Protein Sequence Information Encodes more than the Global Minimum Structure

Dominik Schwarz

Kellogg College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Michaelmas Term 2021

Declaration

I declare that no parts of this thesis or its research herein have been reproduced or accepted for another award or degree or diploma at any other university or learning institution. This thesis contains no other person's work except where stated in the text.

Dominik Schwarz, 21st December 2021

Abstract

Allostery is a conformational or activity change of a protein's active site resulting from a binding event at a distant, allosteric, site. The signal transmission is hypothesised to travel via allosteric networks and knowledge about the exact residues that transmit the signal would be beneficial for developing allosteric drugs. Allosteric drugs should offer high selectivity and/or better treatment through combination therapies.

We investigated if co-evolution techniques could be used to identify allosteric residues. While direct coupling analysis (DCA) methods without machine learning, such as EV-Fold, CCMpred and PSICOV, recalled larger numbers of allosteric residues, machine learning-based techniques like MetaPSICOV2 and RaptorX showed higher precision in predicting physical proximity (contact prediction). From this we conclude that different constraints on the sequence space are likely to be extracted by different co-evolution methods.

Next, we investigated if the co-evolutionary distance predictor DMPfold encodes information on conformational flexibility in the shape of its predicted distance distribution for each residue pair. We analysed a set of pairs of PDB structures (2947 proteins) where the two structures of the same sequence showed different conformations. The pairs were used to approximate residue pair flexibility. We found a statistically significant difference between flexible and rigid residue pairs in terms of their predicted distance distributions. Flexible residue pairs more often had multiple local maxima in their predicted distance distributions whilst rigid pairs more often had just a single maximum. This highlights the potential of co-evolution-based methods to predict conformational ensembles.

In addition to our analyses of co-evolutionary data, we explored other constraints on the sequence space of protein families: rare conformations in protein ensembles as well as folding pathways. Protein kinases are a protein family with a vast amount of structural data available and allow us to observe rare conformations in some kinases that might be accessible by other kinases at an energetic cost. We examined conformational ensembles of kinases that were generated systematically by a novel homology modelling pipeline and assessed the model ensembles' potential for docking studies. In an exploratory docking study with two kinases and five inhibitors we found the generated models to be suitable for further docking calculations.

In the last chapter we describe an initial analysis of folding pathway conservation with TMPfold, a predictor of helical membrane protein folding pathways. We found an indication for folding pathway conservation within families when analysing the predicted helix-helix association energies that build the basis for the folding pathway prediction. Nevertheless, the conservation signal was ambiguous when comparing the predicted pathways directly, suggesting that the predictor itself needs further development before being applied on a larger scale.

Contents

List of Figures

List of Tables

List of Abbreviations

ASD . . . . . Allosteric Database

ATP . . . . . adenosine triphosphate

CASP . . . . . Critical Assessment of protein Structure Prediction

CATH . . . . . database for hierarchical protein classification: Class, Architecture, Topology (fold family), Homologous superfamily

CNS . . . . . Crystallography & NMR System

cryo-EM . . . cryogenic electron microscopy

DCA . . . . . direct coupling analysis

DI . . . . . direct information

DMS . . . . . deep mutational scanning

IDP . . . . . intrinsically disordered protein

K_D . . . . . dissociation constant

MD . . . . . molecular dynamics

MI . . . . . mutual information

MP . . . . . membrane protein

MSA . . . . . multiple sequence alignments

NGS . . . . . next-generation sequencing

NMA . . . . . normal mode analysis

NMR . . . . . nuclear magnetic resonance

PDB . . . . . Protein Data Bank

TM . . . . . transmembrane

SCA . . . . . statistical coupling analysis

XFEL . . . . . X-ray free-electron laser

1

Introduction

Contents

1.1 Motivation

The ultimate goal of my doctorate was an improved understanding of the dynamic nature of proteins and the mechanisms underlying signal transmission inside proteins. In order to do this I examined three interconnected areas: allosteric networks, conformational ensembles and the assembly or membrane proteins.

Understanding the dynamic nature of proteins better will benefit biomedical research in many different ways. Allostery (a conformational change of the active site caused by binding at a distant site) is of great interest to drug discovery as it opens up the possibility of allosteric drugs. These are drugs that act in a non-competitive way and could be beneficial for combination therapies. In this thesis, I describe my work at predicting allostery on a residue-specific level. The second area in my thesis is that of conformational ensembles that are closely linked to allostery as an allosteric binding event can also be described as a shift in the populations of a protein's different conformations. Being able to predict multiple biologically relevant states of a protein would not only help understanding the impact such allosteric binding events can have, but would also allow the investigation of less-populated conformations that may be druggable but have not yet been observed experimentally. One protein family that contains many different observed conformations, and also many high priority drug targets, is the kinase family, therefore I investigate how kinase conformational space relates to potential drug selectivity. Conformational ensembles describe the dynamic nature of proteins when they are folded but folding itself is a dynamic process where mutations can have a significant impact, e.g. in Alzheimer's disease (Polychronidou *et al.*, 2020). The final area I investigated was the folding pathway prediction of helical membrane proteins to examine the relationship of folding and sequence space.

1.2 Allostery & Allosteric Networks

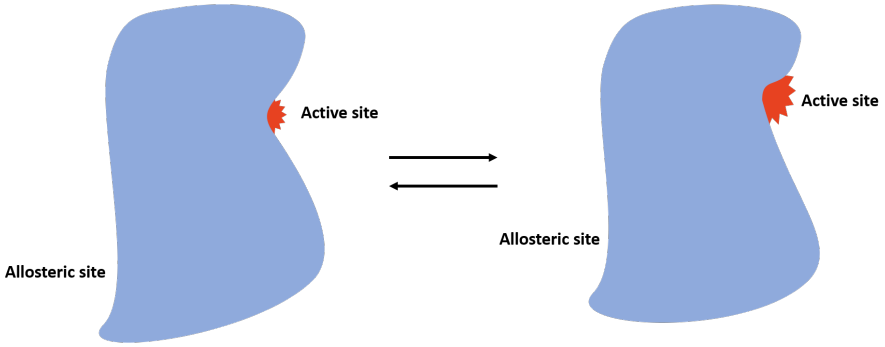
Proteins are the major active components in natural biological systems. They are a polymer chain of amino acids and are thought to fold into unique 3D shapes. Experimental techniques have revealed the structures of proteins and there are now over 180,000 publicly available structures in the Protein Data Bank (PDB) (Berman *et al.*, 2000). Computational techniques now allow us to accurately predict their static structure. However, it is also known that proteins are dynamic molecules with potentially many conformations or even without ordered structure at all. This structural flexibility can be inherent to the protein or dependent on binding events.

1.2.1 Definition

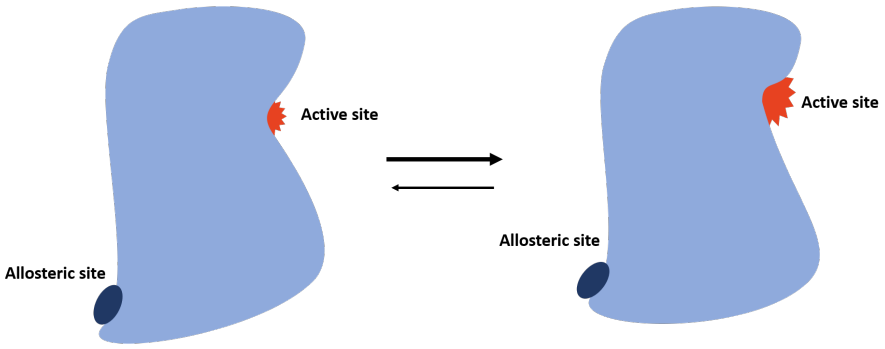
Allostery is thought to be an inherent feature of almost all biological macromolecules (Dokholyan, 2016). In the protein context, allostery is most commonly defined as a conformational change of a protein's or a protein complex's active site that follows a binding event at a different distant site, called the allosteric site (Dokholyan, 2016).

It has been argued that a conformational change is not always necessary and allostery should be defined as a change of activity as a consequence of a distant binding event (Guo and Zhou, 2016). This alternative definition follows from observations that in some cases (almost) no conformational change is seen between active structures and inactive structures of proteins with allosteric inhibitors bound (Tsai *et al.*, 2008). One potential explanation for the difference in activity is a change of entropic cost due to stiffening of parts of the protein. This is an alteration of the protein's dynamics as opposed to its conformation.

Another definition of allostery sees proteins as ensembles of structures with different activities (Figure 1.1a) and an allosteric binding event as a shift of the distribution of these ensembles (Hilser *et al.*, 2012; Tsai and Nussinov, 2014) (Figure 1.1b). This definition combines the allostery definitions of conformational and activity change: even if no structural change was visible after an allosteric event, the proportion of more and less active structures would change and thus, the activity.



(a) Protein as an ensemble of structures



(b) Shift of ensemble distribution upon binding

Figure 1.1: Ensemble view of allostery with conformational change. a) depicts an example protein with an ensemble of two different conformations with varying activity (red star size). Both structures exist in equal quantities. **b)** shows an allosteric binding event (dark blue ellipse) which causes no change in the two structures but a shift in relative abundances. As the more active conformation is present more often, the overall activity of the protein also increases.

One of the first descriptions of allostery was the Monod-Wyman-Changeux (MWC) model (Monod *et al.*, 1965). It described the allosteric events happening in haemoglobin upon oxygen binding. The first oxygen binding to a haemoglobin tetramer causes a conformational change amongst the subunits that allows subsequent oxygens to bind preferentially. This model and others at that time (Koshland *et al.*, 1966) focused on allosteric events on a subunit/domain level, in this thesis I have focused on allostery at the residue level within monomers.

On the level of monomers two main views on allosteric research exist: the thermodynamic view and the structural view (Tsai and Nussinov, 2014). The Allosteric Two State Model (Leff, 1995) is similar to the MWC model but on a monomer level. It describes the thermodynamic implications of an allosteric

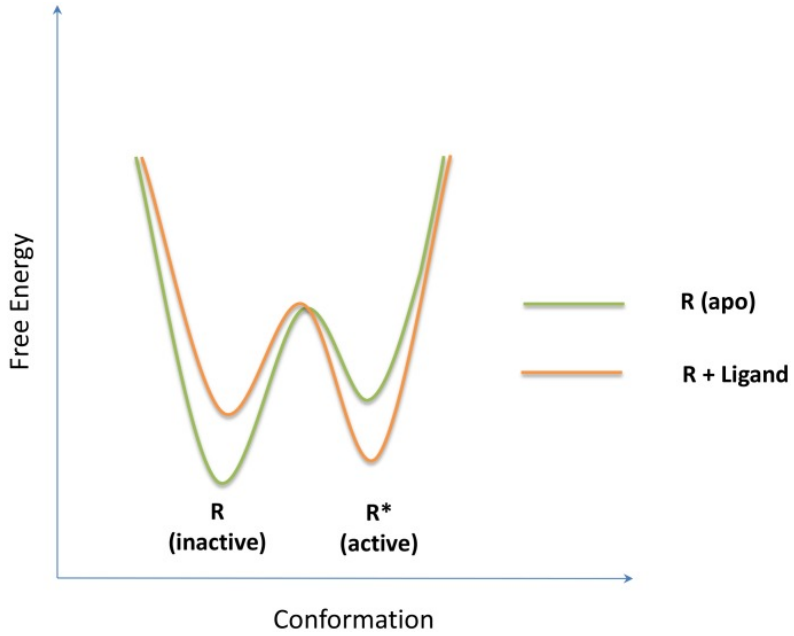
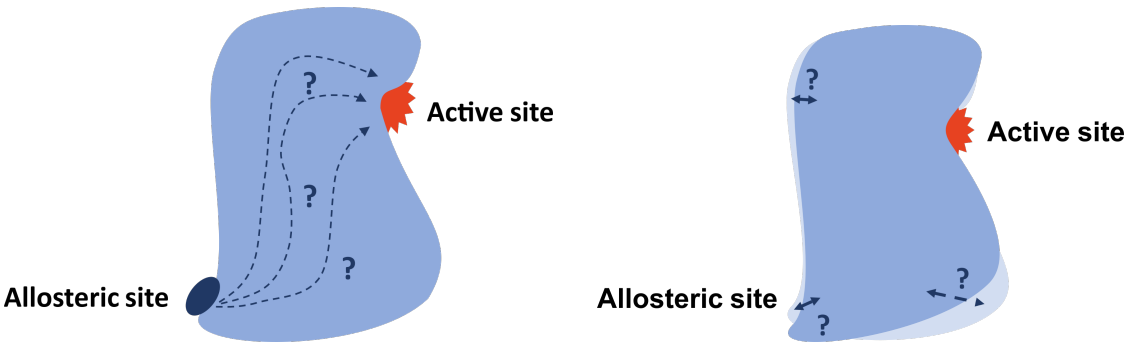


Figure 1.2: Thermodynamic perspective of allostery. Example free energy landscape of a protein modulated by an allosteric ligand. A protein in unbound form (receptor R (apo); green) exists in two states with the inactive conformations populated more as their free energy is lower than the active state's. The binding of an allosteric ligand (R + ligand; orange) changes this energy landscape so that the active state has lower free energy than the inactive form. This leads to the active form being populated more often and thus, overall protein activity to increase. The ligand would therefore be an allosteric activator. Figure from Tsai and Nussinov (2014).

binding event at a single allosteric site on a single active site. Extensions of this model exist, e.g. the relations between two allosteric sites and two functional (active) sites (Hall, 2000; Ehlert and Griffin, 2008). The thermodynamic view in general allows classifications of ligands into allosteric activators, regulators and inhibitors (Figure 1.2) but fails to explain the underlying mechanics of allostery (Tsai and Nussinov, 2014).

The structural view on the other hand, is based on the assumption that an allosteric binding event causes a higher energy state at the allosteric binding site which transfers some of that energy through single or multiple communication channels to the active site (Dokholyan, 2016) (Figure 1.3a). The term allosteric network is here defined as the residues involved in such an allosteric signal transmission. Conformational changes down to the side chain level (Figure 1.3b) may provide a mechanistic explanation for allostery but allow no quantification of



(a) Potential signal transmission channels **(b)** Potential structural changes

Figure 1.3: Signal transmission inside proteins due to an allosteric binding event. a) Potential single or multiple channels of contiguous residues transmit a signal from the allosteric site to the active site. **b)** Structural changes can be observed without knowing the exact transmission pathway, i.e. conformational ensembles give insight into potential endpoints of allosteric regulation.

allosteric efficacy and hence no classification of allosteric ligands. In this thesis, we define allostery using the structural ensemble view (Figure 1.1) and focus on cases where structural change occurs.

1.2.2 Experimental Validation

A small number of experimental techniques exist that analyse allostery but all are resource intensive. Thus, sequence-based prediction of allosteric networks would allow large-scale analysis of allostery and facilitate a systematic study of the phenomenon. Unfortunately, computational methods for allostery prediction lack a general validation strategy which is especially true for the verification of (full-size) allosteric networks. In general, the two main areas of experimental allostery validation are structure determination and biochemical assays coupled to mutagenesis experiments.

Structure determination methods

The main methods of structure determination have been for many years X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy, but cryogenic electron microscopy (cryo-EM) has made fast progress in recent years and has now

overtaken NMR in terms of yearly deposited structures in the PDB (Shoemaker and Ando, 2018).

While NMR is mainly being used to investigate allostery from a thermodynamic perspective (Manley and Loria, 2012; Tsai and Nussinov, 2014) and to study entropy-driven allostery (Xu *et al.*, 2019), X-ray crystallography is used from a structural perspective, e.g. analysing protein structures with and without a bound allosteric ligand. Crystallographic structure determination has made progress in recent years enabling the study of a large number of crystals (>1000) from the same protein and multi-temperature crystallography which allow detection of differently populated conformers within a protein dataset (van den Bedem *et al.*, 2013; Keedy *et al.*, 2015b, 2018). So far, these studies have been performed only with apo-proteins or with ligand fragments bound (Keedy *et al.*, 2018); experiments with full-size allosteric ligands are still missing. This would enable not only the study of a protein's conformational ensemble but potentially also the direct influence of ligand binding on this ensemble.

Cryo-EM has enabled the observation of proteins and protein complexes much closer to their native environment than crystallography and NMR. Samples of protein in solution or even whole viruses or cells (Wolff *et al.*, 2020) are shock-frozen before EM images are taken which captures a protein of interest in many different states. This brings the advantage of recording the dynamic nature of a protein but also provides a major challenge in producing high-resolution structures (as crystallography routinely does). Computational analysis of many different protein images, advances in technical infrastructure and using prior chemical knowledge have improved cryo-EM resolution with 1.22Å resolution being the current record for a rigid protein (Nakane *et al.*, 2020). However, only very few structures solved using cryo-EM have 2Å resolution or better, stressing the need for crystallography when accurate atom- or residue-level data is desired (Yip *et al.*, 2020). Nevertheless, cryo-EM has already been used to investigate allostery (Banerjee *et al.*, 2016; Haselbach *et al.*, 2017; Zhang *et al.*, 2021); similar to using X-ray crystallography, structures are solved with and without ligands to investigate their impact. Special

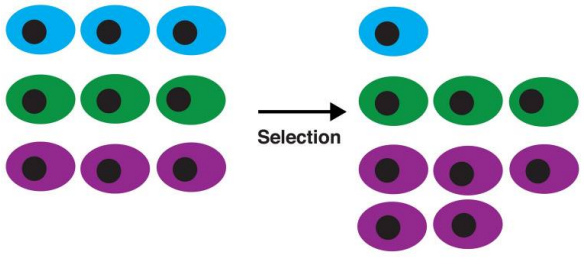
for cryo-EM is not only the capturing of multiple conformations but also the dimensions that can be examined, for example, Haselbach *et al.* (2017) showed an allosteric signal transmission over 150Å.

Still in its infancy are X-ray free-electron laser (XFEL) experiments (Keedy, 2019). XFEL is a crystallographic technique where femtosecond X-ray pulses are used to image a stream of microcrystals, yielding many structural snapshots of a protein of interest providing time resolution of protein dynamics at the femtosecond-scale (Chapman *et al.*, 2011; Shoemaker and Ando, 2018). While the technique is still limited by the ability of a protein to crystallise, time resolution in combination with other techniques such as 'mix-and-inject' (Stagno *et al.*, 2017) allow the study of chemical reactions, or binding and conformational rearrangement events. Even though there have not been dedicated studies on allostery yet, combining XFEL with strong electric field pulses can be used to resolve coupled motions in a protein (Hekstra *et al.*, 2016), a key descriptor of allostery, highlighting the technique's potential.

Mutagenesis experiments

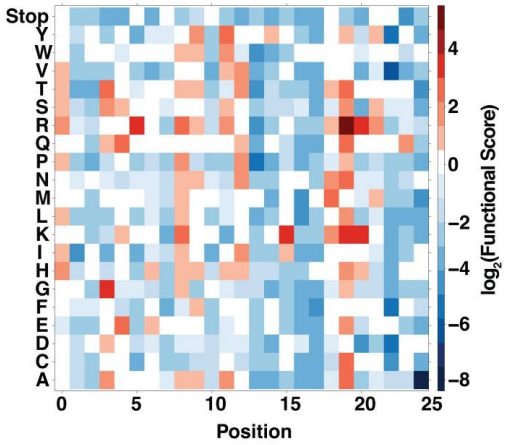
Mutagenesis experiments have developed from single mutations to systematic alanine mutation studies (Yamamoto *et al.*, 2006) up to deep mutational scanning (DMS) (Fowler and Fields, 2014) which is also covered by the terms saturation mutagenesis and multiplexed assays of variant effects (Weile and Roth, 2018). DMS is an experimental technique which tests the fitness of a large number of single and double mutants of a protein of interest (Figure 1.4). Generally, all amino acids of a protein are mutated to all other 19 canonical amino acids (Figure 1.4b) and (a subset of) all possible double mutant combinations as well (Salinas and Ranganathan, 2018).

The technique is based on high-throughput sequencing coupled to a functional assay. A library of single and double mutants is synthesised, expressed (*in vitro* or *in vivo*) and then a selection assay is performed, e.g. selection for binding but selection for signal transmission is also possible. The fitness of each individual member of the library is normally determined by the ratio of sequence reads before



Variant	Mutation	Counts input	Counts selected	Functional score
Blue	A60P	3	1	0.33
Green	WT	3	3	1
Purple	S36T	3	5	1.67

(a) Determination of functional fitness



(b) Functional fitness of all single mutants

Figure 1.4: Deep mutational scanning infers functional fitness for large mutant libraries. a) Selection can occur under various different conditions but functional fitness is generally calculated the same: the ratio between variant counts before and after selection determines the functional fitness score. **b)** Functional fitness is tested for all possible mutants (here: single mutants) and can increase (red) or decrease (blue). Images taken from Fowler and Fields (2014)

and after selection (Figure 1.4a). If the fitness of a double mutant differs from the combined fitness of the corresponding single mutants, the interaction is termed epistatic (Figure 1.5, bottom right) (Fowler and Fields, 2014).

It has been shown that epistasis is an indicator for spacial proximity of a residue pair (Sahoo *et al.*, 2015; Melamed *et al.*, 2013) and two recent publications, Schmiedel & Lehner (Schmiedel and Lehner, 2019) and Rollins *et al.* (Rollins *et al.*, 2019) demonstrated that DMS data can be used to determine the fold of small protein domains. They both treated epistasis data as contact prediction data. Schmiedel & Lehner reported contact precision values of up to 73% for top L predicted contacts (L being protein length) (Schmiedel and Lehner, 2019). Similar to some *de novo* folding techniques based on co-evolution data (Greener *et al.*, 2019; Xu, 2019), this contact prediction was then used as geometric distance constraints for *ab initio* folding, e.g. by simulated annealing molecular dynamics (Schmiedel and Lehner, 2019).

The key limitation of this technique is the number of double mutants that need to be tested experimentally which grows rapidly with protein length (L): $L * (L - 1) * 19 * 19$. For a protein with 300 amino acids this roughly equals a

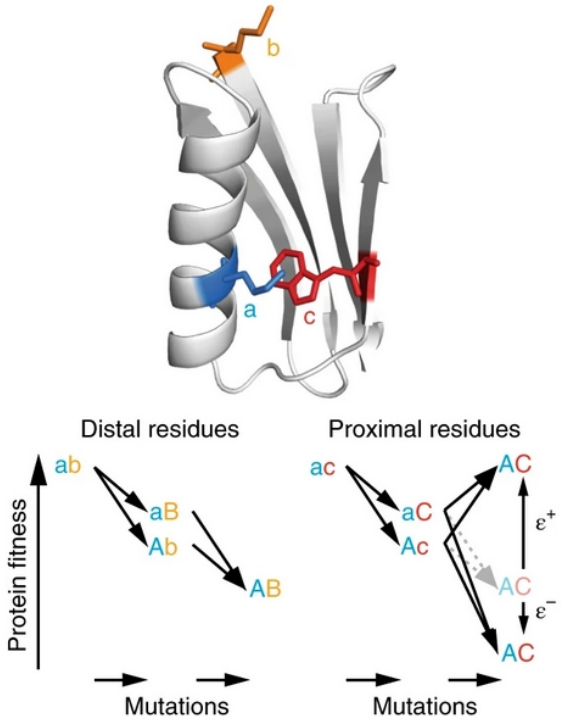


Figure 1.5: Epistasis explained by proximity. The individual fitness of two single mutants is used to calculate an expected fitness for the corresponding double mutant. If the measured fitness of a double mutant deviates from the expected fitness, the two mutated residues are considered to be epistatic (interacting). When selecting for mainly structural constraints, e.g. stability, epistasis is shown to indicate residue pairs in physical proximity (AC, blue and red). When the measured fitness equates to the expected fitness (AB, blue and orange), no interaction is inferred (here: no proximity). Image taken from Schmiedel and Lehner (2019)

library of 16 million mutants. The largest study, that was analysed in the two publications above, investigated 535,917 double mutants (of 555,940 maximum; $L = 56$) (Olson *et al.*, 2014). Rollins *et al.* tested whether a reduction in library size still resulted in successful contact prediction. When selecting only double mutants of pairs of single mutants that were most deleterious, they found that a library size of 5% is sufficient to achieve contact prediction precision close to the one from a non-reduced library (Rollins *et al.*, 2019). This points to the technique's potential applicability for larger proteins.

Both of these DMS studies focused on accurate structure prediction, but as seen by the raw precision values (around 70%), spacial proximity might not be the only cause for epistasis. Allostery, signal transmission inside a protein, or functional

importance (Rollins *et al.*, 2019) might be reasons for false positives in DMS-based contact prediction but might be important for the prediction of allosteric networks.

The authors of both studies state that the technique might be a promising tool for all protein families that do not have sufficient sequence variation for co-evolution-based structure prediction. But, if the coupled assay selected for allosteric signal transmission instead of stability, DMS could become a very valuable technique in validating allosteric networks. For the Ras protein such a selection assay exists already but has only been performed with libraries of single mutants (Bandaru *et al.*, 2017). Even though the Ras study focused on the impact of single mutations, it demonstrated the significant impact of the selection environment on mutation effects. The mutation effects were mostly deleterious in presence of two of the normal regulators of Ras (both nucleotide exchange factors) as opposed to hotspots of activation effects in the absence of these regulators. This points to the importance of assay design, especially for studying signal transmission. Currently only a relatively small number of DMS datasets are available (Weile and Roth, 2018) and most assays select for binding, expression or stability (Kim *et al.*, 2013; Rocklin *et al.*, 2017; Diss and Lehner, 2018; Matreyek *et al.*, 2018; Starr *et al.*, 2020). Nevertheless, at least three datasets exist (Bandaru *et al.*, 2017; Salinas and Ranganathan, 2018; Leander *et al.*, 2020) where DMS was performed and mutants were experimentally tested for allostery, showing the technique's potential for future allostery research.

Allosteric Database

Advances in crystallography, cryo-EM and sequencing methods have brought large-scale allostery validation within reach but few studies have yet focused on allosteric signal transmission. Therefore, we used for some of our analyses a different validation strategy that is based on multiple allostery experiments from different research groups. The Allosteric Database (Shen *et al.*, 2016; Liu *et al.*, 2020) (ASD; http://mdl.shsmu.edu.cn/ASD/) is a collection of information about allosteric modulators, their targets and their interactions. The database also contains *Allosteric pathways* which are equivalent to the allosteric networks defined

above. Each allosteric pathway (or network) is a dataset of experimentally-validated allosterically important residues within a protein. Forty-four of these allosteric pathways have defined residues (verified residues). These pathways/networks contained between one and fifteen verified residues. None of the datasets contain a full-size allosteric network but we choose this database as our validation set as it has one important advantage over other approaches. Validating against multiple different studies lowers the risk of biasing any methodology towards any specific protein or experimental technique.

The ASD's newest update (Liu *et al.*, 2020) contains another dataset potentially interesting for validation: *Allo-Mutations*. It is a dataset of 1312 allosteric mutations (not at the active site) that are associated with one of 33 cancer types. These mutations were verified in large-scale next-generation sequencing (NGS) studies (Tomczak *et al.*, 2015; Tate *et al.*, 2019). If a mutation is associated with more than one cancer this mutation is included multiple times in the overall count, also when the same residue has multiple mutations which are allosteric. This decreases the number of verified residues that would count as validated. Nevertheless, these 1312 mutations are distributed over 133 proteins which is roughly three times as many as in the *Allosteric pathways* dataset. This data is similar to DMS with the combination of sequencing and a functional assay. Sequencing though is less systematic as in DMS and the functional assay verifies involvement in cancer which brings a bias towards phenotypic cell-level effects over allosteric protein-level effects (signal transmission). Furthermore, such a phenotypic mutation could also have an impact on stability (and not signal transmission) which is subsequently causing disease. Mutations that are so severe that they disagree with cell life would also not be detected because they wouldn't be observed in real patient samples. However, the increased number of experimentally-validated residues spread over more proteins is a good argument for including this data into future studies.

1.2.3 Computational Prediction Methods

Molecular dynamics

Several computational methods can be used to investigate allostery. The most common being molecular dynamics (MD) (Feher *et al.*, 2014; LeVine *et al.*, 2018) which simulates the motions of a defined set of atoms or residues. MD simulations sample the dynamic nature of a protein, ranging from side chain rotations (picosecond time scale) to loop motions (nanosecond time scale) and domain motions (microsecond to millisecond time scale) (Henzler-Wildman and Kern, 2007). Conformations of a protein's ensemble that require longer transition times than those time scales cannot yet be sampled, with atomic resolution MD. Coarse-grained MD offers a route to longer time scales (e.g. 50μs (LeVine *et al.*, 2018)) but loses the direct link to the atomic scale. MD simulations are used to sample conformational ensembles or directly calculate allosteric signalling from its trajectories (Hospital *et al.*, 2015).

Normal mode analysis

For studying protein dynamics on a more coarse-grained level, normal mode analysis (NMA) has been used (López-Blanco and Chacón, 2016). NMA analyses elastic networks where amino acids are simulated as nodes of a network which are connected by springs that may have different properties (elasticity, length). NMA samples the slower dynamical motions of a protein (Skjaerven *et al.*, 2009) and may also be used to study perturbations of a system, e.g. mutations. Combining NMA with a structure-based statistical mechanical model of allostery (Guarnera and Berezovsky, 2016) is one example of using NMA for allosteric pathway/network prediction; for example, this is implemented in the ASD as the *'Allo-Pathway'* tool (Liu *et al.*, 2020).

Graph theory analysis

Some techniques predict allosteric sites and allosteric networks based on graph theory analysis (Amor *et al.*, 2016; Wang *et al.*, 2020). For example, for the webserver Ohm (Wang *et al.*, 2020), input structures are analysed based on proximity and bonding between pairs of residues. These features then guide a perturbation propagation

algorithm resulting in allosteric coupling intensities and subsequently in allosteric pathways. The method has been validated against 20 known allosteric proteins and individual residues of some of those proteins have been experimentally mutated to test predictions. The authors also investigated a dependency of their predictions on the input structure. When taking several snapshots from 100ns MD simulations as inputs, they found only a mild variation of their predictions (Wang *et al.*, 2020).

Sequence-based prediction

MD simulations are computationally expensive and all current methods require a structure to perform analysis, therefore sequence-based methods would be particularly useful for studying allostery on a larger scale.

A partially sequence-based methodology was presented by Lakhani and coworkers (Lakhani *et al.*, 2017). They tackled the problem of allosteric network prediction by using co-evolutionary sequence information derived from statistical coupling analysis (SCA; see section *Co-evolution analysis*) (Rivoire *et al.*, 2016) and combined it with structural information. In this thesis, we wanted to explore the potential of more recent techniques of co-evolution analysis for allostery prediction and try to establish an automated pipeline that was independent of crystal structure information. Please see Results section *Co-evolution Analysis for Allosteric Network Prediction / MutS* for a more detailed description of Lakhani *et al.* (2017)'s work and our comparisons to it. In general, the main challenge in developing any allosteric network prediction tool is its validation. Lakhani *et al.* (2017) verified their predicted network by selecting only a subset (13%) of their predicted allosteric residues and by evaluating their effects in short (15ns) MD alanine mutation studies and small molecule docking. MD simulations are useful tools to identify dynamically coupled structural components (atoms, residues, segments or domains) (Stolzenberg *et al.*, 2016), but they are only one source of information which complement experimental validation techniques.

Despite the multiple computational methods described above that have been developed to predict allostery, none have yet been validated on a large-scale experimental dataset.

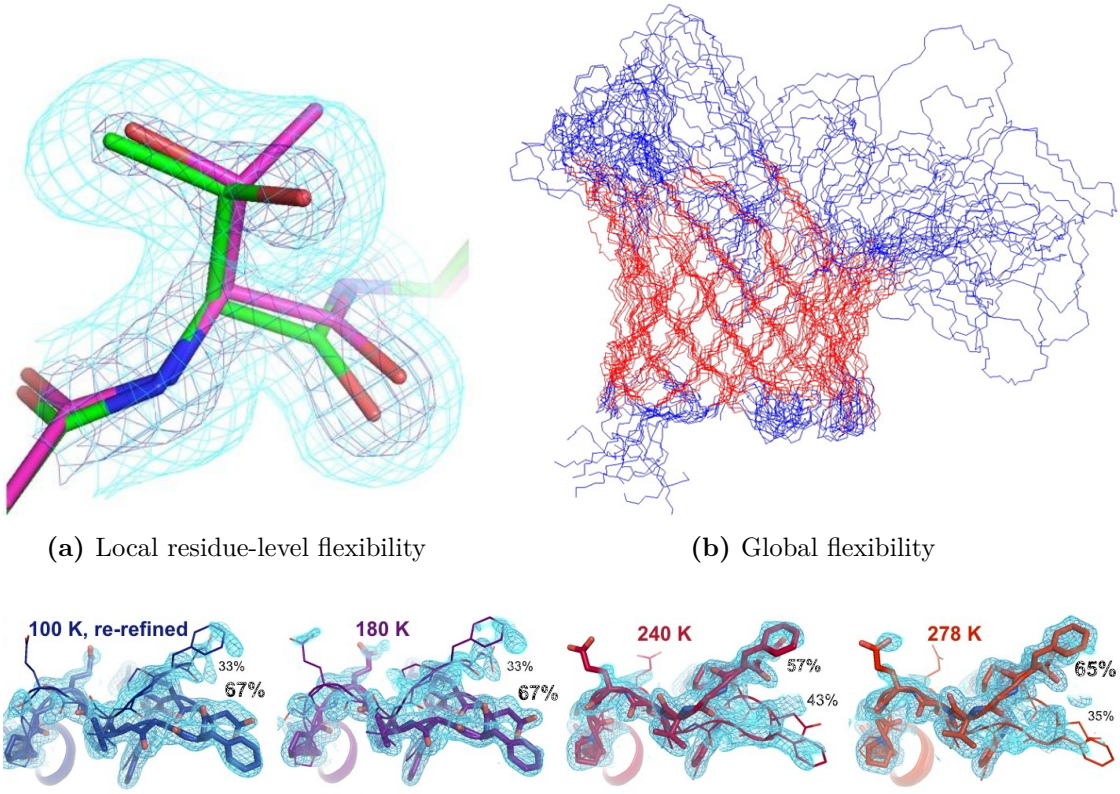
1.3 Protein Flexibility & Conformational Ensembles

1.3.1 Allostery and Protein Flexibility

As explained above in *Allostery & Allosteric Networks / Definition*, allostery can be described in the context of protein conformational ensembles. Proteins exist in a number of biologically relevant states that differ in their activities (Figure 1.1a) and an allosteric binding event may shift the relative free energies of these states (Figure 1.2) so that the occupation of these ensemble states changes (Figure 1.1b) (Hilser *et al.*, 2012; Tsai and Nussinov, 2014). Analysing the conformational ensemble of a protein or even being able to predict the multiple biologically relevant states of it, would be beneficial for understanding the inherent structural changes that occur and what modulates them (allostery). These structural changes can range from side chain movements to whole domain movements (Henzler-Wildman and Kern, 2007) and are always associated with specific residues or regions being flexible (Figure 1.3b). Therefore, knowing which residues change their interactions between different conformers can be informative as to which residues are important in allosteric signalling.

1.3.2 Protein Flexibility Definitions

Protein flexibility in this thesis always relates to ordered proteins as opposed to the flexibility of intrinsically disordered proteins (IDPs). Structured proteins and IDPs vary in several characteristics such as function, amino acid composition and evolutionary rate (van der Lee *et al.*, 2014). Proteins generally exist in a conformational ensemble and different biologically relevant states each possess a spectrum of nearly isoenergetic substates (Frauenfelder *et al.*, 1991). This points out the two general regimes of protein flexibility: local (isoenergetic) flexibility and global



(a) Local residue-level flexibility

(b) Global flexibility

(c) Multiple conformations with different population sizes

Figure 1.6: Different regimes of conformational flexibility. a) Alternative side chain conformations modelled into electron density maps are an example of local flexibility on the residue-level. Image taken from Keedy *et al.* (2015a). **b)** Conformational ensembles determined by NMR constraints display an example of global flexibility. Image taken from Liang and Tamm (2007). **c)** Multiple biologically relevant conformations (local or global) with annotated population sizes represent the ensemble view that condenses the large conformational space. Models here were obtained with multi-temperature crystallography. Image taken from Keedy *et al.* (2018)

.

flexibility. While local flexibility can be seen as flexibility on the residue-level (Figure 1.6a), global flexibility is captured by conformational ensembles (Figure 1.6b).

Local sampling of those isoenergetic substates can be done with various different methods, for example, using flexible side chains in docking studies, but sampling major conformational changes remains challenging. Non-isoenergetic states with more substantial conformational changes might be sampled very rarely (Guimarães *et al.*, 2011), induced by ligands or other binding partners (Lang *et al.*, 2014). Methods for obtaining such biologically relevant states experimentally or for predicting them computationally are described in the following two sections.

Furthermore, we introduce a third concept of flexibility which might be able to close the gap between residue-level flexibility and conformational ensembles: residue pair flexibility (Schwarz *et al.*, 2021). It describes which residue-residue interactions are flexible and thus, are able to change and facilitate switching between different relevant protein conformations. Knowing which residue pairs have flexible interactions will be useful for the prediction of a protein's structural ensemble.

1.3.3 Experimentally Examining Protein Flexibility

Experimentally determining protein flexibility can be achieved with various techniques such as cryo-EM (see section *Allostery & Allosteric Networks / Experimental Validation*), nuclear magnetic resonance (NMR) (Frueh *et al.*, 2013), Förster resonance energy transfer (FRET) (Sanyal *et al.*, 2016), hydrogen-deuterium exchange (HDX) (Hamuro *et al.*, 2003; Zhang, 2020) or X-ray free-electron laser (XFEL) (Keedy *et al.*, 2015b) studies. While for NMR spectroscopy technical difficulties increase for proteins larger than around 20kDa ($\approx$ 200 amino acids) (Frueh *et al.*, 2013), this is also true for cryo-EM and proteins smaller than 64kDa (Shoemaker and Ando, 2018). Although protein size is less of an issue for FRET, HDX and XFEL studies, all these techniques require extensive experimental resources making large-scale application currently infeasible.

While most methods produce snapshots of a conformational ensemble or their data is treated to obtain discrete states, this might not be reflective of reality, e.g. rotor rotation of ATPases (Zhao *et al.*, 2015). Therefore, analyses have been developed to also describe continuous states and changes (Frank and Ourmazd, 2016; Punjani and Fleet, 2021). However, for investigating which residue pairs are interaction-changing or flexible, a description with discrete states could be an appropriate approximation.

Multi-temperature crystallography can yield multiple conformations of a protein in one experimental setup (Figure 1.6c and section *Allostery & Allosteric Networks / Experimental Validation*) but standard X-ray crystallography can also be used to probe different degrees of flexibility. One example of local flexibility that can be

found within a single X-ray structure is the B-factor (Schlessinger and Rost, 2005). The B-factor tries to capture atomic displacement due to thermal fluctuations but often also captures model error (Sun *et al.*, 2019). Another example is the modelling of alternative side chain and backbone conformations which explain a model's underlying electron density better than with just a single conformation (Lang *et al.*, 2014). The final case is when multiple structural models are obtained from different experiments each on the same protein sequence. These different structural shapes can be used to approximate global flexibility.

A few databases that store all the structures related to a single protein sequence exist (Monzon *et al.*, 2016; Hrabe *et al.*, 2016). The database of Conformational Diversity in the Native State of proteins (CoDNaS) (Monzon *et al.*, 2016) is useful as the authors have curated a subset of proteins where they have identified the two PDB (Berman *et al.*, 2000) structures of a protein sequence that have the maximum root-mean-square deviation (RMSD) amongst all pairwise comparisons of a protein's available PDB structures (Monzon *et al.*, 2017). These structure pairs are a proxy for a protein's maximum flexibility.

1.3.4 Computational Prediction Methods

Several computational methods have been developed for the prediction of protein flexibility. They can be divided into two main types: prediction of individual residue flexibility and prediction of structural ensembles.

Residue-level flexibility can be inferred and predicted by several different techniques that include information from crystallographic B-factors (Schlessinger and Rost, 2005), normal mode analysis (NMA) (Jacobs *et al.*, 2001), NMR chemical shift data (Cilia *et al.*, 2014) and molecular dynamics (MD) simulation data (Narwani *et al.*, 2019). Some of these methods can predict flexibility from sequence alone but such residue-level predictions contain no information about different distinct conformations of a protein. In contrast, ensemble predictors based on methods like NMA (Lindahl *et al.*, 2006; Krüger *et al.*, 2012), distance geometry (Greener

et al., 2017), homology modelling (Schwarz *et al.*, 2019), coarse-grained structure-based models (Morcos *et al.*, 2013) or MD simulations (Kuriata *et al.*, 2018) yield information about distinct conformations of a protein but currently no ensemble prediction method can make calculations directly from sequence information.

In addition, most methods that predict residue-level flexibility and all of the ensemble predictors rely on experimentally determined or fully modelled structures to guide analysis. Generally, sequence data is easier to obtain and thus more widely available than structural information. Therefore, it would be desirable to develop ensemble predictors that only need protein sequences as input which would enable the large-scale prediction of conformational ensembles at low computational cost. Recent progress in protein structure prediction (Jumper *et al.*, 2021) makes accurate *de novo* models more widely available and reduces the bottleneck of structural data. However, even the most recent structure prediction tools (Jumper *et al.*, 2021; Minkyung *et al.*, 2021) are not trained to predict a diverse set of conformations but instead, aim to predict only the global minimum structure. They also tend to give less accurate predictions in areas of the protein that are flexible.

1.4 Co-evolution

1.4.1 Definition

Co-evolution of protein residues occurs when the interaction between two (or potentially more) residues is important for a protein's folding, stability or functionality (Neher, 1994) (e.g. allosteric signal transmission). If one of the interacting residues mutates, the other(s) is (are) likely to mutate as well to restore the interaction and satisfy that evolutionary constraint. In contrast to residues that might be crucial for catalytic activity, the interaction of co-evolving residue pairs is conserved, and not their amino acid identity or even chemistry. Residue-residue couplings can be derived from multiple sequence alignments (MSAs) of protein homologs by considering the correlation between columns in the alignment (more details in the next section).

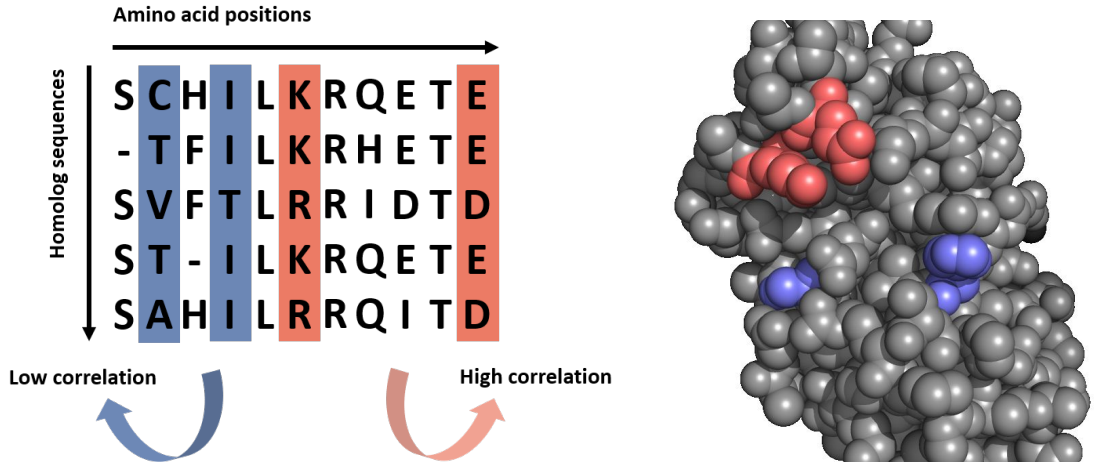
1.4.2 Co-evolution Analysis

Multiple sequence alignments (MSAs) form the basis of all co-evolution analysis methods (Süel *et al.*, 2003; Marks *et al.*, 2011; Jones *et al.*, 2015; Jumper *et al.*, 2021). Figure 1.7a shows a fictional MSA of homologous sequences of an example protein and its structure, Figure 1.7b. Based on correlations between MSA columns (residue positions), a co-evolution analysis method predicts couplings between pairs of residues. Highly correlating residue pairs (red) are assigned higher scores or probabilities to be coupled than residue pairs that are uncorrelated or correlate less (blue). When some evolutionary constraint favours a certain interaction between a pair of residues, for example an attractive interaction between a lysine (K) and a glutamate (E), this interaction is often found in many sequences of a protein family (red). While in this example the amino acid conservation is also high with only two different amino acids per column, key for co-evolutionary conservation is the paired ability to fulfil the evolutionary constraint, e.g. an interaction between arginine (R) and aspartate (D) would also be attractive. Residue pairs with strong co-evolutionary coupling are often in physical proximity (Figure 1.7b red) (Anishchenko *et al.*, 2017) whereas pairs with low correlation are more likely to be further apart (blue).

Statistical coupling analysis

Statistical coupling analysis (SCA) (Süel *et al.*, 2003; Halabi *et al.*, 2009) is an early example of co-evolution analysis that is closely related to mutual information (MI), the 'raw' correlation between two columns in an MSA of homologs (Figure 1.7a).

Mutual information quantifies the information that is gained about a random variable X from observing another random variable Y (Cover and Thomas, 2012). It can be seen as a form of correlation of discrete variables such as amino acid identities. In the context of residue positions in an MSA, this means how often is amino acid A occurring at the same time (in the same sequence) as another amino acid B at a different residue position. Measuring these co-occurrences in two specific columns over all sequences of an MSA leads to the mutual information of these two amino acid positions. In Figure 1.7a, the two red columns will have a higher



(a) Multiple sequence alignment **(b)** 3D structure

Figure 1.7: Co-evolving residue pairs in multiple sequence alignment and 3D structure. a) Multiple sequence alignments form the basis of all co-evolution analysis methods. They assign highly correlating residue pairs (red) higher scores or probabilities to be coupled than residue pairs that correlate less (blue). When some evolutionary constraint favours a certain interaction between a pair of residues, for example an attractive interaction between a lysine (K) and a glutate (E), this interaction is often found in many sequences of a protein family; not conserved in terms of amino acid identities of individual residues but in terms of the paired ability to fulfil the evolutionary constraint, e.g. an interaction between arginine (R) and aspartate (D) would also be attractive. **b)** Pairs with strong co-evolutionary coupling are generally in physical proximity (red) whereas pairs with low correlation are often not close to each other (blue).

mutual information than the blue columns which matches the intuition of the red residue pair mutating (co-evolving) concurrently (0.971 red, 0.721 blue).

Covariance scores are calculated for all pairs of positions (columns) i, j of an MSA, yielding a covariance matrix C. The MI covariance matrix is given by

$$C_{ij}^{MI} = \sum_{aa_A} \sum_{aa_B} p_{ij}^{AB} log \frac{p_{ij}^{AB}}{p_i^A p_j^B}$$

with p_i, p_j and p_{ij} being the individual and joint probabilities (observations) of the residue pair and individual residues with amino acids A and B. The double sum represents all combinations of those two amino acids at that pair of positions. The covariance matrix calculated by SCA is using the same parameters and is given by

$$C_{ij}^{SCA} = \sqrt{\sum_{aa_A} \sum_{aa_B} (p_{ij}^{AB} - p_i^A p_j^B)^2}.$$

SCA additionally weights these raw scores by the conservation of an alignment column and defines 'sector' residues to be the 20% most correlating residues that are

detected by spectral decomposition. Therefore, it yields two levels of information: pairwise scores in the covariance matrix and individual residue scores after spectral decomposition of the covariance matrix.

The downside of both statistical measures is that they quantify the observed correlation of columns but not necessarily the underlying causes for these observations. While SCA has been used for analysing and predicting allostery (Lakhani *et al.*, 2017; Salinas and Ranganathan, 2018), SCA is less good at predicting the physical proximity of residue pairs than co-evolution techniques that were developed later to improve proximity (contact) prediction. The SCA method only achieves around 5-7% precision (in top 75 contact predictions) (Horner *et al.*, 2007).

Direct coupling analysis

Significant improvements in contact prediction have been achieved by a set of methods based on direct coupling analysis (DCA) (Marks *et al.*, 2011; Morcos *et al.*, 2011). For example, the DCA-based method EV-Fold (Marks *et al.*, 2011) achieves about 38% precision in top L (where L is the length of the protein) contact predictions (de Oliveira *et al.*, 2016). Methods using DCA extract pairwise correlations of residue positions but try to discriminate between direct correlations and transitive ones (Figure 1.8). In other words, if residues A and B are correlated and so are residues B and C, then there will also be a correlation between the columns of A and C in a raw MSA. The correlation between A and C is called transitive as it does not reflect a true interaction.

DCA methods have different underlying mathematical frameworks, e.g. mean-field approximation (Morcos *et al.*, 2011) and pseudo-likelihood maximisation (Ekeberg *et al.*, 2013), but they all try to avoid these transitive correlations. Mean-field approximation (Morcos *et al.*, 2011), for example, uses the maximum entropy principle to infer direct information (DI) from a model that takes the frequency counts of an MSA as input and fits residue pair couplings and local biases (background fields). The maximum entropy principle states that a model that makes the fewest assumptions about the underlying data is the best. For explaining the

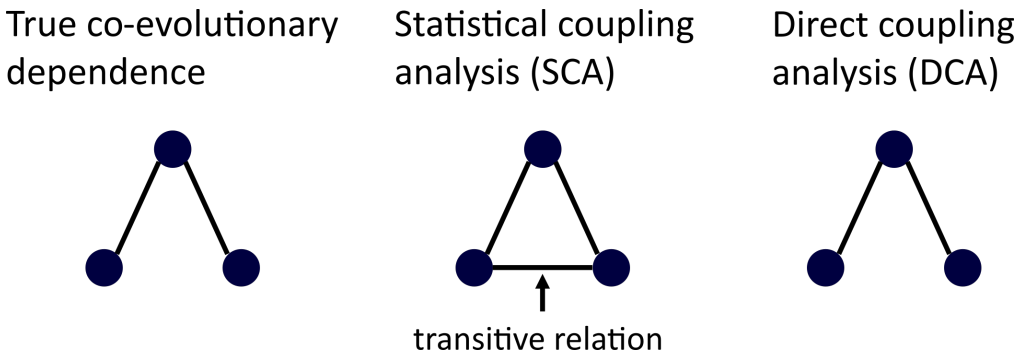


Figure 1.8: Evolutionary coupling inference by co-evolution analysis methods. If all three amino acids (nodes) correlate in an MSA, the SCA method will assign couplings (edges) between all three residue pairs. A true co-evolutionary dependence may only exist between two of them, e.g. when three amino acids are in one line. The first residue (bottom left node) is in contact with the second (the top node) and the second with the third (bottom right node) but the first and the third do not interact (hence, no edge between them). DCA methods try to not infer these transitive couplings using different correction approaches.

observed amino acid frequencies in an MSA this means that couplings between residues are more equally distributed amongst all pairs of residues than the observed correlations measured by mutual information (see Figure 1.9). Direct information (the pairwise score that DCA computes) can be seen as the amount of MI between two residue positions that results from direct coupling alone (Morcos *et al.*, 2011).

DCA's improved precision in contact prediction has led to significant improvements for template-free structure prediction (*ab initio* modelling as opposed to homology modelling) (Marks *et al.*, 2011; Nugent and Jones, 2012; Schaarschmidt *et al.*, 2018). Co-evolution analysis methods using the DCA approach have successfully been applied to other research tasks as well, e.g. domain boundary prediction (Ovchinnikov *et al.*, 2016), prediction of protein-protein interactions (Hopf *et al.*, 2014) or loop modelling (Marks and Deane, 2018). Multiple transitivity-avoiding methods have been developed since the publication of EV-Fold, e.g. CCMpred (Seemayer *et al.*, 2014) and PSICOV (Jones *et al.*, 2012), as well as contact predictors that use the residue-residue coupling analyses and other (e.g. structural) features as input for machine learning predictions. RaptorX (Källberg *et al.*, 2012) and MetaPSICOV (Jones *et al.*, 2015; Buchan and Jones, 2018) are examples for these kind of meta predictors and both show higher precision in contact predictions (Wang

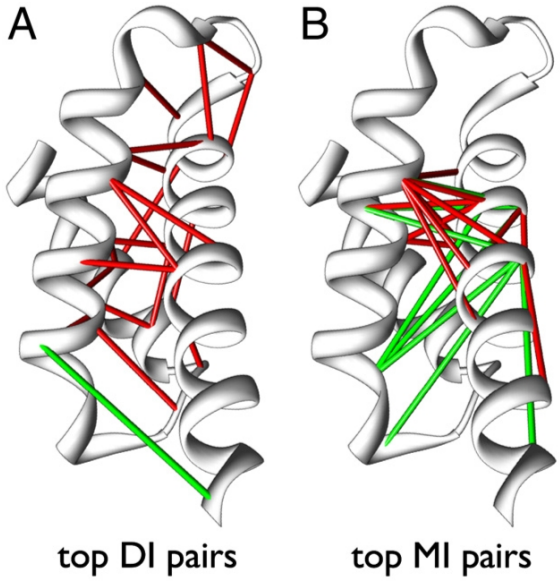


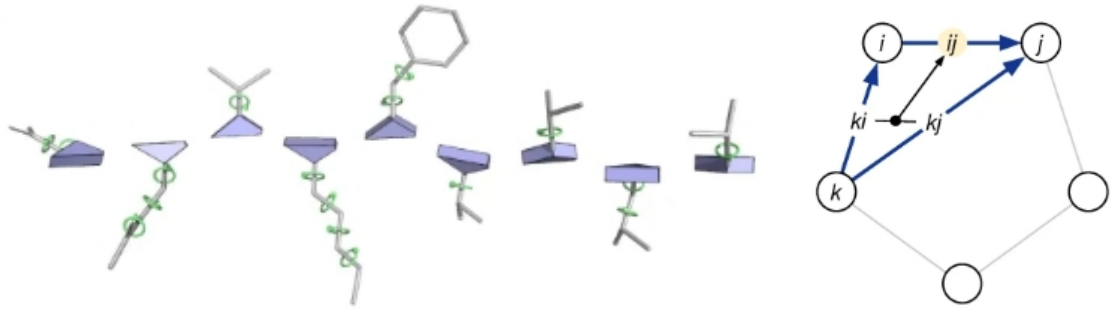
Figure 1.9: DCA predicts more evenly distributed contacts throughout a domain than SCA. Contacts in red are predicted correctly whereas green contacts are false positives. **a)** Direct information (DI) from DCA predicts more true positives that are also more spread out than predictions from **b)** Mutual information (MI). Image taken from Morcos *et al.* (2011).

et al., 2017; de Oliveira and Deane, 2017). MetaPSICOV, for example, reaches 60% precision in top L contact predictions (de Oliveira *et al.*, 2016).

Neural network predictors

The area of protein structure prediction additionally benefited from training neural networks on large sets of available protein structures alongside co-evolutionary information directly inferred from the MSAs (Jones and Kandathil, 2018; Adhikari *et al.*, 2018; Wang *et al.*, 2017). These neural networks were developed to train methods to predict residue-residue distances instead of binary contacts which improved structure prediction even further (Greener *et al.*, 2019; Xu, 2019; Senior *et al.*, 2020; Yang *et al.*, 2020). These distance predictions yield a probability distribution over distance bins instead of assigning a probability of two residues being within the common binary threshold of 8Å (methods for distance prediction from co-evolutionary data are discussed in more detail in chapter 3).

Recently these machine learning-based methods have progressed even further, methods such as AlphaFold2 (Jumper *et al.*, 2021) are now considered potential



(a) Residue gas **(b)** Triangular updates

Figure 1.10: Two key novelties in the AlphaFold2 deep learning algorithm. **a)** AlphaFold2 uses a 'residue gas' to represent amino acid residues in 3D space. Backbone residues (blue triangles) are freely moving rigid bodies with side chains having rotatable bonds (green circles). **b)** Residue pair interactions are updated in a triangular manner; shown here is the 'Triangular multiplicative update based on incoming edges'. Images taken from Jumper *et al.* (2021)

solutions for the protein structure prediction problem (Pereira *et al.*, 2021). The power of AlphaFold2 was demonstrated at the last CASP competition. CASP (Kryshtafovych *et al.*, 2021) is the Critical Assessment of protein Structure Prediction where different research groups try to blindly predict the structure of proteins that had no published structure at the time of prediction.

AlphaFold2 uses a combination of several novel methodologies to create a deep learning algorithm that directly predicts protein structure from sequence data input (Jumper *et al.*, 2021). Sequence information is encoded into an MSA representation and a residue pair representation which update each other in an iterative manner going through the prediction process. The residue pair representation stores information about the pairwise interactions such as predicted distances or other features. These pairwise interactions inform the prediction of a 'residue gas' that contains all amino acid residues as freely moving entities; rigid bodies for backbone atoms combined with angles for the side chain bonds (see Figure 1.10a). Unique to this approach is that it does not enforce protein chain connectivity until the very end of the structure prediction process.

Instead, each residue (pair) interaction is updated based on its neighbourhood which is treated in a novel way. An attention network architecture allows the method to learn for each residue which neighbouring residues are important. This

is done not only in a pairwise manner but with triangular operations based on three residues (see Figure 1.10b). An example would be the triangle inequality that states that the longest edge of a triangle can only be as long as the sum of the other two edges; edges in this context mean pairwise distances. This leads to iterative rotations and translations of the residues in the residue gas forming the basis for structure prediction.

In the structure module of the network this prediction is then brought into a protein-like shape via gradient descent with an Amber force field and peptide bond consistency enforced. The process of updating from sequence information to pairwise interactions and structure prediction is done multiple times, each time learning from the previous prediction. Updating between the different levels of representation are done with an end-to-end technique that facilitates performing all operations in relative space instead of a fixed coordinate space. This is achieved through using an equivariant transformer network architecture (Bouatta *et al.*, 2021).

AlphaFold2 yields *de novo* models to an accuracy that is often considered to be at the level of experimental accuracy (Pereira *et al.*, 2021). However, even this newest generation of co-evolution-based structure prediction tools (Jumper *et al.*, 2021; Minkyung *et al.*, 2021) only aims to predict static protein structures and not multiple biologically relevant conformations. In this thesis, the goal is to analyse the potential to not collapse co-evolutionary predictions into one structure but instead predict multiple conformations.

1.4.3 Causes for Evolutionary Coupling

As can be seen from the above, the focus of co-evolution methods has been on contact and structure prediction. However, as mentioned earlier, co-evolution methods (SCA) have also been used to analyse allostery (Lakhani *et al.*, 2017; Salinas and Ranganathan, 2018). Whether transitivity avoidance through DCA techniques or machine-learning based methods are useful for allostery prediction had not been tested out. The key question being if co-evolutionary data contains any information on allostery at all.

It has been shown that strong co-evolutionary (direct) coupling is a good indicator of an evolutionary constraint on a particular residue pair interaction that leads to spacial proximity (Anishchenko *et al.*, 2017). This proximity can be within the protein chain itself, within other conformations (Morcos *et al.*, 2013; Toth-Petroczy *et al.*, 2016) or between homomeric or heteromeric protein complexes (Ovchinnikov *et al.*, 2014; de Oliveira and Deane, 2017). This likely stems from stability and fold constraints and it has been shown that different types of contact predictors may predict different types of physicochemical bonds that hold domains together (Chonofsky *et al.*, 2019). The average number of bonds per contact is higher in predictions from early DCA methods compared to machine learning-based predictors and the distribution of interaction types, e.g. electrostatic or hydrophobic, that are found within the top predictions, differs as well (Chonofsky *et al.*, 2019). This indicates that, dependent on the co-evolution analysis method used, co-evolutionary signal is caused by different interaction characteristics.

As physical proximity has the strongest co-evolutionary footprint, stability constraints are likely to be the primary reason for couplings, but they could also be caused by (additional) functional constraints which might involve allosteric signalling (which is not mutually exclusive with proximity). The fact that deep mutational scanning (DMS) showed different residue pair interactions to be important depending on the assay's selection conditions (Bandaru *et al.*, 2017), further supports the view that co-evolutionary couplings can have different causes. The mutational effects differed when selection happened in presence or absence of the two co-factors of Ras (whose binding allosterically control activity) (Bandaru *et al.*, 2017).

