MODELING HOMO- AND HETERO-OLIGOMERS USING *IN SILICO* PREDICTION OF PROTEIN QUATERNARY STRUCTURE

Inauguraldissertation
zur
Erlangung der Würde eines Doktors der Philosophie
vorgelegt der
Philosophisch-Naturwissenschaftlichen Fakultät
der Universität Basel

von

Martino Bertoni

aus Italien

2017, Basel

Genehmigt von der Philosophisch-Naturwissenschaftlichen Fakultät auf Antrag von

Fakultätsverantwortliche:
Prof. Dr. Torsten Schwede

Korreferent:
Prof. Dr. Christian von Mering

Basel, 13.12.2016

Prof. Dr. Jörg Schibler
Dekan

“It is good to have an end to journey toward;
but it is the journey that matters, in the end.”

Ursula K. Le Guin

ABSTRACT

Cellular processes often depend on interactions between proteins and the formation of macromolecular complexes. The impairment of such interactions can lead to deregulation of pathways resulting in disease states, and it is hence crucial to gain insights into the nature of the macromolecular assemblies. Detailed structural knowledge about complexes and protein-protein interactions is growing, but experimentally determined three-dimensional multimeric assemblies are outnumbered by complexes supported by non-structural experimental evidence.

In this thesis, we aim to fill this gap by modeling multimeric structures by homology, and we ask which properties of proteins within a family can assist in the prediction of the correct quaternary structure. Specifically, we introduce a description of protein-protein interface conservation as a function of evolutionary distance. This enables us to reduce the noise in deep multiple sequence alignments where sequences of proteins organized in different oligomeric states are interspersed. We also define a distance measure to structurally compare homologous multimeric protein complexes. This allows us to hierarchically cluster protein structures and quantify the diversity of alternative biological assemblies known today in the Protein Data Bank (PDB). We find that a combination of conservation scores, structural clustering, and classical interface descriptors, is able to improve the selection of homologous protein templates leading to reliable models of protein complexes.

CONTENTS

1 INTRODUCTION 1
- 1.1 Proteins 1
  - 1.1.1 Protein structure 1
- 1.2 Experimental structure determination 10
- 1.3 Protein structure prediction 13
  - 1.3.1 Template based modeling 14
  - 1.3.2 Template free modeling 16
  - 1.3.3 Critical Assessment of protein Structure Prediction: CASP 17
- 1.4 Modeling protein-protein interactions 17
  - 1.4.1 Template free docking 18
  - 1.4.2 Template based docking 18
  - 1.4.3 Critical Assessment of Predicted Interactions: CAPRI 19
- 1.5 Thesis aim 19

2 STRUCTURAL SIMILARITY OF PROTEIN COMPLEXES 21
- 2.1 Methods 21
  - 2.1.1 Comparing quaternary structures: QS-score 21
- 2.2 Results 25
  - 2.2.1 Structural similarity in homologous complexes 25
- 2.3 Discussion 26

3 CONSERVATION OF PROTEIN INTERFACES 29
- 3.1 Methods 30
  - 3.1.1 Conservation score 30
  - 3.1.2 PPI fingerprint 32
- 3.2 Results 33
  - 3.2.1 Discriminating crystal contacts vs. biological contacts 33
  - 3.2.2 PPI fingerprint of homologs 35
- 3.3 Discussion 36

4 MODELING OLIGOMERS 38
- 4.1 Methods 38
  - 4.1.1 Template search 38
  - 4.1.2 Template clustering 39
  - 4.1.3 Template ranking 40
- 4.2 Results 49
  - 4.2.1 Template ranking by interface quality prediction 49

4.3 Case studies 50
4.3.1 Modeling fructose bisphosphate aldolase in *Haloferax volcanii* 50
4.3.2 Modeling the urease complex in *Yersinia enterocolitica* 52
4.4 Discussion 57
5 SWISS-MODEL: AUTOMATED OLIGOMERIC MODELING 59
5.1 Methods 59
5.1.1 Oligomeric state prediction 59
5.2 Results 62
5.2.1 Comparison with other modeling servers 62
5.3 Discussion 64
6 CONCLUSION AND OUTLOOKS 66

REFERENCES 68

ACKNOWLEDGMENTS 87

LIST OF FIGURES

Figure 1 Voronoi tree diagram of the macromolecular composition of an *E. coli* 2
Figure 2 ω, ϕ, and ψ dihedral angles 4
Figure 3 Ramachandran plot of the ϕ,ψ protein backbone dihedral angles 5
Figure 4 Showcase of common symmetries in homo-oligomers 7
Figure 5 Classification of protein-protein interactions 9
Figure 6 Example of QS-score for a pair of distances d_1 and d_2. 24
Figure 7 Examples of QS-score comparisons 25
Figure 8 Heterogeneity of quaternary structures available in the PDB repository 27
Figure 9 PPI fingerprint calculation 29
Figure 10 Distribution of interface-surface ratio in random patches 32
Figure 11 PPI fingerprint for conserved homo-dimers 34
Figure 12 PPI fingerprints of the proteins in the Duarte *et al.* dataset 35
Figure 13 PPI fingerprint of fructose bisphosphate aldolase homologs 36
Figure 14 Clustering scheme for homologous assemblies 40
Figure 15 Stoichiometries of target proteins in our TARGET dataset 42
Figure 16 QS-score distribution for all produced models compared to the native structure 44
Figure 17 Distribution of mostly correct and mostly incorrect models 46
Figure 18 Grid search for C and γ parameters 48
Figure 19 Fraction of validation targets in each quality category for top ranked models 49
Figure 20 Fraction of validation targets in each quality category for top ranked models using single features 51
Figure 21 Structural clustering tree of fructose bisphosphate aldolase homologs 53

Figure 22 PPI fingerprint curves of fructose bisphosphate aldolase homologs 54
Figure 23 Urease symmetries and genetic organization 55
Figure 24 Performances of the naïve and logistic regression classifiers 61
Figure 25 ROC analysis of the naïve and logistic regression classifiers 62
Figure 26 Comparison of model quality for three servers participating in CAMEO 64
Figure 27 Example of transitive complex modeling 67

LIST OF TABLES

Table 1 Interface distance measures developed in the last few years 22
Table 2 Analysis of fusion events with the queried *Y. enterocolitica* sequences 56
Table 3 Comparison of the *Y. enterocolitica* X-ray, electron-microscopy and homology model urease structures 57
Table 4 Summary of the modeling performances of SWISS-MODEL Oligo, SWISS-MODEL, and Robetta 63

1 INTRODUCTION

1.1 PROTEINS

Proteins are structural bricks, functional gears, and information mediators that, forged by evolution, enables life as we know it. The study of proteins is hence crucial for the comprehension of the vital processes in any living being. The larger fraction of cellular dry mass is composed of proteins (Figure 1), making them the dominant player in cells.

The secret of their evolutionary success lies in their extreme modularity and in the multifariousness of functions and structures they can perform and assume. Indeed, it is often the tridimensional structure of these chains of amino acids that determines their functioning. It is thus critical to determine the native structure of proteins, pushing for atomic resolution, to fully understand their mechanisms of action. Furthermore, setting up experiments aimed at describing proteins functioning - like mutagenesis on specific sites, mapping disease related polymorphism, or designing specific inhibitors - is greatly aided by the knowledge of the spatial organization and relative orientation of atoms, residues, and polypeptide chains in the protein 3D structures.

1.1.1 *Protein structure*

1.1.1.1 *Primary structure: amino acids*

Proteins are polymers, linear chains combining different modular element called amino acids or residues. The aminoacidic sequence of a protein is referred to as the primary structure of the protein. As indicated by the name, all amino acids are composed of two chemical groups, a positively charged amine ($-NH_2$) and a negatively charged carboxylic acid ($-COOH$). The amine nitrogen (N) and the carbonyl carbon (C) both interact with a central α-carbon (Cα).

Along with these shared groups, covalently linked to the α-carbon, is a third group: the side-chain. This variable group defines the identity and chemical properties of each amino acid, e.g. polarity, hydrophobicity, charge, and steric hindrance.

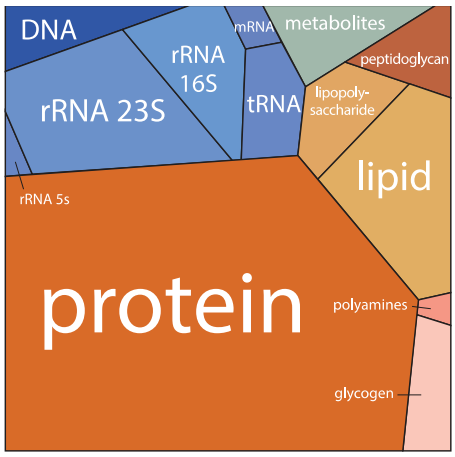
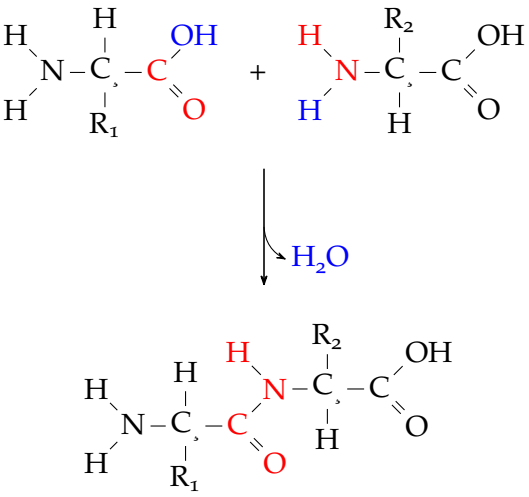


Figure 1: Voronoi tree diagram of the macromolecular composition of an *E. coli* cell growing with a doubling time of 40 min. Each polygon area represents the relative fraction of the corresponding constituent in the cell dry mass. Colors are associated with each polygon such that components with related functional role have similar tints. The Voronoi tree diagram visualization method was developed in order to represent whole genome measurements from micro-arrays or proteome quantitation. Image from http://book.bionumbers.org [1]

Twenty standard amino acids are encoded by triplet codons in the genetic code. The central asymmetric α-carbon induces the chirality of amino acids, so amino acids do not have an inversion plane nor can be superposed mirroring them. All amino acid found in proteins are in the L-configuration (left handed), while natural D-configuration (right handed) amino acids are important for bacterial cell walls or act as brain neurotransmitter.

The two components, basic and acidic, allow the formation of characteristic bond between two amino acids: the peptide bond (Reaction 1). After a condensation reaction, the carbonyl carbon of a first amino acid is covalently bound with the nitrogen of the subsequent amino acid. This bond is a very stable and planar covalent bond. The sequence of $[N - C\alpha - C]_n$ compose the backbone of the protein and is described by the dihedral angle ω between the planes defined by the $N_i - C\alpha_i - C_i$ and $C\alpha_i - C_i - N_{i+1}$ atoms. This dihedral can theoretically assume the cys ($\omega = 0°$) or trans ($\omega = 180°$) conformation, the latter having a favorable energy state due to the steric hindrance of the side-chains that fit better alternating the directionality.

Reaction 1 Condensation reaction forming the peptide bond between two generic amino acids with R_1 and R_2 side-chains



Peptide bond

1.1.1.2 *Secondary structure: α-helix and β-sheet*

Being the ω dihedral fixed, the real contribution in term of degree of freedom for proteins backbone is coming from rotations around the ϕ $[C_{i-1} - N_i - C\alpha_i - C_i]$ and ψ $[N_i - C\alpha_i - C_i - N_{i+1}]$ dihedrals as represented in Figure 2.

The term secondary structure refers to some particular repetitive arrangements of local short sections of the backbone. Still, the presence of side chains restricts the number of possible arrangements to few most common secondary structural elements: α-helices and β-sheets. These elements where first described by Pauling and Corey as structural features stabilized by a regular network of hydrogen bonds [2].

Hydrogen bonds form when a hydrogen atom (donor), linked to a strongly electronegative atom, interacts simultaneously with another atom having a lone pair of electrons (acceptor). In α-helices there is an interaction between the amine N-H hydrogen of the amino acid i and the carbonyl O=C oxygen of the amino acid $i + 4$. There are 3.6 residues per turn of helix and this repeating interaction constitutes an energetic advantage for this structural element.

Hydrogen bonds also stabilize a second kind of secondary structural features: β-sheets. Unlike α-helices, these are not composed of consecutive amino acids but are different adjacent fragments (β-strands) interacting together. The β-sheet is referred

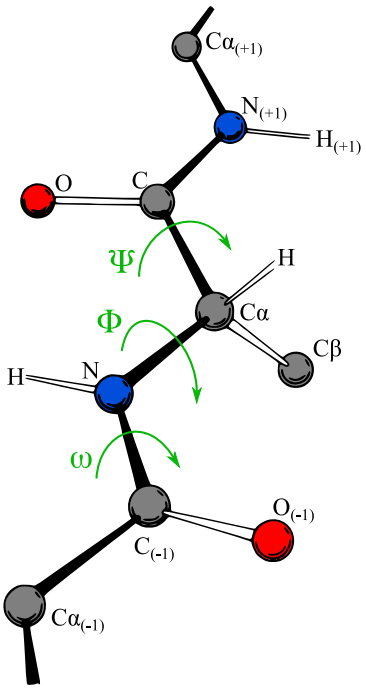


Figure 2: ω, ϕ, and ψ dihedrals. (Image by Dcrjsr under CC BY / Modified from original).

to as "parallel" when all the β-strands have the same orientation from N- to the C- terminus, and "antiparallel" otherwise.

All these secondary structure elements are characterized by specific values of the ϕ and ψ dihedral. A useful way to visualize the rotational freedom of residues is the Ramachandran plot (Figure 3), where the most densely populated areas are exactly those which characterize α-helices and β-sheets. Other secondary structural elements are "turns" or "loops" that tightly or loosely link the more stable secondary structural elements. A last category is "random coils", which are not real structural elements but are rather unstructured fragments.

1.1.1.3 *Tertiary structure: folds*

The tertiary structure of a protein is the real tridimensional displacement of atoms in a protein. This is generally given by an alternation of secondary structural elements that can fold into their energetic minimum spontaneously. The fold of a protein is a specific arrangement of secondary structure elements, and some of these super-secondary structures are recurring in nature even for unrelated sequences. Categorizing folds is not easy as defining secondary structure, as the fold can be seen from different point of view. Databases like CATH [3]

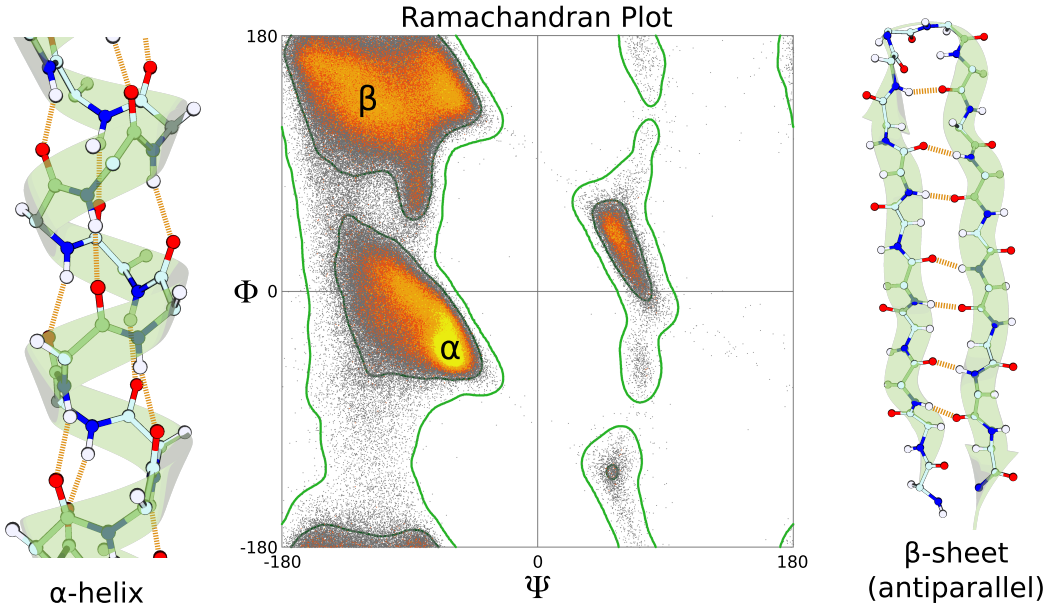


Figure 3: Ramachandran plot of the ϕ,ψ protein backbone dihedral angles for general-case amino acids (no Gly, Ile/Val, Pro, or pre-Pro), from a dataset of 1.5 million residues in 8000 protein chains with resolution <2.0 Å and backbone B-factors ⩽30. The individual-residue data-points are color-coded by the number in each 0.1° bin. The inner contour encloses 98% of the data (the "favored" region, while the outer contour encloses 99.95% of the data, dividing "allowed" from "outlier" regions. (Image by Dcrjsr under CC BY / Modified from original). On the sides examples of the hydrogen bonding network stabilizing secondary structure elements. To the left side for an α-helix and on the right side for an antiparallel β-sheet.

or SCOP [4] try to hierarchically cluster protein folds. While CATH is more directed towards structural classification, SCOP is focused on the evolutionary relationship.

Apart from the peptide bond providing a solid scaffold for the protein backbone, and hydrogen bonds stabilizing secondary structure elements, other covalent or non-covalent interactions can further stabilize the tertiary structure of proteins. The main driving force that pushes unfolded protein to its folded structure is the hydrophobic collapse [5]. When in water solution, non-polar hydrophobic side chains of residues tend to interact reducing the entropy of the polypeptide. This hydrophobic effect is a non-covalent kind of interaction that pushes non-polar residues together in order to minimize the contact surface with the solvent. As secondary structural element comes closer in space disulphide bridges can form. Two sulfur containing amino acids (i.e. two cysteines) can form very strong covalent bond between their sulfur atoms called disulphide bond. This is the strongest type of bond proteins can make (60 Kcal/mol) and acts as main stabilizer of the fold of proteins.

The last kinds of interactions, that tightly pack the already folded protein, are the Van der Waals forces. A Van der Waals interaction is the transient and weak attraction of an atom to another. Every atom has a fluctuating electron cloud that can temporarily yield a dipole. On a very short distance, around 3 Å, this dipole can induce another dipole in neighboring atoms providing a weak (1 Kcal/mol) electrostatic interaction. In complex system like a polypeptide chain the total contribution of many Van der Waals interactions becomes relevant.

1.1.1.4 *Quaternary structure: oligomers*

Quaternary structure is the combination of different polypeptide chains (identical or different) each one with its own tertiary structure. An oligomer, or multimer, is a complex of multiple polypeptide chains, as opposed to monomers that have a single chain. The number of interacting chains can greatly vary from the simple homo-dimeric interaction, involving two identical chains (i.e. originated from the same gene), to heteromeric assemblies (i.e. different genes product) where each component has a defined stoichiometry (e.g. in hemoglobin we have two α and two β subunits).

SYMMETRY A peculiar characteristic of oligomers is their symmetry. While single tertiary elements rarely possess an internal

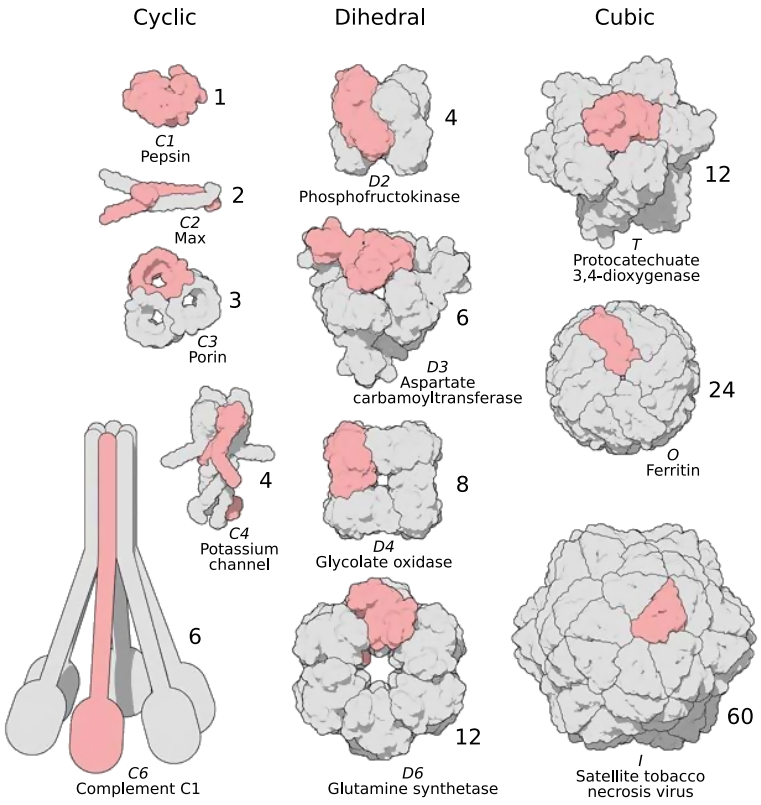


Figure 4: Showcase of common homo-oligomers with beautiful cyclic, dihedral, and cubic symmetries. Monomeric subunit is highlighted in red, The total number of subunits composing the oligomer annotated to its right. Image by David Goodsell adapted from [6].

symmetry, most of the soluble or membrane-bound oligomers have a symmetrical arrangement of their subunits. Goodsell and Olson observe that symmetrical oligomers are favored because of higher stability (each component is less exposed to the solvent) and finite control of assembly, so to avoid deleterious boundless oligomerization of proteins [6]. Given that residues in protein are chiral, only crystallographic point group symmetries are allowed (i.e. mirror and inversion are disallowed) (Examples in Figure 4).

Cyclic groups (C_n) have a single axis of rotational symmetry, forming a ring of n repeated subunits. This arrangement is typical of proteins having a function related with the directionality (e.g. many membrane proteins) or that require the formation of a chamber or a hollow tube (e.g. ion channels). Like cyclic groups, dihedral groups (D_n) have a rotational symmetry axis plus a perpendicular one of two-fold symmetry. With respect to cyclic symmetries, dihedral symmetries have the potential for a much larger interface. The contacts between a subunit in C

symmetries are limited to the two subunits directly to the left and to the right, while in D symmetries subunits tend to be in contact also diagonally. This is a perfect scaffold for allosteric and cooperative interactions, as more binary interactions are available.

Cubic groups contain three-fold symmetry that is combined with a non-perpendicular rotational axis. We have tetrahedral (T) symmetries when the additional rotational axis is two-fold; octahedral (O) when the axis is four-fold; icosahedral (I) when the axis is five-fold. Cubic symmetries are mainly found in proteins specialized in storage and transport and they are also suited for viral capsid providing the hollow shells for viral proteins. Finally helical symmetries (H) derive from the combination of translational and rotational symmetries. This combination results in an unbound repetition of elements that is typically found in structural elements (e.g. fibrils, microtubules, and fibers).

There is no direct correspondence between the crystallographic asymmetric unit and the biologically functional macromolecule. The asymmetric unit might contain part of the biological assembly, coincide with it, or contain multiple biological units. Tools like PISA [7] or PQS [8] help crystallographers in reconstructing the biological unit, often suggesting several alternatives that can be reviewed by authors.

PROTEIN-PROTEIN INTERACTIONS A multitude of forces concur to stabilize Protein-Protein Interactions (PPI). Apart from hydrophobic interactions, Van der Waals forces and hydrogen bonds, a characteristic interaction at interfaces is of electrostatic nature. Amino acids with acidic negatively charged side chains (aspartic acid and glutamic acid) interact with basic positively charged residues (arginine, histidine and lysine) forming a ionic bond, or salt bridge. These residues are scarcer in the protein's core given their bulky side chains, and often, they are on the surface of a monomer where their charge is neutralized by ions in the solvent or, more favorably, by the interaction with an opposite charge residue.

