Assessment of protein assembly prediction in CASP12 & Conformational dynamics of integrin alpha-I domains

Master Thesis

Author(s):
Lafita, Aleix

Publication date:
2017

Permanent link:
https://doi.org/10.3929/ethz-a-010863273

Universität Zürich[UZH]

ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Assessment of protein assembly prediction in CASP12 & Conformational dynamics of integrin α-I domains

Master Thesis

Aleix Lafita

March 2017

Prof. Dr. Amedeo Caflisch
Dr. Guido Capitani
Dr. Spencer Bliven

Laboratory of Biomolecular Research , Paul Scherrer Institute
Department of Biochemistry , University of Zürich
Department of Biosystems Science and Engineering , ETH Zürich

Abstract

Assessment of protein assembly prediction in CASP12 & Conformational dynamics of integrin α-I domains

Aleix Lafita

Protein structures are key to understand the details of biological processes and disease mechanisms. Computational modeling of protein structures is present in different areas of biological studies: structure determination from experimental data, structure prediction from the amino acid sequence, and the study of protein dynamics. This thesis covers two different applications of protein structure modeling: the prediction of protein assemblies and the study of the conformational dynamics of a globular domain.

The first part of this thesis covers the assessment of protein assembly predictions in the CASP12 experiment. Protein assemblies have been included for the first time as a prediction category in the 12th edition of the CASP experiment. As part of the assessment, quaternary structure models have to be scored based on their similarity to their reference structures (targets) and the predictor groups have to be ranked based on their performance. First, we describe the procedure for assigning the biological assembly of target experimental structures, including the contributions to the EPPIC software. Next, a novel scalable quaternary structure alignment algorithm is developed and used to compare prediction models with their target assemblies. Finally, similarity scores are defined to quantify the quality of the assembly models and gain insights into the prediction methods. The algorithms and scores have been integrated into the CASP evaluation pipeline and all the results and rankings, presented and discussed at the final CASP meeting, are available through the *Prediction Center* web site.

The second part of this thesis is about the modulation of integrin signaling though allosteric inhibition. Integrins are heterodimeric transmembrane protein receptors involved in bidirectional signaling across the cell membrane. The C-terminal helix of the I-domain of the α subunit is involved in the integrin signaling mechanism. An allosteric inhibition pocket of the I-domain has been discovered for the subunit type α_L, which involves a conformational change in the C-terminal helix. The pocket has not yet been described for any other type of integrin, although it would also be of pharmacological interest. To investigate the universality of the allosteric inhibition mechanism among the I-domains of integrin α subunits, molecular dynamics simulations are used to study their metastable states. Although the results are inconclusive about whether the mechanism of allosteric inhibition is present in other types of α-I domains, we propose further research lines to continue the investigation.

per l'Annabel, gràcies per fer-me costat durant tot aquest temps

Contents

Abbreviations

CAPRI Critical Assessment of PRediction of Interactions

CASP Critical Assessment of Techniques for Protein Structure Prediction

DNA DeoxyriboNucleic acid

EPPIC Evolutionary Protein Protein Interface Classifier

ETH Eidgenössische Technische Hochschule

GROMACS GROningen MAchine for Chemical Simulations

LFA-1 Lymphocyte Function-associated Antigen 1

MCC Mathews Correlation Coefficient

MFPT Mean Force Passage Time

MIDAS Metal Ion-Dependent Adhesion Site

NMR Nuclear Magnetic Resonance

PCNA Proliferating Cell Nuclear Antigen

PDB Protein Data Bank

PIGS Progress Index-Guided Sampling

PISA Proteins, Interfaces, Structures and Assemblies

PRC Photosynthetic Reaction Center

PSI Paul Scherrer Institute

RAM Replica-Averaged Metadynamics

RMSD Root Mean Squared Deviation

SAPPHIRE States And Pathways Projected with HIgh REsolution

SAS Solvent Accessible Surface

UZH Universität ZüricH

VLA-1 Very Late Antigen 1

Chapter 1

Protein structure

Proteins are composed by a linear sequence of amino acids linked together with peptide bonds. Therefore, proteins are a specific type of polymers, called polypeptides, where the residues are amino acids.

In a process known as protein folding, the linear sequence of amino acids adopts a three-dimensional shape at physiological conditions via residue-residue interactions, the protein structure. In their folded state, proteins participate in biological processes. Thus, knowledge about their structure is often necessary to understand their function at the molecular level.

This chapter is a brief introduction into the most important concepts of protein structures: the levels of structural organization in proteins; the evolution, classification and comparison of protein structures; and experimental structure determination techniques.

1.1 Structural levels

Primary structure

The primary structure of a protein is a one-dimensional feature, defined by the linear sequence of amino acids in the polypeptide chain, from N- to C-terminal. It represents the composition of the protein and the order in which the amino acids are bonded together.

Secondary structure

The secondary structure of proteins is also a one-dimensional feature, defined by the sequence of recurrent sub-structures in the polypeptide chain. Each amino acid in the protein sequence can be in one state from a discrete set of possible secondary structure states. The two main types of secondary

structure states, α-helix and β-sheet, were suggested years before any protein structure could be experimentally determined (Pauling and Corey 1951). Shortly after the publication of the first protein structures, methods to assign secondary structure were developed. One of the most popular methods is DSSP (Kabsch and Sander 1983). The method uses the patterns of hydrogen bonding between backbone atoms to determine the secondary structure elements of a protein.

Tertiary structure

The tertiary structure is the three-dimensional coordinates of all the atoms in a protein chain. The tertiary structure can also be represented by the relative position and orientation of the secondary structure elements of a protein chain. Local segments of tertiary structure, known as structural motifs, are also recurrent in a large number of protein structures. Some structural motifs are the greek key, the helix-turn-helix or the α-β-α.

In addition, parts of protein chains, known as structural domains, are believed to be able to fold and function independently from the rest of the protein chain. Integrins, proteins with a complex domain architecture (domain composition and interactions) key in their biological function, will be described in section 6.1.

Quaternary structure

The quaternary structure of a protein, also called biological unit or biological assembly, is the complete functional structural unit in the cell. It is defined by the number, type and arrangement of protein subunits (polypeptide chains) that participate together in a biological function.

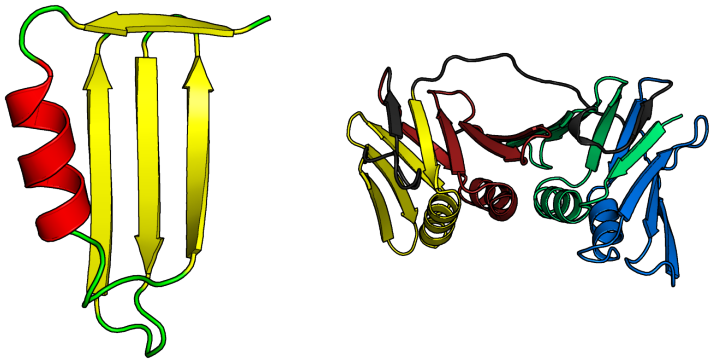
Three identical DNA clamp chains bind together to form a ring around the double stranded DNA helix. It is, therefore, required for DNA clamps to be oligomers to carry out their biological function, since it would be mechanically much more challenging for a single-chain ring to circle the DNA helix.

Since it is one of the main topics of interest of this thesis, quaternary structure will be described in further detail in section 2.1.

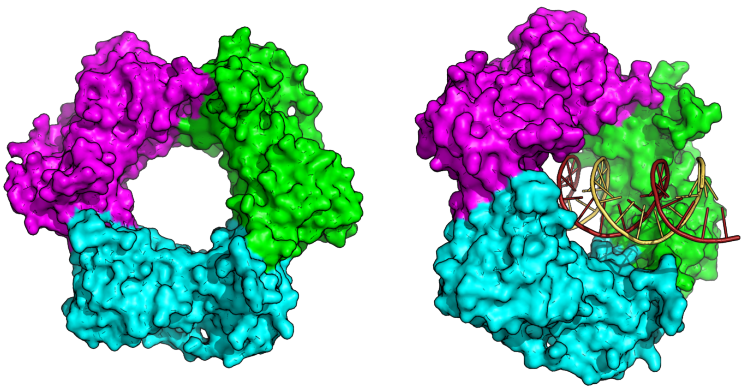
1.2 Structure evolution

Proteins can evolve through the accumulation of single amino acid substitutions, also known as point mutations, caused by underlying nucleotide substitutions in the DNA. Proteins can also evolve by insertions or deletions in their amino acid sequence, which can range from single amino acid residues to full protein domains. One example of is the insertion of the I-domain in some types of integrin α subunits, described in section 6.1.

(a) Primary (bottom) and secondary (top) structure are one dimensional features of proteins.



(b) Tertiary structure of a single domain (left) and that of multiple repeats of the same domain in a protein chain (right).



(c) Quaternary structure formed by the interaction of three protein chains (left) and its binding to the DNA double helix (right).

Figure 1.1: Levels of protein structure of native human PCNA (1VYM), a DNA clamp.

From the point of view of protein structure, point mutations can either be neutral, if they do not alter the structure of the protein, or disruptive, if they change any of the structural properties of the protein. The structure of proteins is more conserved than their amino acid sequence. Therefore, the structural similarity of two proteins can be inferred by their sequence similarity. Some bioinformatics resources, like *Pfam* (Finn et al. 2016b) or *Interpro* (Finn et al. 2016a), allow the prediction and classification of protein structures using their amino acid sequences.

Relevant for the first part of the thesis is that quaternary structure is less conserved than secondary or tertiary structure, with known examples of divergence at high sequence similarities. The explanation is the fact that the free energy of protein binding is not uniform across the surface, but distributed among a few residue hotspots (Clackson and Wells 1995).

1.3 Structure classification

The most popular classification of protein structures is done at the domain level. Therefore, the first step to classify a protein chain is to split it into its component domains. These are then classified in different levels of similarity, following a classification tree. As an example, the coarsest level of the tree could be the types of secondary structure content in the domain, subsequent levels could reflect the content of some small structural motifs, and the finest level of the tree could use the evolutionary relations between the domains. Popular protein domain classification resources are SCOP (Murzin et al. 1995), CATH (Sillitoe et al. 2015) and ECOD (Cheng et al. 2014).

1.4 Structure comparison

The superposition of two structures is their transformation in three-dimensions such that an objective function of the distance between their equivalent positions, e.g. the RMSD, is minimized. We can distinguish two types of protein structure comparison metrics: superposition dependent and superposition independent.

Superposition independent metrics use features extracted independently from each of the structures, like their inter-residue contact map. Some metrics of this type will be introduced in section 5.2 in the context of the CASP12 assessment.

Superposition dependent methods require first, as the name implies, a superposition of the structures prior to measuring their similarity by, for example, the distance between the corresponding residues. The most popular examples types of metrics are the RMSD, described in section 6.2, or the TM-score (Zhang and Skolnick 2004).

Until now we have assumed the knowledge of the corresponding residues of the structures being compared. However, when comparing two protein structures, the corresponding residue positions are often not known. Structural alignment methods compute the equivalent positions of two proteins using their three-dimensional shape and conformation (Bourne and Shindyalov 2003). In chapter 4, a novel structural alignment algorithm at the quaternary structure level will be introduced to compute efficiently the corresponding chain and residue positions of two protein assemblies.

1.5 Structure determination

Structure determination aims at obtaining an accurate high resolution model of a protein structure using experimental data. The three main techniques used are: X-ray crystallography, nuclear magnetic resonance (NMR) and cryo-electron microscopy (Cryo-EM).

X-Ray crystallography

X-ray crystallography requires the formation of protein crystals, the most challenging step. Atoms in the crystals diffract a beam of incident X-rays into many specific directions, known as the diffraction pattern, which allows the reconstruction of their three-dimensional electron density. The positions of the protein atoms can be then fitted to the electron density map to produce an atomic model of the protein structure (Smyth and Martin 2000).

NMR

NMR is a spectroscopic technique to obtain information about the structure and dynamics of proteins in solution. The result of the experiment is a set of NOE measurements, which can be converted into intermolecular distances of protein residues (Wüthrich 2003). This information reduces significantly the number of possible protein conformations, which are explored using sampling techniques, like molecular dynamics simulations, to produce a set of plausible models fulfilling the experimental spatial constraints.

Cryo-EM

Single particle cryo-EM produces a three-dimensional reconstruction of the protein by combining several transmission electron microscopy (TEM) images of similar particles at cryogenic conditions. If the reconstruction reaches atomic resolutions, the positions of the protein atoms can be fitted in a similar manner to X-ray crystallography to generate an atomic model of the protein structure (Jonic and Vénien-Bryan 2009).

Structure database

The Protein Data Bank (PDB) is a database for the storage and annotation of experimentally determined protein structures (Berman et al. 2000). At the time of this thesis, the percentage of structures in the PDB solved by X-ray crystallography was 89%, by NMR 9% and by cryo-EM 1%.

Structure Prediction

The number of protein structures that can be determined experimentally is limited, due to the cost and time scale of the techniques. Structure prediction methods aim at broadening the protein structure knowledge using principles extracted from the already determined protein structures. An overview of the field and some of these methods will be described in section 2.2, as an introduction to the first part of this thesis.

Protein dynamics

Proteins are not static rigid molecules. Oftentimes, the three-dimensional structure alone is not sufficient to understand their mechanism of action. An example is the signaling mechanism of integrins described in section 6.1, the subject of study of the second part of this thesis. Some experimental techniques, like NMR, can characterize the dynamics of protein structures. Another approach is to study protein motions *in silico* using molecular dynamics simulations, a topic that will be covered in section 6.2.

Part I

Assessment of protein assembly prediction in CASP12

Chapter 2

Introduction

This chapter is an introduction into protein quaternary structure and protein structure prediction, with emphasis on the CASP experiment

2.1 Protein quaternary structure

Protein chains associate in macromolecular complexes to carry out their functions in the cell. Hence, a detailed knowledge of the molecular partners and mode of association is key to understand the mechanism of activity of a protein.

The stoichiometry is a representation of the composition of a protein assembly, specifying the type and number of subunits participating in the complex. The symmetry of a protein assembly describes the arrangement of the subunits in the three dimensional space (Levy and Teichmann 2013).

2.1.1 Classification

Quaternary structures can be classified based on different properties: stability, affinity and composition (Keskin, Tuncbag, and Gursoy 2016). Stability wise, protein assemblies can be obligate, if the individual subunits are unstable outside the assembly, or non-obligate, if they can exist independently. Regarding affinity, quaternary structures can be permanent, if they never dissociate in vivo conditions, or transient, if they associate and dissociate temporarily in the cell. Finally, from the composition point of view, assemblies can be homomeric, if they are formed by a single type of subunit (entity), or heteromeric, if they are composed of multiple entities.

2.1.2 Biological relevance

Quaternary structure is relevant in a wide variety of protein functions. We have seen an example in section 1.1 with the importance of the DNA clamp assembly in binding to the DNA double helix. Other examples of the quaternary structure relevance in protein function are cooperativity (Monod, Wyman, and Changeux 1965), like in the oxygen binding of hemoglobin, localization of multiple functions in the cell, like in the Acetyl-CoA carboxylase, or the encapsulation of material, like in viral capsids. In section 5.4, two CASP targets with a quaternary structure relevant for their function will be presented, together with an assessment of the biological relevance of their predicted models.

2.2 Protein structure prediction

2.2.1 Structure prediction problem

Protein structure prediction aims at inferring the three-dimensional structure of a protein, primarily from its amino acid sequence, but also using other information and resources.

Anfinsen's dogma

The Anfinsen's dogma, also known as the thermodynamic hypothesis of protein folding, postulates that a protein native structure is solely determined by its amino acid sequence (Anfinsen 1972). The implication is that, at the physiological conditions of the cell, the native protein structure must be a unique, stable and kinetically accessible minimum of the free energy landscape. Evolution might play a role in selecting for proteins whose structures obey this rule, the major exception being intrinsically disordered proteins.

Levinthal's paradox

If the thermodynamic hypothesis is valid, no other information apart from the amino acid sequence is needed to predict the structure of a protein. However, there is still a challenge to be addressed, known as Levinthal's paradox.

In a thought experiment, Cyrus Levinthal noted that, due to the very large number of degrees of freedom in a polypeptide chain, proteins have an astronomical number of possible conformations (Levinthal 1969). However, proteins manage to fold in the millisecond or microsecond time scale, so they do not have time to exhaustively explore the entirety of their conformational space, hence the paradox. The conclusion is that proteins must follow a folding pathway, that is, sample only a small fraction of all the possible conformations in a series of steps from the unfolded to the folded state.

The implication for protein structure prediction is that, even if the energy for any particular protein conformation can be accurately estimated, it is impossible to sample all possible protein conformations. Efficient conformational sampling methods or simulations of the folding pathway have to be used to circumvent this problem.

2.2.2 Structure prediction approaches

Based on the available information for a particular protein, the prediction of its structure can be divided into two categories: homology modeling or *de novo* modeling.

Homology modeling, also called comparative modeling or template-based modeling, is a technique to extract structural information from homologous structures (templates) in order to predict the three-dimensional structure of a protein (target) from its amino acid sequence (Marti-Renom et al. 2000). Key steps are the identification of remote templates through sophisticated sequence alignment methods, the accurate alignment of the target to the template and the modeling of mismatches and missing segments (Biasini et al. 2014).

For about two thirds of protein sequences, homology modeling is not possible. The prediction of the three-dimensional structure for these proteins has to be done without structural information. There are three main approaches to *de novo* protein structure prediction: *ab initio* folding, fragment based methods and co-evolution methods. *Ab initio* protein folding uses thermodynamic principles (force fields) to score and sample protein conformations. Fragment based methods use statistical potentials derived from the frequencies of protein structural features in the databases (Simons et al. 1997). Co-evolution methods use multiple sequence alignments of the target and its sequence homologs to infer the spatial proximity of residues in the three-dimensional structure (Marks et al. 2011).

2.2.3 Protein assembly prediction

The methods mentioned in the previous section are focused in the prediction of protein tertiary structure, particularly small protein domains. However, as it was described in section 1.1, the functional unit of proteins involves multiple protein domains and chains in close interaction. Methods that predict protein assemblies bring together in the three-dimensional space the structures of component domains and chains of a protein in order to recreate the full functional unit.

The protein assembly prediction problem is related to the protein-protein docking problem. In protein-protein docking, two chains, called ligand and

receptor, are known to interact with an unknown interface. The main difference is that docking is a pairwise problem, where a single interface between two chains has to be predicted, while protein assembly prediction involves an arbitrary number of protein chains with an unknown number of interfaces between them.

2.3 The CASP experiment

2.3.1 History

CASP is a world-wide community-based experiment to independently assess the state of the art of protein structure prediction techniques and drive their further development. The experiment is held every two years since the first edition in 1994 (Moult et al. 1995).

CASP has historically focused on the prediction of tertiary structure. In recent years, however, the awareness in the community of the importance of quaternary structure in the interpretation of the models has grown significantly. The effectiveness of predicting protein assemblies from the amino acid sequence was first explored in CASP9 (Mariani et al. 2011), as part of the template based assessment category. After a second attempt in CASP11 in collaboration with CAPRI (Lensink et al. 2016), the CASP organization decided to include protein quaternary structure prediction as a new assessment category for CASP12, with Dr. Guido Capitani as the lead assessor.

2.3.2 Organization

The CASP experiment is organized in two phases: prediction and assessment (figure 2.1). The prediction phase of the experiment runs from beginning of May until the end of August. Target protein sequences are released through the *Prediction Center* web site every week, around a total of a 100 for edition. Server groups have a short time to submit the models, around 72 hours per target, while human groups have a longer time period to submit their predictions, around 3 weeks. The assessment phase of the experiment runs from the end of the prediction phase until the final meeting of the experiment in December. During that time, assessor teams come up with metrics to evaluate the quality of the models and rank the groups by their overall performance. Every edition of the experiment concludes with a final meeting, where predictor groups and assessor teams present and discuss the results together.

There are two types of predictor groups participating in CASP: server and human. Server groups are automatic prediction pipelines, without human intervention. Human groups are provided the server results, and may use human experience and external resources to guide the modeling process.

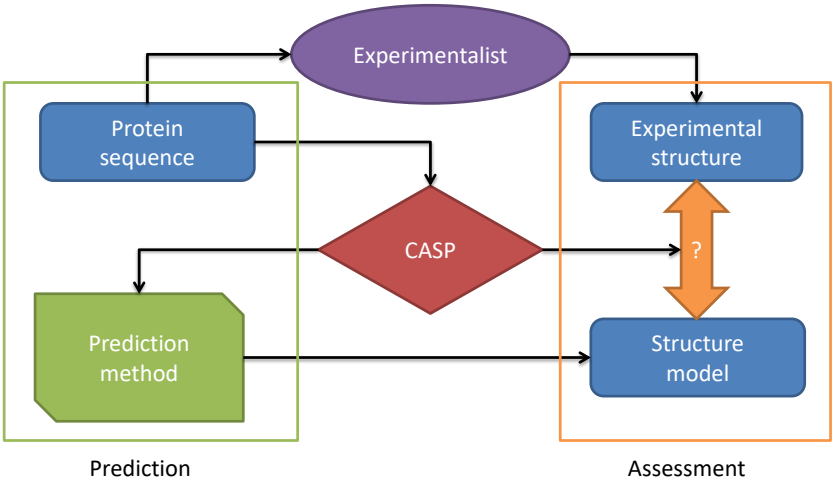


Figure 2.1: Organization of the CASP experiment. The sequences of target proteins are released to predictor groups previous to the release of their experimental structures in the prediction phase. The prediction models are then compared to the experimental structures in each of the categories of the assessment phase.