Furthermore, DCA-based methods that predicted only binary contacts (Marks *et al.*, 2011; Jones *et al.*, 2012; Seemayer *et al.*, 2014) were found to assign a higher coupling score (and rank) for residues that were in contact across all known PDB structures of a multiple sequence alignment, compared to residue pairs that were only in contact in a subset of the structures of that multiple sequence alignment (Zea *et al.*, 2018). A similar observation was also found for protein-protein interaction surfaces where conserved interactions more often showed stronger co-evolutionary

couplings than less conserved ones (Rodriguez-Rivas *et al.*, 2016). Since in these binary predictors a lower rank indicates a lower probability, residue-residue contacts that are only present in some conformations of a protein's ensemble are more likely to be falsely classified as not in contact (at all).

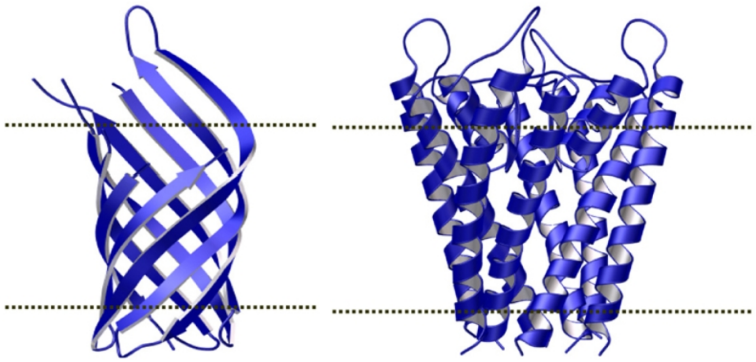
Spatial proximity causes co-evolutionary coupling but current methods only focus on this one feature that leads to accurate static structure prediction. There is an indication that the time frame of that spacial proximity relates to the strength of the co-evolutionary signal: is the residue pair interacting at all times, just in some conformations or even only in presence of a binding partner? The same is true for the reasons of that proximity: is it important for stability, signal transmission or for enabling conformational switching?

1.5 Folding Pathways of Helical Transmembrane Proteins

1.5.1 Folding and Membrane Insertion

About 20-30% of protein-coding genes encode integral membrane proteins (MPs) (Krogh *et al.*, 2001) of which two general classes exist: beta-barrel (Figure 1.11a) and alpha-helical (Figure 1.11b) MPs (White and Wimley, 1999). We focus here on α-helical transmembrane proteins that have more than one helix spanning the membrane (polytopical), because their thermodynamic properties allow a computational prediction of their folding pathway (Lomize *et al.*, 2020).

Membrane proteins' native environments are different from those of soluble proteins. Soluble proteins usually consist of a hydrophobic core and a more polar surface as they are found in the polar solvent water. MPs sit across the membrane which, due to the nature of phospholipids, requires MPs to interact with multiple environments: a hydrophobic layer (the fatty acid tails), a charged layer (the head groups) and the interface between those layers. Therefore, amino acid compositions between soluble and membrane proteins differ (Hill *et al.*, 2011), for example, an α-helix spanning the membrane is much more hydrophobic than an α-helix of a soluble protein (Deber *et al.*, 1986). This has thermodynamic consequences



(a) Beta-barrel **(b)** Helix-bundle

Figure 1.11: Typical polytopical transmembrane protein folds. Polytopical membrane proteins span the membrane multiple times as opposed to bitopical membrane proteins that only span it once, e.g. have one helix inserted into the membrane that anchors a soluble domain. **a)** Beta-barrel transmembrane proteins consist of β-strands forming a cylinder. **b)** The helix bundle transmembrane protein shown here is representative of the third group of membrane proteins: α-helical transmembrane proteins that consist of multiple α-helices spanning the membrane. They form helix bundles with various different helix counts. Images taken from Arinaminpathy *et al.* (2009).

that guide membrane protein folding and assembly of individual transmembrane segments into higher order structures.

Thermodynamic constraints

Figure 1.12 shows estimated energetics of different helix components of the transmembrane helix of glycophorin A (Cymer *et al.*, 2015). This transmembrane helix is hydrophobic and its insertion from an aqueous phase into the membrane has an overall favourable free energy of -12 kcal/mol. Splitting this free energy into the components of inserting a polyglycine helix into a membrane and the difference between glycophorin's side chains relative to glycine highlights the importance of the amino acid composition. Inserting polyglycine into a membrane has an unfavourable free energy (+24 kcal/mol) due to the energy that is necessary for dehydrating the polypeptide backbone. This is only counterbalanced by the hydrophobic side chains of glycophorin A (-36 kcal/mol) that move from a polar environment to a non-polar one.

Furthermore, the free energy difference between inserting the folded polyglycine (+24 kcal/mol) compared to the unfolded peptide (+104 kcal/mol) emphasises the

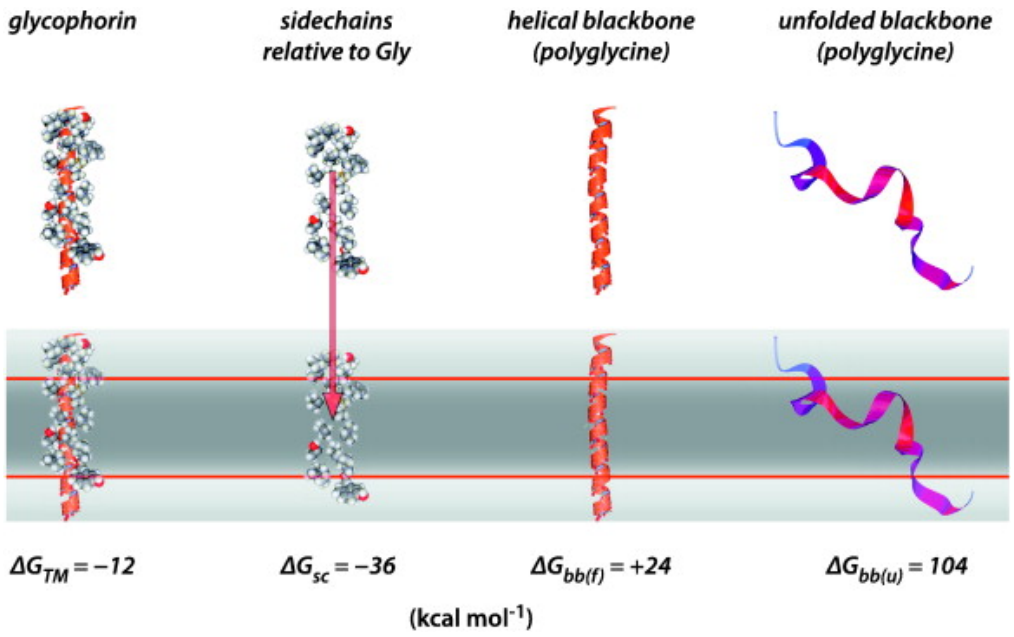


Figure 1.12: Energetics of transmembrane helix insertion components explain α-helix stability in membranes. Estimated Gibbs free energies of glycophorin A (without sugars) membrane insertion are shown for insertion of the full helix (-12 kcal/mol), only its side chains (-36 kcal/mol), only its backbone (+24 kcal/mol) and as an unfolded backbone (+104 kcal/mol). The large unfavourable energy of inserting an unfolded polypeptide directly into the membrane points out the importance of secondary structure (hydrogen bonding) formation for membrane insertion. Inserting just the hydrogen-bonded helix backbone also has an unfavourable (but smaller) energy which comes from the dehydration of the backbone. This energy cost is compensated by a favourable energy contribution of the hydrophobic side chains of glycophorin that are taken from a polar environment to a hydrophobic one. The overall free energy for inserting glycophorin's transmembrane helix is negative and thus, favourable. Image taken from Cymer *et al.* (2015).

role that preformation of secondary structure plays. Hydrogen-bonding between the peptide's backbone atoms (into an α-helix) has a very favourable free energy and is thought to occur before membrane insertion.

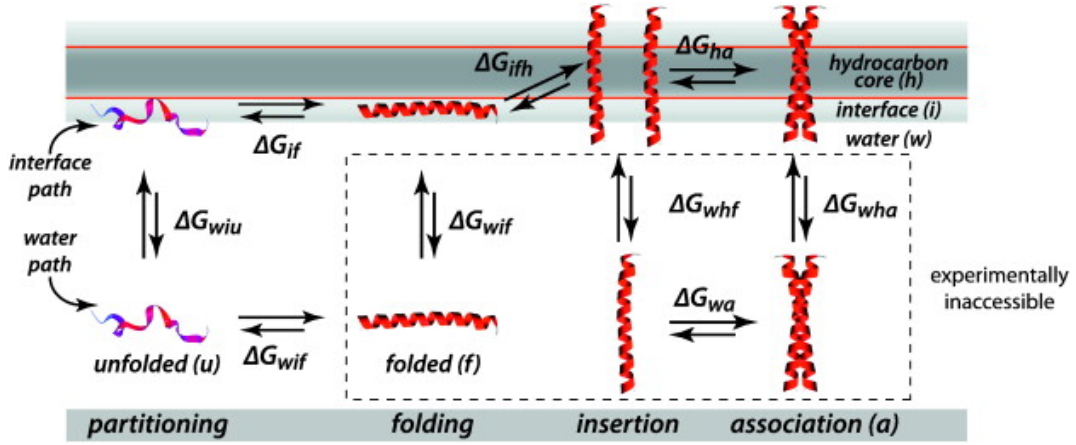
This preformation of secondary structure is thought to occur at the membrane interface (Ladokhin and White, 1999) which leads to a four-step model (White and Wimley, 1999) of membrane protein folding and assembly (Figure 1.13a). While formation of secondary structure and even association of multiple helices are theoretically possible in water, observing these states experimentally is near impossible as thermodynamics favours partitioning to the membrane (interface) (Cymer *et al.*, 2015). Thus, the most likely pathway for most MPs is the interface

path where unfolded MPs interact with the membrane interface first, then fold (form secondary structure), then insert into the membrane and finally associate to higher order helix bundles (Cymer *et al.*, 2015).

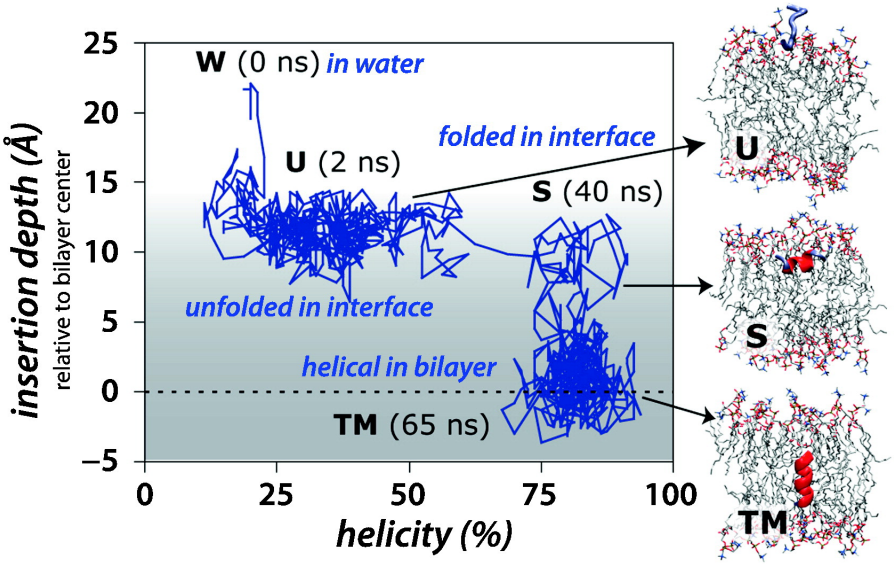
The four-step model (White and Wimley, 1999) has been tested using an unbiased MD simulation where polyleucine was initialised in the water phase near a membrane (Ulmschneider *et al.*, 2011); unbiased means here that the polypeptide was not driven or constrained in any direction by the simulation setup. Figure 1.13b displays the trajectory of the simulation with regard to insertion depth and helicity (foldedness). The unfolded (20% helicity) peptide starts in water (W) but more or less immediately partitions onto the membrane-water interface (U). At this interface, secondary structure starts to form (between 10-60% helicity) until the helix folds completely (>75% helicity), still at the interface (S). Only then, after folding, the helix moves from the membrane interface into the bilayer (TM) where it remains folded but is able to go back and forth between interface and bilayer. This MD simulation matches the interface path proposed by the four-step model, showing its relevance for modelling the thermodynamic constraints of membrane protein folding (Cymer *et al.*, 2015).

Facilitation of membrane protein folding

While thermodynamics govern which states of a protein constitute local or global energy minima, the height of the energy barriers between them and thus, the pathway and speed of folding towards the global minimum structure are determined by kinetics as well (Eaton, 2021). Chaperones bind unfolded polypeptide chains to avoid protein aggregation or to free kinetically-trapped (misfolded) folding intermediates (De Geyter *et al.*, 2020). While this is true for soluble and membrane proteins, membrane proteins experience additional help by translocases/insertases (translocons) that facilitate insertion into the membrane. Their folding and insertion happen co-translationally, which reduces aggregation (Harris *et al.*, 2017), but is minimised further when ribosomes dock to translocons located in the target

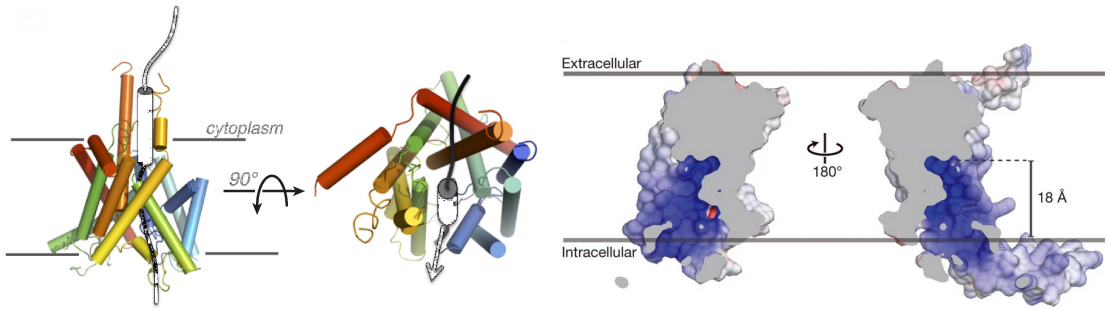


(a) Four-step model of helical transmembrane protein folding



(b) MD simulation of polyleucine partitioning and insertion

Figure 1.13: Four-step model of helical transmembrane protein folding observed in unbiased MD simulation. a) A four-step model (White and Wimley, 1999) describes the thermodynamic constraints of helical transmembrane protein folding in terms of partitioning from solution onto the membrane interface, folding of secondary structure, insertion of hydrophobic helices and association of multiple helices into helix bundles. While only the path with initial partitioning is experimentally accessible, other paths are possible. **b)** An unbiased all-atom MD simulation of a ten leucine helix in water and double lipid bilayer (Ulmschneider *et al.*, 2011) shows rapid partitioning from water (W) to the bilayer interface (U). The peptide forms secondary structure at this interface (increase in helicity, from U to S) before inserting into the membrane only when its fully folded (no increase in helicity, from S to TM). Images taken from Cymer *et al.* (2015).



(a) Transmembrane channel translocon **(b)** Groove translocon

Figure 1.14: Transmembrane helix insertion is facilitated by two different types of translocons. a) Transmembrane channel translocons allow a polypeptide to fully traverse a membrane through its channel and it also allows hydrophobic helices to enter the membrane via its lateral gate. Image taken from Cymer *et al.* (2015). **b)** Some translocons (insertases) do not have transmembrane channels but facilitate membrane insertion via their hydrophilic groove which is thought to lower the energy barrier between the soluble and lipid environments. Image taken from Kumazaki *et al.* (2014).

membrane of the synthesised protein (Cymer *et al.*, 2015). Translocons reduce energy barriers along the folding and insertion pathway (Cymer *et al.*, 2015).

Two types of translocons exist, ones that have a transmembrane channel where unfolded polypeptide chains can pass through (Figure 1.14a) and others that do not have a channel but a hydrophilic groove that does not span the full membrane (Figure 1.14b). Being able to pass a polypeptide through a membrane is important for membrane proteins with domains or segments that are not embedded in the membrane but are soluble, or for secreted proteins. It is thought that this groove reduces the energy barrier of inserting a folded helix from the membrane interface into the bilayer by enabling interactions with water, lipid and charges at the same time (Kumazaki *et al.*, 2014). A similar function is hypothesised for the lateral gate of transmembrane channel translocons which allows access to the membrane from the channel (Cymer *et al.*, 2015).

While different translocons can have different substrates that they support in membrane insertion, they all alter folding kinetics (speed) but not the membrane proteins' thermodynamics (Cymer *et al.*, 2015), i.e. their (global) energy minimum structure. Therefore, the assumption that the folding pathway of membrane proteins can be inferred by analysing a structure near this global minimum (e.g. Figure 1.15a),

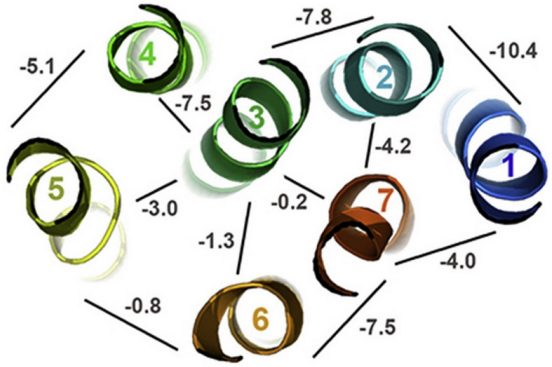
could be valid when just the speed of insertion is altered. The strength of helix-helix interaction energies would then determine in which order helix associations occur.

1.5.2 Prediction of Folding Pathways

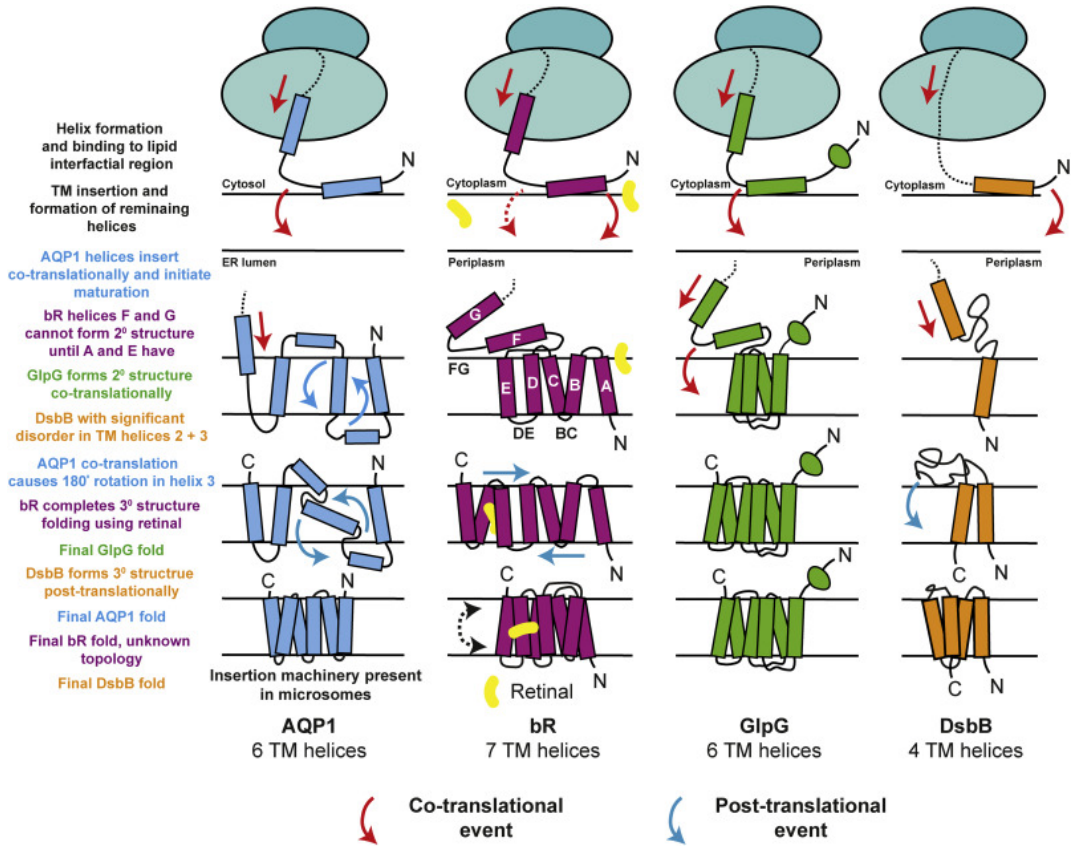
TMPfold (TM Protein Folding) (Lomize *et al.*, 2020) is a computational method developed to predict folding pathways of helical transmembrane proteins based on an input structure. The underlying assumption is that membrane protein folding is mainly thermodynamically controlled and helix bundle interactions determine the folding pathway. The basis of TMPfold is the prediction of helix-helix energies which have been validated against experimental association energies (Lomize *et al.*, 2020). The predicted pairwise association energies (Figure 1.15a) then inform the prediction of a tentative folding pathway based on energy cutoffs.

Considering that the insertion of membrane proteins is supported by different translocases and/or insertases (Cymer *et al.*, 2015) that decrease energy barriers, one could assume that the speed of transmembrane helix insertion was infinite. Then, all helices would be in the membrane at the same time and the strength of their interactions (thermodynamics) would govern their assembly pathway completely. However, as the speed of insertion is not infinite, the folding/assembly pathway is also driven by the sequential nature of protein translation. In other words, if interaction energies between helix one and six (of an example membrane protein) are the most attractive amongst all possible helix pairs, they might not be the helix-helix pair that is interacting first because helix one might have formed a helix bundle with helix two already (because they are translated and inserted directly after one another). TMPfold (Lomize *et al.*, 2020) also takes this sequential nature into account when predicting a folding pathway.

Figure 1.15b displays four experimentally verified folding pathways of helical transmembrane proteins. In principle, the folding pathways of bR (purple) and GlpG (green) are compatible with the TMPfold assumption to predict pathways based on helix-helix interactions of the final state as well as sequential insertion driven by (co-)translation (Pellowe and Booth, 2020). However, the AQP1 (blue)



(a) TMPfold thermodynamics assumption



(b) Experimentally verified folding pathways

Figure 1.15: Exceptions to fully thermodynamics-controlled folding pathway assumption occur. a) TMPfold assumes a fully thermodynamics-controlled folding pathway of transmembrane proteins which allows the prediction of that pathway just based on the interaction energies between helices in the final fold, i.e. a PDB structure. Image taken from Lomize *et al.* (2020). **b)** While co-translational folding governs the sequential insertion into the membrane and bR (purple) and GlpG (green) folding pathways follow the thermodynamic assumptions that are encoded in TMPfold folding pathway predictions, AQP1 (blue) and DsbB (orange) undergo post-translation rearrangements that are not consistent with TMPfold assumptions. Image taken from Pellowe and Booth (2020).

and DsbB (orange) folding pathways are not consistent with these basic assumptions because they undergo post-translational rearrangements. In the case of AQP1, one transmembrane helix, that already has been inserted during translation, flips its orientation in the membrane after translation (Virkki *et al.*, 2014). Interaction energies calculated based on the final state are not reflective of the possible interactions during the folding process (because the helix is inverted). This flip can be explained with a low propensity of that helix to be stable in a membrane by itself (Lomize *et al.*, 2020). For DsbB, two helices do not form and insert into the membrane during translation, only after the final helix has been translated, inserted and associated to the very first helix (Harris *et al.*, 2017). Again, this could be explained with low individual stability estimates for those helices (Lomize *et al.*, 2020) and highlights an area of improvements in folding pathway prediction.

1.6 Thesis Outline

In this thesis, we describe analyses that investigate the relationship between protein sequence space and the functional constraints that are encoded in it. For this, we mainly used co-evolution techniques to examine residue-level information but we also analysed sequence space in the form of families of homologs. We started with the goal of predicting allosteric networks of proteins, analysed if conformational flexibility is contained in sequence data, investigated conformational ensembles of kinases and also assessed a folding pathway predictor for helical membrane proteins.

These analyses investigate basic biological research questions which might not only yield insights on biological principles, but also help to improve analysis methods that extract these kinds of information. More advanced analysis and prediction techniques may benefit drug discovery research in the long-run and thus, society. For example, being able to predict allosteric networks or to identify (allosteric) conformation-switching residues would enable the specific targeting of those residues and help in the discovery of allosteric drugs. Additionally, being able to accurately generate conformational ensembles of proteins would allow us to target rare conformations that could offer increased selectivity. But we also

analysed protein sequence space for other functional features next to those that are related to conformational flexibility. We investigated if constraints on protein folding pathways are encoded in the sequence space because altered folding pathways can lead to diseases like Alzheimer's or Parkinson's. Therefore knowledge of or even the prediction of folding pathways may benefit this kind of biomedical research as well. Utilising the vast amount of available sequence information by extracting functional properties of proteins is the underlying goal of this thesis which has been addressed through several different research angles.

More specifically, in chapter 2, we describe our investigation of allostery prediction based on co-evolutionary information. Studies had indicated that co-evolution analysis could yield allosteric networks and we examined the potential of more recent co-evolution methods for this task and to predict allosteric residues in general. While avoiding transitivity in predicting residue-residue couplings improved allosteric residue recall, machine learning-based contact predictors had increased contact prediction precision but recalled fewer of the allosteric residues compared to DCA methods without machine learning. This suggested that different types of analysis methods can retrieve different kinds of information from sequence data.

In chapter 3, because of the lack of large-scale validation data that directly verifies allostery, we used more widely-available data that is linked to allostery, namely structural data from the PDB. We analysed if co-evolutionary distance predictions encode information on conformational flexibility using a large-scale study of about 3000 proteins. We approximated flexibility with the two most different PDB structures of a protein and investigated if co-evolutionary distance predictions capture information on this flexibility in the shape of the predicted distance distributions for each pair of residues. A statistically significant difference between flexible and rigid residue pairs could be found indicating that flexibility information can be extracted from co-evolutionary data. This work is published in *Bioinformatics* (Schwarz *et al.*, 2021).

In chapter 4, we explain a homology modelling approach that systematically generates conformational ensembles of human kinases and how we tested the models'

suitability for docking studies. The conformational ensembles are generated by using chimeric template structures that each represent a unique combination of the different flexible features of human kinases, allowing the modelling of rare conformations that might be druggable. We picked five inhibitors of two kinases and tested which models of the kinases' ensembles where predicted to bind the inhibitors best. For some of the test cases good poses could be found in models that matched the structural classification of the inhibitor's crystal structure and for others good poses were found in non-matching structures, indicating that the model ensembles can be used for docking calculations when the expected classification state is generally ranked high. This project is published in *Proteins: Structure, Function and Bioinformatics* (Schwarz *et al.*, 2019).

Chapter 5 displays an initial analysis of the first predictor of helical membrane protein folding pathways to investigate the impact of folding pathways on the sequence space. We investigated if folding pathways are conserved within protein families and could find an indication for this in the predicted helix-helix association energies that build the basis for the folding pathway prediction. However, the conservation signal was ambiguous when comparing the predicted pathways directly, suggesting that the predictor itself needs further development before being applied on a larger scale.

2

Co-evolution Analysis for Allosteric Network Prediction

Contents

2.1 Background

As described in the introduction, Lakhani and coworkers (Lakhani *et al.*, 2017) showed that there is some information about allostery in the contact prediction networks that can be derived from statistical coupling analysis (SCA). However, the paper studied only one system (the prokaryotic DNA mismatch repair protein MutS), used several arbitrary parameters and was only validated using molecular dynamics (MD) simulations. This chapter describes my work to build on and improve the methodology described by Lakhani *et al.* in order to build a more robust and generalisable allosteric network prediction pipeline. We identified three main areas to examine for improvements:

1. replacing the statistical coupling analysis with more recent co-evolution analysis methods that more accurately predict residue-residue contacts,

2. testing the use of sequence information alone to remove the need for a solved protein structure. This change would allow allostery predictions on a much larger scale, and

3. automating the allosteric network prediction and selection of important residues for signal transmission by using network centrality measures.

An automatic, accurate allostery prediction pipeline would lead to an increased general understanding of allostery and its mechanisms of signal transmission. Furthermore, prediction on a residue-specific level might enable the control of allosteric functions for the investigation of allosteric drugs acting in a non-competitive way which, for example, could be beneficial for combination therapies (Ni *et al.*, 2020). To enable residue-level predictions we examined the use of co-evolution techniques which have been shown to produce more precise contact predictions between pairs of residues than SCA.

Allostery can be described as the dynamic coupling of two sites that are not within the physical interaction range (Dokholyan, 2016). This coupling needs to be transmitted presumably via an allosteric network of residues. If such a network

is composed of single or multiple communication pathways is unclear (Agarwal *et al.*, 2002; Sol *et al.*, 2006). To identify such a network via co-evolution analysis, this dynamic coupling would be required to leave an evolutionary footprint, seen as evolutionary couplings between residue pairs. Therefore, the first goal of this chapter is to demonstrate that co-evolutionary information contains allosteric signal.

We applied DCA methods to the MutS dataset, generated different residue networks with those and used centrality measures to compare SCA- and DCA-derived networks. While we found some overlap in those networks, it was unclear which methodology was correct in deviating predictions. The MD-based validation approach and the fact that only a single protein was analysed, led us to using a larger validation dataset´with 17 proteins with experimentally verified allosteric residues. The analysis showed the highest recall of those allosteric residues by DCA methods without machine learning, while machine learning predictors achieved the highest precision in contact prediction. This led us to the hypothesis that different co-evolution methods are able to extract different kinds of information from sequence data and in the future, co-evolutionary signal should be separated into structural and functional constraints. We tested this hypothesis by re-analysing our data with a combination of the previous methods and information on the rigidity of residue pairs drawn from predictions of co-evolutionary distance predictors. It could be shown that removing rigid residue pairs from predictions increased the recall of allosteric residues for all methods, highlighting the potential of splitting co-evolutionary signal.

2.2 Methods

2.2.1 Datasets

MutS

MutS is a homodimeric DNA repair protein that recognises DNA mismatches and recruits other mismatch repair proteins as a result. Its function is thought to involve allosteric signalling as the DNA binding site, that recognises a mismatch, and the ATP-binding domain, that in consequence changes its affinity for ATP, are on opposite sides of the MutS protein (Figure 2.3b).

Lakhani *et al.* (2017) predicted an allosteric network (see Figure 2.3a) with statistical coupling analysis (SCA) of a multiple sequence alignment (MSA) of MutS homologs and the PDB structure 1NNE_A. The MSA that Lakhani *et al.* used as input for their SCA was based on a manually-curated sequence alignment of six homologous sequences that served as a reference for aligning the rest of the homologous sequences available at that time. Their MSA contained 164 sequences in total and 142 effective sequences (N_{eff}) at a 90% identity threshold (for more information see Methods section *Multiple Sequence Alignments and Effective Sequences*).

The set of residues that was declared the allosteric network of MutS by Lakhani *et al.* was created by defining nodes as the top 20% correlating residues according to SCA (Lakhani *et al.*, 2017). Those 157 residues were called sector residues and edges were created between them when their minimal heavy atom distance was below 6.8Å (in PDB 1NNE_A) and they were not shielded by other heavy atoms (shadow map algorithm (Noel *et al.*, 2012)). The resulting network generated by Lakhani *et al.* is depicted in Figure 2.3a.

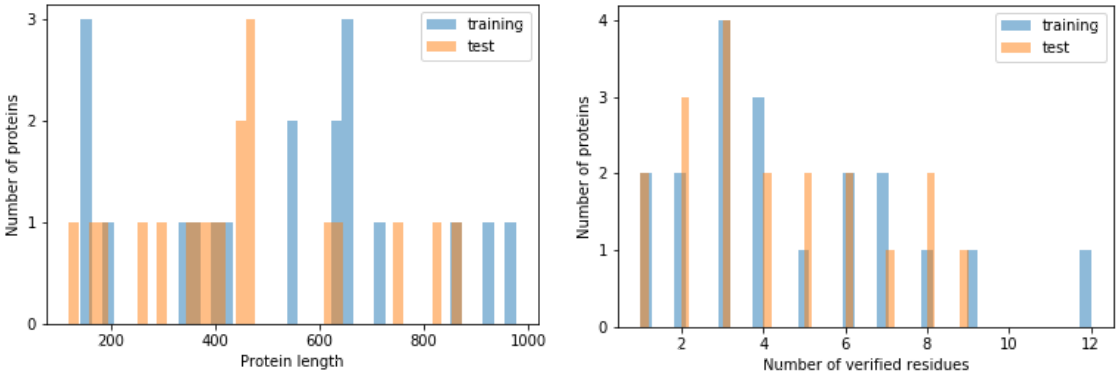
Allosteric Database

The Allosteric Database (ASD) version 3 (Shen *et al.*, 2016) was downloaded from http://mdl.shsmu.edu.cn/ASD/module/download/download.jsp?tabIndex=1 on November 13, 2018. Verified allosteric residues were taken from protein entries which were manually checked at http://mdl.shsmu.edu.cn/ASD/module/pathway /pathway.jsp and extracted from the XF.tar.gz file (XML). In the case of discrepancies between displayed residues and listed residues in the XML file, or when SCA was noted as part of the validation process, primary data sources were checked to ensure that an experimental technique had validated that particular residue.

Proteins with over 1000 amino acids were ignored due to size (memory) limitations of contact prediction methods. The resulting set contained 38 proteins where every protein in the set had at least one verified residue. This set was randomly split into training and test sets by ordering the entries by number of verified residues

and taking every second as training data. The resulting distributions of number of verified residues (Figure 2.1) were checked for difference with a Student's t-test. A p-value of 0.63 indicated no significant difference of their means (training set: 4.7, test set: 4.3). A p-value of 0.34 also indicated no significant difference between protein length means (training set: 524, test set: 450). The final lists of the training and test set proteins can be found in Appendix A Tables A.1 and A.2.

Figure 2.1 updated



(a) Protein length **(b)** Number of verified residues

Figure 2.1: Comparison of protein length and verified residue distributions between training and test set. (a) The protein length distributions of training (blue) and test (orange) set are not significantly different. **(b)** The same is true for the distribution of the number of verified residues per protein. This should ensure that the analyses that depend on the ASD data are not biased towards protein length or proteins with high or low numbers of verified residues.

All results regarding ASD were derived from the training set; the test set has not yet been analysed. The data shown in this chapter is for the 17 (out of 19) training set targets that successfully ran through all six co-evolution methods (see Methods section *Contact/coupling predictors* for more information) and contains 86 verified residues in total (PDB sequence lengths between 101 and 699 residues).

2.2.2 Multiple Sequence Alignments and Effective Sequences

Multiple sequence alignments (MSAs) were created using hhblits (Remmert *et al.*, 2012) with default parameters and the Uniclust30 sequence set (hhsuite.tar.gz; for test case MutS: October 2017; for all other analyses: August 2018), downloaded from http://wwwuser.gwdg.de/~compbiol/uniclust/.

In order to gain an approximate measure of the depth of our MSAs, we used the effective sequence measure. This can be defined in many ways. Here we set the count weight of every sequence with similar sequences in the MSA to 1 / number of sequences that share at least X% identity (Ekeberg *et al.*, 2013), using a sequence identity threshold of 90% (Kamisetty *et al.*, 2013).

2.2.3 Co-evolution Analysis

Definitions of contacts

If the C_β-C_β distance (in the case of glycine C_α) of two residues is 8Å or less, those residues are considered to be in contact (as defined by Marks *et al.*, 2011). This is the standard threshold for 'binary' contact predictors. Predicted contacts that were present in the reference crystal structure, were considered true positives.

Trivial contacts/couplings are defined as those between residues that are four or less residues in the protein sequence apart (as used by MetaPSICOV (Jones *et al.*, 2015)).

Contact/coupling predictors

In contrast to statistical coupling analysis (SCA), all newer co-evolution methods used here (PSICOV, CCMpred, EV-Fold, MetaPSICOV2, RaptorX) are transitivity-avoiding methods and therefore all categorised as direct coupling analysis (DCA)-based methods.

MetaPSICOV2 (Buchan and Jones, 2018) was run with default parameters to generate the coupling predictions of PSICOV (Jones *et al.*, 2012), CCMpred (Seemayer *et al.*, 2014) and EV-Fold (Marks *et al.*, 2011) as well as MetaPSICOV2.

RaptorX (Källberg *et al.*, 2012) contact predictions were generated via the webserver on June 6, 2019; with the target sequence as input.

For MutS, uniprot entry B7AA09_THEAQ (Taq MutS) was used as input to MetaPSICOV2. For the analysis of the Allosteric Database, sequence input to hhblits (Remmert *et al.*, 2012) (through MetaPSICOV2 script) and RaptorX were the sequences of the most similar PDB structures to the target uniprot sequence, some

cut to match resolved parts (see Table A.1). For all sequences but KCNAS_DROME, CHEA_ECOLI and RPGF3_HUMAN, there existed a highly similar PDB structure to the target Uniprot sequence. For KCNAS_DROME the homolog KCNA2_RAT (PDB 2A79_B) was used, for CHEA_ECOLI, CHEA_THEMA (PDB 1B3Q_A) and for RPGF3_HUMAN, RPGF4_MOUSE (PDB 4MGI_E). The lowest sequence identity (over the aligned part) was 44% for the CHEA pair.

SCA predictions for MutS were taken from Lakhani *et al.* (2017) which had used SCA 5.0 implemented in MATLAB. For SCA contact predictions for the proteins in the ASD training set, SCA 6.0 as pySCA version 1 was downloaded from https://github.com/reynoldsk/pySCA on February 22, 2019. The hhblits MSAs created by MetaPSICOV2 runs were used as inputs. To compare verified residue recall, SCA was performed up to the SCAcore.py script (Singular Value Decomposition but no sector identification) and the resulting SCA covariation matrix C_{SCA} entries converted into a descending list.

For 17 out of 19 training set proteins all six contact prediction methods ran correctly. KCNAS_DROME and KIT_HUMAN failed due to a memory error occurring in a conditional script that is automatically called by the MetaPSICOV2 script.

Prediction assessment

To compare the results of the contact predictors we examined the top L predictions (L being the sequence length of the protein) as used in de Oliveira *et al.* (2016). PSICOV, MetaPSICOV2 and RaptorX output no trivial couplings (couplings less than five residue apart) while EV-Fold, CCMpred and SCA do. Trivial couplings were removed from predictions unless otherwise stated. Contact precision was calculated using the matching PDB structures (see Table A.1) as true positive predictions / all predictions up to top L (here, L is the number of the resolved residues that were matched to the target sequence).

2.2.4 Network Analysis

Network analysis was performed using Python Igraph (Csardi and Nepusz, 2006) version 0.7.0 on Python 2.7 and eleven node centrality measures as well as one edge centrality (see Table 2.1) were calculated with the embedded Igraph functions with default parameters unless otherwise stated.

Table 2.1: Centrality metrics for network analysis.

Centrality metric	Details
Degree (un-/weighted)	Number of neighbours of node n; sum of edges to neighbours when weighted. (Freeman, 1978)
Betweenness (un-/weighted)	Number of shortest paths between pairs of all other nodes that go through n; edge weights considered for determining shortest paths. (Freeman, 1977)
Edge betweenness	Number of shortest paths of all node pairs that go through edge e. (Girvan and Newman, 2002)
Eigenvector (un-/weighted)	Measure for importance of n in network based on $Ax = \lambda x$ with A being the weighted or unweighted adjacency matrix, eigenvalue λ, and x the eigenvector containing the centrality values for all nodes. (Bonacich, 1972)
Eccentricity	Maximum of shortest paths to all other nodes in the network. (Hage and Harary, 1995)
Diversity (un-/weighted)	Normalised Shannon entropy of the edges of n. (Eagle *et al.*, 2010)
PageRank (un-/weighted)	Measure of importance of n in network based on random walks. (Brin and Page, 1998)

2.2.5 Allostery Validation

MutS

For MutS, node centralities were sorted in a descending order and the top 157 ranked residues of predicted networks were compared to the 157 sector resides from Lakhani *et al.* (2017). These sector residues are considered to be the most important residues of a protein according to SCA and were taken as the nodes of the allosteric network that Lakhani *et al.* predicted. Edge centralities were sorted in a descending order and from the top of these ranked edge centralities downwards, edge per edge, both node residues were taken into a list of unique residues up to the point where more than 156 residues were included, i.e. depending on the last edge taken into account, 157 or 158 residues were in the final list. Overlap between such a centrality-ranked list and the 157 residues from Lakhani *et al.* is called SCA sector residue recall.

Allosteric Database

Allostery validation via the Allosteric Database relies on residues that were experimentally verified to be involved in allostery. If a co-evolution method contains a verified residue in its top X (length L of the protein if not stated otherwise) coupling predictions, the verified residue counts as recalled. For the residue numbers of the verified residues of each protein in the training set see Table A.1 (KCNAS_DROME and KIT_HUMAN were excluded from the analysis, see Methods section *Contact/coupling predictors*).

All top L coupling predictions have 2*L possible positions for residues of which some must occur more than once (because there are only L residues in a protein). In network terminology, the number of couplings per residue can also be called the degree of a residue. If, for a given protein, only few residues have many predicted couplings to other residues (and therefore appear many times in the top L and have a high degree), this leads to a quite different prediction structure from a case where all residues have about two predicted couplings (and hence, appear only twice in the top L). This prediction structure varies from protein to protein and generally differs between co-evolution methods (see Results section *Contact Network Patterns and Random Controls* Figure 2.7). Random controls must take these differences in prediction structures (coupling patterns) into account because the number of unique residues in the top L will vary drastically and therefore the expected random recall of verified residues will also vary.

Verified residue recall was controlled with 100 random pseudo-predictions for each protein and prediction method. An example for this procedure (one prediction of one protein) is shown in Figure 2.2. Randomisations took into account that no self-pairs or trivial residue pairs were considered in the analyses and thus, were not allowed in random controls either.

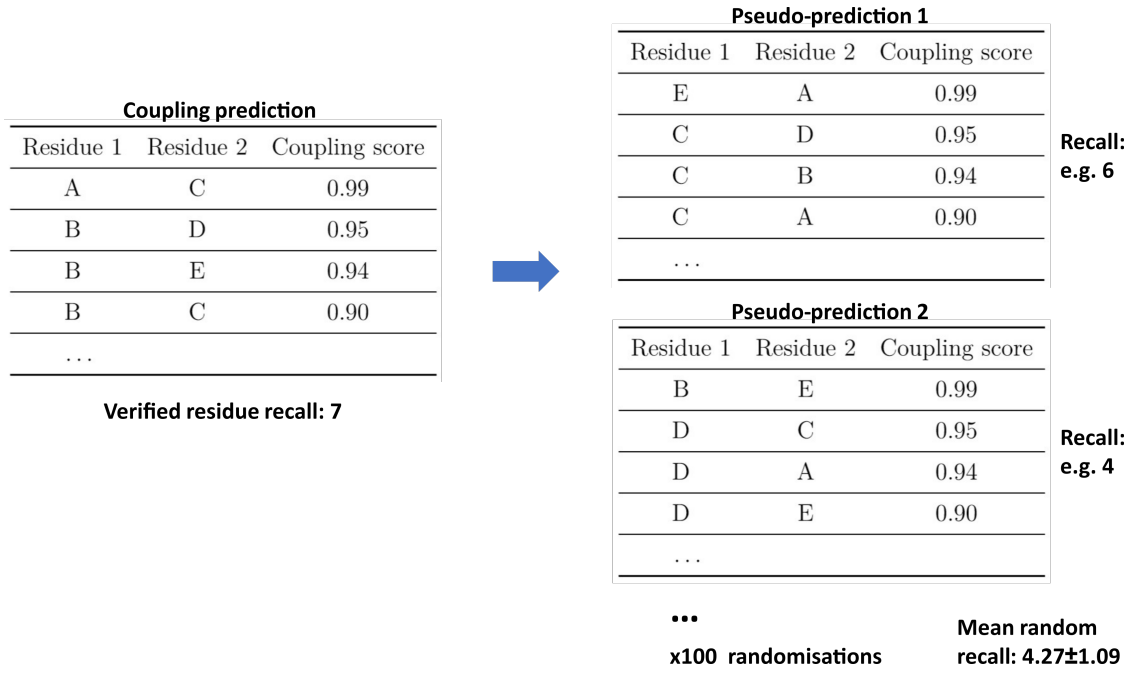


Figure 2.2: Verified residue recall random control example. Four residue pairs of a coupling prediction are displayed here: A-C, B-D, B-E and B-C; this means residue B is present three times, residue C two times and residues A, D and E only once. Random pseudo-predictions with a matching prediction structure are generated so that one random residue is present three times, one two times and three only once. Trivial residue pairs were not formed in pseudo-predictions as they do not exist in the coupling predictions we considered. Such pseudo-predictions were generated 100 times and the mean verified residue recall of these was determined.

2.2.6 Distance Prediction Analysis for Stabilising Residue Pair Approximation

As we show in chapter 3, distance predictions can be informative of the flexibility and stability of residue pairs. In that chapter we demonstrate that having a single local maximum in the predicted distance distribution of a pair of residues is an indicator that these pairs of residues are rigid (i.e. do not move). We used this to identify residue pairs that are likely to be static and potentially stabilising the protein structure in a re-analysis of the Allosteric Database.

Distance predictor

To generate predicted distance distributions for every residue pair in our ASD training set, we used DMPfold (Greener *et al.*, 2019). DMPfold produces distance predictions for all pairs of residues that are five or more residues apart in sequence,

feeds those into a 3D structure modelling programme (CNS), produces updated distance predictions based on the 3D models and then iterates over those two steps another two times. Since we are most interested in distance predictions that are not biased towards a converging 3D model, we only use DMPfold with default parameters but without any 3D model building and updates of distance predictions. Thus, distance predictions used in this work always refer to the initial distance prediction generated before protein structure modelling iterations.

DMPfold was downloaded from https://github.com/psipred/DMPfold on 01/10/2019. It uses following sequence-based input features: sequence profile, mutual information (MI), MI product (MIp), mean contact potential, PSICOV contact scores (Jones *et al.*, 2012), FreeContact (mfDCA) contact scores (Kaján *et al.*, 2014), CCMpred (plmDCA) contact scores (Seemayer *et al.*, 2014), PSIPRED secondary structure, Shannon entropy in multiple sequence alignment columns, SOLVPRED solvent accessibility, log(1 + sequence separation), sequence bounds (channel of ones), DeepCov covariance matrix (Jones and Kandathil, 2018). Those input features for the initial distance prediction were generated with hhblits 3.0.3 (multiple sequence alignment) against Uniclust30 (2018_08).

A distance prediction for a pair of residues refers to the predicted C_β-C_β distance (in the case of glycine C_α) between those two residues and is termed predicted distance distribution here. Each predicted distance distribution is a vector of 20 points, representing the probabilities for each of the distance bins that DMPfold was trained to predict for a given sequence (or multiple sequence alignment). The first bin is ranging from 3.5-4.5Å, followed by seven bins of 0.5Å width up to a distance of 8Å and eleven bins of 1Å width up to 19Å. The last bin contained the probability density for distances over 19Å.

Local maxima analysis

Following our definitions in chapter 3, to define our proxy for stabilising residue pairs, we used all residue pairs that had a single sharp local maximum in their predicted distance distribution. Sharp is defined here as having more than 0.5 of

probability mass within ±1Å around the local maximum. For defining the distance of a probability maximum, the distance was set to a bin's centre; for calculating probability masses in the interval around a maximum, the probability masses were taken proportionally if an interval border was not at a bin border. To detect the number of local maxima, distance probability distributions (predicted distance distributions) were analysed with the 'peakdet' function from Eli Billauer, version 3.4.05 (downloaded from https://gist.github.com/endolith/250860 on 18/10/2019). The function considers a point a local maximum if it has the local maximal value, and was followed (within steps towards the last bin) by a value lower by at least DELTA, which had been chosen to be 0.03 in our analysis of residue pair flexibility. Detecting a local minimum (analogue definition with greater-by-at-least delta) resets the local maximum. The last bin is never considered to be a maximum.

Stabilising residue pair removal

We use the distance predictions generated by DMPfold to identify residue pairs that are likely to be static and hence not involved in allostery. Each residue pair with a single sharp local maximum in its predicted distance distribution was considered to be static and potentially stabilising rather than being involved in allosteric signalling. Therefore, such residue pairs were deleted from the list of all contact predictions to enrich residue pairs that could be involved in allostery. This was done for each protein and all predictions that were generated by the six different contact predictors. From the remainder of contact predictions, a new list of top X (L, L/2, L/5; L being the sequence length; rounding to the closest integer) was computed and used for re-calculating the verified residue recall.

2.3 Results

2.3.1 Contact Networks of MutS

Statistical coupling analysis (SCA), one of the earliest co-evolution methods, was used to predict an allosteric network for MutS (Lakhani *et al.*, 2017). We compared the network derived from SCA with networks derived from more recent co-evolution

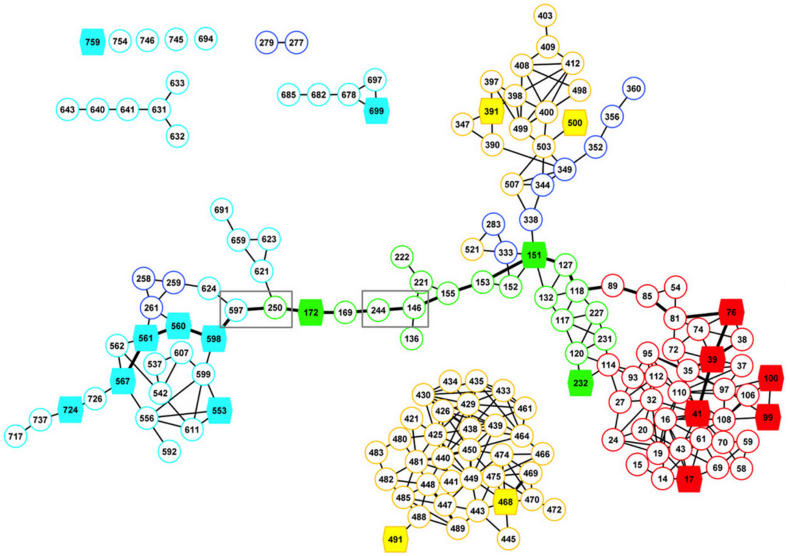
methods. Several direct coupling analysis (DCA) methods were run and protein residue networks were derived from the (contact) predictions.

The MutS protein recognises DNA mismatches and after conformational change of the ATP-binding domain and a resulting change of affinity for ATP, it recruits other mismatch repair proteins. As can be seen in Figure 2.3b, MutS is a homodimer, where each monomer consists of five domains. The depicted residue network in Figure 2.3c represents the PDB structure as a network without trivial contacts and thus, the ground truth for most contact predictors.

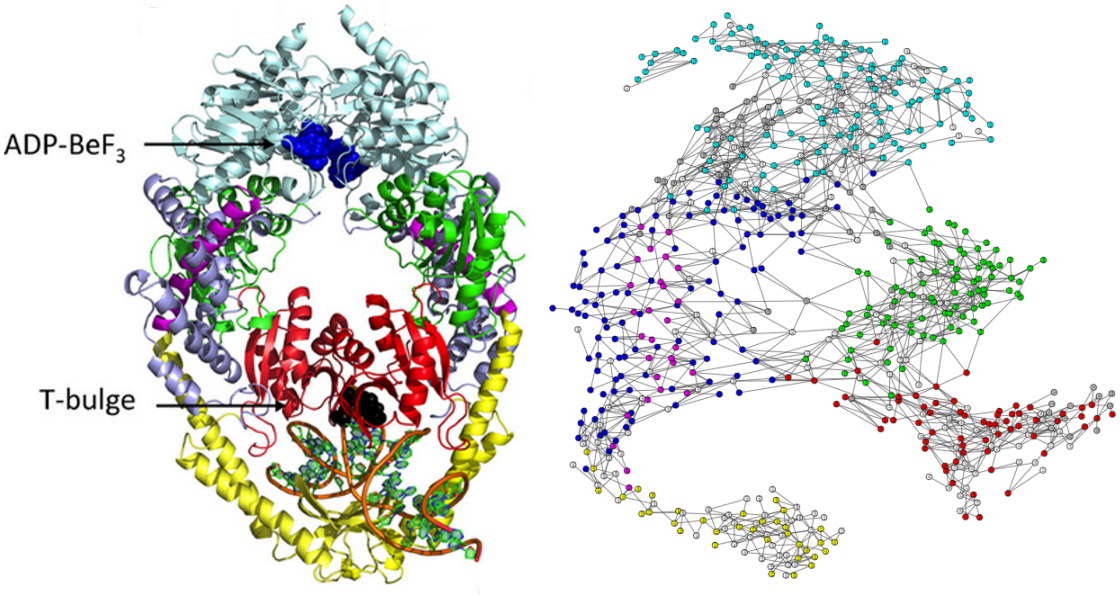
Figure 2.3 updated

Lakhani *et al.* (2017) derived their allosteric network (see Figure 2.3a) from co-evolution data by creating nodes from the top 20% correlating residues according to SCA, called sector residues, and creating edges between them when their minimal heavy atom distance was below 6.8Å in the known PDB structure after considering only heavy atoms that are not shielded by other atoms (shadow map algorithm (Noel *et al.*, 2012)). We initially tested more recent co-evolution techniques without additional structural information to see if they could recapitulate this network.

All co-evolution analysis methods use multiple sequence alignments (MSAs) to extract correlation patterns. A subset of the inferred correlations are then used to define network edges. Depending on the analysis method the subset of predicted positives is defined by a different threshold. SCA takes into account the top ∼20% correlating residues (Lakhani *et al.*, 2017) whereas newer methods such as EV-Fold or CCMpred restrict their subset to top L (where L is the length of the protein), or a fraction of L, residue pairs. Predictors that compute a probability for each residue-residue coupling, e.g. MetaPSICOV2, can use a probability threshold. All of these thresholds are arbitrary and up for exploration but using a common threshold allows better comparison amongst the methods (top L is most commonly used (de Oliveira *et al.*, 2016)). In the following sections we describe comparisons to the results of the MutS study (Lakhani *et al.*, 2017) using the top L residue couplings from four DCA prediction methods (PSICOV, CCMpred, EV-Fold and MetaPSICOV2).



(a) SCA-derived allosteric network



(b) MutS crystal structure

(c) Structure-derived amino acid network

Figure 2.3: MutS representations coloured by domains. (a) Amino acid network derived from SCA sector residues by Lakhani *et al.* (Lakhani *et al.*, 2017) **(b)** MutS dimer with DNA, including mismatch (T-bulge) and ATP analogue (ADP-BeF_3); PDB 1NNE, figure and colouring by Lakhani *et al.* (Lakhani *et al.*, 2017) **(c)** Amino acid network of true, non-trivial contacts derived from PDB 1NNE_A. It is assumed that a signal is passed from the red (T-bulge recognition) domain to the cyan (ATP/ADP binding) domain.

Multiple sequence alignment

Lakhani *et al.* built their MSA using a manually-curated sequence alignment of six homologous sequences (two prokaryotic, four human) as a reference and then

aligned the rest of their collected homologous sequences producing an MSA with 164 sequences. Sequences over 90% identity were removed, yielding 142 effective sequences (N_{eff}). As a greater number of effective sequences in an MSA usually allows more precise contact predictions (Marks *et al.*, 2012; de Oliveira and Deane, 2017), we opted to build a new MSA using an hhblits search (Remmert *et al.*, 2012) generating an alignment that yielded an N_{eff} of 8956 (see Methods *Multiple Sequence Alignments and Effective Sequences* for more details).

Contact predictors

SCA is known to yield low precision in predicting physical proximity/contacts (see *Introduction / Co-evolution Analysis* and Table 2.4). Its contact prediction precision is more in the order of single or low double digit precision (in top L predictions) as opposed to around 50% for DCA methods. Under the assumption of allosteric signal transmission through a network of contiguous residues, using more precise contact predictors becomes obvious, especially with the goal of using sequence information alone.

The precision of the top L contact predictions (with respect to true contacts in the MutS structure 1NNE_A) of the four different DCA contact predictors is shown in Table 2.2 (see Methods section *Co-evolution Analysis* for definitions of predicted and trivial contacts, as well as prediction assessment).

Table 2.2: MutS: precision values of tested contact predictors.

Top L precision [%] for	PSICOV	CCMpred	EV-Fold	MetaPSICOV2
All contacts	-	82.74	92.23	-
Only non-trivial contacts	50.33	57.97	42.42	69.95

As can be seen from the differences in precision of CCMpred and EV-Fold for predictions including and excluding trivial contacts, correct prediction of trivial contacts is easier than for non-trivial, long-range contacts. We tested whether creating networks with and without such trivial contacts has an influence on the overlap between networks from SCA and DCA. MetaPSICOV2 performs best on

the prediction of non-trivial contacts with almost 70% precision, whereas the other three predictors reach between 42 and 58% precision. These precision scores were only calculated with a single structure which provides a good estimate but omits some contacts that are only present in a subset of protein conformations (Anishchenko *et al.*, 2017) or co-evolved residue pairs due to protein-protein interactions (de Oliveira and Deane, 2017). Furthermore, these scores give no indication as to the functional importance of a particular interaction (or contact), which might be crucial for specific applications (such as allosteric network prediction).

Crystal structure-based contact networks

After running the SCA algorithm, Lakhani *et al.* used the top 157 (top ∼20%) correlating residues, out of all 759 MutS residues, as nodes in their network (Figure 2.3a). They then used a crystal structure (PDB 1NNE_A) to draw connections (edges) between these residues (placing edges between residues that were less than 6.8Å apart in the PDB structure). Figure 2.3 shows the crystal structure of the asymmetric homodimer with mismatch DNA and an ATP analogue bound (Figure 2.3b) and next to it the ('true') network of non-trivial contacts of this crystal structure's chain A (Figure 2.3c). Nodes are all 759 residues of the crystal structure and edges were set according to the contact definition (non-trivial only) used by all common predictors (C_β-C_β distances 8Å or less, C_α for glycine). The yellow domain (domain IV) is only connected to the blue domain (domains IIIa) which is connected to all other domains. The purple domain (domain IIIb) is connected predominantly to the blue domain but with a few connections to the yellow domain. The red domain (domain I) is most strongly connected to the green domain (domain III) but also to the blue. The green domain, similar to the blue domain, sits in the middle of the network and is connected to the red, blue and cyan domain (domain V). As the DNA mismatch is recognised by the red domain and the ATP-binding site is located in the cyan domain, the hypothesis for MutS is that a continuous allosteric pathway exists between these two domains.

Whether trivial contacts are crucial in the allosteric signal transmission is an open research question. Since some of the contact predictors (e.g. MetaPSICOV2) only predict non-trivial contacts, we also worked with purely non-trivial residue networks which, in principle, are able to connect all domains to a continuous network (Figure 2.3c). Note a disconnected small cluster of the cyan domain; this represents a terminal α-helix that is only connected to the other residues by trivial contacts.

Predicted contact networks

The first question resulting from working with contact prediction networks is, if they are able to reasonably connect all the domains of MutS. Figure 2.4 shows two networks derived from predicted contacts, one from MetaPSICOV2 (Figure 2.4a) and one from EV-Fold (Figure 2.4b). The network predicted by MetaPSICOV2 contains no trivial contacts and is able to connect all domains into one network. Most of the yellow domain is separated and forms a second cluster which can also be seen in the network predicted by the MutS SCA method (Figure 2.3a). The network predicted by EV-Fold also resembles the general protein structure but does not form a continuous network across all domains. Inter-domain contacts between the red and the green and blue domains are missing as well as intra-domain contacts within the blue domain.

A continuous network is crucial if you assume that an allosteric signal is passed continuously through a network of protein residues (in this case from the red to the cyan domain, see Figure 2.3). We therefore investigated several different approaches to yield networks with more connections, preferably continuous. One way of doing this is including all trivial contacts between residues that were predicted to be in a (non-trivial) network. Another way is adding these trivial contacts but only if they lead to a reduced cluster count, i.e. a larger continuous network. And indeed, in the case when the contact network is derived from CCMpred top L predictions (Figure 2.5a), these additional edges connect the previously disconnected red and green domains with the rest of the network (Figure 2.5b).