Another factor worth considering is the intracellular environment where proteins interact *in vivo*. The cell is a crowded environment where 20-30% of the volume is occupied by macromolecules [9]. This dense heterogeneous environment act as a non-specific kind of force that influence macromolecular association and conformation [10]. Indeed, nature developed specific

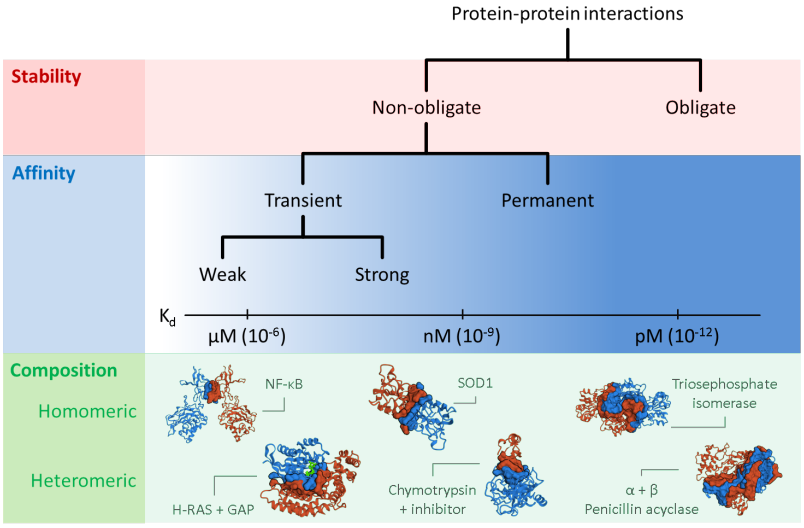


Figure 5: Classification of protein-protein interactions. Image by Ozlem Keskin adapted from [11].

tools, the molecular chaperones, that counteract the dense packing of macromolecules in cells providing a safe environment for nascent/folding proteins avoiding non-native aggregation.

CLASSIFICATION OF PPI Given the high number of possible forces bringing proteins together, it is natural that the modes and types of interaction also greatly vary (Figure 5). On the basis of the stability of the complex, the interaction can be obligate, when the partners involved cannot properly fold in isolation, or non-obligate, when folded monomers can fold independently [12]. Examples of obligate interactions are macromolecular machinery (e.g. proteasome, GroEL) that need a very precise and stable form of interaction for their functioning.

Depending on the lifespan of the interaction, complexes can be classified as permanent, when the interacting partners will not separate anymore (e.g. antibody-antigen, enzyme-inhibitor), or transient, when a spontaneous association/dissociation occurs in vivo. Many examples of the latter can be found in signaling and regulatory pathways, where an alternation of association and dissociation between different partners enables signaling cascades and a quick cell response to external stimuli.

The strength of an interaction is usually referred to as interaction affinity and differentiate between transient and permanent interaction. The affinity between proteins can be influenced by a variety of factor, for example pH, protein concentration, cell crowding, temperature, etc. For a binary interaction

$A + B \rightleftharpoons AB$, the binding affinity represents the force of attraction, between the A and the B proteins. The forward rates (k_{on}) determine the time scale of the association, while the reverse rates (k_{off}) describe the dissociation reaction. k_{on} and k_{off} can be used to find the equilibrium dissociation constant (K_d) with $K_d = \frac{[A][B]}{[AB]} = \frac{k_{off}}{k_{on}}$, where [A], [B] and [AB] are the concentrations of the unbound and bound proteins. The equilibrium dissociation constant, K_d, is related to the Gibbs free energy function $\Delta G = -RT \ln K_d$ and therefore can be used to find the binding free energy. The smaller the dissociation constant, the stronger the interaction is. For example, a complex with a nanomolar (nM) dissociation constant is more tightly bound than complex with a micromolar (μM) or millimolar (mM) dissociation constant.

1.2 EXPERIMENTAL STRUCTURE DETERMINATION

Since the determination of first protein structure in 1958 [13], many steps forward have been done in experimentally solving the structure of proteins at atomic resolution. Techniques like X-ray crystallography, nuclear magnetic resonance, and cryo-electron microscopy are consolidated experimental approaches able to deliver thousands of proteins structures per year.

X-RAY CRYSTALLOGRAPHY X-ray crystallography is one of the most important tool to study the structure of biological macromolecules at atomic resolution. It exploits the regular organization of such molecules when in crystal form. The amount of details of any form of microscopy investigation depend on the wavelength of the electromagnetic radiation used to "light" the sample. Protein are in the nM scale so the wavelength corresponding to X-rays. Protein are expressed, purified, and concentrated in order to grow crystals. As a X-ray beam irradiates the crystal, the electrons in the molecules diffract the beam, and a detector measures the angles and the intensities of the diffracted waves. The diffraction pattern depend on the arrangement of atoms in the crystal, therefore analyzing this pattern the structure of a protein can be deduced. The electron density of the molecule is related to the intensities of the spots in the diffraction pattern by a mathematical relationship know as Fourier transform [14].

In order to reconstruct the electron density in real space both amplitude and phase of the waves are needed. Amplitudes

are measured experimentally, but the phase information is lost. Solving a structure also imply being able to solve the phase problem. Different approaches can be used for this task. For example, in Molecular Replacement [15] the phases are derived by similar proteins of known structure. In Multiple Isomorphous Replacement (MIR) [16] heavy atoms are included in the protein (e.g. selenocysteine), the phase of the native structure must be close to the phase of the the heavy atom alone, which is known.

The fourth generation of light source, X-ray Free Electron Lasers (XFELs), is promising exciting advances in X-ray crystallography. Very short pulses (< 50 femtoseconds) of X-rays billions of times brighter than before will open new doors for the structural biology field. Nano- or micro-sized crystals can,with such level of brightness, generate good diffraction patterns [17]. When coupled with a delivery system (e.g. flow-jet), these advances provided the ground for the nascent field of serial femtosecond crystallography (SFX) that will shed new light on ultra-fast protein reaction dynamics [18].

NUCLEAR MAGNETIC RESONANCE Nuclear Magnetic Resonance (NMR) is a spectroscopic technique allowing structural studies of small proteins in solution. Atom nuclei with an uneven number of protons and neutron (e.g. hydrogen, ^{13}C, or ^{15}N atoms) are characterized by a magnetic momentum. When such nuclei are placed in a magnetic field, they can align with the field (lower energy) or against it (higher energy). Using a radio pulse, state transitions between the low and high energy spin state can be induced (resonance) and detected in the spectrum. Electrons flowing around a magnetic nucleus generate a small magnetic field that opposes the applied field. Because of this local shielding effect, nuclei in different environments will resonate at different field strength or radiation frequencies. The extent of shielding is influenced by local structural features within molecules, hence the variations in response to varying magnetic field or frequencies are called chemical shift.

In structural biology, chemical shifts can be used to predict regions of secondary structure of proteins [19] and also the tertiary structure of proteins [20]. With highly developed techniques the NMR spectra can be splitted in multiple dimensions. The result is, for example, a set of inter-proton distances (exploiting the nuclear Overhauser effect, NOEs) or the relative orientations of the different nuclei in a protein structure (resid-

ual dipolar couplings, RDC). These values can be used as constrains in simulations to obtain an ensemble of possible protein conformations.

While generally less detailed structures are obtained by NMR spectroscopy compared to X-ray crystallography, it is the method of choice when studying the dynamics of proteins, weak interactions, and systems that resist crystallization attempts. In the past, NMR analysis could only target proteins with a molecular mass below 30 kDa. Recent advances enabled NMR study on large proteins or complexes, for example allowing spectra collection on nascent protein folding in the ribosome [21]. Other aspect like post translational modification (PTMs), protein aggregations, and in-cell NMR spectroscopy are the focus of modern NMR [22].

ELECTRON MICROSCOPY As the smaller wave-length of electrons is used as illumination source, electron microscopes can go far beyond the resolution limit of conventional light microscopes, reaching about 10,000,000x magnification. The main issue with looking at biological samples through an electron microscope is the degradation of the sample. Chemical bonds in biological macromolecules can be broken by the high energy of the electron beam. Moreover, electrons are scattered by air molecules, so EM requires a high vacuum in the beam path, which compromises preservation of liquid aqueous samples.

Dehydrating or fixing the samples by negative staining (water is substituted by heavy-metal salt) can secure the sample, but do not preserve its close-to-native state. Samples can be fully preserved with the "cryo-EM" approach where samples are freezed in thin layer of amorphous or vitreous ice [23, 24].

3D structures could be calculated from 2D projections of macromolecules in different directions. The limitation in this approach, called single-particle analysis, is that the relative orientations of the particles are unknown. The low resolution of images makes determining these orientation particularly hard especially for small proteins or in absence of symmetry in the protein.

Since 2013, progresses in cryo-EM single-particle analysis have been so fast that has been termed "the resolution revolution" [25]. The causes for this revolution is a combination of two factors: a new generation of direct electron detectors and an improved image processing procedures correcting sample movements. The synergy between these two factors was unexpected giving a jump in resolution from 15 Å to 3.5 Å. At these res-

olutions, cryo-EM density maps are similar to those obtained by X-ray crystallography allowing *de novo* building of atomic models. This allow the study of membrane protein or sizable macro-molecular complexes in their native conditions at atomic resolution.

EXPERIMENTAL STRUCTURE AVAILABILITY The Protein Data Bank (PDB) [26] was established in 1971 as central archive of all experimentally determined protein structure data. Today the PDB is maintained by an international consortia collectively known as the Worldwide Protein Data Bank (wwPDB). The goal of the wwPDB is to maintain a single archive of macromolecular structural data that is freely and publicly available to the scientific community.

The atomic coordinates are deposited in the archive by experimentalist together with experimental details such as oligomeric state, protein sequence reference, refinement parameters, experimental conditions, etc. Each structure is given a four-letter code (the PDB code, or PDB identifier) that makes it unequivocally referable. More than 120,000 structures are available today. The majority of these are solved by X-ray crystallography (90%), solution NMR (9%), and electron microscopy (1%).

The file format used by the PDB was called the PDB file format. It is historically restricted to 80 columns (as punch card were) and it has limitations in number of atoms and polypeptide chains that can be represented. The main format for the PDB is now the "macromolecular Crystallographic Information file" (mmCIF) [27] that is based on a definition file, avoiding the PDB file limitations. A new format is the "Macromolecular Transmission Format" (MMTF) that is a binary file format much more compact and fast to load and parse.

1.3 PROTEIN STRUCTURE PREDICTION

Although the knowledge about aminoacidic sequences as well as protein structures have grown enormously in the past years, they are not growing at the same scale. Thank to deep sequencing technologies, the UniProtKB/TrEMBL [28] database is currently reporting almost 68 million protein sequences from over five thousands different species, while the available structures in the Protein Data Bank are roughly 124 thousands. That is below 1% of the total known protein sequences. The level of automation for structure determination cannot currently com-

pete with the level of high-throughput sequencing. This uneven amount of knowledge, the so called sequence-structure gap, is increasing over time.

To fill this gap, computational approaches flourished with the aim predicting protein structures. The process by which proteins reach their native conformation is called folding and its mechanisms are not yet fully understood. The number of possible geometrical arrangements of atoms in a protein is astronomically high. It is surprising that proteins can reach their correct conformation in a very short time, in the order or milli- or micro-seconds. This is the so called Levinthal's paradox [29] that raised many questions, catalyzing the attention of the scientific community on protein folding. The commonly accepted hypothesis, that better explain this phenomenon, was formulated by Anfinsen [30] who showed how a denatured protein can be brought back to functionality restoring its environment. The consequences of Anfinsen experiments are two: i) the folding process is driven by thermodynamic stability, i.e. a protein follow a path that minimizes its free energy; ii) the information on a protein structure is contained within its amino acid sequence. This imply that knowing the sequence of a protein we can infer its structure. The whole field of structure prediction is very broad and rich in nuances. In general, the approaches to computationally model protein structures are of two kind: template based and template free.

1.3.1 *Template based modeling*

In their seminal paper [31], Chothia and Lesk compared X-ray structures of evolutionary related proteins. Comparing the structural similarity of proteins core to the sequence similarity they could observe a clear relation between the two: structural similarity increase exponentially with sequence similarity, i.e. structure having similar sequences also have similar structures. Moreover, structure is more conserved than sequence, so even protein with remotely related sequences can assume similar folds. All template based modeling approaches are founded on this principle and hence focus on the prediction of the three-dimensional structure of proteins having homologs of known structure. This kind of modeling methods are also referred to as comparative or homology modeling. The general idea is to exploit the experimentally determined 3D structure of a pro-

tein (template) to compute the structure of a related protein of interest (target). The general procedure follow four steps:

1. Identification of a template for the target sequence.
2. Alignment of target-template sequences.
3. Modeling of the target structure based on template information.
4. Refinement of the model.
5. Evaluation of model quality.

The initial steps of identification and alignment of the target sequence to the template is crucial. When no homologs sufficiently close in sequence the entire procedure is less effective. Local alignment tools as BLAST [32] are used to obtain alignments of the target-template pair. Over a threshold of roughly 30% sequence identity, 90% of the models are accurate, while below 25% sequence identity, only 10% of the models are accurate [33]. When no close homologs are detected, more advanced homology detection algorithms can be used. The most sensitive approach is based on a Hidden Markov Model (HMM) representation of the target sequence. An initial multiple sequence alignment is built for the query sequence and amino acid emission probabilities are computed as well as insertion and deletion states. This HMM query is then aligned to a database of HMM profiles, greatly improving the detection of remote homologs [34, 35].

Following a strictly conservative modeling approach, aligned regions of templates backbone are copied to the model and serve as "raw" starting point. Variable regions (insertion or deletion) are then closed using fragments identified from a library of known structures or modeled *de novo*. Then, sidechains conformations are modeled. Again, identical residues orientations can be directly transferred to the model, while unconserved ones can be modeled using backbone dependent rotamer libraries (e.g. SCWRL software [36]). The refinement step takes care of regularizing the structure, i.e. removing clashes, adjusting angles and bonds and checking the general stereochemistry of the model. Finally, to be complete, a model must also include some confidence or reliability value. A global confidence value can be useful for the ranking of alternative models, while a local per-residue confidence can highlight the most

trustworthy regions of the model for experimental follow-ups (e.g. binding-site accuracy for drug design).

Another approach to comparative modeling is based on the satisfaction of spatial restraints, introduced by Šali [37] and implemented in Modeller [38]. In this case, model generation is approached as an optimization problem, where different restraints are imposed. The restraints are formulated as probability density functions of observables (e.g. atom distances, angles, and dihedrals) derived from different sources (e.g. known structures, force fields, or stereo-chemical considerations). Protein models satisfying the combination of all restraints are generated by conjugate gradient descent of the combined probability density functions.

1.3.2 *Template free modeling*

When no homologs to a target protein are available, template free approaches come into play. Typically, this class of methods perform a conformational search based on the minimization a free energy function approximation. The use of this first principles approach give this class of method the alternative name of *ab initio* or *de novo*. A series of candidate conformations are generated and ranked according to the energy function. This energy function can be used to drive complete folding simulation in Molecular Dynamics (MD) approaches and reveal precious details on the folding process or the dynamic aspects of proteins. For the scope of structure prediction instead, information from experimental structures must be integrated in the form of backbone fragments sampling procedure or knowledge-base empirical potential extracted from databases [39].

ROSETTA [40] is a suite of protein design and prediction softwares, also offering template free functionality. It is using a sampling scheme that is driven by structural fragments coupled with an elaborated energy function [41]. Another approach is to combine the fragment sampling with threading of the target sequence on experimental structures, as implemented in I-TASSER [42]. QUARK [43] instead, models proteins only using small fragments (1-20 residues long) by replica-exchange Monte Carlo simulation under the guide of an atomic-level knowledge-based force field.

1.3.3 *Critical Assessment of protein Structure Prediction: CASP*

Since 1994, the modeling community started an objective evaluation of methods capabilities and bottlenecks. The Critical Assessment of protein Structure Prediction (CASP) [44–54] is a community-wide double blind experiment involving hundreds of prediction teams and delivering an independent assessment of the state of art in the protein structure prediction scene.

The experiment is structured as follows: the experimental community provides sequences of structures about to be solved (by X-ray or NMR). These sequences are sent to all the participating modeling groups, who submit their predictions before any experimental data is released. All the models are then evaluated on different criteria by independent assessors. Targets, methods and assessment teams are usually divided in categories (e.g. template based, template free, refinement, oligomeric assemblies, contact prediction, etc). Methods that performed particularly well in one of the categories are then highlighted at the CASP meeting.

The last CASP editions [53, 54] confirms the higher accuracy for models produced using template information. Slight but consistent improvements have been also achieved in the process of refining structures by physic based molecular dynamics [55]. Also the strive of the modeling community for more biologically meaningful models, led to the opening of a new category for the modeling of oligomers.

1.4 MODELING PROTEIN-PROTEIN INTERACTIONS

Information about protein-protein interactions is growing at a similar pace as of amino acid sequence. Experimental information on interacting partners grows with exponential trend [56–59] as it can be obtained with high-throughput methods [60–62] such as two-hybrid screening (Y2H) or affinity purification of complexes. On the other hand, the number of experimentally determined three-dimensional complexes and oligomeric structures is lagging far behind. Shedding light on the atomic details of such interactions is challenging since the expression of protein complexes is often tightly regulated and obtaining sufficient concentrations for structure determination is not trivial. For this reason it is desirable to gain as much structural details as possible using computational approaches.

1.4.1 *Template free docking*

Historically, one of the first approaches used to model interactions *de novo*, when only structures of the individual components are available, was macromolecular docking. The relative orientation of two proteins is sampled and scored e.g. by exploiting the components' shape [63] or physicochemical complementarity [64]. Extending binary to multi-body docking is problematic since the relative orientation space to be sampled grows exponentially when increasing the number of monomers to combine. Several multimeric docking methods successfully reduced the search space relying on the fact that oligomers are often organized in symmetrical assemblies (e.g. SymmDock [65] and M-ZDOCK [66]) or assembling monomers incrementally and using a greedy approach or linear programming (e.g. LZerD [67] and DockStar [68]). When experimental details of the interaction are available (e.g. EM density maps, cross-linking, SAXS or NMR data, co-evolution analysis, etc.), different "hybrid-modeling" tools can be used (Integrative Modeling Platform (IMP) [69], the Rosetta Suite [70], and HADDOCK [71]) to enforce experimental constrains and model sizable assemblies. At a computational price, these multimeric models can be improved accounting for the dynamic and flexible nature of the multimeric interfaces by molecular dynamic simulations [72].

1.4.2 *Template based docking*

An alternative strategy, homology-based docking, relies on the correct conformation being already discovered. Nature copies itself, and like the limited number of protein folds [73], the number of ways proteins interact is likely limited as well [74, 75]. Indeed, it has been observed that similar binding modes can be identified for almost all known protein-protein interactions [76] and also that the location of the interface is the same between structural homologs [77]. These observations paved the way for homology-based modeling of protein interactions, where uncharacterized interactions are modeled using experimental structures of homologous interacting protomers (interologs) or domains as templates. Speed is the great advantage of approaches based on homology over computational docking approaches, making them scalable to full genomes. In recent years, the scientific community witnessed a flourishing of databases

and online resources that map structural information on protein-protein interactions networks (GWIID [78], Interactome3D [79], PrePPI [80], INstruct [81], PRISM [82]). Altogether, homology-based approaches successfully reduced the gap between known interactions and those that are structurally characterized, providing biologists with an unprecedented amount of detailed structural information.

1.4.3 *Critical Assessment of Predicted Interactions: CAPRI*

Taking its inspiration from CASP, the Critical Assessment of Predicted Interactions (CAPRI) aims at assessing the ability of docking methods to correctly predict interaction between proteins [83–88]. Since its inception in 2001, CAPRI played a central role in advancing the field of macromolecular docking. CAPRI expanded the focus including target of protein-peptide and protein nucleic acids interactions. Moreover, effort in predicting binding affinity [89] and position of relevant interfacial water molecules [90] has been undertaken. In general, docking approaches are especially accurate when no significant conformational changes are required for interface formation.

1.5 THESIS AIM

The general aim of this thesis is to advance methods in protein structure prediction by homology. Today, thanks to the modeling community efforts described in the introduction, some form of structural information is available for the majority of translated amino acid translated in model organisms [91]. Anyhow, we have less structural information about protein-protein interaction, making the problem of predicting structure of interacting proteins more challenging. Our effort in this thesis is hence to tackle the problem of modeling homo- and hetero-oligomers considering their complete quaternary structure.

To reach this goal we first define a distance measure (QS-score) that enables us to compare oligomeric interfaces. This is a required step as we want to measure the similarity of models to native structures that can have different oligomeric architectures.

Independently from geometrical considerations, a critical aspect of protein-protein interfaces is the evolutionary pressure driving formation and stabilization of such interfaces. To account for this, we define a novel approach to describe conser-

vation in protein-protein interfaces (PPI fingerprint). The motivation for this task is that, not every assembly deposited in the PDB is biologically relevant and we need to disregard those artifacts.

We then implement a template based approach, suitable for both homomeric and heteromeric modeling, addressing the problem of template selection developing a ranking method based on the prediction of interface quality. In doing so, we also propose an approach to automate the process of homology modeling including prediction of the oligomeric state of proteins. Finally, this approach is integrated and made available to the research community through the SWISS-MODEL web-server.

2

STRUCTURAL SIMILARITY OF PROTEIN COMPLEXES

PDB entries are often annotated (either by authors, software or both) with multiple potential biological assemblies. These complexes might have diverse stoichiometry and/or alternative interfaces. Several methods to measure interface similarity developed in recent years are summarized in Table 1. These distance metrics have been developed in the context of protein-protein docking, concentrating on binary interactions and not on oligomeric proteins. Decomposing the compared assemblies into binary interactions can result in a factorial number of comparisons and missing interfaces (e.g. comparing a dimer to a tetramer) that cannot be accounted for.

2.1 METHODS

2.1.1 *Comparing quaternary structures: QS-score*

To overcome the limitations of the available interface metrics and to describe the diversity of quaternary structures in the PDB, we developed QS-score (Quaternary Structure score). QS-score is a distance measure that considers the assembly interface as a whole and is suitable for comparing homo- or hetero-oligomers with identical or different stoichiometries, alternative relative orientations of chains, and distinct but related amino acid sequences (i.e. homologous complexes).

To unequivocally identify the residues of all protein chains in complexes, the first step is establishing a mapping between equivalent polypeptide chains of the compared structures. This information is essential to unequivocally identify residues since there are no rules for unique nomenclature of protein chains in complexes. Once the mapping is obtained we can safely compare the interface contacts (i.e. pair of residues interacting across different chains) between complexes.

2.1.1.1 *Chain mapping*

The number of possible mappings between two complexes A and B having a different number of subunits is $\binom{n_a}{n_b}$ where n_A

Table 1: Interface distance measures developed in the last few years. For each we report the measure name, the reference paper, whether is suitable for binary interfaces or multimeric interfaces and a short summary of the method.

Measure	Reference	Binary	Multimeric	Method summary
f_{nat}	CAPRI assessment	✓		Fraction of correctly predicted contacts
L_{rms}	[83, 86, 92–94]	✓		RMSD of ligands (smallest chains)
I_{rms}		✓		RMSD of interface atoms
iRMSD	Aloy *et al.* [95]	✓		RMSD calculated on 14 predefined coordinates (independent chain superposition)
iTM-score	Gao and Skolnick [95]	✓		Geometric distance of interface residues
IS-score		✓		Contacts similarity of interface residues
MM-align	Mukherjee and Zhang [96]	✓	✓	Structural alignment by chain-joining
Q-score	Xu *et al.* [97, 98]	✓		Geometric distance differences between equivalent interfacial residue

is the number of chains in the larger complex A and n_B those of the smaller complex B. In the worst case of two equally sized complexes the number of possible mappings is $n!$. This clearly becomes untreatable when comparing big complexes such as viral capsids.