Usually, each group participating in CASP uses a single prediction method or pipeline, although there are groups using a combination of the results of other severs. Predictor groups are allowed to submit up to 5 models for each target protein, for which the amino acid sequence and some additional information, like the oligomeric state, are released by the CASP organization.

The assessment part of CASP is done independently by an assessor team for each prediction category. The assessor teams are usually research groups experts in the specific topic. They can also be predictor groups who performed well in previous CASP editions. The assessor teams are responsible for evaluating the quality of the prediction models and ranking the predictor groups by their overall performance within the category. In order to keep the assessment independent from the prediction and encourage new ideas, assessor teams are different every edition.

2.3.3 The CASP12 edition

The CASP12 experiment took place in the year 2016. The prediction phase started on the 2nd of May and ended on the 8th of August. The assessment phase, with a first checkpoint on the 11th of September with a meeting for assessors and organizers in Basel (Switzerland), finished at the CASP12 meeting from the 10th to the 13th of December in Gaeta (Italy).

The 12th edition of CASP was divided into the following 8 prediction categories, with their associated assessment teams:

1. **Contacts**: Alexandre Bonvin (University of Utrecht, Netherlands)
2. **High accuracy modeling**: CASP organizers
3. **Topology**: Matteo Dal Peraro (EPFL, Lausanne, Switzerland)
4. **Assembly (quaternary structure and complexes)**: Guido Capitani (Paul Scherrer Institute, Switzerland)
5. **Data assisted prediction**: Matteo Dal Peraro (EPFL, Lausanne, Switzerland)
6. **Refinement**: Francesco Luigi Gervasio (University College London, UK)
7. **Biological relevance**: Russ B Altman (Stanford University, US), Sean Mooney (University of Washington, US), Francesco Luigi Gervasio (University College London, UK) and Guido Capitani (Paul Scherrer Institute, Switzerland)
8. **Accuracy estimation**: CASP organizers

A total of 189 groups participated in CASP12, submitting a total of 54,970 models for the 82 released targets. All the information about the CASP12 experiment can be found at the *Prediction Center* web site: `http://predictioncenter.org/casp12`. The publication of a CASP12 special issue with articles for the results in all prediction categories is expected to be released in 2017.

Chapter 3

Assignment of quaternary structure from protein crystals

This chapter introduces the quaternary structure assignment problem, which was relevant for the generation of the target assemblies (references) from their experimental crystal structures during the CASP assessment. An evolutionary approach to the problem is presented and used in CASP12.

3.1 The assignment problem

Quaternary structure assignment consists in defining the assembly composition and mode of association of protein chains under the physiological conditions in the cell. As described in section 2.1, quaternary structures can be transient. Therefore, the assignment from crystal structures has the aim to identify the biological assembly represented in the crystal lattice.

As mentioned in section 1.5, the majority of protein structures are determined by X-ray crystallography. Crystallographic techniques solve for a lattice of protein subunits, formed by the interaction of multiple biological units, i.e. forming crystal contacts. Recovering the biological assembly from the lattice can be a non-trivial task. Some of the challenges faced in quaternary structure assignment from crystal structures will be presented in section 3.3.

Assembly composition

The composition of a protein assembly can be obtained by experimental techniques. Popular techniques are size exclusion chromatography (SEC) (Fekete et al. 2014) and analytical ultracentrifugation (AUC) (Schuck 2013). A comprehensive list of the experimental techniques available is provided in the review by Dr. Capitani (Capitani et al. 2015).

Mode of association

The mode of association of a protein assembly is defined by the protein-protein interfaces between the subunits of the complex, known as engaged interfaces. To describe the exact mode of association of a quaternary structure, the tertiary structure of each of the subunits in the assembly has to be known at a reasonable resolution. In addition, a careful analysis of the protein-protein interfaces in the crystal lattice is needed to distinguish between engaged interfaces and crystal contacts (Capitani et al. 2015).

Some experimental techniques, like SAXS or cross-linking, allow the study of protein association. Recently, a number of computational techniques, like EPPIC (Duarte et al. 2012) or PISA (Krissinel and Henrick 2007), have been developed to help in studying the problem. A detailed description of the most recent developments in the EPPIC software will be given in section 3.2.

3.2 Probabilistic biological assembly assignment

3.2.1 Interface classification problem

In crystallography, distinguishing between interfaces present in physiological conditions and crystal packing ones is known as the interface classification problem (Capitani et al. 2015). One approach to the problem is to use evolutionary information.

EPPIC

EPPIC is an evolutionary protein interface classifier that uses three types of scores: geometry, core-rim and core-surface (Duarte et al. 2012). The geometry score relies on the calculation of solvent accessible surfaces to estimate the number of fully buried residues upon interface formation. The principle is that stronger interfaces have a higher number of buried residues. The two evolutionary scores, core-rim and core-surface, rely on the principle that residues at the biological interfaces have a higher evolutionary pressure than the rest of the residues at the protein surface. Both scores compare, using multiple sequence alignments of close homologs, the evolutionary conservation of the fully buried residues in the interface against the conservation of other residues at the protein surface or at the interface rim (partially buried residues). In addition to the three scores, EPPIC also reports the interface area.

In previous versions of EPPIC, each of the scores produced a binary interface call, either biologically relevant or crystal contact, using a hard decision threshold optimized using benchmark datasets of interfaces. The final classification decision was made via a majority voting scheme of the three scores described. The major limitation of this classification scheme is that reliable

assignments cannot be distinguished from the unreliable ones, either due to missing information or because of scores near the decision boundary. We develop a classification confidence in this section, which will be part of the upcoming new version of the method.

EPPIC is available both as a web application (http://eppic-web.org) and a command line tool (https://github.com/eppic-team/eppic). The official release of EPPIC 3.0 is scheduled by mid 2017.

Benchmark datasets

Over the past years, datasets of protein-protein interfaces and biological assemblies have been created to benchmark the performance of interface classifiers and methods for biological assembly assignment. The following three datasets will be relevant for this section:

- Duarte-Capitani interface dataset (Duarte et al. 2012): collection of weak biological interfaces and strong crystal contacts in the difficult to classify region.
- Many dataset (Baskaran et al. 2014): large collection of biological interfaces and big crystal contacts.
- Ponstingl assemblies dataset (Ponstingl, Kabir, and Thornton 2003): homomeric and heteromeric symmetric biological assemblies between one and six protein subunits. A dataset of biological interfaces and crystal contacts can be derived from this assembly dataset.

3.2.2 Towards interface classification confidence

From a user perspective and in general for any classification method, it is important to distinguish between reliable and unreliable predictions. The current EPPIC classification scheme only provides a binary call: biological interface or crystal contact. The potential of adding a confidence together with the EPPIC call is to warn users about difficult to classify interfaces and suggest further investigation only in the cases strictly required.

In order to estimate the confidence of EPPIC classification calls we need to introduce a probabilistic classifier. Interface classification is binary, with two possible states: biological interface or crystal contact.

Let us define two variables associated with a particular interface:

- p = probability of the interface being biologically relevant
- q = probability of the interface being a crystal contact

Because it is a two-state problem, and the probabilities need to sum up to 1 for all possible outcomes, it follows that $q = (1 - p)$.

We can also define the log-odds ratio of an interface being biologically relevant as:

$$Log\ odds = log\left(\frac{p}{q}\right) \tag{3.1}$$

Logistic regression

One of the standard methods in machine learning for probabilistic classification is logistic regression. The logistic function $H(x)$ is defined as:

$$H(x) = \frac{1}{1+e^{-x}} \tag{3.2}$$

It is a sigmoid function with a single inflexion point at $H(x)$ equal to 0.5. At large values of the independent variable x, the function approaches asymptotically 1, and at small values of x it approaches asymptotically 0.

The main assumption of logistic regression is that the variable x is linearly dependent on the log-odds ratio of the two state probabilities. To check for this condition, the best logistic function fit to each EPPIC score and the interface area was computed (figure 3.1) with the generalized linear model fitting function (*glm*) of the *stats* package in *R* (Heiberger, Chambers A. And Freeny R. 1992), and using the Many benchmark dataset. Because dataset contains only big interfaces, interface areas below 600 Å^2 are not sampled. However, literature has shown that interfaces of these sizes are predominantly crystal contacts (Baskaran et al. 2014), and the logistic function fit agrees with the fact.

The fraction of biological interfaces for each EPPIC score bin correlates well with the estimated probability of interfaces being biologically relevant, in the score bins with enough samples to reliably compute the fraction. All scores, thus, are suitable features for a logistic regression classifier. However, we can also observe that the peak of all the score distributions falls close to the decision boundary of the logistic function, when the probability is 0.5. Therefore, no single score alone is a reliable interface classifier, so a combination of scores is needed to improve the classification accuracy and reliability (Duarte et al. 2012).

Model training

The independent variable of the logistic function, x in equation 3.2, can also be another function:

$$H(x) = \frac{1}{1+e^{-F(x_1,x_2,...,x_n)}} \tag{3.3}$$

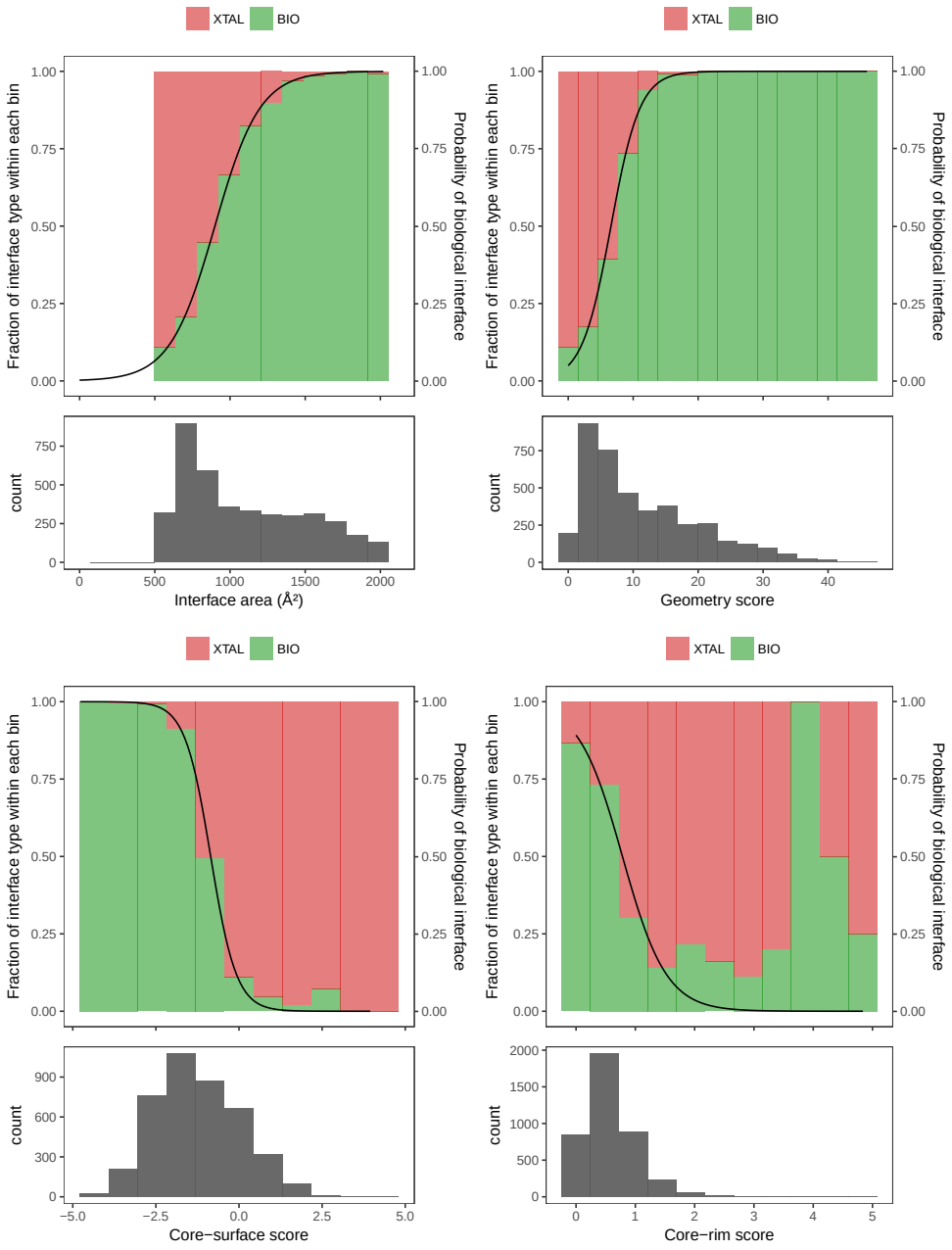


Figure 3.1: Probability of an interface being biologically relevant (BIO) or crystal contact (XTAL) for each of the EPPIC scores and interface area. On top of each subplot, bars are the binned frequencies of BIO and XTAL classes in the Many dataset and lines show the best logistic function fit. On the bottom of each subplot, the underlying score distributions are shown as black histograms.

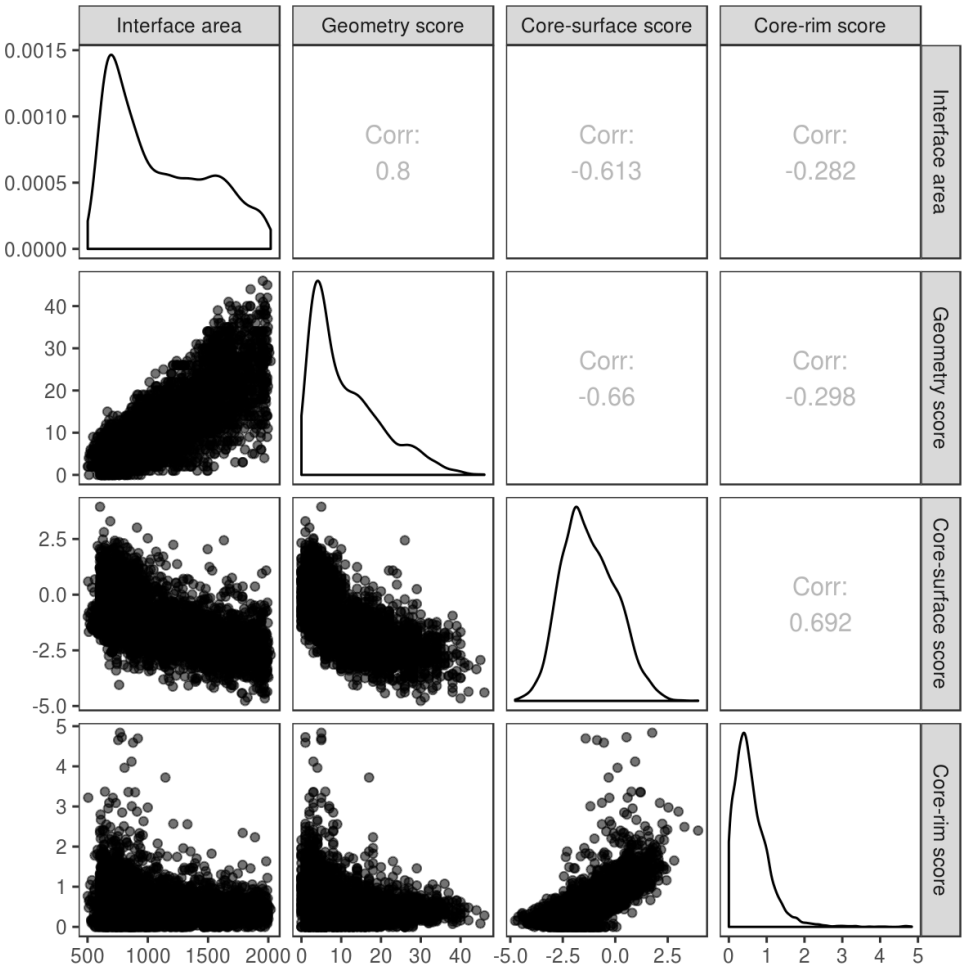


Figure 3.2: Correlation between the EPPIC scores and interface area in the Many dataset.

Here, x_1 to x_n are the features of the model. We will use a linear combination of the features, so the problem becomes a linear regression, where the weight w_i for each feature x_i has to be determined.

$$F(x) = w_0 + w_1x_1 + w_2x_2 + ... + w_nx_n \tag{3.4}$$

First of all, we would like to consider the correlation between the EPPIC scores, shown in figure 5.4. We observe that interface area and geometry score are highly correlated. This is expected, since bigger interfaces can accommodate a higher number of fully buried residues. Another observation is that the core-rim evolutionary score is uncorrelated to interface area and geometry scores, and weakly correlated with the core-surface score.

We trained the model like previously described, using the generalized linear model fitting function (*glm*) of the *stats* package in *R*. We used the Many interfaces dataset, removing the interfaces without evolutionary score predictions due to insufficient sequence homologs. We also removed the interfaces that are present in any of the other datasets, so that we can safely use those other datasets for the validation of the model.

In the training process, we used 10-fold cross-validation splitting randomly the dataset to estimate the out of sample error. This allowed us to select the scores to use as features for the model and prevent overfitting.

The logistic classifier, using the three EPPIC scores and the interface area as features, trained in the Many interfaces dataset obtained a cross-validation accuracy of 90%. Removing the interface area from the model did not have an effect on the accuracy, and the weight of the feature was redistributed among the other scores, mainly to the geometry score. This is consistent with the score correlation analysis. The coefficient of the fitted model for core-rim score turned out to be positive. This contradicts the expected sign of the coefficient, since lower values of the score should be indicative of biological relevance. We attribute this failure to the fact that the score is a ratio, so that differences close the value 0 are smaller than the differences at larger values. Removing the core-rim score from the model did not have an effect on the cross-validation accuracy.

We decided to use only the geometry and core-surface scores for the final logistic classifier. The coefficient of the geometry score (gm) is 0.31, that of the core-surface (cs) score is -2.1, and the intercept is -3.9.

$$p(gm, cs) = \frac{1}{1 + e^{-(-3.9 + 0.31gm - 2.1cs)}} \tag{3.5}$$

Treatment of missing evolutionary information

For some interfaces, there are not enough sequence homologs in the database to compute a reliable evolutionary score. We would like to account for the missing information in the estimation of the classification uncertainty.

Our approach is to calculate the most uncertain value of the core-surface score, i.e. the one that yields a probability of 0.5, and set this value to the interfaces with missing evolutionary scores. This can be understood as the score providing no evidence to the prediction, thus the lower classification confidence reported.

To calculate the most uncertain value of the core-surface score we need to solve for a system of linear equations. The logistic function is equal to 0.5

Table 3.1: Benchmarking statistics for interface classification using a probabilistic score.

Dataset	Accuracy	Sensitivity	Specificity	MCC	Brier score
Many	0.90	0.90	0.91	0.81	0.08
Ponstingl	0.92	0.84	0.97	0.83	0.07
Duarte-Capitani	0.82	0.92	0.71	0.65	0.15

when the exponential term x is equal to 0, so we need to solve for the values of the geometry and core-surface scores in:

$$-3.9 + 0.31gm + -2.1cs = 0 \tag{3.6}$$

We add the condition that both scores have the same importance to be able to have a unique solution to the previous equation:

$$0.31gm = -2.1cs \tag{3.7}$$

The most uncertain value of the core-surface score is -0.93. This value also corresponds to the most uncertain value of the classifier trained using only the core-surface score (figure 3.1). Missing core-surface scores due to the lack of evolutionary information are set to this value for probability calculation purposes.

Benchmarking

To validate the probability estimation we check two properties: the calibration and distribution of probabilities. First, the probabilities should be correctly calibrated; for instance, predictions assigned probability 0.9 should be correct about 90% of the time. Second, predictions should be as certain as possible (close to either 0 or 1). Figure 3.3 shows these two properties for each of the three interface benchmarking datasets described before. We observe that the probabilities are well calibrated for the three benchmark datasets, and that the probability distribution is maximal at the two extremes, which means that most interfaces were classified with high confidence.

The Brier score measures the accuracy of probabilistic predictions by computing the average of the squared probability differences for all the prediction events (Brier 1950). The score ranges from 0, the best score achievable by predicting always with maximal confidence without error, to 1, the worst score possible achieved by predicting always with maximal confidence the wrong class. The Brier scores for each of the benchmarking datasets are shown in table 7.1, together with other standard performance measures.