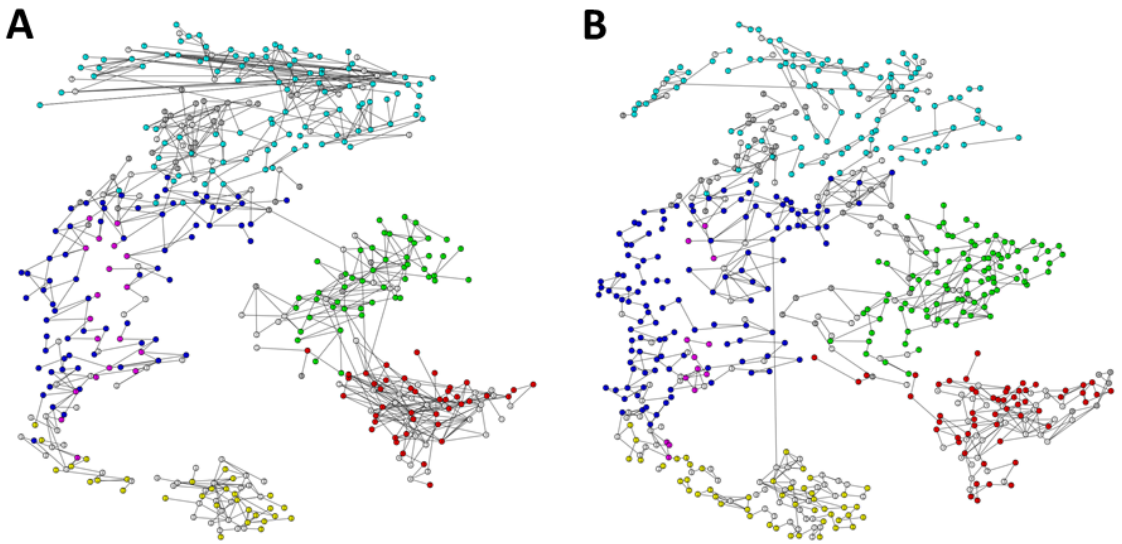


Figure 2.4: Amino acid networks derived from top L contact predictions. a) derived from MetaPSICOV2 and **b)** derived from EV-Fold. Whereas the MetaPSICOV2 network connects all domains, the EV-Fold network does not. In fact, a very long range connection between the yellow and the blue domain is visible (long vertical edge) but is not a true contact.

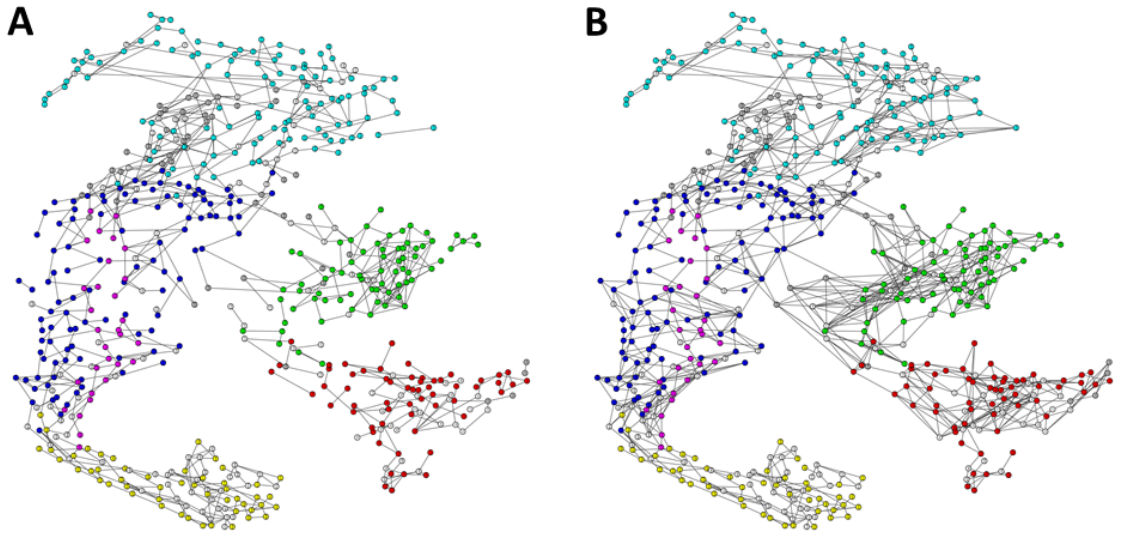


Figure 2.5: Amino acid networks derived from top L contact predictions of CCMpred. a) Network solely by contact predictions. **b)** Network by contact predictions with trivial contacts added if this connected a cluster to the main network. The addition of trivial contacts connects the red and cyan domain (via green and blue) which is hypothesised to be the allosteric signalling route. This points out the potential of this trivial contact addition step.

The biggest difference between the MutS SCA network and the networks predicted by newer co-evolution analysis methods is the number of residues, e.g.

157 (SCA) vs. 457 (MetaPSICOV2). While the SCA network is defined to contain about 20% of all residues, DCA methods are usually analysed in terms of their top L pairwise predictions, and thus have no fixed number of residues in a network derived from that. Theoretically, all residues of a protein could be included in the top L predictions (as L is the sequence length). In order to compare those networks of different sizes, a measure must be implemented that allows the ranking of residues to be able to generate sets containing the same number of residues.

Network centralities are measures that can be calculated for every node or edge in a network and assign a value to a node or an edge. A simple example would be the degree centrality. A node's degree is defined as the number of edges that connect a node to its neighbouring nodes. A slightly more complex example would be the strength centrality which is the same as the degree but every edge does not count as one but as its edge weight. A node that is connected to more nodes by higher edge weights could be considered as a (more) important node in a network and would have a higher score. Centrality measures can therefore be used to rank the importance of residues inside a network and allow us to compare the 157 sector residues derived from SCA to the larger networks from the DCA-based contact predictors.

Overlap of centrality-ranked DCA and SCA residues

In order to rank the residues of any predicted contact networks we calculated several different centrality measures (see Table 2.1). Figure 2.6 shows networks where node size corresponds to centrality (the bigger the node, the bigger its centrality value). Both networks were derived from the top L predictions of MetaPSICOV2. The 'degree' network (Figure 2.6a) shows that the highest degree nodes are found in the red and green domains whereas the 'betweenness' network (Figure 2.6b) shows that the highest betweenness nodes are found connecting the green and blue domains and within the red, green, blue and cyan domains.

Since it is hypothesised that there exists an allosteric pathway between the red and the cyan domain and the ranking by betweenness centrality suggests such a pathway from red to cyan, this could be a useful measure to select allosteric

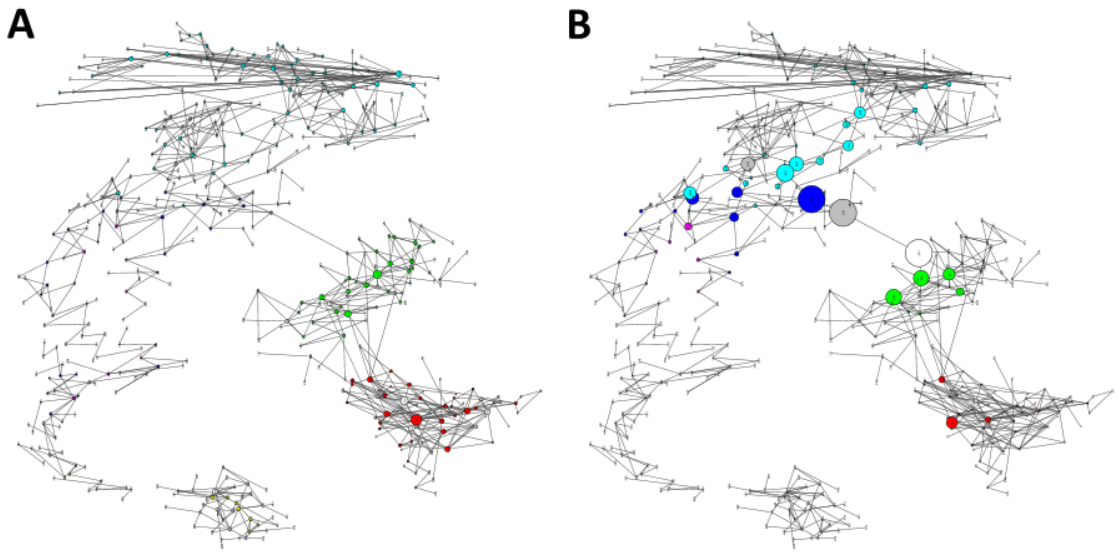


Figure 2.6: Centrality networks derived from top L contact predictions of MetaPSICOV2. Node sizes scaled by centrality values. **a)** Degree centrality. **b)** Betweenness centrality. Centrality measures were calculated to compare DCA networks (e.g. MetaPSICOV2) to the SCA network from Lakhani *et al.*. Residues also present in the SCA network are coloured in white. Only one of the bigger nodes is white, emphasising the general result that the overlap between the DCA networks and the SCA network is generally low.

residues. Residue nodes coloured in white are present in the SCA network. Only one of the larger nodes is coloured in white which means that the pathway suggested by the centrality measure is not the same as the one suggested by the MutS SCA methodology.

It is assumed that some nodes within an allosteric network are crucial for the signal transmission (Sol *et al.*, 2006) but if any centrality measure relates to this, is unclear. Therefore, we calculated a variety of centralities which capture either the number of connections to neighbours or its occurrence in potential pathways within a network. Some centralities allow edge weights to be used in their calculation. In total, we calculated weighted and unweighted scores for the centralities betweenness, strength (unweighted strength is the degree), eigenvector, diversity and pagerank; also the weighted centralities edge betweenness and eccentricity (see Table 2.1).

After ranking all residues according to each centrality score, we determined the number of common residues of the top 157 ranked residues of predicted networks and the 157 sector resides of the SCA method. Table 2.3 shows these recall numbers.

Table 2.3: SCA sector residue recalls by different centrality measure rankings and networks derived from different predictors (or a crystal structure).

SCA sector residue recall by centrality measure (top 157 residues)	Network									
	Crystal structure (1NNE_A)		PSICOV top L		CCMpred top L		EV-Fold top L		MetaPSICOV2 top L	
	all contacts	w/o trivials	-	trivials added	-	trivials added	-	trivials added	-	trivials added
Number of network nodes	759	660	550	550	599	599	619	619	457	457
Number of network edges	3859	1698	759	1032	759	1261	759	1248	759	904
Maximum recall of network	157	150	132	132	122	122	136	136	110	110
Betweenness	44	46	43	48	31	29	40	37	28	35
Betweenness unweighted	44	46	43	43	32	30	38	43	28	32
Edge betweenness	45	45	43	48	31	32	42	37	25	33
Degree (=strength unweighted)	39	41	47	50	27	35	42	34	37	40
Strength	39	41	41	42	26	35	36	34	38	39
Eigenvector	28	31	62	60	18	29	43	44	39	38
Eigenvector unweighted	28	31	62	48	26	34	46	24	39	38
Eccentricity	46	32	41	29	22	38	36	45	43	55
Diversity	40	51	40	45	33	33	34	48	41	40
Diversity unweighted	40	50	48	48	47	47	49	49	48	48
Pagerank	47	40	38	38	33	33	39	43	39	40
Pagerank unweighted	47	40	37	45	30	31	35	40	40	44

The overlap was generally low between these sets of residues. Nevertheless, the comparison of SCA and DCA networks revealed two things. Firstly, different co-evolution methods derive (very) different networks of residues and to improve sequence-based methods for allostery prediction, it will be crucial to discriminate between different influences on evolutionary constraints, e.g. folding, stability and functionality (Dokholyan, 2016).

And secondly, due to the general lack of understanding of the underlying mechanics of allosteric signal transmission, allosteric networks have not yet been rigorously validated. This is also the case for the networks that were derived for MutS using SCA. The SCA-predicted full size allosteric network of MutS has been validated by performing molecular dynamics (MD) alanine mutation studies. In those studies, individual residues were mutated *in silico* to alanine and the MutS mutant's behaviour was then simulated for 15ns with MD. MD trajectories were analysed and different parameters were measured to investigate potential impact on allosteric signalling in MutS: domain RMSDs, occurrence times of important H-bonds from MutS to ATP and DNA as well as the docking energy of ATP in several MD trajectory snapshots. These MD mutation studies were performed on the 21 residues that SCA predicted to be the most correlating and a control group of the ten residues predicted to be least correlating (not part of the predicted allosteric network) (Lakhani *et al.*, 2017). Those residues are spread across the network and it was shown that 19 out of 21 of the tested residues caused a disruption of

the proposed allosteric signal (this was the case for only 2 out of 10 residues of the control group). But the relatively short simulation times (15ns) and the three metrics employed (domain RMSD, H-bond occurrence time, docking energy) do not necessarily indicate a residue's importance for allosteric signalling. The residues chosen were also biased in terms of characteristics between the two sets. The most correlating residues as defined by SCA were mostly buried whereas the selected control residues were mostly on the protein surface.

In order to overcome the limitations of the MutS study, we next focused on evaluating co-evolution methods using a different strategy that included more experimental validation and less bias towards specific methods or proteins.

2.3.2 The Allosteric Database

Following our initial comparisons of MutS networks we used the Allosteric Database (Shen *et al.*, 2016) (ASD) to validate allosteric network predictions in a more unbiased way. Our training set contained 17 proteins with 86 verified residues (allosterically important residues) in total (see Methods section *Datasets*).

Contact network patterns and random controls

The top L predictions (and thus, the resulting networks), of the co-evolution methods used here, have very different structures. The number of predicted residues (nodes) and couplings (edges) and thus, the mean degree per residue varies between DCA methods and SCA (Figure 2.7a and b). SCA generally predicts a subset of residues with lots of connections whereas DCA predicts fewer of these high-degree residues. We controlled for these differences using random pseudo-predictions that match the degree distribution of a given predicted network (see Methods section *Allostery validation* for more details). Such random pseudo-predictions allow an estimate as to the verified residues that can be randomly expected in a prediction.

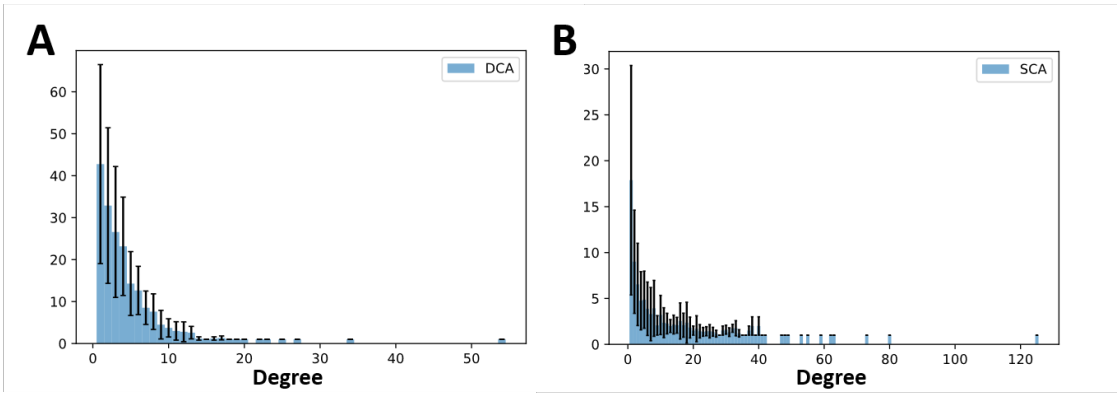


Figure 2.7: Mean degree distributions of DCA and SCA. a) DCA (here: MetaPSICOV2) and **b)** SCA: mean degree histograms from all training set top L predictions. The mean degree distribution of top L predictions represents the prediction structure of a respective co-evolution method. The longer tail of the SCA distribution shows that SCA predicts more nodes with higher degree than DCA. This implies that less unique residues are present in the top L coupling predictions of SCA.

Table 2.4: ASD: verified residue recall and contact precision of tested co-evolution methods.

	Verified residue recall	Random control recall (mean)	Random control recall (std)	Contact precision (mean) [%]	Contact precision (std) [%]
CCMpred	70	57.49	3.99	48	20
PSICOV	68	57.96	4.10	41	20
EV-Fold	61	55.42	4.00	33	17
MetaPSICOV2	55	41.09	4.36	61	20
RaptorX	64	49.33	4.24	78	19
SCA	41	20.52	3.79	4	3

Contact prediction precision and verified residue recall

Networks were generated using the top L predictions for six different co-evolution methods and compared to the verified residues from the ASD (Table 2.4).

None of the methods recalls all 86 verified residues. The highest recall is achieved by co-evolution techniques without machine learning; 70 by CCMpred and 68 by PSICOV although their difference to random controls is relatively small. The difference to random controls is slightly larger for the two methods which include machine learning, MetaPSICOV2 and RaptorX, and substantially better for SCA. But, SCA recalls the least residues in total (41). Both methods that include machine learning use structural information about protein structure (e.g. secondary structure

patterns in contact maps). Hence, their contact predictions are more precise, but it has been shown that this is often due to prediction of contacts in secondary structural elements (Chonofsky *et al.*, 2019) which are likely to be less informative about a potential evolutionary footprint from allostery.

Verified residue recall after removal of stabilising residue pairs

Increased precision in contact prediction appears to not increase information content on allostery but rather decrease it. This led us to the conclusion that co-evolutionary signal should be analysed taking into account different reasons for co-evolutionary coupling. For example, even if all couplings from DCA stem from physical proximity (Anishchenko *et al.*, 2017), it is still unclear if the physical interaction is due to stability constraints or due to being part of an allosteric communication channel or both. We hypothesised that subtracting stability constraints from co-evolutionary data might lead to an increase in allosteric residue recall.

Our work in chapter 3 showed that the co-evolution method DMPfold (Greener *et al.*, 2019), which predicts inter-residue distances instead of binary contacts, can be used to indicate likely rigid residue pairs. DMPfold outputs a predicted distance distribution for all non-trivial residue pairs (see chapter 3 for more details). We showed that flexible residue pairs more often had multiple local maxima in this predicted distance distribution and rigid residue pairs more often just a single local maximum. Given that for allostery prediction we are interested in signal-transmitting (potentially flexible) residues, we examined whether a single local maximum could be a good proxy for a stabilising (non-flexible) residue pair to create a model which accounts for stability as well as allosteric signal transmission being possible constraints in co-evolutionary data. Hence, we re-computed verified residue recall for our ASD training set removing predicted rigid (stabilising) residue pairs. All residue pairs with a sharp single local maximum in their predicted distance distribution were removed from the predictions of each of the six co-evolution methods. This resulted in new sets of top X predictions for each method and therefore different verified residue recalls and random controls.

The verified residue recalls for top L, top L/2 and top L/5 predictions are shown in Figure 2.8. For each method the original recall is shown left and the recall after rigid residue pair removal (labelled 'no1s') right. Blue diamonds mark the verified residue recall achieved by a certain method and red diamonds with error bars display the mean verified residue recall and its standard deviation for the random controls. The random controls give an impression of how many verified residues would appear by chance in the top X (for a given output by a predictor).

All methods improve their verified residue recall when removing (predicted) rigid residue pairs from the contact predictions before re-calculating recall. Matching the methods' general order of precision scores, the two machine learning contact predictors MetaPSICOV2 (labelled 'metapsicov') and RaptorX experience the highest increases in recall, followed by the three non-machine learning DCA methods (CCMpred, PSICOV and EV-Fold) and SCA coming last, improving only minimally. RaptorX and non-machine learning DCA methods seem to be similar, with RaptorX achieving the highest recall (74 out of 86) of all methods when considering the top L residue pair predictions, and the non-machine learning predictors scoring the highest in top L/2 and top L/5 analyses. MetaPSICOV2 shows the greatest relative improvements of all methods but in absolute numbers recalls slightly fewer verified residues than the other DCA methods.

Increased recalls from machine learning methods when removing residue pairs with sharp single local maxima in their predicted distance distributions could be expected. As mentioned in the previous section, higher precision also comes with predicting 'easier' contacts more at the top (Chonofsky *et al.*, 2019). The fact that the machine learning methods improve most when removing residue pairs with more concentrated probability masses in their predicted distance distributions supports the view that those 'easier' residue pairs contain less information on allostery.

2.4 Discussion

The work of Lakhani *et al.* (2017) indicated that there is information about allostery in contact prediction networks but their method used several arbitrary parameters

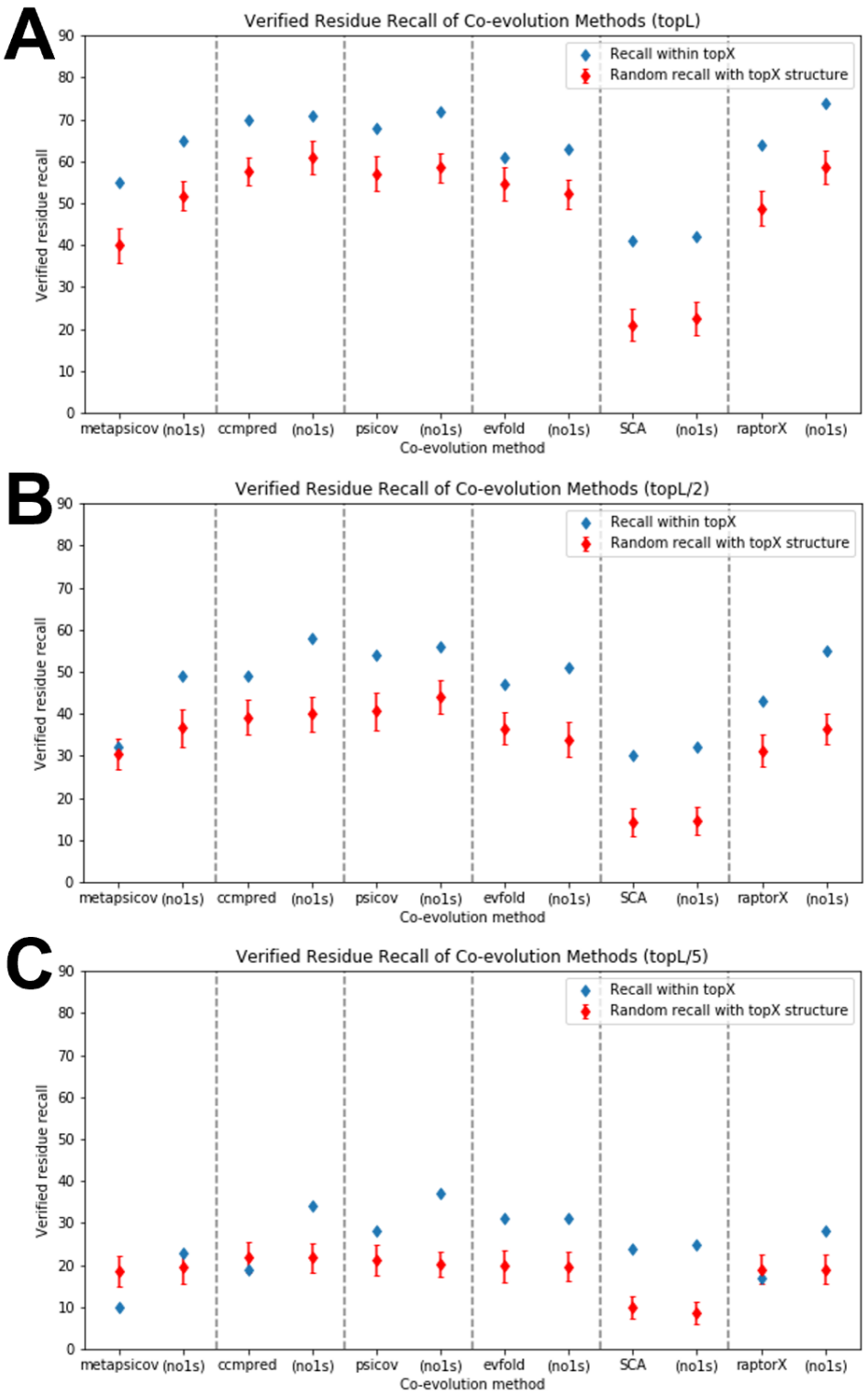


Figure 2.8: Removal of stabilising residue pairs improves verified residue recall. The label 'metapsicov' stands for MetaPSICOV2, 'no1s' refers to the analysis after stabilising residue pair removal. **a)** Improvements in verified residue recall are the largest for machine learning methods (MetaPSICOV2 and RaptorX) but increase for all other methods as well. **b)** Top L/2 and **c)** top L/5 predictions show a similar picture as top L predictions with more prominent improvements for non-machine learning DCA methods.

and had been tested on only one system (MutS; 6.8Å distance cutoff, 20% residues as sector residues, validation was done on 21 residues and 10 control residues and only 15ns of MD per mutation study). Here we aimed to evaluate their allosteric network prediction pipeline and suggest improvements by using more recent co-evolution analysis methods, potential removal of crystal structure dependence and automation using centrality measures.

The first improvement in co-evolution analysis was the creation of a far larger multiple sequence alignment: instead of 142 effective sequences, the use of hhblits against a recent protein sequence database (Uniclust30 from October 2017) yielded an N_{eff} of 8956. This alone should provide a more precise contact prediction (Morcos *et al.*, 2011; Marks *et al.*, 2011).

The next alteration of the pipeline that we explored was the underlying contact predictor. The weakness of statistical coupling analysis (SCA) is transitivity. If two residues are correlated (co-evolving) each with a third residue, they can appear to be correlated as well, even though they are not interacting themselves. This false correlation is a transitive relation and leads to false positive contact predictions (Marks *et al.*, 2012). Direct coupling analysis (DCA) corrects for these transitive correlations which substantially improves contact prediction. Therefore, we created networks from predictions of PSICOV, CCMpred, EV-Fold and MetaPSICOV2 which are all DCA-based and thus transitivity-avoiding methods that have higher precision (de Oliveira *et al.*, 2016) than SCA (see Table 2.4).

The contact networks derived from SCA and DCA methods yield substantially different network patterns in terms of the number of nodes and edges. Whereas the SCA network, per definition (Lakhani *et al.*, 2017), contains only about 20% of the protein's residues, the DCA-based methods predict fewer nodes (residues) with a high number of edges (compare Figure 2.7). This results in DCA-derived networks having almost as many nodes as there are residues in a protein. To distinguish crucial from non-important residues and to make SCA and DCA networks comparable, a ranking of residues becomes necessary. We calculated several centrality measures and ranked residues by their scores. We subsequently compared, for each of the

centralities, the top 157 residues with the 157 sector residues of the Muts SCA method. Depending on the centrality measure and network, a recall of shared residues between 18 and 62 could be achieved (see Table 2.3).

The top L MetaPSICOV2 prediction, ranked by betweenness (see Figure 2.6b), highlights the desired continuous path from the red to the cyan domain of MutS and it also separates a large part of the yellow domain which we assume to be not involved in signal transmission (between the red and cyan domains). But even in this example, we only observe a small overlap with the residues found by the MutS SCA methodology (recall 28, see table 2.3). This raises the question of how useful it is to test against this fixed set of 157 residues, where only 21 have been reportedly analysed for their importance in allostery. We therefore tested our and their methodologies against a larger set of experimentally validated allosteric proteins and residues.

The ASD contains various datasets relating to allosteric sites and interactions and also one subset on allosteric networks. We used this subset as a more general validation set for our comparisons of contact predictors. After selecting training and test sets by similar proteins lengths and number of verified allosteric residues, we generated networks for the training set from the top L coupling predictions of SCA, CCMpred, PSICOV, EV-Fold, MetaPSICOV2 and RaptorX. The achieved contact precisions (see Table 2.4) matched the literature values, implying that the set of proteins is not particularly biased towards one of the methods.

The highest verified residue recall is achieved by CCMpred, a DCA-based method without additional machine learning, whereas SCA recalls the least verified residues but improves most over random. Thus, for SCA the enrichment within the top L predictions is higher than in other methods. SCA is thought to predict functionally-important residues as opposed to correct contacts (Anishchenko *et al.*, 2017) and experimental validation might have been biased towards such generally functional rather than just signal transmission residues. In other words, in a contiguous allosteric network some allosteric residues will be closer to a functional site and others are further away and thus potentially harder to detect, in particular for SCA.

The methods with the best contact predictions, MetaPSICOV2 and RaptorX, outperform the other methods as they use co-evolutionary information alongside structural information learned from the PDB, e.g. secondary structure patterns. But, as their (verified residue) recall values show (see Table 2.4), they might not be best at predicting allosteric residues directly. This led us to two preliminary conclusions. First, co-evolution methods that focus on explaining correlation within an MSA and avoiding transitive couplings (early DCA methods) should be more likely to detect a potential evolutionary footprint from allostery than more recent co-evolution methods that perform better at predicting contacts. And second, reasons for co-evolutionary coupling differ and will have an impact on allosteric residue recall. We hypothesised that stability and potentially allosteric signal transmission could be better predicted when their constraints on co-evolution could be separated. Therefore, we used observations from our work on residue pair flexibility (see chapter 3) to analyse allosteric residue recall when stabilising residue pairs were removed from co-evolutionary contact predictions. More precisely, as a proxy for stabilising residue pairs we used residue pairs with only a single local maximum in their predicted distance distribution generated by the more recent co-evolution technique DMPfold which predicts inter-residue distance distributions instead of binary contacts (Greener *et al.*, 2019).

All contact predictors enhanced their verified residue recall when these predicted stabilising residue pairs were removed. As expected, the methods using machine learning improved the most as they tend to rank residue pairs higher that are part of secondary structural elements (Chonofsky *et al.*, 2019) which are assumed to be more stabilising than residue pairs that are part of loops. And with fewer residue pairs whose co-evolutionary prediction is driven mainly by stabilising constraints, other residue pairs reach the top of the prediction list, which in turn increases recall of allosteric residues. Although recall increases were smaller than for machine learning methods, this is also the case for non-machine learning DCA methods (CCMpred, PSICOV, EV-Fold) highlighting the potential of this approach to using co-evolution data for allosteric network prediction. A methodology for not removing,

but re-scoring residue pairs based on quantitative information on their stability constraint would be the obvious next step.

The main difficulty in the creation of an allostery predictor is the validation of predicted networks as there is very little experimentally validated data that a given residue (or set of residues) is involved in allostery. Even in our validation set derived from the Allosteric Database only 86 residues from 17 proteins could be used to validate against. This lack of data is the key issue for developing a method that aims to be a general tool for allostery research. This is especially true for the prediction of allosteric networks which requires not only the experimental validation data of individual residues but data for almost all residues of a target protein. We believe two recently developed experimental methods might fill this gap in the future.

Deep mutational scanning (DMS) screens the fitness of a large amount of single and double mutants of a protein and assigns a dependency score to each pair of residues, similar to the residue pair coupling in co-evolution analysis (Fowler and Fields, 2014). The method uses high-throughput sequencing coupled to a functional assay to determine the fitness (with regard to the assay) of each mutant in the batch. Whereas the functional assay can be an assay selecting for allosteric signal transmission (Bandaru *et al.*, 2017), the assay has to be designed for each protein of interest individually. Furthermore, the application of this technique is resource intensive as the number of double mutants that need to be tested grows quadratically with protein length. Thus, DMS has the potential to deliver allosteric network data but has not yet produced a large number of datasets for validation.

The second method that has the potential to yield allosteric network data is multi-temperature crystallography (Keedy *et al.*, 2015b; Keedy, 2019). Here the aim is to solve the structures of a protein at different temperatures to generate structural data from different energy regimes ranging from cryogenic up to room temperature. This has successfully observed a temperature-dependent shift of relative population sizes of different conformers of an allosteric protein (Keedy *et al.*, 2018). If this technique was combined with a multi-crystal fragment screening approach (Collins

et al., 2017; Keedy *et al.*, 2018), the methodology could potentially be used to directly examine conformational change due to an allosteric binding event.

As DMS and multi-temperature crystallography have only produced few datasets yet, we could not yet use them for validation. The majority of known allosteric events have been closely linked to conformational change and thus, the flexibility or at least the different conformations of a protein can be extracted from structures deposited in the PDB. This means that a large amount if this conformational flexibility data is available. In the next chapter, we examine if methods that predict inter-residue distance distributions such as DMPfold (Greener *et al.*, 2019) are able to reveal anything about the flexibility of residue pairs in a protein.

3

Co-evolutionary Distance Predictions Contain Flexibility Information

Contents

3.1 Background

The work described in this chapter has been published in *Bioinformatics*, Volume **38**, 65–72 (2021) as Schwarz *et al.* (2021).

As described in the previous chapter there is currently only a small amount of experimentally verified data that explicitly describes allostery. Due to this lack of experimental data directly related to allostery, we decided in this chapter to use data that is more readily available and closely linked to allostery. Although not all allosteric events cause structural changes, a majority of observed events do. If structural changes occur, a good model for describing allostery is a change of activity due to a change of population sizes in the conformational ensemble of a protein. The shifting between different conformations within the conformational ensemble requires interaction changes of individual residue pairs which we term residue pair flexibility.

Whether information on residue pair flexibility is contained in co-evolutionary distance predictions is the subject of this chapter. For a small number of proteins it had already been shown that false positives of binary contact predictors were true positives in a different conformation of the protein (Morcos *et al.*, 2013; Toth-Petroczy *et al.*, 2016) but it had not been analysed if co-evolutionary distance predictions are able to capture this flexibility instead of misclassifying it. If information on interaction-changing residue pairs can be extracted with co-evolution analysis, it would indicate that co-evolutionary data could not only be useful for extracting allosteric information but also for predicting multiple biologically relevant conformations of a protein.

Drawing information on flexibility out of sequence information could be particularly useful for conformational ensemble prediction because current predictors all rely on structural input (see *Introduction / Protein Flexibility & Conformational Ensembles*). Even if a solely sequence-based prediction of protein ensembles was not possible, extracting flexibility information from co-evolutionary data and adding this to existing tools would be beneficial as well.

Co-evolution methods have been used in static protein structure prediction for about ten years and improved from predicting residue contacts to inter-residue

distances and now full-scale models are predicted directly with further advances in neural network architectures. The work described here analysed distance predictions that were generated by DMPfold (Greener *et al.*, 2019). Figure 3.1 shows an example of predicted distance distributions grouped into four different sets based on the number of local maxima that we detected in the distribution.

We used distance predictions because they promise a higher information content than binary contact predictions that only have one probability or score assigned to each residue pair. For those binary contact predictors (Morcos *et al.*, 2011; Marks *et al.*, 2011; Jones *et al.*, 2012; Seemayer *et al.*, 2014), it was found that when two residues were in contact across all known PDB (Berman *et al.*, 2000) structures of a multiple sequence alignment, the predicted residue-residue coupling score was ranked higher than a residue pair that was only in contact in a subset of the structures of that multiple sequence alignment (Zea *et al.*, 2018). A similar observation was also found in the dataset we analysed (see Figure 3.11). Three binary contact predictors serve as part of the input for DMPfold so we could check for the ranking of the residue pairs we classified. As expected, rigid residue pairs were generally ranked higher than flexible residue pairs.

Since in these binary methods a lower rank indicates a lower probability, residue-residue contacts that only exist in some states of a protein's ensemble (which we term interaction-changing or flexible) were more likely to be falsely classified as not in contact (at all). We wanted to test whether the improvements in co-evolution analysis, namely distance predictions, encode information about different conformers of a protein in the shape of their predicted distance distributions. We hypothesised that predicted distance distributions with more than one local maximum indicate residue pairs that can adopt more than one metastable state, and that having more than one local maximum is therefore related to flexibility.

In fact, we found a statistically significant difference between rigid and flexible residue pairs. Flexible residue pairs more often had multiple local maxima than rigid residue pairs. We checked if this difference was driven by individual proteins, biases in our dataset, an imbalance in secondary structure distribution or by

our validation strategy of approximating flexibility with two PDB structures. In all our checks we could confirm a difference between rigid and flexible residue pairs, supporting the view that flexibility information can be drawn from co-evolutionary distance predictions.

3.2 Methods

3.2.1 Dataset

The Database of protein Conformational Diversity in the Native State (CoDNaS) stores multiple PDB structures of single protein sequences (Monzon *et al.*, 2016). For our analysis, the maximum RMSD pairs dataset of Monzon *et al.* (2017) was used. This CoDNaS subset contains 4791 pairs of PDB structures that had the maximum C_α-RMSD amongst the pairwise comparisons between all conformers of a given protein. This subset includes only proteins that had at least five solved structures in the CoDNaS database to increase the reliability of the approximation of conformational diversity. Only X-ray structures with a resolution equal or less than 2.5Å were considered.

For our analysis, a subset of the maximum RMSD pairs dataset was used which contained 3075 proteins. This subset contains only proteins that have no intrinsically disordered regions (unresolved patches of five or more residues) in any of the conformers. The list of PDB structure pairs can be found in Table B.2. PDB structures were pre-processed with clean_pdb.py to remove alternative locations (downloaded from https://github.com/harryjubb/pdbtools/blob/master/clean_pdb.py on 14/11/2019).

For 2947 out of the 3075 proteins, distance predictions, secondary structure assignment and chemical interaction analysis for both PDB structures could be successfully performed. These protein pairs constitute our analysis set and are marked with a 1 in the 'analysis_complete' column of Table B.2.

Topological similarities between the CoDNaS set and the structures of the distance predictor's training set may exist, however, DMPfold was trained on only one protein conformation each for a non-redundant set of proteins and to yield

only one protein conformation. Any overlap between the DMPfold training set and the CoDNaS set would therefore if anything be detrimental as DMPfold would have seen only a single conformation, reducing the likelihood of predicting multiple peaks for any residue pair in that structure.

Unique CATH superfamily subset

To check for a potential bias from overrepresentation of some CATH superfamilies in the CoDNaS dataset, we additionally performed local maxima analysis for a subset of proteins with a maximum of one protein per CATH superfamily. If two or more proteins contained a (sub-)domain from the same CATH superfamily, only one of those proteins was selected at random. The 517 proteins of this subset are marked with a 1 in the 'in_unique_CATH_subset' column of Table B.2.

3.2.2 Co-evolutionary Distance Prediction

DMPfold (Greener *et al.*, 2019) was downloaded from https://github.com/psipred/DMPfold on 01/10/2019. It uses following sequence-based input features: sequence profile, mutual information (MI), MI product (MIp), mean contact potential, PSICOV contact scores (Jones *et al.*, 2012), FreeContact (mfDCA) contact scores (Kaján *et al.*, 2014), CCMpred (plmDCA) contact scores (Seemayer *et al.*, 2014), PSIPRED secondary structure, Shannon entropy in multiple sequence alignment columns, SOLVPRED solvent accessibility, log(1 + sequence separation), sequence bounds (channel of ones), and the DeepCov covariance matrix (Jones and Kandathil, 2018). In its normal application, DMPfold generates distance predictions for all residue pairs and then feeds those into a structure modelling program (CNS). After 3D model building the distance predictions are updated considering the built model (by default this step is done twice but can be varied). DMPfold was run with default parameters but without any 3D model building and updates of distance predictions. Thus, distance predictions used in this work always refer to the initial distance prediction generated before protein structure modelling iterations. Input features for

the initial distance prediction were generated with hhblits 3.0.3 (multiple sequence alignment) against uniclust30 (2018_08).

Trivial contacts/residue pairs are defined as those between residues that are four or less residues in sequence apart. No distance prediction is generated by DMPfold for those residue pairs.

A distance prediction for a pair of residues refers to the predicted C_β-C_β distances (in the case of glycine C_α) between those two residues and is termed predicted distance distribution here. Each predicted distance distribution is a vector of 20 points, representing the probabilities for each of the distance bins that DMPfold was trained to predict for a given sequence (or multiple sequence alignment). The first bin is ranging from 3.5-4.5Å, followed by seven bins of 0.5Å width up to a distance of 8Å and eleven bins of 1Å width up to 19Å. The last bin contained the probability density for distances over 19Å.

Target sequence of a distance prediction was the sequence derived from PDB structure 1 (see 'pdb_id_1' and 'chain_id_1' columns of Table B.2) through PDBParser(permissive=0) from Biopython 1.74 (Cock *et al.*, 2009).

3.2.3 Local Maxima Analysis

Predicted distance distributions were analysed with the 'peakdet' function from Eli Billauer, version 3.4.05 (downloaded from https://gist.github.com/endolith/250860 on 18/10/2019). The function considers a point a local maximum if it has the local maximal value, and was followed (in any probability density step towards 19Å) by a value difference by at least DELTA, which was chosen to be 0.03 (to maximise the fraction having two peaks while allowing at most three peaks). Detecting a local minimum (analogue definition with greater by at least delta) resets the local maximum. Thus, multiple local maxima can be found for each probability distribution. The last bin is never considered to be a maximum. Distance predictions with one maximum and a probability mass >0.5 within the interval of ±1Å around the maximum are considered to be 'sharp'. See Figure 3.1 for randomly selected examples of predictions with 0, 1, 2 or 3 local maxima.

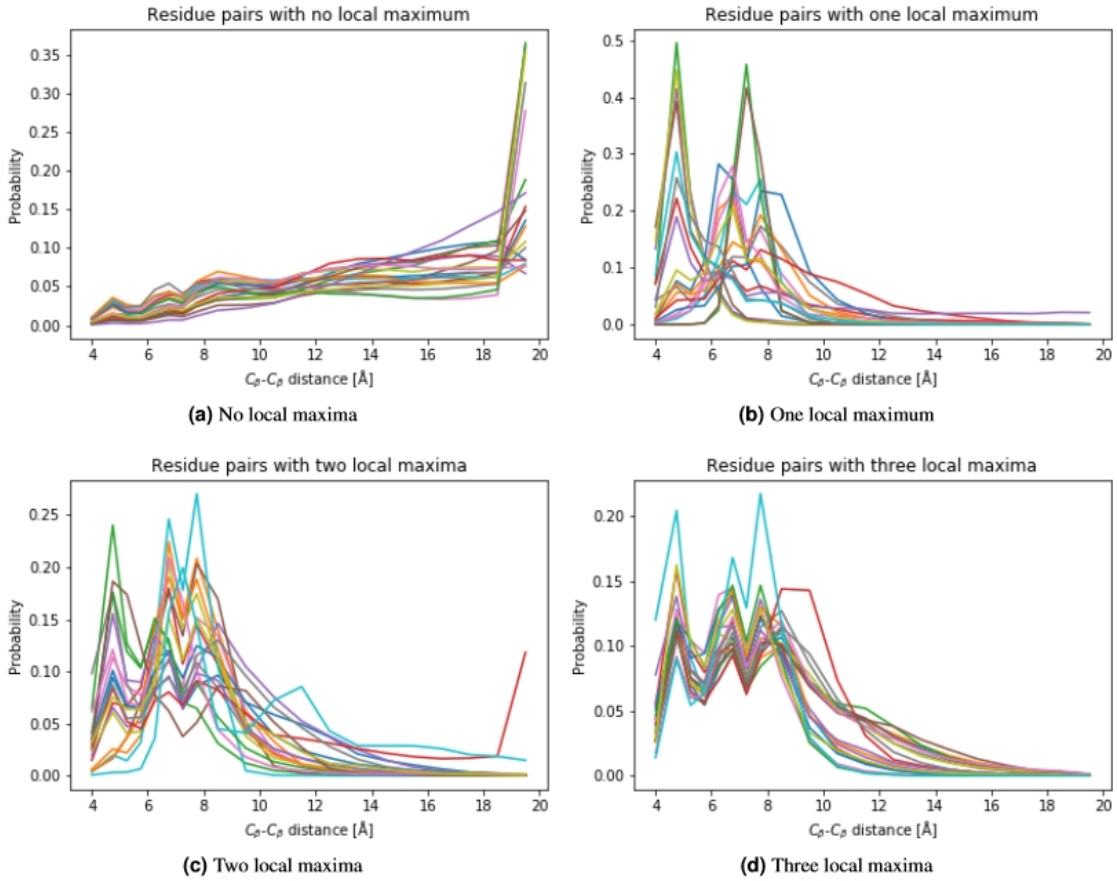


Figure 3.1: Randomly selected predicted distance distributions with no, one, two or three annotated local maxima. All plots show randomly selected distance predictions (of a given local maxima count) from 20 random proteins in the CoDNaS rigid set. **(a)** Distance predictions with no annotated local maximum. Please note, maxima were not assigned to the last bin. DMPfold distributes probability mass to the last bin for a residue-residue distance prediction greater than 19Å. Hence, it is a bin with an infinite interval and not suitable for maxima assignment. **(b)** Distance predictions with one maximum. **(c)** Distance predictions with two local maxima. **(d)** Distance predictions with three local maxima.

Figure 3.1 legend updated

3.2.4 Residue Pair Analysis

We only investigated non-trivial residue pairs that fulfilled the following true contact definition (true positives) and pairs that fulfilled the binary contact definition in at least one PDB structure if no local maximum could be detected at all (false negatives).

Definition of true contacts

The standard threshold for 'binary' contact predictors defines a residue pair to be in contact if the C_β-C_β distance (in the case of glycine C_α) of two residues is 8Å or less (Marks *et al.*, 2011). Adapting this to distance predictions and multiple conformers, we define a true positive contact having at least one local maximum below 8Å and the 'binary' contact definition being satisfied in at least one of the conformers.

Flexibility definition

Residue pairs were classified into two distinct classes, rigid and flexible pairs; C_β-C_β distances (C_α for glycine) and chemical bonds were used for this classification (Table 3.1). Residue pairs not matching those criteria were not classified as it is unclear if those residue pairs represent pairs with medium flexibility or uncertainty of experiments and analyses. See Figures 3.4 and 3.6 for fractions of predicted local maxima of these non-rigid-non-flexible residue pairs.

Table 3.1: Flexibility class definitions.

	Rigid residue pairs	Flexible residue pairs
C_β-C_β distance difference	< 1Å	$>= 2$Å
Chemical bond category	$1+/1+$	$0/1+$

C_β-C_β distance difference

The absolute difference between the C_β-C_β distances of the two structures of the same sequence was determined for each residue pair (in the case of glycine C_α). The absolute difference in distance had to be smaller than 1Å for the rigid classification and greater or equal to 2Å for the flexible classification.

Chemical bond analysis

To increase the confidence in a residue pair's assignment to a flexibility class (rigid/flexible), we determined the presence of at least one chemical bond between those two residues in both PDB structures. A chemical bond is defined as any of the CREDO interactions (Schreyer and Blundell, 2009) detected by Arpeggio

(Jubb *et al.*, 2017); proximal interactions excluded. Arpeggio is a programme that calculates and visualises interatomic interactions from protein structures. It was downloaded from https://github.com/harryjubb/arpeggio. At least one chemical bond had to be present in both PDB structures (1+/1+) to classify for the rigid set and at least one bond in one of the structures but none in the other (0/1+) to classify for the flexible set.

Secondary structure pair types

Secondary structure assignment was determined using DSSP (Kabsch and Sander, 1983) on PDB structure 1 (see 'pdb_id_1' and 'chain_id_1' columns of Table B.2). DSSP outputs an 8-letter code which we converted into two categories: secondary structure element (S) or loop/coil region (L). 'H', 'G', 'I' and 'E' were assigned to S and 'B', 'T', 'S' and ' ' were assigned to L following the definition of Marks and Deane (2018). The secondary structure pair type is constituted by both residues' assignments, leading to three types: 'S-S' (within or between secondary structures), 'S-L' (between secondary structure and loop) and 'L-L' (within or between loops).

3.2.5 Set Comparisons

For the comparison of fraction distributions the Mann-Whitney U test (two-sided) was applied as implemented in the Stats module of SciPy version 1.3.1 (Virtanen *et al.*, 2020). Common language effect size f was calculated as: $f = \frac{U_1}{n_1 n_2}$ with test statistic U and distribution sizes n_1 and n_2.

3.2.6 Rigid Loop Set

Benson and Daggett (2008) defined a set of loops that showed very little movement in MD simulations (mean RMSD below 0.5Å). Lists of the loop definitions and PDB structures used can be found in Table B.1. We followed the set definition of Marks and Deane (2018) (which excluded one original loop because of an uncertainty in labelling). The number of local maxima was determined for each predicted distance distribution of residue pairs involving one of the defined loop residues.

3.3 Results

3.3.1 The Shape of Predicted Distance Distributions is Related to Flexibility

As described earlier, co-evolutionary information and machine-learning have enabled the prediction of residue-residue distances (Greener *et al.*, 2019; Xu, 2019; Senior *et al.*, 2020; Yang *et al.*, 2020). These methods yield a predicted distance distribution for every residue pair and have been used to identify the most likely distance (or distance interval) between a pair of residues in a protein.

Examples of those predicted distance distributions, ordered by the number of local maxima, can be found in Figure 3.1. Most predictions without an annotated local maximum (Figure 3.1a) have probability mass spread out over almost all possible bins but peak in the last bin that contains the probability that the residue pair is further apart than 19Å. Because this last bin captures an infinite distance interval and may also indicate the uncertainty of the prediction, the last bin is never considered to be a maximum. Some distance predictions with one maximum (Figure 3.1b) have a sharp peak where a clear majority of probability mass is centred around the maximum value, while for others probability is much more spread out. Predictions with two or three local maxima (Figure 3.1c and d) were considered together as our work focused on detecting zero, single or multiple local maxima in predicted distance distributions.

Likely distances, distance intervals or the full predicted distance distribution can be used to set distance constraints that have been successfully used to predict static protein structures, e.g. Kryshtafovych *et al.* (2019). Here, we test whether these predicted distance distributions also contain information about residue pair flexibility. Information on which residues change their interactions to facilitate the switching between conformers will be crucial to predict multiple biologically relevant conformations and to enable insights into the mechanisms of allosteric signal transmission. To identify these interaction changes we analysed and compared the two most different PDB structures of a protein sequence. All residue pairs that

are present in both of these structures were classified into rigid, flexible or neither (only non-trivial, true contacts considered; See Methods for more information).

All analysed residue pairs are present in two conformations (two distinct PDB structures), so that for each residue pair two C_β-C_β distances and thus, a C_β-C_β distance difference could be determined (C_α for glycine). Rigid residue pairs were defined as those with a C_β-C_β distance difference below 1Å and flexible pairs were those with a C_β-C_β distance difference equal or above 2Å. Residue pairs with a distance difference between these two cutoffs were not classified into either set. Additional to the distance difference thresholds we applied a bond change criterion. Rigid residue pairs had to have at least one physicochemical bond in both structures and flexible pairs had to display at least one bond in one structure but none in the other. This criterion was applied to ensure interaction change for residue pairs that were classified as flexible.

If the shape of the predicted distance distributions relates to the conformational ensemble of a protein, rigid residue pairs should be characterised by predicted distance distributions with only one local maximum and flexible pairs should show predictions with two or more local maxima.

Figure 3.2 illustrates the expected difference between rigid and flexible residue pairs. The two residues of the rigid residue pair remain in the same orientation to one another in the two structures (Figure 3.2a), the two residues of the flexible pair have changed orientations between the two structures (Figure 3.2b). The distance prediction of the rigid pair shows a distribution with only one local maximum and the two C_β-C_β distances of the two structures are close to that maximum (Figure 3.2c). In contrast to that, the flexible pair's distribution shows two local maxima and the C_β-C_β distances of the two structures differ by more than 4Å (Figure 3.2d).

3.3.2 Flexible Residue Pairs Have more Distance Prediction Local Maxima than Rigid Residue Pairs

To test whether there is a general relationship between the flexibility of a residue pair and its predicted distance distribution, we investigated 2947 proteins from the

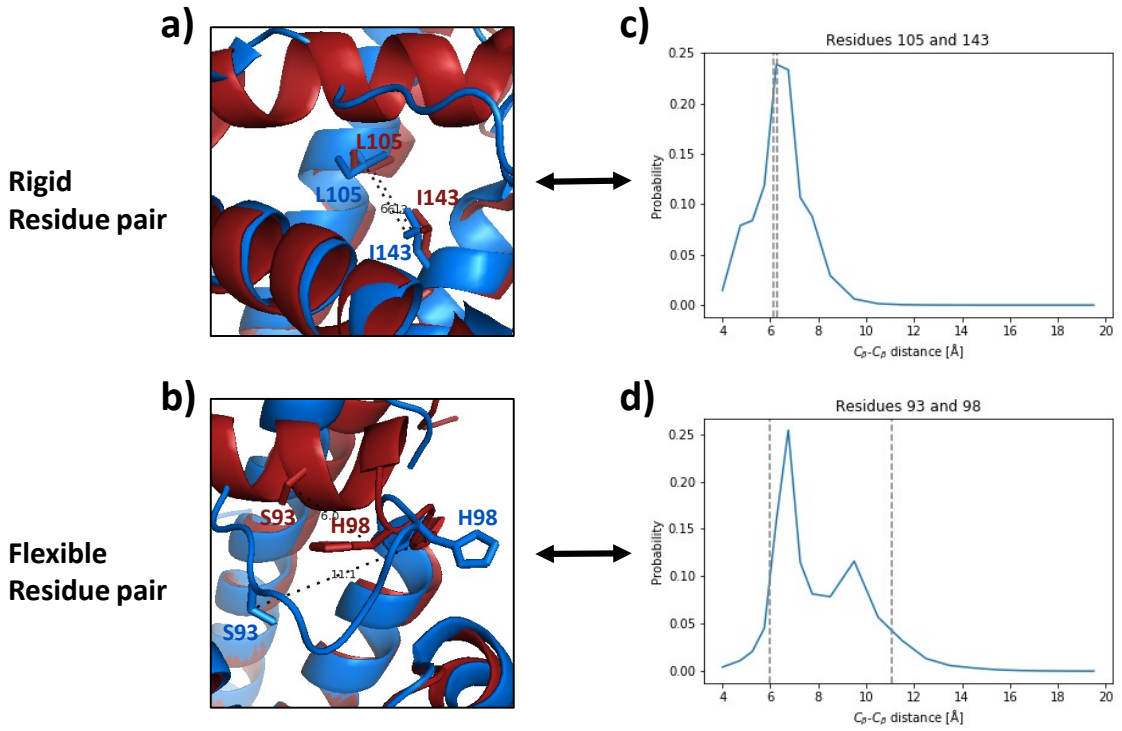


Figure 3.2: A rigid residue pair associated with a single local maximum and a flexible residue pair with multiple local maxima in their predicted distance distributions are shown. a) displays an example of a rigid residue pair of Myoglobin (L105:I143) and **b)** a flexible pair (S93:H98). The two most different PDB structures of Myoglobin were superimposed and in both images PDB 2EB8_A is shown in blue, 2JHO_A in red and the residue pairs in stick representation. **c)** and **d)** depict the predicted distance distributions for the respective residue pairs. Black dotted lines in PDB structures as well as vertical dotted lines in probability distributions represent the two C_β-C_β distances that are associated with the residue pairs in both PDB structures. The rigid residue pair in **a)** is very similar in both structures and hence, its two C_β-C_β distances are close to each other in **c)** next to the single local maximum of the distance prediction. The flexible residue pair in **b)** displays much greater movement and two C_β-C_β distances greater than 4Å apart which roughly represents the difference between the distance prediction's two local maxima in **d)**.

CoDNaS database. We classified each residue pair in each of those proteins into rigid or flexible (or none) and examined the predicted distance distributions for all residue pairs in those sets. Local maxima in the predicted distance distributions were defined as bins that were followed by a local minimum of at least 3% probability below its maximum value (see Methods). For a given set of residue pairs the number of local maxima in each predicted distance distribution was determined and from all those predictions the fraction of predictions with no, one, two or more local maxima was

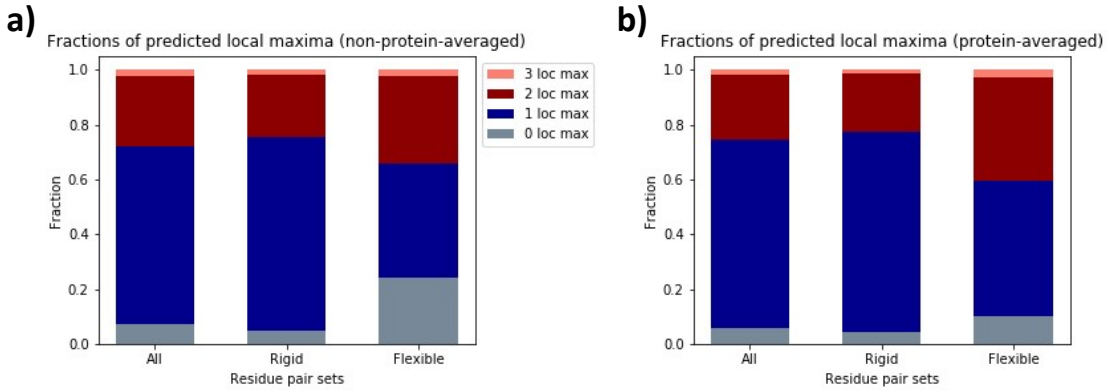
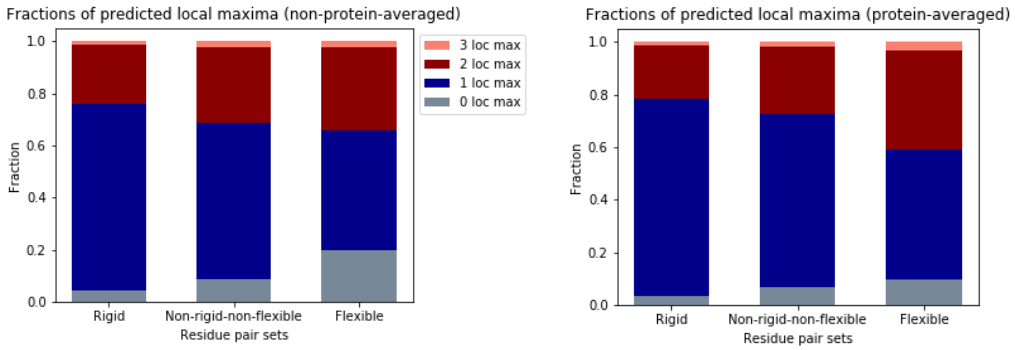


Figure 3.3: Fractions of local maxima counts are different between rigid and flexible residue pairs. (a) Fractions of predicted local maxima across all classified residue pairs of all proteins, showing that flexible residue pairs are both more likely to have no local maximum or multiple local maxima. **(b)** Protein-averaged fractions of predicted local maxima, showing the same overrepresentation of no or multiple local maxima for flexible residue pairs.

calculated. The different fractions of local maxima counts for all residue pairs, rigid residue pairs and flexible residue pairs across all proteins are shown in Figure 3.3a.

Cases where no local maximum could be detected in a distance prediction (grey) implied that either the two residues were predicted to be further than 19Å apart or that the predictor did not predict a highly probable distance (see Methods for definitions), 7% of all residue pairs that we investigated fell into this category. In the set of rigid residue pairs only 5% showed no local maximum but in the set of flexible residue pairs 24%, indicating that predicting residue-residue distances for flexible residue pairs is more challenging. 70% of rigid residue pairs had a single local maximum, whereas this was only 42% for flexible residue pairs.

As there are different numbers of residue pairs in the respective sets: 1,370,737 pairs in total, 711,239 in the rigid set and 8,681 in the flexible set and to show that no individual proteins are driving the differences in the number of local maxima, protein-averaging was performed. For each protein the number of local maxima was calculated for each residue pair ('all') and for the sets of rigid and flexible residue pairs if a protein had at least one residue pair with a single local maximum and at least one other pair with multiple local maxima. Protein-averaged fractions for the residue pair sets are shown in Figure 3.3b. On average, of all residue pairs 25%



(a) Total fractions

(b) Protein-averaged fractions

Figure 3.4: Fractions of local maxima counts are different between rigid and flexible residue pairs, also when considering only unique CATH superfamilies. **(a)** Fractions of predicted local maxima across all residue pairs of proteins from the set of unique CATH superfamilies. As in the full CoDNaS set, flexible residue pairs are both more likely to have no local maximum or multiple local maxima. Unclassified non-rigid-non-flexible residue pairs display intermediate values that lie between rigid and flexible pairs. **(b)** Protein-averaged fractions of predicted local maxima, showing the same overrepresentation of no or multiple local maxima for flexible residue pairs. Also here, non-rigid-non-flexible residue pairs show fractions that are within the difference of rigid and flexible pairs.

± 8% had multiple local maxima predicted. A fraction of 22% ± 7% of the rigid pairs were found to have multiple local maxima compared to 40% ± 17% of the flexible ones. The difference in the fraction of multiple local maxima between rigid and flexible residue pairs is slightly higher when applying protein-averaging but qualitatively the results are the same (Figures 3.3a and 3.3b).

This is also true by avoiding further bias by computing the local maxima fractions for a subset of 517 proteins that consist only of unique CATH superfamilies (see Figure 3.4). The local maxima fractions and their differences are very similar to those of the full CoDNaS set, further supporting the view that the differences between rigid and flexible residue pairs are not due to a sampling bias. In this analysis, we also display the local maxima fractions of residue pairs we did not classify into rigid or flexible to get a better picture of the characteristics of those residue pairs. As expected, they show intermediate values that lie between those of rigid and flexible pairs.