However, when symmetry information is available in the coordinate file or can be deduced, the problem can be reduced to the identification of the mapping between symmetry related groups, which are typically containing a number of treatable subunits. To our knowledge, this currently is the only algorithm taking into account the problem of chain mapping. The steps

performed by the QS-score algorithm to identify the mapping are the following:

1. Polypeptide chains within each complex are grouped by their chemical equivalence (e.g. the two α chains in human hemoglobin)
2. Equivalent groups between the two assemblies to be compared, are identified by global sequence alignment (e.g. hemoglobin chains α in two different structures)
3. Symmetry or pseudo-symmetry of each complex is calculated and chains which can roto-translated reproducing the full assembly are considered as symmetry groups (e.g. in hemoglobin two pairs of α-β chains)
4. The chain mapping between two symmetry groups in different assemblies is identified by superposition. This symmetry group mapping is applied to all symmetry groups.
5. For each symmetry group of step 3 all possible pairs are considered
   a) A symmetry group pair is used as base to superpose complexes
   b) The lowest global RMSD highlight the correct mapping
6. Equivalent residues between the assemblies are indexed by sequence alignment.

2.1.1.2 *Interface contacts*

We consider an interface contact to occur when Cβ atoms (Cα for Glycine) of residues belonging to different chains are at most 12 Å apart. This definition of contact is inspired by Q-score [97] and it allows us to compare structures not having identical side chains. From the inter-complex chain mapping we can deduce also the inter-complex residue mapping aligning the sequences of each chain in the complexes. Each contacting pair of residues (i,j) in the first complex is mapped to a (k,l) pair in the second complex. QS-score is then defined as follow:

$$\text{QS-score} = \frac{\sum_{(i,j)(k,l)} w(\min(d_{(i,j)}, d_{(k,l)}))(1 - \varepsilon|d_{(i,j)} - d_{(k,l)}|)}{\sum_{(i,j)(k,l)} w(\min(d_{(i,j)}, d_{(k,l)}))}$$

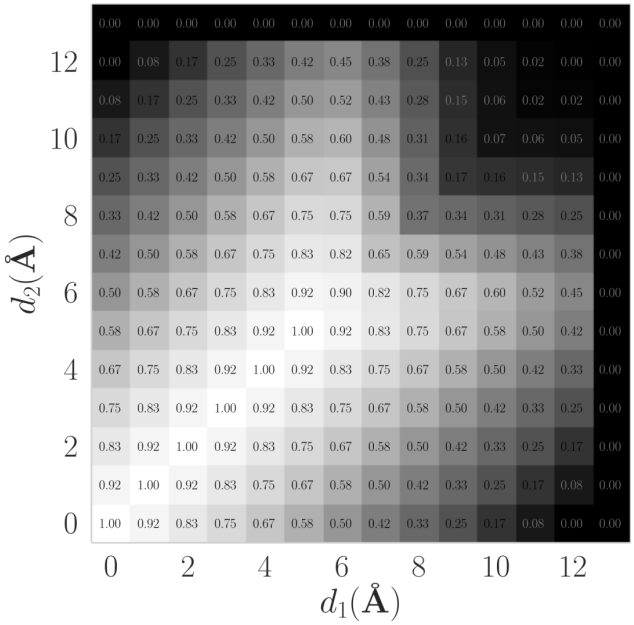


Figure 6: Example of QS-score for a pair of distances d_1 and d_2. The values on the diagonal indicate the weight of the contact pair (the denominator part in 1) that is gradually fading for long range contacts. The off-diagonal values represent the numerator part in 1.

(1)

where d is the Euclidean Cβ distance between the residues, ε the relative error (considering 12 Å as maximal error) and w the weighting function:

$$w(d) = \begin{cases} 1, & \text{if } d \leqslant 5. \\ e^{-2(\frac{d-5}{4.28})^2}, & \text{if } 5 < d \leqslant 12. \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

which expresses the probability of a side-chain interaction given the Cβ distance as derived by Xu *et al.* [97].

If all the distances conserved, QS-score is 1, indicating identical interfaces. When the distances are not equal, the relative error factor pushes the QS-score towards 0 proportionally to the difference in the distances. In case of unmapped contacts a maximal error is considered further penalizing the QS-score (e.g. Figure 6).

When the QS-score is close to 1 it indicates that the compared interfaces are similar, so the complexes have equal stoichiometry and a majority of the interfacial contacts are conserved. On the other end, a QS-score close to 0 indicates a radically

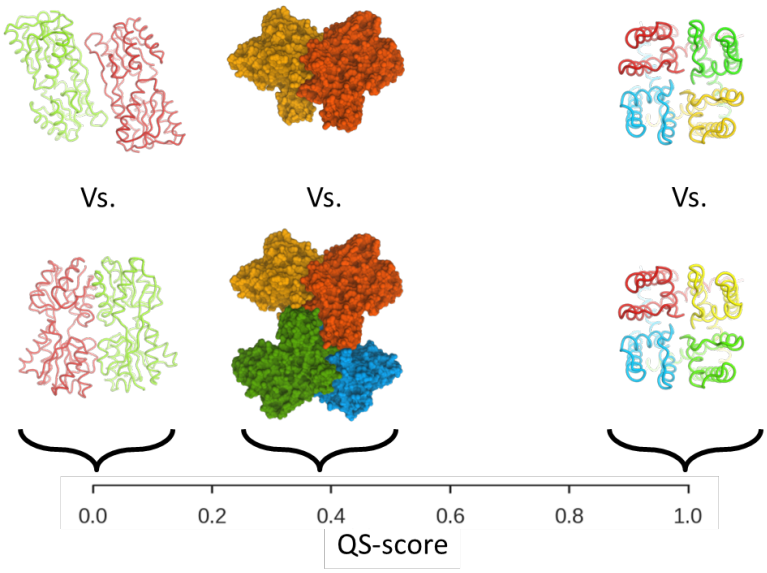


Figure 7: Examples of QS-score comparisons. From left: two possible assemblies of the Lac repressor from *E. coli* (PDB code: 1JYE) are compared resulting in a QS-score of 0 since their interaction mode is not similar (the contact occurs between one side of the monomer or the opposite). Two possible quaternary structures (dimeric and tetrameric) are available for the alkaline phosphatase from *H. salinarum* (PDB code: 2X98), only one dimeric interface is shared between the two forms resulting in a score lower than 0.5. Two structures of the same ion transport channel from *A. butzleri* (PDB codes: 5KLS, 5KLG) where the colors represent the chain names. The chain mapping step solves the disagreement between the otherwise isomorphic structures resulting in a QS-score of 1.

diverse quaternary structure, so the assemblies have different stoichiometries or may represent alternative binding conformations as exemplified in Figure 7.

2.2 RESULTS

2.2.1 *Structural similarity in homologous complexes*

We used QS-score to analyze the structural heterogeneity of all homo- and hetero-oligomeric assemblies deposited in the PDB. Sequences were clustered into groups sharing more than 90% sequence identity and for each sequence cluster we performed structural hierarchical clustering using different QS-score thresholds.

All homo- and hetero-oligomeric structures deposited in the PDB (August 2016) where considered. Chains consisting of small peptides (below 20 amino acids) or Cα traces were discarded. In case a single chain remained after the filtering, this was also removed. This resulted in 90,764 assemblies for 63,902 PDB entries and 356,585 polypeptide chains. The single chain sequences where clustered using CD-HIT [99] (90% sequence identity). To properly handle heteromeric structures, a sequence cluster is defined as the unique set of single chain clusters ids to which each of the complex chains belongs. This resulted in 24,272 clusters of which 13,896 (57%) included multiple assemblies and were further analyzed. All the assemblies in each sequence cluster were compared using QS-score and the resulting distance matrix was used to perform a hierarchical/agglomerative clustering using complete linkage.

491 clusters (3% of the total number of clusters) were discarded mostly due to incompatible symmetry groups between the compared assemblies which led to an intractable number of possible mappings. Figure 8 shows the fraction of sequence clusters being homogeneous (with a single QS cluster) or heterogeneous (with 2 or more QS clusters). Even if the majority of sequence clusters are homogeneous, this analysis clearly shows that sequence neighbors do not always have structurally identical interfaces. Using a QS-score threshold of 0.5, hence grouping structures having similar interfaces and identical stoichiometry, one third of the sequence clusters contain assemblies with interfaces different from each other.

This structural interface diversity between assemblies sharing high sequence identity represents a challenge for QS modeling. All alternative QS options must be considered as potential templates in a protein structure homology modeling approach since a decision based only on sequence similarity cannot distinguish between different oligomeric conformations.

2.3 DISCUSSION

Developing a new protein interface distance measure that consider the entire complex interface allowed us to get a glimpse of the surprising heterogeneity of multimeric structural space. Aloy *et al.* [95] noted that binary domain-domain interactions are structurally conserved above 30-40% sequence identity and Levy *et al.* [100] noted that the symmetry of the complexes is almost invariably conserved over 90% sequence identity.

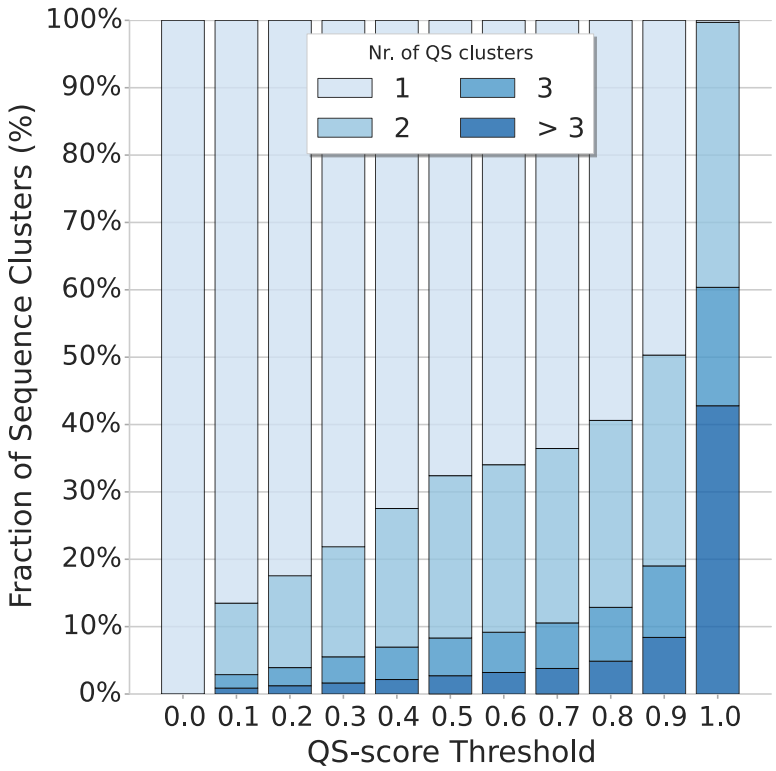


Figure 8: Heterogeneity of quaternary structures available in the PDB repository. Assemblies from the PDB were clustered by sequence identity (90% sequence identity). All the assemblies within one sequence cluster were compared using QS-score. The resulting distance matrix was used to perform hierarchical clustering using different distance thresholds. With a distance threshold (x-axis) of 0 all assemblies are clustered together so that the fraction of sequence clusters (y-axis) having only one QS cluster is 100%. As the threshold is increased the structural heterogeneity of the sequence clusters is evident and the fraction of sequence clusters having multiple QS clusters (in shades of blue) increases

In agreement with these analyses, we clearly show that the majority of sequence neighbors have structurally similar interfaces. Nonetheless, a significant fraction (one third considering a QS-score threshold of 0.5) contains assemblies with interfaces different from each other. While this analysis is agnostic of the actual biologically relevant conformation, it shows that in roughly one third of the cases a similar sequence is not a safe proxy for similar quaternary structure. This does not mean that any attempt to exploit homology relationship is futile, but highlights the necessity of explicitly considering all alternative quaternary structure conformations during the template identification step in homology-based modeling approaches.

3

CONSERVATION OF PROTEIN INTERFACES

Proteins acquire oligomeric organization for a variety of functional and biophysical advantages: modular elements are less prone to coding errors, oligomeric regulation add an additional level of control, large structures are more stable and can perform their function cooperatively [6], and other processes have influenced the evolution of proteins' interface formation [100, 101].

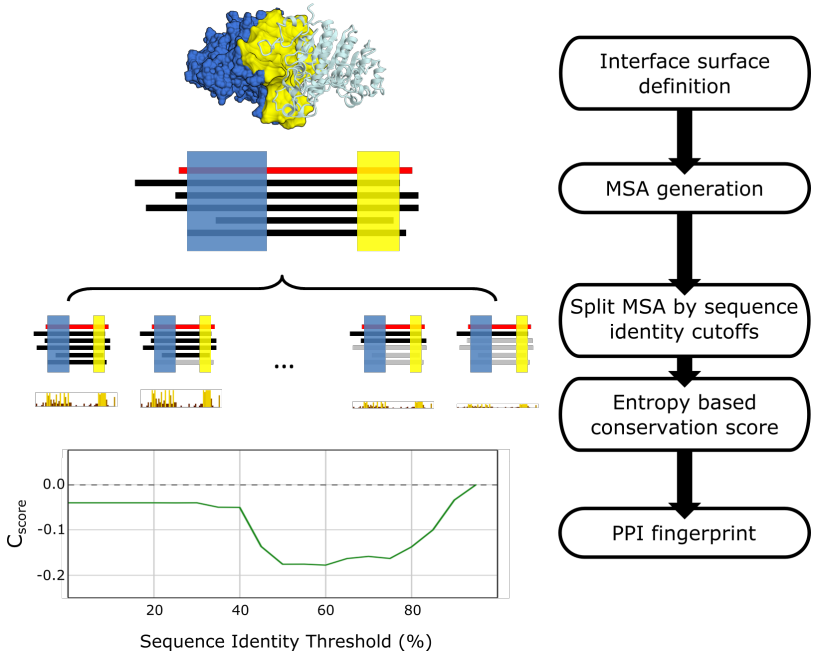


Figure 9: PPI fingerprint calculation. The starting point is a structure for which we define residues belonging to the interface or to the surface of the complex. Then, we generate a MSA representing the protein family of interest. The alignment is then divided using different sequence identity inclusion cutoffs. For each sub-alignment we compute the conservation of interface residues relative to surface residues. This result in a curve, that we call PPI fingerprint, showing the differential conservation signal from close to remote homologs.

During evolution, proteins can vary their oligomeric state by different mechanisms: either by direct mutations occurring at the subunit interface or by indirect mutations allosterically inducing a change in binding modes [102]. Several groups have

analyzed the impact of evolutionary pressure on protein-protein interfaces [103–105]. These analyses rely on an estimation of conservation that is typically derived from a multiple sequence alignment (MSA) of homologous proteins. Residues participating in interfaces are subject to different evolutionary constraints than residues at the protein surface interacting with the solvent.

This creates a confounding factor when proteins organized in different quaternary structures are included in the same alignment. In this chapter, we introduce a refined analysis of interface conservation which captures how interface conservation varies as a function of evolutionary distance within a protein family. We employ this analysis (which we refer to as Protein-Protein Interaction (PPI) fingerprints) for two critical tasks: first, the discrimination of crystal artifacts from biological contacts, which is a crucial step in determining the correct quaternary state of crystal structures to be used as templates in homology modeling; second, the evaluation of interface quality in models to assess the confidence in the predicted quaternary structure. The approach we used for the analyses is presented in Figure 9.

3.1 METHODS

3.1.1 *Conservation score*

INTERFACE AND SURFACE DEFINITION We compute the accessible surface area (ASA) of the monomer and the buried surface area (BSA) of the assembly with the Naccess implementation of the Lee-Richards algorithm [106]. Following the definition of interface core and surface residues in [107], we define surface residues as those having a relative accessibility (rASA) larger than 25% (considering the monomer). Interface residues are those whose relative buried surface area (rBSA) is higher than 25% and that have a rASA below 25% (considering the assembly). The remaining residues are considered as protein's core residues.

MSA GENERATION The MSA is obtained running HHblits [35] against the non-redundant (20% sequence identity) NCBI database with a threshold of 80% as minimum coverage. The MSA alignment is divided using 20 sequence identity inclusion cutoffs (0-100% in steps of 5%). For each of the sub-alignments a conservation score will be independently computed.

INTERFACE CONSERVATION Sequence conservation can be expressed as Relative Entropy [105, 108, 109]:

$$RE_c = \sum_a p_a \log_2 \frac{p_a}{p_{ab}} \tag{3}$$

Where p_a is the probability of an amino acid a to be in the alignment column c and p_{ab} is the background amino acid a probability distribution computed over the entire alignment (gaps are excluded). The Relative Entropy (RE) is computed for each column c of a multiple sequence alignment and normalized in the interval $[0, 1]$ with 0 indicating less conserved residues and 1 more conserved residues. The column-wise RE is computed for each alignment.

We define the degree of conservation of an interface with respect to the surface using log-ratio of the average entropy of interface residues $\langle S \rangle_i$ (weighted by relative ASA, rASA) over the average of those lying in the rest of the surface $\langle S \rangle_s$:

$$\langle S \rangle = \frac{\sum rASA_c RE_c}{\sum rASA_c} \tag{4}$$

$$IS = \ln \frac{1 + \langle S \rangle_i}{1 + \langle S \rangle_s} \tag{5}$$

A negative interface-surface ratio (IS) between interface entropy distribution and surface entropy distribution indicates that residues placed in the interface are less prone to mutate when compared to surface residues.

To test the significance of interface conservation we randomly sample "patches" of surface residues and compute their conservation (excluding the original interface residues). We define an adjacency graph of surface residues considering neighboring residues to have at least one atom within N Å apart each other (where N is dynamically set in order to obtain a connected graph). A surface residue is randomly picked and neighbors are added until the number of residues of the patch equals that of the interface. This process is repeated for a n number of times proportional to the original surface size. The surface residues not included in the patch are used to evaluate the interface-surface ratio, resulting in a distribution $X = (x_1, \ldots, x_n)$ of ra-

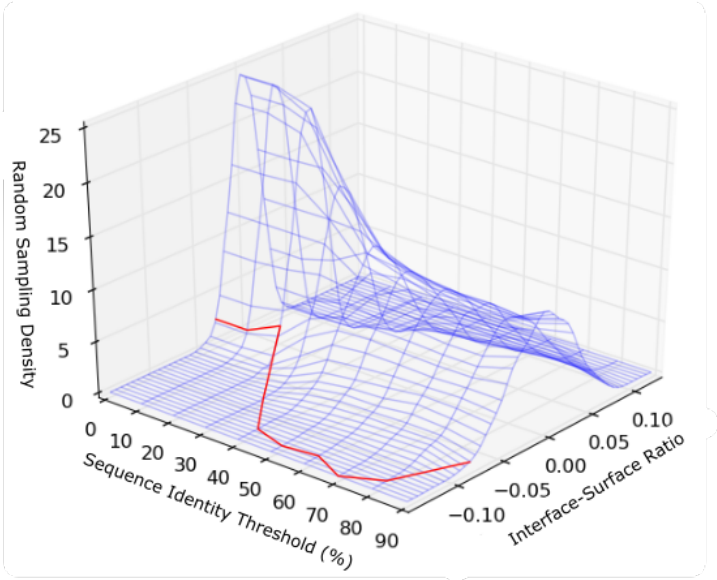


Figure 10: Distribution of interface-surface ratio in random patches. The random sampling of surface patches result in the distribution marked by the blue line. The red line indicates the interface-surface ratio for the actual interface. The further the score is from the random distribution the more significant it is.

tios as represented in Figure 10. We can estimate the P-value of the original interface as:

$$P = \min\left(\int_{\min(X)}^{IS} \hat{f}_h(X)dX, \int_{IS}^{\max(X)} \hat{f}_h(X)dX\right) \tag{6}$$

where IS is the native interface's interface-surface ratio and $\hat{f}_h$ is a kernel density estimation of the probability density function of the random patches conservation. The bandwidth parameter h is computed using Silverman's rule of thumb. Finally our conservation score is defined as:

$$C_{score} = IS(1 - P) \tag{7}$$

where the original interface-surface ratio IS is weighted by the P-value complement. So when an interface is close to the random patch distribution the score will tend to 0.

3.1.2 *PPI fingerprint*

Combining the conservation scores of different sequence identity cutoffs we obtain a curve, which we refer to as PPI fin-

gerprint as it captures the impact of evolutionary pressure on protein-protein interaction sites. As a positive control we computed the PPI fingerprint for a small set of six homo-dimeric proteins [110] where interfaces are conserved. The analyzed families are: alkaline phosphatase (PDB code: 1ALK), copper/zinc superoxide dismutase (PDB code: 1XSO), enolase (PDB code: 1ONE), glutathione S-transferase (PDB code: 1GLQ), streptomices subtilisin inhibitor (PDB code: 2SIC), and triose phosphate isomerase (PDB code: 1TPH).

The resulting PPI fingerprint curves (Figure 11) have values below zero indicating a higher mutation rate of surface residues compared to those at the interface, confirming the overall interface conservation for the protein families. In general, the curves follow a characteristic pattern: when only very similar sequences are considered (80-90% sequence identity threshold) the ratio is close to zero since the low variability in the MSA provides little information on the interface conservation. As we lower the inclusion threshold, the indication for a conserved interface is stronger and eventually reaches a minimum (40-60% sequence identity threshold). When including remote homologs, the ratio tends back to zero, indicating that the signal is weakened by poorly conserved residues in the interface. Notably, the PPI fingerprint of triosephosphate isomerase remains constant once it reaches the minimum. This confirms the high conservation of the interface across the family. Triosephosphate isomerase enzymes are obligate homo-dimer [111] and this might explain the very strong conservation signal found also including remote homologs.

3.2 RESULTS

3.2.1 *Discriminating crystal contacts vs. biological contacts*

We investigated whether PPI fingerprints could be applied to help discriminate between crystal contacts and biologically relevant protein interactions. For this purpose, we computed the PPI fingerprint curves on a recent manually curated dataset of interactions [112]. This dataset is composed of two classes of protein contacts: crystal artifacts (82 interfaces), deriving from the tight packing of proteins in crystals, and biological contacts (83 interfaces), which correspond to biologically relevant interaction of protein chains. The dataset was created with stringent crystallographic quality criteria, including only experimentally

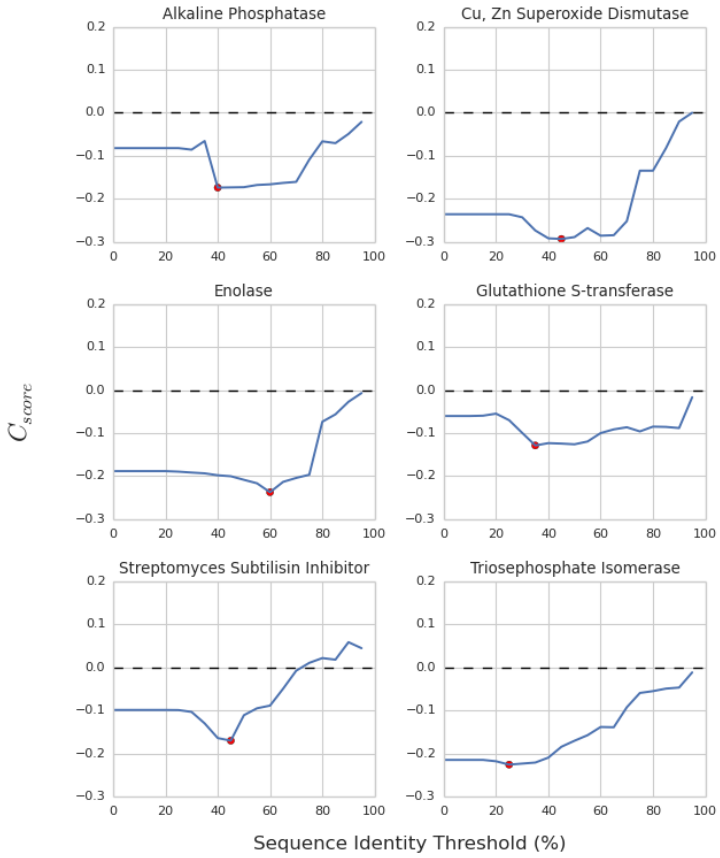


Figure 11: PPI fingerprint for conserved homo-dimers. All the analyzed protein show negative conservation score indicating that the interface is more conserved than the surface of these proteins. The minimum of the PPI fingerprint curve (indicated by the red dot) is found in the 40-60% sequence identity threshold range. These minima are generally showing a lower conservation score than the score obtained using no sequence identity threshold. A notable exception is the triosephosphate isomerase: for this protein, the ratio reaches the minimum at 30% sequence identity ad levels on this value.

confirmed quaternary structures, and focusing on small interfaces (up to 2,000 Å^2) where the discrimination is more difficult.

