumasaka, T., Hori, T., Behnke, C. A., Motoshima, H., Fox, B. A., Le Trong, I., Okada, T., Stenkamp, R. E., Yamamoto, M. & Miyano M. (2000). Crystal structure of rhodopsin: AG protein-coupled receptor. *science*, *289*(5480), 739-745.

Pardo, L., Deupi, X., Dölker, N., López-Rodríguez, M. L., & Campillo, M. (2007). The role of internal water molecules in the structure and function of the rhodopsin family of G protein-coupled receptors. *ChemBioChem*, *8*(1), 19-24.

Rose, A. S., & Hildebrand, P. W. (2015). NGL Viewer: a web application for molecular visualization. *Nucleic acids research*, *43*(W1), W576-W579.

Sun, X., Ågren, H., & Tu, Y. (2014). Functional water molecules in rhodopsin activation. *The journal of physical chemistry B*, *118*(37), 10863-10873.

The PyMol Molecular Graphics System, Version 2.0.5 Schrödinger, LLC

Tehan, B. G., Bortolato, A., Blaney, F. E., Weir, M. P., & Mason, J. S. (2014). Unifying family A GPCR theories of activation. *Pharmacology & therapeutics*, *143*(1), 51-60.

van Beusekom, B., Touw, W. G., Tatineni, M., Somani, S., Rajagopal, G., Luo, J., illiland, G. I., Perrakis, A. & Joosten, R. P. (2018). Homology-based hydrogen bond information improves crystallographic structures in the PDB. *Protein Science*, *27*(3), 798-808.

Vanommeslaeghe, K., Hatcher, E., Acharya, C., Kundu, S., Zhong, S., Shim, J., Darian, E., Guvench, O., Lopes, P., Vorobyov, I. & MacKerell Jr, A. D. (2010). CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *Journal of computational chemistry*, *31*(4), 671-690.

Venkatakrishnan, A. J., Deupi, X., Lebon, G., Heydenreich, F. M., Flock, T., Miljus, T., Balaji, S., Bouvier, M., Veprintsev, D. B., Tate, C. G., Schertler, G. F. & Babu, M. M. (2016). Diverse activation pathways in class A GPCRs converge near the G-protein-coupling region. *Nature*, *536*(7617), 484.

Venkatakrishnan, A. J., Ma, A. K., Fonseca, R., Latorraca, N. R., Kelly, B., Betz, R. M., Asawa, C., Kobilka, B. K. & Dror, R. O. (2019). Diverse GPCRs exhibit conserved water networks for stabilization and activation. *Proceedings of the National Academy of Sciences*, *116*(8), 3288-3293.

Weinert, T., Olieric, N., Cheng, R., Brünle, S., James, D., Ozerov, D., Gashi, D., Vera, L., Marsh, M., Jaeger, k., Dworkowski, F., Panepucci, E., Basu, S., Skopintsev, P., Doré, A. S., Geng, T., Cooke, R. M., Liang, M., Prota, A. E., Panneels, V., Nogly, P., Ermler, U., Schertler, G., Hennig, M., Steinmetz, M. O., Wang, M. & Standfuss, J. (2017). Serial millisecond crystallography for routine room-temperature structure determination at synchrotrons. *Nature communications*, *8*(1), 542.

Yuan, Z., Bailey, T. L., & Teasdale, R. D. (2005). Prediction of protein B-factor profiles. *Proteins: Structure, Function, and Bioinformatics*, *58*(4), 905-912.

Yuan, S., Vogel, H., & Filipek, S. (2013). The Role of Water and Sodium Ions in the Activation of the μ-Opioid Receptor. *Angewandte Chemie International Edition*, *52*(38), 10112-10115.

Yuan, S., Filipek, S., Palczewski, K., & Vogel, H. (2014). Activation of G-protein-coupled receptors correlates with the formation of a continuous internal water pathway. *Nature communications*, *5*, 4733.