Table 3.2: Fraction of secondary structure classification depending on flexibility classification.

	S-S (within/between)[1]	S-L or L-L[2]	Number of proteins[3]
All residue pairs	0.52 ± 0.13	0.48 ± 0.13	2934
Rigid pairs	0.58 ± 0.12	0.42 ± 0.12	2914
Flexible pairs	0.15 ± 0.14	0.85 ± 0.27	1584

[1] Fraction of residue pairs within/between secondary structure elements
[2] Fraction of residue pairs including at least one loop/coil residue
[3] Proteins with at least one classified pair

3.3.3 Predicted Distance Distributions Can Capture Flexibility Independent of Secondary Structure

Table 3.2 shows that the residues in pairs that were classified as flexible were more often part of loop structures than the residues in rigid residue pairs (85% vs. 42%). This is expected since loop structures are often intrinsically flexible (Nilmeier *et al.*, 2011). As it is known that loop structures are hard to predict (Marks and Deane, 2018), a higher fraction of multiple local maxima could just stem from an increased uncertainty of predictions and not from residue pair flexibility that relates to distinct conformations. This would imply that the higher fraction of multiple local maxima observed in flexible residue pairs compared to rigid pairs is due to imbalances in secondary structure type proportions and not driven by residue pair flexibility. Therefore, we analysed the secondary structure composition of all sets of residue pairs and the local maxima fractions of different secondary structure subgroups within the two sets of interest. If the local maxima fractions in the respective rigid and flexible secondary structure subgroups were identical and the overall differences were only caused by different proportions of subgroups, the number of local maxima would predict secondary structure type and not flexibility itself.

Figure 3.5 shows the local maxima fractions dependent on secondary structure subgroups comparing the sets of rigid and flexible residue pairs. Even when stratifying by secondary structure, we still observe differences between the number of local maxima of the rigid and flexible sets. Whereas the subgroups with at least one residue being part of a loop (S-L and L-L) of rigid and flexible sets show similar

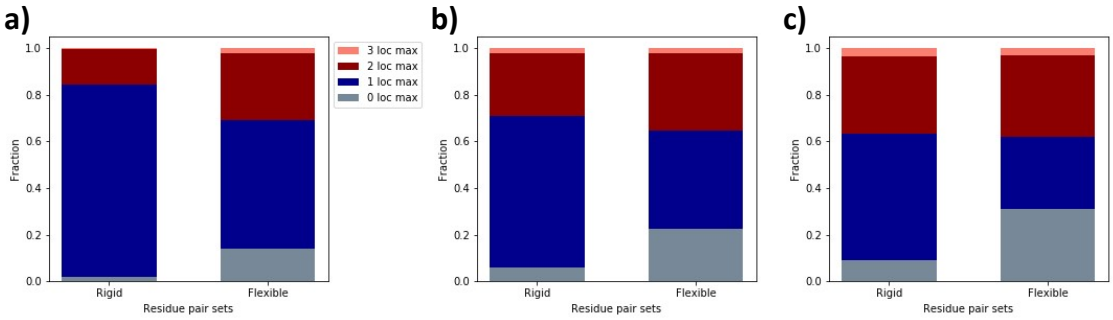
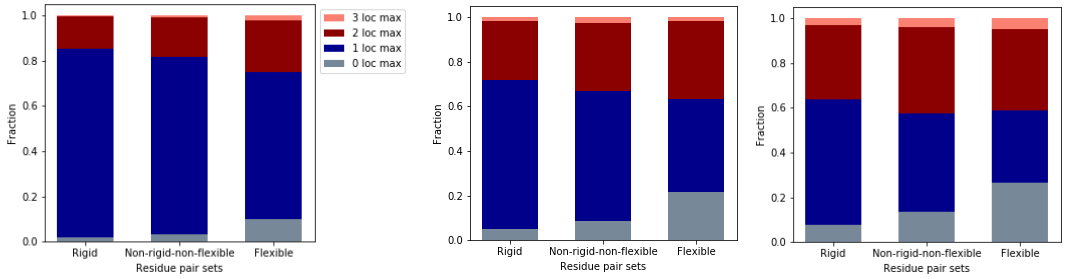


Figure 3.5: Predicted distance distributions capture flexibility independent of secondary structure. Number of local maxima fractions were computed for subgroups depending on their secondary structure annotation: **(a)** 'S-S' when both residues were found to be part of a secondary structure element, **(b)** 'S-L' when one of the residues was in a secondary structure element and the other in a loop, and **(c)** 'L-L' when both were part of a loop/coil region. The different fractions of multiple local maxima (red) between rigid and flexible residue pairs are not only driven by an imbalance of secondary structure element and loop residues: for example the S-S subgroup of the flexible set has a fraction of multiple local maxima twice as big as the rigid set's.

fractions of multiple local maxima (Figures 3.5b and 3.5c), the subgroup where both residues of a pair are part of a secondary structure element (S-S) displays larger differences. In that subgroup, flexible residue pairs have about twice as many predictions with multiple local maxima compared to the rigid pairs (red in Figure 3.5a). This, and the large differences in single local maximum fractions (blue) in all subgroups, confirms that the differences between rigid and flexible residue pairs are not driven only by different proportions of secondary structure in the sets.

As for the overall fractions, we also computed local maxima fractions of the secondary structure subgroups for a subset of proteins with unique CATH superfamilies (see Figure 3.6). The differences in secondary structure subgroups between rigid and flexible residue pairs are also found when controlling for a potential superfamily bias which further supports our finding.

Our definition of rigid depends on the currently solved crystal structures, so it is possible that when a residue pair is classified as rigid it might not be, alternative conformations might not yet have been solved. However, if a residue pair is classified as flexible, this is likely to be correct as it has already been found in at least two different structurally distinct conformations. As residue pairs are more prone to be incorrectly classified as rigid than as flexible, the fraction of multiple local maxima



(a) S-S **(b)** S-L **(c)** L-L

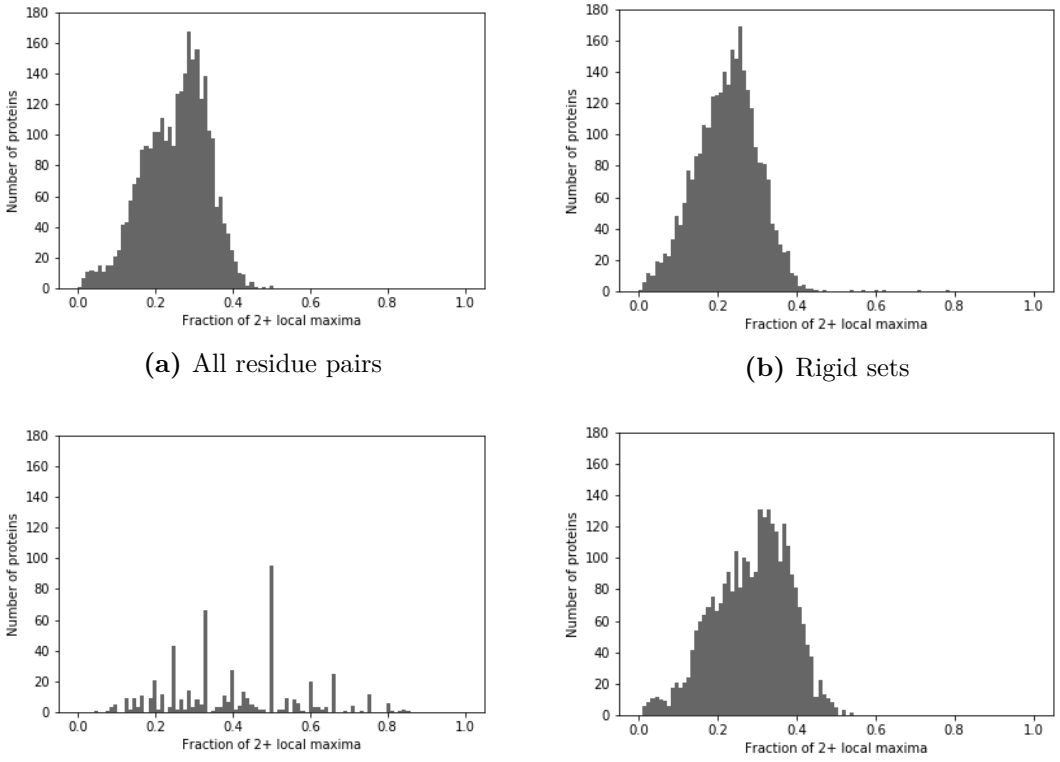
Figure 3.6: Predicted distance distributions capture flexibility independent of secondary structure. Number of local maxima fractions were computed for subgroups depending on their secondary structure annotation: **(a)** 'S-S' when both residues were found to be part of a secondary structure element, **(b)** 'S-L' when one of the residues was in a secondary structure element and the other in a loop, and **(c)** 'L-L' when both were part of a loop/coil region. The different fractions of multiple local maxima (red) between rigid and flexible residue pairs are not only driven by an imbalance of secondary structure element and loop residues: for example the S-S subgroup of the flexible set has a fraction of multiple local maxima twice as big as the rigid set's. Non-rigid-non-flexible residue pairs also display intermediate values here, with one exception in the Loop-Loop subset where unclassified pairs have a multiple local maxima fraction of 0.42 and flexible pairs 0.41.

of rigid residue pairs is likely to be rather overestimated than underestimated. This implies that the difference in multiple local maxima fractions between rigid and flexible residue pairs could increase with more structural data becoming available, further supporting our finding that residue pair flexibility is captured independent of secondary structure imbalances.

3.3.4 Differences in Number of Local Maxima Between Rigid and Flexible Sets Are Statistically Significant

Given that these differences in predicted distance distributions are related to flexibility and rigidity of residue pairs, we tested if these differences are statistically significant. For this, we performed protein averaging and analysed the fraction-per-protein distributions. For each protein the fraction of all residue pairs with multiple local maxima was determined and as well as the fraction within the sets of flexible, rigid or unclassified residue pairs, Figure 3.7 shows these distribution histograms.

The number of proteins contributing to the distribution of all proteins was 2899, 2858 for the rigid sets and 542 for the flexible sets. Note the differences to the



(a) All residue pairs

(b) Rigid sets

(c) Flexible sets

(d) Non-rigid-non-flexible sets

Figure 3.7: Multiple local maxima occur significantly more often in flexible residue pairs than rigid residue pairs. **(a)** Fractions of multiple local maxima for all residue pairs of a protein. **(b)** Fractions of multiple local maxima for the sets of rigid residue pairs. For each protein, all rigid residue pairs were analysed for the number of predicted local maxima. Then, the fraction of predictions with more than one local maximum was plotted in the histogram. This procedure was used to generate the fraction-per-protein distributions for all four subset definitions. **(c)** Fractions of multiple local maxima for the sets of flexible residue pairs. **(d)** Fractions of multiple local maxima for the sets of residue pairs that were not classified as either rigid or flexible. The distributions of all, rigid and flexible residues pairs are significantly different from one another (Mann-Whitney U test): rigid vs. all sets with a p-value $< 10^{-44}$ ($U = 3257618.5$, $n_1 = 2858$, $n_2 = 2899$, effect size $f = 0.39$), flexible vs. all sets with a p-value $< 10^{-87}$ ($U = 1207168$, $n_1 = 542$, $n_2 = 2899$, effect size $f = 0.77$) and flexible vs. rigid sets with a p-value $< 10^{-122}$ ($U = 1268779.5$, $n_1 = 542$, $n_2 = 2858$, effect size $f = 0.82$).

numbers when comparing secondary structure pairs in Table 3.2 (1584 vs. 542). Applying a criterion of having at least one residue pair with a single local maximum and at least one other pair with multiple local maxima in a given protein reduced the number of proteins by about two thirds. We applied this criterion to reduce the potential bias from proteins with a very low number of classified residue pairs.

The distributions of all, rigid and flexible pairs are all significantly different from

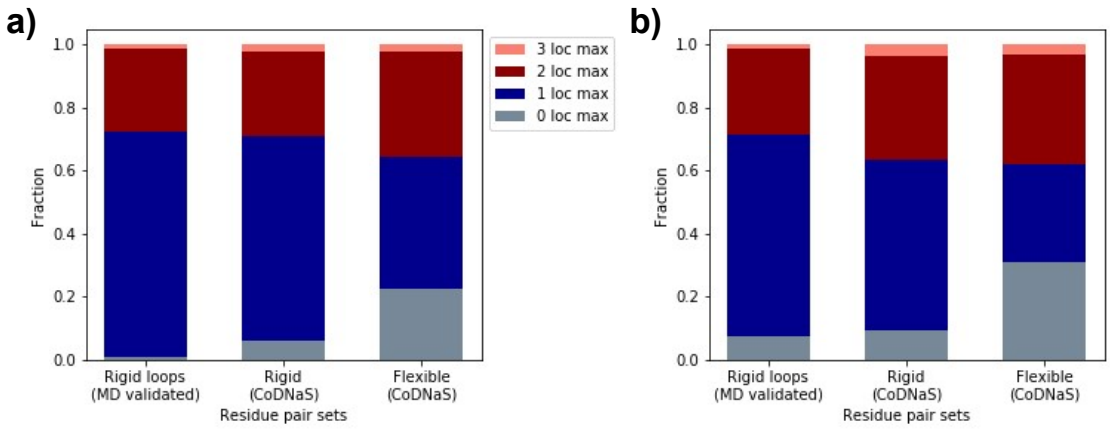


Figure 3.8: Local maxima fractions of the rigid loop set are similar to fractions of the rigid residue pairs of the CoDNaS set. Secondary structure subgroups **(a)** 'S-L' and **(b)** 'L-L' show small fractions of distance predictions with multiple local maxima (red) and large fractions with a single local maximum (blue). The fractions of multiple local maxima are generally similar to the fractions of rigid residue pairs in the CoDNaS set. This rigid loop set that was defined not by comparing two PDB structures but by analysing MD simulations, confirms the finding that rigid residue pairs less often have multiple local maxima in their distance predictions than flexible pairs.

one another (Mann-Whitney U test): all vs. rigid sets with a p-value $< 10^{-44}$, all vs. flexible sets with a p-value $< 10^{-87}$ and rigid vs. flexible with a p-value $< 10^{-122}$. These results not only show that the shape of predicted distance distributions of rigid and flexible residue pairs differ significantly, but also that, when comparing to rigid residue pairs, flexible residue pairs more often have multiple (instead of single) local maxima in their predicted distance distributions.

3.3.5 Examining the Local Maxima Fractions on a Set of Protein Loops Defined as Rigid

In order to see if our findings generalised to other methods of identifying protein flexibility, we investigated a set of 20 loops that had been found to move very little during MD simulations (Benson and Daggett, 2008). In a large-scale flexibility analysis of MD simulations of 250 proteins, these loops had a mean movement of less than 0.5Å and thus, were at the very rigid end of the spectrum of all loops analysed.

We computed the number of predicted local maxima for each residue pair that involved at least one of the loop residues of these 20 rigid loops (496 pairs in total)

and generated per-rigid-loop-averaged fractions. This rigid loop set of residue pairs had on average 72% ± 23% of predicted distance distributions with a single local maximum, 25% ± 13% with multiple maxima and 3% ± 10% without any local maximum. Since the loops of this set had shown little flexibility during simulations, their fractions should be similar to the rigid residue pairs of the CoDNaS set which is indeed what we observe (see Figures 3.8a and 3.8b for a comparision of S-L and L-L secondary structure subgroups). Although generally similar, the fractions of multiple local maxima are even lower and the fractions of single local maxima are even higher in the rigid loops than in the rigid residue pairs of the CoDNaS set. These results agree with the idea described above that some rigid residue pairs within the CoDNaS set might be incorrectly classified as rigid because alternative conformations have not yet been observed in crystal structures.

This set of rigid loops was defined by an alternative validation strategy, not by comparing two PDB structures but by analysing MD simulations, and shows once again that rigid residue pairs have multiple local maxima predicted less often than flexible pairs.

3.3.6 Case Studies Highlight the Need for more Complex Analysis than Simple Local Maxima Counts

The finding that rigid and flexible residue pairs vary in the number of local maxima in predicted distance distributions raises the question on how to use the distance predictions. The ideal scenario would be a straight forward prediction where every predicted distance distribution with multiple local maxima correctly predicts a flexible residue pair. As the confusion matrix in Table 3.3a shows, this scenario is not reality. The imbalance between the number of flexible residue pairs in our dataset compared to the amount of rigid residue pairs is huge (about 1:100) and makes a flexibility prediction based on local maxima counts only feasible if the underlying signal is very clear. Table 3.3b points out that precision is the major issue in a predictive setting. Since the differences in local maxima fractions

between rigid and flexible pairs are not very strong, the imbalance drives up the false positives and therefore impacts precision.

Table 3.3: Local maxima counts in predictive setting

(a) Local maxima counts vs. flexibility class.

Confusion matrix	2+ local maxima	1 maximum
Flexible	2975	3595
Rigid	175658	500154

(b) Predictive performance.

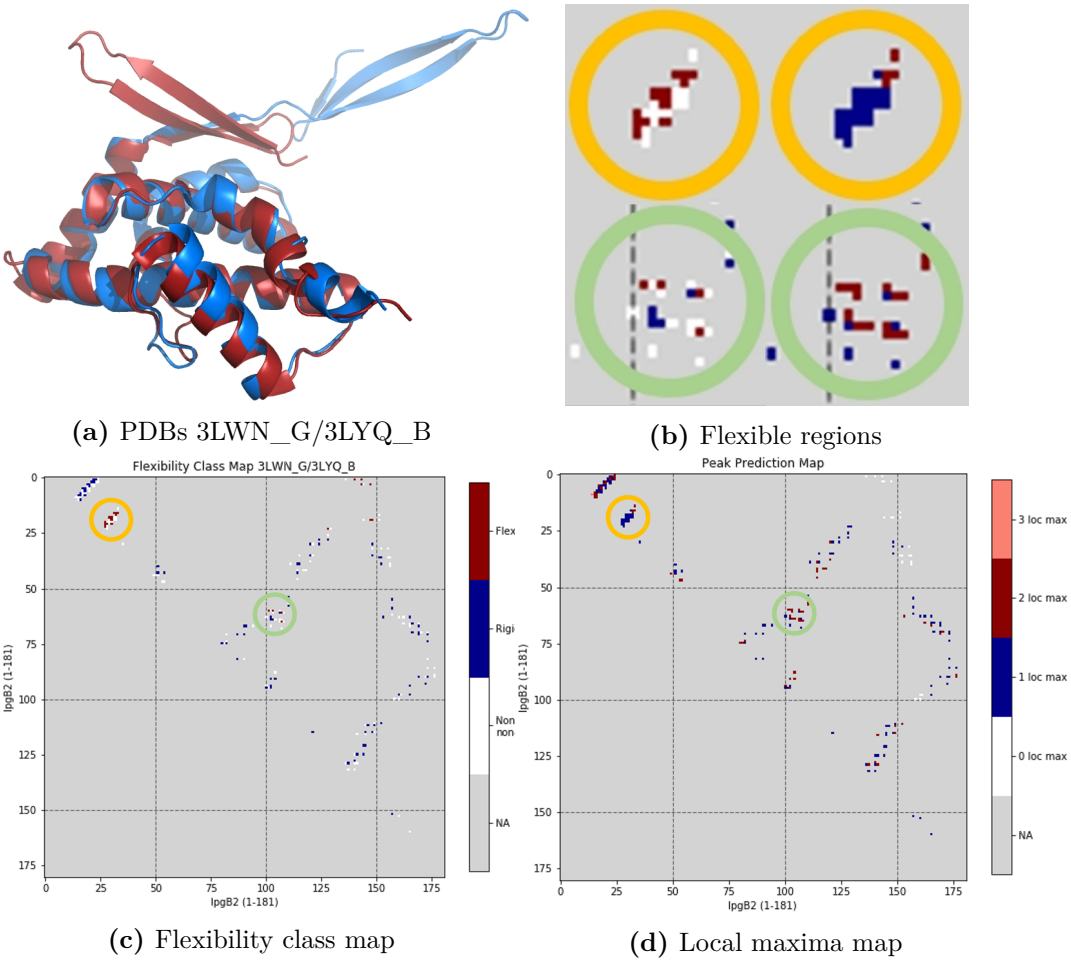
Statistical metric	Score
Accuracy	0.737
Precision	0.017
Recall	0.453
F1 score	0.032

To gain a better understanding of the underlying data and the signal strength, we conducted two case studies.

Case study: 3LWN_G/3LYQ_B

The first case study looks at IpgB2 of *Shigella flexneri* with its maximum-RMSD-pair 3LWN_G (red) and 3LYQ_B (blue) (see Figure 3.9a). The structure pair contains two flexible regions: a large-scale movement of part of the β-sheet (top) and a smaller loop variation (bottom). A flexibility class map shows flexible residue pairs in red, rigid ones in blue and unclassified pairs in white (Figure 3.9c). A local maxima map next to it shows two or more local maxima of a predicted distance distribution of a given residue pair in red colours, single maxima in blue and zero-maxima predictions in white (Figure 3.9d). To facilitate comparison, the areas of interest in both maps have been zoomed in at and highlighted in Figure 3.9b.

The flexibility class map shows two regions of different movement scale. While there are individual red (multiple local maxima) predictions in the β-sheet region (orange circles), the vast majority of residue pairs has only one predicted maximum. Distance prediction in its current form seems to miss such large-scale conformational changes. This comes with no surprise as the predictor was only trained on single conformations of proteins. The highlighted loop region (green circles) shows a mix of flexible, rigid and unclassified residue pairs. The local maxima map displays a mix too, but with a majority of multiple local maxima predictions over single



(a) PDBs 3LWN_G/3LYQ_B

(b) Flexible regions

(c) Flexibility class map

(d) Local maxima map

Figure 3.9: Comparison of flexibility class map and predicted local maxima map show difficulties of predicting large scale motions and potential for predicting small scale variability. **(a)** IpgB2 of *Shigella flexneri* in two different conformations: 3LWN_G (red) and 3LYQ_B (blue). **(b)** The structure pair contains two flexible regions: a large scale movement of part of the β-sheet (top row, orange circles) and a smaller loop variation (bottom row, green circles). The left circles are zoomed in from the flexibility class map and the right circles from the predicted local maxima map. **(c)** Flexibility class map showing flexible residue pairs in red, rigid ones in blue and unclassified pairs in white. **(d)** Local maxima map showing two or more local maxima of a predicted distance distribution of a given residue pair in red colours, single maxima in blue and zero-maxima predictions in white.

maxima, highlighting the potential of current predictors to catch such smaller variations already.

Case study: 1ODB_F/2WCB_B

The second case study shows three residue pairs from the maximum-RMSD-pair 1ODB_F/2WCB_B and their respective predicted distance distributions (Figure

3.10). All three have in common that they were classified as flexible and have a zero-maxima prediction.

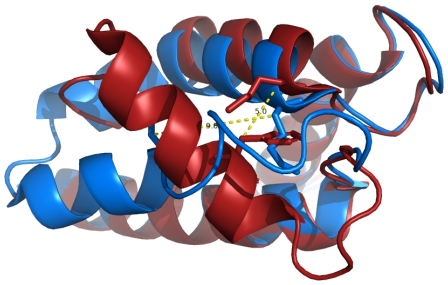
Please note: the last bin captures probability for all distances above 19Å and thus, also to a degree the uncertainty of a distance prediction. Because no likely distance can be assigned to this last bin, it was never considered to be a maximum. Furthermore, local maxima could only occur when they were followed by a local minimum with at least 0.03 probability difference.

While for all three zero-maxima predictions the highest probability can be found in or around the last bin, the majority of probability mass is in other bins, also around distances found in the PDB structures. This highlights that there can also be useful information in zero-maxima predictions and simply using the number of predicted local maxima does not seem to be the best way to exploit the findings of our work; a more complex extraction of this data is necessary.

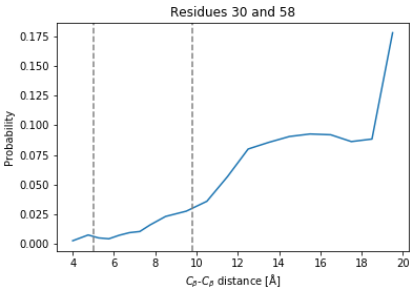
3.4 Discussion

Switching between different structurally distinct states is common in proteins carrying out various functions like catalysis or molecular recognition. For this switching between conformations, flexibility and the change of specific residue interactions is necessary. Natural or drug-induced allosteric events may influence these interaction changes and a protein's resulting activity which motivates research that helps understanding these fundamental mechanisms.

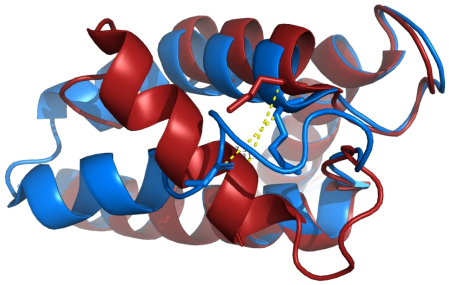
It had been reported that binary contact predictors tend to simply assign lower probabilities to interaction-changing residue pairs (Zea *et al.*, 2018), an effect we also found in our large-scale analysis of about 3000 proteins (Figure 3.11). Flexible residue pairs were generally ranked lower than rigid residue pairs. A lower binary score implies that a flexible residue pair, that has different interactions in different conformations, is more likely to be classified as not in contact at all, even though it is in contact in some conformations of a protein's ensemble. We wanted to test if newer co-evolution methods which predict residue-residue distances capture information about these changing interactions in a more exploitable way.



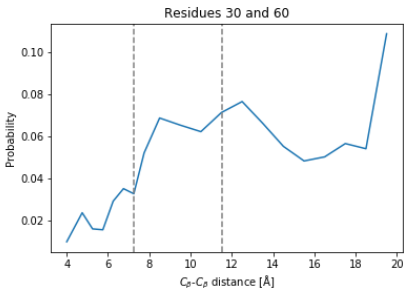
(a) Flexible residue pair 30/58



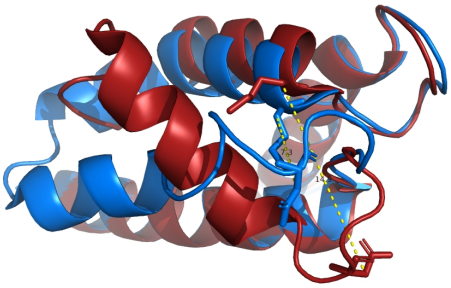
(b) Zero-maxima prediction for 30/58



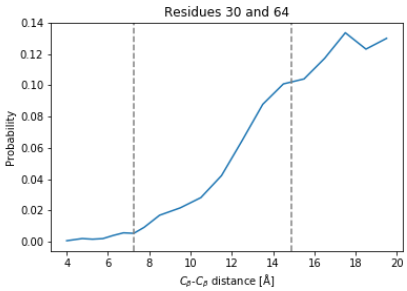
(c) Flexible residue pair 30/60



(d) Zero-maxima prediction for 30/60



(e) Flexible residue pair 30/64



(f) Zero-maxima prediction for 30/64

Figure 3.10: Example residue pairs from flexible region of 1ODB_F/2WCB_B show that zero-maxima predictions may contain information on residue pair flexibility. (a)-(f) All three example residue pairs are flexible pairs that have zero-maxima predictions. While the highest probability can be found in or around the last bin, the majority of probability mass is in other bins, also around distances found in the PDB structures (dashed vertical lines). This is especially true for the residue pair 30/60 (**(c)+(d)**) and highlights that there can also be useful information in zero-maxima predictions.

Specifically, in the work of this chapter we investigated if rigid and flexible residue pairs have different numbers of local maxima in their predicted distance distributions. For this we analysed 2947 proteins of the CoDNaS database. For every protein the two most different available PDB structures were taken and residue pairs that were present in both structures were grouped into rigid and flexible sets. We analysed and defined residue pair flexibility according to the C_β-C_β distance

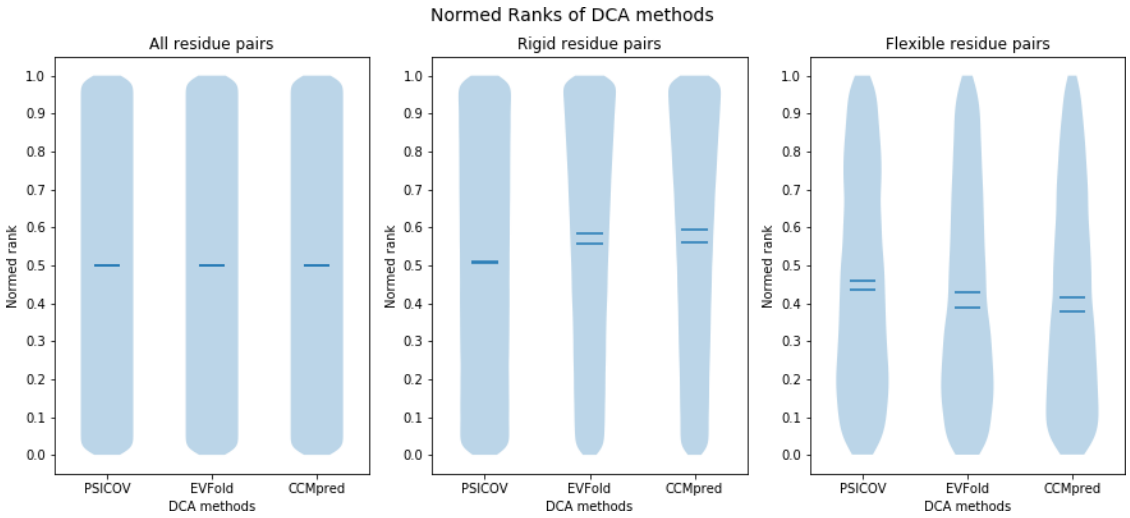


Figure 3.11: Non-machine learning contact predictors based on direct coupling analysis (DCA) rank flexible residue pairs lower than rigid residue pairs. Analysis includes all proteins of our CoDNaS dataset. 'All residue pairs' (left) is the set of residue pairs that are true contacts according to our definition for two PDB structures (see Methods). As the ranks in all three plots are normalised to just the true contacts of a given protein, ranks are uniformly distributed for All residue pairs and all three DCA methods. A rank of 1 indicates a prediction with the highest probability amongst all true contacts and a rank of 0 a prediction with the lowest probability. For rigid residue pairs predictions from EVfold and CCMpred are skewed towards 1 and predictions from PSICOV are close to uniformly distributed. For flexible residue pairs predictions from all three DCA methods are skewed towards 0 and means and medians are all below the ones for rigid residue pair predictions. This implies that flexible residue pairs are generally ranked lower (and thus predicted to be less likely in contact) than rigid residue pairs by all three investigated DCA methods.

difference between the two different conformations of a residue pair and added a bond change criterion to ensure interaction change.

When analysing the number of local maxima in the predicted distance distributions of rigid and flexible residue pairs, the two groups were found to have different fractions of single and multiple local maxima. Rigid residue pairs displayed a larger fraction of single local maxima than flexible pairs, whereas flexible residue pairs had a larger fraction of multiple local maxima than rigid residue pairs. Each local maximum is potentially related to a distinct distance that stems from a biologically relevant conformation of a protein (see Figure 3.2). However, two case studies (see Figures 3.9 and 3.10) suggest that using the predicted distances directly might not be the best way of exploiting the flexibility information that

is contained in distance predictions. As current predictors are trained only on and for single conformations, it seems more appropriate to specifically train with multiple structures of related sequences.

Since flexible residue pairs are more commonly found as part of loop structures which are often intrinsically flexible, we tested if this imbalance of secondary structure proportions between rigid and flexible sets was the sole driver of differences between residue pair sets. We found that this was not the case and the difference between sets was also present in secondary structure subgroups. Hence, the information on residue pair flexibility that is captured by distance predictions is not only due to the uncertainty of predictions for areas of loop/coil secondary structure.

After confirming that distance predictions capture residue pair flexibility and not only secondary structure, we performed protein averaging of the fractions of single and multiple local maxima. This yielded distributions of fractions on a per-protein basis which allowed us to demonstrate that the differences between rigid and flexible residue pairs are statistically significant.

We then tested if our results held when considering protein flexibility as defined by molecular dynamics. We predicted distance distributions for all residue pairs within and in contact with a set of loops which were found to be rigid in MD simulations. Although the validation of rigidity differed from our approach of comparing two PDB structures of one protein, the resulting fractions of single and multiple local maxima were similar to the set of rigid residue pairs in the CoDNaS set. This alternative dataset supports our findings of differences between distance predictions of rigid and flexible residue pairs. As discussed above, some rigid residue pairs might have been incorrectly classified as rigid because alternative protein conformations, that might show a flexible behaviour of that residue pair, have not yet been solved. The even higher fraction of predicted distance distributions with a single local maximum of residue pairs that were classified as rigid by MD simulations, compared to residue pairs classified by PDB structure pairs, further strengthens this view.

The fact that the average number of local maxima and thus, the number of predicted distances that are associated with these local maxima varies between

rigid and flexible residue pairs could be used to improve multi-conformer modelling. Current structure prediction tools aim to predict the static structure of a protein (Senior *et al.*, 2020; Greener *et al.*, 2019; Xu, 2019; Yang *et al.*, 2020) and multiple output structures are an occasional byproduct (Greener *et al.*, 2019). The results of this study imply that distance predictions contain information about residue pair flexibility even when the distance predictor was only trained on one protein shape and to generally yield only a single conformation. This highlights the potential of predictors specifically designed for predicting more than one conformation. While there is a statistical difference, distance predictions are not able to predict residue pair flexibility on their own, but they could be useful information to feed into such a predictor. A recently published study (del Alamo *et al.*, 2021) indicates that, as part of CASP14, AlphaFold2 had predicted one model of an alternative protein conformation that had only been suggested in double electron-electron resonance spectroscopy experiments but had not been observed in a crystal structure. This further supports our finding that sequence information contains information on conformational flexibility and emphasises our conclusion that the potential to predict biologically meaningful protein ensembles can and should be actively pursued.

In the next chapter, I examined a particular conformational ensemble, a structural ensemble of kinases that I had built previously. During my DPhil, we performed docking studies with the generated kinase models to demonstrate merit of those models and could subsequently publish our work in *Proteins* (Schwarz *et al.*, 2019). The systematic modelling and docking study is described in the next chapter of this thesis.

4

Modelling Conformational Ensembles

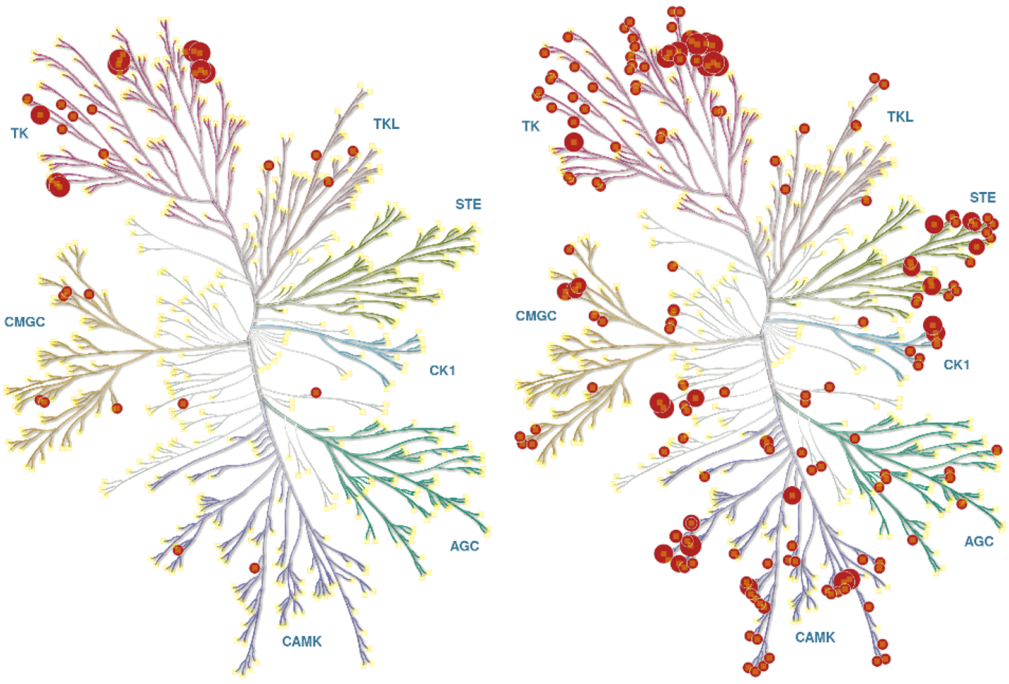
Contents

4.1 Background

In this chapter, we investigate homology model-based structural ensembles that were built during my master's degree at the University of Heidelberg. Modelling proteins in multiple biologically relevant conformations is not only helpful for investigating allosteric mechanisms but also for drug discovery approaches (Cleves and Jain, 2020). In my master's, I worked on generating conformational ensembles for kinases, which constitute a large group of potential drug targets (Cohen, 2002). The pipeline building and modelling was carried out in Heidelberg but during my DPhil I used this model set as a case study to investigate if such modelled conformational ensembles can lead to functional insight. The general modelling approach is described here only briefly and the docking study in more detail; for more information on the modelling please refer to the published paper in *Proteins*, Volume **87**, 943–951 (2019) (Schwarz *et al.*, 2019).

The human kinome contains about 500 typical kinases (Möbitz, 2015) which are a class of important drug targets as they are key players in cellular signal transduction (Taylor and Kornev, 2011) and mutations to or mis-regulation of cause various diseases, e.g. cancer (Cohen, 2002). All kinases catalyse a phosphorylation reaction and therefore all have similar binding pockets. This means most small molecule kinase inhibitors bind to more than one kinase (see Figure 4.1) which results in selectivity and intellectual property issues when developing drugs against kinases (Zuccotto *et al.*, 2010). One strategy to address this is to develop drug candidates that fully or partially bind to a hydrophobic side pocket of the main binding cavity. This side pocket of kinases is not accessible in active conformations but only in inactive states of a kinase (see Figure 4.2b). Furthermore, it is less conserved than the main binding pocket and therefore offers the potential to be more selective and to lead to novel intellectual property.

Typical kinases consist of a static C-terminal lobe (C-lobe; bottom in Figure 4.2a) and a more flexible N-terminal lobe (N-lobe; top in Figure 4.2a) that contains an αC-helix (aC-helix; pink in Figure 4.2a) and a phosphate-stabilising loop (P-loop; blue in Figure 4.2a) which both can adopt different conformations (see Figure 4.3).

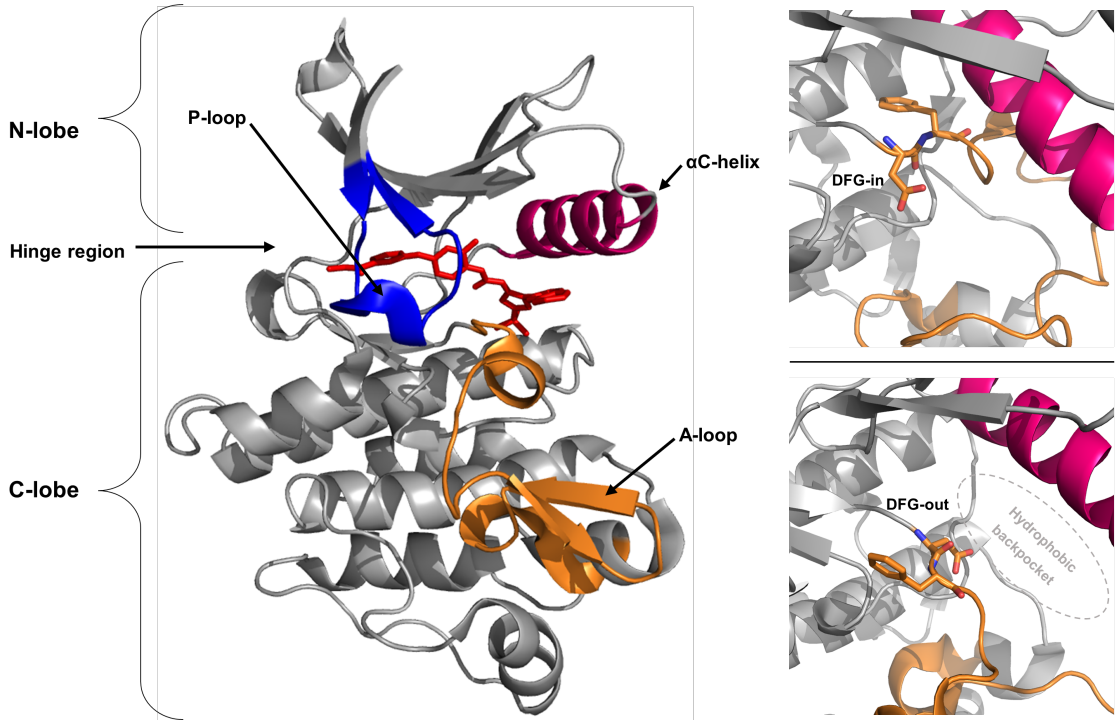


(a) Targets of Imatinib **(b)** Targets of Sunitinib

Figure 4.1: Key- and off-targets of one selective and one promiscuous drug. A kinome tree here represents all typical human kinases. Bigger circles indicate lower K_D values (higher affinities). **a)** displays the kinases targeted by Imatinib and **b)** the targets hit by Sunitinib. Much fewer targets are hit by Imatinib than Sunitinib which makes Imatinib more selective than Sunitinib (or Sunitinib more promiscuous). Figure created with KinMap (www.kinhub.org/kinmap/) (Eid *et al.*, 2017).

The most variable element of kinases is the activation loop (A-loop; orange in Figures 4.2 and 4.3) which carries a highly conserved aspartate-phenylalanine-glycine-motif (DFG). This DFG motif is activity-determining for kinases as its orientation is crucial for catalytic activity; if the residues are in an inwards orientation (DFG-in) a kinase may be active and if the motif is in an outwards orientation (DFG-out) the kinase is inactive (see Figure 4.2b). Kinases in the DFG-out state provide multiple druggable variants (Fabbro *et al.*, 2015) but compared to DFG-in structures, DFG-out structures are underrepresented in the PDB. Thus, structure-based drug design efforts for DFG-out inhibitors may benefit from an efficient approach to generate conformational ensembles of DFG-out structures.

The conformational plasticity of the DFG-out state can be mainly described by different conformations of ATP-binding site-forming elements: the A-loop (which



(a) Typical kinase domain structure **(b)** DFG-in and DFG-out

Figure 4.2: Kinase structure overview. a) A typical kinase consists of a C-terminal lobe (bottom) and an N-terminal lobe (top) connected by a hinge region. The main pocket lies in between those two lobes (red ligand). P-loop (blue) and αC-helix (pink) are part of the N-lobe whereas the A-loop (orange) is the only flexible element of the C-lobe. **b)** The activity-determining element is the DFG motif which can exist in an active DFG-in position (top) or an inactive DFG-out configuration (bottom) which allows access to a hydrophobic backpocket.

carries the DFG motif), the P-loop and the αC-helix (Figure 4.3).

The models used in our study were generated using a defined pipeline (Schwarz *et al.*, 2019). In stage one, all available human kinase structures were analysed to determine the different classes of these flexible elements and to check for their distribution throughout the kinome. A class of a flexible element was defined either by previous work (Möbitz, 2015; Xu *et al.*, 2011; Brooijmans *et al.*, 2010) or by us after clustering observed conformations. If a class had been observed in a majority of kinase families, it was assumed to be suitable for kinome-wide modelling (all of the selected elements' classes are shown in Figure 4.3). Accordingly, the modelling pipeline that was built systematically generates homology models of kinases in several DFG-out conformations representing its conformational plasticity. Other

work had addressed this issue before (Kufareva and Abagyan, 2008; Xu *et al.*, 2011; Ung and Schlessinger, 2015) and the difference and novelty of our modelling approach comes from the systematic creation of template structures that represent the major states (and combinations) of the flexible structural elements (Figure 4.3).

The major structural feature of kinase binding pockets is whether the DFG motif's phenylalanine is pointing inwards or outwards and thus giving access to the DFG-out pocket underneath the αC-helix (Möbitz, 2015). Access to this pocket is also determined by the A-loop conformation. Most commonly an outward pointing phenylalanine is associated with a closed A-loop conformation ('closed type 2' and 'closed A-under-P') but an open A-loop conformation ('open DFG-out') also exists amongst the 'DFG-out' structures (Möbitz, 2015). Access to the DFG-out pocket is then restricted which is important for the type of inhibitor that can bind. Inhibitors that target DFG-in conformations are generally called 'type I' inhibitors, whereas ligands that reach into the hydrophobic side pocket, that opens up in most DFG-out conformations, are called 'type II' inhibitors. Small molecules that bind kinases at a site distant to the ATP-binding pocket are allosteric 'type III' inhibitors (Möbitz, 2015).

Even though only about 15% of human kinases have at least one DFG-out structure in the PDB and about 45% of kinases are covered with DFG-in structures (Schwarz *et al.*, 2019), with more than 500 human kinases (Möbitz, 2015), 15% is still lots of data for a protein family. This makes kinases ideal for a systematic exploration of their conformational ensembles. Studies indicate that many more than the kinases that are already found with a DFG-out crystal structure can be targeted by type II inhibitors (Zhao *et al.*, 2014) or are likely to be able to adopt a universal conformational space (Ung *et al.*, 2018). Some kinase states may be scarcely-populated but it can be crucial to model those as well because they might offer increased selectivity compared to highly-populated conformations (Guimarães *et al.*, 2011). Taking this into account, the systematic homology modelling pipeline (described in more detail in Methods section *Initial Data Generation*) generates kinase models in novel DFG-out conformations that had not been seen in the

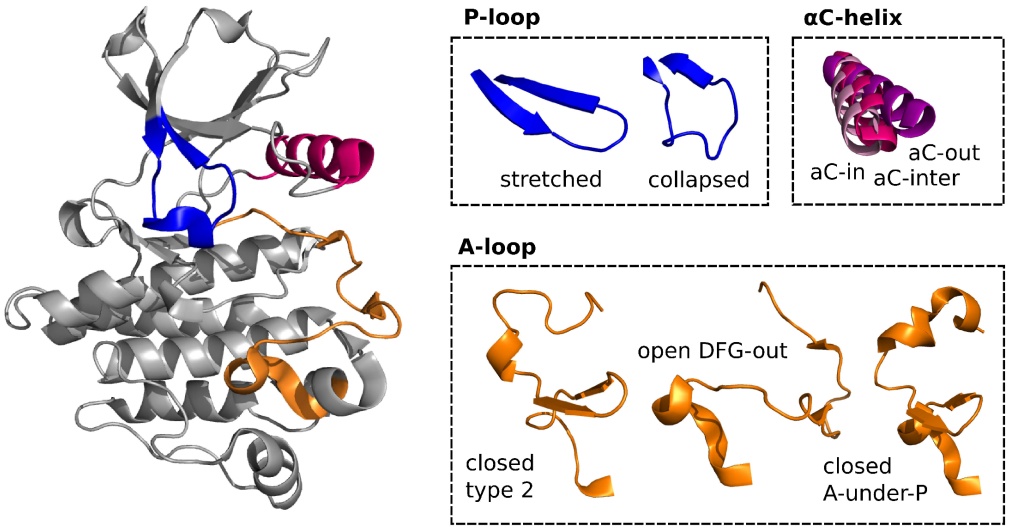


Figure 4.3: Major states of flexible kinase elements. With the P-loop found in two major conformations, the αC-helix in three, and the A-loop also in three, 18 (2*3*3) systematically different kinase models are created by our homology modelling pipeline.

PDB before (Schwarz *et al.*, 2019). Our subsequent exploratory docking study with FDA-approved type II inhibitors showed suitability of the homology models for further *in silico* studies.

4.2 Methods

4.2.1 Initial Data Generation

As described above, kinase homology model data was generated by determining flexible structural elements that are distributed in a kinome-wide manner, selecting template structures containing these elements and then building chimeric full-length kinase templates out of these.

Two different P-loop classes and three αC-helix classes were identified (see Figure 4.3). As both elements are part of the N-lobe, six PDB structures were selected that each carry a unique combination of the two elements' classes: 4QQ5_A (collapsed, aC-in), 2G2H_B (collapsed, aC-inter), 5HX6_A (collapsed, aC-out), 3VHK_A (stretched, aC-in), 4PMM_A (stretched, aC-inter) and 2W5B_A (stretched, aC-out). Just the templates' N-lobe sections were taken for building chimeric full-length templates (blue in Figure 4.4). The C-lobe was taken from an existing or modelled

DFG-in structure since it is rigid and very similar in DFG-in and DFG-out structures (green in Figure 4.4). The A-loop was cut out of all the DFG-in C-lobes and replaced with an A-loop of one of the three A-loop classes we had selected for modelling (orange in Figure 4.4). A-loop templates were taken from 3V5Q_A (closed type 2), 2HZI_A (open DFG-out) and 3BEA_A (closed A-under-P). The combination of six different N-lobes and three different A-loops resulted in eighteen chimeric templates for modelling the entire kinome in different DFG-out variants.

Molecular dynamics simulations revealed that conformational transitions between the different DFG-out states generally do not occur within trajectories of a few hundred nanoseconds length (for more details see Schwarz *et al.* (2019)). This underlined the potential of the developed homology modelling pipeline to generate relevant conformations of DFG-out kinase structures for subsequent *in silico* screening or binding site analysis studies.

4.2.2 Selection of Docking Test Set

PKIDB (Carles *et al.*, 2018) (http://www.icoa.fr/pkidb/) was used to select the FDA-approved (phase=4) type II (type=2) protein kinase inhibitors of ABL1 and KDR respectively, that had a published crystal structure (Table 4.1).

Table 4.1: Docking set of FDA-approved type II inhibitors for ABL1 and KDR

INN name	Ligand ID	PDB code	Canonical Smiles
Imatinib	STI	2HYY_A	Cc1ccc(cc1Nc2nccc(n2)c3cccnc3)NC(=O)c4ccc(cc4)CN5CCN(CC5)C
Nilotinib	NIL	3CS9_A	Cc1ccc(cc1Nc2nccc(n2)c3cccnc3)C(=O)Nc4cc(cc(c4)n5cc(nc5)C)C(F)(F)F
Ponatinib	0LI	3OXZ_A	Cc1ccc(cc1C#Cc2cnc3n2nccc3)C(=O)Nc4ccc(c(c4)C(F)(F)F)CN5CCN(CC5)C
Sorafenib	BAX	3WZE_A	CNC(=O)c1cc(ccn1)Oc2ccc(cc2)NC(=O)Nc3ccc(c(c3)C(F)(F)F)Cl
Axitinib	AXI	4AG8_A	CNC(=O)c1ccccc1Sc2ccc3c(c2)[nH]nc3/C=C/c4ccccn4

4.2.3 Receptor and Ligand Preparation

One PDB structure per ligand and target pair was chosen as a reference (see Table 4.1), waters were removed from all crystal structures and hydrogens added to all oxygen and nitrogen atoms. All ABL1 complexes (and homology models) were aligned to 2HYY_A and all KDR complexes (and homology models) to 3WZE_A.

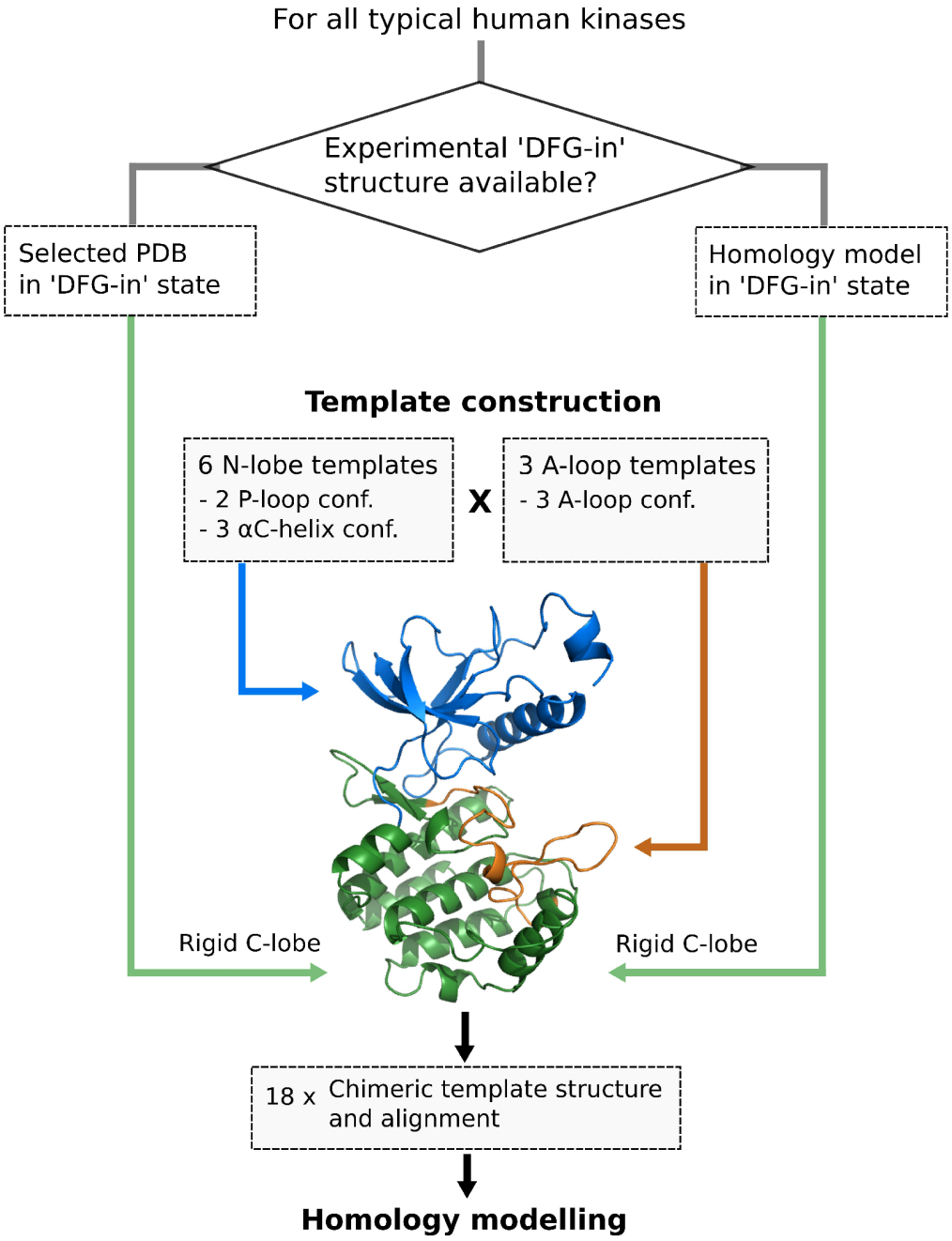


Figure 4.4: Homology modelling pipeline for modelling the entire kinome in different DFG-out conformations. Template structures are constituted by a C-lobe of a corresponding DFG-in structure (green), an N-lobe of one of the six selected N-lobe template structures (blue), and an A-loop of one the three selected A-loop templates (orange), resulting into eighteen chimeric template structures per kinase.

All receptor structures and ligands were prepared with Open Babel 2.4.1 (O'Boyle *et al.*, 2011) into *pdbqt* format. Note that the KDR homology model 9 was not

generated by the homology modelling program and that the reference structure of 0LI into ABL1 (PDB 3OXZ) is from *mus musculus*.

4.2.4 Re- and Crossdocking with Crystal Structures

Ligands were re-docked into their receptor crystal structure (see Table 4.1) and cross-docked into the crystal structures of the same kinase, so PDBs 3OXZ_A, 2HYY_A, 3CS9_A for ABL1 and 3WZE_A, 4AG8_A for KDR. Re- and crossdocking was performed with Smina (Koes *et al.*, 2013) without flexible side chains, the search space ('–autobox_add') around the respective ligand was set to 8Å (default: 4Å) and the exhaustiveness parameter's was set to 32 (default: 8). A maximum of 9 poses per docking were output. All-atom RMSDs were calculated against the crystal structure pose.

4.2.5 Docking Calculations

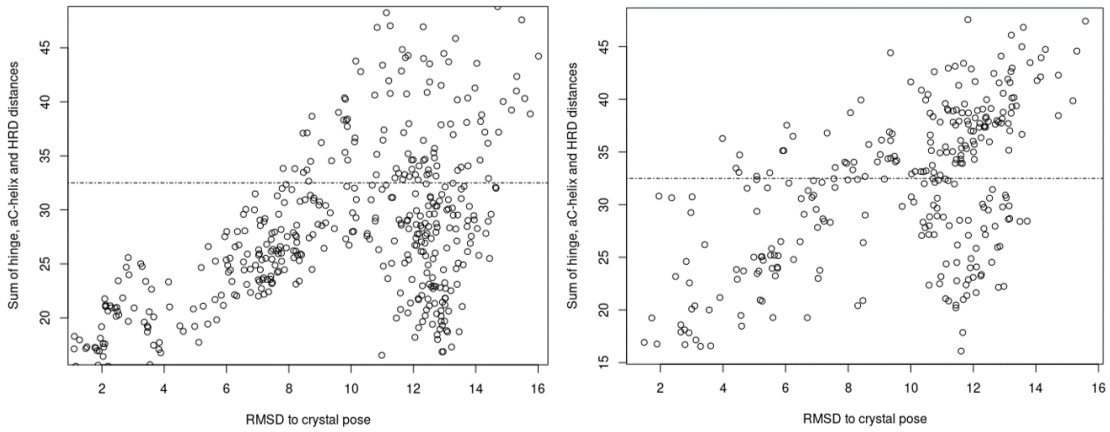
Docking was performed with Smina (Koes *et al.*, 2013) which is based on AutoDock Vina 1.1.2 (Trott and Olson, 2010). As homology models were generated without the presence of any ligand, the '–flexdist' argument was set to 2.5Å. Thus, every amino acid's side chain that possessed an atom within 2.5Å of the respective ligand's coordinates in the crystal structure (set by '–flexdist_ligand' parameter) was treated as flexible. The search space ('–autobox_add') around the respective ligand was set to 5Å (default: 4Å). The exhaustiveness parameter's default of 8 was used. A maximum of 9 poses per homology model were output.

4.2.6 Docking Analysis

All docking poses of an inhibitor were initially filtered for docking poses within the core binding cavity. This was achieved by measuring distances of the closest inhibitor atom to C_α atoms of the 'hinge donor' (ID 123 with respect to PDB 1ATP), αC-helix's Glu (ID 91), and HRD's His (ID 164). To prevent a bias towards poses between the αC-helix and HRD motif, the hinge's distance was double weighted.

Dependency of RMSD to the reference crystal structure pose and the measured distance sums are displayed in Figure 4.5.

Figure 4.5 updated



(a) ABL1 inhibitors **(b)** KDR inhibitors

Figure 4.5: Dependency of RMSD and minimal distances to binding pocket. Docking poses of **a)** STI, NIL and 0LI into ABL1 and **b)** of AXI and BAX into KDR. The filtering threshold (dashed line at 32.5Å) was applied to exclude binding poses obviously outside the binding pocket. Poses below this threshold were generally inside the binding pocket and either in orientation of the crystal pose (between 1 and 10Å RMSD) or in an inverted orientation (10 to 15Å RMSD).

Visual inspection revealed that poses with an RMSD around 12Å (and low distance sums) are those inside the binding pocket but in an inverted orientation compared to the crystal structure's binding pose. These poses were kept to resemble a real-world docking scenario. A distance cut-off of 32.5Å (dashed line) was chosen to exclude poses obviously outside of the binding pocket. This filtering step reduced the number of poses found for STI from 155 to 118, for NIL from 133 to 99, for 0LI from 151 to 119, AXI from 147 to 42 and for BAX from 153 to 107. All remaining docking poses were ordered by ascending energy and the top 5 docking poses of each inhibitor were examined in the presented analysis (see also Tables 4.2 and 4.3). A 'good matching' pose was defined to have a maximum of 2.2Å RMSD to the crystal pose.

4.3 Results

The homology modelling pipeline successfully generated DFG-out ensembles for over 95% of the human protein kinases (for more details see Schwarz *et al.* (2019)). There are many potential applications of the generated conformational ensembles, including *in silico* screening and binding site analysis studies. To test whether our model ensembles are useful for docking studies, we selected FDA-approved type II inhibitors of the two kinases ABL1 and KDR (5 cases in total; Table 4.1) and docked them into the generated DFG-out ensembles for those two kinases. We considered two things when assessing our models, are any of the (top-ranked) docking poses similar to the crystal structure pose the inhibitor had been seen in? And since all five inhibitors were observed binding a closed A-loop conformation (in a PDB structure), do we find more top-ranked docking poses in homology models with closed A-loop conformations?