Our results indicate that PPI fingerprints calculated from the crystal contacts group have a constant median around zero, while in the biologically relevant class we clearly observe a significant shift towards negative values (Figure 12) that makes discrimination easier. We compared the conservation score distributions for crystal and biological interfaces using the Mann-Whitney test: the p-values for distributions between 35-55% inclusion thresholds are significantly lower than those obtained using the full MSA, in agreement with the finding by Duarte *et al.* [112].

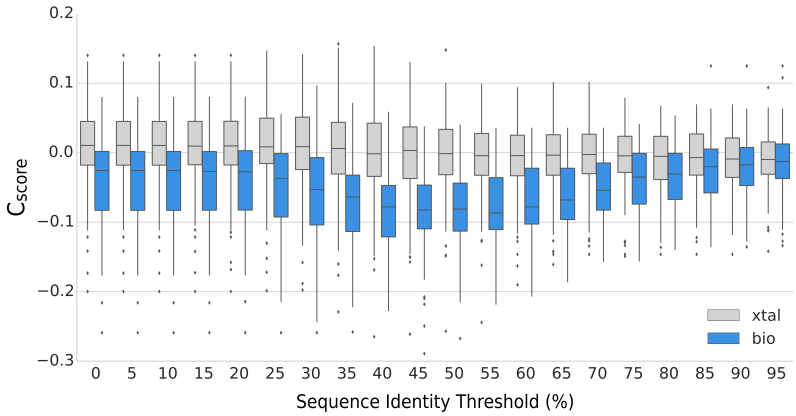


Figure 12: PPI fingerprints of the proteins in the Duarte *et al.* dataset. Biological interfaces (bio) are shown in blue, crystal contacts (xtal) in gray. Using different sequence identity inclusion thresholds (x-axis) to generate the MSAs, we see the conservation score (y-axis) helping to discriminate between crystal contacts and biological relevant interfaces. Using an inclusive MSA (0-25% sequence identity) the two distributions overlap to a large extent (Mann-Whitney p-values between 8.12×10^{-7} and 3.82×10^{-8}), while in the range between 35-55% they are clearly separable (Mann-Whitney p-values between 7.47×10^{-11} and 4.56×10^{-13}).

3.2.2 *PPI fingerprint of homologs*

In several protein families we find a mixture of different oligomeric conformations. For example, in the fructose bisphosphate aldolase (FBA) family we find a mixture of dimers and tetramers (blue and green dots in Figure 13A). The described PPI fingerprint approach can be used to compute the interface conservation as function of sequence divergence for the pro-

teins of known structure in the family. The resulting PPI fingerprint curves are grouped depending on the stoichiometry of the complex (blue and green curves in Figure 13B).

Both the groups have values below zero indicating a higher mutation rate of surface residues compared to those at the interface, confirming the overall interface conservation for the protein family in both oligomeric states. Interestingly, when remote homologs below 40% sequence identity are included the dimers' curve has a stronger conservation signal than the tetramers' one. When including only close homologs (above 60% sequence identity) the picture changes and a stronger evolutionary support is attributed to the tetramers. That is, alternative oligomeric states will have different PPI fingerprints and thus provide additional criterion for quaternary structure prediction.

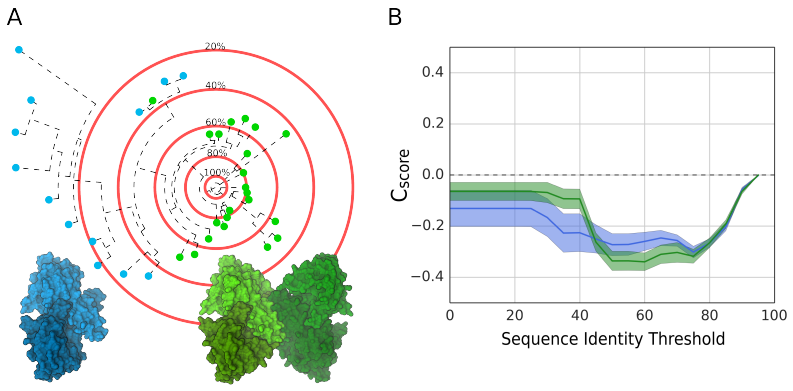


Figure 13: PPI fingerprint of fructose bisphosphate aldolase homologs. (A) The idealized sequence space of fructose bisphosphate aldolase represented as a phylogenetic tree rooted on a specific sequence. In this class of protein we can find either dimers (PDB code: 4A21, in blue) or tetramers (PDB code: 3EKL, in green). The red concentric circles represent the sequence identity thresholds used to calculate the interface conservation score (C_{score}). (B) The PPI fingerprint curves of homologs with dimeric (blue) or tetrameric (green) quaternary structures (standard error is used for the error area).

3.3 DISCUSSION

Our findings on the behavior of interface conservation expressed as a function of evolutionary distance (PPI fingerprint) are in agreement with the results obtained by Duarte *et al.* [112] and

provide a fine-grained description of protein family interaction landscape. This information, orthogonal to interaction energy considerations, helps in the differentiation between biologically relevant interactions and crystal contacts. Differently from other approaches with the same intent (e.g. EPPIC [112] and PISA [7]), we consider the interface as a whole thus avoiding the problem of combining different binary interfaces.

When the PPI fingerprint concept is applied to homology modeling, it provides additional criteria to support one quaternary structure hypothesis over another, as illustrated in the FBA case. Proteins assuming different oligomeric conformations in different organisms can be recognized, thus providing a hint on which families need deeper analysis for the determination of quaternary structure.

4

MODELING OLIGOMERS

We aim to exploit structural information available from interacting homologs (interologs) in order to predict the structure of a complete protein assembly, including its quaternary structure. Classical automated homology modeling approaches are designed to accept one or two amino acid sequences as input.

This approach is perfectly valid when the intent is to predict homo-oligomers or hetero-dimers, but to account for heterogeneous complexes (e.g. proteasome) this paradigm needs to be extended to accept multiple amino acid sequences. In general, homology modeling pipelines comprise the following steps:

1. Template search
2. Template clustering
3. Template ranking
4. Model generation
5. Model assessment

In the case of oligomeric modeling the three initial steps, resulting in the selection of templates to be modeled, need to be adapted as the definition of template must take into account heteromeric targets.

Model generation and assessment steps do not need specific adaptations, as the expected input is a structure and no assumptions are made on its composition. In this chapter we describe how templates search, clustering, and ranking are performed and validated.

4.1 METHODS

4.1.1 *Template search*

Each of the N query target sequences is independently searched against the PDB using tools such as BLAST [32] or HHblits [35]. This results in N sets of possible target-template alignments each referring to a specific PDB structure. The intersection between all the sets highlights possible heteromeric template candidates. We filter these candidates applying stringent criteria:

1. Each query input sequence must have at least one homolog chain in the template
2. Different target sequences cannot be mapped to overlapping fragments of equivalent chains in the template structure
3. The fraction of the template structure that is mapped to target sequences must be topologically connected (i.e. chains must physically interact to form a complex)

This leaves us with a set of heteromeric templates. Each heteromeric template is composed of different target-template alignments covering all the query target sequences are mapped to all the identical chains in a biological assembly.

4.1.2 *Template clustering*

Grouping similar interfaces and organizing them in a template library is a crucial starting point for an efficient structure or sequence based search [113]. This explains the multitude of databases (SCOPPI [114], ProtCID [98], 3DID [115], DOMMINO [116], and DOCKGROUND [117] amongst others) that collect and organize, using different criteria, available binary interfaces deposited in the PDB. The same level of attention has not been devoted to interfaces between multiple partners and, as of today, 3DComplex [118] is the only resource focused on a whole-complex perspective.

The authors of 3DComplex implemented a customized hierarchical grouping of assemblies based on their topology (a simplified graph representation of the direct interaction between subunits), their composition in terms of SCOP superfamilies and several layers representing different sequence similarity clustering. While 3DComplex proved to be useful to study the evolution of oligomeric complexes [100], it is not suitable to function as a template library due to the incomplete coverage of the PDB repository (SCOP annotation covers only 38% of the PDB repository).

Hence, we defined an ad hoc hierarchical clustering aware of entire complex topology as well as interatomic contacts occurring at the interface. The clustering is based on three hierarchical levels which represent structural organization of biological complexes as represented in Figure 14.

The first level describes the nature of the interacting subunits and is characterized by three possible states: we distinguish templates composed by a single polypeptide chain, labeled as "mono"; templates composed by two or more different chains, labeled as "hetero"; templates with two or more identical chains, labeled as "homo". The second level is based on the stoichiometry of the complex, so the amount of chains with a specific sequence. The last level clusters templates using an agglomerative hierarchical clustering approach based on QS-score distance measure. All pairwise distances are computed and complete linkage hierarchical clustering is performed.

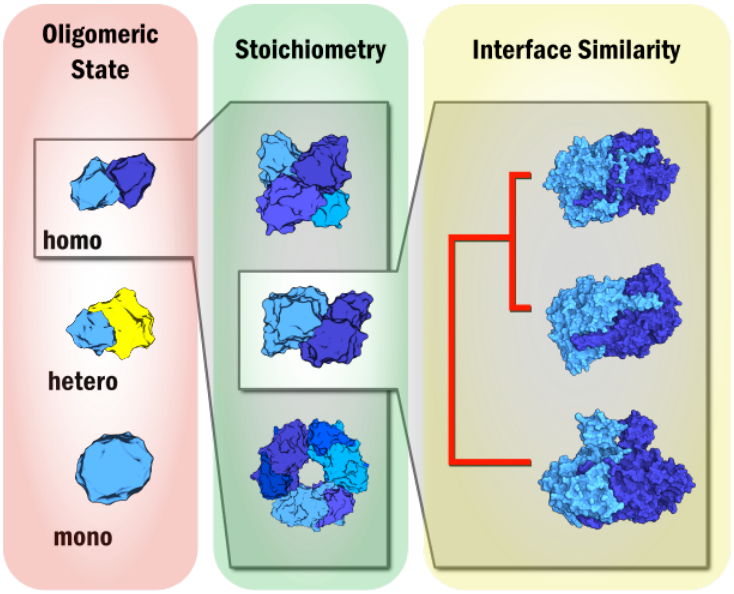


Figure 14: Clustering scheme for homologous assemblies. Templates are clustered depending on (i) their oligomeric state, (ii) their stoichiometry and (iii) on the structural similarity of interfaces based on pairwise QS-score distance measures.

4.1.3 *Template ranking*

Once we compiled a list of possible templates, we have to select those that are likely to result in useful models. The simplest selection rule would be to select templates with the highest sequence identity to the target sequence.

However, sequence identity might be a poor criterion for the modeling of protein assemblies as we showed that even at high sequence identity, homologs can assume different assembly architectures. Hence, we used a supervised machine learning ap-

proach on a dataset of known structures to train an interface quality predictor that will perform the ranking using a set of template features. The steps to train the predictor are the following:

1. Generate a dataset of targets with known structure
   a) Search templates for each target
2. Generate models
   a) Measure the distance between models and the native structure
3. Train an interface quality predictor
   a) Measure the templates features
   b) Split the dataset for cross-validation
4. Validate and assess the accuracy of the ranking

4.1.3.1 *Dataset generation and template search*

We compiled a dataset (TARGET) of non-redundant proteins with experimentally validated quaternary structures. The homo-oligomers dataset is derived from the PiQSi database [119]. PiQSi comprises 20,000 annotated biological units that we reduced by culling the sequences with PISCES [120] on a 25% sequence identity basis. We visually inspected entries with multiple binding modes to select those that are described in the respective paper. For hetero-oligomers we started from the complete list of PDB entries annotated as hetero-complexes.

As an initial filter we removed complexes that are marked as hetero-oligomers because of their interaction with antibodies or short peptides (below 20 amino acids). We filtered out complexes with an average per binary interface BSA below 250 $Å^2$ and having unconnected components. We then culled the set in order to get high quality representatives of unique interactions (with a resolution of at least 3.0 Å).

To reduce the redundancy we clustered the subunit sequences by a 30% sequence identity threshold using CD-HIT [99] and we grouped complexes whose chains belonged to the same set of clusters. We kept only the most inclusive assemblies (i.e. subcomplexes were discarded). Finally, we structurally clustered the complexes using CATH [3] domain annotation retaining only those that had a unique set of domains at the topological level.

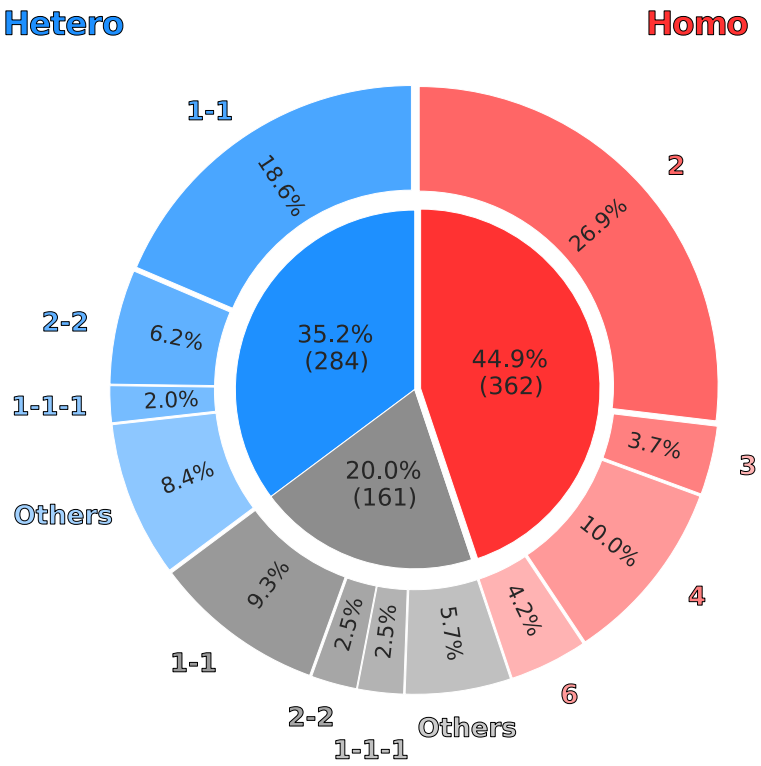


Figure 15: Stoichiometries of target proteins in our TARGET dataset. We have 807 targets in our dataset. Homo-oligomers are represented in shades of red, while hetero-oligomers in shades of blue. In shades of gray the heteromeric targets for which no template could be identified. Each wedge of the pie chart is annotated with the fraction of the total dataset for the most common stoichiometries.

The resulting dataset (807 targets) is equally composed of homo-oligomers (362 targets) and hetero-oligomers (445 targets) of varying stoichiometries as reported in Figure 15. For each of the TARGET dataset proteins we performed an extensive template search using SWISS-MODEL [121], while for heteromeric target we used the procedure described in Section 4.1.1.

To avoid bias from homologs too close to the target proteins, we removed target-template pairs having a sequence identity higher than 95%. The largest fraction of complexes deposited in the PDB - which as of today contains about 120,000 entries - is composed of homo-oligomers (more than 40,000 entries), whereas hetero-oligomers are scarcer (in the order of 14,000 structures). It is hence not surprising that for all homomeric targets at least one template could be identified, while for 36% (162) of the heteromeric targets no homologous complex was identified.

4.1.3.2 *Model generation and distance to native structure*

All the potential templates obtained from the template search were then used to generate models of the target protein and collected in our MODEL dataset. The models did not undergo any refinement. Un-aligned regions and side-chains were removed from the template structure with the exception of Cβ atoms (i.e. only backbone and Cβ atoms of the template are transferred to the model).

Each model was annotated with the QS-score to the native structure and the set of features that will be described in Section 4.1.3.3. For the sake of an unbiased learning step, all models are grouped by target. This way, during cross-validation, the set of targets can be randomly divided in testing and validation sets avoiding similar models of a same target to be used at the same time for testing and validation.

Since, for each model, the experimental reference structure is known, we can directly compare and measure their QS-score to the native structure (i.e. the fraction of correctly modeled interface residues). The accuracy of the produced models is reported in Figure 16.

Models with an incorrect stoichiometry have QS-scores consistently below 0.5 while correct stoichiometries distribute preferentially toward high QS-scores values peaking at around 0.7. The number of completely incorrect models with very low QS-score is anyway high. It is hence important to rank the tem-

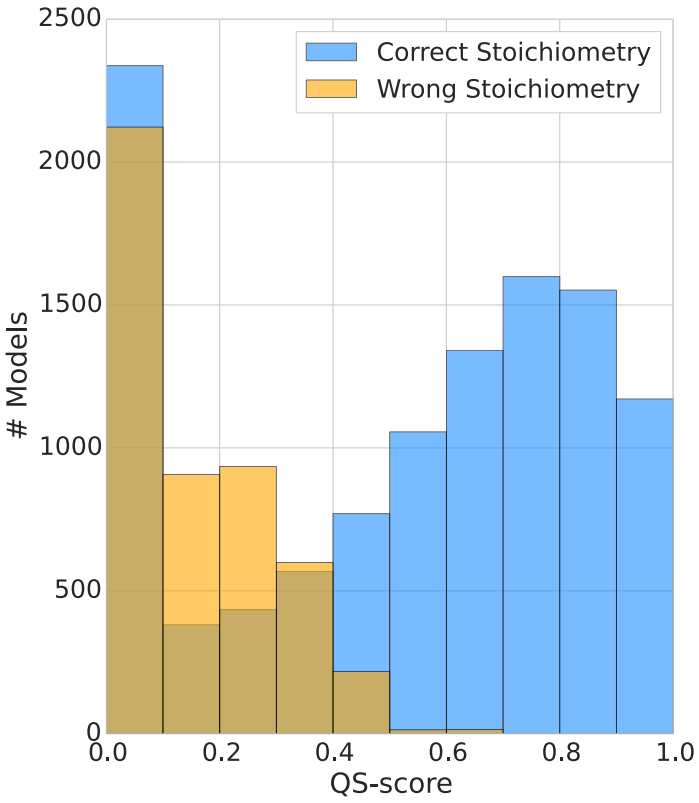


Figure 16: QS-score distribution for all produced models compared to the native structure. For both, model with a correct (in blue) or incorrect (in yellow) stoichiometry, a sizable fraction of models have an interface different from the native one as they are based on a template having the wrong quaternary structure.

plates and favor those templates leading to correctly modeled interfaces.

4.1.3.3 *Interface quality predictor*

Machine learning techniques have been frequently adopted in the context of quaternary structure prediction and prevalently applied to the problem of discriminating crystal vs. biological contacts [122–124] and for the prediction of PPI interfaces [125]. In this study, we employ a supervised learning approach using support vector machines (SVM) to predict the expected model-target QS-score given a set of template features. SVMs are scalable to large datasets and they can capture non-linear relationships using kernel functions.

The complete dataset that will be used for machine learning is composed of more than 20,000 models from a total of 657 different complexes. Our aim is to identify which features of the obtained target-template alignment would aid in the selection of templates leading to a correct quaternary structure model.

For this purpose we measure four kinds of properties: (1) sequence properties, (2) MSA properties, (3) QS consensus properties and (4) interface composition properties (Figure 17). Sequence properties include sequence identity, similarity, and an agreement measure of secondary structure and accessibility prediction. These features are computed for the different structural regions of the template: (i) the entire structure, (ii) the template's interface residues, (iii) the core residues, and (iv) the surface residues.

The MSA properties are derived from the target's family alignment. These include average profile entropy and the template e-value obtained from the HHblits run as well as the previously described PPI fingerprint. For the latter, we rely on the template interface fraction that is mapped on the target sequence for which we compute the PPI fingerprint curve. We summarize the resulting PPI fingerprint curve by the minimum of the curve, its area, the absolute maximum, and the conservation score obtained considering the full MSA.

To derive QS consensus properties, we first cluster templates hierarchically by (i) oligomeric state (i.e. being monomers, homo or hetero-oligomers), by (ii) stoichiometry and (iii) using the previously described QS-score measure. The QS consensus properties are then calculated as a template's cluster size relative to the total number of homologs considering the different levels (i-iii) of clustering.

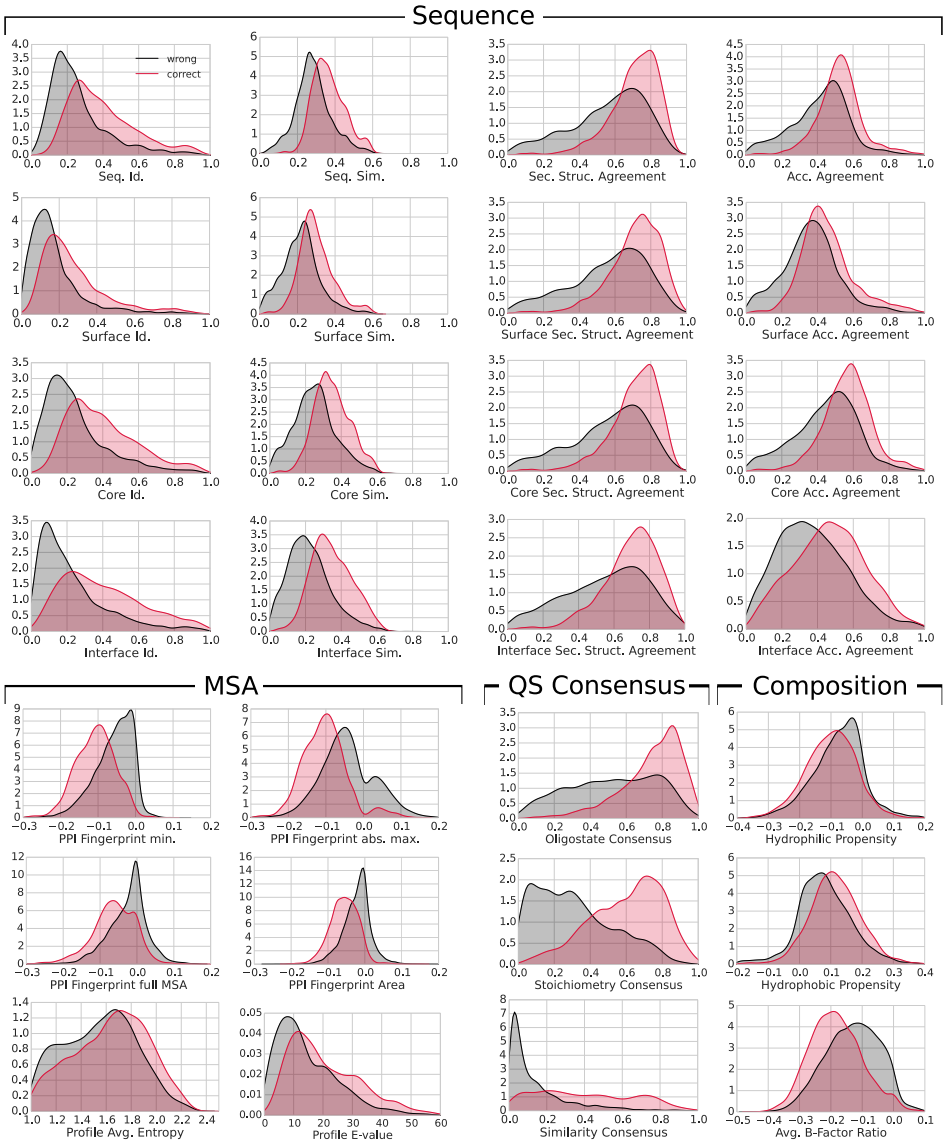


Figure 17: Distribution of mostly correct (red) and mostly incorrect (black) models with respect to various features. Mostly correct models are those having a QS-score with the known target structure of $\geqslant 0.5$ and mostly incorrect models are those with QS-score < 0.5.