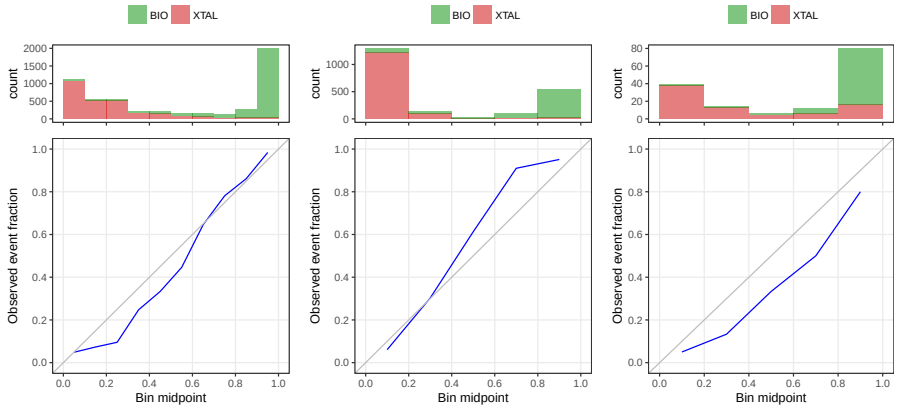


Figure 3.3: Probability distributions and calibration curves for interfaces in the DC (left), Many (middle) and Ponstingl (right) datasets.

3.2.3 Probabilistic assembly scoring

Assemblies as graphs

Protein assemblies can be defined by protein chains and the interfaces between them. In a graphical representation of assemblies, protein chains are nodes connected by the interfaces between them, the edges (figure 3.4). An interface that is part of an assembly is known as an engaged interface for that assembly. An interface is induced in an assembly if it is engaged, but removing it does not change the assembly composition. In graphical terms, removing an induced interface edge from the assembly graph does not change the number of components in the graph.

Assumptions

In deriving a score for individual assemblies in a crystal, we make the following assumptions:

1. There is exactly one biological assembly per structure.
2. The biological assembly is preserved in the crystal.
3. Interfaces engaged in an assembly are independent of each other.
4. Assembly formation is driven by the minimum number of interfaces required to maintain the assembly.

The first two assumptions are general for any quaternary structure assignment procedure from crystal structures. The first assumption is not valid when the biological assembly of the protein involves transient interactions. The second assumption is not valid if the crystallization conditions disrupted the biological assembly of the protein. The third assumption means that

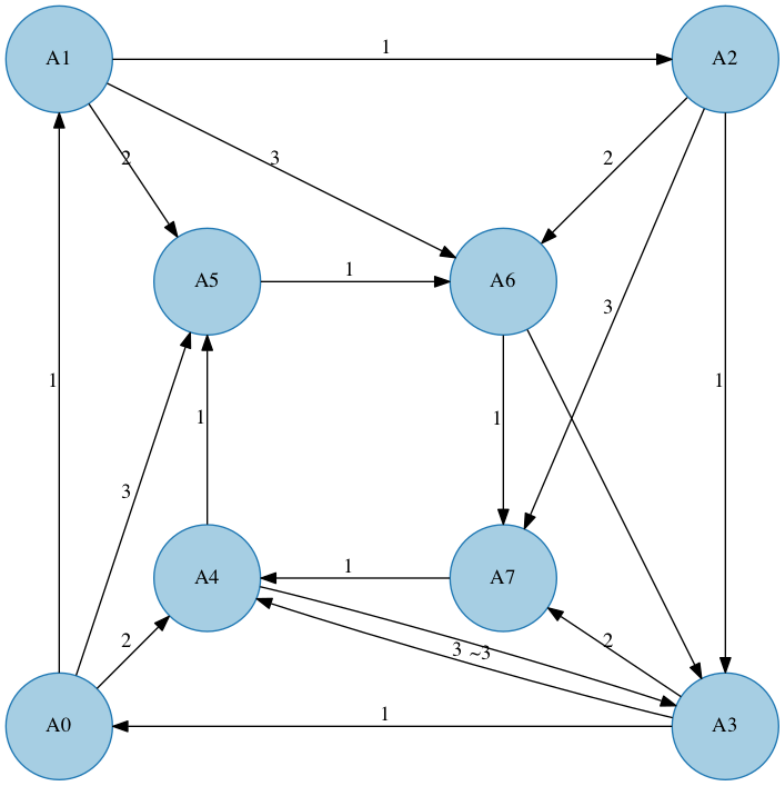


Figure 3.4: Example of an assembly representation as a graph, where nodes are protein chains and edges are interfaces. The graph represents a homomeric dihedral D2 assembly, where 8 copies of a chain A interact through 3 different types of engaged interfaces. All the interfaces in the assembly are induced, since removing one of them does not change the assembly composition (number of graph components). However, at least two of them are required to maintain the protein assembly.

the introduction and evolution of another interface does not depend on the other interfaces in the assembly. This assumption allows the simplification of probability calculations. The last assumption is related to the concept of induced interfaces, and it means that an assembly is maintained only by the biggest interfaces. Other interfaces not required for the assembly, the induced interfaces, are thus a direct a result of the formation of the bigger interfaces.

Assembly scoring algorithm

In a protein crystal structure, we can define the set of all protein interfaces as S. Each possible assembly in the crystal is defined by a subset $S_a \subseteq S$ of the set of crystal interfaces. Thus, each protein assembly can be defined by a boolean vector of length $|S|$, where entry s_i is true if interface i is engaged or false if the interface is not engaged in the assembly.

For each interface in the crystal, we have computed a probability of the interface being biological relevant, p, and the associated probability of the interface being a crystal contact, q. If we consider each interface as a random event the probability of an assembly is the joint probability of the individual interface events.

$$P(assembly) = P(s_1, s_2, ..., s_n) \tag{3.8}$$

We know the required outcome of each interface in the crystal to define a particular assembly, and we have assumed complete independence between the interface events (assumption 3). Therefore, we can use the chain rule to split the joint probability into the product of the individual interface probabilities. For a pair of interfaces the joint probability can be expressed as:

$$P(s_1, s_2) = p_{s_1} p_{s_2} \tag{3.9}$$

By definition, a set of engaged interfaces with or without an induced interface defines the same assembly. Therefore, if there are induced interfaces in the assembly, multiple subsets of the crystal interfaces S define the same assembly. To compute the probability of the assembly we add the probabilities of all the subsets that define the same assembly.

The probabilities of all possible combinations of interfaces in the crystal (all subsets of S) sum up to 1, since they defines the entire probability space. Thus, the score for each assembly in the crystal is also a probability, since it is a fraction of the total probability density.

Some combinations of crystal interfaces lead to assemblies that do not resemble the biological assemblies. These assemblies are known as invalid assemblies, and there is a set of rules to define valid biological assemblies in a crystal lattice (publication in preparation). Invalid assemblies also account for a fraction of the probability space, but from a theoretical point of view, their probability should be 0, since they do not obey the rules of biological assemblies and should always be crystallographic artifacts by definition. Therefore, once all the possible and valid assemblies of a crystal structure are generated, we add a normalization step so that the sum of the probabilities of all the valid assemblies sums up to 1.

Benchmarking

To validate the score for quaternary structure assignment, we used the Ponstingl assemblies dataset. We performed a similar analysis to the interface probability validation. The probability distribution and calibration plots are

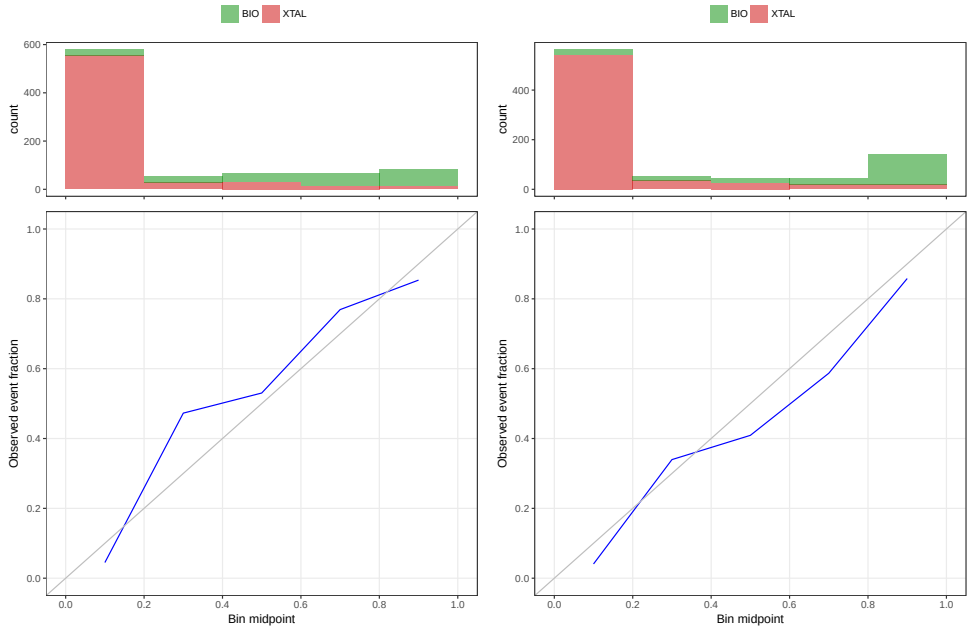


Figure 3.5: Probability distribution and calibration curves for the Ponstingl assembly dataset using non-normalized (left) and normalized (right) scores.

shown in figure 3.5. The calibration curve has an associated Brier score of 0.09 for the non-normalized scoring and 0.08 for the normalized one. Both the probability calibration and the distribution are better for the normalized scoring, suggesting that the assumption of invalid assemblies having zero probability of being biologically relevant is valid, at least for this dataset.

Evaluating the accuracy of assembly assignments is more difficult than for interfaces. For each structure, there is a single biological assembly from a variable number of possible assemblies. Once the method chooses one assembly as positive, the biologically relevant, all the other ones are automatically assigned as negative. This implies that the number of false negatives will always be equal to the number of false positives. Therefore, the precision measure alone suffices to evaluate the performance of the method as a classifier. In the Ponstingl assemblies dataset the precision of EPPIC is 75%.

However, assignment errors can be of different severity and this is not taken into account by precision. For instance, the assignment of monomer as the biological assembly of a tetramer is worse than a dimeric assignment. A confusion matrix of the properties of actual and predicted assemblies is needed for that purpose. In figure 3.6 the symmetry and macromolecular size confusion matrices of the EPPIC assignment in the Ponsingl datatset are shown. In the figure, we observe that errors are more frequent in higher oligomeric assemblies, especially those with dihedral symmetry. This is probably due

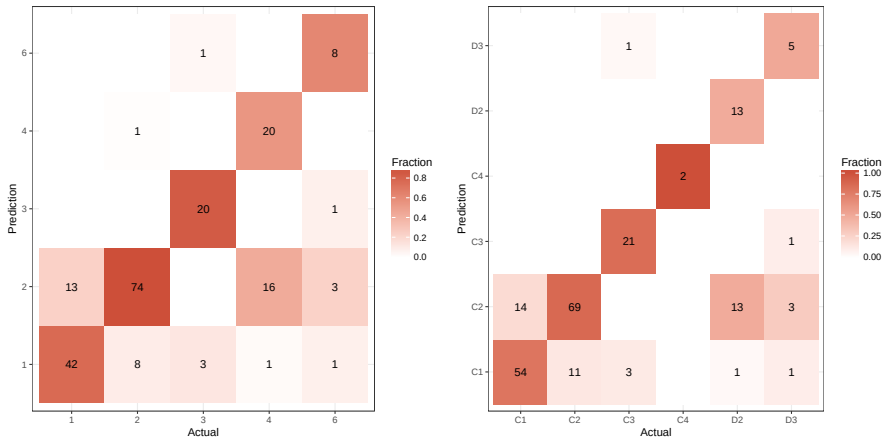


Figure 3.6: Confusion matrix plot for macro molecular size (left) and symmetry (right) of assembly assignments in the Ponstingl dataset.

to the fact that one of the two required interfaces of dihedral symmetries is generally weaker and, thus, a borderline interface classification.

3.3 Quaternary structure assignment in CASP12

The first step of the CASP assessment is to define the standard of truth that predictions should aim at reproducing, i.e. the target structure models. An accurate quaternary structure assignment protocol has to be defined in order to minimize the possible annotation mistakes in the targets.

3.3.1 Challenges

Structures obtained by electron microscopy (EM) or nuclear magnetic resonance (NMR) were considered to be in the correct quaternary structure state. On the other hand, the biological assembly assignment of crystal structures was carefully analyzed, since it can be sometimes challenging.

For some target proteins, the composition of the biological assembly was unknown. The authors of the structure could not provide experimental evidence, so we had to rely entirely on the computational evaluation of the assemblies in the crystal to produce the target structure model.

For other target proteins, the authors provided experimental evidence for a particular assembly composition. However, the crystal lattice contained more than one possible biological assembly with the same stoichiometry. An example is target *T0871o*, shown in figure 3.7, for which size exclusion chromatography showed evidence of a dimeric quaternary structure, but the crystal lattice contains two possible dimeric assemblies with a different interface.

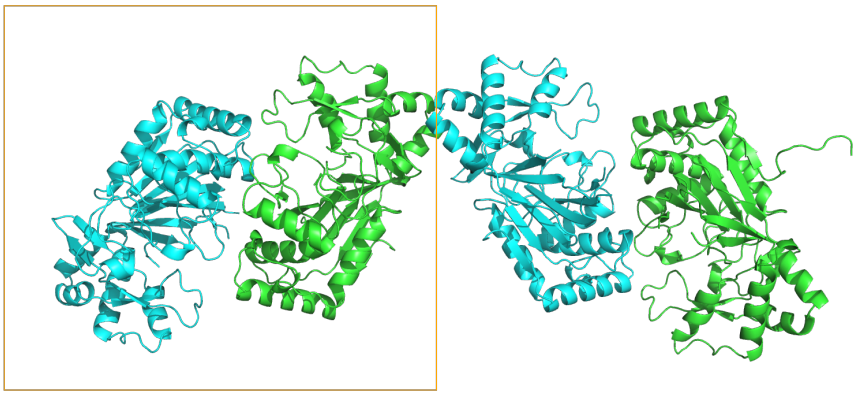


Figure 3.7: Two contiguous asymmetric units (orange square) of the crystal structure of target *T0871o*. There are two possible dimeric assemblies in the lattice.

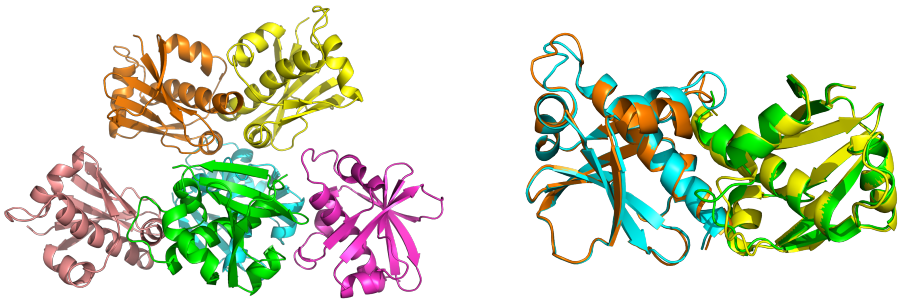


Figure 3.8: Asymmetric unit of the crystal structure of target assembly *T0875o* (left) and superposition of AB (green, cyan) and DF (orange, yellow) dimers (right). The superposition RMSD is 1.6 Å.

The presence of more than one possible assembly in the crystal is an underdetermination problem, while the presence of multiple copies of the same assembly is an overdetermination problem. In the latter, one of the assemblies has to be selected as the reference structure, since there can be differences between the assemblies due to crystal packing or modeling effects. One such example is target assembly *T0875o* (figure 3.8). Three copies of the dimeric assembly are present in the asymmetric unit of the crystal, and their superposition reveals an RMSD of 1.6 Å.

3.3.2 Assignment protocol

The main computational tool we used to assign and produce the quaternary structure of the CASP12 target proteins was EPPIC (Duarte et al. 2012), with the new features described in section 3.2. For EPPIC assignments with low confidence, we additionally analyzed the structures with PISA (Krissinel

and Henrick 2007), which uses a thermodynamic approach complementary to the evolutionary approach of EPPIC. Sequence and structural homologs were also checked, if available.

In all cases, we requested experimental evidence for the oligomeric state of the target proteins from the authors of the target experimental structures. All the experimental evidence we obtained came from size exclusion chromatography (SEC) data.

The experimental information and the computational tools were never incompatible, although some cases had unreliable computational assignments. In those cases, we requested confirmation from the experimentalists after describing the computational analysis.

If multiple copies of the biological assembly were present in the asymmetric unit, we selected the stronger assembly, in terms of the strongest computational evidence for being biological.

3.3.3 Target assemblies overview

The quaternary structure of the oligomeric targets of the CASP12 edition is shown in table 3.2. The assemblies were representative of the PDB in macromolecular size, symmetry and number of entities.

The prediction difficulty of target assemblies was assigned to one of the following categories:

- Easy: templates with detectable sequence similarity and the same quaternary structure exist. Equivalent to tertiary structure template-based modeling.
- Medium: partial templates, sharing a subset of chains and interfaces with the target and with detectable sequence similarity, and/or structural templates with no sequence similarity exist.
- Hard: no quaternary structure templates exist (novel assembly). Equivalent to tertiary structure free modeling.

3.4 Conclusions

In this chapter we have presented the contributions to EPPIC, a computational method to assign the biological assembly of proteins from their crystal structures. The estimation of the assignment confidence developed in this thesis is a valuable information for crystallographers, because it suggests additional investigation for unreliable cases only.

In this chapter we have also combined experimental information and computational tools in a protocol to assign the reference biological assemblies

Table 3.2: Overview of CASP12 target assemblies.

Target	Name	Stoich.	Sym.	Difficulty
T0860o	Raptor adenovirus fibre head	A3	C3	EASY
T0861o-T0862o-T0870o	Cysk-Toxin-Immunity Complex	A2B2C2	C2	MEDIUM
T0866o	MlaD	A6	C6	HARD
T0867o	Lizard adenovirus 2 fibre 1 head	A3	C3	EASY
T0868-T0869	CdiA-CT and CdiI complex	AB	C1	HARD
T0873o	TtnD	A4	D2	MEDIUM
T0875o	LV4 2A2	A2	C2	HARD
T0880o	Mouse adenovirus 2 fibre head	A3	C3	MEDIUM
T0881o	Goose adenovirus 4 fibre head	A3	C3	EASY
T0884-T0885	CPX19 CdiAtox-CdiI2 complex	AB	C1	HARD
T0888o	Lizard adenovirus 2 fibre 2 head	A3	C3	MEDIUM
T0889o	D-sorbitol dehydrogenase	A4	D2	EASY
T0893o	CckA histidine kinase DHp-CA	A2	C2	EASY
T0894-T0895	CDI204 E1-E2 complex	AB	C1	HARD
T0897-T0898	ANL-KT U1-U2 complex	AB	C1	HARD
T0903o-T0904o	LGN-Inscuteable complex	A2B2	C2	MEDIUM
T0906o	Fructose biphosphatase	A8	D4	EASY
T0909o	LH3	A3	C3	EASY
T0912o	BT1002	A2	C2	HARD
T0913o	F4ZCI3	A6	D3	HARD
T0914-T0915	CPX209 complex	AB	C1	HARD
T0917	Red sea protein	A2	C2	EASY
T0921-T0922	ScaB Cohesin-Dockerin complex	AB	C1	EASY
T0929o	AP205 capsid coat protein	A2	C2	HARD
T0930o	STRA6 retinol uptake receptor	A2	C2	HARD
T0931o	DM77-3428	A2	C2	MEDIUM
T0932o	TIPRL	A2	C2	HARD
T0933o	FliD	A6	C6	HARD
T0934o	Bd0886	A2	C2	HARD
T0945o	DPAGT1	A2	C2	HARD

of CASP12 target proteins. Since it was the first CASP edition with a quaternary structure prediction category, we had to design the protocol and face some unexpected challenges, but we hope that our work can be useful for the quaternary structure prediction assessments of future CASP editions.

Chapter 4

A scalable algorithm for the alignment of quaternary structures

This chapter describes a novel approach to align quaternary structures that is scalable with the number of chains. We analyze the algorithm complexity and performance for a few example protein assemblies. In addition, we provide the source code and an executable of the algorithm implementation used in the CASP12 assessment pipeline.

4.1 The alignment problem

The alignment of quaternary structures requires building sets of equivalences both at the protein chain and residue levels. This is due to the possible different chain ordering in the input structure files of the alignment. The chain equivalences are used to input the same chain ordering for the two structures to a regular structure alignment algorithm in order to obtain the alignment at the residue level.

The main open challenge is, thus, the calculation of the optimal chain equivalences, since regular tertiary structure alignment algorithms can be used for the residue equivalences.

4.1.1 Brute force approach

The brute force approach is to try all possible chain permutations of one structure, equivalent to evaluating all possible equivalences between the chains of the two aligned structures. The chain mapping that yields the largest number of aligned residues is selected as the result.

Although this is a valid approach, the number of chain mappings that have to be evaluated scales exponentially with the number of chains of the smallest

structure. If we define n as the number of chains of the smallest structure, the possible number of chain permutations is $n!$.