4.3.1 Re- and Cross-docking

Before docking the inhibitors into the generated homology models, we checked whether our docking pipeline was working with crystal structures. For this, we performed re-docking of the five inhibitors into their respective crystal structures and cross-docking into the other crystal structures of the same kinase (ABL1/KDR). While re-docks must work as there is a known solution to the docking problem (the crystal pose), cross-docks may fail due to different kinase conformations in the different PDB structures. However, as all three structures of ABL1 are classified the same structural classes (A-loop: 'closed type 2', P-loop: 'collapsed' and αC-helix: 'αC-in') and both KDR structures are also classified in the same structural classes ('closed A-under-P', 'stretched' and 'αC-in'), cross-docking should in these cases succeed.

Figure 4.6 shows the top-ranked docking pose for each ABL1 inhibitor and receptor structure. Re-docking of Ponatinib (0LI) back into PDB 3OXZ_A, Imatinib (STI) into 2HYY_A and Nilotinib (NIL) into 3CS9_A is shown on the diagonal (Figure 4.6a, e and i). For all three ABL1 re-docks a good pose was ranked top by

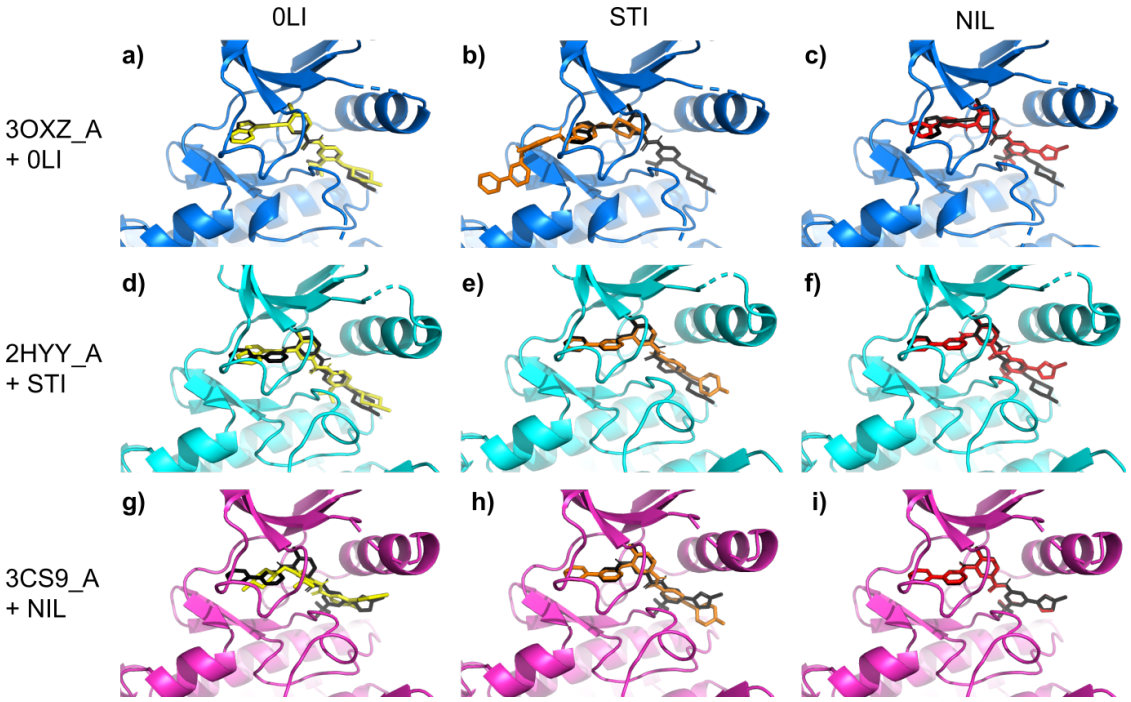


Figure 4.6: Re- and cross-docking of inhibitors 0LI, STI and NIL into their receptor structures PDBs 3OXZ_A, 2HYY_A and 3CS9_A. Rows show a crystal structure and its ligand: 3OXZ_A in blue, 2HYY_A in cyan, 3CS9_A in magenta and crystal structure ligand poses always in black. Columns add top-ranked docking poses for 0LI in yellow, STI in orange and NIL in red. Thus, **a)**, **e)** and **i)** are re-docks into a ligand's own crystal structure and the others are cross-docks into other crystal structures of the same kinase. RMSDs to crystal structure poses (diagonal) of the docked ligand: **a)** 0.68Å, **b)** 12.73Å, **c)** 1.69Å, **d)** 1.05Å, **e)** 1.25Å, **f)** 0.71Å, **g)** 12.00Å, **h)** 1.32Å and **i)** 0.60Å. All re-docks recovered the crystal pose with their top binding pose. Cross-docks worked well for all pairs but STI into 3OXZ_A (**b)**) and 0LI into 3CS9_A (**g)**). However, for both pairs the pose ranked second was in good agreement with the crystal pose (2.03Å for STI/3OXZ_A; 1.49Å for 0LI/3CS9_A).

the docking software: 0LI/3OXZ_A with 0.68Å RMSD, STI/2HYY_A with 1.25Å and NIL/3CS9_A with 0.60Å. For the cross-docks, good poses were found at the top position as well, with the exception of STI into 3OXZ_A (Figure 4.6b) and 0LI into 3CS9_A (Figure 4.6g) where the pose ranked second was a good match. Even though all flexible elements of the three receptor structures were classified the same, subtle structural differences could explain the slight mis-ranking.

KDR re-docking of Sorafenib (BAX) into PDB 3WZE_A and Axitinib (AXI) into PDB 4AG8_A also worked with RMSDs of 0.67Å and 0.43Å (Figure 4.7a and d). The same is true for cross-docking BAX into 4AG8_A (RMSD 0.95Å; Figure

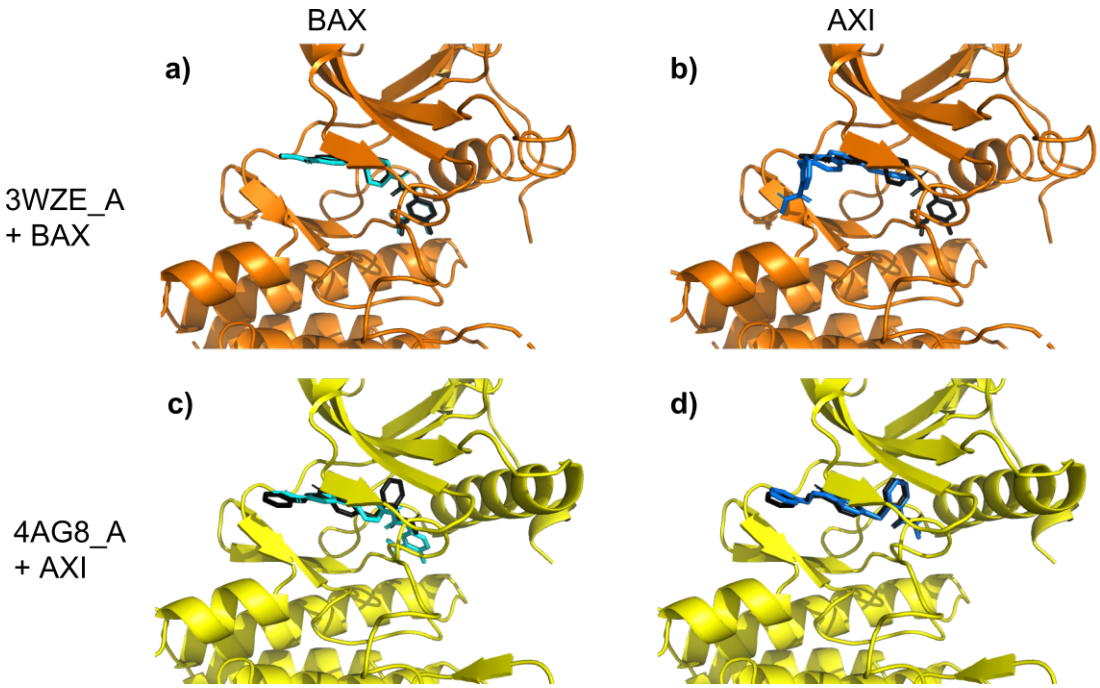


Figure 4.7: Re- and cross-docking of inhibitors BAX and AXI into their receptor structures PDBs 3WZE_A and 4AG8_A. Rows show a crystal structure and its ligand: 3WZE_A in orange, 4AG8_A in yellow and crystal structure ligand poses always in black. Columns add top-ranked docking poses for BAX in cyan and AXI in blue. Thus, **a)** and **d)** are re-docks into a ligand's own crystal structure and the others are cross-docks into another crystal structure of the same kinase. RMSDs to crystal structure poses (diagonal) of the docked ligand: **a)** 0.67Å, **b)** 10.53Å, **c)** 0.95Å and **d)** 0.43Å. Both re-docks recovered the crystal pose with their top binding pose. Cross-docking BAX into 4AG8_A (**c)**) worked well, while for AXI into 3WZE_A (**b)**) no matching pose could be recovered within the top 9 poses.

4.7c) but not for AXI into 3WZE_A. No similar pose to the known crystallographic pose was found in the top 9 poses.

As all ABL1 and KDR re-docks generated correct poses and only one out of eight cross-docks failed, we felt that our pipeline was suitable for examining the behaviour of kinase inhibitors docked into structural ensembles of ABL1 and KDR.

4.3.2 Ensemble Docking

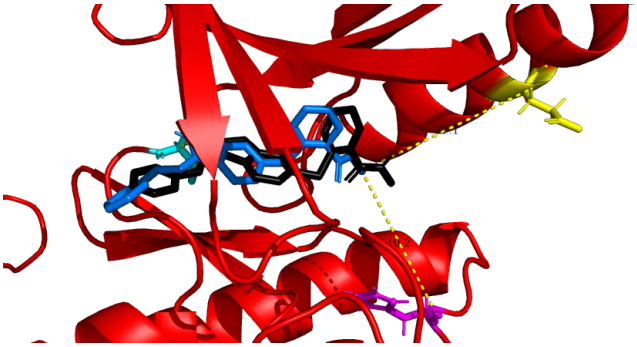
Ranking ligand poses accurately is difficult (Boyles *et al.*, 2020), therefore here we examine the top 5 binding poses per kinase ensemble.

In order to remove poses that can be trivially identified as outside the binding pocket, we measured the distances between three anchor residues in the protein

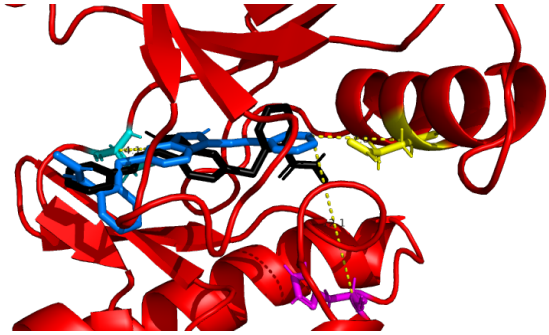
model and the nearest ligand atom. The three anchor points were the C_αs of the 'hinge donor' (cyan), αC-helix's glutamic acid (yellow) and HRD's histidine (magenta). Figures 4.8a, b and c display examples of docking poses inside, inside but inverted (compared to the crystal structure pose) and outside the binding pocket. To create a metric we summed up the distance from the αC-helix's Glu to its nearest ligand atom, HRD's His to its nearest ligand atom and twice the distance from the 'hinge donor' to its nearest ligand atom. The double weighting of the 'hinge donor' distance was used because the αC-helix's Glu and the HRD's His lie on one side of the pocket and the 'hinge donor' on the other side. Without double weighting poses would be favoured that are mostly underneath the αC-helix (between the Glu and His anchor points). The threshold for excluding poses over a certain sum of distances was determined by plotting this sum of distances against ligand RMSDs (see Methods Figure 4.5) and set to 32.5Å. Poses in Figure 4.8a, b and c have distance sums of 30.66Å, 27.15Å and 54.26Å, correctly identifying the pose outside the binding pocket. The remaining poses after filtering were analysed to see if any good matches to the known crystal pose were present in the top poses.

As access to the side pocket underneath the αC-helix is restricted in open A-loop conformations (i.e. 'DFG-in' and 'open DFG-out') and since none of the tested type II inhibitors is known to bind in such an open conformation, all inhibitors are expected to bind preferentially into models with closed A-loop conformations (closed type 2 or closed A-under-P).

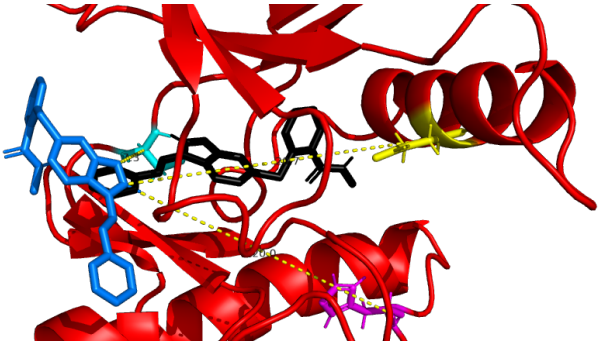
Docking poses with good agreement to the crystal structure (i.e. RMSD $<$ 2.2 Å) were found within the top 5 docking poses for four out of the five cases (Tables 4.2 and 4.3) and at least in the case of the ABL1 structures many of the corresponding receptor models have the same structural features / classification as the respective crystal structure (asterisk in Table 4.2). For the three ABL1 cases, the receptor structures are all classified as A-loop: 'closed type 2', P-loop: 'collapsed' and αC-helix: 'αC-in'. For two of the respective docking cases (Imatinib (STI) and Nilotinib (NIL)), a good docking pose and the correct model class were



(a) Pose inside the binding pocket



(b) Pose inverted but inside binding pocket



(c) Pose outside the binding pocket

Figure 4.8: Novel KDR conformations with examples of docking poses inside, inverted and outside the binding pocket. KDR homology models (HM) are shown in red with the AXI crystal structure pose of PDB 4AG8_A in black and docking poses in blue. Distance measurements of the closest ligand atom to the three anchor points (C_αs of 'hinge donor' (cyan), αC-helix's Glu (yellow) and HRD's His (magenta)) are indicated with dashed yellow lines. **a)** KDR HM7 (closed type 2, collapsed, αC-out) with AXI docking pose 2 shows an inhibitor orientation inline with the crystal pose (2.36Å RMSD). **b)** Pose 5 of docking AXI into KDR HM6 (open DFG-out, collapsed, αC-inter) is inside the binding pocket but displays an inverted orientation compared to the crystal pose (10.58Å RMSD). **c)** AXI docking pose 8 for KDR HM4 (closed type 2, collapsed, αC-inter) is an example of a pose outside the binding pocket which is removed by the anchor point distance filtering (see Methods section *Docking analysis* for more details) (12.66Å RMSD to crystal pose).

Table 4.2: Top docking poses into ABL1

Ligand	Pose name	Classification of homology model[1]	RMSD[2] [Å]	Affinity [kcal/mol]
STI	HM_10_pose_1	closed type 2, stretched, αC-in	12.42	-11.4
	HM_1_pose_1	closed type 2, collapsed, αC-in *	3.41	-11.2
	HM_1_pose_2	closed type 2, collapsed, αC-in *	2.12	-11.0
	HM_1_pose_3	closed type 2, collapsed, αC-in *	2.12	-11.0
	HM_4_pose_1	closed type 2, collapsed, αC-inter	1.52	-11.0
NIL	HM_10_pose_1	closed type 2, stretched, αC-in	11.82	-12.3
	HM_1_pose_1	closed type 2, collapsed, αC-in *	3.85	-12.0
	HM_10_pose_3	closed type 2, stretched, αC-in	12.43	-12.0
	HM_10_pose_4	closed type 2, stretched, αC-in	6.40	-12.0
	HM_1_pose_2	closed type 2, collapsed, αC-in *	2.12	-11.8
0LI	HM_8_pose_1	open DFG-out, collapsed, αC-out	13.11	-11.3
	HM_5_pose_1	open DFG-out, collapsed, αC-inter	6.93	-10.9
	HM_8_pose_2	open DFG-out, collapsed, αC-out	5.42	-10.9
	HM_18_pose_1	closed A-under-P, stretched, αC-out	7.95	-10.6
	HM_5_pose_2	open DFG-out, collapsed, αC-inter	7.06	-10.5
	... (ommiting ranks 6 to 13)			
	HM_16_pose_1	closed type 2, stretched, αC-out	1.12	-9.6

[1] Models that have the same classification with respect to A-loop, P-loop and αC-helix as the crystal structure of the respective ligand (Table 4.1) are marked with an asterisk.
[2] To respective crystal pose (all atom RMSD).

found within the top 5 docking poses and for all three cases all receptor models of the top poses have at least a closed A-loop conformation.

For KDR docking the respective crystal structures are both classified as 'closed A-under-P', 'stretched' and 'αC-in'. For the case of Sorafenib (BAX), a good docking pose was obtained for the 'closed type 2' model which is similar in shape to the 'closed A-under-P' conformation and 3 out of the top 5 selected structures have a closed A-loop conformation. In the case of Axitinib (AXI) a matching pose was found in the top 5 docking solutions but the respective receptor model has a different A-loop orientation than the crystal structure.

Overall, the results imply that the generated conformational ensembles of homology models can be employed for docking calculations, especially when the top poses are found for the same receptor classification state. Having an ensemble that

Table 4.3: Top docking poses into KDR

Ligand	Pose name	Classification of homology model	RMSD[2] [Å]	Affinity [kcal/mol]
AXI	HM_4_pose_1	closed type 2, collapsed, αC-inter	6.70	-10.0
	HM_1_pose_1	closed type 2, collapsed, αC-in	5.60	-9.4
	HM_8_pose_1	open DFG-out, collapsed, αC-out	1.95	-9.4
	HM_1_pose_2	closed type 2, collapsed, αC-in	10.35	-8.9
	HM_11_pose_1	open DFG-out, stretched, αC-in	10.59	-8.8
BAX	HM_8_pose_1	open DFG-out, collapsed, αC-out	6.90	-10.6
	HM_1_pose_1	closed type 2, collapsed, αC-in	11.44	-10.3
	HM_18_pose_1	closed A-under-P, stretched, αC-out	5.22	-10.1
	HM_1_pose_2	closed type 2, collapsed, αC-in	1.72	-10.0
	HM_8_pose_2	open DFG-out, collapsed, αC-out	5.15	-9.9

[2] To respective crystal pose (all atom RMSD).

represents the major states of a protein but also the less-populated conformations of its structural landscape will be particularly useful for families with many similar binding pockets and thus, selectivity issues. Generating many different conformations of a protein can be achieved in many different ways (homology modelling is just one of them) but validating the biological relevance of the resulting conformers is always hard. Kinases may be the protein family where this is potentially the easiest because of the amount of data that has been generated for kinases.

4.4 Discussion

The developed homology modelling pipeline systematically generates homology models (HMs) of kinases in the DFG-out state by building chimeric templates that represent all combinations of the flexible features that we identified before (Figure 4.3). Only individual flexible feature classes that were observed throughout multiple kinase families were considered for this combinatoric approach, resulting in eighteen different HM classes (3 (A-loops) * 2 (P-loops) * 3 (αC-helix states)). For over 95% of the kinome, structural ensembles could be generated this way. This systematic modelling approach efficiently samples the major conformations of kinases which are of potential value for drug design efforts as it allows the generation

of kinase conformations that have not been observed in a crystal structure of a certain kinase yet. Even though a kinase conformation has not been seen, a kinase may be able to adopt such a conformation at an energetic penalty (Möbitz, 2015; Haldane *et al.*, 2016) and thus, it may exist in small population sizes or when induced by an inhibitor (Ung *et al.*, 2018).

The results of our exploratory docking study indicate that this type of homology modelled conformational ensemble can be employed for docking calculations and analysis of the models may provide insights into selectivity-determining features which might be only addressable in scarcely populated conformations (Guimarães *et al.*, 2011).

Homology modelling may be a good approach for generating conformational ensembles for protein families with a diverse set of known structures. The kinase protein family is potentially the ideal family in which to attempt the systematic generation of ensembles as it has many known structures in various conformations because it is of high interest to pharmaceutical research. This research interest has led to a bias towards some kinases (tyrosine kinases (TK)) but overall has produced a large amount of structural data relating to kinases which provide a form of ground truth for generating conformational ensembles. This wealth of structural data does not exist for other protein families (yet) and sequence-based *de novo* ensemble prediction would offer a wider applicability. Being able to predict multiple biologically relevant conformations of any protein would be desirable for drug discovery research and it might be achieved soon considering the recent advances in static protein structure prediction (e.g. Jumper *et al.*, 2021).

Having such ensembles will be beneficial for many areas but it is missing one important aspect from a basic research perspective: how did the protein reach this folded ensemble? The protein folding problem is separate and distinct from the protein structure prediction problem and like the allostery and flexibility problems previously considered, it is necessary to understand basic biological principles and the functions of proteins. The next chapter describes an initial analysis of the predicted folding pathways of helical transmembrane proteins.

5

Transmembrane Protein Folding Pathway Prediction

Contents

5.1 Background

Crucial for understanding a protein's function or misfunction is its 3D structure, but this shape represents only one data point in the protein's structural and functional space. The structure observed experimentally is the starting point for many other analyses that investigate protein function; for example the generation of structural ensembles for allostery analysis or drug discovery studies. The 3D structure of a protein is believed to be encoded in its sequence (Anfinsen, 1973)

and with co-evolution analysis, and most successfully with AlphaFold2 (Jumper *et al.*, 2021), it has been demonstrated that static structure can often be inferred from sequence information. As described in the introduction, the protein structure prediction problem is now largely solved but while nature 'knows' how to fold the primary sequence into the final fold, its mechanisms and pathways are not yet well understood. The pathway of folding is crucial for some diseases, e.g. Alzheimer's (Polychronidou *et al.*, 2020), as well as being a fundamental process in biology.

As laid out above, structural information can be extracted from sequence data, including dynamic (flexibility) information (see chapter 3). In this chapter, we investigate whether folding (pathway) information can also be extracted from sequence information. Our case study for this is the conservation of folding pathways of α-helical membrane proteins.

The folding of membrane proteins is controlled by a relatively small number of thermodynamic constraints as they are produced in an aqueous environment (by ribosomes in the cytosol) but perform their function in or next to the membrane. Due to the high content of hydrophobic amino acids in membrane proteins, they are generally inserted into membranes in a co-translational manner to reduce protein aggregation. Thermodynamics suggest that unfolded hydrophobic helices associate to the membrane interface where they form their secondary structure (see Introduction). This formation of secondary structure generally occurs before insertion into the membrane as it is energetically favourable compared to the insertion of unfolded polypeptide. Nevertheless, translocases and insertases (translocons) exist to reduce energy barriers and facilitate the correct localisation, insertion and folding into a protein's target membrane.

As thermodynamics favours secondary structure formation and enzymes speed up membrane insertion of α-helices, the final steps of α-helical membrane protein folding can be assumed to be mainly dependent on helix-helix associations within the lipid bilayer. TMPfold (TM Protein Folding) (Lomize *et al.*, 2020) is a package that uses this assumption to predict folding pathways based on helix-helix association free energies inferred from an input structure. In this chapter, we describe an initial

analysis of TMPfold predictions with the aim of investigating conservation of folding pathways. Predicted pathways as well as the underlying helix-helix association energies were compared between protein pairs within families and between families with the same number of transmembrane helices. While a conservation trend could be seen in some of these comparisons when analysing predicted pathways, the signal was ambiguous and data dependent. Comparisons of helix-helix energies displayed a much clearer signal of conservation, highlighting the need for more development on the methodology of inferring pathways from association energies.

5.2 Methods

5.2.1 Dataset

The Orientations of Proteins in Membranes (OPM) database contains transmembrane and membrane-associated proteins and peptides, and provides additional annotation and classification alongside the deposited PDB structure (Lomize *et al.*, 2012). The database's main feature is the prediction of the positioning with respect to the membrane based on optimisation of free energy of the transfer from water to the lipid bilayer (Lomize *et al.*, 2006). For our work the classification feature was important as it allowed us to select only multi-pass transmembrane proteins with α-helical structure.

Only proteins described by OPM as alpha-helical, polytopical TM proteins were analysed. PDB structures of these proteins were downloaded from https://storage.googleapis.com/opm-assets/pdb/tar_files/Alpha-helical_polytopic.tar.gz on 27/06/2021.

Protein families were defined according to the OPM 'families' CSV downloaded from https://lomize-group-opm.herokuapp.com//families?fileFormat=csv on 8/7/2021. When referenced in this chapter, an OPM family ID refers to the internal ID used by the server's API and CSV tables; the OPM website's displayed numbering scheme is referenced as 'ordering' in CSV files and not identical to the internal IDs.

For each family a JSON file was downloaded via lomize-group-opm.herokuapp.com/families/[opm_family_id]/primary_structures on 8/7/2021. Those contain protein

names, PDB codes, resolution, determination technique and Uniprot codes. No NMR structures and only proteins with PDB structures with an X-ray or EM resolution equal or better than 3Å were considered for analysis. Potential PDB structures were further filtered to only contain unique Uniprot codes within a family.

Transmembrane segments were defined following the OPM 'subunits' CSV downloaded from https://lomize-group-opm.herokuapp.com//structure_subunits ?fileFormat=csv on 8/7/2021. The order of segments in TMPfold input files is taken from this subunits table.

5.2.2 Transmembrane Folding Pathway Prediction

TMPfold (Lomize *et al.*, 2020) was downloaded from https://console.cloud.google.com /storage/browser/opm-assets/tmpfold_fmap_code on 10/6/2021 and run with default parameters. Input files for each PDB structure contained all segments listed in the subunits CSV.

5.2.3 Folding Pathway Analysis

Determination of transmembrane helix count

For each family the most common transmembrane (TM) helix count was determined amongst all segments (chains) of PDB structures that passed prefiltering (resolution and unique Uniprot ID) and ran successfully through TMPfold. If a structure contained two or more segments of a specific TM count, this contributed only once towards determining the most common TM helix count, e.g. if structure XYZ had chain A with TM helix count 5, chain B with 2 and chain C with 5, the structure was considered to have TM helix count 2 or 5; 5 is not counted twice.

If two TM helix counts are equally present in the family's structures, the greater one was selected, e.g. if 15 structures had TM helix count 7 and 15 structures TM count 5, that family was assigned to TM helix count 7 and only structures with segments of that count were further analysed. Furthermore, only families that had at least three PDB structures (with a segment that had the family's most common TM helix count) were further analysed. For each PDB structure that

contained multiple segments with the family's most common TM count, the first (order defined by 'subunits' CSV) segment with this TM count was selected to be analysed. All analysed PDB structures of TM helix counts with two or more different families can be found in Table C.1.

Pathway and energy matrix comparisons

For the quantification of differences of predicted pathway and helix-helix association energies, pathways were converted into matrices and a distance metric was defined. Predicted stepwise folding pathways (see Figure 5.1a) were represented by a matrix with rows and columns being the transmembrane helices (i,j) and the values being the folding step's number in which those two helices associate. Association is either the formation of a two-helix bundle (folding nuclei), the addition of a single helix to an existing helix bundle or joining of two helix bundles. For example, if helix 1 and helix 2 associate in the predicted pathway's first step, the matrix cells '1,2' and '2,1' contain a 1 (step 1) and if helix 3 joins this helix bundle in the second pathway step, cells '1,3', '2,3', '3,1' and '3,2' contain a 2 (step 2). This procedure is done for all predicted pathway steps and yields a single matrix representation of the folding pathway.

Protein segments with the same TM helix count are compared by computing the sum of absolute differences (SAD) between their two matrices. To yield a distance metric between zero and one, the SAD is normalised by the sum of absolute values of both matrices (e.g. for matrices A and B):

$$d(A,B) = \frac{\sum |A_{i,j} - B_{i,j}|}{\sum |A_{i,j}| + \sum |B_{i,j}|}$$

with i, j being the TM helix indices from one to the TM helix count. The same distance metric was used for the comparison of two helix-helix energy matrices (values directly taken from TMPfold output).

Intra- and inter-family distance distributions

For the analysis of distance distributions of families with the same TM helix count, matrix-matrix distances were grouped into intra- and inter-family comparisons. To test for the difference between those two distance distributions (per TM count), the Mann-Whitney U test (two-sided) was applied as implemented in the Stats module of SciPy version 1.3.1 (Virtanen *et al.*, 2020). Common language effect size f was calculated as: $f = \frac{U_1}{n_1 n_2}$ with test statistic U and distribution sizes n_1 and n_2.

5.3 Results

5.3.1 TMPfold Predictions

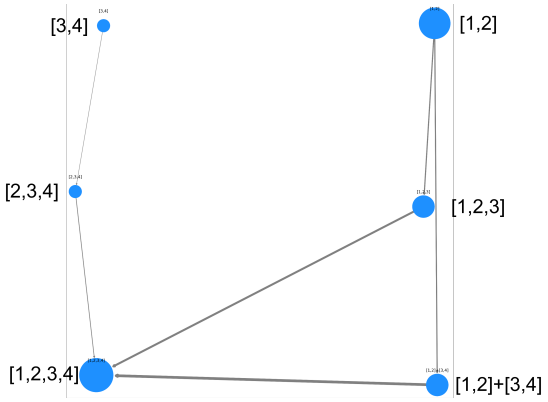
TMPfold predicts pairwise helix-helix association energies and uses these to propose a folding/assembly pathway for a given helical membrane protein structure (as an example, PDB 6W1J_A in Figure 5.1a). These pathways are output in a stepwise manner but can also be represented as a state network where a state shows which helices have associated already (OPM family 22 is shown as an example in Figure 5.1b). The visual state network representation can highlight some conserved folding pathways within a protein family, for example in OPM family 22 ('Ligand-gated ion channel of neurotransmitter receptors') the majority of pathways form the 1,2-helix bundle before forming the 3,4-helix bundle. Subsequent steps are less conserved with either adding helix 3 directly or forming the 3,4-bundle before joining this with the 1,2-bundle.

5.3.2 Matrix Distances for Measuring Pathway Similarity

While a visual representation is good for a qualitative view of the data, more robust metrics are needed to analyse the data in a quantitative way. For this, we represented a predicted folding pathway as a matrix with rows and columns being the transmembrane helices and the cell values being the folding step number in which those two helices associate. For example, if helix 1 and 2 associate in the pathway's first step, the matrix cells '1,2' and '2,1' contain a 1 and if helix 3 joins

```
Helix-helix association free energies:
       1     2     3     4
  1   0.0  -5.6   1.0  -2.7
  2  -5.6   0.0  -4.9   0.6
  3   1.0  -4.9   0.0  -0.5
  4  -2.7   0.6  -0.5   0.0
TM helix assembly pathway:
Step  1  -5.6  1+2
Step  2  -9.6  (1+2)+3
Step  3 -12.2  ((1+2)+3)+4
Pathway: (((1+2)+3)+4)
```

(a) TMPfold output example (PDB 6W1J_A)



(b) Predicted pathways of OPM family 22

Figure 5.1: TMPfold predictions example. a) TMPfold yields helix-helix association energy predictions (matrix) and proposes a stepwise assembly (folding) pathway. **b)** OPM family 22 ('Ligand-gated ion channel of neurotransmitter receptors') has seven protein structures that remained after filtering for resolution and unique Uniprot IDs. The seven predicted folding pathways are shown here in a state-wise representation (as opposed to the stepwise output of TMPfold seen in a)); helices within square brackets are associated as a bundle (e.g. [2,3,4]) and the plus indicates that two bundles are co-existing but have not associated yet (e.g. [1,2]+[3,4]). Point size is proportional to the to the number of predicted pathways that share that state. This means that most predicted pathways in family 22 first form the 1,2-helix bundle before forming the 3,4-helix bundle. Subsequent steps are more diverse with either adding helix 3 ([1,2,3]) or forming the 3,4-bundle ([1,2]+[3,4]) before joining this with 1,2 ([1,2,3,4]).

this helix bundle in the second pathway step, cells '1,3', '2,3', '3,1' and '3,2' contain a 2. This procedure is done for all predicted pathway steps and yields a single matrix representing the stepwise folding pathway.

For proteins with the same number of transmembrane helices these matrices can be compared by computing a distance metric. The sum of absolute differences (SAD) were used for this purpose and normalised to distances between zero and one (see Methods). When comparing the pathway matrices of all proteins of one family with all matrices of another family a distribution of distances can be obtained. The same distribution of distances for all comparisons within a protein family allows an analysis if (predicted) folding pathways are more conserved within a family than between families. Figure 5.2 shows the within-family and between-family distributions for all OPM families with three transmembrane helices (that had at least three structures that passed the filtering steps).

Figure 5.2ab split into separate figures

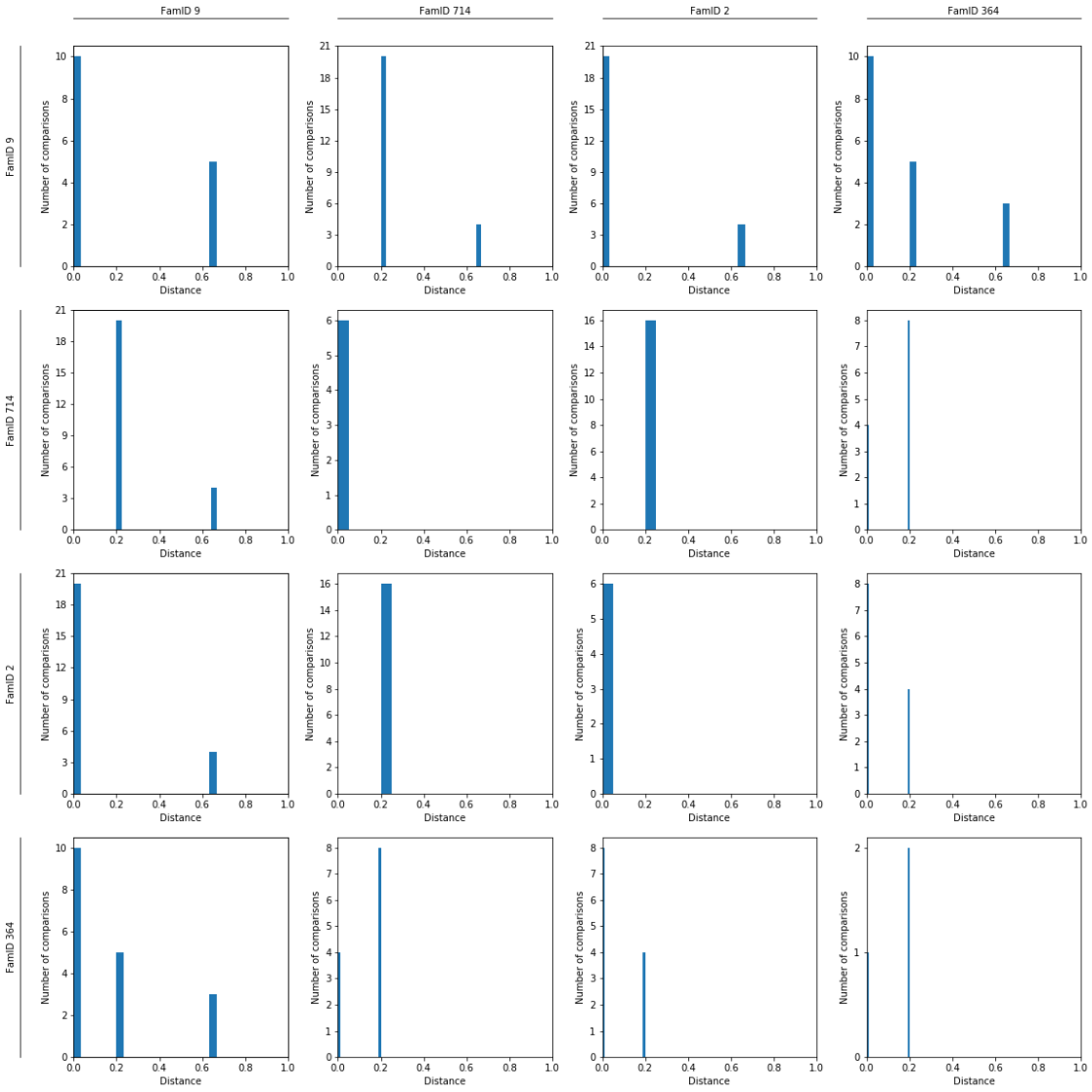


Figure 5.2: Pathway matrix comparisons of families with 3 TM helices. Normalised distances (sum of absolute differences; x-axes) are computed between pathway matrices of all proteins of one family and all proteins of another family (diagonal has within family comparisons). Order of OPM family IDs from left to right and top to bottom: 9, 714, 2, 364. Pathway matrices are a stepwise representation of the predicted folding pathway with an integer (step number) indicating the helices that associate in a given step, i.e. if helix 1 and 2 associate in the pathway's first step, the cells '1,2' and '2,1' contain a 1; if helix 3 joins this helix bundle in the second pathway step, cells '1,3', '2,3', '3,1' and '3,2' contain a 2; the diagonal has zeros.

Within-family comparisons are on the diagonal (from top left to bottom right) and it can be seen that families 714 and 2 only have proteins with identical predicted folding pathways (all matrix-to-matrix distances are zero). Furthermore, between-family distributions tend to have a higher mean than within-family distributions.

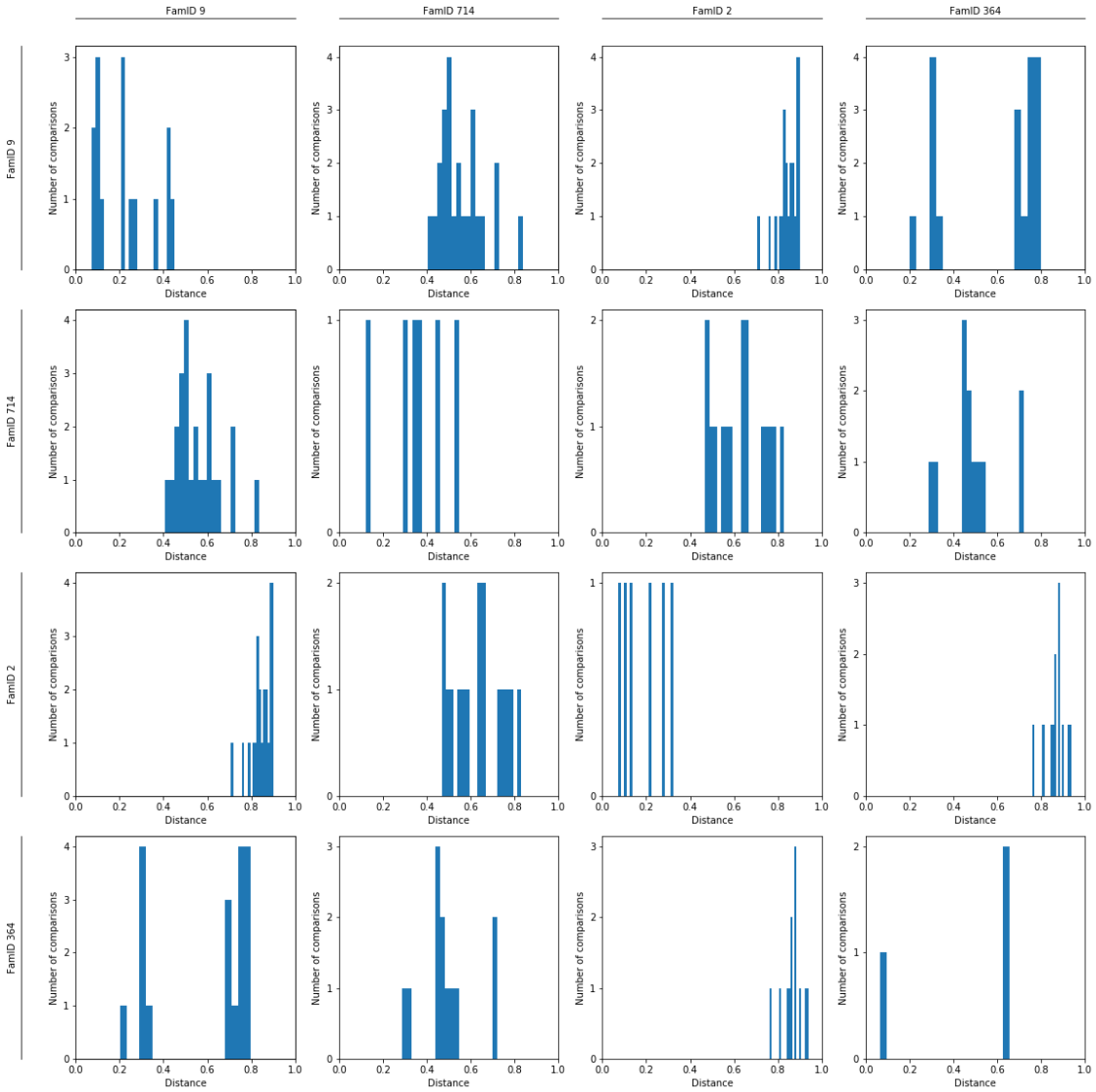


Figure 5.3: Energy comparisons of families with 3 TM helices. Normalised distances (sum of absolute differences; x-axes) are computed between energy matrices of all proteins of one family and all proteins of another family (diagonal has within family comparisons). Order of OPM family IDs from left to right and top to bottom: 9, 714, 2, 364. Energy matrices are taken from TMPfold output (see Figure 5.1a) and contain the predicted helix-helix association energies.

While the order of helix associations defines the folding pathway, it is also a condensation of information as the underlying data are the predicted helix-helix association energies. For example, small differences in helix-helix energies can have a great impact on the predicted pathway when near a threshold value. Therefore, we wanted to investigate matrix differences between the helix-helix energy matrices as well. Due to the non-integer nature of the energy matrices, more subtle differences

between proteins can be observed. For example, when comparing pathway matrices, the families 714 and 2 have identical pathways (within-family distances equal zero, see Figure 5.2), but when comparing energy matrices both families show distance distributions with a low mean but not all distances equal to zero (see Figure 5.3). Although the difference in means between within- and between-family distributions is more visible in the plot comparing energy matrices, a statistical analysis checking whether a potential difference is statistically significant is needed.

5.3.3 Statistical Significance of Differing Matrix Distances

At least two OPM families with at least three proteins per family could be found for transmembrane helix counts 3, 4, 5, 6, 7, 10, 11, 12 and 14. The distance distribution plots for TM helix count 3 are shown above (Figure **??**) and further distance distributions ordered by TM helix count can be found in Figures C.1-C.8. All of these distributions are part of the statistical analysis that tested if within-family matrix comparisons are different than between-family matrix comparisons. For this, grouped by TM helix count, the within-family ('intra') distributions were pooled and all of the between-family ('inter') distributions as well. The number of protein-to-protein comparisons per pool is shown in Table 5.1. The amount of data between TM helix counts differs greatly with some groups having hundreds of intra- and inter-family comparisons, e.g. TM helix counts 6 or 7, and others having only about a dozen, e.g. TM helix count 5.

The different distance means of intra- and inter-family pools can be seen in Table 5.2, for pathway as well as energy matrix comparisons. The Mann-Whitney U test was used to test if randomly drawing from the intra- or inter-family distributions yields a greater matrix-to-matrix distance in either of the investigated data types. The hypothesis that intra-family distances are smaller than inter-family ones could be supported in some cases and had to be rejected in others. More specifically, for comparisons of pathway matrices, only TM helix counts 6, 7 and 11 showed a p-value $< 10^{-5}$ and even in those cases, TM helix count 7 has only an effect size of 0.38 (0.5 indicates no difference). The picture changes when analysing the energy

Table 5.1: OPM families with more than three proteins ordered by TM helix count.

TM helix count	OPM family IDs	intra_n	inter_n
3	2; 9; 364; 714	30	106
4	22; 254; 265; 279; 567	39	171
5	3; 532	9	12
6	5; 15; 99; 237; 301; 682; 701	188	847
7	13; 14; 242; 667; 691; 823	471	1069
10	25; 27; 30; 302; 675	15	90
11	4; 21	55	50
12	11; 23; 24; 78; 281; 327; 656; 1010	70	560
14	18; 333	16	20

Table 5.2: Intra- and inter-family matrix distances of predicted folding pathways and helix-helix association energies.

TM helix count	Predicted pathways					Predicted helix-helix energies				
	intra_fam mean	inter_fam mean	p-value	U	f	intra_fam mean	inter_fam mean	p-value	U	f
3	0.12	0.17	6.9E-03	1122.5	0.35	0.27	0.67	3.7E-13	205	0.06
4	0.08	0.10	7.0E-02	2729	0.41	0.28	0.38	4.1E-05	1930	0.29
5	0.06	0.06	9.1E-01	56	0.52	0.15	0.17	2.7E-01	38	0.35
6	0.13	0.40	3.4E-28	38823.5	0.24	0.18	0.56	5.4E-91	4613	0.03
7	0.10	0.12	5.5E-13	193756.5	0.38	0.19	0.35	1.5E-168	29225.5	0.06
10	0.23	0.26	3.6E-01	574	0.43	0.28	0.58	1.1E-06	143	0.11
11	0.19	0.67	9.2E-19	0	0	0.10	0.65	1.2E-18	0	0
12	0.09	0.12	1.4E-05	13360.5	0.34	0.25	0.58	2.9E-40	527	0.01
14	0.29	0.31	5.3E-01	140	0.44	0.22	0.67	3.8E-07	0	0

matrices, with TM helix counts 3, 6, 7, 10, 11, 12 and 14 having p-values $< 10^{-5}$ and of those, the maximal effect size being 0.11 (TM helix count 10).

While for TM helix count 4, at least for energy matrix comparisons, the p-value is $< 10^{-4}$, for TM helix count 5 the p-values of pathway and energy comparisons are both not below 0.1. It is possible that for an unknown reason this group of proteins shows no conservation of its association pathway or propensity, but it is also the group with the least amount of data (only 9 intra-family and 12 inter-family comparisons).

5.4 Discussion

TMPfold is the first software for the prediction of membrane protein folding pathways (Lomize *et al.*, 2020). These pathway predictions are based on helix-helix association free energies inferred from an input structure. Pairwise association energies and the sequential nature of co-translational membrane protein folding inform TMPfold pathway predictions. While the helix-helix energy predictions have been experimentally validated, verifying folding pathway predictions relies on case studies of a small number of proteins, with agreement in some cases and deviations in others. Based on this evidence, the authors term them tentative folding pathways. Nevertheless, the software provides a first tool to study membrane protein folding on a larger scale. In this chapter, we have described an initial analysis on the conservation of helix-helix association energies and pathways.

It is known that some proteins do not fold following the assumptions that TMPfold has built-in, i.e. all transmembrane (TM) helices insert into the membrane and associate sequentially unless some pairwise association energy is above a threshold so that a helix bundle forms in a non-sequential order. Examples for exceptions are the proteins AQP1 and DsbB which both undergo post-translational rearrangements; in the case of AQP1, one helix flips its orientation in the membrane (Virkki *et al.*, 2014) and in the case of DsbB, two helices associate to the membrane interface but only insert into the membrane after all helices have been translated (Harris *et al.*, 2017). These deviations from TMPfold assumptions could be explained by the individual helix propensities to be unstable or stable in the membrane on their own. Individual membrane insertion energies are also calculated by TMPfold but are not yet considered for the folding pathway prediction (Lomize *et al.*, 2020). It will be interesting to analyse future updates that explain rare folding pathways better than the current version.

Due to these known deviations from the predicted pathways, TMPfold should not be used to draw conclusions for individual proteins but can be used to analyse trends in membrane protein families. We were interested if folding pathways are conserved within protein families or in other words, if pathways or their underlying energetics

are encoded in the sequence space of a family. To be able to compare individual proteins, we have converted predicted folding pathways into matrices representing the stepwise pathway and have also used the predicted helix-helix association energy matrices directly. Proteins can be compared to other proteins by calculating a distance between the two proteins' pathway or energy matrices. For this, we used the sum of absolute differences (SAD), a measure used in computer vision to evaluate the similarity between two image regions (Niitsuma and Maruyama, 2010). A normalised distance was calculated for matrix comparisons of within-family (intra) protein pairs as well as between-family (inter) pairs for families with the same transmembrane helix count. The distributions of pairwise distances were analysed for a statistical difference between intra- and inter-family groups (for a given TM helix count); their means and test statistics can be found in Table 5.2.

When comparing predicted pathways, the results show that for a few TM helix counts (3, 6, 7, 11 and 12) pairwise distances are smaller for intra-family comparisons than for inter-family ones (p-value $< 10^{-2}$). The common effect size f (see Methods) indicates the fraction where intra-family distances were greater than the inter-family distances. Thus, an effect size of 0.5 implies no difference in distance distributions and a value near zero indicates a strong signal that pairwise distances within families are smaller than between-family distances. While statistically significant differences could be found for five TM helix counts, only the difference for TM helix count 11 has an effect size of zero (meaning all intra-family distances were smaller than all inter-family ones). The other four tests showed effect sizes between 0.24 and 0.38, indicating no strong signal.

We were interested if differences also showed when comparing helix-helix energy matrices, the underlying data for TMPfold's pathway prediction. The tests on energy matrices displayed statistical significance for all TM helix counts but count 5. Effect sizes of those range from 0 to 0.11, with the exception of TM count 4 which has the highest p-value of $4.1 * 10^{-5}$ and an effect size of 0.29. Overall, the signal of conservation is clearer for energy matrix comparisons than for predicted pathways which is particularly true for sets with only few data points, e.g. TM

helix count 5 with only 9 intra- and 12 inter-family protein comparisons. This suggests two preliminary conclusions: association energetics might be conserved in sequence space but folding pathway predictions lack accuracy.

Pathway predictions are likely to be more error-prone than the predicted helix-helix energies because only the energy predictions have been validated against experimental data. Prediction of individual helix stabilities and pairwise interaction energies are an important step towards accurate folding pathway prediction, but inferring folding pathways just from the thermodynamic constraints is missing other factors that can influence membrane protein folding. For example, additional to post-translational re-arrangements (as seen with AQP1 and DsbB), it has been shown that monomers of multimeric assemblies support the folding process of other monomers (Feige and Hendershot, 2013) or that translocons not only facilitate membrane insertion but also assembly formation (Sadlish *et al.*, 2005). This highlights the need for more development and validation of membrane protein folding predictions. Including individual helix stability predictions into folding considerations is a first step but more is needed, for example, the validation of pathway predictions; not only in a more systematic manner across multiple protein families but also its robustness when using different input structures of the same protein. A useful outcome would add error bars to helix-helix association energy predictions or suggest multiple tentative folding pathways which are annotated with probability scores.

6
Conclusions

Contents

Our analyses were driven by the underlying goal of identifying functional features, other than the global minimum structure of a protein, in its sequence space. We used co-evolution analysis methods to examine residue-level information or investigated sequence space on the protein family level. The results of all of our analyses show that more functional features of proteins are encoded in their sequences than just their static structure. Our conclusions drawn for the different research areas are laid out here in more detail.

6.1 Allostery and Co-evolutionary Information

The initial research question of this thesis was drawn from a study by Lakhani *et al.* (2017) in which a structure of the MutS protein and statistical coupling analysis (SCA, an early co-evolution method) was used to predict the MutS allosteric network.

We examined whether the prediction pipeline of Lakhani *et al.* could be improved using more recent co-evolution techniques. We tested several different direct coupling analysis (DCA) methods that correct for transitive relations in co-evolution data and thus, have a much higher contact prediction precision than SCA. We generated multiple residue networks using different thresholds and parameters and analysed twelve network centralities to weight these networks. Analysis of the ability of the networks to recall predicted allosteric residues gave ambiguous results.

As the original SCA methodology had only been validated using short MD simulations, we decided to diversify our allostery validation. For this, we used data on experimentally-verified allosteric residues from different experiments listed in the Allosteric Database (ASD). The analysed dataset contained 17 proteins with 86 verified residues. We tested if the residue networks of different DCA-based methods recalled these verified residues and identified two trends: the highest recall was achieved by 'raw' DCA methods without machine learning and the highest contact prediction precision was seen for DCA methods with machine learning. This led us to the conclusion that different co-evolution methods may be able to detect different kinds of protein properties.

It has been suggested that the co-evolutionary footprint in a protein stems almost exclusively from physical proximity (Anishchenko *et al.*, 2017) but this proximity might be constrained for different reasons. Fold stability is likely to be a very important cause of co-evolutionary coupling but the conformational flexibility needed to perform a protein's function might be crucial as well. This could not only include important residue contacts that are present only in some of the biologically-relevant conformations, but it might also include the flexibility needed to transmit allosteric signals.

With this in mind, we hypothesised that stability constraints should be separated from allosteric signalling or other functional constraints to enhance their signal-to-noise-ratio. Using the consensus sequence of a protein family could be a direction to investigate as consensus sequences are considered to yield the most stable proteins (Porebski and Buckle, 2016; Sternke *et al.*, 2019) but other measures, especially

on the residue pair-level, could be investigated as well. We tested this hypothesis using findings from our analysis on co-evolutionary information and its link to flexibility. We found that rigid residue pairs have sharp peaks in co-evolutionary distance predictions. In addition to the DCA-based methods we had run on our ASD dataset, we generated predicted distance distributions for all analysed residues pairs with the distance predictor DMPfold (Greener *et al.*, 2019). If a residue pair had a single peak in its predicted distance distribution, we considered this pair as rigid (stabilising) and ignored it when determining the recall of verified allosteric residues of the DCA-based methods. This generally improved the recall of all of our tested methods which supported our view that co-evolutionary coupling can have different causes and that these constraints can be extracted from sequence data.

While these are promising initial results, the size of our improved allostery validation set (with 17 proteins and 86 verified residues) highlights an issue for allostery research in general: validation data is sparse. Even though allostery is thought to be a very common phenomenon amongst biological macromolecules (Dokholyan, 2016), allosteric mechanisms are diverse and limited experimental data exists for them. Deep mutational scanning (DMS) performed with selection assays testing for allostery could be a valuable tool to gain data on all residues and residue pairs of a protein. But it is still resource intensive, especially, because every protein needs a custom allostery-testing assay. Currently, no experimental methods validating allosteric residues are applied to proteins on a large scale. The ASD's most recent update (Liu *et al.*, 2020) does provide an increase in validation set size. Its *Allo-Mutations* dataset is a dataset of 1312 allosteric mutations that are associated with one of 33 cancer types. These mutations cover 133 proteins and were verified in next-generation sequencing (NGS) studies which highlights its potential to further grow in size with the NGS techniques' growth in the future.

6.2 Flexibility and Co-evolutionary Information

Due to the lack of large-scale validation data that directly relates to allostery, we decided to use more readily available data that is closely linked to allostery.

In the ensemble view of allostery, proteins exist in an ensemble of conformations with varying activities and different population sizes, where an allosteric binding event causes a shift of populations sizes and thus, a shift of activity. Knowledge about the conformational ensemble of a protein is therefore also informative of allostery. Conformational ensembles are difficult to obtain completely but they can be approximated by using multiple biologically-relevant structures. We have used the two most different structures known for a given protein to estimate this conformational flexibility, namely the maximum-RMSD-pair dataset from the Database of protein Conformational Diversity in the Native State (CoDNaS) (Monzon *et al.*, 2017). Using the structure pairs for each protein, each pair of residues in a protein can be observed in two different states which were used to classify residue pairs into rigid, flexible or neither.

Morcos *et al.* (2013), amongst others, found that information on conformational flexibility is contained in co-evolutionary data by showing that some false positives in co-evolutionary contact predictions were true positives in alternative conformations of the protein (Morcos *et al.*, 2013; Toth-Petroczy *et al.*, 2016; Anishchenko *et al.*, 2017). However, this has only been reported for binary contact predictions. The publication of co-evolutionary distance predictors that predicted distance distributions between pairs of residues instead of binary contacts, allowed us to raise the question if this information on conformational flexibility could be extracted in a useful way instead of just leading to false positives. We hypothesised that predicted distance distributions would show multi-modality (multiple local maxima) when a residue pair was flexible (changing its interactions between the two structures we analysed).

Our analysis showed a relationship between residue pair flexibility and the number of local maxima in predicted distance distributions with flexible residue pairs more often having two or more peaks and rigid residue pairs just a single peak. While we could confirm that this effect was not just driven by imbalances in our protein set, secondary structure or the validation strategy, the signal was not strong enough to use the number of local maxima to directly predict flexible

residue pairs. The imbalance of rigid and flexible residue pairs in our dataset was about 1:100 making precise predictions difficult. In two case studies we could see that the number of local maxima is condensing the information content of predicted distance distributions too much and potentially useful information is discarded. This, and the fact that current methods are all trained with and for the prediction of single static structures led us to the conclusion that co-evolution methods need to be trained with multiple structures of the same protein to be able to predict conformational flexibility directly.

Our result that flexibility information is contained in co-evolutionary data (Schwarz *et al.*, 2021) is another indication that more than static structural information is contained in sequence information and could be extracted with the right methods.

6.3 Predicting Conformational Ensembles

In chapter 4, we describe a homology modelling pipeline that generates conformational ensembles of kinases and their suitability for docking studies. The pipeline is novel because it is using a systematic approach for generating conformational ensembles that aims to represent even scarcely-populated conformations. Targeting rare conformations can offer increased selectivity (Guimarães *et al.*, 2011); an issue that is important for kinase drug discovery. This was achieved by analysing the available kinase structures and identifying the different conformations of the flexible kinase elements (A-loop, P-loop, αC-helix) that are observed in a majority of kinase families. Template structures were then generated that represent every possible combination of those flexible elements; in total 18 systematically-different conformations of 95% of typical human kinases were generated.

Generating conformational ensembles with homology modelling is a good approach for protein families with a diverse set of available structures. The kinase family has many known structures in a variety of conformations due to the family's relevance for pharmaceutical research. Therefore, it was the ideal family to test the systematic generation of conformational ensembles as it provides a form of

ground truth. We used this characteristic here for systematic homology modelling but the family will also be valuable for training and testing of new structure predictors that aim to generate multiple biologically-relevant conformations than just the global minimum structure.

6.4 Folding Pathway Conservation

While the static protein structure prediction problem is now largely solved with the latest generation of deep learning-based structure predictors, the protein folding problem is not. It is assumed that nature does not try to solve this as one global combinatorial problem but rather by solving many local problems (Dill *et al.*, 2008). Nevertheless, its underlying mechanisms are poorly understood and folding and aggregation-driven diseases lack effective treatments. Investigating the folding problem remains an important area of research and for the last project of my thesis, we looked into whether helical membrane proteins had conserved folding pathways. For this, we used the first (and only) membrane protein folding pathway predictor (TMPfold) which predicts helix-helix association energies from an input structure and uses those to suggest a folding pathway.

We compared the pathways of pairs of proteins with the same number of transmembrane helices (TM helix count) and analysed if there was a difference between intra- and inter-family comparisons. A conservation signal for about half of the TM helix counts could be found. Generally, TM helix counts with large amounts of data showed conservation while this could not be unambiguously determined for TM helix counts with few data points. This changed when calculating the differences between the energy matrices instead of predicted pathways. For all but one TM helix count conservation was observed. This led us to the conclusion that folding information may indeed be encoded in sequence data but the folding pathway predictor needs further improvement and validation. The predictor is based on assumptions of thermodynamic constraints and sequential insertion of transmembrane helices but reality is likely to be more complex than solely taking

into account the pairwise association energies. Including individual helix stabilities and uncertainty information are likely to be the first steps forward.

6.5 Future Perspectives

The main conclusion from this thesis, drawn from the examination of co-evolution data and its relation to allostery and flexibility, is that co-evolutionary couplings are driven by several different biophysical constraints and that different methods extract different kinds of information from sequence data. Developing a model that takes into account these different constraints would be useful to separate them. This might be key to their future use because stability is likely to be the major factor driving co-evolutionary couplings making other constraints hard to analyse because of an unfavourable signal-to-noise-ratio. Thus, a quantification of stability constraints on a residue pair level could be critical to enable the analysis of other, lower-influence constraints.

An example where such an approach was used is a recent study by Xu *et al.* (2021) where X-ray crystallography data is analysed for correlated motions in total X-ray scattering. Total X-ray scattering consists of the intense Bragg peaks and weaker, diffuse scattering underneath and around Bragg peaks. Bragg peaks are used to obtain high-resolution information on atom positions in the crystal lattice whereas the diffuse scattering was noise. Improvements in detectors and analysis software has enable them to split the intense and weak signals and extract information from the 'noise' to infer correlated motions of atoms in the lattice.