Composition features are defined as in [126] by comparing the relative hydrophobic and hydrophilic composition of interface and surface residues. The composition in terms of temperature factors (B-factors) is also considered as it was shown to have discriminative power between crystal contacts and biological interfaces [127]. All the different properties are weighted (scaled) according to the coverage of the target sequence (i.e. the fraction of target residues mapped on the template).

4.1.3.4 *SVM hyper-parameter selection*

As we do not expect linear relationships between QS-score and the features we are considering, we used a Radial Basis Function (RBF) kernel for the SVM. Two parameters C and γ need to be considered when using RBF kernels. C is the parameter for the margin cost function. It is affecting the trade-off between frequency of error and stability of the prediction. A low value of C makes the margins of the trained function smoother, while a high value of C aims at making as few errors as possible.

γ is a parameter, specific for Gaussian kernels, that relates to the σ ($\gamma = \frac{1}{2\sigma^2}$) of the Gaussian function $K(x, x') = e^{(-\gamma\|x-x'\|^2)}$. A small γ underlies a Gaussian with a large variance, so a support vector can influence another even at high distances. An high value of γ will instead tend to overfit the data.

Our dataset of models was divided in a train-test set (70%) and a validation set (30%). The train-test dataset was further divided in a train set (70%) and a test set. A grid search in combination with a 10-fold cross-validation was performed on the train-test set to fine tune the C and γ hyper-parameters and avoid overfitting.

C and γ were scanned in the logarithmic (base 10) range from 1×10^{-7} to 1×10^{3}. Since we are interested in the ability of the predictor to rank templates, we used Spearman's rank correlation coefficient as fitness function. For each pair of parameters values we report the results of the 10 fold cross-validation for train and test sets (Figure 18).

We have clear over-fitting when the fitness on the train set is higher than the one of the test set. When $\gamma \geqslant 0.1$ the performance on the train set reaches correlation values as high as 1.0, but the performance on the test is very low. With low values of γ the prediction performances on train and test are close, indicating that the learned function is able to generalize well on unseen data. With $\gamma < 0.1$ the influence of the C parameter be-

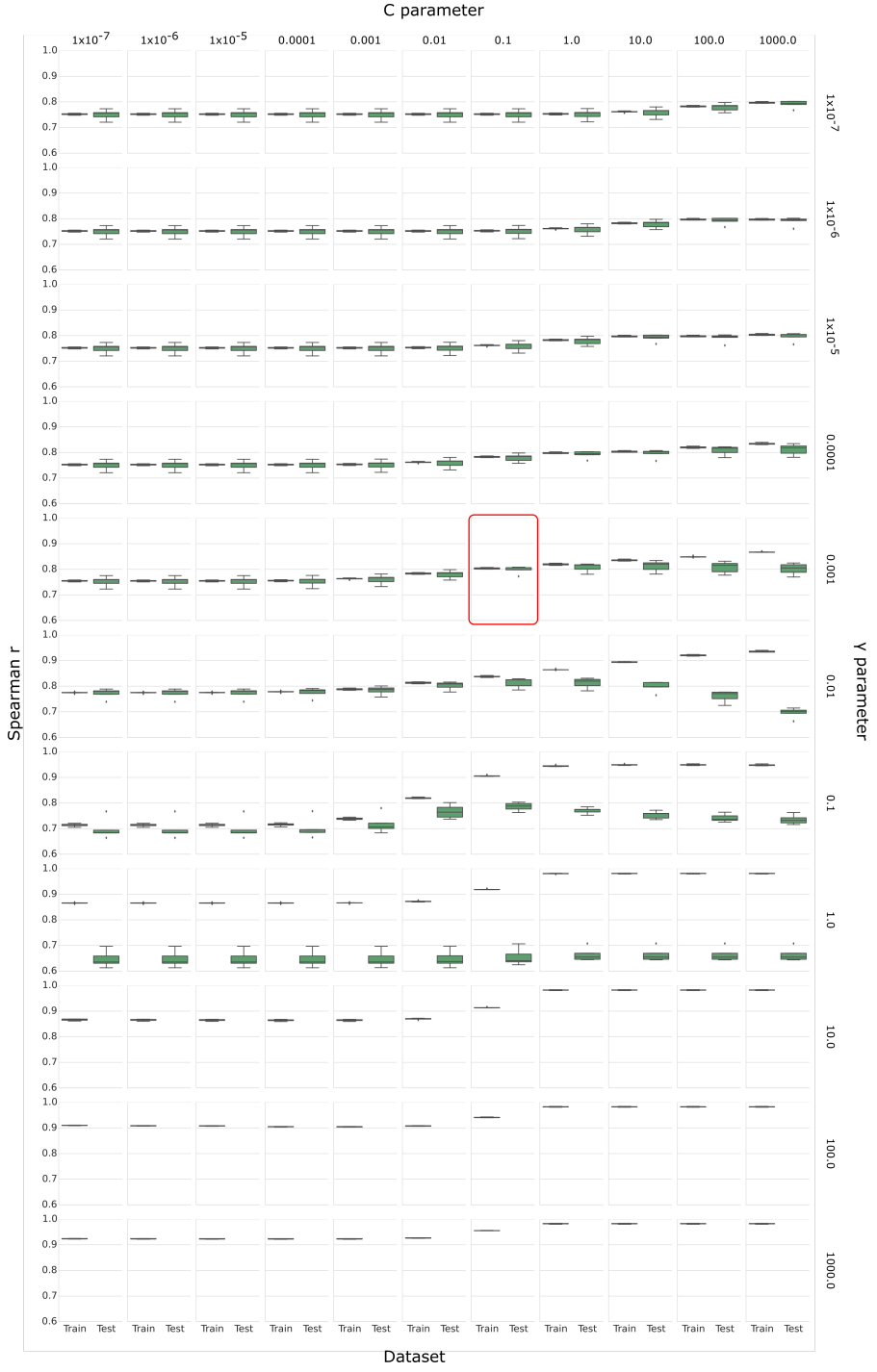


Figure 18: Grid search for C and γ parameters. Each plots report the Spearman r performance of 10 predictors for the training and testing sets. The C and γ values are reported on the top and right margin. The selected pair of parameters is the one in the red highlighted plot.

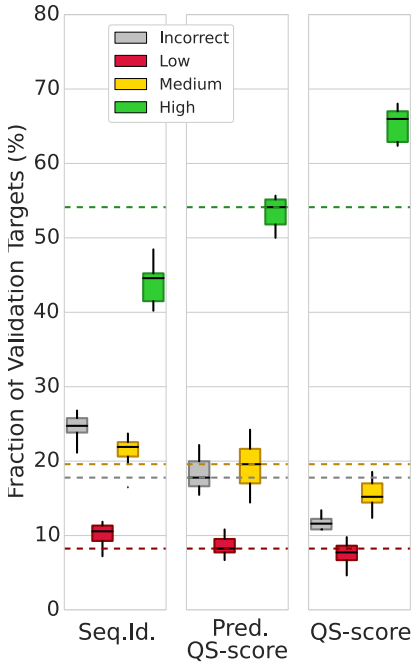


Figure 19: Fraction of validation targets in each quality category for top ranked models. The CAPRI like evaluation is used. Three ranking criteria are considered: the naïve sequence identity (Seq. Id.), our SVM prediction (Pred. QS-score) and the perfect ranking based on the QS-score distance from the native structure (QS-score). The fraction of validation targets are computed for the ten different cross-validation iterations.

comes relevant. With very low C values ($C < 0.0001$), allowing more errors, the correlation coefficient is stable at 0.75. Increasing the C parameter we allow less errors and the fitness reaches 0.8 before over-fitting again.

The values of the parameters which maximize Spearman correlation without overfitting the training data is $C = 0.1$ and $\gamma = 0.001$. The resulting predictors were used to rank templates of the validation set.

4.2 RESULTS

4.2.1 *Template ranking by interface quality prediction*

To assess the ability of the predicted QS-score to correctly rank the models we used an evaluation scheme analogous to the one used the CAPRI experiment (Critical Assessment of Prediction of Interactions) [15]: the quality of models with a QS-score below 0.1 is deemed as "incorrect", between 0.1 and 0.3 as "low", between 0.3 and 0.7 as "medium", and higher than 0.7 as "high". For each validation target the model generated from the top scoring template, in terms of predicted QS-score, was compared to the reference structure and assigned to one of the quality categories.

The results are summarized in Figure 19 where the SVM-predicted QS-score is compared to other ranking criteria: the sequence identity criteria would always rank first the model whose template has the highest sequence identity to the target sequence whereas the QS-score criteria ranks models according to their distance from the native structure (i.e. the perfect but hypothetical ranker). Looking at the latter criterion, we can observe that a consistent fraction of the validation target can be modeled with a high quality (median of 66%). The naïve idea of selecting the models with highest sequence identity provides a high quality model only in 45% of the cases. Our SVM prediction approach improves the ranking significantly with a median of 54%.

This improvement is highlighted by the lower fraction of incorrect, low, and medium quality models. To have an idea of the importance of each feature we trained predictors using only single features (Figure 20). The combination of all the features outperforms all the single feature predictors; however, most of the sequence descriptors can correctly rank 45% of the validation targets.

4.3 CASE STUDIES

4.3.1 *Modeling fructose bisphosphate aldolase in Haloferax volcanii*

Fructose bisphosphate aldolase (FBA) is a crucial enzyme in the glycolysis pathway splitting the hexose ring of fructose 1,6-bisphosphate (FBP) into two triose sugars: glyceraldehyde 3-phosphate (GAP) and dihydroxyacetone phosphate (DHAP). FBAs are divided into two classes depending on their mechanism of action: class I aldolases form reaction intermediates by covalently linking the DHAP to a conserved lysine in the active site; class II aldolases instead rely on the presence of a metal cofactor [128]. The quaternary structure of class I aldolases (found mostly in eukaryotes) is homo-tetrameric, while class II aldolases (found in prokaryotes and lower eukaryotes) can be found with different stoichiometries, the most common being homo-dimer or homo-tetramer [129–131].

We illustrate the application of our approach on the example of a class II FBA from *Haloferax volcanii* (UniProt AC: D4GYE0). No crystal structures of this specific enzyme or of homologs having closely related amino acid sequence are available. The result of templates structural clustering is reported in Figure

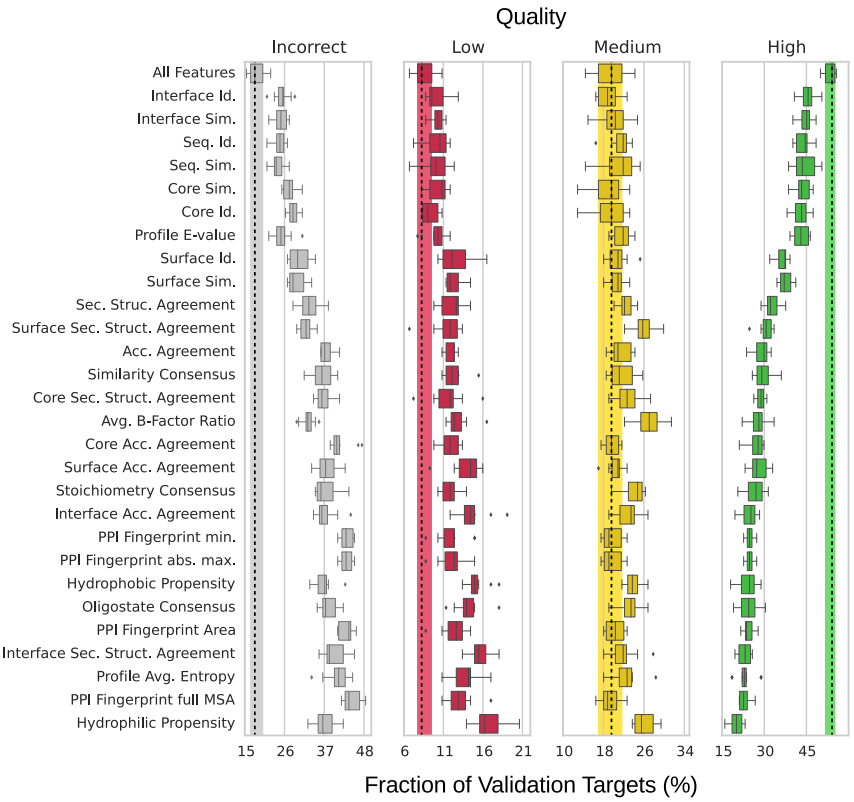


Figure 20: Fraction of validation targets in each quality category for top ranked models using single features for learning. The features are sorted in descending order based on the median of the high quality category performance. As reference the result obtained using all the features is reported and the vertical bar spans from the 25th to the 75th quartile with the median highlighted by a vertical dashed line.

21A in a decision tree style. Sequence identity highlights two clusters of dimeric and tetrameric templates, but does not allow for a finer differentiation as all the highlighted templates spans in the range between 25-35%.

A more indicative feature is the PPI fingerprint curve for these two groups (Figure 2215). The dimeric and tetrameric interfaces follow two different patterns. The conservation score obtained using a complete MSA is almost equal for both the dimeric and tetrameric options, with tetramers being slightly more conserved. The minimum for both the curves is between 30% and 40% sequence identity which is the typical distance between most of the FBAs. Using more stringent sequence identity thresholds (40-80%) the indication for dimeric interface conservation is stronger reaching lower absolute values. Thus we can state that the dimeric interface is more conserved than the tetrameric interface among close homologs even in absence of direct structural evidence.

The QS-score predictor we trained is able to capture the discussed trend and assign a higher score to dimeric templates (predicted QS-score higher than 0.5 are indicated by the green thread on the decision tree). This protein was indeed proven to be homo-dimeric [132] by gel filtration chromatography and molecular weight consideration. Notably, no aldolases were included in training or validation set; nonetheless our predictor is able to generalize on this unseen protein family and correctly assigns high predicted QS-scores to dimeric templates. This example illustrates that models built using the quaternary structure of templates having a high predicted QS-score will most likely agree with the experimentally determined structure.

4.3.2 *Modeling the urease complex in Yersinia enterocolitica*

The urease enzyme (EC number: 3.5.1.5) catabolizes urea into carbonic acid and ammonia. This is a neutralization reaction which effectively protects the bacteria from acidic environments [133, 134]. Here, we want to charactherize the oligomeric assembly of urease in *Yersinia enterocolitica*. This organism is psychrophilic, gram-negative enterobacteria naturally present in our environment. It is responsible for infections in humans ranging from mild enteritis through food ingestion, to more severe lymph node infections, arthritis and septicemia.

The urease enzyme is found throughout animal kingdom with the exception of mammals. Urease genes are well con-

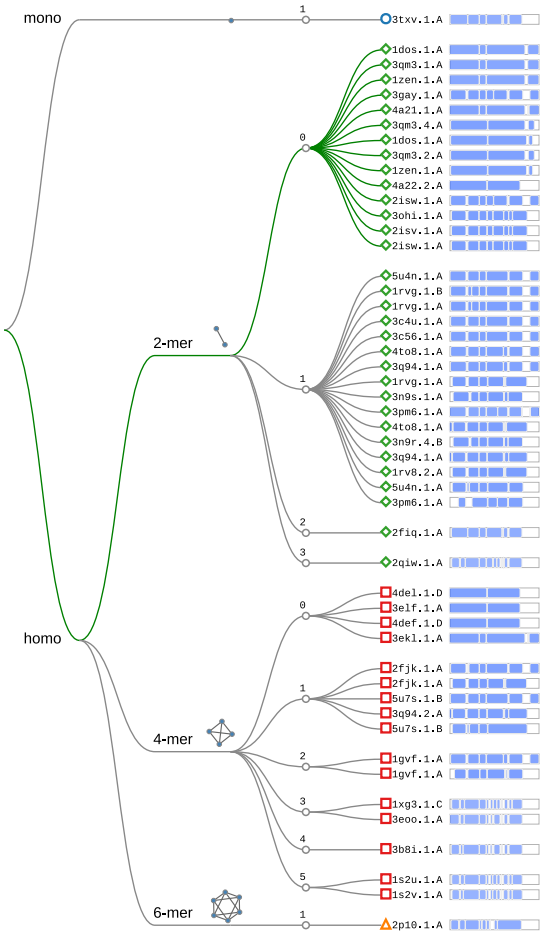


Figure 21: Structural clustering tree of *H. volcanii* FBA homologs with known structure. Each leaf is a template labeled with the PDB code and a bar indicating sequence identity and coverage. The decision tree follows the levels of clustering: oligomeric state, stoichiometry (the topology of the complexes is also shown as a small graph), and QS-score clustering. The green thread indicates templates with a predicted QS-score higher than 0.5. The highlighted cluster includes both dimers (orange) and tetramers (lilac) that do not clearly segregate just considering sequence identity.

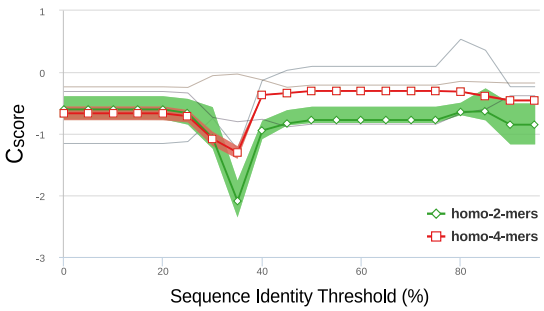


Figure 22: The PPI fingerprint curves of the dimeric (orange) and tetrameric (lilac) sets (the area plot spans between the 25th and 75th percentiles). The dimeric forms have a stronger interface conservation signal with respect to the tetrameric form. This is observable using different evolutionary distance thresholds, notably taking into account the entire MSA would not highlight a diverse conservation pattern.

served, as illustrated by 57.5% sequence identity between the distant Jack-bean (*Canavalia ensiformis*) and *Proteus mirabilis* sequences [135]. While bacterial ureases are encoded by two or three genes [136], the plant and fungal urease enzymes have a single gene probably as consequence of a single gene fusion event [137].

Gene fusion is a key determinant in the evolution of protein architecture. Following a recombination event, gene fusion can result in a new hybrid gene encoding for a multi-domain protein. Gene fusion events impact regulation or the function of existing genes and have been linked to tumor genesis and leukemia [138]. The detection of fused-genes has been used to predict protein-protein interactions [139, 140]. Interestingly, it has also been noted that fusion events tend to optimize assembly pathways of protein complexes [101], revealing a tight connection between quaternary structure and evolution.

The quaternary structure of ureases reflects their genetic organization (Figure 23) and thus, plants and fungi can only form homo-oligomers, while bacterial ureases are found as heterodimeric or heterotrimeric complexes. The Jack-bean (PDB code: 3LA4) [141] and the Pigeon Pea (*Cajanus cajan*, PDB code: 4G7E) [142] ureases are at the moment the unique representatives of plant urease structures and their single peptide chain forms a stack of trimers in a dihedral D3 symmetry. The available crystal structures of bacterial ureases reveal that the *Klebsiella aerogenes* [143] and *Sporosarcina pasteurii* [144][69] ureases arrange as a single trimer with cyclic C3 symmetry containing the α, β and γ-

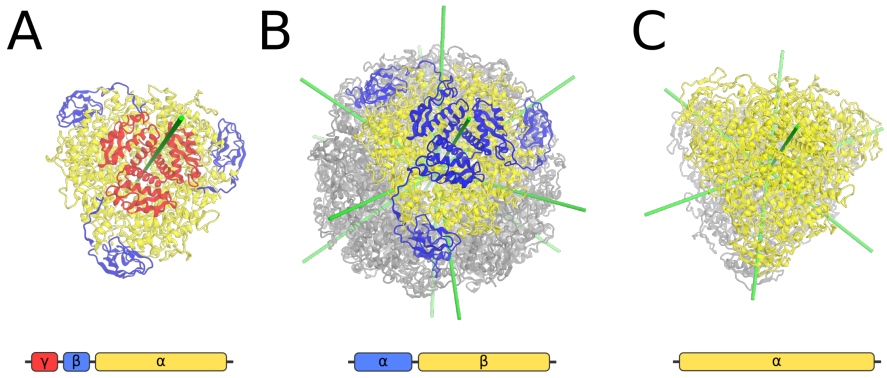


Figure 23: Urease symmetries and genetic organization. (A) The *K. aerogenes* urease (three genes) adopts a cyclic C3 symmetry with a $(\alpha\beta\gamma)_3$ stoichiometry; (B) the *H. pylori* has a tetrahedral T symmetry where the γ and β genes are fused (counter intuitively the fusion product is called α and the larger subunit is referred to as β) with a $((\alpha\beta)_3)_4$ stoichiometry; the Jack-bean urease has a dihedral D3 symmetry with a single gene and a $((\alpha)_3)_2$ stoichiometry.

subunits. Conversely, the *Helicobacter pylori* [145] and *Helicobacter mustelae* [146] ureases are tetramers of trimers arranged in a tetrahedral (T) symmetry, with each trimer containing two polypeptide chains organized with a cyclic C3 symmetry.

Previous pipelines for homology modeling incorporated the assumption that one target sequence should correspond to one homologous template chain. The *Y. enterocolitica* urease is an example where multiple target sequences may correspond to different domains of a single fused protein. Here, the relevance of gene fusion events was taken into account to model urease structures, because ignoring the possibility of gene fusion events in the template search step would result in missing homologs with the correct symmetry. The approach for template search described in Section 4.1.1 allows the use of fused proteins as templates.

One of the conditions a valid hetero template has to satisfy is that different target sequences must not be mapped to overlapping fragments of the same template chain. So if the template is a multi-domain protein resulting from a fusion event, the different target sequences will be mapped to the independent domains and the template structure divided accordingly. The heteromeric template search of our modeling server returned homologs with a variable number of subunits (Table 2). It is worth noting that only templates composed of three genes would be identified without explicitly accounting for gene fusion events.

Our interface quality predictor indicates that both a C3 and a T symmetry complex would fit, giving a higher ranking to C3 symmetry templates. On the contrary, according to multi-angle laser light scattering (MALS) result obtained by researchers of our wet lab (Nicolet *et al.*, in preparation), the *Y. enterocolitica* urease mass peak maximum is at 1.025 MDa in agreement with the reported 1.1 MDa of *H. pylori* [145], which suggest a $((\alpha\beta)_3\gamma)_4$ stoichiometry as in the *Helicobacteraceae* templates. Out of the two homologs having the stoichiometry supported by experimental data, we selected the *H. pylori* template since it shares the highest sequence identity with the *Y. enterocolitica* protein sequences.

Table 2: Analysis of fusion events with the queried *Y. enterocolitica* sequences. The sequence identity for the *Y. enterocolitica* α, β and γ-subunits is reported in the last three columns.

Organism	PDB code	Nr. of genes	Fusion events	Symmetry	Stoichiometry	Seq.Id. (%)		
						alpha	beta	gamma
S. pasteurii	4AC7	3	0	C3	$(\alpha\beta\gamma)_3$	57.9	51.8	60.6
E. aerogenes	4EP8	3	0	C3	$(\alpha\beta\gamma)_3$	59.5	54.0	60.6
H. pylori	1E9Z	2	1	T (4*C3)	$((\alpha\beta)_3)_4$	57.8	53.8	52.0
H. mustelae	3QGA	2	1	T (4*C3)	$((\alpha\beta)_3)_4$	57.1	46.4	50.0
C. ensiformis	3LA4	1	2	D3 (2*C3)	$((\alpha)_3)_2$	55.9	47.2	51.5

Structures of the *Y. enterocolitica* have been determined by EM (by researchers of the group of Henning Stahlberg, Biozentrum, University of Basel, CH) and X-ray crystallography (by researchers of the group of Timm Maier, Biozentrum, University of Basel, CH) (Nicolet et al., in preparation). We computed the structural similarity of the homology model to the X-ray and cryo-EM experimental structures by the means of RMSD, QS-score and lDDT [147] for both the entire tetrahedral complex and the C3 subunit is shown in Table 3.