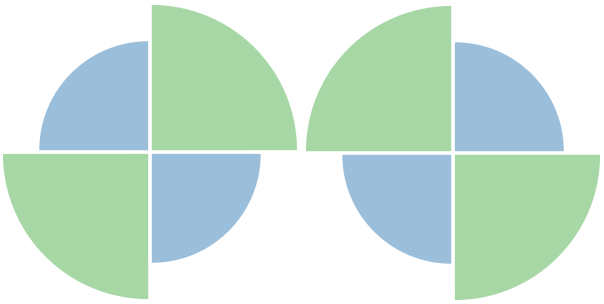
Although the number of chains of protein assemblies is usually small, algorithms using the brute force approach will fail to evaluate a significant portion of the assemblies in the PDB due to its poor scalability. As an estimate, the alignment of two assemblies with 10 chains would take 3.6 million evaluations, or around 20 days at an average time of 1 second per structure alignment evaluation. A quick search in the PDB reveals over 4,000 structures with biological assemblies with 10 chains or more. Therefore, a different approach with better scalability is needed to do database wide comparisons of quaternary structures. This need also applies directly to the CASP experiment, were tens of thousands of models are evaluated every edition.

Popular algorithms to align protein assemblies use the brute force approach to generate the optimal chain equivalences between the structures. They recognize that the scalability of the algorithms is the main disadvantage and propose it as a future research line.

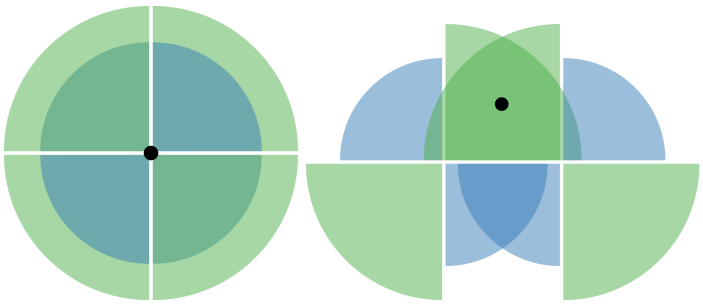
VAST+ (Madej et al. 2014) is entirely a web-based tool, so it could not be considered as an option for the CASP model evaluation pipeline. *MM-align* (Mukherjee and Zhang 2009) does have a command line tool that could have been incorporated into the CASP model evaluation pipeline, but we realized that the tool does not perform the alignment at the chain level (the combinatorial approach described in the article), and the final alignment is incomplete if the chains of the structures are not in the same order. Therefore, there was a big need for an efficient command line tool to align quaternary structures.

4.1.2 A simplifying observation

When we consider the transformation matrix needed to superpose two protein assemblies, we can make an observation: each chain of the assembly is transformed by the same operation in order to superpose to its equivalent chain in the other assembly. Therefore, it is possible to approximate the optimal global transformation matrix by using the local transformation matrix from a single chain pair superposition. Applying the local transformation matrix to each other chain of the assembly will superpose them to their equivalent chain in the other assembky, so a measure between two chains (like RMSD or centroid distance) can be used to select the pairs of closely superposed chains and generate the optimal chain equivalences. Figure 4.1 graphically shows the idea: only a single pair of equivalent chains is needed to uniquely determine all other chain equivalences, given that the local transformation operation is a good approximation of the global one.



(a) Toy 2D geometric complex, in two different orientations rotated by 90° to each other.



(b) Superposition of the complexes using the global centroid (left) or the local centroid of a single chain pair (right). A 90° rotation applied to either of the two centroids results in the same superposition of all the chains in the complex.

Figure 4.1: Schematic representation of the global transformation operation approximated by the local superposition of a 2D geometric complex.

4.2 The *QS-align* algorithm

The *QS-align* algorithm exploits the observation mentioned in the previous section to efficiently build a quaternary structure alignment.[1] Since we have no intitial indication about a correct chain pair, all the possible pairs of equivalent chains between the two assemblies are tried as a starting point. If the number of chains in each structure is n, the possible starting chain pairs is $\binom{n}{2}$, so the number of evaluations scales quadratically with the number of chains. This is a huge improvement in efficiency over the brute force approach, which scales exponentially.

For each starting point, the algorithm increments the number of equivalent chains by iteratively adding pairs at a close distance (measured with RMSD, chain centroid distance or chain orientation), until no more matchings are possible. The final alignment reported is the one with the maximum number of aligned chains, while having a lower RMSD. *QS-align* uses the Combina-

[1] Not to be confused with the QS-score (Bertoni submitted) discussed in 5.3.2

Table 4.1: Quaternary structure alignment examples with *QS-align*. Abbreviations: *N. Ch./Res.* Number of Chains/Residues; *A. Ch./Res.* Aligned Chains/Residues

Name	Structures	N. Ch.	N. Res.	A. Ch	A. Res.	RMSD	Time
PCNA	3hi8+3ifv	3+3	741+741	3	639	1.78	6s
PRC	1dxr+2jiy	4+3	1190+849	3	718	1.57	48s
Phycocyanin	2vml+2bv8	12+12	2004+2004	12	1920	1.24	8s
Cytochrome	1bcc+1kb9	9+11	2048+2180	8	1561	2.25	124s

torial Extension (CE) algorithm (Shindyalov and Bourne 1998) to perform the structural alignments at the residue level, using the C_α atoms as representative points for each residue.

Example alignments of quaternary structures are shown in table 4.1. We can observe that the calculation times are reasonable given the complexity of the problem. The slowest alignment is the bacterial cytochrome, which takes around 2 minutes to complete. All the chain equivalences are correct and complete in the alignments, compared to their manual alignment.

4.2.1 Scaling performance

To evaluate the scaling of the algorithm implementation as a function of the number of chains per input assembly we use a series of alignments between two phycocyanin structures (2BVL and 3JDB), dihedral pseudosymmetric heterododecamers. The comparison of these two assemblies is interesting because they are big heteromers with inconsistent chain naming and ordering, so the alignment at the chain level is needed and the brute force approach would not be feasible.

We generated computing times as a function of the number of chains removing pairs of equivalent chains of the input structure files (figure 4.2). The computing time scales almost linearly up to 12 chains per input structure and the time required to align two assemblies with 12 chains is only double the amount of computational time for the alignment of two single chain structures. These results show that the alignment is scalable (low complexity) and efficient (low scaling factor).

The computing time of the brute force approach on 12 chains $T_{bf}(12)$ can be theoretically extrapolated from the *QS-align* computing time $T_s(12)$, the number of evaluations of the brute force approach $E_{bf}(12)$ and the QS-align approach $E_s(12)$:

$$T_{bf}(12) = \frac{T_s(12) - T_s(1)}{E_s(12)} E_{bf}(12) = \frac{4.5s}{\binom{12}{2}}\, 12! > 1000\ years \tag{4.1}$$

Therefore, the comparison of these two assemblies can only be done using the *QS-align* approach, since the brute force approach would not finish in

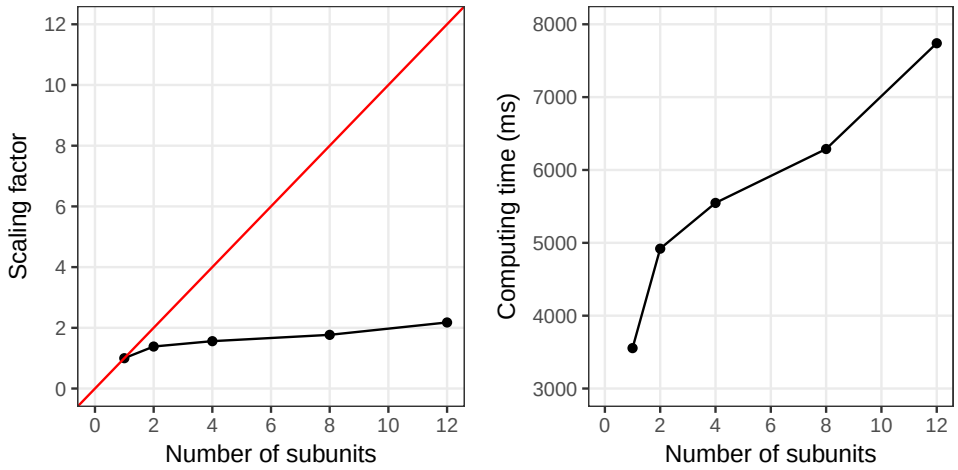


Figure 4.2: Scalability of the *QS-align* algorithm. Scaling factor compared to the alignment of two single chain complexes (left) and absolute computing time (right) as a function of the number of chains per input structure. The phycocyanin alignment (2vml vs. 2bv8) was used for the calculation.

a reasonable time. Unfortunately, we could not compare the scaling performance against any of the algorithms using the brute force approach, since the command line version of *MM-align* does not try all the possible chain combinations (mentioned in section 4.1) and the alignment is thus incomplete.

4.2.2 Availability

An implementation of the *QS-align* algorithm has been included in the BioJava library (Prlić et al. 2012), since version 5. BioJava is the leading open-source structural bioinformatics library, hosted on GitHub (https://github.com/biojava/biojava). Further documentation about the algorithm and how to use it programatically is available and kept up to date in the BioJava tutorial (https://github.com/biojava/biojava-tutorial).

Running the algorithm programatically in Java, using the source code, allows users to customize the parameters, like the score thresholds, and options, like the choice of distance measure or structure alignment algorithm, and create their own output format. However, many users might not have the time or skills to learn to use the algorithm programatically. Therefore, a simple command line tool that uses the default parameter options and outputs a summary tab-delimited line of the result is also provided. The tool can be downloaded from the releases section of the repository: https://github.com/lafita/qs-align.

Java 8 is required to run the executable JAR file, which requires a query and a

target structure, with an optional output file for the results. This is specified by the help message of the tool:

```
usage: java -jar QsAlign.jar [options]
-h,--help Print usage information
-t,--target <file> Model of the first Structure [required]
-q,--query <file> Model of the second Structure [required]
-o,--output <file> Path to the output file [default: stdout]
```

As an example, to align a query structure against a target, we need to run:

```
java -jar qs-align.jar -t target.pdb -q query.pdb -o result.tsv
```

Applications in CASP12

This tool has been successfully used in the CASP model evaluation pipeline to characterize systematically the assembly models. The number of equivalent chains (chain coverage), the length of the residue alignment (residue coverage) and the RMSD of the alignment were useful information for the assessment. The results of the tool for all CASP12 models are available on-line at the *Prediction Center* web site, inside the CASP12 multimer results page: http://predictioncenter.org/casp12/results.cgi?tr_type=multimer.

4.3 Conclusions

In this chapter we have developed a new algorithm for the alignment of quaternary structures based on an additional assumption that reduces the complexity of the problem. The algorithm allows efficient comparisons of protein assemblies, as we have demonstrated with examples, and is scalable as a function of the number of chains. We have also created and made available an implementation of the algorithm, which has been successfully used for the CASP12 quaternary structure prediction assessment.

As with sequence and tertiary structure alignment algorithms, the applications of this algorithm are very diverse. It allows systematic searches over large databases of quaternary structures. On the other hand, the chain matching produced by the algorithm can be used as input to other algorithms that need the chain equivalences between structures to operate.

We plan to continue the maintenance and development of the algorithm in the BioJava library. We would also like to improve the command line tool, adding more options and parameters available to the user for customization of the alignments. Other possible future developments include the creation of a graphical interface, or even a web interface, in order to show the alignment of the assemblies directly in a molecular viewer.

Chapter 5

Quality assessment of quaternary structure models

This chapter introduces scores to evaluate the quality of the assembly prediction models. The scores are computed for every model submitted to the CASP12 experiment using an evaluation pipeline and the results are used to assess the performance of the prediction methods.

5.1 Assessment goal

The goal of the CASP assessment is to (i) quantify the quality of the prediction models in a relative and global scale, (ii) evaluate to what degree are models useful to understand the protein function compared to the experimental structures, and (iii) provide insights into the prediction methodologies to identify established and promising approaches.

5.2 Model quality scoring

5.2.1 Interface representation

Protein-protein interfaces can be represented as a bipartite graphs of residue interactions (figure 5.1. Edges are defined with a measure of the interaction of two residues, for example a distance cutoff. From the interface graph two sets can be defined:

- R: set of contacting residues, nodes of the graph.
- C: set of residue-residue contacts, edges of the graph.

Quality measures of assembly prediction models can be defined as the accurate reproduction of the two sets in the target assemblies.

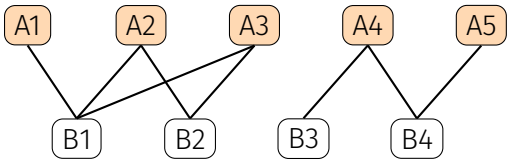


Figure 5.1: Interface as a bipartite graph. Orange squares represent residues in chain A, while white squares are residues in chain B. Edges represent geometric proximity between the residues.

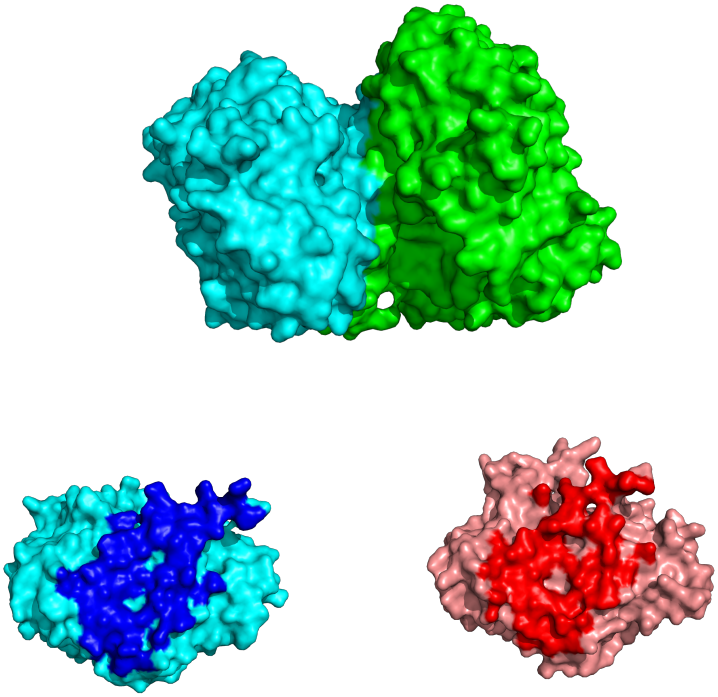


Figure 5.2: Visual comparison of interface patches. Dimeric target *T0917o* (top), chain A (bottom left) and corresponding chain of a model (bottom right). Interface patches of the target and the model are highlighted. Although the overall position of the patch is correctly predicted, some differences can be appreciated.

5.2.2 Interface patch score

An interface patch is defined as the set of residues $\mathbf{R}$ of one chain with at least one heavy atom within a distance threshold (5 Å) to a heavy atom of another chain. We would like to measure how different the interface patches of the model $\mathbf{R}_M$ and the target $\mathbf{R}_T$ are. A visual representation of the score is shown in figure 5.2.

We used a common set comparison metric, the Jaccard distance, which for

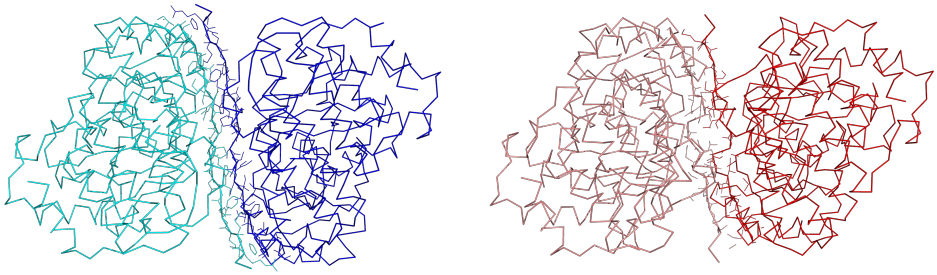


Figure 5.3: Visual comparison of interface contacts. Backbone representation of dimeric target *T0917o* (left) and its best model (right), colored by chain. The contacting residues at the interface of each dimer are shown in lines.

the comparison of a model M against the target T can be defined as:

$$J_D(M,T) = 1 - \frac{|\mathbf{R}_M \bigcap \mathbf{R}_T|}{|\mathbf{R}_M \bigcup \mathbf{R}_T|} \in [0,1] \tag{5.1}$$

The interface patch score, thus, is a measure of the dissimilarity between residues at the interface. The score can have a value between 0 and 1, and lower values mean that the model reproduces more accurately the residues at the target interface (higher quality).

5.2.3 Interface contact score

The interface contacts $\mathbf{C}$ are defined as the set of residue-residue pairs between different chains with at least one heavy atom closer than a distance threshold (5 Å). An example of the differences measured by the score is shown in figure 5.2.

We used a common performance metric for binary predictions, the F1-measure, which is the harmonic mean of the precision and recall of contact predictions of the model compared to the target.

The precision P is the fraction of contacts in the model $\mathbf{C}_M$ that are actually present in the target $\mathbf{C}_T$.

$$P(M,T) = \frac{|\mathbf{C}_M \bigcap \mathbf{C}_T|}{|\mathbf{C}_M|} \tag{5.2}$$

The recall R is the fraction of target contacts $\mathbf{C}_T$ that are correctly reproduced by the model $\mathbf{C}_M$.

$$R(M,T) = \frac{|\mathbf{C}_M \bigcap \mathbf{C}_T|}{|\mathbf{C}_T|} \tag{5.3}$$

The contact similarity score combines the precision and recall of contact predictions with a harmonic mean, called F1 score, defined as:

$$F1(M,T) = 2 \cdot \frac{P(\mathbf{C}_M, \mathbf{C}_T) \cdot R(\mathbf{C}_M, \mathbf{C}_T)}{P(\mathbf{C}_M, \mathbf{C}_T) + R(\mathbf{C}_M, \mathbf{C}_T)} \cdot 100\% \in [0, 100] \tag{5.4}$$

The interface contact score, thus, is a measure of the similarity between residue contacts in the interface. The score can have a value between 0 and 100, where higher values mean that the model reproduces more accurately the interface contacts of the target (higher quality).

5.2.4 Interface score comparison

Interface patches are easier to predict than specific contacts at the chain interfaces. Therefore, although both scores are correlated, the contact score reflects the high resolution details and the patch score coarser features.

The actual correlation between the two interface scores for all the submitted models is shown in figure 5.4. We can observe that for high accuracy models (lower right), the two scores are highly correlated. However, for low accuracy models (top left), most of the difficult targets, the patch score can still be an informative measure while the contact score is too low to discriminate the better models.

An example of the discriminative difference between scores are models submitted to dimeric target assembly *T0945o*. As shown in figure 5.5, the contact score for the model cannot distinguish between completely wrong assembly predictions and those that identify the interface patch correctly and only fail in the chain orientations.

5.2.5 Symmetry deviation

We used the quaternary structure symmetry detector algorithm from *BioJava* (Prlić et al. 2012) to determine the symmetry axes of the assemblies. A description of the algorithm can be found in the symmetry chapter of the BioJava tutorial: https://github.com/biojava/biojava-tutorial/blob/master/structure/symmetry.md.

The symmetry deviation measure is defined as the RMSD between the chains of an assembly constrained to the symmetry axes. In other words, chains are rotated using the symmetry axes and superposed to a single chain position (figure 5.6).

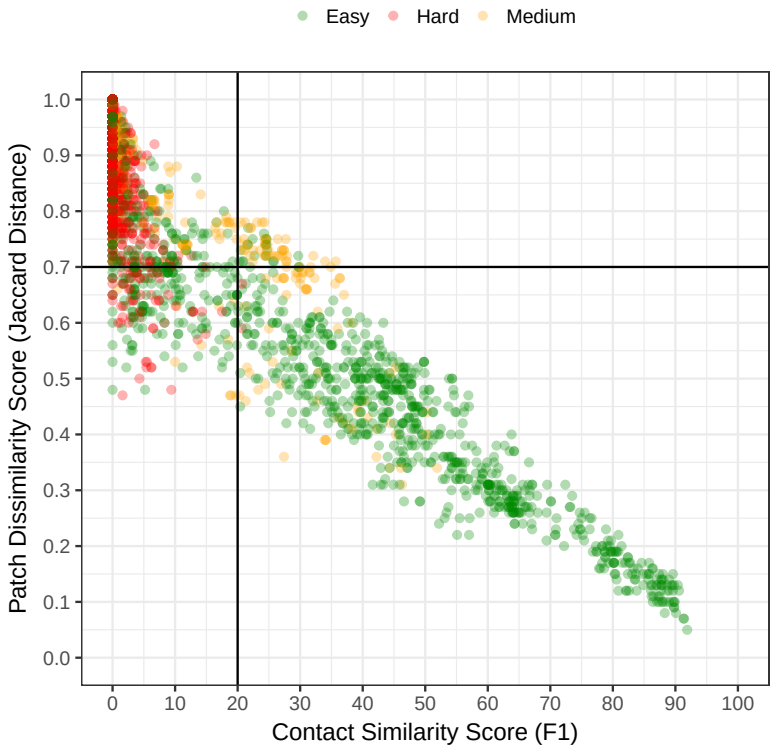


Figure 5.4: Correlation of interface patch (Jaccard distance) and interface contact (F1) scores for all the oligomeric models submitted to CASP12. Models colored by target difficulty. Black lines indicate the threshold values at which, at the left of contact score and at the top of patch score, score differences are considered uninformative.