Supplementary information

receptor	code	Crystals	(%)	Chains	(%)	Waters	(%)
Adenosine receptor A2a	AA2AR	29	23.8	30	18.5	996	46.7
Rhodopsin	OPSD	21	17.2	34	21.0	290	13.6
Adrenergic receptor β1	ADRB1	18	14.8	36	22.2	288	13.5
Endothelin receptor B	EDNRB	4	3.3	4	2.5	59	2.8
C5a anaphylatoxin chemotactic receptor 1	C5AR1	2	1.6	3	1.9	48	2.3
Muscarinic acetylcholine receptor M2	ACM2	5	4.1	5	3.1	45	2.1
Orexin receptor 2	OX2R	2	1.6	2	1.2	44	2.1
CXC chemokine receptor 4	CXCR4	2	1.6	3	1.9	37	1.7
Angiotensin receptor 1	AGTR1	1	0.8	2	1.2	32	1.5
Free fatty acid receptor 1	FFAR1	3	2.5	3	1.9	32	1.5
Mu opioid receptor	OPRM	2	1.6	2	1.2	30	1.4
Adrenergic receptor β2	ADRB2	6	4.9	6	3.7	28	1.3
Delta opioid receptor	OPRD	1	0.8	1	0.6	28	1.3
CC chemokine receptor 5	CCR5	2	1.6	2	1.2	24	1.1
Proteinase-activated receptor 1	PAR1	1	0.8	1	0.6	21	1.0
Dopamine D4 receptor	DRD4	2	1.6	2	1.2	20	0.9
CC chemokine receptor 2	CCR2	2	1.6	3	1.9	19	0.9
GPCR homolog US28	US28	1	0.8	1	0.6	15	0.7
P2Y purinoreceptor 1	P2RY1	2	1.6	3	1.9	13	0.6
Tachykinin receptor 1	NK1R	2	1.6	2	1.2	10	0.5
Muscarinic acetylcholine receptor M4	ACM4	1	0.8	2	1.2	10	0.5
CC chemokine receptor 9	CCR9	1	0.8	2	1.2	9	0.4
P2Y purinoreceptor 1	P2Y12	2	1.6	2	1.2	8	0.4
Cannabinoid receptor 1	CNR1	1	0.8	1	0.6	6	0.3
Proteinase-activated receptor 2	PAR2	2	1.6	2	1.2	5	0.2
Nociceptin receptor	OPRX	1	0.8	2	1.2	5	0.2
5-Hydroxytryptamine receptor 2B	5HT2B	2	1.6	2	1.2	3	0.1
Lysophospholipid acid receptor 1	LPAR1	1	0.8	1	0.6	2	0.1
Apelin receptor	APJ	1	0.8	1	0.6	2	0.1
Sphingosine 1-phosphate receptor 1	S1PR1	1	0.8	1	0.6	2	0.1
Neurotensin receptor 1	NTR1	1	0.8	1	0.6	2	0.1
		122		162		2133	

Table S4.6-1. List of crystalized receptors with internal waters, number of crystals deposited in the PDB, number of Chains for each receptor and number of internal waters with which each receptor contributes to the database with the percentage respect the total. For a detailed view of each chain see http://lmc.uab.cat/HW/gpcr_table.