We used these three distance measures to investigate different aspects of the structures. The RMSD distance measure, computed using a Cα superposition of complexes, is known to be size dependent and susceptible to outliers, but remains a widely used indicator of the overall similarity. Consequently, we added an interface similarity measurement (QS-score) and a superposition-free metric representing the agreement between interatomic distances in two structures (lDDT).

Table 3: Comparison of the *Y. enterocolitica* X-ray, electron-microscopy and homology model urease structures. We report the RMSD of the Cα trace, the QS-score and lDDT between the different structures, considering the complete T symmetry complex and the C3 subunit.

T symmetry complex	RMSD	QS-score	lDDT
X-ray vs. Model	3.19	0.81	0.91
EM vs. Model	3.50	0.79	0.90
X-ray vs. EM	0.86	0.95	0.99

C3 subunit	RMSD	QS-score	lDDT
X-ray vs. Model	2.41	0.88	0.92
EM vs. Model	2.71	0.87	0.91
X-ray vs. EM	0.85	0.95	0.99

We confirmed that the two experimental structures are nearly isomorphic (with an RMSD below 1 Å), share identical interfaces and have conserved interatomic distances. The RMSD of the model compared to both the X-ray and EM structures was below 4 Å, a small difference considering the size of the complex. The RMSD falls below 3 Å if only the C3 symmetry subunit is used as reference. The interatomic contact network is dominated by the intra-chain contacts for both the T and C3 complexes. The only difference we observed between the model and the experimental structures was in the fraction of shared interface contacts of the T and C3 complexes. The interfaces between chains in the C3 subunits are almost perfectly modeled with a QS-score between the X-ray structure and the homology model of 0.88. The interfaces between different C3 subunits considering the full T symmetry complex are a little less accurate are with a QS-score between the X-ray structure and the homology model of 0.81.

4.4 DISCUSSION

Comparative modeling of the complete architecture of homo- and hetero-oligomers starting only from their amino acid sequences is feasible and effective. To our knowledge, this is the first attempt to predict protein assemblies for a large scale cu-

rated dataset taking into account their entire quaternary structure beyond binary interactions. The models produced with the described approach have a high quality interface in 54% of the cases, which is halfway from the sequence identity baseline to the theoretical maximum given the current structural information in the PDB.

The main limitation of this method is that of relying on available templates of homologous complexes. This is most evident in the case of hetero-oligomers where we could not identify templates for 20% of the initial dataset. Thanks to the large effort of structural biology, structures of macromolecular complex are continuously unveiled at unprecedented levels of detail. This will be reflected on our approach, enabling it to model more and more precise protein-protein interfaces and assemblies.

5

SWISS-MODEL: AUTOMATED OLIGOMERIC MODELING

SWISS-MODEL is a fully automated protein structure homology modeling server with the aim of making protein modeling accessible to all biochemists and molecular biologists worldwide [121, 148–150]. We developed a modified version of the SWISS-MODEL server (available at `http://oligo.swissmodel.expasy.org`) including the interface quality predictor and the modified pipeline presented in the previous sections. As templates are identified and clustered, the features discussed in Section 4.1.3.3 are measured and serve as input for the predictor.

Given these template features, the predictor returns the expected interface quality (predicted QS-score) of a template. In the previous chapter, we showed that the predicted interface quality provides good indications for template ranking. While this is a sufficient piece of information to automate hetero-oligomer modelling, in the case of homo-oligomer a final decision must be taken on whether to model the monomer or the homo-oligomer.

5.1 METHODS

5.1.1 *Oligomeric state prediction*

We can define the oligomeric state prediction problem as a binary classification problem, where a target sequence can be either be classified as monomer or homo-oligomer. Also in this case a supervised learning approach can be followed, this time with the aim of training a binary classifier.

DATASET We compiled a dataset ("HOMO-MONO") of non-redundant proteins with experimentally validated quaternary structures. The dataset is derived from the PiQSi database [119]. PiQSi comprises 20,000 annotated biological units that we reduced by culling the sequences with PISCES [120] on a 25% sequence identity basis. We included only X-ray structures with at least 2.5 Å resolution and R-factor $\leqslant$0.3 having between 40 and 10000 residues. We visually inspected entries with multiple assemblies to select those which are described in the respective

paper. The set of homo-oligomers (362) is the same as in Section 4.1.3.1. The set of monomers is composed of 112 proteins.

FEATURES For each target protein in the HOMO-MONO dataset we perform a complete template search as described in Section 4.1.1. While in training the QS-score predictor we had many data points (20,000 templates), in this case we have a limited number of data points that is equal to the number of targets (474). Hence, we need to summarize the outcome of the template search in few features to avoid over-fitting.

We used the following features: (i) the highest predicted QS-score among templates, (ii) the oligomeric state of the highest predicted QS-score template, (iii) the fraction of homo-oligomeric templates, and (iv) the fraction of monomeric templates.

PERFORMANCE We compare the performances of a simple logistic regression with a naïve classification based only on the predicted QS-score value. Logistic regression is a regression model where the dependent variable is dichotomous (homomeric/monomeric). In this model, the probabilities of one class against the other are modeled using a logistic function. The naïve approach predicts a target to be an oligomer when the predicted QS-score is higher than a fixed threshold (0.5).

We evaluated the two classifiers using common metrics: accuracy, precision, recall, F1-score, and Matthew's correlation coefficient (MCC). We define true positives (TP) cases when the target is homomeric and the model as well, false positives (FP) when the target is monomeric but is predicted as oligomeric, true negatives (TN) when the target is a monomer and also the prediction is a monomer), and false negatives (FN) when the target is an oligomer but the prediction was monomeric. Given these definition we can compute the different metrics:

$$\mathrm{Accuracy} = \frac{TP + TN}{P + N}$$

$$\mathrm{Precision} = \frac{TP}{TP + FP}$$

$$\mathrm{Recall} = \frac{TP}{TP + FN}$$

$$\mathrm{F1-score} = \frac{2TP}{2TP + FP + FN}$$

$$\mathrm{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

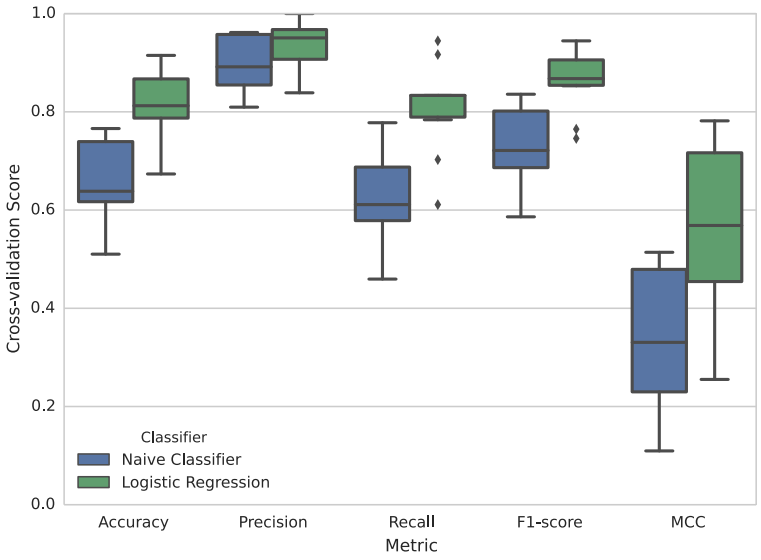


Figure 24: Performances of the naïve and logistic regression classifiers. Different scoring metrics are evaluated on a 10 fold cross-validation of the HOMO-MONO dataset. The logistic regression outperforms the naïve classifier on all measured metrics.

On a 10 fold cross-validation it is clear that even a simple logistic regression based classifier performs better than the naïve approach (Figure 24).

This is also confirmed by the Receiver Operating Characteristics (ROC) analysis on the full dataset (Figure 25). In ROC analysis we plot the recall or True Positive Rate (TPR) against the False Positive Rate ($FPR = \frac{FP}{FP+TN}$). This is a convenient way to visualize the trade-off between benefits (true positives) and costs (false positives). Some classifiers produces, along the binary prediction, a probability or confidence value. For the naïve classifier this probability can be represented by the predicted QS-score, while for logistic regression is given by the logistic function. Using different threshold on this probability value we can draw a curve in the ROC space. The more the curve tends to the top left corner the better (i.e. the predictor correctly classify all the TP without any FP prediction).

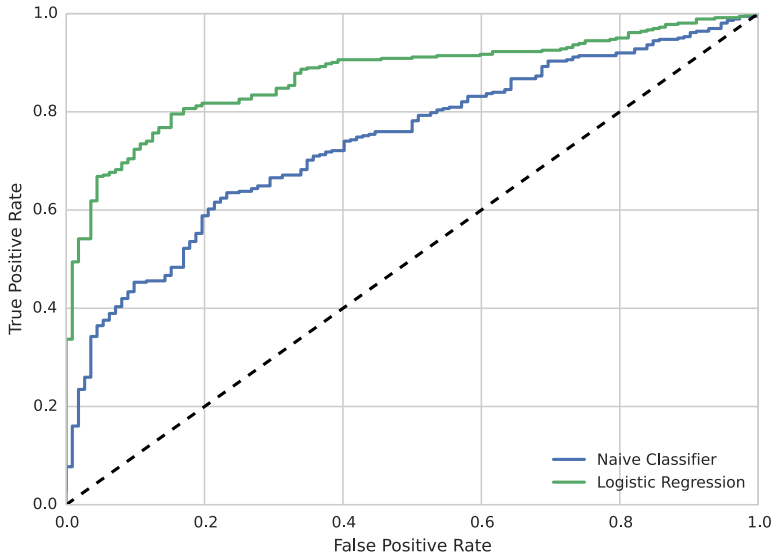


Figure 25: Receiver Operating Characteristics (ROC) analysis of the naïve and logistic regression classifiers. Logistic regression is more accurate than the naïve classifier.

5.2 RESULTS

5.2.1 *Comparison with other modeling servers*

The assessment of our automated modeling pipeline is provided by the Continuous Automated Model EvaluatiOn performed by CAMEO [151]. The CAMEO server retrieves on a weekly basis the sequences of new PDB entries that will be released the following week. The sequences are submitted to several structure prediction servers and, when the actual structure is published, the models are evaluated. Not many publicly available servers perform quaternary structure prediction.

We could analyze the quality of models produced by the classical SWISS-MODEL server [121] and Robetta [152]. A modified version of the SWISS-MODEL server including the pipeline presented in this study (SWISS-MODEL Oligo) was used for a retrospective analysis running the template search on corresponding previous releases of the PDB. The decision on whether to predict a monomer or an oligomer is based on the logistic regression previously described.

We compared models produced by these servers from August 2015 to August 2016. In this period a total of 813 targets have been submitted by CAMEO to predictors. For each

server we report in Table 4 the number of models returned, the number of true positives (i.e. the target is homomeric and the model as well), false positives (i.e. the target is monomeric but is predicted as oligomeric), true negatives (i.e. the target is a monomer and also the prediction is a monomer), false negatives (i.e. the target is an oligomer but the prediction was monomeric).

With respect to other servers, our approach is able to correctly recognize 10-11% more homo-oligomers (TP) as such at the expenses of just 4% more false positives. This corresponds to a decreased ability to recognize monomers as such (TN, 4% less), but also less oligomers are wrongly predicted as monomers (FN, 10-11% less). Overall we improve the MCC by 0.1.

Table 4: Summary of the modeling performances of SWISS-MODEL Oligo, SWISS-MODEL, and Robetta. From 2015-07-31 to 2016-08-01 a total of 813 targets (427 monomeric and 386 homomeric) have been submitted by CAMEO to these servers. For each server we report the number of models returned, the number of true positives (i.e. the target is homomeric and the model as well), false positives (i.e. the target is monomeric but is predicted as oligomeric), true negatives (i.e. the target is a monomer and also the prediction is a monomer), false negatives (i.e. the target is an oligomer but the prediction was monomeric). The percentages refer to the targets modeled by each server. In the last column the Matthews correlation coefficient is reported.

Server	Models	TP	FP	TN	FN	MCC
SWISS-MODEL Oligo	797	257 (32%)	64 (8%)	355 (44%)	121 (15%)	0.54
SWISS-MODEL	800	173 (21%)	28 (3%)	390 (48%)	209 (26%)	0.44
Robetta	789	167 (21%)	40 (5%)	379 (48%)	203 (25%)	0.40

The predictions of these three servers had a total of 111 common homo-oligomeric targets. For this fraction of targets we could compare the quality of the prediction made by the different servers. The models produced by each server are compared to the native structure using QS-score and a structural-similarity based measure, TM-score, obtained using MM-align [96]. MM-align does not perform chain mapping, so the structures with renamed chains obtained from QS-score are used instead of the original models.

The method we propose outperforms the other servers in terms of interface quality (QS-score) and in global structural similarity (TM-score) without being explicitly trained on this last distance measure (Figure 26).

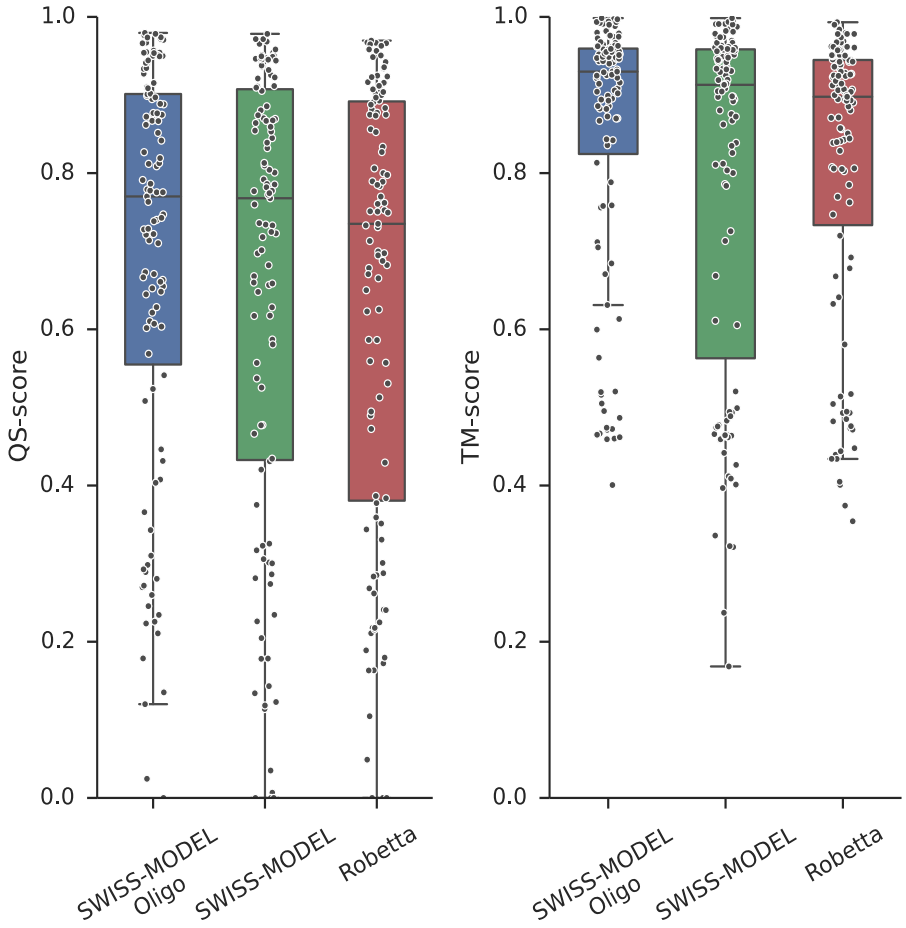


Figure 26: Comparison of model quality for three servers participating in CAMEO. The approach described in the current study (SWISS-MODEL Oligo) is compared to the classic SWISS-MODEL and Robetta servers. The top models produced by each server are compared to the native structure using two distance measures: QS-score (representing interface accuracy) and TM-score (representing global fold accuracy). Models produced by SWISS-MODEL Oligo have a quaternary structure and interface contacts closer to the native structure with respect to the other servers.

5.3 DISCUSSION

We showed that the prediction of interface quality must be complemented by the prediction of the oligomeric state of a protein. Starting from a target amino acid sequence and accounting for

homologs information, we could substantially improve the ability to correctly classify protein as homomeric or monomeric.

This increased ability to predict oligomers is not achieved at the expenses of quality of the model. We could show, comparing our prediction approach to other servers, that the quality of the modeled interface and the global structure of the oligomer are also of improved.

The method we developed is publicly available at http://oligo.swissmodel.expasy.org and can readily aid molecular biologists and biochemists by providing an overview of homologs' quaternary structural space along with the prediction made by our method.

6

CONCLUSION AND OUTLOOKS

The amount of known protein sequences and protein interaction is growing exponentially leaving detailed knowledge about protein structure and protein-protein interaction behind. In this thesis we have presented a novel approach to counteract this disparity. We described how both homo- and hetero-oligomeric structures can be modeled.

We defined a novel measure of quaternary structure similarity (QS-score), which overcome limitations of available distance measures. This measure is robust and fast enough to allow an analysis on the full set of available experimental oligomeric structures. This analysis revealed that differently from binary domain-domain interactions [95] and assembly symmetries [100] a significant fraction of closely related homologs contains assemblies with whole-complex interface structures different from each other. It is imperative to account for this diversity, when the intent is to exploit quaternary structures of homologs.

Composition and conservation of protein-protein interaction has been a long standing interest of structural biologist. The huge wealth of sequences that is causing the sequence-structure gap on one side, allows us to be more rigorous in defining multiple sequence alignments on the other. Filtering out remote homologs, we could show differential evolutionary signal coming from alternative quaternary structures in protein families. As it was done in EPPIC [112], this signal can successfully be exploited to discriminate crystal and biological contacts.

We developed and validated an approach to model homo- and hetero-oligomers. Scrupulous estimation of templates interface quality allowed us to improve template ranking over sequence only approach, approaching theoretical ranking performances and performing better than state of art tools. This approach was readily made available for the scientific community in the context of the SWISS-MODEL web server. Clearly the low amount of available experimental structures of heteromeric complexes, does not allow us to find templates and build models for the complete dataset of targets. While for binary complexes this problem can be only tackled by template based modeling of domain-domain interactions or using free

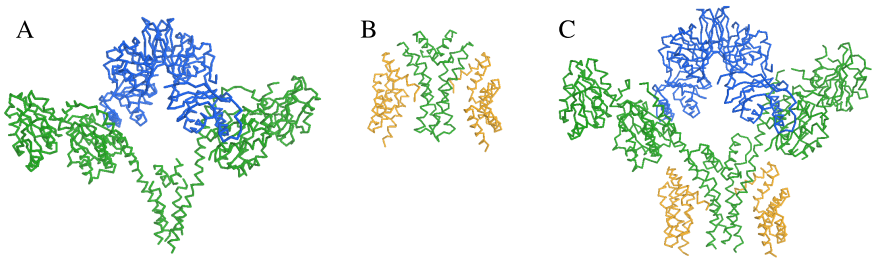


Figure 27: Example of transitive complex modeling. (A) (PDB code: 4N7R; released in 2014) A hetero-tetramer containing two glutamyl-tRNA reductase (GluTR; in green) chains in complex with its binding partner (GluBP; in blue). (B) (PDB code: 4YVQ; released in 2015) The C-terminal domain of GluTR in complex with the FLU protein (in yellow). (C) (PDB code: 5CHE; released in 2016) The complete assembly of GluTR with its regulatory proteins. The combination of the two older structures could have been used, in principle, to derive the complete complex.

docking software, for higher order interaction a stringent homology based approach might still be possible.

In analogy to multi-template modeling, the interface information coming from several interologs can be combined. The idea is to check whether some form of transitive property can be used in homology modeling: given an A′-B′ and a B″-C″ complexes, combine their structure to model the ternary complex A-B-C as illustrated in Figure 27. This should not be limited to the ternary example presented, but be applicable between any pair of complexes that share at least one homologous chain.

Obtaining a highly accurate interaction description, at atomistic level, of supramolecular assemblies would be of critical importance in several areas spanning from applied pharmacology to descriptive systems biology. Describing novel interfaces, explaining pleiotropic effect of disease mutations, can be among the possible application of this approach.

REFERENCES

[1] R. Milo and R. Phillips. *Cell Biology by the Numbers*. Garland Science, 2015. ISBN: 0815345372. URL: http://book.bionumbers.org/.

[2] L. Pauling and R. B. Corey. "Configurations of Polypeptide Chains With Favored Orientations Around Single Bonds: Two New Pleated Sheets." In: *Proc. Natl. Acad. Sci. U. S. A.* 37.11 (1951), pp. 729–40. ISSN: 0027-8424 (Print) 0027-8424 (Linking). URL: http://www.ncbi.nlm.nih.gov/pubmed/16578412.

[3] C. A. Orengo, A. D. Michie, S. Jones, D. T. Jones, M. B. Swindells, and J. M. Thornton. "CATH–a hierarchic classification of protein domain structures." In: *Structure* 5.8 (1997), pp. 1093–108. ISSN: 0969-2126 (Print) 0969-2126 (Linking). DOI: 10.1016/s0969-2126(97)00260-8. URL: http://www.ncbi.nlm.nih.gov/pubmed/9309224.

[4] A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia. "SCOP: a structural classification of proteins database for the investigation of sequences and structures." In: *J. Mol. Biol.* 247.4 (1995), pp. 536–40. ISSN: 0022-2836 (Print) 0022-2836 (Linking). DOI: 10.1006/jmbi.1995.0159. URL: http://www.ncbi.nlm.nih.gov/pubmed/7723011.

[5] C. N. Pace, B. A. Shirley, M. McNutt, and K. Gajiwala. "Forces contributing to the conformational stability of proteins." In: *FASEB J.* 10.1 (1996), pp. 75–83. ISSN: 0892-6638 (Print) 0892-6638 (Linking). URL: http://www.ncbi.nlm.nih.gov/pubmed/8566551.

[6] D. S. Goodsell and A. J. Olson. "Structural symmetry and protein function." In: *Annu. Rev. Biophys. Biomol. Struct.* 29.1 (2000), pp. 105–53. ISSN: 1056-8700 (Print) 1056-8700 (Linking). DOI: 10.1146/annurev.biophys.29.1.105. URL: http://www.ncbi.nlm.nih.gov/pubmed/10940245.

[7] E. Krissinel and K. Henrick. "Inference of macromolecular assemblies from crystalline state." In: *J. Mol. Biol.* 372.3 (2007), pp. 774–97. ISSN: 0022-2836 (Print) 0022-2836 (Linking). DOI: 10.1016/j.jmb.2007.05.022. URL: http://www.ncbi.nlm.nih.gov/pubmed/17681537.

[8] K. Henrick and J. M. Thornton. "PQS: a protein quaternary structure file server." In: *Trends Biochem. Sci.* 23.9 (1998), pp. 358–61. ISSN: 0968-0004 (Print) 0968-0004 (Linking). DOI: 10.1016/s0968-0004(98)01253-5. URL: http://www.ncbi.nlm.nih.gov/pubmed/9787643.

[9] R. John Ellis. "Macromolecular crowding: obvious but underappreciated." In: *Trends Biochem. Sci.* 26.10 (2001), pp. 597–604. ISSN: 09680004. DOI: 10.1016/s0968-0004(01)01938-7.

[10] Allen P. Minton. "Implications of macromolecular crowding for protein assembly." In: *Curr. Opin. Struct. Biol.* 10.1 (2000), pp. 34–39. ISSN: 0959440X. DOI: 10.1016/s0959-440x(99)00045-7.