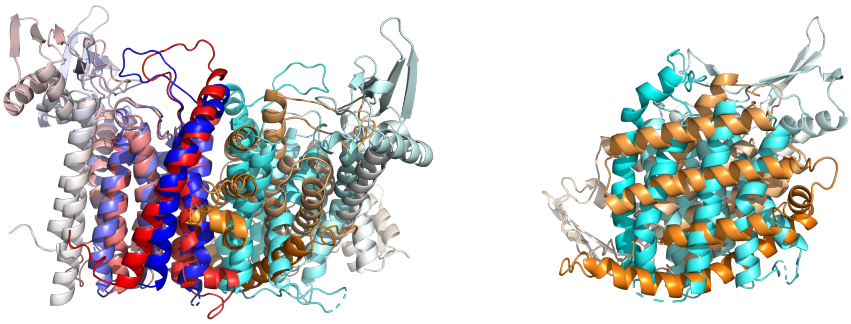


Figure 5.5: A model with significant patch score but insignificant contact score. Superposition of target dimer *T0945o* (blue/cyan) and its best model (red/orange) based on a single chain (left). The non-superposed chains, as seen from the interface (right). The model reproduces accurately the binding patch of the chain ($J_D = 0.52$). However, the interface is rotated, so the contact score is in the uninformative region ($F1 = 5.4$).

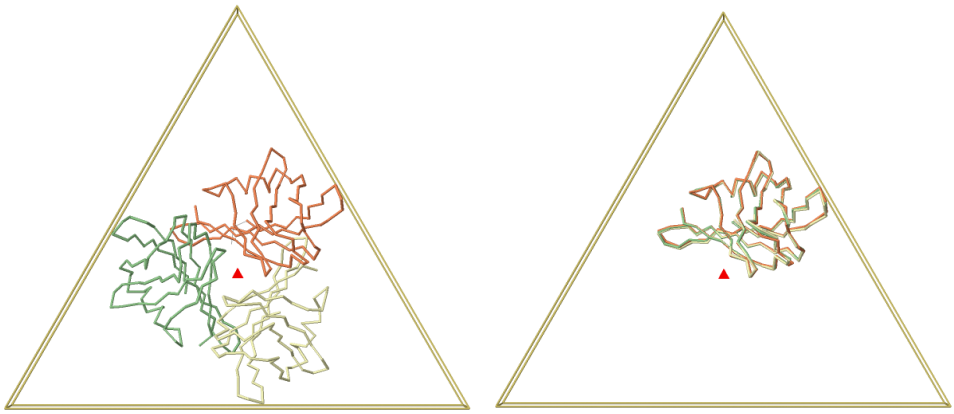


Figure 5.6: Example of the symmetry deviation measure. Cyclic symmetry axis and box of target *T0867o* (left) and superposition of the three chains constrained to the symmetry axis (right).

5.3 Prediction performance

5.3.1 Performance per target

The scores defined in section 5.2, together with some other metrics, were calculated for all the prediction models in the CASP evaluation pipeline. The scores for each model are available on-line at the *Prediction Center* website, at the CASP12 multimer results page: http://predictioncenter.org/casp12/results.cgi?tr_type=multimer.

An interesting analysis is to plot the distribution of scores of the models for each target, to identify those targets that could be successfully modeled and those for which no good predictions were submitted.

Reasonable quality predictions of the interface patch score are achieved for the majority of targets, even for some of the difficult ones (figure 5.7). On the other hand, no reasonable quality predictions of the interface contacts were submitted for the hard targets (figure 5.8), which shows the extreme difficulty of the problem.

5.3.2 Baseline performance

The comparison of the performance of the prediction methods against a baseline performance, represented by a very simple or easy to implement solution to the problem, has proven to be very useful for the analysis of method sophistication in the past. In CASP9, only one server group performed better than a naive predictor baseline, defined as copying the quaternary structure of the highest sequence similarity protein to the target (Mariani et al. 2011). In CASP11, groups improved over the sequence similar-

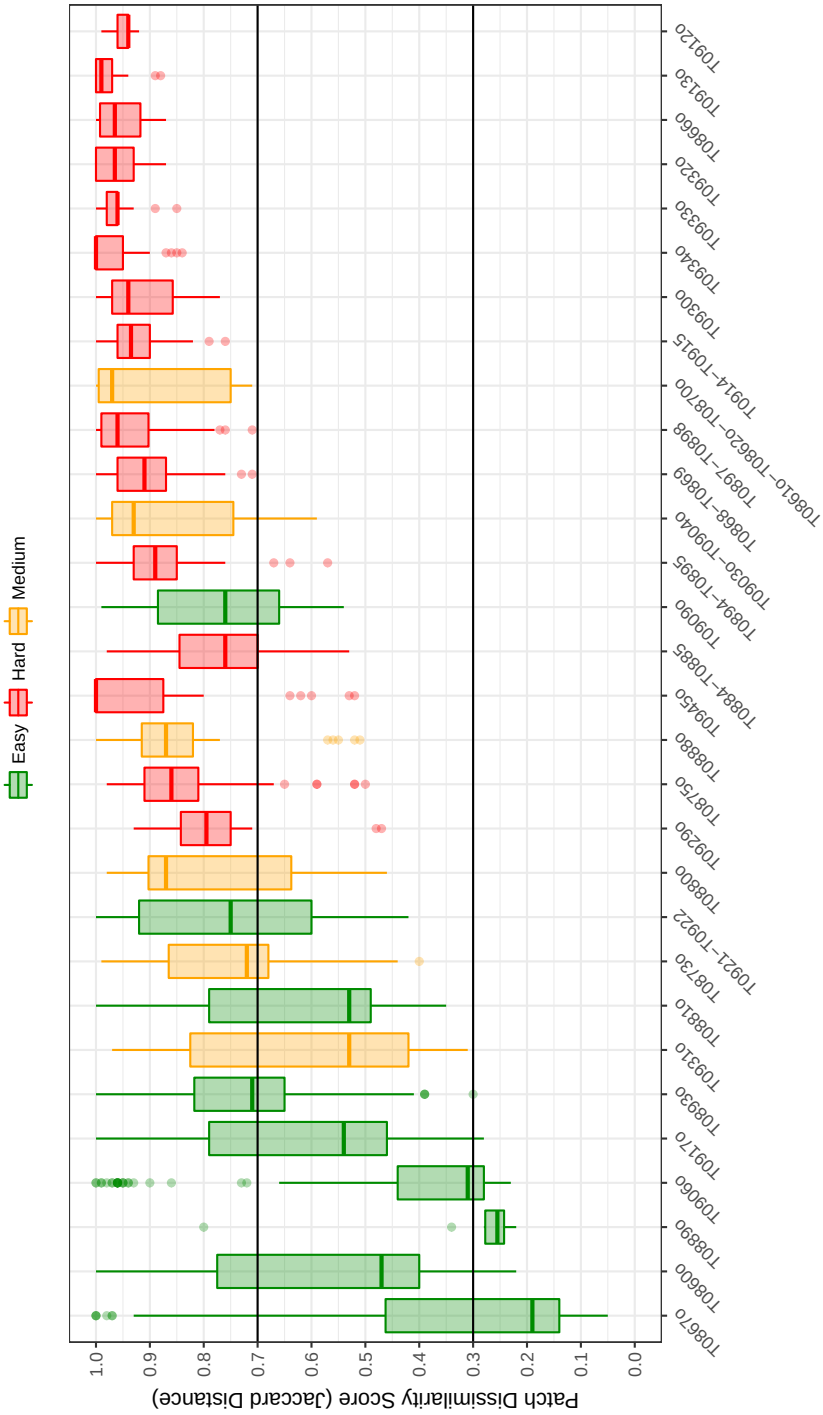


Figure 5.7: Quality of interface patch predictions for all models for each target. Targets colored by their prediction difficulty. The horizontal black lines roughly separate uninformative scores (top) and scores indicative of high quality predictions (bottom).

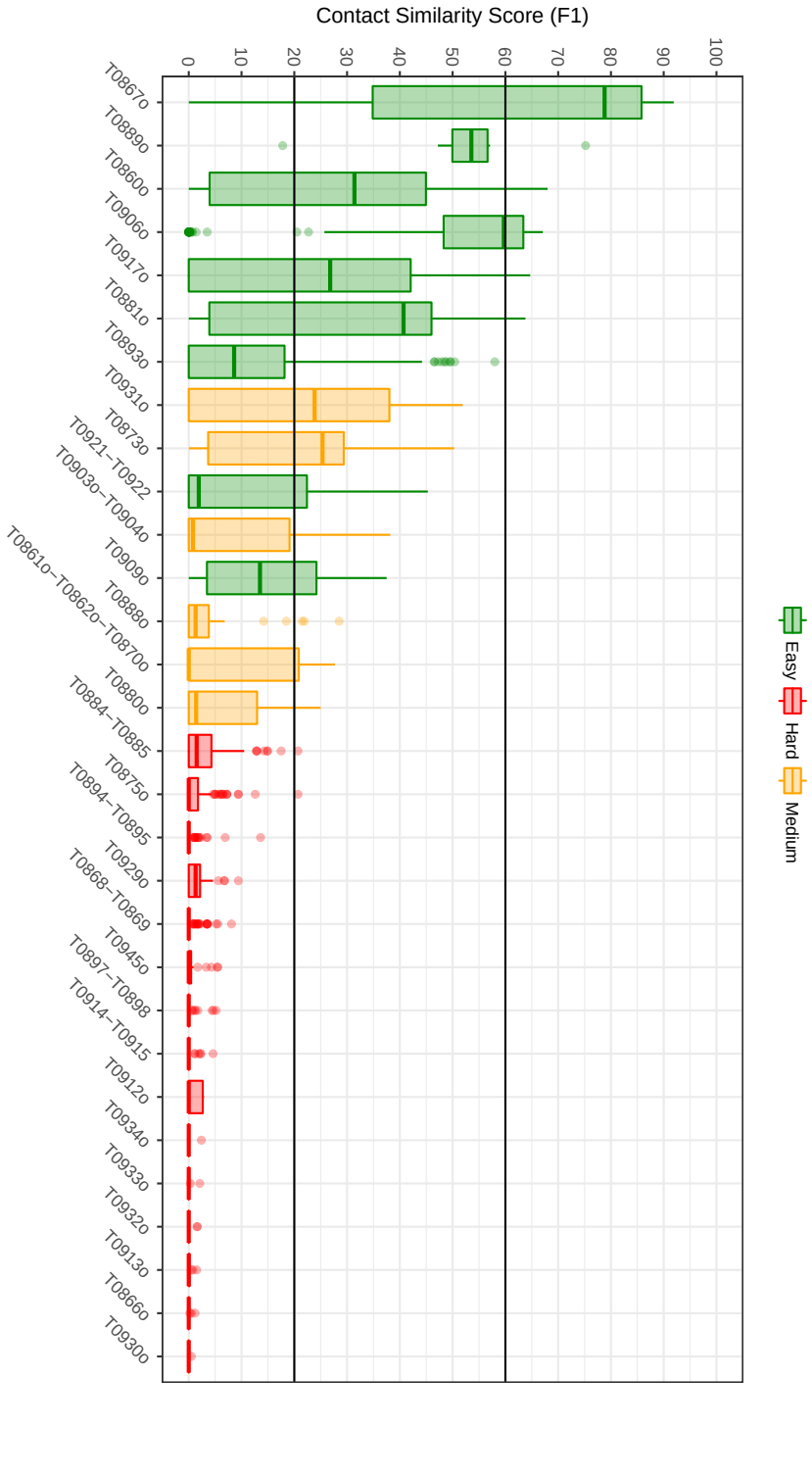


Figure 5.8: Quality of interface contacts prediction of all models for each target. Targets are colored by their prediction difficulty. The horizontal black lines roughly separate uninformative scores (bottom) and scores indicative of high quality predictions (top).

ity baseline for almost all targets, and the improvement was significant for some of the subset of difficult dimers (Lensink et al. 2016).

To define a performance baseline we used the *QS-score* metric. *QS-score* is an interface similarity score, like the interface contact score described in section 5.2, but developed specifically to search and cluster for quaternary structure similarity among protein assemblies (Bertoni submitted). The method performs an alignment of the sequences of the equivalent chains of the two input assemblies and computes a weighted fraction of shared C_β contacts in all the interfaces. We computed a baseline performance using the *QS-score* of the top scoring sequence template, as determined by HHSearch, for each target.

In the assembly prediction category of the CASP12 edition we observe that most groups perform consistently over the baseline, with few exceptions (figure 5.9). Furthermore, the improvement over the available templates for some targets is very impressive, like *T0917o* or *T0860o*. As observed in CASP11, CAPRI groups manage to predict acceptable models in the harder targets, while CASP methods still struggle in the prediction of oligomeric assemblies when no templates are available. This can be explained by the fact that CAPRI groups are specialized in the docking problem, while the predominant focus of CASP groups continues to be tertiary structure prediction.

5.3.3 Symmetry constrains

An important property of quaternary structure is symmetry. We wanted to analyze how prediction methods consider symmetry in the generation of their assembly models. We plotted the symmetry deviation of all the models submitted by each group, without checking for the correct assembly or symmetry. From figure 5.10 we can identify (i) groups that constrain the models to be perfectly symmetric (e.g. *Grudinin*), (ii) groups that never produce asymmetric models, but that allow differences (flexibility) in the tertiary structure of the subunits (e.g. *FONT*), and (iii) groups that do not consider symmetry in their modeling, allowing some asymmetric assembly models (e.g. *Bates_BMM*).

5.4 Biological relevance assessment

The goal of the biological relevance assessment is to identify targets where the quaternary structure has a clear biological relevance, and evaluate to what degree are models useful to understand the function of the protein, compared to the experimental structure. Because biological relevance is hard to quantify, we focused our assessment in two target assemblies with a particular functional relevance of their quaternary structure.

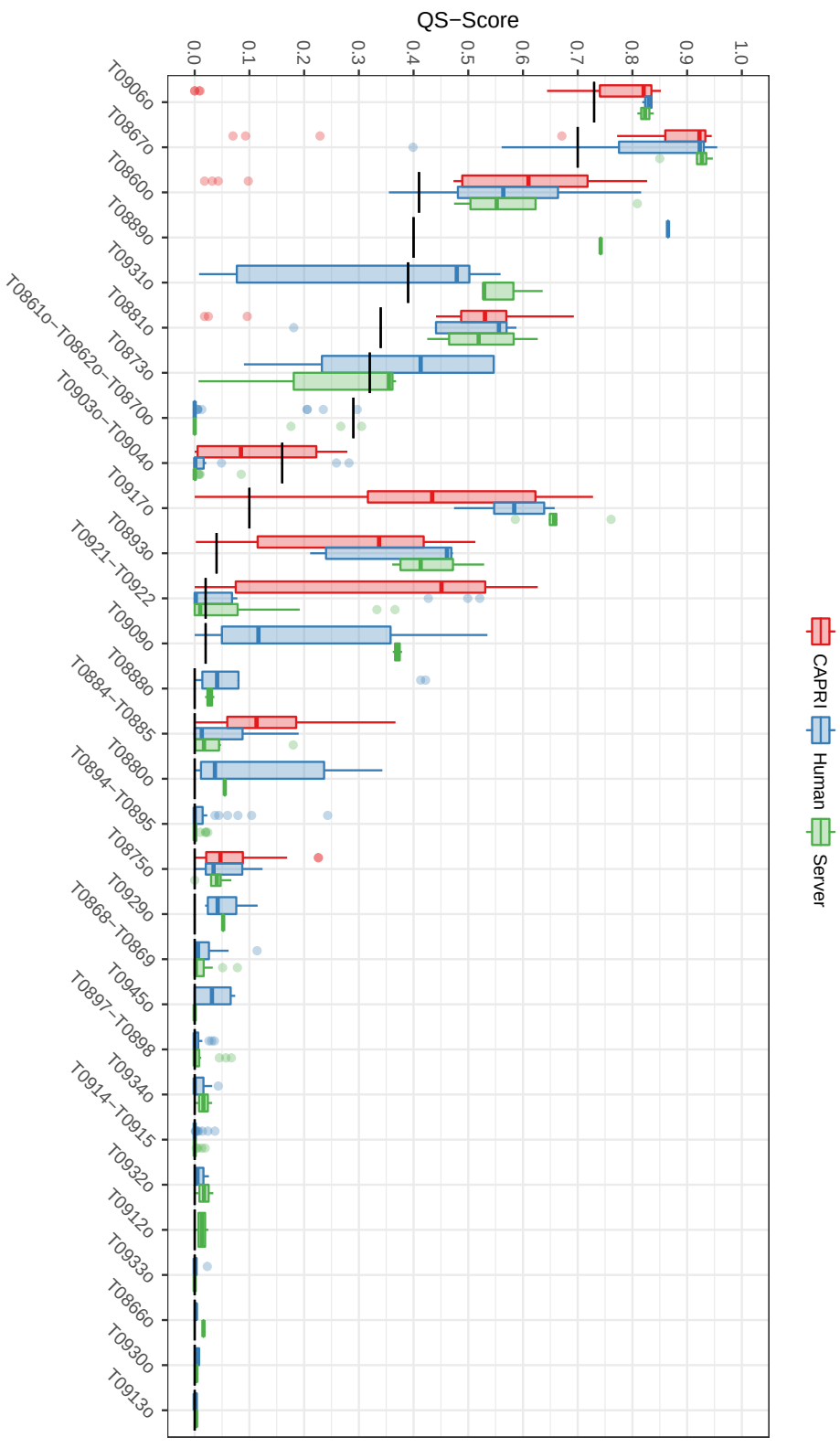


Figure 5.9: Baseline performance analysis of CASP12 oligomeric predictions. Distribution of the score for the best model from each group, assessed with QS-score. Baseline performance for each target is shown as horizontal black lines. Predictions are colored by the type of group: CAPRI groups, CASP human groups or CASP server groups.

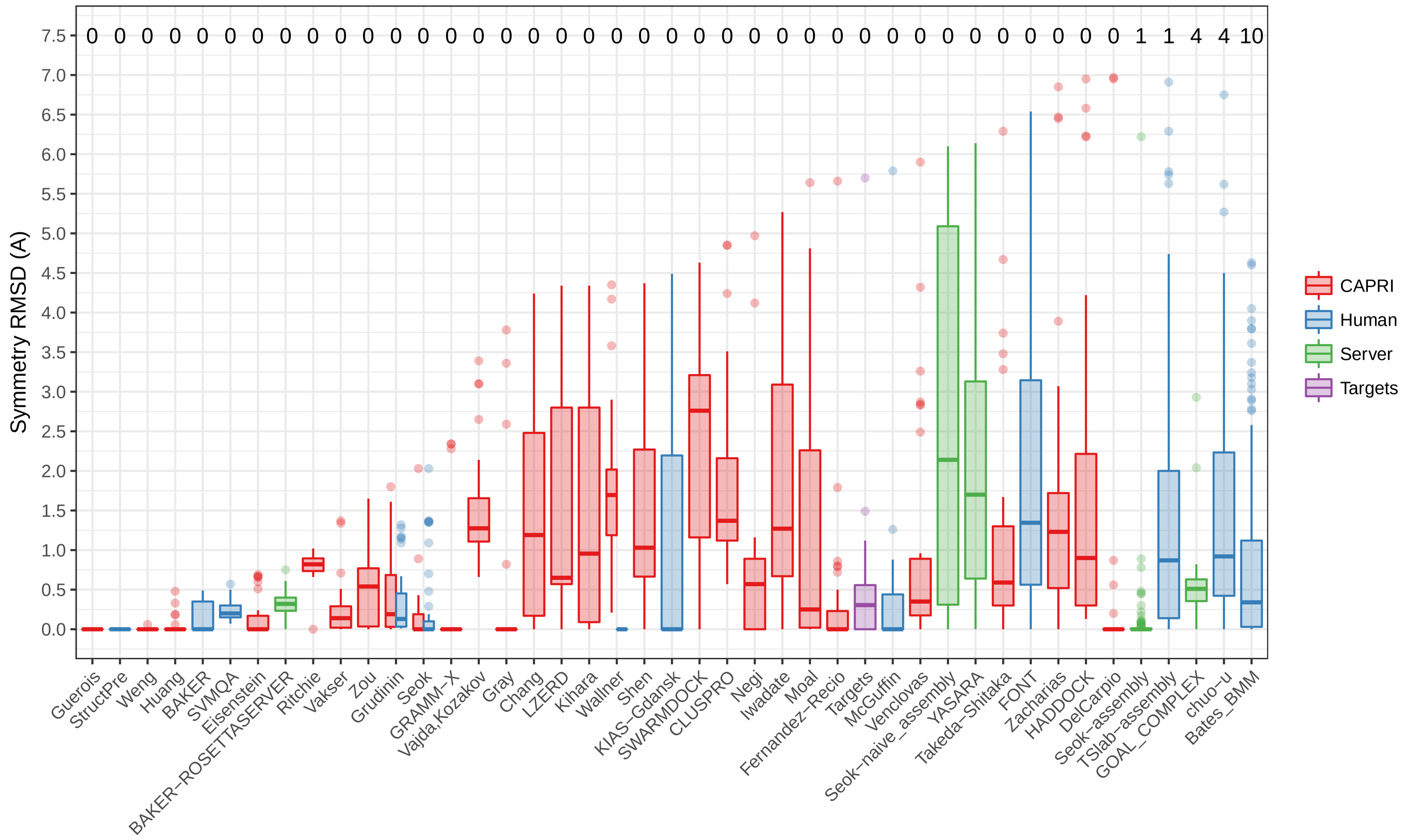


Figure 5.10: Symmetry deviation, measured as the RMSD, of models submitted by each group for each of the symmetric target assemblies. Numbers on top indicate the total number of asymmetric models submitted per group. Groups sorted in ascending order firstly by the number of asymmetric models submitted and secondly by the maximum allowed symmetry RMSD.