Splitting the co-evolutionary signal directly into its different components would be ideal but can only be done when the underlying constraints are better understood. Therefore, stability estimates are needed first as proximity is the major cause for co-evolutionary couplings (Anishchenko *et al.*, 2017). Approximating this, for example by using the consensus sequence of a protein family as the base line for stability, could be a direction that is not resource intensive, but experimental validation would be necessary.

Deep mutational scanning (DMS) is potentially the most promising method to deliver stability information on a residue pair level. Furthermore, DMS can be performed with different selection assays that yield information on different constraints of a residue pair. One assay selecting for stability and another one selecting for allosteric signal transmission would be the perfect set up to validate a two-constraint model that explains co-evolutionary couplings.

The other main conclusion from my work on conformational flexibility in co-evolutionary data is that protein structure predictors should be trained with multiple structures of the same sequence and aim to yield more than a single global minimum structure. We had considered using a fragment-based structure predictor in combination with distance predictions to predict multiple diverging structures but the performance of AlphaFold2 (in static structure prediction) highlighted that deep learning-based methods (Jumper *et al.*, 2021; Minkyung *et al.*, 2021) are able to achieve much more accurate structures (Pereira *et al.*, 2021).

AlphaFold2 uses non-deterministic behaviour in two steps: when using or omitting a random small number of templates and when masking the underlying multiple sequence alignment (MSA) that a structure is predicted for. It will be interesting to see when the masking is not done randomly but in a systematic way, for example, clustering of sequences and determining structures for each cluster or only allowing sequences that display a certain conformation or protein-protein interaction. One could imagine the generation of a latent space for each of these different maskings which are fed by different clusters of sequences but would then be applied to the target sequence in question (even though that sequence might not be part of that cluster). This could yield a diverse set of structures which might be energetically accessible for most proteins of a family, even though they might not be the global minimum structure and thus, occupied less often but still relevant.

The newest generation of structure predictors are only trained to generate a global minimum structure but they already predict alternative conformations occasionally (del Alamo *et al.*, 2021). Therefore, it seems likely that updates in the future will consider conformational flexibility and yield *de novo* model ensembles. This progress

could then be used to develop a metric that quantifies the proximity (or stability) contribution of co-evolutionary signal which in turn benefits allostery research.

Bibliography

Adhikari, B., Hou, J., and Cheng, J. (2018). DNCON2: improved protein contact prediction using two-level deep convolutional neural networks. *Bioinformatics*, **34**(9), 1466–1472.

Agarwal, P. K., Billeter, S. R., Rajagopalan, P. T. R., Benkovic, S. J., and Hammes-Schiffer, S. (2002). Network of coupled promoting motions in enzyme catalysis. *Proceedings of the National Academy of Sciences of the United States of America*, **99**(5), 2794.

Amor, B. R. C., Schaub, M. T., Yaliraki, S. N., and Barahona, M. (2016). Prediction of allosteric sites and mediating interactions through bond-to-bond propensities. *Nature Communications*, **7**(1), 12477.

Anfinsen, C. B. (1973). Principles that Govern the Folding of Protein Chains. *Science*, **181**(4096), 223–230.

Anishchenko, I., Ovchinnikov, S., Kamisetty, H., and Baker, D. (2017). Origins of coevolution between residues distant in protein 3D structures. *Proceedings of the National Academy of Sciences of the United States of America*, **114**(34), 9122–9127.

Arinaminpathy, Y., Khurana, E., Engelman, D. M., and Gerstein, M. B. (2009). Computational analysis of membrane proteins: the largest class of drug targets. *Drug discovery today*, **14**(23-24), 1130–1135.

Bandaru, P., Shah, N. H., Bhattacharyya, M., Barton, J. P., Kondo, Y., Cofsky, J. C., Gee, C. L., Chakraborty, A. K., Kortemme, T., Ranganathan, R., and Kuriyan, J. (2017). Deconstruction of the Ras switching cycle through saturation mutagenesis. *eLife*, **6**.

Banerjee, S., Bartesaghi, A., Merk, A., Rao, P., Bulfer, S. L., Yan, Y., Green, N., Mroczkowski, B., Neitz, R. J., Wipf, P., Falconieri, V., Deshaies, R. J., Milne, J. L. S., Huryn, D., Arkin, M., and Subramaniam, S. (2016). 2.3 A resolution cryo-EM structure of human p97 and mechanism of allosteric inhibition. *Science*, **351**(6275), 871–875.

Benson, N. C. and Daggett, V. (2008). Dynameomics: Large-scale assessment of native protein flexibility. *Protein Science*, **17**(12), 2038–2050.

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

Bonacich, P. (1972). Factoring and weighting approaches to status scores and clique identification. *The Journal of Mathematical Sociology*, **2**(1), 113–120.

Bouatta, N., Sorger, P., and AlQuraishi, M. (2021). Protein structure prediction by AlphaFold2: are attention and symmetries all you need? *Acta Crystallographica Section D*, **77**(8), 982–991.

Boyles, F., Deane, C. M., and Morris, G. M. (2020). Learning from the ligand: using ligand-based features to improve binding affinity prediction. *Bioinformatics*, **36**(3), 758–764.

Brin, S. and Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, **30**(1), 107–117.

Brooijmans, N., Chang, Y.-W., Mobilio, D., Denny, R. A., and Humblet, C. (2010). An enriched structural kinase database to enable kinome-wide structure-based analyses and drug discovery. *Protein Sci*, **19**(4), 763–774.

Buchan, D. W. A. and Jones, D. T. (2018). Improved protein contact predictions with the MetaPSICOV2 server in CASP12. *Proteins: Structure, Function, and Bioinformatics*, **86**, 78–83.

Carles, F., Bourg, S., Meyer, C., and Bonnet, P. (2018). PKIDB: A Curated, Annotated and Updated Database of Protein Kinase Inhibitors in Clinical Trials. *Molecules*, **23**(4), 908.

Chapman, H. N., Fromme, P., Barty, A., White, T. A., Kirian, R. A., Aquila, A., Hunter, M. S., Schulz, J., DePonte, D. P., Weierstall, U., Doak, R. B., Maia, F. R. N. C., Martin, A. V., Schlichting, I., Lomb, L., Coppola, N., Shoeman, R. L., Epp, S. W., Hartmann, R., Rolles, D., Rudenko, A., Foucar, L., Kimmel, N., Weidenspointner, G., Holl, P., Liang, M., Barthelmess, M., Caleman, C., Boutet, S., Bogan, M. J., Krzywinski, J., Bostedt, C., Bajt, S., Gumprecht, L., Rudek, B., Erk, B., Schmidt, C., Hömke, A., Reich, C., Pietschner, D., Strüder, L., Hauser, G., Gorke, H., Ullrich, J., Herrmann, S., Schaller, G., Schopper, F., Soltau, H., Kühnel, K.-U., Messerschmidt, M., Bozek, J. D., Hau-Riege, S. P., Frank, M., Hampton, C. Y., Sierra, R. G., Starodub, D., Williams, G. J., Hajdu, J., Timneanu, N., Seibert, M. M., Andreasson, J., Rocker, A., Jönsson, O., Svenda, M., Stern, S., Nass, K., Andritschke, R., Schröter, C.-D., Krasniqi, F., Bott, M., Schmidt, K. E., Wang, X., Grotjohann, I., Holton, J. M., Barends, T. R. M., Neutze, R., Marchesini, S., Fromme, R., Schorb, S., Rupp, D., Adolph, M., Gorkhover, T., Andersson, I., Hirsemann, H., Potdevin, G., Graafsma, H., Nilsson, B., and Spence, J. C. H. (2011). Femtosecond X-ray protein nanocrystallography. *Nature*, **470**(7332), 73–77.

Chonofsky, M., Oliveira, S. H. P. d., Krawczyk, K., and Deane, C. M. (2019). The evolution of contact prediction: Evidence that contact selection in statistical contact prediction is changing. *bioRxiv*, page 660191.

Cilia, E., Pancsa, R., Tompa, P., Lenaerts, T., and Vranken, W. F. (2014). The DynaMine webserver: predicting protein dynamics from sequence. *Nucleic acids research*, **42**(Web Server issue), 264–70.

Cleves, A. E. and Jain, A. N. (2020). Structure- and Ligand-Based Virtual Screening on DUD-E+: Performance Dependence on Approximations to the Binding Pocket. *Journal of Chemical Information and Modeling*.

Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., and de Hoon, M. J. L. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, **25**(11), 1422–1423.

Cohen, P. (2002). Protein kinases - the major drug targets of the twenty-first century? *Nat Rev Drug Discov*, **1**(4), 309–315.

Collins, P. M., Ng, J. T., Talon, R., Nekrosiute, K., Krojer, T., Douangamath, A., Brandao-Neto, J., Wright, N., Pearce, N. M., and von Delft, F. (2017). Gentle, fast and effective crystal soaking by acoustic dispensing. *Acta crystallographica. Section D, Structural biology*, **73**(Pt 3), 246–255.

Cover, T. M. and Thomas, J. A. (2012). *Elements of Information Theory*. Wiley.

Csardi, G. and Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal Complex Systems*, page 1695.

Cymer, F., von Heijne, G., and White, S. H. (2015). Mechanisms of Integral Membrane Protein Insertion and Folding. *Journal of Molecular Biology*, **427**(5), 999–1022.

De Geyter, J., Portaliou, A. G., Srinivasu, B., Krishnamurthy, S., Economou, A., and Karamanou, S. (2020). Trigger factor is a bona fide secretory pathway chaperone that interacts with SecB and the translocase. *EMBO reports*, **21**(6), e49054.

de Oliveira, S. and Deane, C. (2017). Co-evolution techniques are reshaping the way we do structural bioinformatics. *F1000Research*, **6**, 1224.

de Oliveira, S. H. P., Shi, J., and Deane, C. M. (2016). Comparing co-evolution methods and their application to template-free protein structure prediction. *Bioinformatics*, **33**(3), btw618.

Deber, C. M., Brandl, C. J., Deber, R. B., Hsu, L. C., and Young, X. K. (1986). Amino acid composition of the membrane and aqueous domains of integral membrane proteins. *Archives of biochemistry and biophysics*, **251**(1), 68–76.

del Alamo, D., Govaerts, C., and Mchaourab, H. S. (2021). AlphaFold2 predicts the inward-facing conformation of the multidrug transporter LmrP. *Proteins: Structure, Function, and Bioinformatics*, **n/a**(n/a).

Dill, K. A., Ozkan, S. B., Shell, M. S., and Weikl, T. R. (2008). The protein folding problem. *Annual review of biophysics*, **37**, 289–316.

Diss, G. and Lehner, B. (2018). The genetic landscape of a physical interaction. *eLife*, **7**.

Dokholyan, N. V. (2016). Controlling Allosteric Networks in Proteins. *Chemical Reviews*, **116**, 6463–6487.

Eagle, N., Macy, M., and Claxton, R. (2010). Network Diversity and Economic Development. *Science (New York, N.Y.)*, **328**, 1029–1031.

Eaton, W. A. (2021). Modern Kinetics and Mechanism of Protein Folding: A Retrospective. *The journal of physical chemistry. B*, **125**(14), 3452–3467.

Ehlert, F. J. and Griffin, M. T. (2008). Two-State Models and the Analysis of the Allosteric Effect of Gallamine at the M2 Muscarinic Receptor. *Journal of Pharmacology and Experimental Therapeutics*, **325**(3), 1039–1060.

Eid, S., Turk, S., Volkamer, A., Rippmann, F., and Fulle, S. (2017). KinMap: a web-based tool for interactive navigation through human kinome data. *BMC Bioinformatics*, **18**, 16.

Ekeberg, M., Lövkvist, C., Lan, Y., Weigt, M., and Aurell, E. (2013). Improved contact prediction in proteins: Using pseudolikelihoods to infer Potts models. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, **87**(1), 012707.

Fabbro, D., Cowan-Jacob, S. W., and Moebitz, H. (2015). Ten things you should know about protein kinases: IUPHAR Review 14. *British journal of pharmacology*, **172**(11), 2675–2700.

Feher, V. A., Durrant, J. D., Van Wart, A. T., and Amaro, R. E. (2014). Computational approaches to mapping allosteric pathways. *Current opinion in structural biology*, **25**, 98–103.

Feige, M. and Hendershot, L. (2013). Quality Control of Integral Membrane Proteins by Assembly-Dependent Membrane Integration. *Molecular Cell*, **51**(3), 297–309.

Fowler, D. M. and Fields, S. (2014). Deep mutational scanning: a new style of protein science. *Nature Methods*, **11**(8), 801–807.

Frank, J. and Ourmazd, A. (2016). Continuous changes in structure mapped by manifold embedding of single-particle data in cryo-EM. *Methods (San Diego, Calif.)*, **100**, 61–67.

Frauenfelder, H., Sligar, S. G., and Wolynes, P. G. (1991). The Energy Landscapes and Motions of Proteins. *Science*, **254**(5038), 1598–1603.

Freeman, L. C. (1977). A Set of Measures of Centrality Based on Betweenness. *Sociometry*, **40**(1), 35–41.

Freeman, L. C. (1978). Centrality in social networks conceptual clarification. *Social Networks*, **1**(3), 215–239.

Frueh, D. P., Goodrich, A. C., Mishra, S. H., and Nichols, S. R. (2013). NMR methods for structural studies of large monomeric and multimeric proteins. *Current opinion in structural biology*, **23**(5), 734–739.

Girvan, M. and Newman, M. E. J. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, **99**(12), 7821 LP – 7826.

Greener, J. G., Filippis, I., and Sternberg, M. J. E. (2017). Predicting Protein Dynamics and Allostery Using Multi-Protein Atomic Distance Constraints. *Structure (London, England : 1993)*, **25**(3), 546–558.

Greener, J. G., Kandathil, S. M., and Jones, D. T. (2019). Deep learning extends de novo protein modelling coverage of genomes using iteratively predicted structural constraints. *Nature Communications*, **10**(1), 3977.

Guarnera, E. and Berezovsky, I. N. (2016). Allosteric sites: remote control in regulation of protein activity. *Current Opinion in Structural Biology*, **37**, 1–8.

Guimarães, C. R. W., Rai, B. K., Munchhof, M. J., Liu, S., Wang, J., Bhattacharya, S. K., and Buckbinder, L. (2011). Understanding the Impact of the P-loop Conformation on Kinase Selectivity. *Journal of Chemical Information and Modeling*, **51**(6), 1199–1204.

Guo, J. and Zhou, H.-X. (2016). Protein Allostery and Conformational Dynamics. *Chemical Reviews*, **116**(11), 6503–6515.

Hage, P. and Harary, F. (1995). Eccentricity and centrality in networks. *Social Networks*, **17**(1), 57–63.

Halabi, N., Rivoire, O., Leibler, S., and Ranganathan, R. (2009). Protein sectors: evolutionary units of three-dimensional structure. *Cell*, **138**(4), 774–86.

Haldane, A., Flynn, W. F., He, P., Vijayan, R., and Levy, R. M. (2016). Structural propensities of kinase family proteins from a Potts model of residue co-variation. *Protein Science*, **25**(8), 1378–1384.

Hall, D. A. (2000). Modeling the functional effects of allosteric modulators at pharmacological receptors: an extension of the two-state model of receptor activation. *Molecular pharmacology*, **58**(6), 1412–23.

Hamuro, Y., Coales, S. J., Southern, M. R., Nemeth-Cawley, J. F., Stranz, D. D., and Griffin, P. R. (2003). Rapid analysis of protein structure and dynamics by hydrogen/deuterium exchange mass spectrometry. *Journal of biomolecular techniques : JBT*, **14**(3), 171–182.

Harris, N. J., Reading, E., Ataka, K., Grzegorzewski, L., Charalambous, K., Liu, X., Schlesinger, R., Heberle, J., and Booth, P. J. (2017). Structure formation during translocon-unassisted co-translational membrane protein folding. *Scientific Reports*, **7**(1), 8021.

Haselbach, D., Schrader, J., Lambrecht, F., Henneberg, F., Chari, A., and Stark, H. (2017). Long-range allosteric regulation of the human 26S proteasome by 20S proteasome-targeting cancer drugs. *Nature Communications*, **8**, 15578.

Hekstra, D. R., White, K. I., Socolich, M. A., Henning, R. W., Šrajer, V., and Ranganathan, R. (2016). Electric-field-stimulated protein mechanics. *Nature*, **540**(7633), 400–405.

Henzler-Wildman, K. and Kern, D. (2007). Dynamic personalities of proteins. *Nature*, **450**(7172), 964–972.

Hill, J. R., Kelm, S., Shi, J., and Deane, C. M. (2011). Environment specific substitution tables improve membrane protein alignment. *Bioinformatics*, **27**(13), i15–i23.

Hilser, V. J., Wrabl, J. O., and Motlagh, H. N. (2012). Structural and Energetic Basis of Allostery. *Annual Review of Biophysics*, **41**(1), 585–609.

Hopf, T. A., Schärfe, C. P. I., Rodrigues, J. P. G. L. M., Green, A. G., Kohlbacher, O., Sander, C., Bonvin, A. M. J. J., and Marks, D. S. (2014). Sequence co-evolution gives 3D contacts and structures of protein complexes. *eLife*, **3**.

Horner, D. S., Pirovano, W., and Pesole, G. (2007). Correlated substitution analysis and the prediction of amino acid structural contacts. *Briefings in Bioinformatics*, **9**(1), 46–56.

Hospital, A., Goñi, J. R., Orozco, M., and Gelpí, J. L. (2015). Molecular dynamics simulations: advances and applications. *Advances and applications in bioinformatics and chemistry : AABC*, **8**, 37–47.

Hrabe, T., Li, Z., Sedova, M., Rotkiewicz, P., Jaroszewski, L., and Godzik, A. (2016). PDBFlex: exploring flexibility in protein structures. *Nucleic acids research*, **44**(D1), 423–8.

Jacobs, D. J., Rader, A. J., Kuhn, L. A., and Thorpe, M. F. (2001). Protein flexibility predictions using graph theory. *Proteins: Structure, Function, and Bioinformatics*, **44**(2), 150–165.

Jones, D. T. and Kandathil, S. M. (2018). High precision in protein contact prediction using fully convolutional neural networks and minimal sequence features. *Bioinformatics*, **34**(19), 3308–3315.

Jones, D. T., Buchan, D. W. A., Cozzetto, D., and Pontil, M. (2012). PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics*, **28**(2), 184–190.

Jones, D. T., Singh, T., Kosciolek, T., and Tetchner, S. (2015). MetaPSICOV: combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins. *Bioinformatics*, **31**(7), 999–1006.

Jubb, H. C., Higueruelo, A. P., Ochoa-Montaño, B., Pitt, W. R., Ascher, D. B., and Blundell, T. L. (2017). Arpeggio: A Web Server for Calculating and Visualising Interatomic Interactions in Protein Structures. *Journal of molecular biology*, **429**(3), 365–371.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A. W., Kavukcuoglu, K., Kohli, P., and Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, **596**(7873), 583–589.

Kabsch, W. and Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**(12), 2577–2637.

Kaján, L., Hopf, T. A., Kalaš, M., Marks, D. S., and Rost, B. (2014). FreeContact: fast and free software for protein contact prediction from residue co-evolution. *BMC bioinformatics*, **15**, 85.

Källberg, M., Wang, H., Wang, S., Peng, J., Wang, Z., Lu, H., and Xu, J. (2012). Template-based protein structure modeling using the RaptorX web server. *Nature Protocols*, **7**(8), 1511–1522.

Kamisetty, H., Ovchinnikov, S., and Baker, D. (2013). Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era. *Proceedings of the National Academy of Sciences of the United States of America*, **110**(39), 15674–9.

Keedy, D. A. (2019). Journey to the center of the protein: allostery from multitemperature multiconformer X-ray crystallography. *Acta Crystallographica Section D Structural Biology*, **75**(2), 123–137.

Keedy, D. A., Fraser, J. S., and van den Bedem, H. (2015a). Exposing Hidden Alternative Backbone Conformations in X-ray Crystallography Using qFit. *PLOS Computational Biology*, **11**(10), e1004507.

Keedy, D. A., Kenner, L. R., Warkentin, M., Woldeyes, R. A., Hopkins, J. B., Thompson, M. C., Brewster, A. S., Van Benschoten, A. H., Baxter, E. L., Uervirojnangkoorn, M., McPhillips, S. E., Song, J., Alonso-Mori, R., Holton, J. M., Weis, W. I., Brunger, A. T., Soltis, S. M., Lemke, H., Gonzalez, A., Sauter, N. K., Cohen, A. E., van den Bedem, H., Thorne, R. E., and Fraser, J. S. (2015b). Mapping the conformational landscape of a dynamic enzyme by multitemperature and XFEL crystallography. *eLife*, **4**.

Keedy, D. A., Hill, Z. B., Biel, J. T., Kang, E., Rettenmaier, T. J., Brandão-Neto, J., Pearce, N. M., von Delft, F., Wells, J. A., and Fraser, J. S. (2018). An expanded allosteric network in PTP1B by multitemperature crystallography, fragment screening, and covalent tethering. *eLife*, **7**.

Kim, I., Miller, C. R., Young, D. L., and Fields, S. (2013). High-throughput Analysis of in vivo Protein Stability. *Molecular & Cellular Proteomics*, **12**(11), 3370–3378.

Koes, D. R., Baumgartner, M. P., and Camacho, C. J. (2013). Lessons Learned in Empirical Scoring with smina from the CSAR 2011 Benchmarking Exercise. *Journal of Chemical Information and Modeling*, **53**(8), 1893–1904.

Koshland, D. E., Némethy, G., and Filmer, D. (1966). Comparison of Experimental Binding Data and Theoretical Models in Proteins Containing Subunits. *Biochemistry*, **5**(1), 365–385.

Krogh, A., Larsson, B., von Heijne, G., and Sonnhammer, E. L. (2001). Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *Journal of molecular biology*, **305**(3), 567–580.

Krüger, D. M., Ahmed, A., and Gohlke, H. (2012). NMSim web server: integrated approach for normal mode-based geometric simulations of biologically relevant conformational transitions in proteins. *Nucleic acids research*, **40**(Web Server issue), 310–6.

Kryshtafovych, A., Schwede, T., Topf, M., Fidelis, K., and Moult, J. (2019). Critical assessment of methods of protein structure prediction (CASP)—Round XIII. *Proteins: Structure, Function, and Bioinformatics*, **87**(12), 1011–1020.

Kryshtafovych, A., Schwede, T., Topf, M., Fidelis, K., and Moult, J. (2021). Critical assessment of methods of protein structure prediction (CASP)—Round XIV. *Proteins: Structure, Function, and Bioinformatics*, **89**(12), 1607–1617.

Kufareva, I. and Abagyan, R. (2008). Type-II Kinase Inhibitor Docking, Screening, and Profiling Using Modified Structures of Active Kinase States. *J Med Chem*, **51**(24), 7921–7932.

Kumazaki, K., Chiba, S., Takemoto, M., Furukawa, A., Nishiyama, K.-i., Sugano, Y., Mori, T., Dohmae, N., Hirata, K., Nakada-Nakura, Y., Maturana, A. D., Tanaka, Y., Mori, H., Sugita, Y., Arisaka, F., Ito, K., Ishitani, R., Tsukazaki, T., and Nureki, O. (2014). Structural basis of Sec-independent membrane protein insertion by YidC. *Nature*, **509**(7501), 516–520.

Kuriata, A., Gierut, A. M., Oleniecki, T., Ciemny, M., Kolinski, A., Kurcinski, M., and Kmiecik, S. (2018). CABS-flex 2.0: a web server for fast simulations of flexibility of protein structures. *Nucleic Acids Research*, **46**(W1), W338–W343.

Ladokhin, A. S. and White, S. H. (1999). Folding of amphipathic α-helices on membranes: energetics of helix formation by melittin11Edited by D. Rees. *Journal of Molecular Biology*, **285**(4), 1363–1369.

Lakhani, B., Thayer, K. M., Hingorani, M. M., and Beveridge, D. L. (2017). Evolutionary Covariance Combined with Molecular Dynamics Predicts a Framework for Allostery in the MutS DNA Mismatch Repair Protein. *The Journal of Physical Chemistry B*, **121**(9), 2049–2061.

Lang, P. T., Holton, J. M., Fraser, J. S., and Alber, T. (2014). Protein structural ensembles are revealed by redefining X-ray electron density noise. *Proceedings of the National Academy of Sciences of the United States of America*, **111**(1), 237–242.

Leander, M., Yuan, Y., Meger, A., Cui, Q., and Raman, S. (2020). Functional plasticity and evolutionary adaptation of allosteric regulation. *Proceedings of the National Academy of Sciences of the United States of America*, **117**(41), 25445–25454.

Leff, P. (1995). The two-state model of receptor activation. *Trends in pharmacological sciences*, **16**(3), 89–97.

LeVine, M. V., Cuendet, M. A., Razavi, A. M., Khelashvili, G., and Weinstein, H. (2018). Thermodynamic Coupling Function Analysis of Allosteric Mechanisms in the Human Dopamine Transporter. *Biophysical Journal*, **114**(1), 10–14.

Liang, B. and Tamm, L. K. (2007). Structure of outer membrane protein G by solution NMR spectroscopy. *Proceedings of the National Academy of Sciences*, **104**(41), 16140 LP – 16145.

Lindahl, E., Azuara, C., Koehl, P., and Delarue, M. (2006). NOMAD-Ref: visualization, deformation and refinement of macromolecular structures based on all-atom normal mode analysis. *Nucleic acids research*, **34**(Web Server issue), 52–6.

Liu, X., Lu, S., Song, K., Shen, Q., Ni, D., Li, Q., He, X., Zhang, H., Wang, Q., Chen, Y., Li, X., Wu, J., Sheng, C., Chen, G., Liu, Y., Lu, X., and Zhang, J. (2020). Unraveling allosteric landscapes of allosterome with ASD. *Nucleic Acids Research*, **48**(D1), D394–D401.

Lomize, A. L., Pogozheva, I. D., Lomize, M. A., and Mosberg, H. I. (2006). Positioning of proteins in membranes: a computational approach. *Protein science : a publication of the Protein Society*, **15**(6), 1318–1333.

Lomize, A. L., Schnitzer, K. A., and Pogozheva, I. D. (2020). TMPfold: A Web Tool for Predicting Stability of Transmembrane α-Helix Association. *Journal of Molecular Biology*, **432**(11), 3388–3394.

Lomize, M. A., Pogozheva, I. D., Joo, H., Mosberg, H. I., and Lomize, A. L. (2012). OPM database and PPM web server: resources for positioning of proteins in membranes. *Nucleic acids research*, **40**(Database issue), 370–6.

López-Blanco, J. R. and Chacón, P. (2016). New generation of elastic network models. *Current Opinion in Structural Biology*, **37**, 46–53.

Manley, G. and Loria, J. P. (2012). NMR insights into protein allostery. *Archives of biochemistry and biophysics*, **519**(2), 223–231.

Marks, C. and Deane, C. M. (2018). Increasing the accuracy of protein loop structure prediction with evolutionary constraints. *Bioinformatics*, **35**(15), 2585–2592.

Marks, D. S., Colwell, L. J., Sheridan, R., Hopf, T. A., Pagnani, A., Zecchina, R., and Sander, C. (2011). Protein 3D structure computed from evolutionary sequence variation. *PloS one*, **6**(12), e28766.

Marks, D. S., Hopf, T. A., and Sander, C. (2012). Protein structure prediction from sequence variation. *Nature Biotechnology*, **30**(11), 1072–1080.

Matreyek, K. A., Starita, L. M., Stephany, J. J., Martin, B., Chiasson, M. A., Gray, V. E., Kircher, M., Khechaduri, A., Dines, J. N., Hause, R. J., Bhatia, S., Evans, W. E., Relling, M. V., Yang, W., Shendure, J., and Fowler, D. M. (2018). Multiplex assessment of protein variant abundance by massively parallel sequencing. *Nature Genetics*, **50**(6), 874–882.

Melamed, D., Young, D. L., Gamble, C. E., Miller, C. R., and Fields, S. (2013). Deep mutational scanning of an RRM domain of the Saccharomyces cerevisiae poly(A)-binding protein. *RNA*, **19**(11), 1537–1551.

Minkyung, B., Frank, D., Ivan, A., Justas, D., Sergey, O., Rie, L. G., Jue, W., Qian, C., N., K. L., Dustin, S. R., Claudia, M., Hahnbeom, P., Carson, A., R., G. C., Andy, D., H., P. J., V., R. A., A., v. D. A., C., E. A., J., O. D., Theo, S., Christoph, B., Tea, P.-K., K., R. M., Udit, D., K., Y. C., E., B. J., Christopher, G. K., V., G. N., D., A. P., J., R. R., and David, B. (2021). Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, **373**(6557), 871–876.

Möbitz, H. (2015). The ABC of protein kinase conformations. *Biochim Biophys Acta*, **1854**(10 Pt B), 1555–1566.

Monod, J., Wyman, J., and Changeux, J.-P. (1965). On the nature of allosteric transitions: A plausible model. *Journal of Molecular Biology*, **12**(1), 88–118.

Monzon, A. M., Rohr, C. O., Fornasari, M. S., and Parisi, G. (2016). CoDNaS 2.0: a comprehensive database of protein conformational diversity in the native state. *Database : the journal of biological databases and curation*, **2016**, baw038.

Monzon, A. M., Zea, D. J., Fornasari, M. S., Saldaño, T. E., Fernandez-Alberti, S., Tosatto, S. C. E., and Parisi, G. (2017). Conformational diversity analysis reveals three functional mechanisms in proteins. *PLOS Computational Biology*, **13**(2), e1005398.

Morcos, F., Pagnani, A., Lunt, B., Bertolino, A., Marks, D. S., Sander, C., Zecchina, R., Onuchic, J. N., Hwa, T., and Weigt, M. (2011). Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proceedings of the National Academy of Sciences of the United States of America*, **108**(49), 1293–301.

Morcos, F., Jana, B., Hwa, T., and Onuchic, J. N. (2013). Coevolutionary signals across protein lineages help capture multiple protein conformations. *Proceedings of the National Academy of Sciences*, **110**(51), 20533 LP – 20538.

Nakane, T., Kotecha, A., Sente, A., McMullan, G., Masiulis, S., Brown, P. M. G. E., Grigoras, I. T., Malinauskaite, L., Malinauskas, T., Miehling, J., Uchański, T., Yu, L., Karia, D., Pechnikova, E. V., de Jong, E., Keizer, J., Bischoff, M., McCormack, J., Tiemeijer, P., Hardwick, S. W., Chirgadze, D. Y., Murshudov, G., Aricescu, A. R., and Scheres, S. H. W. (2020). Single-particle cryo-EM at atomic resolution. *Nature*, **587**(7832), 152–156.

Narwani, T. J., Etchebest, C., Craveur, P., Léonard, S., Rebehmed, J., Srinivasan, N., Bornot, A., Gelly, J.-C., and de Brevern, A. G. (2019). In silico prediction of protein flexibility with local structure approach. *Biochimie*, **165**, 150–155.

Neher, E. (1994). How frequent are correlated changes in families of protein sequences? *Proceedings of the National Academy of Sciences of the United States of America*, **91**(1), 98–102.

Ni, D., Li, Y., Qiu, Y., Pu, J., Lu, S., and Zhang, J. (2020). Combining Allosteric and Orthosteric Drugs to Overcome Drug Resistance. *Trends in pharmacological sciences*, **41**(5), 336–348.

Niitsuma, H. and Maruyama, T. (2010). Sum of Absolute Difference Implementations for Image Processing on FPGAs. In *2010 International Conference on Field Programmable Logic and Applications*, pages 167–170.

Nilmeier, J., Hua, L., Coutsias, E. A., and Jacobson, M. P. (2011). Assessing protein loop flexibility by hierarchical Monte Carlo sampling. *Journal of chemical theory and computation*, **7**(5), 1564–1574.

Noel, J. K., Whitford, P. C., and Onuchic, J. N. (2012). The Shadow Map: A General Contact Definition for Capturing the Dynamics of Biomolecular Folding and Function. *The Journal of Physical Chemistry B*, **116**(29), 8692–8702.

Nugent, T. and Jones, D. T. (2012). Accurate de novo structure prediction of large transmembrane protein domains using fragment-assembly and correlated mutation analysis. *Proceedings of the National Academy of Sciences of the United States of America*, **109**(24), 1540–7.

O'Boyle, N. M., Banck, M., James, C. A., Morley, C., Vandermeersch, T., and Hutchison, G. R. (2011). Open Babel: An open chemical toolbox. *Journal of Cheminformatics*, **3**(1), 33.

Olson, C. A., Wu, N. C., and Sun, R. (2014). A comprehensive biophysical description of pairwise epistasis throughout an entire protein domain. *Current biology : CB*, **24**(22), 2643–51.

Ovchinnikov, S., Kamisetty, H., and Baker, D. (2014). Robust and accurate prediction of residue–residue interactions across protein interfaces using evolutionary information. *eLife*, **3**, e02030.

Ovchinnikov, S., Kim, D. E., Wang, R. Y.-R., Liu, Y., DiMaio, F., and Baker, D. (2016). Improved de novo structure prediction in CASP11 by incorporating coevolution information into Rosetta. *Proteins: Structure, Function, and Bioinformatics*, **84**, 67–75.

Pellowe, G. A. and Booth, P. J. (2020). Structural insight into co-translational membrane protein folding. *Biochimica et Biophysica Acta (BBA) - Biomembranes*, **1862**(1), 183019.

Pereira, J., Simpkin, A. J., Hartmann, M. D., Rigden, D. J., Keegan, R. M., and Lupas, A. N. (2021). High-accuracy protein structure prediction in CASP14. *Proteins: Structure, Function, and Bioinformatics*, **89**(12), 1687–1699.

Polychronidou, E., Avramouli, A., and Vlamos, P. (2020). Alzheimer's Disease: The Role of Mutations in Protein Folding BT - GeNeDis 2018. pages 227–236, Cham. Springer International Publishing.

Porebski, B. T. and Buckle, A. M. (2016). Consensus protein design. *Protein engineering, design & selection : PEDS*, **29**(7), 245–51.

Punjani, A. and Fleet, D. J. (2021). 3D variability analysis: Resolving continuous flexibility and discrete heterogeneity from single particle cryo-EM. *Journal of Structural Biology*, **213**(2), 107702.

Remmert, M., Biegert, A., Hauser, A., and Söding, J. (2012). HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nature Methods*, **9**(2), 173–175.

Rivoire, O., Reynolds, K. A., and Ranganathan, R. (2016). Evolution-Based Functional Decomposition of Proteins. *PLOS Computational Biology*, **12**(6), e1004817.

Rocklin, G. J., Chidyausiku, T. M., Goreshnik, I., Ford, A., Houliston, S., Lemak, A., Carter, L., Ravichandran, R., Mulligan, V. K., Chevalier, A., Arrowsmith, C. H., and Baker, D. (2017). Global analysis of protein folding using massively parallel design, synthesis, and testing. *Science*, **357**(6347), 168–175.

Rodriguez-Rivas, J., Marsili, S., Juan, D., and Valencia, A. (2016). Conservation of coevolving protein interfaces bridges prokaryote–eukaryote homologies in the twilight zone. *Proceedings of the National Academy of Sciences*, **113**(52), 15018 LP – 15023.

Rollins, N. J., Brock, K. P., Poelwijk, F. J., Stiffler, M. A., Gauthier, N. P., Sander, C., and Marks, D. S. (2019). Inferring protein 3D structure from deep mutation scans. *Nature Genetics*, **51**(7), 1170–1176.

Sadlish, H., Pitonzo, D., Johnson, A. E., and Skach, W. R. (2005). Sequential triage of transmembrane segments by Sec61α during biogenesis of a native multispanning membrane protein. *Nature Structural & Molecular Biology*, **12**(10), 870–878.

Sahoo, A., Khare, S., Devanarayanan, S., Jain, P. C., and Varadarajan, R. (2015). Residue proximity information and protein model discrimination using saturation-suppressor mutagenesis. *eLife*, **4**.

Salinas, V. H. and Ranganathan, R. (2018). Coevolution-based inference of amino acid interactions underlying protein function. *eLife*, **7**.

Sanyal, S., Coker, D. F., and MacKernan, D. (2016). How flexible is a protein: simple estimates using FRET microscopy. *Molecular bioSystems*, **12**(10), 2988–2991.

Schaarschmidt, J., Monastyrskyy, B., Kryshtafovych, A., and Bonvin, A. M. J. J. (2018). Assessment of contact predictions in CASP12: Co-evolution and deep learning coming of age. *Proteins: Structure, Function, and Bioinformatics*, **86**(S1), 51–66.

Schlessinger, A. and Rost, B. (2005). Protein flexibility and rigidity predicted from sequence. *Proteins*, **61**(1), 115–126.

Schmiedel, J. M. and Lehner, B. (2019). Determining protein structures using deep mutagenesis. *Nature Genetics*, **51**(7), 1177–1186.

Schreyer, A. and Blundell, T. (2009). CREDO: A Protein-Ligand Interaction Database for Drug Discovery. *Chemical Biology & Drug Design*, **73**(2), 157–167.

Schwarz, D., Merget, B., Deane, C., and Fulle, S. (2019). Modeling conformational flexibility of kinases in inactive states. *Proteins: Structure, Function and Bioinformatics*.

Schwarz, D., Georges, G., Kelm, S., Shi, J., Vangone, A., and Deane, C. M. (2021). Co-evolutionary Distance Predictions Contain Flexibility Information. *Bioinformatics (Oxford, England)*.

Seemayer, S., Gruber, M., and Söding, J. (2014). CCMpred–fast and precise prediction of protein residue-residue contacts from correlated mutations. *Bioinformatics (Oxford, England)*, **30**(21), 3128–30.

Senior, A. W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., Qin, C., Žídek, A., Nelson, A. W. R., Bridgland, A., Penedones, H., Petersen, S., Simonyan, K., Crossan, S., Kohli, P., Jones, D. T., Silver, D., Kavukcuoglu, K., and Hassabis, D. (2020). Improved protein structure prediction using potentials from deep learning. *Nature*, **577**(7792), 706–710.

Shen, Q., Wang, G., Li, S., Liu, X., Lu, S., Chen, Z., Song, K., Yan, J., Geng, L., Huang, Z., Huang, W., Chen, G., and Zhang, J. (2016). ASD v3.0: unraveling allosteric regulation with structural mechanisms and biological networks. *Nucleic Acids Research*, **44**(D1), D527–D535.

Shoemaker, S. C. and Ando, N. (2018). X-rays in the Cryo-Electron Microscopy Era: Structural Biology's Dynamic Future. *Biochemistry*, **57**(3), 277–285.

Skjaerven, L., Hollup, S. M., and Reuter, N. (2009). Normal mode analysis for proteins. *Journal of Molecular Structure: THEOCHEM*, **898**(1-3), 42–48.

Sol, A. d., Fujihashi, H., Amoros, D., and Nussinov, R. (2006). Residues crucial for maintaining short paths in network communication mediate signaling in proteins. *Molecular Systems Biology*, **2**, 2006.0019.

Stagno, J. R., Liu, Y., Bhandari, Y. R., Conrad, C. E., Panja, S., Swain, M., Fan, L., Nelson, G., Li, C., Wendel, D. R., White, T. A., Coe, J. D., Wiedorn, M. O., Knoska, J., Oberthuer, D., Tuckey, R. A., Yu, P., Dyba, M., Tarasov, S. G., Weierstall, U., Grant, T. D., Schwieters, C. D., Zhang, J., Ferré-D'Amaré, A. R., Fromme, P., Draper, D. E., Liang, M., Hunter, M. S., Boutet, S., Tan, K., Zuo, X., Ji, X., Barty, A., Zatsepin, N. A., Chapman, H. N., Spence, J. C. H., Woodson, S. A., and Wang, Y.-X. (2017). Structures of riboswitch RNA reaction states by mix-and-inject XFEL serial crystallography. *Nature*, **541**(7636), 242–246.

Starr, T. N., Greaney, A. J., Hilton, S. K., Ellis, D., Crawford, K. H. D., Dingens, A. S., Navarro, M. J., Bowen, J. E., Tortorici, M. A., Walls, A. C., King, N. P., Veesler, D., and Bloom, J. D. (2020). Deep Mutational Scanning of SARS-CoV-2 Receptor Binding Domain Reveals Constraints on Folding and ACE2 Binding. *Cell*, **182**(5), 1295–1310.

Sternke, M., Tripp, K. W., and Barrick, D. (2019). Consensus sequence design as a general strategy to create hyperstable, biologically active proteins. *Proceedings of the National Academy of Sciences*, **116**(23), 11275–11284.

Stolzenberg, S., Michino, M., LeVine, M. V., Weinstein, H., and Shi, L. (2016). Computational approaches to detect allosteric pathways in transmembrane molecular machines. *Biochimica et Biophysica Acta (BBA) - Biomembranes*, **1858**(7), 1652–1662.

Süel, G. M., Lockless, S. W., Wall, M. A., and Ranganathan, R. (2003). Evolutionarily conserved networks of residues mediate allosteric communication in proteins. *Nature Structural Biology*, **10**(1), 59–69.

Sun, Z., Liu, Q., Qu, G., Feng, Y., and Reetz, M. T. (2019). Utility of B-Factors in Protein Science: Interpreting Rigidity, Flexibility, and Internal Motion and Engineering Thermostability. *Chemical Reviews*, **119**(3), 1626–1665.

Tate, J. G., Bamford, S., Jubb, H. C., Sondka, Z., Beare, D. M., Bindal, N., Boutselakis, H., Cole, C. G., Creatore, C., Dawson, E., Fish, P., Harsha, B., Hathaway, C., Jupe, S. C., Kok, C. Y., Noble, K., Ponting, L., Ramshaw, C. C., Rye, C. E., Speedy, H. E., Stefancsik, R., Thompson, S. L., Wang, S., Ward, S., Campbell, P. J., and Forbes, S. A. (2019). COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Research*, **47**(D1), D941–D947.

Taylor, S. S. and Kornev, A. P. (2011). Protein kinases: evolution of dynamic regulatory proteins. *Trends Biochem Sci*, **36**(2), 65–77.

Tomczak, K., Czerwińska, P., and Wiznerowicz, M. (2015). The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemporary oncology (Poznan, Poland)*, **19**(1A), 68–77.

Toth-Petroczy, A., Palmedo, P., Ingraham, J., Hopf, T. A., Berger, B., Sander, C., and Marks, D. S. (2016). Structured States of Disordered Proteins from Genomic Sequences. *Cell*, **167**(1), 158–170.

Trott, O. and Olson, A. J. (2010). AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of computational chemistry*, **31**(2), 455–61.

Tsai, C.-J. and Nussinov, R. (2014). A Unified View of "How Allostery Works". *PLoS Computational Biology*, **10**(2), e1003394.

Tsai, C.-J., del Sol, A., and Nussinov, R. (2008). Allostery: absence of a change in shape does not imply that allostery is not at play. *Journal of molecular biology*, **378**(1), 1–11.

Ulmschneider, J. P., Smith, J. C., White, S. H., and Ulmschneider, M. B. (2011). In Silico Partitioning and Transmembrane Insertion of Hydrophobic Peptides under Equilibrium Conditions. *Journal of the American Chemical Society*, **133**(39), 15487–15495.

Ung, P. M. and Schlessinger, A. (2015). DFGmodel: predicting protein kinase structures in inactive states for structure-based discovery of type-II inhibitors. *ACS Chem Biol*, **10**(1), 269–278.

Ung, P. M.-U., Rahman, R., and Schlessinger, A. (2018). Redefining the Protein Kinase Conformational Space with Machine Learning. *Cell chemical biology*, **25**(7), 916–924.

van den Bedem, H., Bhabha, G., Yang, K., Wright, P. E., and Fraser, J. S. (2013). Automated identification of functional dynamic contact networks from X-ray crystallography. *Nature methods*, **10**(9), 896–902.

van der Lee, R., Buljan, M., Lang, B., Weatheritt, R. J., Daughdrill, G. W., Dunker, A. K., Fuxreiter, M., Gough, J., Gsponer, J., Jones, D. T., Kim, P. M., Kriwacki, R. W., Oldfield, C. J., Pappu, R. V., Tompa, P., Uversky, V. N., Wright, P. E., and Babu, M. M. (2014). Classification of intrinsically disordered regions and proteins. *Chemical reviews*, **114**(13), 6589–6631.

Virkki, M. T., Agrawal, N., Edsbäcker, E., Cristobal, S., Elofsson, A., and Kauko, A. (2014). Folding of Aquaporin 1: Multiple evidence that helix 3 can shift out of the membrane core. *Protein Science*, **23**(7), 981–992.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, I., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., Vijaykumar, A., Bardelli, A. P., Rothberg, A., Hilboll, A., Kloeckner, A., Scopatz, A., Lee, A., Rokem, A., Woods, C. N., Fulton, C., Masson, C., Häggström, C., Fitzgerald, C., Nicholson, D. A., Hagen, D. R., Pasechnik, D. V., Olivetti, E., Martin, E., Wieser, E., Silva, F., Lenders, F., Wilhelm, F., Young, G., Price, G. A., Ingold, G.-L., Allen, G. E., Lee, G. R., Audren, H., Probst, I., Dietrich, J. P., Silterra, J., Webber, J. T., Slavič, J., Nothman, J., Buchner, J., Kulick, J., Schönberger, J. L., de Miranda Cardoso, J. V., Reimer, J., Harrington, J., Rodríguez, J. L. C., Nunez-Iglesias, J., Kuczynski, J., Tritz, K., Thoma, M., Newville, M., Kümmerer, M., Bolingbroke, M., Tartre, M., Pak, M., Smith, N. J., Nowaczyk, N., Shebanov, N., Pavlyk, O., Brodtkorb, P. A., Lee, P., McGibbon, R. T., Feldbauer, R., Lewis, S., Tygier, S., Sievert, S., Vigna, S., Peterson, S., More, S., Pudlik, T., Oshima, T., Pingel, T. J., Robitaille, T. P., Spura, T., Jones, T. R., Cera, T., Leslie, T., Zito, T., Krauss, T., Upadhyay, U., Halchenko, Y. O., Vázquez-Baeza, Y., and Contributors, S. . (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*, **17**(3), 261–272.

Wang, J., Jain, A., McDonald, L. R., Gambogi, C., Lee, A. L., and Dokholyan, N. V. (2020). Mapping allosteric communications within individual proteins. *Nature Communications*, **11**(1), 3862.

Wang, S., Sun, S., Li, Z., Zhang, R., and Xu, J. (2017). Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model. *PLOS Computational Biology*, **13**(1), e1005324.

Weile, J. and Roth, F. P. (2018). Multiplexed assays of variant effects contribute to a growing genotype–phenotype atlas. *Human Genetics*, **137**(9), 665–678.

White, S. H. and Wimley, W. C. (1999). MEMBRANE PROTEIN FOLDING AND STABILITY: Physical Principles. *Annual Review of Biophysics and Biomolecular Structure*, **28**(1), 319–365.

Wolff, G., Limpens, R. W. A. L., Zevenhoven-Dobbe, J. C., Laugks, U., Zheng, S., de Jong, A. W. M., Koning, R. I., Agard, D. A., Grünewald, K., Koster, A. J., Snijder, E. J., and Bárcena, M. (2020). A molecular pore spans the double membrane of the coronavirus replication organelle. *Science (New York, N.Y.)*, **369**(6509), 1395–1398.

Xu, D., Meisburger, S. P., and Ando, N. (2021). Correlated Motions in Structural Biology. *Biochemistry*, **60**(30), 2331–2340.

Xu, J. (2019). Distance-based protein folding powered by deep learning. *Proceedings of the National Academy of Sciences of the United States of America*, **116**(34), 16856–16865.

Xu, M., Yu, L., Wan, B., Yu, L., and Huang, Q. (2011). Predicting inactive conformations of protein kinases using active structures: conformational selection of type-II inhibitors. *PLoS One*, **6**(7), e22644.

Xu, Y., Zhang, D., Rogawski, R., Nimigean, C. M., and McDermott, A. E. (2019). Identifying coupled clusters of allostery participants through chemical shift perturbations. *Proceedings of the National Academy of Sciences*, **116**(6), 2078 LP – 2085.

Yamamoto, K., Abe, D., Yoshimoto, N., Choi, M., Yamagishi, K., Tokiwa, H., Shimizu, M., Makishima, M., and Yamada, S. (2006). Vitamin D Receptor: Ligand Recognition and Allosteric Network [†]. *Journal of Medicinal Chemistry*, **49**(4), 1313–1324.

Yang, J., Anishchenko, I., Park, H., Peng, Z., Ovchinnikov, S., and Baker, D. (2020). Improved protein structure prediction using predicted interresidue orientations. *Proceedings of the National Academy of Sciences*, **117**(3), 1496 LP – 1503.

Yip, K. M., Fischer, N., Paknia, E., Chari, A., and Stark, H. (2020). Atomic-resolution protein structure determination by cryo-EM. *Nature*, **587**(7832), 157–161.

Zea, D. J., Monzon, A. M., Parisi, G., and Marino-Buslje, C. (2018). How is structural divergence related to evolutionary information? *Molecular Phylogenetics and Evolution*, **127**, 859–866.

Zhang, Y., Krieger, J., Mikulska-Ruminska, K., Kaynak, B., Sorzano, C. O. S., Carazo, J.-M., Xing, J., and Bahar, I. (2021). State-dependent sequential allostery exhibited by chaperonin TRiC/CCT revealed by network analysis of Cryo-EM maps. *Progress in biophysics and molecular biology*, **160**, 104–120.

Zhang, Z. (2020). Complete Extraction of Protein Dynamics Information in Hydrogen/Deuterium Exchange Mass Spectrometry Data. *Analytical Chemistry*, **92**(9), 6486–6494.

Zhao, J., Benlekbir, S., and Rubinstein, J. L. (2015). Electron cryomicroscopy observation of rotational states in a eukaryotic V-ATPase. *Nature*, **521**(7551), 241–245.

Zhao, Z., Wu, H., Wang, L., Liu, Y., Knapp, S., Liu, Q., and Gray, N. S. (2014). Exploration of Type II Binding Mode: A Privileged Approach for Kinase Inhibitor Focused Drug Discovery? *ACS Chemical Biology*, **9**(6), 1230–1241.

Zuccotto, F., Ardini, E., Casale, E., and Angiolini, M. (2010). Through the "gatekeeper door": exploiting the active kinase conformation. *Journal of medicinal chemistry*, **53**(7), 2681–2694.

Appendices

A

Supporting Information for Analysis of the Allosteric Database

Table A.1: ASD: Training set

Uniprot ID	Uniprot name	Length	Verified residues	PDB ID	Uniprot coverage	PDB sequence
P11473	VDR_HUMAN	427	Y143,Y147,F150,I238,E271,R274,W286,Y295,H397,Y401,V418,F422	4G2I_A	118-427	DSLRPKLSEEQQRHIAILLDAHHKTYDPTYSDFCQFRPPVRVNDGGGSVTLELSQLSMLPHLADLVSYSIQKVIGFAKMIPGFRDLTSEDQ IVLLKSSAIEVIMLRSNESPTMDDMSWTCGNQDYKYRVSDVTKAGHSLELIEPLIKFQVGLKKLNLHEEEHVLLMAICIVSPDRPGVQDAA LIEAIQDRLSNTLQTYIRCRHPPPGSHLLYAKMIQKLADLRSLNEEHSKQYRCLSFQPECSMKLTPLVLEVFGNEIS
P29466	CASP1_HUMAN	404	S339,N337,D336,T334,S333,S332,R286,T388,E390	6F6R_AB	118-297,317-404	QDNPAMPTSSGSEGNVKLCSLEEAQRIWKQKSAEIYPIMDKSSRTRLALIICNEEFDSIPRRTGAEVDITGMTMLLQNLGYSVDVKKNLTA SDMTTELEAFAHRPEHKTSDSTFLVFMSHGIREGICGKKHSEQVPDILQLNAIFNMLNTKNCPSLKDKPKVIIIQACRGDSPGVVWFKDAI KKAHIEKDFIAFCSSTPDNVSWRHPTMGSVFIGRLIEHMQEYACSCDVEEIFRKVRFSFEQPDGRAQMPTTERVTLTRCFYLFPGH
Q06187	BTK_HUMAN	659	W395,L460,T474,M449,F540,H519,D579,R544	4OT5_A	378-659	GSQQNKNAPSTAGLGYGSWEIDPKDLTFLKELGTGQFGVVKYGKWRGQYDVAIKMIKEGSMSEDEFIEEAKVMMNLSHEKLVQLYGVCTKQ RPIFITEYMANGCLLNYLREARHAFQTQQLLEMCKDVCEAMEYLESKQFLHRDLAARNCLVNDQGVVKVSDFGLSRYVLDDEYTSSVGSK FPVRWSPPEVLMYSKFSSKSDIWAFGVLMWEIYSLGKMPYERFTNSETAEHIAQGLRLYRPHLASAAVYTIMYSCWHEKADERPTFKILLS NILDVMDEES
P68871	HBB_HUMAN	147	V55,K67,L97,D100,P101,N103,F104	5UCU_B	2-147	VHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSEL HCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH
P08510	KCNAS_DROME	655	S460,A465,V476,A391,E395,T469,T442	2A79_B	356-489	MSLAILRVRLVRVFRIFKLSRHSKGLQILGQTLKASMRELGLLIFFLFIGVILFSSAVYFAEADERDSQFPSIPDAFWWAVVSMTTVGYG DMVPTTIGGKIVGSLCAIAGVLTIALPVPVIVSNFNYFYHRET
P07363	CHEA_ECOLI	654	M322,R325,M326,L507,T508,L509	1B3Q_A	261-640	SQTVRVDIEKLDNLMDLMIGELVIARSRILETLKKYNIKELDESLSHLSRITLDLQNVVMKIRMVPISFVFNRFPRMVRDLAKKMNKEVNFI MRGEDTELDRTFVEEIGEPLLHLLRNAIDHGIEPKEERIAKGKPPIGTLILSARHEGNVVIEVEDDGRGIDKEKIRKRAIEKGLIDESKA ATLSDQELLNFLFVPGFSTKEKVSEVSGRGVGMDVVKNVVESLNGSMGIESEKDKGTKVTIRLPLTLAIACALLVKVGNSTVAIPIANIDT ILSISKEDIQRVQDRDVIVIRGEVIPVYRLWEVLQIEHKEELEEMEAVIVRVGNRKYGIVVDDLLGQDDIVIKSLGKVFSEVKEFSGAAIL GDGSIALIINVSGI
P0A6Y8	DNAK_ECOLI	638	D326,N415,Q442,G443,E444,R445	5NRO_A	1-606	MGKIIGIDLGTTNSCVAIMDGTTPRVLENAEGDRITTPSIIAYTQDGCTLVGQPAKRQAVTNPQNTLFAIKRLIGRRFQDEEVQRDVSIMPF KIIAADNGDAWVEVKGQKMAPPQISAEVLKKMKKTAEDYLGEPVTEAVITVPAYFNDAQRQATKDAGRIAGLEVKRIINEPTAAALAYGLD KGTGNRTIAVYDLGGGAFDISIIEIDEVDGEKTFEVLATNGDTHLGGEDFDSRLINYLVEEFKKDQGIDLRNDPLAMQRLKEAAEKAKIEL SSAQQTDVNLPYITADATGPKHMNIKVTRAKLESLVEDLVNRSIEPLKVALQDAGLSVSDIDDVILVGGQTRMPMVQKKVAEFFGKEPRKD VNPDEAVAIGAAVQGGVLTGDVKDVLLLDVTPLSLGIETMGGVMTTLIAKNTTIPTKHSQVFSTAEDNQSAVTIHVLQGERKRAADNKSLG QFNLDGINPAPRGMPQIEVTFDIDADGILHVSAKDKNSGKEQKITIKASSGLNEDEIQKMVRDAEANAEADRKCEELVQTRNQGDHLLHST RKQVEEAGDKLPADDKTAIESALTALETALKGEDKAAIEAKMQELAQVSQKLMEIAQQQHHHHHH
P35520	CBS_HUMAN	551	T257,T260,R266,C52,H65	4COO_A	1-515,526-551	MPSETPQAEVGPTGCPHRSGPHSAKGSLEKGSPEDKEAKEPLWIRPDAPSRCTWQLGRPASESPHHHTAPAKSPKILPDILKKIGDTPMVR INKIGKKFGLKCELLAKCEFFNAGGSVKDRISLRMIEDAERDGTLKPGDTIIEPTSGNTGIGLALAAAVRGYRCIIVMPEKMSSEKVDVLR ALGAEIVRTPTNARFDSPESHVGVAWRLKNEIPNSHILDQYRNASNPLAHYDTTADEILQQCDGKLDMLVASVGTGGTITGIARKLKEKCP GCRIHGVDPEGSILAEPEELNQTEQITTYEVEGIGYDFIPTVLDRTVVDKWFKSNDEEAFTFARMLIAQEGLLCGGSAGSTVAVAVKAAQEL QEGQRCVVILPDSVRNYMTKFLSDRWMLQKGFLKEEDLTEKKPWWWHLRVQELGLSAPLTVLPTTCGHTIEHLREKGFDQAPVVDEAGVI LGMVTLGNMLSSLLAGKVQPSDQVGKVIYKQFKQIRLTDTLGRLSHILEMDHFALVVHEQQRQMVFGVVTADDLLNFVAAQERDQK
P0A731	MGSA_ECOLI	152	D91,D71,D101,D20	1IK4_A	1-152	MELTTRTLPARKHIALVAHDHCKQMLMSWVERHQPLLEQHVLYATGTTGNLISRATGMNVNAMLSGPMGGDQQVGALISEGKIDVLIFFWD PLNAVPHDPDVKALLRLATVWNIPVATNVATADFIIQSPHFNDAVDILIPDVQRYLADRLK
Q07812	BAX_HUMAN	192	L63,K64,R65,R94	4BD2_A	1-171	MDGSGEQPRGGGPTSSEQIMKTGALLLQGFIQDRAGRMGGEAPELALDPVPQDASTKKLSESLKRIGDELDSNMELQRMIAAVDTDSPREV FFRVAADMFSDGNFNWGRVVALFYFASKLVLKALSTKVPELIRTIMGWTLDFLRERLLGWIQDQGGWDGLLSYFGTPTWQSS
O95398	RPGF3_HUMAN	923	L315,A322,F342,E350	4MGI_E	213-923	PDAHMRMILRKPPGQRTVDDLEHIYDELLHIKALSHLSTTVKRELAGVLIFESHAKGGTVLFNQGEEGTSWYIILKGSVNVVIYGKGVVCT LHEGDDFGKLALVNDAPRAASIVLREDNCHFLRVDKEDFNRILRDVEANTVRLKEHDQDVLVLEKVPAGNRAANQGNSQPQQKYTVMSGTP EKILEHFLETIRLEPSLNEATDSVLNDFVMMHCVFMPNTQLCPAIVAHYHAQPSQQTEQERMDYALNVKERNVRIVLQWAAMYGDLLQEDD VAMAFLEEFYVSVSDDARMMAAFKEQLPELEKIVKQISEDAKAPQKKHKVLLQQFNTGDERAQKRQPIRGSDEVLFKVYCIDHTYTTIRVP VAASVKEVISAVADKLGSGEGLIIVKMNSGGEKVVLKSNDVSVFTTLTINGRLFACPREQFDSLITPLPEQEGPTTGTVGTFELMSSKDLAY QMTTYDWELFNCVHELELIYHTFGRHNFKKTTANLDLFLRRFNEIQFWVVTEVCLCSQLSKRVQLLKKFIKIAAHCKEYKNLNSFFAIVMG LSNVAVSRLALTWEKLPSKFKKFYAEFESLMDPSRNHRAYRLTAAKLEPPLIPFMPLLIKDMTFTHEGNKTFIDNLVNFEKMRMIANTART VRYYRSQPFNPDAAQANKNHQDVRSYVRQLNVIDNQRTLSQMSHRLEP
P01130	LDLR_HUMAN	860	H211,H583,H607	1N7D_A	22-720	AVGDRCERNEFQCQDGKCISYKWVCDGSAECQDGSDESQETCLSVTCKSGDFSCGGRVNRCIPQFWRCDGQVDCDNGSDEQGCPPKTCSQD EFRCHDGKCISRQFVCDSDRDCLDGSDEASCPVLTCGPASFQCNSSTCIPQLWACDNDPDCEDGSDEWPQRCRGLYVFQGDSSPCSAFEFH CLSGECIHSSWRCDGGPDCKDKSDEENCAVATCRPDEFQCSDGNCIHGSRQCDREYDCKDMSDEVGCVNVTLCEGPNKFKCHSGECITLDK VCNMARDCRDWSDEPIKECGTNECLDNNGGCSHVCNDLKIGYECLCPDGFQLVAQRRCEDIDECQDPDTCSQLCVNLEGGYKCQCEEGFQL DPHTKACKAVGSIAYLFFTNRHEVRKMTLDRSEYTSLIPNLRNVIALDTEVASNRIYWSDLSQRMICSTQLDRAHGVSSYDTVISRDIQAP DGLAVDWIHSNIYWTDSVLGTVSVADTKGVKRKTLFREQGSKPRAIVVDPVHGFMYWTDWGTPAKIKKGGLNGVDIYSLVTENIQWPNGIT LDLLSGRLYWVDSKLHSISSIDVNGGNRKTILEDEKRLAHPFSLAVFEDKVFWTDIINEAIFSANRLTGSDVNLLAENLLSPEDMVLFHQL TQPRGVNWCERTTLSNGGCQYLCLPAPQINPHSPKFTCACPDGMLLARDMRSCLTEAEAAVA
P69905	HBA_HUMAN	142	D95,P96,N98	6BWP_A	2-142	VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKL RVDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR
P31016	DLG4_RAT	724	V362,H372,V386	5MZ7_A	303-402	GGEEDIPREPRRIVIHRGSTGLGFNIIGGEDGEGIFISFALAGGPADLSGELRKGDQILSVNGVDLRNASHEQAAIALKNAGQTVTIHAQY KPEEYSRFEA
Q7NDN8	GLIC_GLOVI	359	D133,E219,D220	6HZW_A	43-359	GQDMVSPPPPIADEPLTVNTGIYLIECYSLDDKAETFKVNAFLSLSWKDRRLAFDPVRSGVRVKTYEPEAIWIPEIRFVNVENARDADVVD ISVSPDGTVQYLERFSARVLSPLDFRRYPFDSQTLHIYLIVRSVDTRNIVLAVDLEKVGKNDDVFLTGWDIESFTAVVKPANFALEDRLES KLDYQLRISRQVFSYIPNIILPMLFILFISWTAFWSTSYEANVTLVVSTLIAHIAFNILVETNLPKTPYMTYTGAIIFMIVLFYPVAVIEV TVQHYLKVESQPARAASITRASRIAFPVVFLLANIILAFLFFGF
P96169	SGLT_VIBPH	543	N64,Y263	2XQ2_A	1-543	MSNIEHGLSFIDIMVFAIYVAIIIGVGIAVSRDKKGTQKSTEDYFLAGKSLPWWAVGASLIAANISAEQFIGMSGSGYSIGLAIASYEWMS AITLIIVGKYFLPIFIEKGIYTIPEFVEKRFNKKLKTILAVFWISLYIFVNLTSVLYLGGLALETILGIPLMYSILGLALFALVYSIYGGL SAVVWTDVIQVFFLVLGGFMTTYMAVSFIGGTDGWFAGVSKMVDAAPGHFEMILDQSNPQYMNLPGIAVLIGGLWVVANLYYWGFNQYIIQR TLAAKSVSEAQKGIVFAAFLALIVPFLVVLPGIAAYVITSDPQDLMASLGDIAATNLPSAANALDKAYPWLTQFLPVGVKGVVFAALAAAIVS SLASMLNSTATIFTMDIYKEYISPDSGDHKLVNVGRTAAVVALIIAAIIAPMLGGIGQCFQYIQEYTGLVSPGILAVFLLGLFWKKTTSKG AIIGVVASIPFALFLKFMPLSMPFMDQMLYTLLFTMVVIAFTSLSTSINDDDPKGISVTSSMFVTDRSFNIAYGIMIVLAVLYTLFW
P00734	THRB_HUMAN	622	D462,W590	5JZY_H	364-622	IVEGSDAEIGMSPWQVMLFRKSPQELLCGASLISDRWVLTAAHCLLYPPWDKNFTENDLLVRIGKHSRTRYERNIEKISMLEKIYIHPRYN WRENLDRDIALMKLKKPVAFSDYIHPVCLPDRETAASLLQAGYKGRVTGWGNLKETWTANVGKGQPSVLQVVNLPIVERPVCKDSTRIRIT DNMFCAGYKPDEGKRGDACEGDSGGPFVMKSPFNNRWYQMGIVSWGEGCDRDGKYGFYTHVFRLKKWIQKVIDQFGE
P06746	DPOLB_HUMAN	335	Y265	6CTU_A	1-335	MSKRKAPQETLNGGITDMLTELANFEKNVSQAIHKYNAYRKAASVIAKYPHKIKSGAEAKKLPGVGTKILEKIDEFLATGKLRKLEKIRQD DTSSSINFLTRVSGIGPSAARKFVDEGIKTLEDLRKNEDKLNHHQRIGLKYFGDFEKRIPREEMLQMQDIVLNEVKKVDSEYIATVCGSFR RGAESSGDMDVLLTHPSFTSESTKQPKLLHQVVEQLQKVHFITDTLSKGETKFMGVCQLPSKNDEKEYPHRRIDIRLIPKDQYYCGVLYFT GSDIFNKNMRAHALEKGFTINEYTIRPLGVTGVAGEPLPVDSEKDIFDYIQWKYREPKDRSE
P10721	KIT_HUMAN	976	D816	4HVS_A	551-934	PMYEVQWKVVEESNGNNYSYIDPTQLPYDHKWEFPRNRLSFGKTLGAGAFGKVVEATAQGLIKSDAAMTVAVKMLKPSAHSTEREALMSEL KVLSYLGNHENIVNLLGACTHGGPTLVITEYCCYGDLLNFLRRKRDSFISSDTSPASMEDDENALDLEDLLSFSYQVAKGMAFLASKNCIH RDLAARNILLTHGRITKICDFGLARDIKNDSNYVVKGNARLPVKWMAPESIFNSVYTFESDVWSYGIFLWELFSLGSSPYPGMPVDSKFYK MIKEGFRMSSPEHAPAEMYDIMKTCWDADPDKRPTFKQIVQDIEKQISESTNH