[11] Ozlem Keskin, Nurcan Tuncbag, and Attila Gursoy. "Predicting Protein–Protein Interactions from the Molecular to the Proteome Level." In: *Chem. Rev.* 116.8 (2016). PMID: 27074302, pp. 4884–4909. DOI: 10.1021/acs.chemrev.5b00683. eprint: http://dx.doi.org/10.1021/acs.chemrev.5b00683.

[12] S. Jones and J. M. Thornton. "Principles of protein-protein interactions." In: *Proc. Natl. Acad. Sci. U. S. A.* 93.1 (1996), pp. 13–20. ISSN: 0027-8424 (Print) 0027-8424 (Linking). URL: http://www.ncbi.nlm.nih.gov/pubmed/8552589.

[13] John C Kendrew, G Bodo, Howard M Dintzis, RG Parrish, Harold Wyckoff, and David C Phillips. "A three-dimensional model of the myoglobin molecule obtained by x-ray analysis." In: *Nature* 181.4610 (1958), pp. 662–666.

[14] William Henry Bragg and William Lawrence Bragg. *X rays and crystal structure*. Bell, 1915.

[15] MICHAEL G Rossmann. "The molecular replacement method." In: *Acta Crystallogr. A* 46.2 (1990), pp. 73–82.

[16] DW Green, VM Ingram, and MF Perutz. "The structure of haemoglobin. Sign determination by the isomorphous replacement method." In: *Proc R Soc London, Ser A*. Vol. 225. 1162. The Royal Society. 1954, pp. 287–307.

[17] Richard Neutze, Remco Wouts, David van der Spoel, Edgar Weckert, and Janos Hajdu. "Potential for biomolecular imaging with femtosecond X-ray pulses." In: *Nature* 406.6797 (2000), pp. 752–757.

[18] Matteo Levantino, Giorgio Schirò, Henrik Till Lemke, Grazia Cottone, James Michael Glownia, Diling Zhu, Mathieu Chollet, Hyotcherl Ihee, Antonio Cupane, and Marco Cammarata. "Ultrafast myoglobin structural dynamics observed with an X-ray free-electron laser." In: *Nature communications* 6 (2015).

[19] Gabriel Cornilescu, Frank Delaglio, and Ad Bax. "Protein backbone angle restraints from searching a database for chemical shift and sequence homology." In: *J. Biomol. NMR* 13.3 (1999), pp. 289–302.

[20] Kurt Wüthrich. "The way to NMR structures of proteins." In: *Nat. Struct. Biol.* 8.11 (2001), pp. 923–5.

[21] Shang-Te Danny Hsu, Lisa D Cabrita, Paola Fucini, Christopher M Dobson, and John Christodoulou. "Probing protein folding on the ribosome by solution state NMR spectroscopy." In: *J. Biomol. Struct. Dyn.* 26.6 (2009), pp. 846–846.

[22] Guifang Wang, Ze-Ting Zhang, Bin Jiang, Xu Zhang, Conggang Li, and Maili Liu. "Recent advances in protein NMR spectroscopy and their implications in protein therapeutics research." In: *Anal. Bioanal. Chem.* 406.9-10 (2014), pp. 2279–2288.

[23] Marc Adrian, Jacques Dubochet, Jean Lepault, and Alasdair W McDowall. "Cryo-electron microscopy of viruses." In: *Nature* 308 (1984), pp. 32–36. DOI: 10.1038/308032a0.

[24] J Dubochet, FP Booy, R Freeman, AV Jones, and CA Walter. "Low temperature electron microscopy." In: *Annu. Rev. Biophys. Bioeng.* 10.1 (1981), pp. 133–149.

[25] Werner Kühlbrandt. "The resolution revolution." In: *Science* 343.6178 (2014), pp. 1443–1444.

[26] Helen M Berman, John Westbrook, Zukang Feng, Gary Gilliland, Talapady N Bhat, Helge Weissig, Ilya N Shindyalov, and Philip E Bourne. "The protein data bank." In: *Nucleic Acids Res.* 28.1 (2000), pp. 235–242.

[27] Joel L Sussman, Dawei Lin, Jiansheng Jiang, Nancy O Manning, Jaime Prilusky, O Ritter, EE Abola, et al. "Protein Data Bank (PDB): database of three-dimensional structural information of biological macromolecules." In: *Acta Crystallogr Sect D Biol Crystallogr* 54.6 (1998), pp. 1078–1084.

[28] UniProt Consortium et al. "UniProt: a hub for protein information." In: *Nucleic Acids Res.* (2014), gku989.

[29] Cyrus Levinthal. "How to fold graciously." In: *Mossbauer spectroscopy in biological systems* 67 (1969), pp. 22–24.

[30] Christian B Anfinsen. *Studies on the principles that govern the folding of protein chains*. 1972.

[31] Cyrus Chothia and Arthur M Lesk. "The relation between the divergence of sequence and structure in proteins." In: *The EMBO journal* 5.4 (1986), p. 823.

[32] Stephen F. Altschul, Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman. "Basic local alignment search tool." In: *J. Mol. Biol.* 215.3 (1990), pp. 403–410. ISSN: 00222836. DOI: 10.1016/s0022-2836(05)80360-2.

[33] Burkhard Rost. "Twilight zone of protein sequence alignments." In: *Protein Eng.* 12.2 (1999), pp. 85–94.

[34] Johannes Söding. "Protein homology detection by HMM–HMM comparison." In: *Bioinformatics* 21.7 (2005), pp. 951–960.

[35] M. Remmert, A. Biegert, A. Hauser, and J. Soding. "HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment." In: *Nat. Methods* 9.2 (2012), pp. 173–5. ISSN: 1548-7105 (Electronic) 1548-7091 (Linking). DOI: 10.1038/nmeth.1818. URL: http://www.ncbi.nlm.nih.gov/pubmed/22198341.

[36] Georgii G Krivov, Maxim V Shapovalov, and Roland L Dunbrack. "Improved prediction of protein side-chain conformations with SCWRL4." In: *Proteins: Struct., Funct., Bioinf.* 77.4 (2009), pp. 778–795.

[37] Andrej Sali and Tom Blundell. "Comparative protein modelling by satisfaction of spatial restraints." In: *Protein structure by distance analysis* 64 (1994), p. C86.

[38] Benjamin Webb and Andrej Sali. "Comparative protein structure modeling using Modeller." In: *Current protocols in bioinformatics* (2014), pp. 5–6.

[39] Torsten Schwede. *Computational structural biology: Methods and applications*. World scientific, 2008.

[40] Andrew Leaver-Fay, Michael Tyka, Steven M Lewis, Oliver F Lange, James Thompson, Ron Jacak, Kristian Kaufman, P Douglas Renfrew, Colin A Smith, Will Sheffler, et al. "ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules." In: *Methods in enzymology* 487 (2011), p. 545.

[41] Kim T Simons, Rich Bonneau, Ingo Ruczinski, and David Baker. "Ab initio protein structure prediction of CASP III targets using ROSETTA." In: *Proteins: Structure, Function, and Bioinformatics* 37.S3 (1999), pp. 171–176.

[42] Yang Zhang. "I-TASSER server for protein 3D structure prediction." In: *BMC bioinformatics* 9.1 (2008), p. 1.

[43] Dong Xu and Yang Zhang. "Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field." In: *Proteins: Structure, Function, and Bioinformatics* 80.7 (2012), pp. 1715–1735.

[44] John Moult, Jan T Pedersen, Richard Judson, and Krzysztof Fidelis. "A large-scale experiment to assess protein structure prediction methods." In: *Proteins: Struct., Funct., Bioinf.* 23.3 (1995).

[45] John Moult, Tim Hubbard, Stephen H. Bryant, Krzysztof Fidelis, and Jan T. Pedersen. "Critical assessment of methods of protein structure prediction (CASP): Round II." In: *Proteins: Struct., Funct., Bioinf.* 29.S1 (1997), pp. 2–6. ISSN: 1097-0134. DOI: 10.1002/(SICI)1097-0134(1997)1+<2::AID-PROT2>3.0.CO;2-T.

[46] John Moult, Tim Hubbard, Krzysztof Fidelis, and Jan T Pedersen. "Critical assessment of methods of protein structure prediction (CASP): round III." In: *Proteins: Struct., Funct., Bioinf.* 37.S3 (1999), pp. 2–6.

[47] John Moult, Krzysztof Fidelis, Adam Zemla, and Tim Hubbard. "Critical assessment of methods of protein structure prediction (CASP): Round IV." In: *Proteins: Struct., Funct., Bioinf.* 45.S5 (2001), pp. 2–7. ISSN: 1097-0134. DOI: 10.1002/prot.10054.

[48] John Moult, Krzysztof Fidelis, Adam Zemla, and Tim Hubbard. "Critical assessment of methods of protein structure prediction (CASP)-round V." In: *Proteins: Struct., Funct., Bioinf.* 53.S6 (2003), pp. 334–339. ISSN: 1097-0134. DOI: 10.1002/prot.10556.

[49] John Moult, Krzysztof Fidelis, Burkhard Rost, Tim Hubbard, and Anna Tramontano. "Critical assessment of methods of protein structure prediction (CASP)—Round 6." In: *Proteins: Struct., Funct., Bioinf.* 61.S7 (2005), pp. 3–7. ISSN: 1097-0134. DOI: 10.1002/prot.20716.

[50] John Moult, Krzysztof Fidelis, Andriy Kryshtafovych, Burkhard Rost, Tim Hubbard, and Anna Tramontano. "Critical assessment of methods of protein structure prediction—Round VII." In: *Proteins: Struct., Funct., Bioinf.* 69.S8 (2007), pp. 3–9. ISSN: 1097-0134. DOI: 10.1002/prot.21767.

[51] John Moult, Krzysztof Fidelis, Andriy Kryshtafovych, Burkhard Rost, and Anna Tramontano. "Critical assessment of methods of protein structure prediction—Round VIII." In: *Proteins: Struct., Funct., Bioinf.* 77.S9 (2009), pp. 1–4. ISSN: 1097-0134. DOI: 10.1002/prot.22589.

[52] John Moult, Krzysztof Fidelis, Andriy Kryshtafovych, and Anna Tramontano. "Critical assessment of methods of protein structure prediction (CASP)—round IX." In: *Proteins: Struct., Funct., Bioinf.* 79.S10 (2011), pp. 1–5. ISSN: 1097-0134. DOI: 10.1002/prot.23200.

[53] John Moult, Krzysztof Fidelis, Andriy Kryshtafovych, Torsten Schwede, and Anna Tramontano. "Critical assessment of methods of protein structure prediction (CASP) — round x." In: *Proteins: Struct., Funct., Bioinf.* 82 (2014), pp. 1–6. ISSN: 1097-0134. DOI: 10.1002/prot.24452.

[54] John Moult, Krzysztof Fidelis, Andriy Kryshtafovych, Torsten Schwede, and Anna Tramontano. "Critical assessment of methods of protein structure prediction: Progress and new directions in round XI." In: *Proteins: Struct., Funct., Bioinf.* 84 (2016), pp. 4–14. ISSN: 1097-0134. DOI: 10.1002/prot.25064.

[55] Vivek Modi and Roland L. Dunbrack. "Assessment of refinement of template-based models in CASP11." In: *Proteins: Struct., Funct., Bioinf.* 84 (2016), pp. 260–281. ISSN: 1097-0134. DOI: 10.1002/prot.25048.

[56] Lukasz Salwinski, Christopher S Miller, Adam J Smith, Frank K Pettit, James U Bowie, and David Eisenberg. "The database of interacting proteins: 2004 update." In: *Nucleic Acids Res.* 32.suppl 1 (2004), pp. D449–D451.

[57] Luana Licata, Leonardo Briganti, Daniele Peluso, Livia Perfetto, Marta Iannuccelli, Eugenia Galeota, Francesca Sacco, Anita Palma, Aurelio Pio Nardozza, Elena Santonico, et al. "MINT, the molecular interaction database: 2012 update." In: *Nucleic Acids Res.* 40.D1 (2012), pp. D857–D861.

[58] Sandra Orchard, Mais Ammari, Bruno Aranda, Lionel Breuza, Leonardo Briganti, Fiona Broackes-Carter, Nancy H Campbell, Gayatri Chavali, Carol Chen, Noemi Del-Toro, et al. "The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases." In: *Nucleic Acids Res.* (2013), gkt1115.

[59] Damian Szklarczyk, Andrea Franceschini, Stefan Wyder, Kristoffer Forslund, Davide Heller, Jaime Huerta-Cepas, Milan Simonovic, Alexander Roth, Alberto Santos, Kalliopi P Tsafou, et al. "STRING v10: protein–protein interaction networks, integrated over the tree of life." In: *Nucleic Acids Res.* (2014), gku1003.

[60] Albertha JM Walhout and Marc Vidal. "High-throughput yeast two-hybrid assays for large-scale protein interaction mapping." In: *Methods* 24.3 (2001), pp. 297–306.

[61] Laurent Terradot, Nathan Durnell, Min Li, Ming Li, Jeremiah Ory, Agnes Labigne, Pierre Legrain, Frederic Colland, and Gabriel Waksman. "Biochemical Characterization of Protein Complexes from the Helicobacter pylori Protein Interaction Map Strategies for Complex Formation and Evidence for Novel Interactions Within Type IV Secretion Systems." In: *Molecular & Cellular Proteomics* 3.8 (2004), pp. 809–819.

[62] Nevan J Krogan, Gerard Cagney, Haiyuan Yu, Gouqing Zhong, Xinghua Guo, Alexandr Ignatchenko, Joyce Li, Shuye Pu, Nira Datta, Aaron P Tikuisis, et al. "Global landscape of protein complexes in the yeast Saccharomyces cerevisiae." In: *Nature* 440.7084 (2006), pp. 637–643.

[63] E. Katchalski-Katzir, I. Shariv, M. Eisenstein, A. A. Friesem, C. Aflalo, and I. A. Vakser. "Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques." In: *Proc. Natl. Acad. Sci. U. S. A.* 89.6 (1992), pp. 2195–2199.

[64] H. A. Gabb, R. M. Jackson, and M. J. Sternberg. "Modelling protein docking using shape complementarity, electrostatics and biochemical information." In: *J. Mol. Biol.* 272.1 (1997), pp. 106–20.

[65] D. Schneidman-Duhovny, Y. Inbar, R. Nussinov, and H. J. Wolfson. "PatchDock and SymmDock: servers for rigid and symmetric docking." In: *Nucleic Acids Res.* 33 (2005), pp. 363–7.

[66] B. Pierce, W. Tong, and Z. Weng. "M-ZDOCK: a grid-based approach for Cn symmetric multimer docking." In: *Bioinformatics* 21.8 (2005), pp. 1472–8.

[67] J. Esquivel-Rodriguez, V. Filos-Gonzalez, B. Li, and D. Kihara. "Pairwise and multimeric protein-protein docking using the LZerD program suite." In: *Methods Mol. Biol.* 1137 (2014), pp. 209–34.

[68] N. Amir, D. Cohen, and H. J. Wolfson. "DockStar: a novel ILP-based integrative method for structural modeling of multimolecular protein complexes." In: *Bioinformatics* 31.17 (2015), pp. 2801–7.

[69] D. Russel, K. Lasker, B. Webb, J. Velazquez-Muriel, E. Tjioe, D. Schneidman-Duhovny, B. Peterson, and A. Sali. "Putting the pieces together: integrative modeling platform software for structure determination of macromolecular assemblies." In: *PLoS Biol.* 10 (2012), e1001244.

[70] A. Leaver-Fay et al. "ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules." In: *Methods Enzymol.* 487 (2011), pp. 545–74.

[71] S. J. de Vries, M. van Dijk, and A. M. Bonvin. "The HADDOCK web server for data-driven biomolecular docking." In: *Nat Protoc* 5.5 (2010), pp. 883–97.

[72] E. Spiga, M. T. Degiacomi, and M. Dal Peraro. "New strategies for integrative dynamic modeling of macromolecular assembly." In: *Adv Protein Chem Struct Biol* 96 (2014), pp. 77–111.

[73] C. Chothia. "One thousand families for the molecular biologist." In: *Proteins* 357.6379 (1992), pp. 543–4.

[74] P. Aloy and R. B. Russell. "Ten thousand interactions for the molecular biologist." In: *Nat. Biotechnol.* 22.10 (2004), pp. 1317–21.

[75] M. Gao and J. Skolnick. "Structural space of protein-protein interfaces is degenerate, close to complete, and highly connected." In: *Proc. Natl. Acad. Sci. U. S. A.* 107.52 (2010), pp. 22517–22.

[76] P. J. Kundrotas, Z. Zhu, J. Janin, and I. A. Vakser. "Templates are available to model nearly all complexes of structurally characterized proteins." In: *Proc. Natl. Acad. Sci. U. S. A.* 109.24 (2012), pp. 9438–41.

[77] Q. C. Zhang, D. Petrey, R. Norel, and B. H. Honig. "Protein interface conservation across structure space." In: *Proc. Natl. Acad. Sci. U. S. A.* 107.24 (2010), pp. 10896–901.

[78] P. J. Kundrotas, Z. Zhu, and I. A. Vakser. "GWIDD: a comprehensive resource for genome-wide structural modeling of protein-protein interactions." In: *Hum. Genomics* 6 (2012), p. 7.

[79] R. Mosca, A. Ceol, and P. Aloy. "Interactome3D: adding structural details to protein networks." In: *Nat. Methods* 10.1 (2013), pp. 47–53.

[80] Q. C. Zhang, D. Petrey, J. I. Garzon, L. Deng, and B. Honig. "PrePPI: a structure-informed database of protein-protein interactions." In: *Nucleic Acids Res.* 41 (2013), pp. D828–33.

[81] M. J. Meyer, J. Das, X. Wang, and H. Yu. "INstruct: a database of high-quality 3D structurally resolved protein interactome networks." In: *Bioinformatics* 29.12 (2013), pp. 1577–9.

[82] A. Baspinar, E. Cukuroglu, R. Nussinov, O. Keskin, and A. Gursoy. "PRISM: a web server and repository for prediction of protein-protein interactions and modeling their 3D complexes." In: *Nucleic Acids Res.* 42 (2014), W285–9.

[83] J. Janin, K. Henrick, J. Moult, L. T. Eyck, M. J. Sternberg, S. Vajda, I. Vakser, and S. J. Wodak. "CAPRI: a Critical Assessment of PRedicted Interactions." In: *Proteins* 52.1 (2003), pp. 2–9. ISSN: 1097-0134 (Electronic) 0887-3585 (Linking). DOI: 10.1002/prot.10381. URL: http://www.ncbi.nlm.nih.gov/pubmed/12784359.

[84] Raúl Méndez, Raphaël Leplae, Marc F. Lensink, and Shoshana J. Wodak. "Assessment of CAPRI predictions in rounds 3–5 shows progress in docking procedures." In: *Proteins: Struct., Funct., Bioinf.* 60.2 (2005), pp. 150–169. ISSN: 1097-0134. DOI: 10.1002/prot.20551.

[85] Joël Janin. "The targets of CAPRI rounds 6–12." In: *Proteins: Struct., Funct., Bioinf.* 69.4 (2007), pp. 699–703. ISSN: 1097-0134. DOI: 10.1002/prot.21689.

[86] M. F. Lensink and S. J. Wodak. "Docking and scoring protein interactions: CAPRI 2009." In: *Proteins* 78.15 (2010), pp. 3073–84. ISSN: 1097-0134 (Electronic) 0887-3585 (Linking). DOI: 10.1002/prot.22818. URL: http://www.ncbi.nlm.nih.gov/pubmed/20806235.

[87] Joël Janin. "The targets of CAPRI rounds 20–27." In: *Proteins: Struct., Funct., Bioinf.* 81.12 (2013), pp. 2075–2081. ISSN: 1097-0134. DOI: 10.1002/prot.24375.

[88] Marc F. Lensink et al. "Prediction of homoprotein and heteroprotein complexes by protein docking and template-based modeling: A CASP-CAPRI experiment." In: *Proteins: Struct., Funct., Bioinf.* 84 (2016), pp. 323–348. ISSN: 1097-0134. DOI: 10.1002/prot.25007.

[89] Rocco Moretti, Sarel J Fleishman, Rudi Agius, Mieczyslaw Torchala, Paul A Bates, Panagiotis L Kastritis, Joao PGLM Rodrigues, Mikaël Trellet, Alexandre MJJ Bonvin, Meng Cui, et al. "Community-wide evaluation of methods for predicting the effect of mutations on protein–protein interactions." In: *Proteins: Struct., Funct., Bioinf.* 81.11 (2013), pp. 1980–1987.

[90] Marc F Lensink, Iain H Moal, Paul A Bates, Panagiotis L Kastritis, Adrien SJ Melquiond, Ezgi Karaca, Christophe Schmitz, Marc Dijk, Alexandre MJJ Bonvin, Miriam Eisenstein, et al. "Blind prediction of interfacial water positions in CAPRI." In: *Proteins: Struct., Funct., Bioinf.* 82.4 (2014), pp. 620–632.

[91] Torsten Schwede. "Protein modeling: what happened to the "protein structure gap"?" In: *Structure* 21.9 (2013), pp. 1531–1540.

[92] J. Janin. "Protein-protein docking tested in blind predictions: the CAPRI experiment." In: *Mol. Biosyst.* 6.12 (2010), pp. 2351–62. ISSN: 1742-2051 (Electronic) 1742-

2051 (Linking). DOI: 10.1039/c005060c. URL: http://www.ncbi.nlm.nih.gov/pubmed/20725658.

[93] M. Gao and J. Skolnick. "New benchmark metrics for protein-protein docking methods." In: *Proteins* 79.5 (2011), pp. 1623–34. ISSN: 1097-0134 (Electronic) 0887-3585 (Linking). DOI: 10.1002/prot.22987. URL: http://www.ncbi.nlm.nih.gov/pubmed/21365685.

[94] Joël Janin, Shoshana J. Wodak, Marc F. Lensink, and Sameer Velankar. "Assessing Structural Predictions of Protein-Protein Recognition: The CAPRI Experiment." In: *Reviews in Computational Chemistry*. Wiley-Blackwell, 2015, pp. 137–173. ISBN: 1934-5372 http://id.crossref.org/isbn/9781118889886 http://id.crossref.org/isbn/9781118407776. DOI: 10.1002/9781118889886.ch4.

[95] P. Aloy, H. Ceulemans, A. Stark, and R. B. Russell. "The relationship between sequence and interaction divergence in proteins." In: *J. Mol. Biol.* 332.5 (2003), pp. 989–998. ISSN: 0022-2836. DOI: 10.1016/j.jmb.2003.07.006. URL: <GotoISI>://WOS:000185575300002.

[96] S. Mukherjee and Y. Zhang. "MM-align: a quick algorithm for aligning multiple-chain protein complex structures using iterative dynamic programming." In: *Nucleic Acids Res.* 37.11 (2009), e83. ISSN: 1362-4962 (Electronic) 0305-1048 (Linking). DOI: 10.1093/nar/gkp318. URL: http://www.ncbi.nlm.nih.gov/pubmed/19443443.

[97] Q. Xu, A. A. Canutescu, G. Wang, M. Shapovalov, Z. Obradovic, and Jr. Dunbrack R. L. "Statistical analysis of interface similarity in crystals of homologous proteins." In: *J. Mol. Biol.* 381.2 (2008), pp. 487–507. ISSN: 1089-8638 (Electronic) 0022-2836 (Linking). DOI: 10.1016/j.jmb.2008.06.002. URL: http://www.ncbi.nlm.nih.gov/pubmed/18599072.

[98] Q. Xu and Jr. Dunbrack R. L. "The protein common interface database (ProtCID)–a comprehensive database of interactions of homologous proteins in multiple crystal forms." In: *Nucleic Acids Res.* 39.Database issue (2011), pp. D761–70. ISSN: 1362-4962 (Electronic) 0305-1048 (Linking). DOI: 10.1093/nar/gkq1059. URL: http://www.ncbi.nlm.nih.gov/pubmed/21036862.

[99] W. Li and A. Godzik. "Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences." In: *Bioinformatics* 22.13 (2006), pp. 1658–9. ISSN: 1367-4803 (Print) 1367-4803 (Linking). DOI: 10.1093/bioinformatics/btl158. URL: http://www.ncbi.nlm.nih.gov/pubmed/16731699.