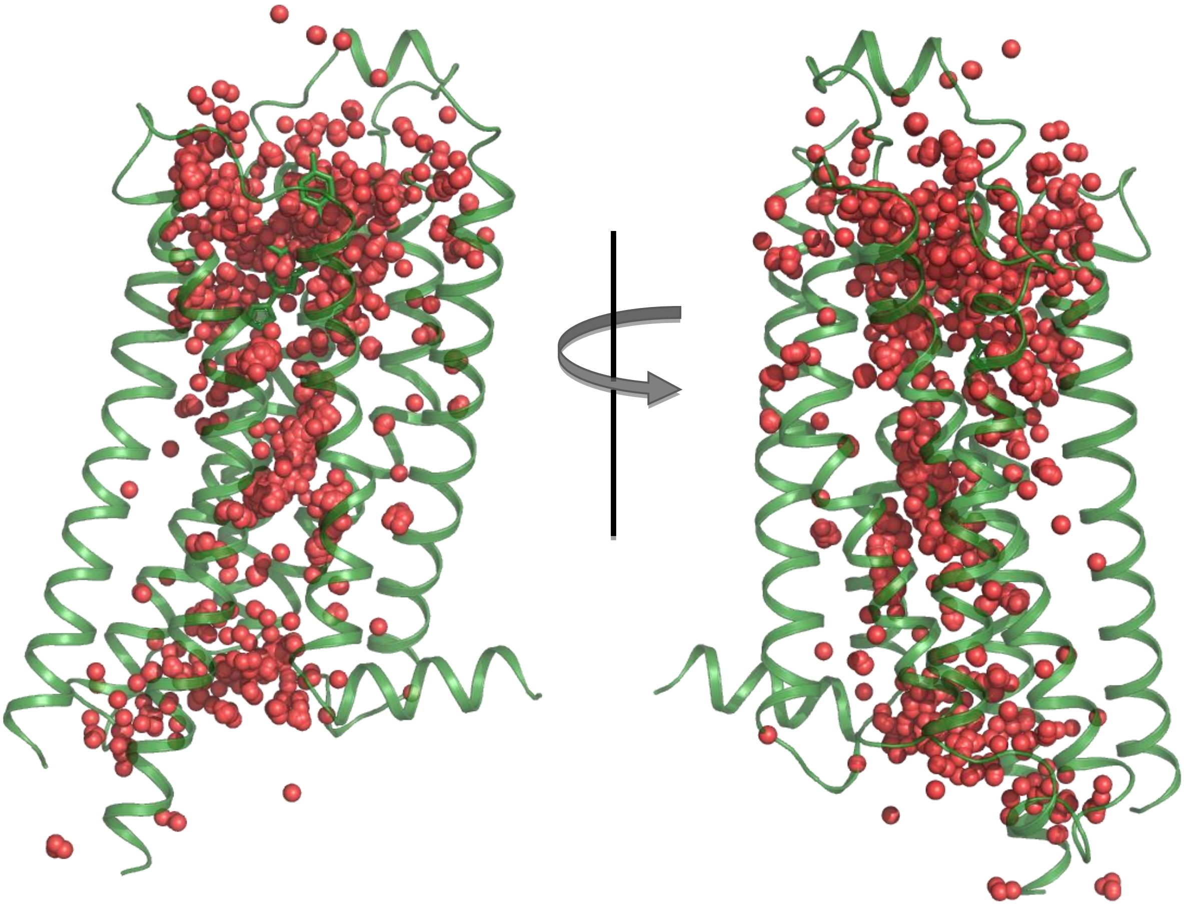


Figure S4.6-1. A representing structure of a GPCR with all internal waters (represented as red spheres) crystallized until February 2019. Shown are two views from the membrane rotated 180º .

5. Conclusions

5. Conclusions

1) Two non-redundant databases of TM segments of alpha and beta membrane protein structures (TMalphaDB and TMbetaDB) have been developed.

2) The analysis of TMalphaDB and TMbetaDB, has revealed specific patterns of in terms of inter- residue composition and inter-residue interactions of alpha helical segments compared to beta-barrel membrane proteins and of alpha-helical globular.

3) In alpha-helical proteins, there are preferences of amino acids to be in the protein core or exposed to the lipid bilayer and to occupy specific positions along the TM segment.

4) Analysis of inter-residue interactions in TM segments of membrane proteins shows that almost all interactions involve aliphatic residues and Phe, whereas, there is lack of polar-polar, polar-charged and charged-charged interactions.

5) Met and Cys often interact with Leu, Ile, Val, Phe, and other Met or Cys. The characterization of their strength using *ab-initio* calculations in small-molecule model systems, predicted that Met-Met, Met-Phe, Cys-Phe, Met-aliphatic and Cys-aliphatic interactions are stronger in magnitude than aliphatic-aliphatic interactions.

6) A web application tool to quantify structural distortions induced by specific sequence motifs in membrane proteins, and to elucidate their role in the 3D structure have been developed. This specific structural information has direct implications in homology modelling of the growing sequences of membrane proteins lacking experimental structure.

7) GPCRS-SAS, a web application that performs frequency, covariance and correlation analyses for sequence positions or motifs in GPCRs has been developed. The tool takes advantage of the structural similarity in the TM domain of GPCRs and allows performing comparisons and statistical analyses of sequence positions or motifs within the TM helices and helix 8 for receptors of classes A, B, C and F.

8) Homolwat, a web application able to introduce internal water molecules in GPCR structures using based on internal water molecules experimentally determined in homologous structures has been developed.