5.4.1 CckA histidine kinase

Target T0893o is a histidine kinase, known as CckA. Histidine kinases are dimeric bifunctional enzymes, mediating both phosphorylation and dephosphorylation of downstream targets. The function of the enzyme switches by domain rearrangements (Dubey et al. 2016).

The quaternary structure of histidine kinases is well known, with many experimental structures available. The catalytic core contains a dimerization domain, called dimerization histidine phosphotransfer (DHp) domain, that includes a conserved histidine position acting as a phosphate acceptor (autophosphorylation). The connectivity of the four helices of the DHp domain and its interactions with the catalytically active (CA) domain of the protein determine the mechanism of action of the enzyme (Bhate et al. 2015).

The four helical dimer of the DHp domain can be in two different connectivity types: helix B left of helix A, or helix B right of helix A. In addition, the conserved histidine phosphate acceptor (H322) can be in two different geometries: cis, if the ATP of the same subunit is close, or trans, if the ATP of the dimeric partner is close (Bhate et al. 2015).

The CckA has a DHp connectivity of *helix B right of helix A*, similar to CpxA, but an unusual *cis* histidine phosphate acceptor geometry. The authors of the experimental structure indicate that this might be due to distinct crystal contacts, since the CA domain is very flexible (Dubey et al. 2016), but we will also include this feature in our assessment. Finally, we will consider the *exposure of histidine H322* in both subunits (not part of the DHp interface) as a fundamental requirement for a functionally valid model.

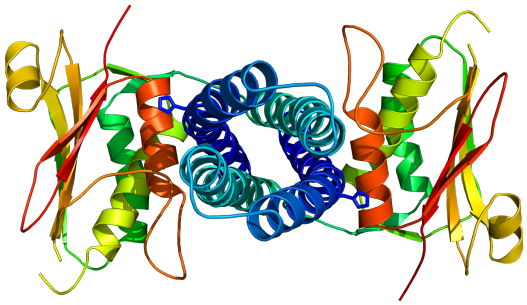
We manually looked at the best quaternary structure models of each group and annotated the important functional features we identified. The results showed some predictions reproducing correctly the DHp connectivity and others with the right histidine geometry, but no model reproduced both features from the experimental structure. The target structure and some representatives models are shown in figure 5.11.

5.4.2 STRA6 Receptor

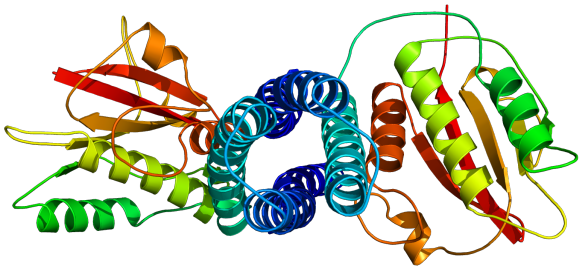
STRA6 is a dimeric integral membrane receptor for retinol uptake. It binds the retinol binding protein (RBP), which transports the highly hydrophobic retinol through the bloodstream and translocates the molecule into the lipid bilayer (Chen et al. 2016).

The STRA6 receptor generates a cleft in the dimeric interface of its assembly. Inside the cleft, the outer membrane layer is bended outwards, generating a space inside which the retinol molecule can be inserted (figure 5.12). In

a) Target T0893o
H322 exposed, B right of A, cis



b) YASARA
H322 exposed, B right of A, trans



c) BAKER-ROSETTASERVER
H322 exposed, B left of A, cis

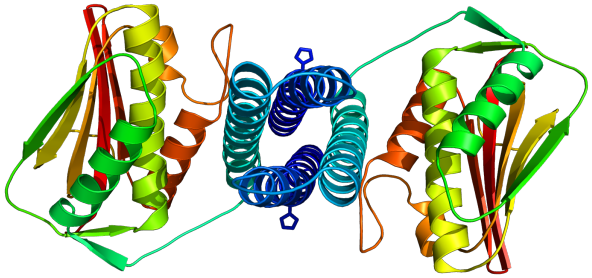


Figure 5.11: Quaternary structure of target T0893o and its representative models, with functional feature annotations. Structures colored from blue (N-terminal) to red (C-terminal). Helix A is in blue, helix B is in cyan.

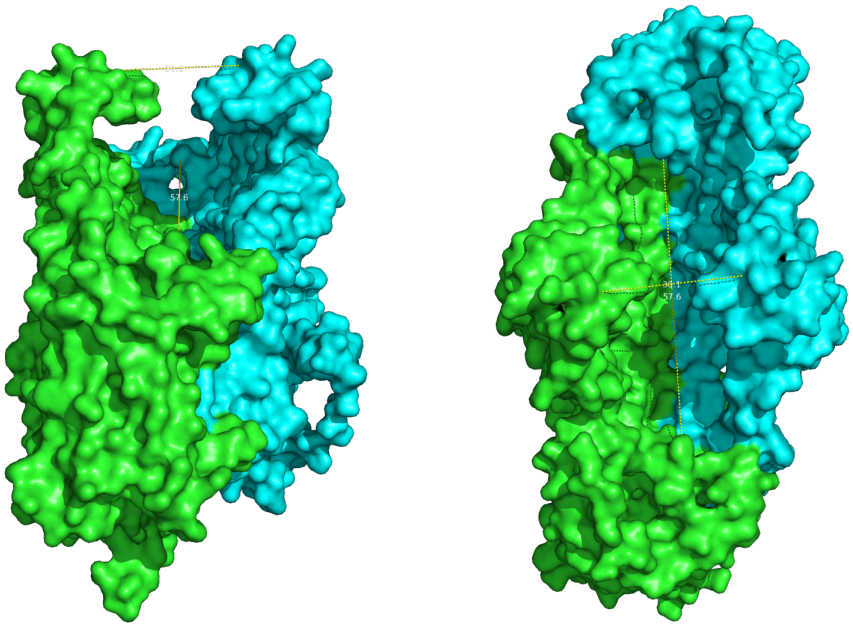


Figure 5.12: Side view (left) and top view (right) of the STRA6 receptor outer cleft. Two measurements between the same residues of each subunit are shown in yellow dashed lines.

addition, the coordination of residues from both subunits in the dimer is needed to form the RBP-binding motif (Chen et al. 2016).

The two most important functional features related to quaternary structure are the geometry of the outer cleft and the position of the RBP-binding motif residues. We have measured in the experimental structure the distances between the same residues in the two subunits of the dimer facing the inside of the outer cleft. For example, the phenilalanines of the top of the outer cleft (involved in the RBP-binding motif) are separated by around 30 Å. Good models should reproduce as close as possible the cleft geometry of the experimental structure.

STRA6 had no sequence similarity to any known membrane transporter, channel or receptor at the time of the CASP12 experiment. Thus, the prediction of its structure was very difficult and none of the groups produced acceptable tertiary structure models. The quaternary structure prediction was totally unsuccessful, with none of the groups predicting a single correct interface contact and with interface patch scores always below the acceptable baseline of 0.75. With the given quality of the models, a functional assessment could not be done for this particular target, in spite of its particularly interesting functional relevance.

5.5 Conclusions

As first assessors of the quaternary structure prediction category of CASP, we established a set of scores and analysis tools for the quality evaluation of protein assembly prediction models. We have also analyzed the overall performance of the predictions and the added value of the methods using a baseline performance to account for the prediction difficulty of each target assembly. Finally, we described the biological relevance of two oligomeric targets and evaluated the functional relevance of the models, where possible.

The two scores we used for model quality evaluation captured all the spectrum of prediction accuracy, thanks to their different levels of detail. They also correlated with our visual assessment of the models. The overall performance of the methods was consistently higher than the baseline performance, with a few outstanding performances that were discussed in the CASP12 final meeting. Although there is still room for improvement, the increasing participation and the discussion of promising new ideas and approaches to the problem are good signs.

We hope that CASP continues to encourage the prediction of protein assemblies and that our work in the model evaluation pipeline can be used by the future assessor teams.

Part II

Conformational dynamics of integrin α-I domains

Chapter 6

Introduction

This chapter is an introduction to the basics of integrin biology, the structure of integrins and the use of molecular dynamics simulations to study conformational dynamics of proteins.

6.1 Integrins

Integrins are integral transmembrane receptors involved in cell signaling. They mediate interactions of cells with the extracellular matrix and other cells. Some of their ligands are fibronectin, collagen and laminin (Humphries 2000). They signal across the membrane in both directions, using long-range allosteric conformational changes (Hynes 2002), and trigger intracellular pathways, allowing to respond quickly to changes at the cell surface. One example of their function is platelet signaling to initiate coagulation.

6.1.1 Integrin structure

Integrins are obligate heterodimers, composed of subunits α and β. They consist of a head piece formed by the head domains of both subunits, from which two legs emerge and end in a single transmembrane helix and a short cytoplasmic tail (figure 6.1). The integrin head piece comprises a seven bladed propeller, the α-P domain, in close interaction with the β-I domain (Gullberg 2014).

There are 18 types of α subunits and 8 types of β subunits. For each type of α subunit, there can be one or multiple possible types of β subunit partners, and vice versa. For example, the lymphocyte function-associated antigen 1 (LFA-1) is an integrin formed by the α_L and the β_2 subunits, while the very late antigen 1 (VLA-1) integrin is formed by the α_1 and β_1 subunits. The α_1 subunit always oligomerizes with the β_1 type, while the partner of α_L subunit is always β_2 (Srichai and Zent 2010).

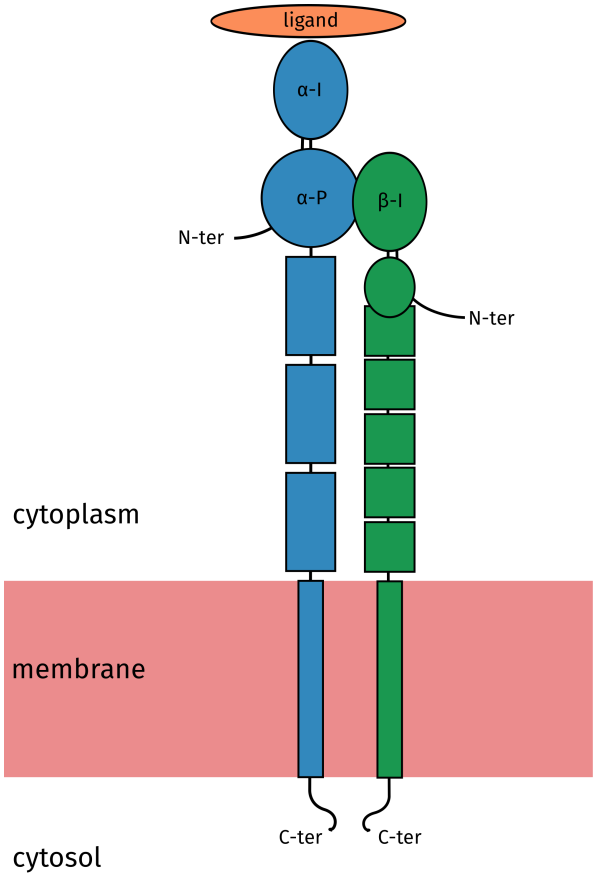


Figure 6.1: Schematic representation of an integrin structure. The two subunits, α (blue) and β (green), with different domain architecture, form a heterodimer with close interaction at the head piece.

6.1.2 The α-I domain

In humans, 9 out of the 18 integrin α subunit types, including α_1 and α_L, contain an additional domain, called α-I or αA domain, inserted on top of the α-P propeller domain (figure 6.1). The α-I domain plays a central role in ligand binding and specificity.

The domain is categorized as a member of the von Willebrand Factor (vWF) A domain superfamily. It contains a central parallel β-sheet of 6 strands surrounded on both sides by 7 amphipathic α-helices. On top of the central β-sheet there is a ligand binding site, called MIDAS (Metal Ion-Dependent Adhesion Site), where a metal ion is coordinated by three residues coming from different loops and three water molecules. (Gullberg 2014).

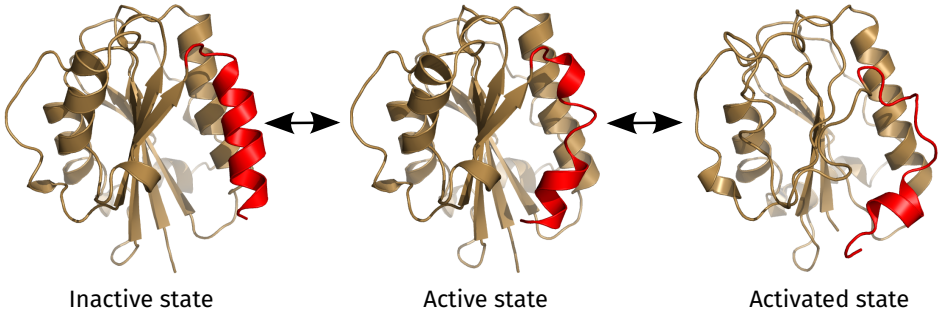


Figure 6.2: The three relevant conformational states of the C-terminal helix of the α_L-I domain described in the literature. Alternative names for the states in the literature refer to their ligand binding affinity, from left to right: allosterically inhibited (AI), low affinity (LA) and intermediate affinity (IA). Also from left to right, PDB codes: 1ZOO, 1ZON and 1MQ8.

6.1.3 Allosteric modulation of the α-I domain

It has been described in the literature that ligand binding alters the conformation of α_1-I, α_2-I and α_L-I domains in the same way. A switch in the ion coordination of the MIDAS upon ligand binding causes a downward movement of the C-terminal helix (helix-7) of about 10 Å, which is responsible for transducting the signal. The conformational change corresponds to the transition between the active and activated states, shown in figure 6.2 (Gullberg 2014).

For the α_L-I domain, an allosteric inhibition mechanism that involves a second conformational change of the C-terminal helix was discovered (Kallen et al. 1999). A small molecule, lovastatin, can be inserted inside the domain, between the central β-sheet and the C-terminal helix, stabilizing the inactive state (figure 6.3). The inhibitor binding involves, however, an opening of the pocket, which is closed in the active and activated states. The current biological model is that the activated state of the α_L-I domain can only be reached from the active state of the domain, so the inhibitor effectively disables the signal transduction even if the natural ligand binds (figure 6.2).

Since the discovery of the allosteric pocket, more inhibitors for the α_L-I domain have been found. However, screenings for inhibitors for the other types of α subunits have been unsuccessful. A scientific question derived from these observations is whether the allosteric modulation mechanism present in the α_L-I domain is universal among the other types of α subunits.

6.2 Molecular dynamics simulations

Some molecular motions of protein structures happen at short time-scales and small sizes (figure 6.4). These events are difficult to study experimen-

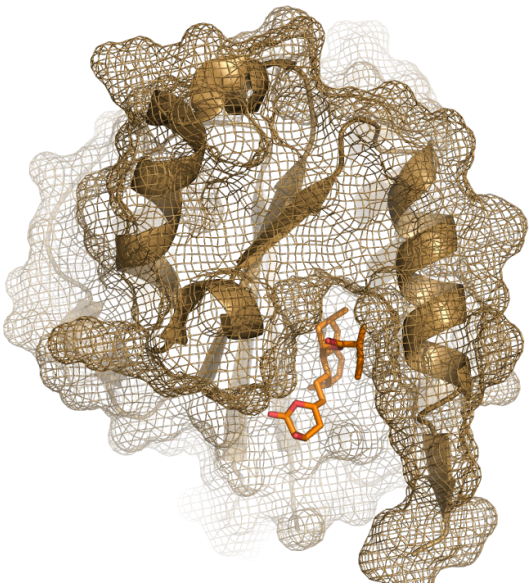


Figure 6.3: Allosteric inhibitor (lovastatin) bound to the α_L-I domain (1CQP).

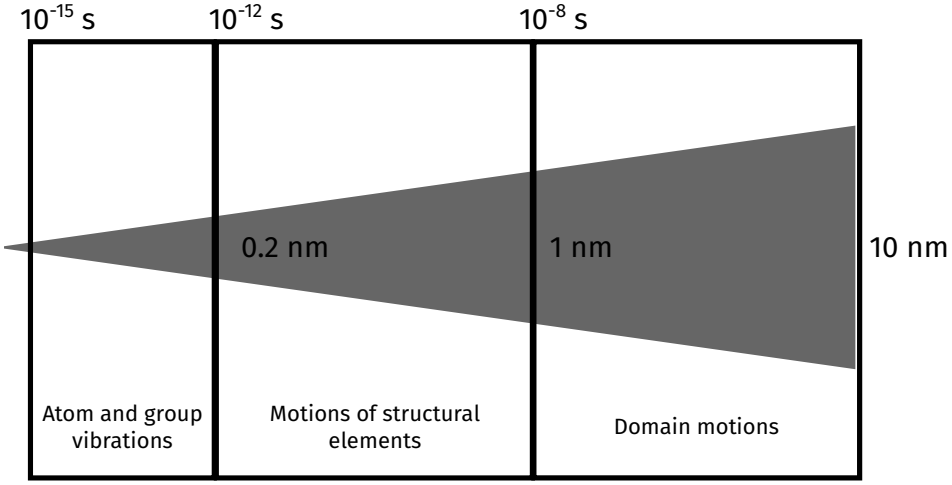


Figure 6.4: Time and spatial scales of protein molecular motions. The conformational change involved in the integrin α_L-I domain corresponds to the motion of a structural element (middle box).

tally, due to the time and spatial resolutions of the measurements. Computer simulations can offer the level of molecular detail needed in studying these types of protein dynamics. However, abundant and sophisticated computational resources are usually needed, especially to reach significant statistical sampling.

Molecular dynamics simulation is a computational method for studying the

physical behavior of a molecular system. The atomic interactions and movements are computed for a limited period of time, resulting in a dynamical evolution of the system. In the most common version, the simulation is performed using Newton's equations of motion for a system of interacting particles, where potential energies are defined using molecular mechanics force fields (Karplus and McCammon 2002).

Molecular dynamics simulations have been extensively applied to the modeling of biomolecules, especially proteins. In the present study, molecular dynamics simulations will be used to characterize the conformational states of integrin α-I domains and their transitions, with special focus on the C-terminal helix.

6.2.1 Integrin simulation precedents

Studying integrin structures using molecular dynamics simulations has been done before. There are two relevant articles worth mentioning in this introduction, because their focus is also the conformational dynamics of the α-I domain.

Jin, Andricioaei, and Springer 2004 used short molecular dynamics simulations of the α-I domain with a pull spring at the C-terminus of the helix-7 to study the activation mechanism, the transition between the active and activated conformations. They compared the α_L, α_M, α_1 and α_2 subunit types and found some differences in the dynamics, which involved an intermediate state in subunits α_L and α_M, but not in α_1 or α_2. They justify these differences in the dynamics with a Phe to Glu substitution at the top of the helix.

Kukic et al. 2015 used replica-averaged metadynamics (RAM) simulations with NMR restraints of the α_L-I domain to quantify the relative population frequencies of the three known conformational states of the helix-7 (figure 6.2). However, when they remove the restraints from the simulations, they do not observe any transition events between the states, indicating that either the time scales of these events are much larger than the simulations or that they do not occur spontaneously (i.e, ligand binding is required).

These two studies use additional (unphysical) forces to induce the conformational changes of the α-I domain. Because we would like to observe spontaneous transitions between the conformational states, we use unrestrained molecular dynamics.

6.2.2 Trajectory analysis

Monitoring

In order to monitor the evolution of the system along the simulation trajectory, properties can be followed as a function of time.