[100] E. D. Levy, E. Boeri Erba, C. V. Robinson, and S. A. Teichmann. "Assembly reflects evolution of protein complexes." In: *Nature* 453.7199 (2008), pp. 1262–5. ISSN: 1476-4687 (Electronic) 0028-0836 (Linking). DOI: 10.1038/nature06942. URL: http://www.ncbi.nlm.nih.gov/pubmed/18563089.

[101] J. A. Marsh, H. Hernandez, Z. Hall, S. E. Ahnert, T. Perica, C. V. Robinson, and S. A. Teichmann. "Protein complexes are under evolutionary selection to assemble via ordered pathways." In: *Cell* 153.2 (2013), pp. 461–70. ISSN: 1097-4172 (Electronic) 0092-8674 (Linking). DOI: 10.1016/j.cell.2013.02.044. URL: http://www.ncbi.nlm.nih.gov/pubmed/23582331.

[102] T. Perica, C. Chothia, and S. A. Teichmann. "Evolution of oligomeric state through geometric coupling of protein interfaces." In: *Proc. Natl. Acad. Sci. U. S. A.* 109.21 (2012), pp. 8127–32. ISSN: 1091-6490 (Electronic) 0027-8424 (Linking). DOI: 10.1073/pnas.1120028109. URL: http://www.ncbi.nlm.nih.gov/pubmed/22566652.

[103] A. H. Elcock and J. A. McCammon. "Identification of protein oligomerization states by analysis of interface conservation." In: *Proc. Natl. Acad. Sci. U. S. A.* 98.6 (2001), pp. 2990–4. ISSN: 0027-8424 (Print) 0027-8424 (Linking). DOI: 10.1073/pnas.061411798. URL: http://www.ncbi.nlm.nih.gov/pubmed/11248019.

[104] D. R. Caffrey, S. Somaroo, J. D. Hughes, J. Mintseris, and E. S. Huang. "Are protein-protein interfaces more conserved in sequence than the rest of the protein surface?" In: *Protein Sci.* 13.1 (2004), pp. 190–202. ISSN: 0961-8368 (Print) 0961-8368 (Linking). DOI: 10.1110/ps.03323604. URL: http://www.ncbi.nlm.nih.gov/pubmed/14691234.

[105] M. Guharoy and P. Chakrabarti. "Conservation and relative importance of residues across protein-protein interfaces." In: *Proc. Natl. Acad. Sci. U. S. A.* 102.43 (2005), pp. 15447–52. ISSN: 0027-8424 (Print) 0027-8424 (Linking). DOI: 10.1073/pnas.0505425102. URL: http://www.ncbi.nlm.nih.gov/pubmed/16221766.

[106] B. Lee and F. M. Richards. "The interpretation of protein structures: estimation of static accessibility." In: *J. Mol. Biol.* 55.3 (1971), pp. 379–400. ISSN: 0022-2836 (Print) 0022-2836 (Linking). DOI: 10.1016/0022-2836(71)90324-x. URL: http://www.ncbi.nlm.nih.gov/pubmed/5551392.

[107] E. D. Levy. "A simple definition of structural regions in proteins and its use in analyzing interface evolution." In: *J. Mol. Biol.* 403.4 (2010), pp. 660–70. ISSN: 1089-8638 (Electronic) 0022-2836 (Linking). DOI: 10.1016/j.jmb.2010.09.028. URL: http://www.ncbi.nlm.nih.gov/pubmed/20868694.

[108] K. Wang and R. Samudrala. "Incorporating background frequency improves entropy-based residue conservation measures." In: *BMC Bioinformatics* 7.1 (2006), p. 385. ISSN: 1471-2105 (Electronic) 1471-2105 (Linking). DOI: 10.1186/1471-2105-7-385. URL: http://www.ncbi.nlm.nih.gov/pubmed/16916457.

[109] J. A. Capra and M. Singh. "Predicting functionally important residues from sequence conservation." In: *Bioinformatics* 23.15 (2007), pp. 1875–82. ISSN: 1367-4811 (Electronic) 1367-4803 (Linking). DOI: 10.1093/bioinformatics/btm270. URL: http://www.ncbi.nlm.nih.gov/pubmed/17519246.

[110] William S. J. Valdar and Janet M. Thornton. "Protein-protein interfaces: Analysis of amino acid conservation in homodimers." In: *Proteins: Structure, Function, and Genetics* 42.1 (2001), pp. 108–124. ISSN: 0887-3585 1097-0134. DOI: 10.1002/1097-0134(20010101)42:1<108::aid-prot110>3.0.co;2-o.

[111] V. Mainfroid, P. Terpstra, M. Beauregard, J. M. Frere, S. C. Mande, W. G. Hol, J. A. Martial, and K. Goraj. "Three hTIM mutants that provide new insights on why TIM is a dimer." In: *J. Mol. Biol.* 257.2 (1996), pp. 441–56. ISSN: 0022-2836 (Print) 0022-2836 (Linking). DOI: 10.1006/jmbi.1996.0174. URL: http://www.ncbi.nlm.nih.gov/pubmed/8609635.

[112] J. M. Duarte, A. Srebniak, M. A. Scharer, and G. Capitani. "Protein interface classification by evolutionary analysis." In: *BMC Bioinformatics* 13.1 (2012), p. 334. ISSN: 1471-2105 (Electronic) 1471-2105 (Linking). DOI: 10.1186/1471-2105-13-334. URL: http://www.ncbi.nlm.nih.gov/pubmed/23259833.

[113] A. Szilagyi and Y. Zhang. "Template-based structure modeling of protein-protein interactions." In: *Curr. Opin. Struct. Biol.* 24 (2014), pp. 10–23. ISSN: 1879-033X (Electronic) 0959-440X (Linking). DOI: 10.1016/j.sbi.2013.11.005. URL: http://www.ncbi.nlm.nih.gov/pubmed/24721449.

[114] C. Winter, A. Henschel, W. K. Kim, and M. Schroeder. "SCOPPI: a structural classification of protein-protein interfaces." In: *Nucleic Acids Res.* 34.Database issue (2006), pp. D310–4. ISSN: 1362-4962 (Electronic) 0305-1048 (Linking). DOI: 10.1093/nar/gkj099. URL: http://www.ncbi.nlm.nih.gov/pubmed/16381874.

[115] R. Mosca, A. Ceol, A. Stein, R. Olivella, and P. Aloy. "3did: a catalog of domain-based interactions of known three-dimensional structure." In: *Nucleic Acids Res.* 42.Database issue (2014), pp. D374–9. ISSN: 1362-4962 (Electronic) 0305-1048 (Linking). DOI: 10.1093/nar/gkt887. URL: http://www.ncbi.nlm.nih.gov/pubmed/24081580.

[116] X. Kuang, A. Dhroso, J. G. Han, C. R. Shyu, and D. Korkin. "DOMMINO 2.0: integrating structurally resolved protein-, RNA-, and DNA-mediated macromolecular interactions." In: *Database (Oxford)* 2016 (2016). ISSN: 1758-0463 (Electronic) 1758-0463 (Linking). DOI: 10.1093/database/bav114. URL: http://www.ncbi.nlm.nih.gov/pubmed/26827237.

[117] D. Douguet, H. C. Chen, A. Tovchigrechko, and I. A. Vakser. "DOCKGROUND resource for studying protein-protein interfaces." In: *Bioinformatics* 22.21 (2006), pp. 2612–8. ISSN: 1367-4811 (Electronic) 1367-4803 (Linking). DOI: 10.1093/bioinformatics/btl447. URL: http://www.ncbi.nlm.nih.gov/pubmed/16928732.

[118] E. D. Levy, J. B. Pereira-Leal, C. Chothia, and S. A. Teichmann. "3D complex: a structural classification of protein complexes." In: *PLoS Comput. Biol.* 2.11 (2006), e155. ISSN: 1553-7358 (Electronic) 1553-734X (Linking). DOI: 10.1371/journal.pcbi.0020155. URL: http://www.ncbi.nlm.nih.gov/pubmed/17112313.

[119] E. D. Levy. "PiQSi: protein quaternary structure investigation." In: *Structure* 15.11 (2007), pp. 1364–7. ISSN: 0969-2126 (Print) 0969-2126 (Linking). DOI: 10.1016/j.str.2007.09.019. URL: http://www.ncbi.nlm.nih.gov/pubmed/17997962.

[120] G. Wang and Jr. Dunbrack R. L. "PISCES: a protein sequence culling server." In: *Bioinformatics* 19.12 (2003), pp. 1589–91. ISSN: 1367-4803 (Print) 1367-4803 (Linking). DOI: 10.1093/bioinformatics/btg224. URL: http://www.ncbi.nlm.nih.gov/pubmed/12912846.

[121] M. Biasini et al. "SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information." In: *Nucleic Acids Res.* 42.Web Server issue (2014), W252–8. ISSN: 1362-4962 (Electronic) 0305-1048 (Linking). DOI: 10.1093/nar/gku340. URL: http://www.ncbi.nlm.nih.gov/pubmed/24782522.

[122] Y. Ofran and B. Rost. "ISIS: interaction sites identified from sequence." In: *Bioinformatics* 23.2 (2007), e13–6. ISSN: 1367-4811 (Electronic) 1367-4803 (Linking). DOI: 10.1093/bioinformatics/btl303. URL: http://www.ncbi.nlm.nih.gov/pubmed/17237081.

[123] J. Bernauer, R. P. Bahadur, F. Rodier, J. Janin, and A. Poupon. "DiMoVo: a Voronoi tessellation-based method for discriminating crystallographic and biological protein-protein interactions." In: *Bioinformatics* 24.5 (2008), pp. 652–8. ISSN: 1367-4811 (Electronic) 1367-4803 (Linking). DOI: 10.1093/bioinformatics/btn022. URL: http://www.ncbi.nlm.nih.gov/pubmed/18204058.

[124] P. Block, J. Paern, E. Hullermeier, P. Sanschagrin, C. A. Sotriffer, and G. Klebe. "Physicochemical descriptors to discriminate protein-protein interactions in permanent and transient complexes selected by means of machine learning algorithms." In: *Proteins* 65.3 (2006), pp. 607–22. ISSN: 1097-0134 (Electronic) 0887-3585 (Linking). DOI: 10.1002/prot.21104. URL: http://www.ncbi.nlm.nih.gov/pubmed/16955490.

[125] T. Hamp and B. Rost. "Evolutionary profiles improve protein-protein interaction prediction from sequence." In: *Bioinformatics* 31.12 (2015), pp. 1945–50. ISSN: 1367-4811 (Electronic) 1367-4803 (Linking). DOI: 10.1093/bioinformatics/btv077. URL: http://www.ncbi.nlm.nih.gov/pubmed/25657331.

[126] Q. Dong, X. Wang, L. Lin, and Y. Guan. "Exploiting residue-level and profile-level interface propensities for usage in binding sites prediction of proteins." In: *BMC Bioinformatics* 8.1 (2007), p. 147. ISSN: 1471-2105 (Electronic) 1471-2105 (Linking). DOI: 10.1186/1471-2105-

8-147. URL: http://www.ncbi.nlm.nih.gov/pubmed/17480235.

[127] Q. Liu, Z. Li, and J. Li. "Use B-factor related features for accurate classification between protein binding interfaces and crystal packing contacts." In: *BMC Bioinformatics* 15 Suppl 16.Suppl 16 (2014), S3. ISSN: 1471-2105 (Electronic) 1471-2105 (Linking). DOI: 10.1186/1471-2105-15-S16-S3. URL: http://www.ncbi.nlm.nih.gov/pubmed/25522196.

[128] J. J. Marsh and H. G. Lebherz. "Fructose-bisphosphate aldolases: an evolutionary history." In: *Trends Biochem. Sci.* 17.3 (Mar. 1992), pp. 110–113.

[129] K. Nakahara, H. Yamamoto, C. Miyake, and A. Yokota. "Purification and characterization of class-I and class-II fructose-1,6-bisphosphate aldolases from the cyanobacterium Synechocystis sp. PCC 6803." In: *Plant Cell Physiol.* 44.3 (Mar. 2003). [PubMed:12668779], pp. 326–333.

[130] T. Izard and J. Sygusch. "Induced fit movements and metal cofactor selectivity of class II aldolases: structure of Thermus aquaticus fructose-1,6-bisphosphate aldolase." In: *J. Biol. Chem.* 279.12 (Mar. 2004). [DOI:] [PubMed:14699122], pp. 11825–11833. DOI: 10.1074/jbc.M311375200.

[131] A. Galkin, L. Kulakova, E. Melamud, L. Li, C. Wu, P. Mariano, D. Dunaway-Mariano, T. E. Nash, and O. Herzberg. "Characterization, kinetics, and crystal structures of fructose-1,6-bisphosphate aldolase from the human parasite, Giardia lamblia." In: *J. Biol. Chem.* 282.7 (Feb. 2007). [DOI:] [PubMed:17166851], pp. 4859–4867. DOI: 10.1074/jbc.M609534200.

[132] A. Pickl, U. Johnsen, and P. Schonheit. "Fructose degradation in the haloarchaeon Haloferax volcanii involves a bacterial type phosphoenolpyruvate-dependent phosphotransferase system, fructose-1-phosphate kinase, and class II fructose-1,6-bisphosphate aldolase." In: *J. Bacteriol.* 194.12 (June 2012). [PubMed Central:PMC3370872] [DOI:] [PubMed:22493022], pp. 3088–3097. DOI: 10.1128/JB.00200-12.

[133] T. F. De Koning-Ward and R. M. Robins-Browne. "Contribution of urease to acid tolerance in Yersinia enterocolitica." In: *Infect. Immun.* 63.10 (1995), pp. 3790–5. ISSN:

0019-9567 (Print) 0019-9567 (Linking). URL: http://www.ncbi.nlm.nih.gov/pubmed/7558281.

[134] N. Bhagat and J. S. Virdi. "Molecular and biochemical characterization of urease and survival of Yersinia enterocolitica biovar 1A in acidic pH in vitro." In: *BMC Microbiol.* 9 (2009), p. 262. ISSN: 1471-2180 (Electronic) 1471-2180 (Linking). DOI: 10.1186/1471-2180-9-262. URL: http://www.ncbi.nlm.nih.gov/pubmed/20017936.

[135] B. D. Jones and H. L. Mobley. "Proteus mirabilis urease: nucleotide sequence determination and comparison with jack bean urease." In: *J. Bacteriol.* 171.12 (1989), pp. 6414–22. ISSN: 0021-9193 (Print) 0021-9193 (Linking). URL: http://www.ncbi.nlm.nih.gov/pubmed/2687233.

[136] H. L. Mobley, M. D. Island, and R. P. Hausinger. "Molecular biology of microbial ureases." In: *Microbiol. Rev.* 59.3 (1995), pp. 451–80. ISSN: 0146-0749 (Print) 0146-0749 (Linking). URL: http://www.ncbi.nlm.nih.gov/pubmed/7565414.

[137] R. Ligabue-Braun, F. C. Andreis, H. Verli, and C. R. Carlini. "3-to-1: unraveling structural transitions in ureases." In: *Naturwissenschaften* 100.5 (May 2013). [DOI:] [PubMed:23619940], pp. 459–467. DOI: 10.1007/s00114-013-1045-22.

[138] P. A. Edwards. "Fusion genes and chromosome translocations in the common epithelial cancers." In: *J. Pathol.* 220.2 (2010), pp. 244–54. ISSN: 1096-9896 (Electronic) 0022-3417 (Linking). DOI: 10.1002/path.2632. URL: http://www.ncbi.nlm.nih.gov/pubmed/19921709.

[139] A. J. Enright, I. Iliopoulos, N. C. Kyrpides, and C. A. Ouzounis. "Protein interaction maps for complete genomes based on gene fusion events." In: *Nature* 402.6757 (1999), pp. 86–90. ISSN: 0028-0836 (Print) 0028-0836 (Linking). DOI: 10.1038/47056. URL: http://www.ncbi.nlm.nih.gov/pubmed/10573422.

[140] E. M. Marcotte. "Detecting Protein Function and Protein-Protein Interactions from Genome Sequences." In: *Science* 285.5428 (1999), pp. 751–753. ISSN: 00368075 10959203. DOI: 10.1126/science.285.5428.751.

[141] A. Balasubramanian and K. Ponnuraj. "Crystal structure of the first plant urease from jack bean: 83 years of journey from its first crystal to molecular structure." In: *J.*

Mol. Biol. 400.3 (2010), pp. 274–83. ISSN: 1089-8638 (Electronic) 0022-2836 (Linking). DOI: 10.1016/j.jmb.2010.05.009. URL: http://www.ncbi.nlm.nih.gov/pubmed/20471401.

[142] A. Balasubramanian, V. Durairajpandian, S. Elumalai, N. Mathivanan, A. K. Munirajan, and K. Ponnuraj. "Structural and functional studies on urease from pigeon pea (Cajanus cajan)." In: *Int. J. Biol. Macromol.* 58 (2013), pp. 301–9. ISSN: 1879-0003 (Electronic) 0141-8130 (Linking). DOI: 10.1016/j.ijbiomac.2013.04.055. URL: http://www.ncbi.nlm.nih.gov/pubmed/23624166.

[143] E. Jabri, M. B. Carr, R. P. Hausinger, and P. A. Karplus. "The crystal structure of urease from Klebsiella aerogenes." In: *Science* 268.5213 (1995), p. 998. URL: http://science.sciencemag.org/content/268/5213/998.abstract.

[144] Stefano Benini, Wojciech R. Rypniewski, Keith S. Wilson, Silvia Miletti, Stefano Ciurli, and Stefano Mangani. "A new proposal for urease mechanism based on the crystal structures of the native and inhibited enzyme from Bacillus pasteurii: why urea hydrolysis costs two nickels." In: *Structure* 7.2 (1999), pp. 205–216. ISSN: 09692126. DOI: 10.1016/s0969-2126(99)80026-4.

[145] N. C. Ha, S. T. Oh, J. Y. Sung, K. A. Cha, M. H. Lee, and B. H. Oh. "Supramolecular assembly and acid resistance of Helicobacter pylori urease." In: *Nat. Struct. Biol.* 8.6 (2001), pp. 505–9. ISSN: 1072-8368 (Print) 1072-8368 (Linking). DOI: 10.1038/88563. URL: http://www.ncbi.nlm.nih.gov/pubmed/11373617.

[146] E. L. Carter, D. E. Tronrud, S. R. Taber, P. A. Karplus, and R. P. Hausinger. "Iron-containing urease in a pathogenic bacterium." In: *Proc. Natl. Acad. Sci. U. S. A.* 108.32 (2011), pp. 13095–9. ISSN: 1091-6490 (Electronic) 0027-8424 (Linking). DOI: 10.1073/pnas.1106915108. URL: http://www.ncbi.nlm.nih.gov/pubmed/21788478.

[147] V. Mariani, M. Biasini, A. Barbato, and T. Schwede. "lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests." In: *Bioinformatics* 29.21 (2013), pp. 2722–8. ISSN: 1367-4811 (Electronic) 1367-4803 (Linking). DOI: 10.1093/bioinformatics/btt473. URL: http://www.ncbi.nlm.nih.gov/pubmed/23986568.

[148] Konstantin Arnold, Lorenza Bordoli, Jürgen Kopp, and Torsten Schwede. "The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling." In: *Bioinformatics* 22.2 (2006), pp. 195–201.

[149] Florian Kiefer, Konstantin Arnold, Michael Künzli, Lorenza Bordoli, and Torsten Schwede. "The SWISS-MODEL Repository and associated resources." In: *Nucleic acids research* 37.suppl 1 (2009), pp. D387–D392.

[150] Nicolas Guex, Manuel C Peitsch, and Torsten Schwede. "Automated comparative protein structure modeling with SWISS-MODEL and Swiss-PdbViewer: A historical perspective." In: *Electrophoresis* 30.S1 (2009), S162–S173.

[151] J. Haas, S. Roth, K. Arnold, F. Kiefer, T. Schmidt, L. Bordoli, and T. Schwede. "The Protein Model Portal–a comprehensive resource for protein structure and model information." In: *Database (Oxford)* 2013.0 (2013), bat031. ISSN: 1758-0463 (Electronic) 1758-0463 (Linking). DOI: 10.1093/database/bat031. URL: http://www.ncbi.nlm.nih.gov/pubmed/23624946.

[152] D. E. Kim, D. Chivian, and D. Baker. "Protein structure prediction and analysis using the Robetta server." In: *Nucleic Acids Res.* 32.Web Server issue (2004), W526–31. ISSN: 1362-4962 (Electronic) 0305-1048 (Linking). DOI: 10.1093/nar/gkh468. URL: http://www.ncbi.nlm.nih.gov/pubmed/15215442.

ACKNOWLEDGMENTS

This thesis would not have been possible without the supervision of Prof. Torsten Schwede, who expertly navigated me throughout these four years of research and discovery. His clear vision guided me past obstacles, from small but subtle details to the pursuit of a "bigger picture", without dodging personal and life-impacting tips.

My deepest thanks also go to Prof. Christian von Mering and Dr. Guido Capitani for being part of my committee and helping me with priceless advises, recommendations, cues, and hints that steered me towards relevant questions. To Guido, all my wishes for a prompt recovery.

In addition, I would like to thank the Werner Siemens Foundation (WSF) for the fundings and especially Prof. Joachim Seelig, Prof. Christoph Handschin, and Angie for building an international environment of PhD students and organizing many interesting study-trips.

I also want to express my gratitude to all people in the group: Lorenza for the supervision, pleasant chat, and encouragement, Tjaart for the countless corrections to my writings, and both Jürgen and Mark for the good laughs, the scientifically pregnant discussions, and all the personal help they provided. A special recognition to my office-mate Bienchen who prevented me from messing up the git repository and doing many other explosive mistakes. Thanks to Andrew for supervision on all that web/django/js related stuff. Many thanks to Gabriel (as well as Valentina, Marco, Tiziano, Tobi, and Thomas) with whom I shared the pain and pleasure of doing a PhD. It would have been much more difficult without the friendship of Dario, beer drinking sessions and concerts with him and Stefan helped a lot. Thanks to Alessandro for the fun and microwave operations. Also, thanks to Valerio for mentoring me at the very beginning. A global "thank you" to the Berneche group, Simon, Florian, Niklaus, and Olivier for providing an alternative and interesting point of view. I am indebted to the whole sciCORE team, Pablo, Thierry, Geoffrey, Martin, Konsti, Jani, Ruben and Eva not only for the technical support but also for the funny evenings at the Cargo bar. My recognition also goes to Rita,

Sarah, and Yvonne for their support in organizing and for all that coffee.

I need to mention also the contribution coming from exchange, discussions, and retreats with many PhD and master buddies: Dominik, Fabian, Luca, Ricardo, Arantxa, Keith, Kate, Joka, Kathrin, Johanna, Dominik, Cedric, Vahap, Mario, Minkyoung, Max, Peter, Simon, Anne, Christine, Julian, Andreas, Said, Ralf, Foivos and Maria. Moreover, many longtime friends helped with emotional/musical/psychological/illogical support: Mauro, Cecco, Vecca, Löch, Tommy, Prando, Meme, and many more.

My profound gratitude to my family for their unflagging love and support. Thanks to Pina for being a sweet mother-in-law and to Antonio for all the mushrooms and lessons on how to properly salt water. A special mention to my brother Filippo who always found time to correct my drafts, share interests, curiosities, and exposing oddities and idiosyncrasies of our epistemological systems. With him, the crazy Justine, and many friends we spent memorable times in Amsterdam.

None of this would have happened without my parents, Roberto and Cristiana, being at my side. They instilled in me the passion for science, the thirst for knowledge, and the fascination for Nature.

Finally, I deeply thank my beloved one, Francesca. She risked everything following me to Basel, in a purely selfless act of Love. Whatever I will do, wherever I will go, I will always feel at home with you by my side.