Table A.2: ASD: Test set

Uniprot ID	Uniprot name	Length	Verified residues
P14416	DRD2_HUMAN	443	C118,S193,T205,L379,N418,V83,M117,Y199,L414
P51178	PLCD1_HUMAN	756	K86,F87,K32,K30,K57,K43,K102,K127
Q14416	GRM2_HUMAN	872	C500,C519,C504,C522,C525,C537,C553,C540
Q08881	ITK_HUMAN	620	W356,M411,M410,L421,F501,H480,D540
P0ABQ4	DYR_ECOLI	159	D27,F31,M42,L54,T113,G121
P21888	SYC_ECOLI	461	H40,R42,M294,H297,E354,R427
P0AG30	RHO_ECOLI	419	R347,E333,D328,K298,R269
P03023	LACI_ECOLI	360	V94,V95,V96,S97,M98
Q8RT53	Q8RT53_GEOSE	293	R226,W228,V263,N262
P0A7B8	HSLV_ECOLI	176	L72,V77,L89,L92
P41789	NTRC_SALTY	469	T82,S85,D86
P02711	ACHA_TORMA	461	Y214,Y222,D224
P9WP48	CSOR_MYCTO	119	Y35,E81,H61
P29678	MP2K1_RABIT	393	S218,S222,S298
Q9JJZ8	CNGA3_MOUSE	631	R377,F488
P60010	ACT_YEAST	375	K113,E195
Q9X0C6	HIS6_THEMA	253	V48,L50
P19821	DPO1_THEAQ	832	I614
Q01196	RUNX1_HUMAN	453	R164

B

Supporting Information for Co-evolutionary Distance Predictions Contain Flexibility Information

Table B.1: MD-validated rigid loop set

pdb_id	chain_id	loop_start	loop_end
1FKB	A	64	70
1GPR	A	14	20
1E1Q	A	44	50
1OPS	A	34	43
1GPR	A	151	157
1WHI	A	11	17
1VMO	A	26	32
1OPS	A	17	22
1OPS	A	50	59
1OPS	A	26	32
1RIS	A	81	86
3FIB	A	215	229
2HNP	A	257	263
1CHU	A	271	277
1FQN	A	29	38
1DYW	A	131	136
1WHI	A	88	103
1E1Q	A	64	70
1UBQ	A	16	22
1GE8	A	54	64

Table B.2: PDB structure pairs of the CoDNaS maximum-RMSD-pairs rigid subset (3075 pairs). PDB 1 sequence is the 'target_sequence' that was used for DMPfold distance predictions. '1' in the 'analysis_complete' column indicates all targets used in this work (2947 pairs). 'in_unique_CATH_subset' indicates the structure pairs analysed after removing redundant domains (517 pairs).

pdb_id_1	chain_id_1	pdb_id_2	chain_id_2	analysis _complete	in_unique _CATH_subset
2EB8	A	2JHO	A	1	1
1HMY	A	2C7Q	A	1	1
1BSR	B	3DJX	A	1	1
2A2S	B	3PGT	B	1	1
1DZB	Y	1JTP	M	1	1
13PK	C	16PK	A	1	1
150L	D	256L	A	1	1
2B70	A	2RB2	X	1	0
1DAA	A	2DAA	A	1	1
1ETS	H	3PMA	D	1	1
3UA6	B	3UA7	B	1	0
1A0S	R	1OH2	P	1	1
1A16	A	1M35	B	1	1
1A2D	B	1LIE	A	1	1
1BRS	D	1X1U	E	1	1
1A2J	A	4WEY	A	1	0
1A26	A	2PAW	A	1	1
1FPH	L	1QJ7	A	1	0
1A2K	B	1GY6	A	1	1
1B2X	B	1YVS	A	1	1
1A3U	A	1NUC	A	1	1
1A2Y	B	1VFA	B	1	1
1A30	B	3I8W	A	1	1
1A58	A	1C5F	A	1	1
1OJJ	B	1OJK	A	1	1
1OIB	A	1PBP	A	1	1
1A49	C	1A5U	H	1	1
1A4B	B	1A4C	C	1	1
1A4G	B	1A4Q	A	1	1
1B2L	A	1SBY	A	1	1
3ZOW	R	4JCG	A	1	0
1A5D	B	1ZIE	A	1	1
1A5O	A	1EJX	A	1	1
2CLE	B	2CLO	B	1	1
1A6V	J	1A6W	H	1	0
1A6Y	B	1GA5	A	1	1
1CZ0	B	1EVX	A	1	1
1BYI	A	1DBS	A	1	1
1G5N	A	2IE6	A	1	1
1VIK	B	2QMP	A	1	0
2CIK	A	4PRP	A	1	0
1LRJ	A	1XEL	A	1	0
1AUS	V	1UPM	S	1	1
1T5K	B	2J55	A	1	0
2CHA	F	3RU4	D	1	0
1CA0	H	1N8O	C	1	0
7ABP	A	8ABP	A	1	1
2RBU	X	2RBX	X	1	1
1Q95	G	1TUG	B	1	1
1MG0	C	8ADH	A	1	0
4BWU	B	4BXV	A	1	0
1MJJ	A	1MJU	L	1	0
1FIF	B	1FIH	A	1	1
1FK1	A	1MZL	A	1	1
1AHF	B	1AHX	A	1	1
1AHH	B	1AHI	A	1	0
2AHJ	A	2ZPG	A	1	1
2AHJ	D	3A8L	B	1	1
2ASV	A	2AZD	B	1	1
1SJS	A	9ICD	A	1	1
1FXV	A	1PNM	A	1	1
1M79	A	4HOF	B	1	1
1O8E	A	1PLU	A	1	1
2NGR	A	4ITR	C	1	0
1C1H	A	1N0I	A	1	1
1IVR	A	8AAT	A	1	0
4AKE	A	4JZK	B	1	0
4JFD	A	4JFP	D	1	0
1AKZ	A	1Q3F	A	1	1
1AL6	A	6CSC	A	1	1
1NX1	B	4PHN	B	1	1
1QDS	A	2Y63	A	1	1
1F4E	A	2KCE	A	1	1
1I07	B	1I0C	B	1	1
1AOZ	B	1ASO	A	1	0
3RV5	C	4GJE	A	1	0
4E7V	S	4M4H	A	1	0
2WEA	A	2WED	A	1	0
2BU1	C	2C4Q	B	1	1
1QQ6	A	1QQ7	B	1	1

Table B.2 continued from previous page

pdb_id_1	chain_id_1	pdb_id_2	chain_id_2	analysis _complete	in_unique _CATH_subset
1AS6	B	1AS7	A	1	0
1AQH	A	1B0I	A	1	1
2IH4	A	2NP7	A	1	0
1BSM	A	1BT8	B	1	1
1ELM	P	2ABZ	B	1	1
1HSR	A	2E3B	A	1	0
1IAA	A	1QJI	A	1	1
4I60	A	4JHQ	B	1	0
1AVH	B	2XO3	A	1	0
2ZVV	Y	2ZVW	M	1	0
1AXG	D	4NG5	A	1	0
1GPU	B	1NGS	B	1	1
1AY7	A	1LNI	B	1	0
1YNR	C	3VYM	A	1	1
1BBC	A	3U8D	B	1	1
1AYY	C	2GL9	C	1	0
1AYY	B	2GAW	B	1	1
1CZM	A	3LXE	A	1	1
1CA5	A	1WD0	A	1	1
1AZZ	D	1ECY	A	1	1
1BSF	B	1BSP	B	1	0
1JEV	A	1RKM	A	1	0
1B09	E	3PVN	F	1	1
1EAI	B	2DE8	A	1	0
3Q76	B	4NZL	A	1	0
1B0W	A	1BRE	C	1	0
1FHH	A	1R0J	A	1	1
1RN4	A	4RNT	A	1	0
3ZBU	A	3ZC3	B	1	0
1B3Y	A	1BG4	A	1	0
1B49	A	1B5E	A	1	0
1I8J	A	1L6Y	B	1	0
1B4F	F	1F0M	A	1	1
1OPA	A	1OPB	A	1	0
1NDP	B	1S5Z	F	1	1
1B66	A	1GTQ	B	1	1
2EDC	A	2PKY	X	1	1
7HVP	B	8HVP	A	1	0
1D4L	B	3BXS	A	1	0
1B6T	B	1GN8	A	1	1
1B71	A	1LKM	A	1	0
2WE4	D	2WE5	A	1	1
3NEL	B	3NEM	A	1	1
1J18	A	5BCA	D	1	0
1B93	A	1IK4	E	1	1
1B9S	A	1IVB	A	1	0
4CPA	B	7CPA	A	1	0
1BB6	A	1LMP	A	1	0
1BBR	L	1UCY	J	1	0
1L0L	H	1PP9	U	1	1
2HTN	H	3E1L	L	1	1
1JLR	B	1JLS	B	1	1
1BD9	B	2QYQ	A	1	1
1NTM	C	2A06	P	1	1
1L0L	E	2A06	E	1	1
1BE9	A	1BFE	A	1	1
1BG8	C	1DJ8	B	1	1
1RIW	C	2HNT	F	1	0
1J35	A	1X2T	A	1	0
1IXX	D	1X2W	B	1	0
3BJL	B	4BJL	B	1	0
4X19	N	4X1C	N	1	0
1HT3	A	1PFG	A	1	1
1IPS	A	3ZKU	A	1	1
2O0W	A	2O17	A	1	0
1ORB	A	1RHS	A	1	1
2ULL	A	3QGJ	A	1	0
3N0P	B	3NCB	B	1	0
1POP	A	4KP9	B	1	1
1BQ4	D	4PGM	D	1	1
1HAW	A	2ZON	B	1	0
2UX6	A	2UXG	A	1	0
1T4B	B	1T4D	B	1	0
3BAW	A	4GQQ	A	1	0
1OHP	B	1QJG	E	1	0
1RZL	A	1UVB	A	1	0
1RDN	1	1RDO	2	1	0
1BV9	A	1MES	B	1	0
1PHR	A	1Z12	A	1	0
1KXQ	C	1KXV	A	1	0
1XNB	A	3HD8	B	1	1
1BWA	A	1MEU	B	1	0
1FXW	A	1WAB	A	1	1
1V13	A	1V15	A	1	1
1CE8	H	1JDB	I	1	1
1FZM	B	1S7V	B	1	0
2ZQ7	A	2ZQ9	A	1	1
2P42	B	2P49	B	1	0
1ZP5	A	3DPF	B	1	0
1UAC	L	2DQF	D	1	0

Table B.2 continued from previous page

pdb_id_1	chain_id_1	pdb_id_2	chain_id_2	analysis _complete	in_unique _CATH_subset
1UA6	H	1UAC	H	1	0
3FL1	A	3FL3	B	1	0
1C0F	S	3A5M	S	1	1
2VDR	L	3ZE2	F	1	0
1C3X	B	1QE5	C	1	1
1C4F	A	1Q4A	A	1	1
1C5E	B	1TCZ	E	1	1
1C6N	A	1L89	A	1	0
1C6Y	B	1C6Z	A	1	0
1C78	B	1C79	A	1	1
1QL3	A	1QL4	C	1	0
1C7N	H	1C7O	G	1	0
1C81	A	1TIP	B	1	0
1EGU	A	1OJP	A	1	1
1D00	B	1D0J	A	1	1
1CB2	B	1QJW	B	1	1
1CB7	D	1I9C	D	1	1
4KO2	T	4KO4	S	1	0
1CC7	A	3K7R	G	1	1
1U74	A	2BCN	C	1	0
1FQO	A	1FSF	A	1	1
1CXI	A	2CXG	A	1	0
1B8C	B	1B8L	A	1	0
1LDS	A	4NQV	B	1	0
1IKG	A	1SCW	A	1	0
1AYI	A	1UNK	C	1	1
1CEL	B	2V3I	A	1	0
1CEV	B	2CEV	A	1	1
3ESA	B	3ESB	A	1	0
1CFM	A	1EWH	B	1	1
3GK4	X	3LLE	A	1	0
1DCD	A	1DXG	A	1	0
1CGU	A	5CGT	A	1	0
1CHO	I	1PPF	I	1	1
2DVD	B	2PEL	D	1	0
1E1L	A	1E6E	A	1	0
1HQN	A	1HX6	C	1	1
1FIK	A	2PBD	P	1	1
1KDT	B	2CMK	A	1	0
1OXL	A	1Y38	B	1	0
1J2A	A	1VAI	A	1	0
2FHT	A	2FJ2	A	1	0
1MJM	A	1MJO	D	1	1
3E8C	L	4WB6	J	1	0
1CMS	A	1CZI	E	1	0
1J16	A	1J17	T	1	0
1COM	C	1DBF	C	1	1
1COZ	A	1N1D	A	1	0
1RYC	A	2EUU	A	1	0
1NBB	B	1RCP	A	1	1
3OZU	A	3OZV	A	1	0
4ONF	L	4ONG	L	1	0
3NBS	A	3NBT	B	1	0
1CSB	D	1HUC	C	1	1
1CSB	B	2IPP	B	1	0
3CSC	A	6CTS	A	1	0
2QA9	E	2SGD	E	1	0
1CSP	A	2ES2	A	1	0
1CY5	A	4RHW	D	1	1
1CXA	A	1L9B	C	1	0
1NB5	I	3KSE	D	1	1
1CZU	A	1OFV	A	1	1
1EWY	C	1FXA	B	1	1
4KQH	A	4KQI	A	1	0
1HQ5	B	1RLA	B	1	0
1D5N	A	3K9S	B	1	0
1D5W	C	1DBW	A	1	0
1PAF	B	1QCJ	B	1	1
1FCJ	A	1OAS	B	1	0
1OAC	A	2WGQ	B	1	1
1V3J	A	1V3L	B	1	0
1D7L	A	1IUV	A	1	1
1ZC9	A	2DKB	A	1	0
2GQW	A	2YVJ	P	1	0
1KRP	A	1QSL	A	1	1
1DAP	A	1F06	B	1	0
2GD1	P	3CMC	R	1	0
1DCO	H	1DCP	G	1	1
1ZAW	V	1ZAX	Z	1	0
1DD6	A	1JJT	B	1	1
4EIZ	B	4KJJ	A	1	0
1G1M	B	1M34	E	1	0
1K7C	A	1PP4	A	1	0
1RRW	A	2NM2	D	1	1
1YXD	B	2ATS	A	1	0
1DIO	L	1EGM	A	1	1
1GHI	A	1GHP	A	1	0
1PT9	B	1U31	A	1	1
1EAY	B	1F4V	A	1	0
1DJN	B	1O94	A	1	0

Table B.2 continued from previous page

pdb_id_1	chain_id_1	pdb_id_2	chain_id_2	analysis _complete	in_unique _CATH_subset
1DJR	F	1HTL	G	1	1
1DKM	A	1DKN	A	1	0
1EYN	A	4E7C	A	1	1
1DLQ	B	1DMH	A	1	1
1DO6	A	1DQI	B	1	1
1JUB	A	1OVD	B	1	0
2VV7	B	2VV8	A	1	1
3KAN	C	4Q3F	B	1	1
1WC7	L	1WCB	A	1	0
1DQQ	B	1NBZ	B	1	0
1DR6	A	1DR7	A	1	0
1RA1	A	1RG7	A	1	0
1DS4	A	1DSP	A	1	0
1ICU	D	1OO5	B	1	1
1DTE	B	4EA6	B	1	0
1MLZ	A	1QJ5	B	1	0
1RNJ	A	1SYL	A	1	1
1DVF	A	1VFA	A	1	0
1DW1	B	1DW3	A	1	0
1FEJ	C	3NUO	B	1	0
1DW9	A	1DWK	D	1	1
1DWP	A	1DWQ	B	1	0
1DYW	A	2IGV	A	1	0
1DZ3	A	1QMP	B	1	0
1DZP	A	1HQP	A	1	0
1X6P	A	1X6Z	A	1	1
1E0W	A	1OD8	A	1	0
1UR8	B	1W1Y	A	1	0
2REQ	C	6REQ	A	1	0
1REQ	D	2REQ	D	1	0
1GNT	A	1OA1	A	1	1
1E4N	B	1V08	A	1	0
1K7D	B	1PNM	B	1	1
1E3S	C	1E3W	D	1	0
1H3T	B	1H3X	A	1	0
1DWI	M	1DWJ	M	1	0
1IJU	D	1IJV	A	1	0
1QHZ	A	2V38	A	1	0
1E5L	A	1E5Q	H	1	0
1O82	D	1O83	C	1	1
1E8M	A	1H2Z	A	1	0
1E8T	B	1USR	B	1	0
1Q0E	B	2SOD	O	1	1
2IXT	A	3D43	B	1	0
1EAO	B	1EAQ	A	1	1
3BN9	A	3NPS	A	1	0
1JM0	E	1JMB	C	1	1
1ECC	B	1ECJ	A	1	0
1NDO	A	1O7H	A	1	0
1O7N	B	4HM0	B	1	0
1EHK	B	3S8G	B	1	0
1EJ3	A	1UHI	B	1	0
2BUI	A	2VB7	C	1	1
1EK5	A	1EK6	B	1	0
1ELQ	A	1ELU	B	1	0
1EMC	D	1EML	A	1	0
1XYO	A	4HKW	A	1	0
1EO2	A	2BUW	A	1	0
1EOC	B	2BUX	B	1	0
1EP5	B	1EP6	A	1	0
1X8T	A	2AAY	A	1	0
1EQZ	A	2ARO	E	1	1
1KT3	A	1KT7	A	1	0
1ERM	A	1S0W	B	1	0
2GOO	D	2H62	B	1	1
1M6Z	C	1M70	A	1	0
1EUH	A	1QI6	D	1	1
1T68	X	2AL0	X	1	0
1N1E	B	1N1G	A	1	0
2GLQ	A	3MK2	A	1	1
3DR9	A	4GZG	B	1	0
1EYD	A	1STG	A	1	0
1EYT	A	1IUA	A	1	1
1EYY	D	1EZ0	D	1	0
1P84	B	3CX5	B	1	1
1EZV	C	1P84	C	1	0
1P84	E	3CX5	P	1	0
3CX5	U	3CXH	U	1	0
3CX5	V	3CXH	V	1	0
1JCV	A	1SDY	B	1	0
1F1U	B	1F1V	A	1	1
1F38	C	1L3I	F	1	0
1F47	B	1S1S	B	1	1
1DGB	B	1DGH	C	1	0
1F4M	A	1F4N	A	1	1
1K94	B	1K95	A	1	0
1KDK	A	1KDM	A	1	0
3TGI	I	3TGJ	I	1	1
1F61	B	1F8M	B	1	0
1F6R	F	1F6S	E	1	0

Table B.2 continued from previous page

pdb_id_1	chain_id_1	pdb_id_2	chain_id_2	analysis _complete	in_unique _CATH_subset
1TSU	A	2FNS	B	1	0
1F7D	A	1F7O	A	1	0
1F8Q	A	1MRH	A	1	0
1F8R	B	1F8S	G	1	0
1F9O	A	1G9G	A	1	1
1F9Q	A	1RHP	A	1	0
1F9Z	A	1FA7	B	1	0
1FD3	B	1FD4	H	1	1
1DFP	A	1DSU	A	1	0
1FEE	A	3IWL	A	1	0
1FEA	D	1FEC	B	1	0
1FFF	C	2QCI	B	1	0
1VQ7	C	1VQ9	C	1	1
1YN7	A	3WS3	A	1	0
2JCK	A	2VFZ	B	1	1
1FG6	C	1FG8	D	1	0
1PZT	A	1PZY	D	1	0
3H6S	D	3KFQ	B	1	0
1FHA	A	2CEI	A	1	0
1FIC	A	1FID	A	1	1
1FIQ	C	3B9J	C	1	1
1FIU	C	4ABT	B	1	1
5NLL	A	5ULL	A	1	0
1FMA	D	1JWB	D	1	0
1GDN	A	2VU8	E	1	0
1L5H	B	3K1A	B	1	1
1FQI	A	1FQJ	E	1	1
1FR7	B	1FSR	B	1	0
1FS8	A	2E81	A	1	0
1HDA	A	3PIA	C	1	0
1HDA	B	2QSS	D	1	0
1W5A	B	1W5B	A	1	1
1FTR	D	2FHJ	C	1	1
1P7L	C	1RG9	B	1	1
1FUO	B	1YFE	A	1	1
1FV2	A	1FV3	A	1	0
1TT8	A	2AHC	B	1	1
1EF2	B	4EP8	B	1	1
1FWL	B	1H72	C	1	1
1WSB	A	1XYY	A	1	0
1FX5	A	1JXN	C	1	0
1AX2	A	1AXZ	A	1	0
1XMF	A	1XVB	B	1	1
1FZ6	D	1XVE	C	1	0
1XVD	F	1XVE	E	1	1
1FZG	D	3E1I	D	1	1
1NS6	A	2ZLV	A	1	0
1G0F	A	4HF3	A	1	0
1GE5	A	1GE6	A	1	0
1G1A	D	1KEW	B	1	0
1G1Q	B	1G1S	B	1	0
1I80	B	1N3I	A	1	0
1G3K	C	1OFH	I	1	0
1HY3	A	4JVL	B	1	0
1G48	A	1I9O	A	1	0
1RID	A	1Y8E	B	1	1
2CUZ	A	2CV1	B	1	0
1G6O	B	2PT7	F	1	0
2AD6	A	4AAH	A	1	1
2AD7	B	4AAH	D	1	1
1GGX	B	1ZGO	D	1	0
1G87	B	1K72	A	1	0
1SMN	A	4E3Y	A	1	1
1DQR	B	1KOJ	A	1	1
1RZ8	C	1YYM	Q	1	0
2I5Y	G	2I60	P	1	1
1GCE	A	1ONH	A	1	0
1GA1	A	1GA6	A	1	0
3BX1	A	4CFZ	A	1	0
1GCP	B	1GCQ	C	1	0
1LH6	A	2GDM	A	1	0
1GEF	E	1IPI	B	1	1
3A13	J	3KDN	B	1	1
1GES	A	1GEU	B	1	0
1IR2	T	1UZH	E	1	0
1GK8	I	2V6A	J	1	0
3ZI7	A	4BAG	A	1	0
1GMU	D	1GMW	A	1	1
1OA0	B	1UPX	A	1	0
2XFP	A	4A79	B	1	0
1GOU	A	1GOY	A	1	0
2CEQ	A	2CER	A	1	0
1GPI	A	1Z3T	A	1	0
1H7E	A	1H7H	B	1	0
1GQI	B	1GQL	A	1	0
1GQN	A	1QFE	B	1	0
1RK2	B	1RKA	A	1	1
1OS7	D	1OTJ	B	1	1
1GS5	A	2X2W	A	1	0
6GSX	A	6GSY	B	1	0

Table B.2 continued from previous page

pdb_id_1	chain_id_1	pdb_id_2	chain_id_2	analysis _complete	in_unique _CATH_subset
2UYU	A	2V9N	B	1	1
1GTW	A	1GU5	B	1	0
3L1T	D	3TOR	D	1	0
1GU9	L	1KNC	B	1	1
1GUN	E	1GUT	C	1	0
1GUP	A	1GUQ	D	1	1
2IH9	B	3FU9	B	1	0
1GWY	B	1O72	B	1	1
1O17	A	1ZYK	C	1	1
1GXP	F	1GXQ	A	1	1
2CE8	D	2CE9	A	1	1
1GXZ	B	1GY0	A	1	1
1GY7	B	1GYB	A	1	0
1GYJ	A	1GYX	B	1	0
2W08	B	4AYU	B	1	0
1OUS	A	1OXC	A	1	1
1XT3	B	2BHI	B	1	1
2BKY	A	3WBM	D	1	1
1QMT	A	4OXF	A	1	0
1QH1	A	1QH8	C	1	0
1H1L	B	1QH8	D	1	0
1UBT	L	1WUI	L	1	1
1H2R	S	1WUI	S	1	1
3EQ9	A	4AN1	A	1	0
1H32	B	2OZ1	H	1	0
3EDJ	B	3EDK	A	1	0
1H4G	A	1QH7	B	1	0
2R9P	B	4U30	D	1	0
1H5R	C	1H5T	A	1	0
1HAN	A	1LGT	A	1	0
1S5X	A	4IRO	C	1	0
6HBI	B	7HBI	B	1	0
1HCJ	A	1W7T	A	1	0
1HCR	A	1JJ6	C	1	1
1HDO	A	1HE2	A	1	0
1HIA	X	2KAI	A	1	0
1HIA	Y	2KAI	B	1	0
1HSI	A	2MIP	C	1	0
1HIV	B	1HXB	B	1	0
1HIZ	A	1R85	A	1	0
1HJS	B	1HJU	D	1	0
1HJX	D	1NWT	D	1	0
3RER	B	4HT9	B	1	1
2HMQ	B	2HMZ	D	1	1
1FX9	A	2B01	A	1	0
2PXB	A	2PXE	A	1	1
2LIP	A	4LIP	E	1	0
1HQK	D	1NQV	D	1	1
1K5N	A	1UXW	A	1	0
1LTS	C	1LTT	C	1	1
1HUV	A	1P5B	A	1	0
1HV9	A	2OI5	B	1	0
1HVA	A	4JSW	A	1	0
1KQY	A	1KQZ	A	1	0
4H5N	B	4H5T	A	1	0
1HX1	B	3FZM	B	1	1
1PVF	B	2VNQ	A	1	1
1HZ5	B	1HZ6	A	1	1
2H5I	B	2H65	B	1	1
1I3R	B	1IEA	D	1	0
1YPI	B	2YPI	B	1	0
1I4U	A	1OBQ	B	1	0
2QL5	D	3IBF	B	1	0
1I6P	A	1T75	D	1	1
1I6W	B	1R50	A	1	0
3O18	A	4GY3	A	1	1
1I8F	F	1LNX	B	1	0
1I8M	A	1XF3	A	1	0
1SK0	A	1SK2	A	1	0
1IBY	D	1IC0	E	1	0
4STD	C	7STD	A	1	0
1RTE	B	1S61	A	1	0
1J98	A	1JQW	A	1	1
1IE4	C	1KGI	D	1	1
1S3T	B	3UBP	B	1	0
3UBP	C	4AC7	C	1	1
1IGQ	A	1IGU	B	1	1
1OUZ	A	2HT0	A	1	1
1IIM	A	1IIN	A	1	0
1J9J	A	1J9L	B	1	1
1IVD	B	1IVG	B	1	0
1INN	A	1VGX	B	1	0
1IQ8	B	1IT8	A	1	1
1IQZ	A	1IR0	A	1	1
1IRI	D	1JIQ	D	1	0
2BKB	A	2NYB	B	1	0
1ISI	A	1ISJ	B	1	0
1IT2	A	1IT3	C	1	0
1ITO	A	2DCC	A	1	0
1IUO	A	1UK9	A	1	0

Table B.2 continued from previous page

pdb_id_1	chain_id_1	pdb_id_2	chain_id_2	analysis _complete	in_unique _CATH_subset
1IV2	E	1IV3	B	1	1
1IVN	A	1U8U	A	1	0
5UPJ	B	6UPJ	A	1	0
1IW0	B	4GOH	A	1	1
2A6H	O	2BE5	E	1	1
1IYK	A	1NMT	B	1	1
3AA1	A	3AAA	A	1	0
1IZO	B	2ZQX	C	1	1
1J0M	A	1J0N	A	1	0
1WLV	A	1WM6	F	1	1
1O3Y	A	2J59	D	1	0
1J2T	D	1J2U	C	1	1
4NYO	E	4NYP	B	1	0
1J2W	A	1UB3	B	1	0
1J36	B	1J37	B	1	0
2Y2C	A	2Y2D	A	1	1
1J3J	D	1J3K	D	1	0
1J3P	B	1J3R	A	1	1
1J3U	A	3R6Q	C	1	0
1J4S	D	1J4T	G	1	1
1L8T	A	3TM0	A	1	1
1J8D	B	1K1E	L	1	0
1KU8	A	4R6R	C	1	0
2J8W	B	2J9B	A	1	0
1JLU	S	4IAD	S	1	0
1JC0	C	1JC1	B	1	0
1JC4	D	1JC5	A	1	0
1JCQ	B	2IEJ	B	1	1
4HT2	A	4KP5	D	1	0
1JD5	A	1Q4Q	E	1	1
1JE8	F	1ZG1	E	1	0
1JFV	A	2FXI	A	1	0
1JFA	A	1JFG	B	1	1
1GED	A	1ROM	A	1	0
4TXO	C	4TXV	C	1	0
1JFZ	A	1RC5	B	1	1
1JG1	A	1JG3	A	1	0
1JIF	B	1QTO	A	1	0
4H30	A	4H84	B	1	0
2NU8	E	2SCU	B	1	1
1JKU	C	1JKV	B	1	0
1K2X	A	1T3M	A	1	0
2FJB	A	2FJE	C	1	0
2FJA	D	2FJE	B	1	0
1DK7	A	1DKD	D	1	1
3LLY	A	3LM1	A	1	0
1TQ7	A	3BEI	A	1	1
1JPC	A	1MSA	D	1	1
1R3Q	A	3GVR	A	1	1
1JPL	A	1JUQ	A	1	1
1JPM	C	1TKK	H	1	1
1KEV	D	1PED	D	1	0
2AOT	B	2AOU	B	1	0
1JR8	B	1JRA	A	1	1
1JRK	C	1K26	A	1	0
3TF4	B	4TF4	A	1	0
3P1D	A	4NYW	A	1	1
1JTM	A	1JTN	B	1	0
1IQA	C	3QBQ	A	1	1
1JVB	A	1R37	B	1	0
1LGV	A	1LHZ	B	1	0
1K4D	C	2HVK	C	1	1
1NXY	A	1PZO	A	1	0
1PND	A	6PCY	A	1	0
1JXV	E	2HVD	A	1	0
1JYK	A	1JYL	B	1	0
1JYN	D	1JYV	C	1	0
1Y6R	A	1Z5P	A	1	0
1JZN	E	1MUQ	D	1	0
1KX9	A	1N8V	B	1	1
1HI3	A	2BEX	C	1	0
1K35	A	1P5D	X	1	1
1K3T	C	3DMT	B	1	0
2ATK	A	3OR7	A	1	0
1T7I	B	3EL5	B	1	0
1K75	A	1KAR	B	1	0
1K7L	F	2HFP	B	1	0
1K7T	A	1K7V	B	1	1
1K9X	B	1KA2	A	1	0
1K9Z	A	1KA0	A	1	1
2IBZ	A	3CX5	L	1	0
4DSY	C	4KKZ	A	1	1
1KBV	A	1KBW	E	1	0
1NIA	A	2BW4	A	1	0
3FXP	A	3SSB	B	1	1
1KEP	A	1KER	B	1	0
4PHJ	A	4PHM	B	1	0
1KG2	A	1MUY	A	1	1
1MR7	Y	3DHO	C	1	1
3AYA	B	3AYD	A	1	0

Table B.2 continued from previous page

pdb_id_1	chain_id_1	pdb_id_2	chain_id_2	analysis _complete	in_unique _CATH_subset
1ADE	B	1QF4	A	1	1
4J9Z	B	4QNH	B	1	0
2W7J	A	2W7L	A	1	0
1KWW	C	1KWX	B	1	0
2JHU	A	2JHV	D	1	1
1YEJ	H	1YEK	H	1	0
1KP9	B	1KPG	A	1	0
1KQ1	H	3QSU	E	1	0
1KQB	C	1KQD	B	1	0
2XKI	A	4AVE	A	1	0
1KSO	B	3NSO	B	1	0
1XBX	A	1XBZ	A	1	0
2D8O	A	2VU7	A	1	1
1K9V	F	3ZR4	F	1	0
1TL9	A	2R9F	A	1	0
1UXQ	A	1UXU	A	1	0
2IBZ	D	3CXH	D	1	0
1KYV	E	1KYY	A	1	0
1KYW	C	1KYZ	E	1	0
3AID	B	4F74	A	1	0
1KZU	H	2FKW	S	1	1
1L6L	2	2OU1	I	1	0
1L7H	A	2QWE	A	1	0
1NS0	A	1NS4	B	1	0
1L8J	A	1LQV	B	1	1
1T1N	A	3NG6	A	1	0
1LAU	E	1UDH	A	1	0
1LBV	A	1LBX	B	1	0
1BHF	A	1BHH	A	1	1
1LES	A	2LAL	C	1	0
1LOA	E	1LOC	C	1	0
2RBQ	A	2RBR	A	1	0
2LIS	A	2LYN	D	1	1
2A8F	B	2AIB	B	1	1
1LK5	A	1LK7	C	1	0
1LQB	A	3ZUN	J	1	1
3JQ5	A	3JQL	A	1	0
1LO2	Y	1LO3	H	1	0
1LO0	L	1LO3	X	1	0
1LOA	B	1LOF	B	1	0
1I81	C	1MGQ	A	1	0
1LP9	E	2JCC	L	1	0
1LP9	M	2JCC	F	1	0
1TRH	A	3RAR	A	1	0
1LQ9	A	1N5T	B	1	0
1LQP	A	1NNR	A	1	0
1LR5	A	1LRH	D	1	0
2HMT	A	2HMU	B	1	0
3B6L	A	4CJ2	A	1	0
2F9O	C	2F9P	D	1	0
1LU9	A	1LUA	C	1	0
4S0W	B	5LZM	A	1	0
3PVF	A	3PY2	A	1	0
1LZ4	A	208L	A	1	0
1ZBG	A	1ZPK	B	1	0
1K6K	A	1R6O	A	1	1
1M4T	D	1OU6	C	1	0
1M1P	E	1M1R	A	1	1
1M3D	H	1T60	S	1	1
1M3D	I	1T60	I	1	0
1M44	B	1M4G	A	1	0
1M56	G	3FYI	A	1	1
3DTU	D	3FYI	B	1	0
1M5F	B	1XHY	A	1	0
1MWC	A	1PMB	B	1	0
1M7E	B	1P3R	A	1	1
1THZ	B	2B1I	A	1	0
3MBA	A	4MBA	A	1	0
1M6W	A	1MP0	B	1	0
1RBY	B	4EW2	A	1	1
1MFI	A	2GDG	B	1	0
4MJY	A	4MKU	A	1	0
1MG2	I	1MG3	E	1	0
2GC4	P	2MTA	C	1	0
1IZ7	A	1K5P	A	1	0
1OAO	B	3I01	D	1	0
1OAO	D	3I01	P	1	0
1MKA	B	4KEH	B	1	0
1MNP	A	3M8M	A	1	0
1MO1	D	1MU4	A	1	1
1MO9	B	3Q6J	A	1	0
1TJO	B	1TKP	D	1	0
4AB0	B	4CQK	D	1	1
1XK4	E	4GGF	K	1	0
1YTT	B	2MSB	B	1	0
1MT1	H	1N13	F	1	1
1MT3	A	1MTZ	A	1	0
1N49	C	2FNT	B	1	0
4I36	A	4I7E	B	1	0
1XG3	D	1XG4	C	1	0

Table B.2 continued from previous page

pdb_id_1	chain_id_1	pdb_id_2	chain_id_2	analysis _complete	in_unique _CATH_subset
2W3B	B	4M6J	A	1	0
1Z59	A	1Z5C	B	1	0
1ZV2	A	2DWS	A	1	0
1AP5	B	1XIL	A	1	0
1PDK	A	3ME0	A	1	0
3KMJ	A	3KMT	C	1	0
1N5W	A	1N62	D	1	0
1N5W	E	1N62	B	1	0
1N5W	C	1N62	F	1	1
1N7Z	B	1N80	B	1	1
2Q8X	B	3MSG	A	1	0
1FZC	C	3E1I	F	1	0
1RMT	D	1RMY	A	1	0
1OBZ	B	1W9E	A	1	0
1GU7	B	4WAS	B	1	0
1NAQ	D	3X3U	C	1	1
1NB5	D	8PCH	A	1	0
1NCA	N	4DGR	A	1	0
1ND0	F	1RUM	H	1	0
1NDH	A	3W5H	A	1	1
1NDL	C	1NSQ	B	1	0
2WDQ	I	2WU2	A	1	0
2WDQ	J	2WU2	J	1	0
2WDQ	D	2WU2	L	1	1
1NF6	E	1NFV	L	1	0
1NHE	C	1NQI	C	1	0
1NFF	B	1NFQ	A	1	0
1NHW	B	2NQ8	A	1	0
1NHG	D	2NQ8	C	1	0
1NHK	R	2NCK	L	1	0
1EYU	A	1F0O	B	1	1
3EXH	F	3EXI	B	1	0
1NJF	B	1NJG	B	1	0
3BAS	B	3BAT	C	1	0
2AYS	A	4OQO	A	1	0
1NMO	F	1NMP	C	1	1
3MPS	D	3PWF	A	1	0
1PD5	K	1Q23	J	1	1
3RNK	A	3SBW	A	1	0
1NR2	B	1NR4	B	1	0
3BBB	E	3BBF	C	1	0
1NTE	A	1R6J	A	1	0
1Y21	A	2IMQ	X	1	1
1NW4	D	1Q1G	B	1	0
4JAD	A	4JAF	B	1	0
1NXM	B	2IXL	D	1	0
1NXS	A	2A9P	A	1	0
1ZK2	A	1ZK3	C	1	0
1A09	B	1A1E	A	1	0
1O81	B	1OC8	A	1	0
1O9O	B	1O9P	A	1	1
4APT	A	4AQP	D	1	0
1OAR	N	1OAU	O	1	0
1OBR	A	4DUK	A	1	0
1OCJ	A	1OCN	A	1	0
2OCC	N	3AG4	A	1	0
2OCC	B	3AG3	O	1	0
1OCR	P	2OCC	P	1	0
3AG1	Q	3WG7	Q	1	1
2OCC	R	3ABK	R	1	1
2OCC	S	3ASO	S	1	1
2OCC	W	3ABK	W	1	1
2OCC	X	3ABM	X	1	1
1OCR	Y	2OCC	Y	1	1
2OCC	Z	3AG4	Z	1	1
2CES	B	2J75	A	1	0
1QFO	C	2BVE	A	1	0
1ODB	F	2WCB	B	1	0
1OE8	B	2CA8	A	1	0
1BQP	A	1HKD	A	1	0
1OGE	B	1OGF	A	1	1
1OHV	D	1OHW	B	1	0
1OJW	B	1OK3	A	1	0
1OJX	E	1OK6	A	1	0
1OKB	B	4LYL	O	1	0
4WQ4	C	4YDU	C	1	0
1UU4	A	1UU5	A	1	0
3B87	B	3B88	A	1	1
1ORE	A	1ZN9	A	1	0
2AJ8	D	2AJC	C	1	0
1OUM	C	1OV6	A	1	0
1OU9	C	1OUL	A	1	1
1OV8	C	1QHQ	A	1	0
1OVO	D	3OVO	A	1	0
1OWN	A	1QNF	A	1	0
1OX7	A	1P6O	B	1	1
1OXB	B	1OXK	L	1	0
1OXN	E	1OXQ	B	1	0
1OXW	B	4PK9	A	1	1
1OYB	A	3TX9	A	1	0

Table B.2 continued from previous page

pdb_id_1	chain_id_1	pdb_id_2	chain_id_2	analysis_complete	in_unique_CATH_subset
261L	A	262L	B	1	0
1P51	D	1P78	A	1	0
2VJL	B	2VJM	A	1	1
1P7G	F	3EVK	C	1	0
1MDU	A	1YAG	G	1	0
2PAL	A	2PVB	A	1	0
1FDB	A	7FDR	A	1	0
1PCF	G	2C62	A	1	1
1PDV	A	1PDW	H	1	0
1R69	A	2OR1	L	1	0
1PFK	A	2PFK	C	1	1
1PH8	B	1PHJ	B	1	1
1PI5	B	2FFY	A	1	0
1PL7	C	1PL8	D	1	0
1PM7	B	2IXC	C	1	0
1PMM	C	3FZ7	D	1	0
4HZE	A	4I06	C	1	0
1PR9	A	1WNT	D	1	0
1TQU	A	3EJS	A	1	0
1PS7	C	1PTM	B	1	0
1EYW	A	3CAK	A	1	0
1ONC	A	2I5S	X	1	0
1PU5	A	2AG4	B	1	1
1PU6	B	1PU8	A	1	0
1PWX	C	1ZO8	M	1	0
3PAZ	A	4RH4	A	1	0
1JWF	A	1JWG	B	1	0
1PY3	A	3DH2	B	1	0
1PYT	B	3KGQ	A	1	0
4AOK	A	4CRY	A	1	0
1QW8	A	1QW9	B	1	0
1PZE	A	1PZH	D	1	0
1Q05	A	1Q07	B	1	1
1Q08	A	1Q0A	A	1	0
1Q0K	L	1Q0M	B	1	1
3R36	A	3R37	B	1	0
1Q5H	B	1Q5U	X	1	0
1Q72	L	1RIU	L	1	0
3LNX	F	3LNY	A	1	0
1QVO	D	1X7Q	A	1	0
1QB5	D	1QCB	D	1	1
1QBJ	C	3F21	B	1	0
1QCN	A	2HZY	B	1	1
1QF9	A	1UKE	A	1	0
1QIL	B	1TS3	A	1	0
1QIY	H	1QJ0	D	1	0
1YAS	A	3C6Z	A	1	0
1QN9	B	1VOK	A	1	1
1QNN	A	1UER	D	1	0
1QNP	A	1QNQ	A	1	0
1QPO	F	1QPR	E	1	0
1QQF	A	1QSJ	D	1	0
4CKV	X	4CL7	C	1	0
2C2Z	A	2Y1L	C	1	1
2NQH	A	2NQJ	B	1	1
1U6J	L	3IQE	F	1	1
3AH1	A	3AH4	B	1	1
1ZNE	A	2OZQ	A	1	0
1QZG	B	1QZH	B	1	0
1R1P	C	1R1Q	B	1	0
1R1T	A	1R23	A	1	0
2PL6	B	2PL7	A	1	1
1R2R	A	4OWG	B	1	0
1R45	D	1R4B	A	1	0
3G6U	A	3G8X	B	1	0
1R8H	F	2AYE	B	1	1
1UR0	A	2CCR	B	1	0
1ZON	A	1ZOP	A	1	1
2GJ8	D	2GJA	A	1	0
2E1C	A	2ZNZ	E	1	0
2CJ5	A	2CJ8	B	1	1
1NG1	A	2CNW	C	1	0
1RJD	B	1RJF	C	1	0
2CLV	H	2CLZ	A	1	0
3JUT	F	3K1X	B	1	0
1RPI	B	3OTS	A	1	0
1THW	A	4EK0	A	1	0
1RRE	F	3WU6	E	1	0
1RTP	1	1RWY	C	1	0
1IGB	A	2PRQ	A	1	0
1RTV	A	2IXH	B	1	0
1RU7	L	1RVZ	B	1	1
1RV0	M	1RVT	K	1	0
4RHN	A	5RHN	A	1	0
1S0Y	G	3EJ3	A	1	0
3EJ7	H	3EJ9	B	1	0
1S2O	A	1TJ5	A	1	0
1S3L	B	1S3M	B	1	0
1S4N	A	1S4P	B	1	0
1RTX	A	2HZ1	A	1	0

Table B.2 continued from previous page

pdb_id_1	chain_id_1	pdb_id_2	chain_id_2	analysis _complete	in_unique _CATH_subset
1VQ7	P	1VQ9	P	1	1
1VQ7	Q	1VQ9	Q	1	1
3H64	D	3H67	D	1	0
2GP7	D	2GPU	A	1	1
2XMM	B	2XMV	A	1	0
1SBQ	B	1U3G	A	1	1
1OYV	B	1SBC	A	1	0
2NMY	B	2QD7	B	1	0
1SGC	A	3SGA	E	1	0
1SI9	B	1TR0	L	1	0
1VLB	A	3L4P	A	1	0
1SIX	A	1SNF	A	1	0
2Z8Q	A	4DHV	B	1	0
1SJC	B	1SJD	D	1	0
1SHB	A	1SPR	D	1	0
1SLA	B	1SLB	D	1	0
2GFJ	A	2H6A	B	1	0
1SN2	C	1SN5	A	1	0
1SNZ	B	1SO0	C	1	0
1NTM	A	2A06	A	1	0
1SVL	A	1SVM	E	1	0
1SW4	A	1SW5	C	1	0
1SWX	A	3RWV	A	1	1
1SZ9	B	1SZA	C	1	0
1T0S	B	2RDB	B	1	0
1T0S	C	2RDB	C	1	1
1GTG	1	1GTJ	2	1	0
1T4A	B	1TWJ	B	1	1
1T6Q	C	1T6U	J	1	0
2FTM	B	8PTI	A	1	0
1T8P	B	2H4Z	B	1	0
3L0B	A	3L0C	B	1	0
3BAA	A	3BAB	A	1	1
1TAG	A	1TND	B	1	0
2NTC	A	3QK2	A	1	1
2O99	C	2O9A	A	1	1
1TDI	A	2VCV	C	1	0
3E3S	A	3VHF	A	1	0
3LEV	L	3MOA	L	1	0
1TO4	C	1TO5	D	1	0
4AV5	D	4BUQ	B	1	1
1TU7	B	1TU8	B	1	0
3F7I	A	3GTA	B	1	0
2F36	A	2QS3	A	1	0
1TKL	B	1TLB	W	1	1
1TXQ	A	2HQH	D	1	1
1WU9	B	2HKQ	A	1	0
1TY0	C	1TY2	B	1	0
1TZC	B	1X9I	A	1	0
1TZW	A	1TZX	A	1	1
1U0V	A	1U0W	B	1	0
1U17	B	1U18	A	1	0
1U1I	A	3QVS	A	1	0
3QUI	B	4J6X	F	1	0
1U1V	A	1U1W	B	1	1
1U20	A	2A8R	B	1	0
1FHW	A	1FHX	B	1	0
1U73	B	1Z76	B	1	0
1U79	C	1Y0O	C	1	1
1U7O	A	1U7P	A	1	0
1NLJ	A	3O1G	A	1	0
1UF4	B	1UF5	A	1	1
2APR	A	6APR	E	1	0
1V7C	C	3AEY	A	1	0
1UJM	B	1Y1P	A	1	0
3CPI	H	3CPJ	G	1	0
1ULD	B	1ULG	B	1	0
1UM0	D	1UMF	C	1	1
1AJ7	L	2RCS	L	1	0
1UM9	B	1UMC	D	1	0
2V15	G	2XKQ	L	1	0
1A4Y	E	1ANG	A	1	0
1USF	B	3ZOH	A	1	1
1USG	A	1USK	A	1	0
2JKL	F	2W5P	A	1	1
1UTJ	A	1UTM	A	1	0
2XIU	A	2XJ3	B	1	0
1W4V	E	1W89	F	1	0
1UW9	M	1UWA	T	1	0
1UWC	A	1UZA	B	1	0
1UXE	A	4K6T	B	1	1
1UYX	A	1UZ0	A	1	1
1UZZ	C	1V00	C	1	0
4RPG	A	4RPL	C	1	0
1W1X	A	2CML	C	1	0
1D0I	I	1GU0	G	1	1
1V2I	A	1V3E	B	1	0
1V4U	C	1V4X	C	1	0
1V4U	D	1V4X	D	1	0
1V5H	A	2DC3	B	1	0

Table B.2 continued from previous page

pdb_id_1	chain_id_1	pdb_id_2	chain_id_2	analysis _complete	in_unique _CATH_subset
1WVM	A	1Y9Z	B	1	0
1V4P	B	1WXO	B	1	1
2DVN	B	2DVO	A	1	1
1V8G	A	2ELC	C	1	0
3T4U	B	3T52	E	1	0
2HUQ	A	2OWK	B	1	1
1VCK	A	4NBF	F	1	0
1VCL	A	2Z49	A	1	0
2FV0	B	2FV1	B	1	0
1VE1	A	2EFY	B	1	0
1WPV	B	3BOY	B	1	1
1VEF	B	1WKH	A	1	0
1VF2	A	1VF4	A	1	0
1VFS	A	1VFT	B	1	1
4K49	D	4K4B	H	1	0
4K4C	A	4K4D	A	1	0
3WHB	B	3WHC	E	1	0
1VIE	A	2GQV	A	1	1
1YRI	A	3TRV	A	1	0
3VZY	B	3VZZ	A	1	1
1VJG	A	1Z8H	D	1	0
3A8I	B	3A8J	D	1	1
3M82	E	3M83	C	1	0
1VP2	B	3S86	A	1	0
1VPN	B	1VPS	C	1	1
1RZM	B	3PG9	F	1	0
1X31	A	3AD9	A	1	0
1X31	B	3AD9	B	1	0
1VRQ	C	1X31	C	1	1
1VRQ	D	3ADA	D	1	1
2VUB	H	3HPW	B	1	1
1FIN	D	1OKV	D	1	1
1VZG	B	1VZI	A	1	0
1W0D	D	2G4O	C	1	0
1W29	C	2C92	A	1	0
1W1D	A	1W1H	B	1	0
1W37	D	1W3T	B	1	0
1W3R	A	2VPA	A	1	0
1ATJ	F	1HCH	A	1	0
2QNP	B	3T11	B	1	0
1W8Y	B	2VGK	D	1	0
1KY6	A	1KYU	A	1	0
4NG9	L	4X8T	L	1	0
1XXO	B	2AQ6	B	1	0
1W9T	B	1W9W	A	1	0
1WBF	A	2D3S	C	1	0
1EUN	C	1FQ0	B	1	0
1WC0	A	1WC5	C	1	1
1WC7	H	1WCB	B	1	0
2CWG	A	2UVO	B	1	0
3VSC	A	3VSD	B	1	0
4TPW	A	4TQB	B	1	0
1WMA	A	3BHI	A	1	0
1WMR	A	1X0C	A	1	0
1WMY	B	1WMZ	C	1	0
1WN6	B	2Z3H	D	1	0
1WNB	D	1WND	C	1	0
1WOH	F	1WOI	C	1	0
1WOW	A	1WOX	B	1	0
1WP4	C	2CVZ	D	1	0
1WS7	D	1WS8	C	1	0
1WS9	A	2CX9	C	1	0
1WTJ	B	2CWF	A	1	1
2EF6	D	2P2K	A	1	0
3THE	B	4GWC	B	1	0
1WVQ	B	2GL0	A	1	1
1WW6	C	1WW7	C	1	0
1WW8	A	2DVM	D	1	0
1WW9	A	3VMG	B	1	0
1WWI	A	1WWS	A	1	0
1WX5	A	3AWZ	A	1	1
1WXD	B	2D5C	A	1	0
1WY5	B	2E89	D	1	0
1WYT	A	1WYU	G	1	0
1WYT	D	1WYU	H	1	0
1X3K	A	3A5B	A	1	0
1X3M	A	4FWP	A	1	1
1X6M	A	1XA8	C	1	1
2GC0	A	2GC1	A	1	0
3LW8	A	3LXR	A	1	0
3FAI	A	3IOG	A	1	0
1XA1	D	1XA7	B	1	0
1XA3	A	1XA4	A	1	0
4JYG	A	4JYI	B	1	0
1U56	A	4IT2	B	1	1
1XCC	A	3TB2	B	1	0
1BM3	L	1OPG	L	1	0
1XKY	B	1XL9	B	1	0
1XME	C	3S8G	C	1	0
1XNV	B	1XO6	C	1	1

Table B.2 continued from previous page

pdb_id_1	chain_id_1	pdb_id_2	chain_id_2	analysis _complete	in_unique _CATH_subset
1XO7	D	1XQ7	C	1	0
1T9S	A	1TBF	A	1	1
1XQ5	D	3BJ2	B	1	0
1UFQ	D	1UJ2	A	1	0
1XSJ	E	1XSK	D	1	0
2BKQ	D	2BKR	A	1	0
1XX5	B	2GIZ	A	1	1
1XW6	D	1XWK	C	1	0
1XWB	D	1XWC	A	1	0
1XX1	D	2F9R	A	1	1
2O3Z	A	3P76	A	1	1
1XXN	A	2DCY	E	1	0
1XXQ	A	1XXR	D	1	0
1XYG	A	2Q49	D	1	0
3PKB	A	3PKD	A	1	0
1Y2T	A	1Y2V	B	1	0
2OF8	B	2OF9	A	1	1
1Y5Y	D	1Y60	E	1	1
1Y7B	B	1YI7	C	1	0
4OIX	A	4OJ1	B	1	0
1YAR	M	1YAU	N	1	0
1YBT	D	1YBU	C	1	0
1YCE	i	2WGM	u	1	1
1YEP	D	4MA9	E	1	0
1YFU	A	1YFW	A	1	0
1YHK	A	1YHM	B	1	0
1YI8	B	1YID	A	1	0
1YIZ	B	2R5C	A	1	0
1YJ1	B	2FCM	A	1	0
3MI1	A	3MI5	F	1	0
1YKK	B	1YKL	L	1	0
1YLJ	A	3G32	B	1	0
1YMQ	A	2RBK	A	1	0
1YQ4	A	2H88	N	1	0
1YQ4	B	2H88	B	1	0
2H88	P	2H89	C	1	0
1YQ4	D	2H89	D	1	0
1YR6	A	1YRB	B	1	0
1YTA	B	2IGI	A	1	0
1YUM	D	1YUN	B	1	0
2NUF	B	2NUG	A	1	0
1Z02	C	1Z03	E	1	0
1Z0E	B	1Z0G	F	1	0
1Z0S	D	1Z0U	A	1	1
2FDW	D	2FDY	B	1	0
1Z2W	B	1Z2X	A	1	0
1Z5G	C	1Z5U	D	1	0
2JEW	A	2PIZ	C	1	0
2GRO	B	2GRP	B	1	1
1Z83	B	2C95	B	1	0
1Z8K	B	2Q4I	C	1	1
1Z9L	A	1Z9O	C	1	0
1ZAB	A	2FR6	B	1	0
2QUV	B	3LGE	C	1	0
3EY1	A	3TWH	A	1	0
1ZCB	A	3AB3	A	1	0
1BTL	A	3C7V	C	1	0
4ND7	A	4NDA	A	1	0
2VON	B	2VOO	B	1	0
3VFU	A	4JRX	A	1	0
1ZHQ	B	1ZHS	G	1	0
1ZI6	A	4P93	A	1	0
2PWG	B	2PWH	A	1	0
1Z77	A	2IEK	A	1	0
3F9W	A	3F9X	C	1	1
1ZMP	C	4E83	A	1	0
1U8F	R	4WNC	C	1	0
1ZPA	A	2O4S	B	1	0
2WVA	Z	2WVH	Z	1	0
1ZR3	B	3IIF	B	1	1
2A4T	A	2HUK	A	1	0
4JE6	A	4JE7	B	1	0
1ZZI	A	1ZZK	A	1	1
2F62	B	2F67	B	1	0
2A10	C	2A18	C	1	0
2A15	A	2Z7A	B	1	0
2A2M	A	2A2O	F	1	0
2A4O	A	2CXV	A	1	0
2A50	A	2A56	C	1	0
2A53	B	2A54	D	1	0
1G24	A	1GZF	B	1	0
2BZN	H	2C6Q	B	1	0
2A8Y	J	3T94	B	1	0
2A92	A	2AA3	D	1	0
2GH5	B	3SQP	B	1	0
1GHD	B	1GK1	B	1	0
3GWC	A	3HZG	D	1	0
2AGW	H	2OK4	D	1	1
2AJV	H	2AJX	H	1	0
2AJU	L	2AJY	L	1	0