The root mean square deviation ($RMSD$) between two sets of points is a measure of the average distance (d) between corresponding positions. The RMSD of a set of N pairs of corresponding positions between two point sets is defined as:

$$RMSD = \sqrt{\frac{1}{N}\sum_{i=1}^{N} d_i^2} \tag{6.1}$$

In order to compare two protein structures using the RMSD, the set of representative atoms has to be chosen. This is usually the C_α atom of each amino acid residue, but it can also include C_β atoms or other heavy atoms. Then, a superposition of the two structures need to be performed, using the chosen representative atoms, in order to account for shifts and rotations in the reference frame. Once the structures are superposed, the distances between the corresponding atomic positions can be computed to obtain the final RMSD. One particularity of the superposition dependent measures, like the RMSD, is that the superposition and the measure can be done on different sets of atoms. In our study case, the α-I domain, the changes in the C-terminal helix are both internal and rigid-body movements with respect to the rest of the protein, so we can perform the superposition on the protein backbone excluding the helix and compute the RMSD only for the helix backbone.

The solvent-accessible surface (SAS), or accessible surface area (ASA), is the area of a molecule that is accessible to the solvent. SAS is calculated using a sphere of a particular radius to probe the surface of the molecule. This measure is useful to analyze changes in the orientation of side chains or the opening of pockets, accessible to water or potential ligands.

Interatomic distances can also be defined to describe the relative movements of the parts of the structure like, for example, salt bridges or specific conformational changes.

SAPPHIRE plot

The SAPPHIRE (States And Pathways Projected with HIgh REsolution) plot of a molecular dynamics trajectory provides a picture of the kinetically distinct states (termed basins) of the system (Blöchliger, Vitalis, and Caflisch 2013).

The method relies on a the choice of a metric, a definition of a pairwise distance between snapshots of the system, to reorder the original trajectory in a new progress index. The progress index is plotted against the number of transitions between two partitions, defined by a cut function, of the entire snapshot collection. The idea is that transitions within a basin are more frequent than transitions between basins, so that the final curve will resemble the free energy profile of the system.

The SAPPHIRE plot has been successfully applied to analyze molecular dynamics trajectories of biomolecules (Blöchliger, Vitalis, and Caflisch 2014). For the purposes of this study, the SAPPHIRE plot is a key tool for its power to summarize long molecular dynamics trajectories into a small set of representative states of the system, minimizing the risk of missing relevant information.

Chapter 7

Molecular dynamics simulations of integrin α-I domains

7.1 Goal of the simulations

The modulation of LFA-1 signaling though the allosteric pocket in the C-terminal helix of the α-I domain, as it was described in section 6.1, is well studied and widely accepted. However, attempts to find small molecule ligands that can allosterically modulate the signaling of other types of integrins have been unsuccessful. The main goal of this study is to evaluate whether the allosteric mechanism of signal modulation of LFA-1 is also possible in other types of integrins. In particular, we will study the I-domain of the α_1 subunit and compare its dynamics with the I-domain of α_L. More specifically, we would like to know if the allosteric pocket opening, which corresponds to the transition from the active to the inactive states, can happen spontaneously without the presence of the inhibitor.

7.2 System preparation

The preparation of the system is the most important and time consuming part of molecular dynamics simulations. The setup of a protein simulation consists in the following steps: choosing a starting structure, usually experimentally determined, and processing it; recreating the physiological environment by adding a solvent and other relevant molecules, usually water and ions; parameterizing the system with a force field; relaxing the system with an energy minimization scheme; and equilibrating the system at the physiological conditions of temperature and pressure.

There is software available to help and automatize the preparation of molecular dynamics systems. For the preparation of the integrin α-I domain sys-

Table 7.1: Summary of the properties of the two solvated α-I domain systems.

System	alpha-1	alpha-L
Structure	1QC5	1ZOO
Box volume (nm3)	499	486
Number of atoms	50039	48278
Water molecules	15640	15077
Ions	38 Na + 36 Cl	38 Na + 36 Cl

tem, we used the tools provided with the GROMACS package (Berendsen, Spoel, and Drunen 1995).

Starting structure

From all the integrin α-I domain structures, we decided to use the structure 1QC5 for the α_1-I domain and 1ZOO for the α_L-I domain with the open helix conformation stabilized by a crystal contact at the C-terminal. Both are APO structures with a Mg^{2+} ion at the MIDAS site. The chain A of the asymmetric unit was selected for both structures.

Since there were no missing residues in the structures, modeling of missing protein segments was not needed. The processing steps required were the addition of hydrogen atoms to all residues and capping groups to the N and C terminal residues of the structure.

Both structures were processed using GROMACS *pdb2gmx* command. An acetyl group was selected for the N-terminal and an amine group for the C-terminal capping. Hydrogens were added with the default GROMACS settings, which assigns negative charges to acidic residues and positive charges to basic residues in order to resemble the neutral pH conditions. Histidines are kept neutral and protonation positions are assigned based on their chemical environment.

Solvation

The α-I is a cytoplasmic domain of the integrin structure. Since it is a soluble globular domain, the physiological environment consists mainly of water molecules and ions.

A cubic solvation box at a minimum distance to the protein of 1.2 nm was created, with the *spc216* starting configuration and the modified TIP3P water model to be used with Charmm. Next, chloride and sodium ions were added to account for the ionic strength and neutralize the charges of the protein. A summary of the properties of the two systems after the solvation step is shown in table 7.1.

Parameterization

For the system parameterization, we used the *CHARMM36* force field (Huang et al. 2016).

The α-I domain contains the MIDAS site with an ion in a octahedral coordination, carrying out a key function in the activity of the domain. Force fields are not able to parameterize correctly this type of interactions, so the original coordination would be lost in the simulation if additional restrains are not applied. Therefore, we applied position restrains of 1000 $kJ\ mol^{-1}nm^{-2}$ for the entire simulation to the 7 atoms involved in the octahedral coordination: 3 water oxygens, the Ser-42 and Ser-44 side-chain oxygens, one of the Asp-143 side chain oxygens and the magnesium ion.

Energy minimization

The initial state of the system has not been optimized for the force field used in the simulation. Therefore, it is likely that some parts of the system are in a high energy state, as defined by the physical parameters of the force field. If the simulation were started immediately, those parts of the system with high energy would create very high and unphysical forces that could potentially break the simulation.

The energy minimization is a procedure to relax parts of the system with a very high energy. It can be thought as an initial fitting of the system to the force field. The procedure iteratively reduces the forces acting on the atoms of the system (equivalent to reducing the energy), until all the forces are below a maximum force threshold.

We used the steepest descent minimization. Less than 200 steps for both the α_1-I and the α_L-I system were required to achieve a maximum force below the threshold, indicating that the initial systems had no problematic regions from a physical point of view.

Equilibration

Equilibration consists in applying position restraints to the protein heavy atoms, while allowing the solvent and other molecules to move freely so that they can adapt to the protein. We used two steps of equilibration: a first one at fixed volume (NVT) and a second one at fixed pressure (NPT) conditions.

For the NVT equilibration, we used the V-rescale (a modified Berendsen thermostat) temperature coupling scheme with two coupling groups, protein and non-protein, with a time constant of 0.1 ps and a reference temperature of 300 K. For the NPT equilibration, we used the Parrinello-Rahman pressure coupling scheme, with a time constant of 2 ps and a reference pressure of 1

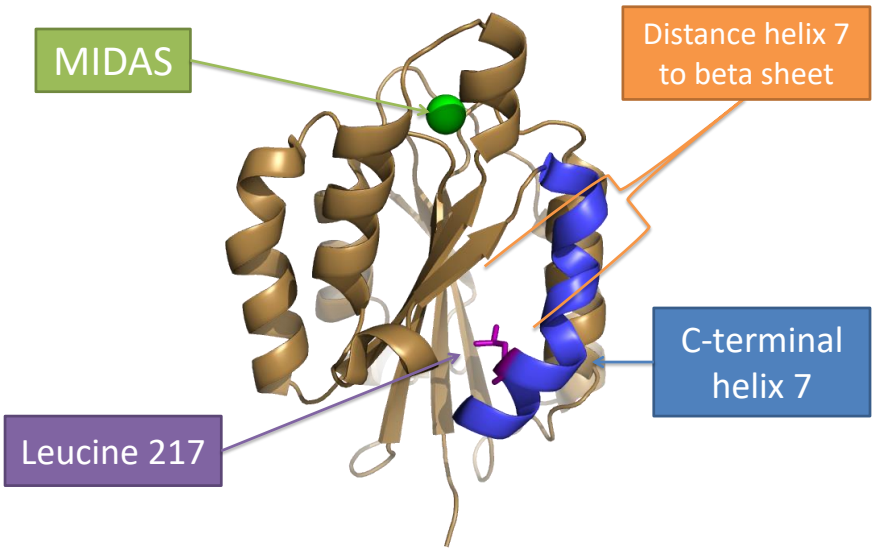


Figure 7.1: Properties of the α_1-I domain followed in the simulation: MIDAS site, Leucine at position 217, C-terminal helix-7 and distance between the helix-7 and the central β-sheet.

bar, in addition to the same temperature coupling scheme as the NVT equilibration. Both equilibration steps converged fast, on the order of 10 ps, to the desired reference temperature and pressure.

Production simulations

The production simulations were carried out at NPT conditions, with the same parameters described for the NPT equilibration. For the long range electrostatics, we used the Particle Mesh Ewald (PME) and a cut-off for the short-range Van der Waals and electrostatic interactions of 1 nm.

7.3 Simulations of the conformational dynamics

7.3.1 Simulations of α_1-I domain

The goal of the following simulations is to observe a spontaneous opening event of the binding pocket between the C-terminal helix and the central β-sheet of the domain.

Long simulation

The first simulation that was performed was a sequential and long simulation, starting from the energy minimized and equilibrated system. The evolution of some system properties is shown in figure 7.2. The RMSD of the protein backbone stays more or less constant throughout the simulation, while the RMSD of the C-terminal helix increases for about 300 ns and goes back to the initial conformation. The Leucine 217 is at the middle of the C-terminal helix and facing inside the hydrophobic pocket. As it can be seen in the plot, neither the SAS of the residue nor the distance of the C-terminal helix to the central β-sheet increase, indicating that the allosteric pocket did not spontaneously open at any point of the simulation.

Simulations from snapshots

Two snapshots of the long trajectory with high RMSD to the starting structure were used as starting conformations of two other simulations. The idea was to increase the sampling of alternative conformations reached in the previous simulation. To select the snapshots, a preliminary SAPPHIRE plot of the long trajectory was generated, and representative snapshot configurations of the structure were selected. The SAPPHIRE plot contained two basins, from which the representative snapshots were located at the original trajectory times of 210.6 and 292.4 ns.

The evolution of the properties of both simulations is also shown in figure 7.2. As before, the C-terminal helix goes back to the starting conformation at the end of the simulation, and there are no signs of the opening of the pocket.

Distributed parallel simulations

If we assume a single free-energy barrier for the transition between the closed and open helix-7 conformations of the α-I domain, we can compute the probability of observing an event in our simulation (Paci et al. 2003). The transition event t is exponentially distributed, with the distribution parameter being the half-life ($t_{1/2}$), λ or τ:

$$P(t) \sim exp(\lambda) \tag{7.1}$$

$$t_{1/2} = \frac{ln(2)}{\lambda} = \tau ln(2) \tag{7.2}$$

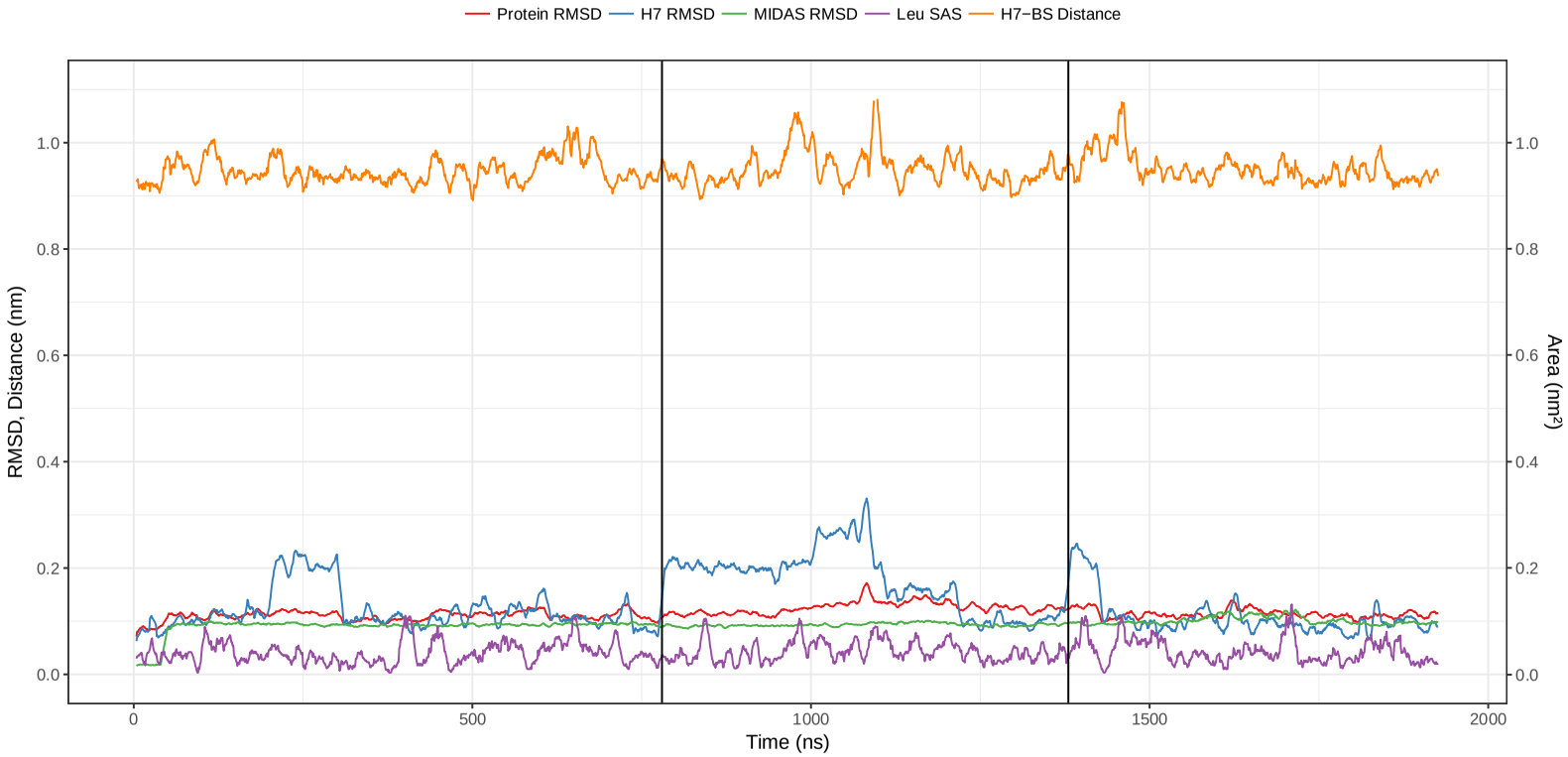


Figure 7.2: Trajectory of α_1-I domain (1QC5) starting from the energy minimized and equilibrated experimental structure. Black vertical lines indicated points were the simulation was restarted from a previous snapshot (from snapshots at time 292.4 ns (first) and 210.6 ns (second) of the simulation. Evolution of the RMSD of the backbone C_α atoms of the protein and the C-terminal helix (H7), and of the restrained atoms of the MIDAS site; the distance from the C-terminal helix (H7) to the central beta sheet (BS) of the domain; and the solvent accessible surface (SAS) of the Leucine 217. Trajectory saved every 100ps of simulation. An average sliding window of size 10 is applied to smooth the data points.

Table 7.2: Probability of observing at least one transition event in 40 parallel simulations for different lengths (columns) and estimated half-lifes (rows).

	10 ns	50 ns	100 ns	200 ns
1 us	0.24	0.75	0.94	1.00
10 us	0.03	0.13	0.24	0.43
100 us	0.00	0.01	0.03	0.05

The probability of a transition event t being observed in a single simulation of time x is computed as:

$$P(t \leq x) = 1 - e^{-\lambda x} \tag{7.3}$$

If we run N independent simulations in parallel, starting from the same structure but generating random starting velocities, the probability of observing at least one transition event in any of the simulations is:

$$P(t \leq x|N) = 1 - (1 - P(t \leq x))^N \tag{7.4}$$

Shown in table 7.2 are the probabilities for different estimates of simulation length and half-life. The probability of observing a transition if the half-life of the event is in the order of magnitude of 100 μs is very low. However, if it were in the order of magnitude of 10 μs or lower, the likelihood of observing such an event with the distributed approach is reasonable.

We ran 32 simulations of 200ns and 2 simulations of 100ns each, for a total of 6.6 μs of molecular dynamics simulation. Visual inspection of the trajectories with the help of similar plots to figure 7.2 indicated that no opening event of the allosteric pocket at the C-terminal helix had occurred. In spite of that, the simulations can be used, together with the other trajectories, to generate a SAPPHIRE plot and study other properties of the conformational dynamics of the domain.

SAPPHIRE plot analysis

All the trajectories of the α_1-I domain, summing up to a total of 8.5 μs of simulation time, were subsampled to a snapshot every 100 ps and concatenated to create the SAPPHIRE plot shown in figure 7.3. The pairwise distance measure selected to compare snapshots is the RMSD of the C_α atoms of the C-terminal helix (residue indexes 207 to 221), after superposition on the C_α atoms of the rest of the protein structure. For the clustering step of the progress index construction we used a tree with eight levels, a threshold radius of clusters of 1 Å and a coarsest level in the tree of 2 Å.

The SAPPHIRE plot contains three metastable states. The first and biggest one corresponds to the starting configuration of the C-terminal helix, consisting of the closed pocket with a kink in the middle of the helix. The second basin in the middle of the plot is very similar to the first one, but the kink in the helix is replaced by a wider helix turn, as it can be appreciated in the DSSP annotation changes. The third basin at the right of the plot maintains the helix conformation at the C-terminal end of the helix, but the top part of the helix unfolds completely. The first and second basins are sampled multiple times throughout the simulations,while the third one is only sampled once (the recurrence observed is due to the restart of the trajectory).

7.3.2 Simulation of the α_L-I domain

The goal of this simulation is to follow the dynamics of the open allosteric pocket. We start with the structure of the α_L-I with the C-terminal helix in the open conformation (PDB code 1zoo).

The evolution of the properties of the system is shown in figure 7.4. We can observe how the allosteric pocket closes after 100ns of simulation: the RMSD to the closed conformation decreases to less than 2 Å, the SAS of the Leucine facing inside the pocket falls below the 20 Å^2 and the C-terminal helix gets closer to the central β-sheet. However, after 300ns, the C-terminal helix adopts a conformation different to the open or closed pocket ones, since the RMSD to both of the conformations is high.

7.4 Conclusions

We have not observed a spontaneous transition of the C-terminal helix of the α_1-I domain from the closed to the open pocket conformations in a cumulative sampling of nearly 10 μs of molecular dynamics simulation. Therefore, the question that initially motivated this study remains unanswered: we could not determine whether the pocket for allosteric signal modulation exists in other types of integrins other than the LFA-1. However, we have observed two other metastable states of the C-terminal helix conformation, one robustly sampled and another sampled a single time, that could be relevant for further integrin studies.

Two possible reasons could be behind our inability to observe the opening of the C-terminal helix: sampling and ligand induced fit. Sampling enough of the conformational space of proteins is challenging because it requires a lot of computational time. As presented in the previous section, the likelihood of observing slow molecular motions is very small in the time scales of our simulations. It can be that the length of the individual simulations is not enough for the time scale of the pocket opening event. The other possible reason could be that the ligand that binds in the pocket can play

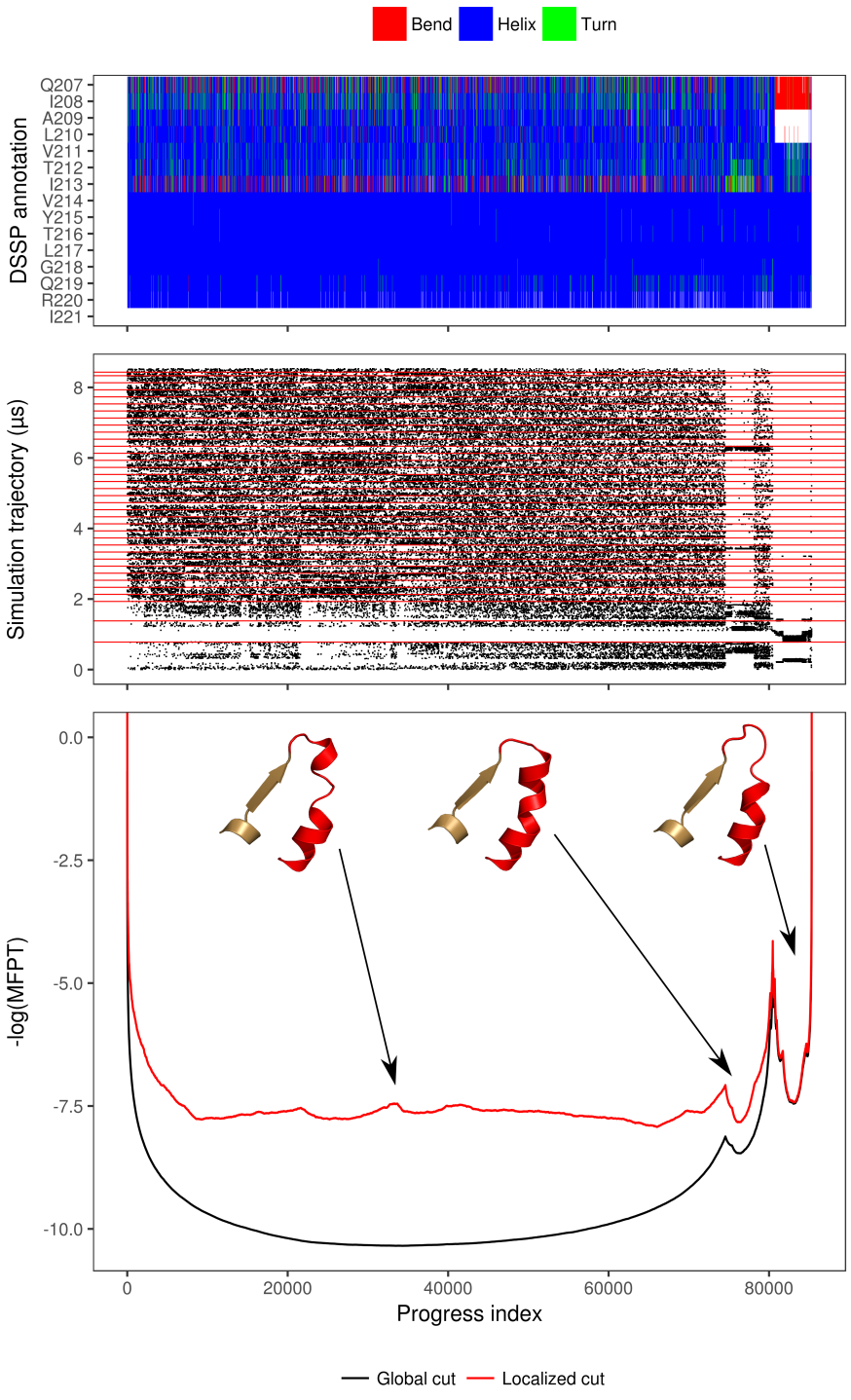


Figure 7.3: SAPPHIRE plot of all α_1-I domain (1QC5) simulations. Secondary structure (top), simulation trace (middle) and SAPPHIRE plot (bottom). Trajectory saved every 100ps. Red horizontal lines in the simulation trace indicate break points, where trajectories are concatenated, for a final simulation time of 8.5 μs. The cartoon representations of each basin are from the snapshots at the minimum value of the SAPPHIRE plot within the basin.

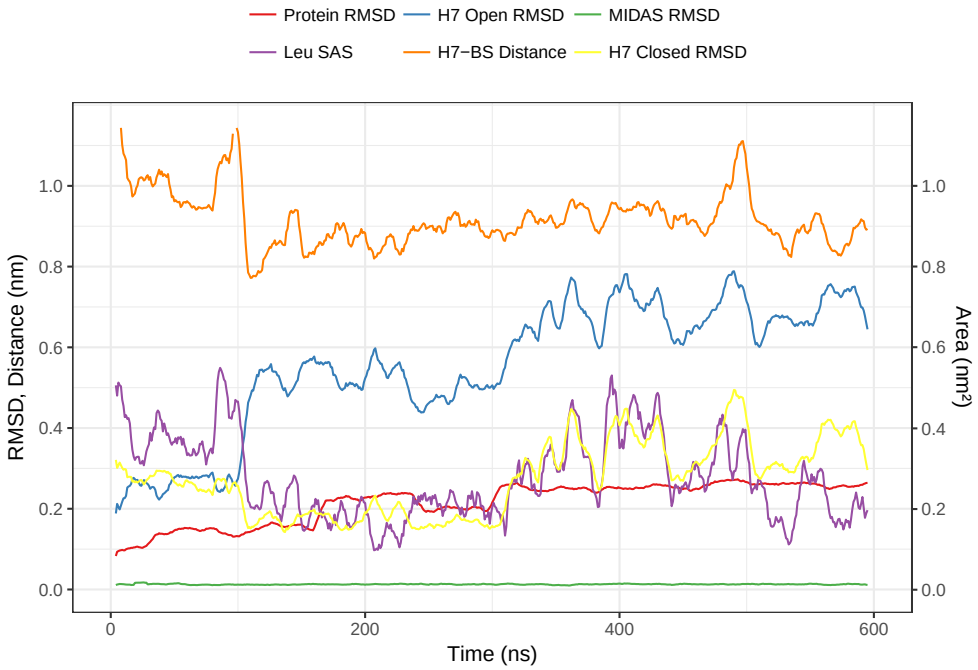


Figure 7.4: Trajectory of α_L-I domain (1zoo) starting from the experimental structure, with the C-terminal helix in the open conformation. Evolution of the RMSD of the backbone C_α atoms of the protein and the C-terminal helix (H7), compared to the open (1zoo) and the closed (1zon) C-terminal helix (H7), and the restrained atoms of the MIDAS site; the distance from the C-terminal helix (H7) to the central beta sheet (BS) of the domain; and the solvent accessible surface (SAS) of the Leucine 302. Trajectory saved every 100ps of simulation. An average sliding window of size 10 is applied to smooth the data points.

a central role in inducing the pocket opening. This would challenge our assumption that spontaneous transitions between the conformations happen, but it would be consistent with previous molecular dynamics studies that could not see the transition without the use of restraints and our observation of the rapid pocket closing of the α_L subunit in the absence of the ligand. Therefore, an alternative hypothesis could be that, in solution, the pocket opens only in the presence of the ligand.

In the future, an enhanced sampling technique could be used to sample a bigger fraction of the conformational space of the C-terminal helix. One option could be the Progress Index-Guided Sampling (PIGS) (Bacci, Vitalis, and Caflisch 2015), which is an automated method to do what we manually recreated by restarting two simulations from two high RMSD snapshots of the trajectory. Another option could be umbrella sampling with an initial targeted molecular dynamics simulation to force the opening of the pocket. Finally, introducing small hydrophobic molecules near or inside the pocket could be tried to recreate the exact conditions of the ligand binding at the

pocket and allosteric inhibition.

Appendix A

Publications

A.1 Assessment of protein assembly prediction in CASP12

The results obtained during the assessment of quaternary structure prediction models in the CASP12 experiment were presented at the CASP12 meeting in Gaeta, Italy. A publication will be submitted as part of a special issue on CASP12 in Proteins. This will cover the methods and results presented in chapters 3, 4 and 5.

A.2 Finding valid quaternary assemblies in protein crystals

The new features and improvements of the EPPIC software and web application, including the part described in section 3.2, will be published following the official release, scheduled for the end of March of 2017. The publication is currently in preparation.

A.3 BioJava 5

The contributions to the open-source BioJava library since version 4, including the algorithm described in chapter 4 together with the contributions made by other developers, will be presented in a common publication following the release of the version 5 of the library, scheduled for the first half of the year 2017.

A.4 Exploring internal symmetry and structural repeats with CE-Symm

Although this work has not been presented in this thesis, since it was carried out during an earlier internship also at Dr. Capitani's lab, we are preparing a new research article of the CE-Symm tool (Myers-Turnbull et al. 2014), used to analyze internal symmetry in protein structures. A modification of the algorithm was used to analyze the symmetry deviations of the quaternary structure models, presented in section 5.3.

List of Figures

List of Tables

Acknowledgements

This thesis could not have been possible without the help and contributions of many people, so I would like to acknowledge everyone that was involved in it.

First of all, I would like to thank Guido for the opportunity to do my thesis with him. You have been a fantastic example for me and I can only wish you all the best for the future.

I would also like to thank Spencer for being such a great supervisor and spending time to explain and discuss many interesting science with me. It has been a pleasure to work with you and I hope we can meet again in the future.

I am very grateful to Amedeo for mentoring me during my entire Master degree and for proposing me an additional research topic for the thesis. I appreciate your scientific input and I have enjoyed a lot the time in your group.

I am also very thankful to Cassiano for spending endless time reviewing and supervising my work, always with constructive comments and a positive attitude. I have learned a lot from you and I wish you all the best for your project.

I have had a memorable time at the Paul Scherrer Insitute, so I would like to thank all my colleagues and friends. Particularly, I would like to thank Prof. Dr. Michel Steinmetz for his advice and support in the most difficult situations.

I would like to thank all the colleagues at the Caflisch group (University of Zürich) - you made it really easy for me to integrate in your group.

I would also like to thank the CASP organizers, other assessor teams and predictor groups, with special mention to Dr. Andriy Kryshtafovych and Prof. Dr. Torsten Schwede.

Many thanks to the colleagues at the RCSB PDB and BioJava developers: Dr. Jose Duarte, Dr. Andreas Prlic and Dr. Peter Rose.

Finally, special thanks to Dr. Claus Erhardt (Novartis) and Dr. Gabriele Weitz-Schmidt (AlloCyte) for the discussions about the integrin project.

And last but not least, I would like to thank my parents, Xavier and Guadalupe, and my sister, Carme, for always remembering me how good it feels to be back home.

Bibliography

Anfinsen, Christian B (1972). ``The Formation and Stabilization of Protein Structure''. In: *Biochemical Journal* 128.4, pp. 737–749.

Bacci, Marco, Andreas Vitalis, and Amedeo Caflisch (2015). ``A molecular simulation protocol to avoid sampling redundancy and discover new states''. In: *Biochimica et Biophysica Acta - General Subjects* 1850.5, pp. 889–902.

Baskaran, Kumaran et al. (2014). ``Open Access A PDB-wide , evolution-based assessment of protein – protein interfaces''. In: *BMC Structural Biology* 14, pp. 1–11.

Berendsen, H. J C, D. van der Spoel, and R. van Drunen (1995). ``GROMACS: A message-passing parallel molecular dynamics implementation''. In: *Computer Physics Communications* 91.1-3, pp. 43–56.

Berman, Helen M et al. (2000). ``The protein data bank.'' In: *Nucleic acids research* 28.1, pp. 235–242.

Bertoni, Martino (submitted). ``When two is not enough: modeling protein quaternary structure of homo- and hetero-oligomers beyond binary interactions''. In:

Bhate, Manasi P et al. (2015). ``Signal Transduction in Histidine Kinases: Insights from New Structures''. In: *Structure* 23.6, pp. 981–994.

Biasini, Marco et al. (2014). ``SWISS-MODEL: Modelling protein tertiary and quaternary structure using evolutionary information''. In: *Nucleic Acids Research* 42.W1.

Blöchliger, Nicolas, Andreas Vitalis, and Amedeo Caflisch (2013). ``A scalable algorithm to order and annotate continuous observations reveals the metastable states visited by dynamical systems''. In: *Computer Physics Communications* 184.11, pp. 2446–2453.

— (2014). ``High-resolution visualisation of the states and pathways sampled in molecular dynamics simulations.'' In: *Scientific reports* 4, p. 6264.

Bourne, Philip E and Ilya N Shindyalov (2003). ``Structure comparison and alignment.'' In: *Methods of Biochemical Analysis* 44, pp. 321–337.

Brier, Glen W (1950). ``Verification of forecasts expersses in terms of probaility.'' In: *Monthly Weather Review* 78.1, pp. 1–3.

Capitani, Guido et al. (2015). ``Understanding the fabric of protein crystals: Computational classification of biological interfaces and crystal contacts''. In: *Bioinformatics* 32.4, pp. 481–489.

Chen, Yunting et al. (2016). ``Structure of the STRA6 receptor for retinol uptake''. In: *Science* 353.6302, aad8266–aad8266.

Cheng, Hua et al. (2014). ``ECOD: An Evolutionary Classification of Protein Domains''. In: *PLoS Computational Biology* 10.12.

Clackson, Tim and James A. Wells (1995). ``A hot spot of binding energy in a hormone-receptor interface''. In: *Science* 267.5196, pp. 383–386.

Duarte, Jose M et al. (2012). ``Protein interface classification by evolutionary analysis.'' In: *BMC bioinformatics* 13.1, p. 334.

Dubey, Badri N et al. (2016). ``Cyclic di-GMP mediates a histidine kinase/phosphatase switch by noncovalent domain cross-linking''. In: *Science Advances* 2.9, e1600823–e1600823.

Fekete, Szabolcs et al. (2014). ``Theory and practice of size exclusion chromatography for the analysis of protein aggregates.'' In: *Journal of pharmaceutical and biomedical analysis* 101, pp. 161–73.

Finn, Robert D et al. (2016a). ``InterPro in 2017-beyond protein family and domain annotations.'' In: *Nucleic acids research* 45, gkw1107.

Finn, Robert D. et al. (2016b). ``The Pfam protein families database: Towards a more sustainable future''. In: *Nucleic Acids Research* 44.D1, pp. D279–D285.

Gullberg, Donald (2014). *I Domain Integrins*, p. 188.

Heiberger, J., Chambers A., and Freeny R. (1992). ``Statistical Models in S''. In: *... Experiments. Wadsworth & Brooks/Cole, ...* Ed. by J.M Chambers and T.J Hastie, pp. 34–47.

Huang, Jing et al. (2016). ``charmm36m: an improved force field for folded and intrinsically disordered proteins''. In: *Nature Methods* 14.

Humphries, M.J. (2000). ``Integrin Structure''. In: *Biochemical Society Transactions* 28.4, pp. 311–340.

Hynes, Richard O. (2002). *Integrins: Bidirectional, allosteric signaling machines*.

Jin, Moonsoo, Ioan Andricioaei, and Timothy A. Springer (2004). ``Conversion between three conformational states of integrin I domains with a C-terminal pull spring studied with molecular dynamics''. In: *Structure* 12.12, pp. 2137–2147.

Jonic, Slavica and Catherine Vénien-Bryan (2009). *Protein structure determination by electron cryo-microscopy*.

Kabsch, Wolfgang and Christian Sander (1983). ``Dictionary of protein secondary structure: Pattern recognition of hydrogen bonded and geometrical features.'' In: *Biopolymers* 22, pp. 2577–2637.

Kallen, J et al. (1999). ``Structural basis for LFA-1 inhibition upon lovastatin binding to the CD11a I-domain.'' In: *Journal of molecular biology* 292.1, pp. 1–9.

Karplus, Martin and Ja Andrew McCammon (2002). ``Molecular dynamics simulations of biomolecules.'' In: *Nature Structural Biology* 9.9, pp. 646–652.

Keskin, Ozlem, Nurcan Tuncbag, and Attila Gursoy (2016). *Predicting Protein-Protein Interactions from the Molecular to the Proteome Level*.

Krissinel, Evgeny and Kim Henrick (2007). ``Inference of Macromolecular Assemblies from Crystalline State''. In: *Journal of Molecular Biology* 372.3, pp. 774–797.

Kukic, Predrag et al. (2015). ``Structure and dynamics of the integrin LFA-1 I-domain in the inactive state underlie its inside-out/outside-in signaling and allosteric mechanisms''. In: *Structure* 23.4, pp. 745–753.

Lensink, Marc F. et al. (2016). ``Prediction of homoprotein and heteroprotein complexes by protein docking and template-based modeling: A CASP-CAPRI experiment''. In: *Proteins: Structure, Function and Bioinformatics*.

Levinthal, Cyrus (1969). ``How to fold graciously''. In: *Mössbauer Spectroscopy in Biological Systems Proceedings* 24.41, pp. 22–24.

Levy, Emmanuel D. and Sarah Teichmann (2013). ``Structural, evolutionary, and assembly principles of protein oligomerization''. In: *Progress in Molecular Biology and Translational Science* 117, pp. 25–51.

Madej, Thomas et al. (2014). ``MMDB and VAST+: Tracking structural similarities between macromolecular complexes''. In: *Nucleic Acids Research* 42.D1.

Mariani, Valerio et al. (2011). ``Assessment of template based protein structure predictions in CASP9''. In: *Proteins: Structure, Function and Bioinformatics* 79.SUPPL. 10, pp. 37–58.

Marks, Debora S. et al. (2011). ``Protein 3D structure computed from evolutionary sequence variation''. In: *PLoS ONE* 6.12.

Marti-Renom, M.A. et al. (2000). ``Comparative protein structure modeling of genes and genomes''. In: *Annu Rev Biophys Biomol Struct* 29, pp. 291–325.

Monod, J, J Wyman, and J P Changeux (1965). ``on the Nature of Allosteric Transitions: a Plausible Model.'' In: *Journal of molecular biology* 12.1, pp. 88–118.

Moult, John et al. (1995). ``A large scale experiment to assess protein structure prediction methods''. In: *Proteins: Structure, Function, and Bioinformatics* 23.3, pp. ii–iv.

Mukherjee, Srayanta and Yang Zhang (2009). ``MM-align: A quick algorithm for aligning multiple-chain protein complex structures using iterative dynamic programming''. In: *Nucleic Acids Research* 37.11.

Murzin, Alexey G. et al. (1995). ``SCOP: A structural classification of proteins database for the investigation of sequences and structures''. In: *Journal of Molecular Biology* 247.4, pp. 536–540.

Myers-Turnbull, Douglas et al. (2014). ``Systematic detection of internal symmetry in proteins using CE-symm''. In: *Journal of Molecular Biology* 426.11, pp. 2255–2268.

Paci, Emanuele et al. (2003). ``Analysis of the distributed computing approach applied to the folding of a small peptide''. In: *Proceedings of the National Academy of Sciences of the United States of America* 100.14, pp. 8217–8222.

Pauling, Linus and Robert B Corey (1951). ``Configurations of Polypeptide Chains With Favored Orientations Around Single Bonds: Two New Pleated Sheets.'' In: *Proceedings of the National Academy of Sciences of the United States of America* 37.11, pp. 729–40.

Ponstingl, Hannes, Thomas Kabir, and Janet M. Thornton (2003). ``Automatic inference of protein quaternary structure from crystals''. In: *Journal of Applied Crystallography* 36.5, pp. 1116–1122.

Prlić, Andreas et al. (2012). ``BioJava: An open-source framework for bioinformatics in 2012''. In: *Bioinformatics* 28.20, pp. 2693–2695.

Schuck, Peter (2013). *Analytical ultracentrifugation as a tool for studying protein interactions*.

Shindyalov, I N and P E Bourne (1998). ``Protein structure alignment by incremental combinatorial extension (CE) of the optimal path''. In: *Protein Engineering Design and Selection* 11.9, pp. 739–747.

Sillitoe, Ian et al. (2015). ``CATH: Comprehensive structural and functional annotations for genome sequences''. In: *Nucleic Acids Research* 43.D1, pp. D376–D381.

Simons, Kim T et al. (1997). ``Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions.'' In: *Journal of molecular biology* 268.1, pp. 209–25.

Smyth, M S and J H J Martin (2000). ``Review x Ray crystallography''. In: *J Clin Pathol: Mol Pathol* 53.1, pp. 8–14.

Srichai, Manakan Betsy and Roy Zent (2010). ``Integrin structure and function''. In: *Cell-Extracellular Matrix Interactions in Cancer*, pp. 19–41.

Wüthrich, Kurt (2003). ``NMR studies of structure and function of biological macromolecules (Nobel Lecture)''. In: *Angewandte Chemie - International Edition*. Vol. 42. 29, pp. 3340–3363.

Zhang, Yang and Jeffrey Skolnick (2004). ``Scoring function for automated assessment of protein structure template quality''. In: *Proteins: Structure, Function and Genetics* 57.4, pp. 702–710.

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Declaration of originality

The signed declaration of originality is a component of every semester paper, Bachelor's thesis, Master's thesis and any other degree paper undertaken during the course of studies, including the respective electronic versions.

Lecturers may also require a declaration of originality for other written papers compiled for their courses.

I hereby confirm that I am the sole author of the written work here enclosed and that I have compiled it in my own words. Parts excepted are corrections of form and content by the supervisor.

Title of work (in block letters):

ASSESSMENT OF PROTEIN ASSEMBLY PREDICTION IN CASP12 & CONFORMATIONAL DYNAMICS OF INTEGRIN ALPHA-I DOMAINS

Authored by (in block letters):

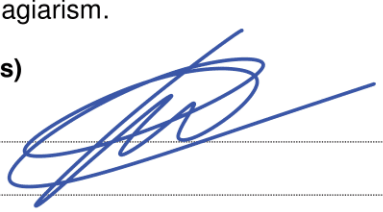
For papers written by groups the names of all authors are required.

Name(s):	**First name(s):**
LAFITA	ALEIX

With my signature I confirm that

- I have committed none of the forms of plagiarism described in the 'Citation etiquette' information sheet.
- I have documented all methods, data and processes truthfully.
- I have not manipulated any data.
- I have mentioned all persons who were significant facilitators of the work.

I am aware that the work may be screened electronically for plagiarism.

Place, date	**Signature(s)**
Zürich, 1st March 2017	

For papers written by groups the names of all authors are required. Their signatures collectively guarantee the entire content of the written paper.