Table B.2 continued from previous page

pdb_id_1	chain_id_1	pdb_id_2	chain_id_2	analysis _complete	in_unique _CATH_subset
1OX0	A	1OXH	C	1	0
2AOC	A	2NNP	B	1	0
2APW	A	2AQ1	G	1	0
2AQ1	D	3BZD	B	1	0
3BYT	E	3BZD	A	1	0
1S0O	B	3FDS	A	1	0
4IEO	A	4IEW	A	1	0
2AVK	A	3AGU	A	1	0
2AVS	A	3NU5	B	1	0
2AX3	A	3RU2	A	1	0
2AXY	C	2PQU	A	1	0
2HB4	A	4EJK	B	1	0
2AZC	A	2HB2	A	1	0
2B0J	A	3F47	A	1	0
4IPC	A	4NB3	A	1	0
2B3B	C	2B3F	E	1	0
4U6W	A	4U75	A	1	0
2B43	C	4LQ9	A	1	1
2B4H	B	2B4I	A	1	0
2B4R	Q	2B4T	P	1	0
2VWH	A	2VWP	A	1	1
2B6E	D	3LZ7	C	1	0
3LIT	B	3LIY	B	1	0
3FBB	D	3FBE	B	1	0
2B98	B	2B99	C	1	0
2B9Y	A	2BAC	A	1	0
2BB6	C	2BBC	A	1	0
3VXP	D	4F7T	D	1	0
2BDG	A	4KEL	A	1	0
2G09	A	2Q4T	A	1	0
2BES	A	2BET	E	1	1
2BGJ	A	2VNH	A	1	0
1KRM	A	1QXL	A	1	0
2BH0	A	4FER	A	1	1
2EHQ	A	2EJ6	B	1	0
2BHY	A	2BHZ	A	1	0
4AZJ	B	4AZK	A	1	0
2BIW	C	4OU9	A	1	0
2BJ3	A	2BJ8	A	1	1
2BL1	A	2J8M	B	1	0
2BL2	J	2DB4	H	1	1
2BLE	A	2BWG	A	1	0
2C9X	B	2CA4	B	1	0
2BM2	D	4MQA	B	1	0
2BM4	B	2BM5	A	1	0
2BMD	A	2BME	A	1	0
2BO4	F	2BO6	A	1	0
1VL9	A	2B96	A	1	0
1HTO	X	1HTQ	X	1	1
2BVJ	B	4B7D	A	1	0
2BW8	A	2BWC	B	1	0
3OXU	D	4ONT	F	1	0
3D9D	C	3FCJ	D	1	0
2C29	D	3C1T	C	1	0
2G0N	A	2OV2	H	1	0
2C31	B	2JI6	A	1	0
2C3G	A	2C3H	B	1	0
2C3Q	C	2C3T	C	1	0
2C3V	B	2C3W	B	1	0
1XW5	A	3GUR	D	1	0
2C4V	A	4B6S	B	1	0
2CI5	B	2CI6	A	1	1
2C78	A	4H9G	A	1	0
2C7M	A	2C7N	I	1	0
2C8B	X	2C8D	A	1	0
2C8E	E	2C8F	F	1	0
2C8G	D	2C8H	A	1	0
2YAV	F	2YAW	F	1	0
2CBJ	A	2V5C	B	1	1
2CCB	A	2CJC	A	1	1
2CDY	B	2CE4	B	1	0
2CE2	X	2EVW	X	1	0
2CGK	B	2UYT	A	1	0
1EEI	H	1FGB	D	1	0
2CJH	A	2JJG	A	1	0
2W39	A	2W52	A	1	0
3ZUI	A	3ZUO	A	1	0
2CMG	B	2CMH	A	1	0
2CNS	B	2CNT	D	1	0
2CNZ	A	2CO6	A	1	0
2CQS	B	3QFZ	A	1	0
2CSL	F	2CVL	F	1	0
2CU3	B	2HTM	F	1	0
3CQ1	A	3CQ2	B	1	1
2CUW	A	2DGB	B	1	0
2E2G	C	2NVL	B	1	0
2CV9	B	2Z06	B	1	0
2E1A	B	2Z4P	A	1	1
2CWK	A	2DYA	A	1	0
2CWX	A	2D69	B	1	0

Table B.2 continued from previous page

pdb_id_1	chain_id_1	pdb_id_2	chain_id_2	analysis _complete	in_unique _CATH_subset
2CX4	H	4GQF	A	1	0
2CX5	A	2Z0X	A	1	1
2Z3N	B	2Z3P	A	1	0
2CZD	B	2CZE	B	1	0
2CZ8	C	2DEG	C	1	0
2DKE	A	3I94	A	1	0
2R1M	A	2R1N	A	1	0
3MM1	A	3MM3	A	1	0
2D4M	A	3TPN	A	1	0
2EJF	D	2EVB	A	1	0
2D6K	B	2D6L	X	1	0
2DG5	C	2E0X	C	1	0
2DXB	G	3VYG	A	1	0
2DD4	H	3VYG	K	1	0
2DD5	L	2ZZD	F	1	0
2DDB	A	2EPF	D	1	0
2FW3	A	4EP9	A	1	0
2DEC	B	2DF8	A	1	0
3BPP	A	3VIV	A	1	0
2DG0	L	2DG1	D	1	1
2DHO	A	2I6K	B	1	0
2XWQ	A	2XWS	A	1	0
3W1A	B	3W7L	B	1	0
2V8M	B	4BFN	A	1	0
3IEK	D	3IEM	B	1	0
2DS5	A	2DS8	B	1	0
2DTE	B	2DTX	A	1	0
2DU9	A	2EK5	B	1	0
2DUP	B	2DUQ	A	1	0
2DVT	A	2DVX	C	1	0
2DWZ	D	3AJI	D	1	1
2ZOA	A	3D03	D	1	0
2I6O	A	2I6P	A	1	1
2E2O	A	2E2Q	B	1	0
2E4G	A	2OAL	B	1	0
2E4P	A	2E4Q	A	1	0
3QUG	B	3VUA	F	1	0
2OO1	C	3S92	A	1	0
2EFN	A	2PN6	A	1	0
2E8Z	A	2E9B	B	1	0
4IG5	A	4QYQ	A	1	0
3GF2	A	3GFI	A	1	0
2ECU	B	2ED4	A	1	0
2EGZ	A	2YSW	C	1	0
2EK1	H	2EK6	B	1	0
2EKY	E	2EPI	B	1	1
2EQ8	A	2EQ9	K	1	0
3N7H	A	4FQT	A	1	0
2EV2	B	2EV4	A	1	0
2EW6	A	4E9B	A	1	1
2EZT	B	2EZU	A	1	0
2F0R	B	4EJE	B	1	1
2F0X	A	3F5O	E	1	0
2IB7	D	2IB9	A	1	0
3SA6	B	4DJO	A	1	0
2F3N	A	2F44	B	1	0
2F59	C	2I0F	A	1	0
2B4D	B	2B5G	A	1	0
2F5Z	O	2F60	K	1	1
2F78	A	2F7A	B	1	0
2F6N	B	3QZV	A	1	0
2F6L	A	2FP1	B	1	0
1H68	A	3QDC	A	1	1
2F98	C	2F99	C	1	0
3TBG	A	4WNU	D	1	0
2FAO	A	2FAQ	B	1	0
2FBV	A	2FC0	A	1	0
2FG4	A	2FG8	F	1	0
3KRD	Z	3MI0	L	1	0
3N0Z	B	3N10	A	1	1
2FK3	F	2FMA	A	1	1
2FK7	A	2FK8	A	1	0
2FLK	A	2FMK	A	1	0
1LW5	D	1M6S	A	1	0
2NNR	A	3E1Z	A	1	0
2FPD	D	2FPE	D	1	0
2FPW	A	2FPX	B	1	0
2FQ6	A	2GQN	B	1	0
2FW6	A	2FW7	B	1	1
1C52	A	1R0Q	A	1	0
3K98	A	3K99	A	1	1
2FYT	A	4HSG	A	1	0
2FZ6	D	2GVM	B	1	0
2G6X	C	2G6Y	B	1	0
4QEU	A	4QEW	A	1	0
4HRQ	B	4LD2	A	1	0
2GBM	A	2GBN	A	1	0
2GC7	M	3RLM	D	1	0
2H1S	D	2Q50	C	1	0
2BOK	L	2VVC	K	1	1

Table B.2 continued from previous page

pdb_id_1	chain_id_1	pdb_id_2	chain_id_2	analysis _complete	in_unique _CATH_subset
2GDQ	A	2GGE	B	1	0
2GFV	A	3G11	A	1	0
2GG4	A	2GG6	A	1	0
2GGU	B	2ORJ	C	1	0
4DX5	B	4DX7	C	1	1
3D04	C	3DOZ	F	1	0
2GMW	B	3L1V	A	1	0
2GN0	C	2GN2	A	1	0
2GQ9	A	2GQA	A	1	0
2GSE	A	2VM8	B	1	0
2OTU	G	2OTW	A	1	0
2GSG	B	2OTW	D	1	0
2GSU	A	2RH6	B	1	0
3BEX	E	3BF3	B	1	0
3VX0	A	6TAA	A	1	0
2GWC	D	2GWD	A	1	0
3LS3	A	3LSA	D	1	0
1URP	B	2DRI	A	1	0
2H1X	A	2H6U	E	1	0
2H61	B	3D10	A	1	0
2H6B	A	3E5X	D	1	0
2H77	A	2H79	A	1	0
2QTG	A	3LGS	B	1	0
3CZ1	A	3FE8	A	1	0
4ERA	A	4ERI	B	1	0
2I57	B	3IUD	C	1	0
2HDV	B	2HDX	C	1	0
2HE8	A	2HEJ	A	1	1
1RLM	B	1RLO	C	1	0
2HK0	C	2HK1	A	1	0
2HK7	A	2HK8	F	1	1
1GT3	B	1HN2	A	1	0
2IDH	H	2OEI	A	1	0
2HP6	A	2HPB	B	1	0
2HPO	A	4Q4I	A	1	0
2OTB	B	2OTE	A	1	0
2HQS	E	2W8B	G	1	1
3ARN	C	3EHW	Z	1	0
2HS1	A	3NU4	A	1	0
2HT5	A	2HTQ	A	1	0
2HTA	A	2HTB	D	1	0
2HTY	D	2HU4	H	1	0
2HUI	B	2HUU	A	1	0
3LWP	C	3LWV	C	1	1
2HY5	A	2HYB	J	1	1
2HY5	B	2HYB	E	1	0
2HY5	C	2HYB	I	1	0
2J0S	C	3EX7	A	1	1
2I26	N	2I27	O	1	0
2I2S	B	3TAY	A	1	0
2I3G	B	2NQT	A	1	0
2I4D	B	2I4W	A	1	0
2I4L	C	2I4O	A	1	0
1R56	C	4CW6	A	1	1
3ZNL	E	4BGW	A	1	0
3CX5	R	3CXH	R	1	1
1EZV	G	3CXH	S	1	1
2IC7	B	2P2O	A	1	0
2IDE	H	3JQM	B	1	1
2IFC	C	2R9E	C	1	0
2G2N	C	2G2P	B	1	0
2IHO	A	3EF2	D	1	0
2IIU	B	2OLT	C	1	0
2IJ0	B	2QIL	C	1	0
2IL3	B	2IMK	A	1	0
2INU	C	2INV	C	1	0
2GBP	A	2QW1	A	1	0
2IPU	L	2IQ9	L	1	0
2IRU	A	4MKY	D	1	0
2IS8	C	3MCH	B	1	1
2ISJ	G	2ISK	F	1	0
2ISY	B	2ISZ	B	1	0
3QZO	D	3QZP	B	1	0
1O86	A	4APH	A	1	0
2IUY	B	2IV3	C	1	0
2VC0	A	2W24	B	1	0
2VLT	B	2VLV	B	1	0
2IJX	A	2NTI	C	1	1
4WJN	A	4WJQ	C	1	0
2J12	B	2J1K	Z	1	0
2J1K	I	2J2J	E	1	0
2J1R	A	2J1U	A	1	0
2J5G	J	2J5S	A	1	0
2J5Z	B	2J64	A	1	0
2J71	A	2J73	B	1	0
2J7V	C	2J8Y	D	1	0
4U95	E	4U96	D	1	0
2J9D	L	2J9E	C	1	0
1U5B	B	1V1R	B	1	0
1PRC	C	2WJM	C	1	1

Table B.2 continued from previous page

pdb_id_1	chain_id_1	pdb_id_2	chain_id_2	analysis _complete	in_unique _CATH_subset
2JBZ	A	2JCA	C	1	1
2JCQ	A	2JCR	A	1	0
2JCX	A	2JCY	B	1	0
2VOT	A	2VQT	B	1	0
2JG8	A	2WNV	A	1	0
2JG8	B	2WNV	E	1	0
2JG9	C	2WNV	F	1	0
2JH7	A	3F53	A	1	0
2JHH	F	2JHL	F	1	0
2JIO	A	2V3V	A	1	1
2JJT	B	2UV3	A	1	0
2JJT	D	2VSC	D	1	0
2JLQ	A	2JLY	B	1	0
3TCK	A	3TCN	B	1	1
1PEY	C	1SRR	B	1	0
2V88	A	2V89	A	1	0
3N7D	A	3N7E	B	1	0
2W80	H	2W81	F	1	0
4J4O	A	4MGV	A	1	0
4L2M	A	4MAX	B	1	0
1UWP	X	2QJ7	A	1	0
1G7B	B	2G54	D	1	0
2LBD	A	4LBD	A	1	0
3QL9	A	3QLC	B	1	0
3IFW	A	4JKJ	A	1	1
4J5R	A	4J5S	B	1	0
3M0R	A	3M0U	A	1	0
3UCT	A	3UCW	D	1	0
4RKG	A	4RKH	D	1	0
3B5G	B	3BDX	C	1	0
4A1I	H	4A1K	A	1	1
2NQO	B	2QM6	D	1	0
3QM8	A	3QMA	A	1	0
2NST	A	2NT9	B	1	0
2NTB	B	2NTP	A	1	0
2NUW	B	2NUY	A	1	0
2NXD	B	4OBH	D	1	0
2NYH	A	2P8I	D	1	1
3L6J	A	3TF8	B	1	0
2O26	A	2O27	A	1	1
3RHM	A	3RHO	D	1	0
2O4Q	B	3CS2	A	1	0
2O7E	F	2O7F	C	1	0
2O70	C	2O74	A	1	1
3EDX	D	3HKI	B	1	0
2OE0	A	2OE3	A	1	0
2OEK	A	2OEM	A	1	0
2OFC	B	2OFD	A	1	0
2OG1	B	2PU5	B	1	0
2OGA	A	2OGE	C	1	0
2OGX	B	4NDQ	B	1	0
2OHH	B	2OHJ	A	1	0
2OKD	A	2OKE	C	1	0
3BXE	B	3BXF	A	1	0
2OKW	E	2OKY	B	1	0
2OLJ	B	2OUK	A	1	0
3PI1	B	3PI3	A	1	0
2OMU	B	2OMZ	B	1	1
3SN0	A	3SN4	A	1	0
2GP3	B	2GP5	A	1	0
3ES8	G	3GD6	A	1	0
2ORM	E	3M21	B	1	0
3RF4	C	3RF5	B	1	0
3TS5	C	3TUY	F	1	0
2OU7	A	3FC2	A	1	0
2OV7	A	2VPL	A	1	0
3OVW	B	4OVW	B	1	0
1CWY	A	1ESW	A	1	0
2OWQ	B	4LZB	C	1	0
2OXG	Y	2OXH	F	1	0
2OYC	A	2P27	A	1	0
2ATK	B	4UUJ	B	1	0
2P1N	E	2P1P	B	1	0
2P3N	C	2P3V	D	1	0
2P3Z	A	3BOX	A	1	0
2P5Q	A	2P5R	B	1	0
2P5U	A	2P5Y	A	1	0
4DE1	A	4DE3	B	1	0
2P7H	D	2P7I	B	1	0
2P88	H	2P8C	A	1	0
1J8T	A	1MMT	A	1	1
4M80	A	4M82	A	1	0
2PBP	A	2QQ3	L	1	0
2QQ1	C	3MCJ	C	1	0
2PD3	A	2PD4	C	1	0
2PGN	B	2PGO	A	1	0
3NST	A	3NW4	A	1	0
2PI2	H	2PQA	B	1	0
2PKF	B	4UBE	A	1	0
2POS	C	2PR0	B	1	0

Table B.2 continued from previous page

pdb_id_1	chain_id_1	pdb_id_2	chain_id_2	analysis _complete	in_unique _CATH_subset
2PP0	C	2PP1	A	1	0
2PQL	A	2QEO	B	1	0
1U24	B	1U25	B	1	0
2PUX	A	3HK3	A	1	0
2PV1	A	2PV2	D	1	0
2IWM	D	3MJI	C	1	1
2OXY	A	3FL5	A	1	0
4J4D	C	4J4G	A	1	0
2PZL	B	2PZM	B	1	0
2Q0L	A	3ISH	B	1	0
2Q0Q	F	2Q0S	C	1	0
2Q28	A	2Q29	B	1	0
2Q5W	D	2QIE	G	1	0
2Q73	C	2Q9L	C	1	0
1VQ7	M	1VQN	M	1	1
1VQ9	Z	1YHQ	Z	1	1
2QAC	A	4AOM	A	1	0
2QAP	B	2QDH	D	1	0
2VKL	A	2W19	D	1	1
2QCU	A	2R4E	B	1	0
2Q08	L	3HK9	E	1	1
2QFV	C	2QFY	B	1	0
2QHE	A	3BJW	G	1	0
2QKT	A	2QKU	B	1	0
3DE8	D	3DE9	A	1	0
2QMC	C	3FNM	A	1	0
1MT1	K	1N13	E	1	0
3UH1	A	3UHA	B	1	0
2QTN	A	3E27	B	1	0
2QTQ	C	2RHA	A	1	1
2QTX	B	4X9C	D	1	0
2QUA	A	2QUB	I	1	0
2QWO	A	2QWP	A	1	0
2QYS	B	2R6J	A	1	0
3COP	A	3CX4	A	1	0
2R0F	A	2R0H	C	1	0
2R1Q	A	2RB3	B	1	1
3JXB	C	3JXD	R	1	0
3C2X	A	4OUG	B	1	0
2R3W	B	3BC4	A	1	0
1HS6	A	4L2L	A	1	0
2R5N	A	2R8O	A	1	0
2R5Q	D	2R8N	A	1	0
3PUX	G	4KI0	G	1	1
2R73	C	2R74	B	1	0
2R7A	D	2RG7	D	1	0
2R80	B	3DHR	F	1	0
2R8E	D	2R8Y	G	1	0
2R91	D	2R94	A	1	0
2RC9	A	4KCD	B	1	0
2RCH	B	3DSI	A	1	0
4QYN	B	4QZU	D	1	0
2RF1	B	2RIN	B	1	0
2RFJ	A	4FLP	A	1	0
2RFY	B	2RG0	A	1	0
3AHV	A	3F5L	B	1	0
2RH5	C	4IKE	B	1	0
2RHH	A	2RHO	A	1	0
2RIF	D	2RIH	A	1	1
2RJ6	A	2RJ7	A	1	0
2RJG	D	2RJH	C	1	0
1RBF	A	3OQY	B	1	0
2DRY	B	2DS0	A	1	0
1IE7	A	4CEX	A	1	0
2WH8	B	2WHF	A	1	0
2UXH	B	2UXI	A	1	0
2UXQ	D	4AOV	A	1	0
2UY3	A	2UY5	A	1	0
2UYK	B	2UYN	C	1	0
2V0I	A	4KQL	A	1	0
2V17	L	3L1O	L	1	0
3AWR	A	3AX5	C	1	1
2V32	C	2V41	D	1	0
3AF7	X	3F36	A	1	0
2V2Z	B	2VF3	B	1	0
3A75	D	3WHQ	B	1	0
2F61	B	2NT1	A	1	0
2V57	B	2WGB	A	1	0
2V78	A	2VAR	C	1	0
4OE7	C	4PTN	B	1	0
2V9S	D	2V9T	B	1	1
4KET	C	4KF1	D	1	0
2GHH	X	2GHK	X	1	0
1ZZ0	B	1ZZ1	D	1	1
2VCT	G	2WJU	H	1	0
2VE8	A	2VE9	F	1	0
3QZX	A	3ZJI	A	1	0
2WUD	A	2WUG	B	1	0
2VFR	A	2VFS	A	1	0
2VG2	D	2VG3	A	1	1

Table B.2 continued from previous page

pdb_id_1	chain_id_1	pdb_id_2	chain_id_2	analysis _complete	in_unique _CATH_subset
2BFY	C	2VRX	D	1	0
2VIH	B	2VJV	A	1	1
2VJI	A	2X85	A	1	0
2VJY	C	2VK4	B	1	0
2VKJ	B	2VKO	D	1	0
1OC3	C	2VL3	C	1	0
1EMV	A	2VLQ	A	1	0
2VM1	D	2VM2	C	1	0
2VM9	A	2VME	C	1	0
2ZB4	A	2ZB8	A	1	0
2VNO	A	2VNR	A	1	0
4JNI	U	4JNL	U	1	0
2WR9	C	2WRA	A	1	0
2VOG	A	2VOH	A	1	1
4O1T	A	4O35	A	1	0
2VS7	D	4D6N	A	1	1
3H71	B	3H72	A	1	0
2VWS	A	2VWT	C	1	0
3K83	A	3OJ8	B	1	0
3FEA	A	3U15	D	1	1
2VZC	A	2VZD	A	1	1
2VZE	B	3GPC	B	1	1
2VZS	A	2X05	B	1	0
2VZP	A	2VZR	B	1	0
2W1U	D	2WDB	A	1	0
2W3E	A	2W3F	B	1	0
3CVB	B	3CVD	A	1	0
2W94	C	2W95	A	1	0
3O2B	C	3O2H	A	1	1
2W9S	F	2W9T	B	1	0
2WC6	A	2WCK	A	1	0
2WDZ	C	3LQF	D	1	0
2WEE	A	2YES	A	1	0
2WHS	C	2WHT	A	1	0
2XQT	D	2XQU	C	1	0
2WIQ	B	2WIS	A	1	0
1PRC	L	3G7F	L	1	1
1PRC	M	1VRN	M	1	0
3B3X	B	4A5R	A	1	0
4A6C	A	4CPR	B	1	0
2WLC	A	2WLD	C	1	0
3IPO	A	3IPP	B	1	0
4IGM	E	4IH3	F	1	0
2WOX	B	4CBB	E	1	0
2WNX	A	2WOB	C	1	1
2WO1	B	2WO2	A	1	0
2WOC	B	2WOD	A	1	1
2WPC	C	2WPF	A	1	0
2WWM	T	4C4K	T	1	0
3U2Y	K	3U32	M	1	0
2WPI	E	2WPM	E	1	0
2EIE	A	2VZ1	A	1	0
2WQI	D	2WTT	N	1	0
2WS1	B	2WS6	J	1	0
2WSV	A	2WT0	A	1	0
4G3J	A	4G7G	A	1	0
2WVX	C	2WW3	C	1	0
2CEX	B	2CEY	A	1	0
4WKT	A	4WKV	A	1	0
2WYT	A	2WYZ	F	1	0
1ZRU	B	2BSD	B	1	1
2X2E	D	3ZYC	A	1	0
2X2H	C	2X2J	C	1	0
2X2S	D	2X2T	A	1	0
2X2V	G	4CBK	F	1	0
2X35	A	2X4Y	O	1	1
2X6Q	A	2X6R	A	1	0
2XCB	A	2XCC	B	1	1
2XCD	A	4APZ	k	1	0
2XCI	B	2XCU	D	1	0
1BMG	A	4F7E	B	1	0
1CWH	A	2X2D	B	1	0
1YQM	A	1YQR	A	1	0
1BYQ	A	1YER	A	1	0
2XI2	B	2XI3	B	1	0
2XI5	D	2XI7	B	1	0
3FVR	D	3FVT	M	1	0
2XM3	C	2XQC	A	1	0
2XN0	B	2XN1	D	1	0
2XPE	B	3ZRE	A	1	0
2XRX	P	2YFJ	F	1	0
2XV6	D	2XXC	B	1	0
2XUW	A	2YAF	A	1	0
2XVA	B	2XVM	A	1	0
2XVI	A	2XVJ	C	1	0
3H6W	B	3KGC	B	1	0
2XXB	A	2XXJ	B	1	0
1S50	A	4H8I	B	1	0
2XXW	B	2XXX	C	1	0
2BG5	D	2XZ9	A	1	0

Table B.2 continued from previous page

pdb_id_1	chain_id_1	pdb_id_2	chain_id_2	analysis _complete	in_unique _CATH_subset
2CNN	A	3EDQ	A	1	0
3B3F	C	3B3G	B	1	0
2Y2H	A	2Y2P	A	1	0
2Y3D	B	2Y3G	C	1	0
2Y42	C	4F7I	A	1	0
2Y5N	A	2Y5Z	C	1	0
2Y4X	B	2Y4Y	D	1	0
2OCC	U	3AG3	H	1	1
2OCC	V	3WG7	I	1	1
2Y7K	C	2Y7P	A	1	0
2YA4	B	2YA7	C	1	0
2YAZ	D	2YB0	E	1	1
2YB7	B	2YFZ	A	1	0
2YBV	L	3ZXW	H	1	0
2V4I	G	2VZK	G	1	1
2YH2	B	3ZWQ	B	1	0
2YIH	A	2YJQ	B	1	0
2YIL	D	2YIO	A	1	0
1SU7	A	1SUF	A	1	0
4A1U	A	4TUH	E	1	0
2YJJ	E	2YJK	G	1	0
2YJW	A	2YK2	A	1	0
2YKU	B	2YKV	C	1	0
2YLN	A	3ZSF	H	1	0
4CE6	A	4CNQ	B	1	0
4J77	B	4J86	A	1	0
2YOO	B	4UAX	A	1	0
2YOR	B	2YP1	D	1	0
3ZIY	A	4AX3	C	1	0
3AP5	A	3AP7	A	1	0
2YZ5	A	2Z4G	B	1	0
2EAK	B	2EAL	B	1	0
3JSW	B	4E90	A	1	0
4PVO	B	4UA4	B	1	0
2YZ7	C	3W8D	A	1	0
2YZC	F	2YZD	A	1	0
2YZJ	C	2ZDC	A	1	0
2Z0Z	A	2ZXV	B	1	0
3EI6	B	3EI7	A	1	0
2Z2N	A	2Z2O	C	1	0
2Z8E	B	2Z8F	A	1	0
2Z8S	A	2ZUX	B	1	0
2Z9U	A	2Z9X	B	1	0
4KYH	A	4OPN	B	1	0
2ZBU	D	2ZBV	A	1	1
2ZDP	A	3LGM	B	1	0
2ZFZ	E	3CAG	F	1	1
2ZGM	B	3AFK	B	1	0
1FT1	B	1QBQ	B	1	0
2ZIZ	B	3DHY	A	1	0
2WKO	F	3GZO	H	1	0
2CL5	B	4P7J	A	1	0
2ZN9	B	3WXA	A	1	0
3GM6	A	4Q1O	B	1	0
2ZVP	X	2ZYT	X	1	0
2ZRQ	A	3A3N	A	1	0
3B06	A	3VKJ	A	1	0
4E9V	C	4E9Y	B	1	0
2ZX0	A	2ZX2	B	1	0
2ZXY	A	3X15	J	1	0
2ZYK	B	2ZYO	A	1	0
3IA4	D	3IA5	A	1	0
4DEN	A	4G1R	A	1	0
3A0C	D	3A0E	A	1	0
3A15	D	3A16	C	1	0
3A22	B	3A23	A	1	0
3A45	A	3A46	B	1	1
3AXH	A	3AXI	A	1	0
3A50	A	3A51	D	1	0
2E52	A	3WVK	B	1	0
2VRR	A	3UIP	A	1	0
3AFA	G	3AZG	C	1	0
3QBL	D	3VVK	B	1	0
3A8I	E	3AB9	A	1	0
3A9R	A	3A9S	B	1	1
3ABR	C	3ABS	A	1	0
3ACZ	C	3AEP	D	1	0
3A2M	B	3A2N	A	1	0
1OMW	G	3PVW	G	1	1
3AHN	A	3AHO	B	1	0
3AHU	B	3HSB	B	1	0
3AI0	A	3VIO	A	1	0
3AI1	B	3AI2	H	1	0
3AIK	C	3AIL	A	1	0
3AK1	A	3AK3	B	1	0
3AK4	C	3WDS	D	1	0
3AK8	K	3AK9	C	1	0
3ALT	D	3ALU	A	1	0
3AMD	B	3MMU	D	1	0
3AMN	A	3AMQ	D	1	0

Table B.2 continued from previous page

pdb_id_1	chain_id_1	pdb_id_2	chain_id_2	analysis _complete	in_unique _CATH_subset
2DFX	E	2DJH	A	1	0
3ATH	C	3AV7	C	1	0
3APU	A	3APX	A	1	0
1X0J	B	2DVS	A	1	0
3ARP	A	3ARZ	A	1	0
3ASG	A	3ASH	A	1	0
3ASQ	B	3ASS	A	1	1
3ASY	B	3ASZ	B	1	0
3ATQ	A	4OPD	A	1	0
3AW1	B	4TWW	B	1	0
3AUS	A	3AUU	B	1	0
4BFW	A	4BFZ	B	1	0
3AYT	C	3AYV	B	1	0
3AYX	A	3AYY	C	1	0
3AYX	B	3AYY	D	1	0
3AZZ	B	3B01	A	1	0
3B08	H	3B0A	E	1	0
3VLZ	A	3VM1	A	1	1
3B1C	C	3B1E	B	1	0
3B1J	B	3B20	A	1	0
3B1V	A	3B1W	A	1	0
3VG2	A	3VG3	A	1	0
3B4O	B	3EX9	A	1	0
3B7Q	B	3B7Z	A	1	1
3D20	A	3NUJ	B	1	0
3BBE	B	3BBH	A	1	0
3BD3	A	3BD5	A	1	0
3BE5	D	3BE6	C	1	0
3BFC	A	3BFG	C	1	0
3BGT	D	3BH3	A	1	1
2A89	B	3QSE	A	1	0
3BJ1	C	3BJ3	A	1	0
1HT6	A	1P6W	A	1	0
3BVI	D	3BVK	F	1	0
3BWQ	C	3BWR	A	1	0
3BZ4	G	3GGW	C	1	0
3C1Z	A	3C23	A	1	0
3GHB	M	4M1D	L	1	0
3C9R	A	3C9S	B	1	1
3CA0	A	3CA5	A	1	0
3CB0	C	4IRA	A	1	0
3CFF	H	3CFH	G	1	0
3CFA	M	3CFF	R	1	0
3CGC	B	3CGE	B	1	0
3CGM	A	4ODN	A	1	0
3CL6	A	3CL8	B	1	1
4I6R	A	4IWR	E	1	0
3BUV	A	3CAV	A	1	0
1Z98	M	4JC6	A	1	1
1JNX	X	4IGK	B	1	1
3M6P	B	3PN3	B	1	0
3CQ5	B	3CQ6	C	1	0
3CQD	A	3UQD	C	1	0
3EEL	B	3QLZ	B	1	0
3CSR	A	3CT5	A	1	0
3CT7	C	3CTL	E	1	0
1FHD	A	2XYL	A	1	0
3I3H	A	4K06	B	1	0
3CYP	C	3IMP	H	1	0
3CYT	O	5CYT	R	1	0
4KNN	A	4QSJ	B	1	0
3D1X	A	3NU6	B	1	0
3D2Q	B	3D2S	C	1	0
3W7S	A	3W7U	A	1	0
4BZW	B	4C01	F	1	0
3D46	H	3D47	C	1	0
3C4P	A	3MKE	A	1	0
3D4I	C	3D6A	B	1	0
3GQP	C	3GQR	E	1	0
3GQP	B	3GQR	F	1	0
3D53	C	3EMJ	A	1	1
3F98	C	3F9C	B	1	0
3D63	A	3GVF	A	1	0
3D77	A	3D78	A	1	0
3D8N	A	3D8T	A	1	1
3D9K	A	3D9L	B	1	0
3DAB	G	3FDO	A	1	0
4QUT	A	4TT2	A	1	0
1NDD	B	1XT9	B	1	0
3DCC	A	3OKV	A	1	0
3BHT	D	3TNW	B	1	0
3DFO	D	3DFP	C	1	0
3DFS	C	3DFT	D	1	0
3DGF	C	3GL9	D	1	0
3DHG	A	3Q3O	A	1	0
3DHG	B	3DHH	B	1	0
3DHG	C	3I5J	C	1	0
3GE3	E	3I63	E	1	0
3DKR	A	3DYI	A	1	0
3LSW	A	4F39	A	1	0

Table B.2 continued from previous page

pdb_id_1	chain_id_1	pdb_id_2	chain_id_2	analysis _complete	in_unique _CATH_subset
3DLZ	A	4QTC	A	1	0
3DMF	A	3DMH	A	1	0
3G19	A	3GQ0	B	1	0
3DPZ	A	3DQ6	A	1	0
3DS0	A	3DS5	D	1	0
3DUU	B	3DV4	B	1	0
3DUS	C	3DUU	A	1	0
2ZNV	B	3A9K	A	1	0
3DVH	C	3DVP	B	1	1
3DWA	E	3DWP	D	1	0
3DXD	A	3DXE	A	1	0
3DYO	B	3DYP	D	1	0
1SKR	B	2TRX	B	1	0
3E0N	B	3E0P	B	1	0
3HSS	A	3HYS	B	1	0
3E6I	A	3T3Z	D	1	0
1R28	B	1R2B	A	1	1
3E5P	B	3E6E	C	1	0
3E5R	O	3V1Y	B	1	0
3E6S	F	4IXK	F	1	0
3E8M	D	4HGO	C	1	0
3E86	B	3E8G	A	1	0
3EBS	B	3T3Q	A	1	0
3EER	A	3I07	B	1	1
3EG3	A	3EGU	A	1	0
4NCZ	C	4R57	H	1	0
3EII	B	3EJA	C	1	0
3EIX	A	3LI2	B	1	0
3EJB	A	3EJE	G	1	0
3EJX	F	3EKM	D	1	0
3EKP	A	3EL0	A	1	0
3ELX	A	3EM0	A	1	0
3ENV	A	3ENW	B	1	0
3ETR	L	3NVW	A	1	0
3EUE	A	3EUF	A	1	0
3EUO	B	3EUT	A	1	0
3EVA	A	3EVF	A	1	0
4NVB	A	4NVK	B	1	0
3EZ8	A	3H3K	A	1	0
3EZN	A	3GP5	A	1	0
3F1O	A	3H82	A	1	0
3F1O	B	3H82	B	1	0
3F2G	B	3F2H	A	1	0
3F4F	C	3HHQ	M	1	0
3F51	B	3F52	E	1	0
3F6R	D	3F90	E	1	0
3F7T	A	3SZO	B	1	0
1D2V	B	4DL1	A	1	0
1D5L	C	4DL1	L	1	1
3H17	A	3H1B	A	1	0
3FAS	B	3FAT	C	1	0
3NIF	A	3ZDY	C	1	1
1BRL	D	1BSL	A	1	1
3FMI	C	3FPA	B	1	0
3FJG	D	3FJM	A	1	0
3FJK	D	3HOM	B	1	0
3L25	E	3L26	A	1	0
4KDM	D	4KDN	B	1	0
3H3N	X	3H3O	B	1	0
3FMA	C	3K3V	A	1	0
3FS8	A	3FSC	B	1	0
3FT1	A	3FT9	A	1	0
3FV3	B	3TNE	B	1	0
3FV6	B	3FWS	A	1	0
3FVL	E	3FX6	C	1	0
3KIK	D	4FK5	B	1	0
3FZI	A	3G2C	A	1	0
3G0J	B	3MB4	B	1	0
4NRC	A	4RVR	A	1	0
3G28	A	3G29	B	1	1
3G17	F	4YCA	A	1	1
3G5K	C	3G5P	B	1	0
3GAC	C	3GAD	E	1	0
3BGS	A	3K8O	S	1	0
3GC6	B	3GHH	A	1	1
3GDF	D	3GDG	A	1	0
3GGG	D	3GGP	C	1	0
3C2A	H	4M1D	I	1	0
3GJ1	C	3GJ2	D	1	0
3GJ4	C	3GJ7	C	1	0
4H2N	A	4H2R	B	1	0
3GUI	A	3GUP	B	1	0
3GVM	A	3GWK	E	1	0
3GWZ	B	3GXO	D	1	0
3GXG	B	3GXH	B	1	0
3IGU	A	4DO6	B	1	0
4CGP	A	4CYO	A	1	0
3H7J	A	3H7Y	B	1	0
3HVU	D	3KB8	A	1	0
1E9I	D	2FYM	C	1	0

Table B.2 continued from previous page

pdb_id_1	chain_id_1	pdb_id_2	chain_id_2	analysis _complete	in_unique _CATH_subset
3H8F	F	3H8G	B	1	0
3H8O	A	3H8X	A	1	1
3H9Y	B	3H9Z	D	1	0
3HBR	A	4S2K	D	1	0
3HDS	C	3HF5	B	1	0
4DYC	A	4DYR	B	1	0
3HF3	A	3HGJ	D	1	0
3HFR	B	3ISV	A	1	1
3GXP	B	3S5Z	A	1	0
3HGI	A	3HJ8	A	1	0
3HIS	A	3HIV	B	1	0
3PD6	D	3PDB	D	1	0
3HLW	A	3HVF	B	1	0
3HNK	B	3HNL	A	1	0
3B9S	C	3DJI	D	1	0
3HPV	A	3HPY	C	1	0
3HRY	C	3K33	C	1	0
3HSQ	A	3I3X	B	1	0
3HTF	A	3HU9	A	1	0
3HTA	B	3HTI	A	1	0
1FYK	A	2HTS	A	1	0
1E8I	A	3CCK	A	1	0
2OMF	A	3POX	D	1	1
3HXV	A	3HXY	A	1	0
3HY7	B	3HY9	A	1	0
3KGG	A	3O4P	A	1	0
2JAI	B	2JAJ	A	1	0
3I2K	A	3PUH	B	1	0
3I3B	D	3I3E	C	1	0
3I69	A	3IK9	D	1	0
3I6C	B	3IK8	A	1	0
3I7U	D	3I7V	B	1	0
3I8S	C	3I8X	B	1	0
1BZ1	A	1Z8U	D	1	0
3ITC	A	3K5X	A	1	0
3IE5	A	4N3E	F	1	1
3IEB	A	3JR2	C	1	0
3IG9	D	3IGE	A	1	0
4E5O	A	4EIN	B	1	0
3II4	A	4M52	C	1	0
3IJL	A	3IJQ	B	1	0
3INV	B	3IRM	C	1	0
4GFM	A	4GMY	A	1	0
3IP7	A	3IP9	A	1	0
3IR2	B	3V4J	B	1	0
3IRS	C	3K4W	H	1	0
2ZD1	A	4WE1	A	1	0
2PD7	B	2PDR	B	1	0
3IS3	A	3QWF	C	1	0
4TOG	C	4TOH	B	1	0
3IT0	B	3IT2	B	1	0
3IT5	B	3IT7	A	1	0
4TSV	A	5TSW	F	1	0
3LDN	B	3LDP	A	1	0
1EWM	A	3I06	A	1	0
3IVB	A	3IVV	A	1	0
3IVZ	A	3KLC	A	1	0
3IWC	A	3IWD	C	1	0
3IWB	D	3IWC	B	1	0
4BLC	C	8CAT	A	1	0
3JQW	C	3JQX	B	1	1
3JS3	C	4H3D	B	1	0
3P2X	B	4CY7	B	1	0
3UW9	A	4MR3	A	1	0
3A8Y	A	3AY9	A	1	0
3K1Y	C	3K20	A	1	0
4J6G	C	4MSV	B	1	0
2F7W	A	2F7Y	B	1	0
3K6U	A	3K6W	A	1	0
3NH3	X	3NN6	X	1	0
3K7O	A	3K7S	C	1	0
3K87	B	3K88	A	1	0
4DMC	B	4DNO	C	1	0
3KE4	A	3KE5	C	1	1
3KE8	B	3KE9	A	1	0
3RUS	C	3RUV	D	1	0
3TV1	A	3TW3	A	1	1
2FDG	A	3I3Q	A	1	0
4IXH	C	4QJ1	A	1	0
3KNE	A	3M2Z	A	1	0
3KJE	A	3KJH	A	1	0
3KJJ	L	3KJK	K	1	0
1FM6	V	4DM8	D	1	0
3KOS	A	4WKM	D	1	0
3KOX	B	3KP1	D	1	0
3KOX	H	3KP1	F	1	0
1H81	C	1H83	B	1	0
3KQL	B	3KQN	A	1	0
3KRF	B	3KRP	C	1	0
3KS5	A	3KS6	B	1	0

Table B.2 continued from previous page

pdb_id_1	chain_id_1	pdb_id_2	chain_id_2	analysis _complete	in_unique _CATH_subset
3KSF	H	3KSI	A	1	0
3KUT	B	3PKN	A	1	0
3KU4	A	3KVY	A	1	0
3L4I	A	3L6Q	A	1	0
3KWM	C	4IO1	B	1	0
3KZM	A	3KZN	A	1	1
3KZE	C	4GVD	A	1	0
3M7W	A	4GUJ	B	1	0
4FAV	B	4K3I	B	1	0
3RLM	C	4L1Q	C	1	0
3L4T	A	3L4X	A	1	0
3L5M	A	3N19	D	1	0
3TJ8	B	3TJA	B	1	0
3LBB	B	3LBE	C	1	0
2WO5	A	2YGY	C	1	0
3RFF	A	3UC5	A	1	0
3LCM	C	4F8Y	A	1	0
3LCZ	D	3LD0	c	1	0
3LIG	A	3LIH	A	1	0
3LEG	A	3LEI	A	1	0
3LE5	B	3LFG	C	1	0
3LIS	A	4JCX	A	1	0
3LL4	B	3OI7	D	1	0
2G3R	A	2IG0	A	1	0
3LJF	C	4L2D	B	1	0
2GUD	B	2HYR	A	1	0
3LOC	A	4JYK	A	1	0
1CDH	A	1CDJ	A	1	0
3LQH	A	3LQI	C	1	0
3LUG	B	3LUK	A	1	0
3LWN	G	3LYQ	B	1	0
3LXH	B	4C9L	A	1	0
3LXT	D	3M0F	A	1	0
3LXZ	D	3PR8	B	1	0
3LZO	B	3LZP	A	1	0
1XEV	B	2FOV	A	1	0
3F0V	X	3FQF	A	1	0
3M1X	A	3M4S	F	1	0
3M27	A	3M29	A	1	0
3M3K	C	4F1Y	A	1	0
3M4D	G	3M4E	C	1	1
3M5M	B	3SUD	D	1	0
3M7B	A	3M7E	A	1	0
3M7H	A	4GC2	B	1	0
3M8O	L	3QNX	A	1	0
3M9W	A	3MA0	A	1	0
3MBH	A	3MBJ	A	1	0
3LRT	B	3NAG	A	1	0
4RDV	A	4RZB	A	1	0
3MEU	A	3MEV	B	1	0
3MHF	D	3MHG	C	1	0
1HVH	B	1QBU	B	1	0
3MJZ	L	3MLC	C	1	0
3MKC	A	3NZG	C	1	0
3QXX	A	3QY0	A	1	0
3MLS	O	3MLT	A	1	0
3MM5	A	3MMC	D	1	0
3MM5	B	3MM7	E	1	0
3MN1	A	3NRJ	I	1	0
3MNV	A	3MO1	A	1	0
3MNV	D	3MNW	B	1	0
3MPI	C	3MPJ	B	1	0
3MQG	D	3MQH	E	1	0
3MRE	A	3MRN	A	1	0
3N31	A	4JTB	A	1	0
3MSR	A	3OVG	A	1	0
1YG6	B	2FZS	B	1	0
3MTU	D	3MUD	D	1	0
3MV0	4	3MV1	3	1	0
3MWK	A	3NBF	B	1	0
4AAB	A	4AAF	B	1	0
2RJV	A	2RJW	B	1	1
3N2X	C	3NEV	B	1	0
3N39	C	3N3A	C	1	0
3MS5	A	4CWD	A	1	0
4KFD	B	4KFF	B	1	0
3N9Z	B	3NA1	B	1	0
3NBK	D	3PNB	B	1	0
3NBC	B	3NBE	B	1	0
3NBW	A	3NC2	B	1	0
3NUG	A	3RWB	C	1	0
3NDY	C	3NDZ	A	1	0
3NDY	G	3NDZ	F	1	1
3NFK	A	3NFL	D	1	0
4OLU	L	4S1Q	L	1	0
3NGS	B	3NGU	B	1	0
3NIP	F	3NIQ	A	1	0
2FD7	A	2FD9	A	1	1
3NIL	D	3NIM	B	1	0
3NMP	B	3NMV	A	1	0

Table B.2 continued from previous page

pdb_id_1	chain_id_1	pdb_id_2	chain_id_2	analysis _complete	in_unique _CATH_subset
3NJK	A	3NJL	A	1	0
3NOF	B	3O6T	C	1	0
3NRC	A	4J3F	F	1	0
3NRP	B	3NRQ	B	1	0
3NSS	B	3TI3	B	1	0
3OZR	A	4PYQ	B	1	0
3NYC	A	3SM8	A	1	0
3NU8	A	3NUB	B	1	0
4IWW	A	4IX0	A	1	0
3O28	A	4U2R	C	1	0
1GOM	A	1GOO	A	1	0
4H3H	B	4H3K	E	1	0
3O2S	A	4H3H	A	1	0
3O3N	C	3O3O	A	1	0
3O3N	D	3O3O	B	1	0
1ITB	A	2NVH	A	1	0
4DRH	D	4DRI	A	1	0
3O5O	A	4TW6	A	1	0
4FGK	A	4FGL	B	1	0
3O9L	C	3OAG	A	1	0
3O9L	D	3OAD	B	1	0
3O9Z	B	3OA0	A	1	0
1OM3	M	3OAU	L	1	0
3OD9	A	4PJ2	B	1	0
3OFJ	A	3OFK	C	1	0
1PVS	B	3CWA	B	1	0
3OHU	A	3OHV	B	1	0
3OKM	A	3OKN	A	1	0
3OL9	M	4K4S	E	1	0
3OND	B	3ONE	A	1	0
2EKE	B	2GJD	A	1	0
3OTY	A	4FAE	A	1	0
3ORU	A	3SIY	A	1	0
3OT4	A	3UAO	B	1	1
4PDL	A	4PDR	B	1	0
3OUO	C	3OV9	B	1	0
3OVP	A	3OVR	B	1	0
3OWO	A	3OX4	D	1	1
3OWR	C	3P69	A	1	0
3OXU	A	4ONT	B	1	0
3OXW	D	3OXX	D	1	0
3OZF	A	3OZG	D	1	0
3P12	C	3P13	D	1	0
3KUW	A	3KVU	C	1	0
1H50	A	3KFT	A	1	0
2NNQ	A	3FR5	A	1	0
3P7F	A	3P7G	C	1	0
4LJC	D	4LJD	A	1	0
3A78	A	4ITE	A	1	0
3P9C	A	3P9I	B	1	0
3PB6	X	3PB9	X	1	0
4C4X	A	4C4Y	A	1	0
3PF7	A	3PM5	C	1	0
3PFB	A	3S2Z	A	1	0
4L89	A	4M3S	A	1	0
3Q7G	A	3QMK	B	1	0
3PN7	A	3TS5	D	1	0
3PN7	E	3TUY	B	1	0
3PNL	B	4LRZ	B	1	0
3POB	A	3POE	A	1	1
3POP	C	3PQB	D	1	0
1HBM	E	1HBU	B	1	1
1HBU	F	3M2V	C	1	1
3PQA	D	3RHD	A	1	0
3PR3	A	3QKI	C	1	0
3PRL	B	3RHH	C	1	0
3PHW	C	3PHX	A	1	0
3PTK	B	3PTM	A	1	0
3PTR	B	3PUS	B	1	0
4EFQ	B	4EFR	D	1	0
3PUL	B	3TDF	B	1	0
3PUN	A	3PVD	B	1	0
3PVY	C	3PW1	B	1	0
3PWM	A	3VF7	A	1	0
3PZG	A	3PZN	A	1	0
3Q18	A	3QAG	A	1	0
1ANF	A	1JW5	A	1	0
3Q2O	B	3R5H	A	1	0
3Q3X	A	3RUO	A	1	0
3SOC	A	4ASX	A	1	0
3Q7V	A	3Q82	B	1	0
3Q8U	E	3Q8V	H	1	0
3ZR0	A	3ZR1	B	1	0
3Q9B	L	3Q9F	I	1	0
4AD6	B	4CWB	A	1	1
3QDT	B	3QDW	B	1	0
3QFM	A	3QFN	B	1	0
4FOG	D	4FOX	B	1	0
3QMB	A	3QMH	A	1	0
3LWE	B	3SVM	A	1	0

Table B.2 continued from previous page

pdb_id_1	chain_id_1	pdb_id_2	chain_id_2	analysis _complete	in_unique _CATH_subset
3QPF	B	3QSP	A	1	0
3QU9	A	3QUT	A	1	0
3QWA	B	3QWB	A	1	0
3QXE	E	3QZ9	A	1	0
3QXE	F	3QZ5	D	1	0
3QXT	B	3QXU	D	1	0
3QYN	A	3US0	B	1	0
3R2M	A	3R2O	A	1	0
4J5B	A	4J5D	X	1	0
3R50	C	3R51	B	1	0
3R5R	D	3R5W	D	1	0
3R7D	A	3R7F	A	1	0
3R95	B	3R9E	A	1	0
3RE5	A	3RE6	A	1	0
3RFT	B	3RFV	A	1	0
3RG8	A	3RGG	A	1	0
3R7B	D	3R7S	B	1	0
4MUU	A	4POV	A	1	0
1ZMD	G	2F5Z	B	1	0
1N0T	B	1NNK	A	1	0
3RNZ	A	3RO0	C	1	1
3ROJ	B	3RPL	C	1	0
3RPN	B	3RPP	B	1	0
1F12	A	1F17	B	1	0
3RSH	A	4I08	A	1	0
3RSY	B	3S4B	B	1	0
3RVL	B	3RVM	A	1	0
2FZD	A	2PFH	A	1	0
4YGL	A	4YGN	A	1	0
3S0A	A	3S0F	A	1	0
3S0C	D	3S1U	B	1	0
3S1I	A	3S1J	C	1	0
3S1L	B	3S2G	H	1	0
3S3L	B	3T8E	B	1	0
3S53	A	3S54	A	1	0
3S4T	C	4QRO	G	1	0
3S6X	C	3S6Z	B	1	0
1L6X	A	2IWG	D	1	0
3S7N	A	3S7O	A	1	0
3S7Z	A	3S81	D	1	0
3U92	B	3U94	A	1	0
3SAL	A	4QN5	B	1	0
3T08	C	3T09	D	1	0
3SEW	A	3SEY	C	1	0
2W9G	A	2W9H	A	1	0
3SH7	A	3SH9	B	1	0
3ASL	A	3SOW	A	1	0
3SHT	C	3SHV	B	1	0
3SK1	D	3SK2	A	1	0
3SLP	A	4WUZ	B	1	1
1GG2	B	1TBG	D	1	0
3SPK	A	4GYE	A	1	0
3NAW	B	3NB2	D	1	0
3SS0	A	3SSH	A	1	0
3SSV	A	3SSY	A	1	0
3SV5	A	3SVB	A	1	0
3ST3	B	3ST4	C	1	0
3STT	B	3STV	A	1	0
3SUE	D	3SV7	A	1	0
3SVR	B	3SVS	B	1	0
3T2Z	B	3T31	A	1	0
3SY3	B	3TJ7	A	1	0
4L1O	A	4L2O	E	1	0
3SZI	G	3SZJ	A	1	0
3SPD	B	3SPL	C	1	0
3T0A	D	3T0D	C	1	0
1A4M	C	3MVI	B	1	0
3T2O	C	3T2Q	D	1	0
3OYW	A	3W58	D	1	0
3T4L	A	3T4T	B	1	0
3T9A	A	4Q4C	A	1	0
3T6N	A	3V78	A	1	0
3T6V	C	3T71	A	1	0
3T93	D	3T9V	B	1	0
4PTV	A	4PTW	A	1	0
3TCF	C	3TCH	A	1	0
3TD4	F	3TD5	F	1	0
3TDJ	B	4IY5	B	1	0
3TEE	A	3VKI	D	1	0
3TFH	A	3TFJ	B	1	0
3TKG	A	3TL9	B	1	0
1AYA	B	1AYD	A	1	0
3TLC	B	4IIY	B	1	1
3TM8	B	3TMB	A	1	0
3TNL	C	3TOZ	E	1	0
3TNN	B	4H8W	L	1	0
3TOL	C	3TOM	C	1	0
3TOY	D	3TTE	A	1	0
3TSU	A	3TTC	A	1	0
3TTS	F	3TTY	F	1	0

Table B.2 continued from previous page

pdb_id_1	chain_id_1	pdb_id_2	chain_id_2	analysis _complete	in_unique _CATH_subset
3G75	B	3G7B	A	1	0
3TUU	G	4HNN	D	1	0
3TWR	B	3TWT	A	1	1
3TZY	B	3TZZ	A	1	0
3U13	A	4E88	A	1	0
3U3K	A	3U3M	A	1	0
3U54	B	4O7N	A	1	0
3U8A	C	3U8C	D	1	0
3UB6	B	3UB7	A	1	0
3UCB	B	3UHL	A	1	0
3UCK	B	3UCO	A	1	0
3UD1	A	3UD2	A	1	0
3UDO	A	3UDU	B	1	0
3UEC	A	3UEF	C	1	0
3UEJ	A	3UGL	A	1	0
1S46	A	1ZS2	A	1	0
3UG3	F	3UG5	D	1	0
3UIE	B	4FXP	A	1	0
3UJ9	A	4FGZ	B	1	0
1TE6	B	2AKZ	B	1	0
3UJO	A	3UJQ	D	1	0
3UTE	D	3UTG	B	1	0
4A75	G	4ALP	A	1	0
3UNM	B	3UNN	A	1	1
2RFX	A	3VH8	A	1	0
4FN2	B	4GIV	B	1	0
3USC	L	3USE	M	1	0
1XQK	B	1XQL	A	1	0
3UWU	B	3UWY	A	1	0
3UXK	D	4M6U	B	1	0
4F5T	A	4J2V	A	1	1
3V1A	A	3V1D	B	1	1
3VB1	A	4F7F	D	1	0
4DLY	A	4DLZ	A	1	0
3V4N	A	3V4X	A	1	0
3U9J	B	3U9M	E	1	0
4LGZ	B	4LH0	A	1	0
3VAJ	A	3VAL	B	1	0
3K73	Q	3L6O	O	1	0
3VBI	E	3VBJ	C	1	0
2XWC	A	4A63	A	1	0
3VD9	C	4DUX	D	1	0
3VDA	B	3VDB	A	1	0
3VDJ	A	3VDL	C	1	0
3VEN	A	3VET	A	1	0
3VEP	A	3VFZ	B	1	0
3VKB	A	3VKC	B	1	0
3VKZ	A	3VL7	A	1	0
1YPQ	A	1YPU	B	1	0
1Y3P	A	1Y3Q	A	1	0
3VMY	D	3VMZ	C	1	0
4G0P	A	4G0X	A	1	0
3VNG	A	3VNH	A	1	1
3VNI	B	3VNL	A	1	0
3VOY	A	3VP1	A	1	0
3VPG	D	3VPH	A	1	0
3VPI	A	3VPJ	B	1	0
3VSG	A	3VSI	C	1	0
3VSG	D	3VSI	D	1	1
3VST	D	3VSV	B	1	0
1OGX	A	3VGN	A	1	0
3VTX	A	3VTY	A	1	0
3VUX	E	3VUY	D	1	0
3VV5	B	3VVF	B	1	0
3VVG	B	4DM1	A	1	0
3VVX	A	3VVY	C	1	0
3VW9	B	3W0U	A	1	0
4TSN	C	4TSO	A	1	0
4IH9	B	4IHA	A	1	0
3VYH	B	4OB3	B	1	0
4IT9	A	4ITB	B	1	0
3VZJ	B	3VZM	A	1	0
3VZL	C	3VZO	A	1	0
3VZP	B	3VZS	D	1	0
2QD3	A	3HCO	A	1	0
3W25	A	3W26	A	1	0
3W41	A	3W43	A	1	0
1SU4	A	2AGV	A	1	0
3W6E	A	3W6F	C	1	0
3W79	A	3W7A	D	1	0
3ELL	A	3MOK	A	1	1
3W8W	A	3W8Z	B	1	0
1LBT	B	1TCC	A	1	0
3WB1	C	3WB2	D	1	0
3WBB	C	3WBF	C	1	0
3WBQ	A	3WBR	B	1	0
3WCA	B	3WSB	A	1	0
3WD1	A	3WD2	A	1	0
3WDK	A	3WDM	D	1	0
3WDW	A	3WDX	A	1	0

Table B.2 continued from previous page

pdb_id_1	chain_id_1	pdb_id_2	chain_id_2	analysis _complete	in_unique _CATH_subset
3WE7	A	3WL3	C	1	0
3WGY	B	3WGZ	A	1	0
4IKO	A	4JX9	A	1	0
3WHO	C	3WR2	D	1	0
3WID	A	3WIE	D	1	0
3WJB	B	3WJC	A	1	0
1ZD3	A	3I28	A	1	0
1IEX	A	3WLP	A	1	1
4LNM	B	4RJY	C	1	0
3WMP	E	3WMQ	A	1	0
3WNN	B	3WNO	A	1	0
3WOI	B	3WOK	A	1	0
3WQD	B	3WQF	A	1	0
1VGO	B	2Z55	E	1	0
3WRR	A	4PP4	A	1	0
1MI7	R	1WRP	R	1	1
3WRZ	B	3WS1	C	1	0
3WSE	D	3WSF	E	1	0
3WU2	A	4UB8	a	1	0
4UB6	Y	4UB8	y	1	0
1HMT	A	2HMB	A	1	0
3WWB	A	3WWF	A	1	0
3WXK	A	3WXL	D	1	0
3WZQ	C	3X00	D	1	0
3X0X	H	3X0Y	A	1	0
4XTL	A	4XTN	D	1	0
3ZBK	A	3ZBN	D	1	0
3ZSQ	A	3ZT4	A	1	0
3ZCX	A	3ZDF	A	1	0
3ZE8	A	3ZE9	A	1	0
3ZFN	A	3ZFR	A	1	0
3ZHG	B	4CAJ	C	1	0
2GDN	A	4Q8I	A	1	0
3ZLQ	D	4BEL	D	1	0
3ZLK	A	4B2X	D	1	0
3ZNI	G	4V3K	A	1	0
3ZNJ	N	3ZNU	C	1	0
3ZNT	A	4WM9	A	1	0
3ZQM	I	3ZQO	J	1	0
4UFH	A	4UFL	A	1	0
3ZRK	X	3ZRM	Y	1	0
3ZRP	B	3ZRR	B	1	0
3ZST	B	3ZT7	B	1	0
3ZTP	C	3ZTS	J	1	0
3ZTW	A	3ZWD	B	1	0
3ZV8	A	3ZVA	A	1	0
3ZW0	C	3ZW2	C	1	0
3ZW8	A	3ZWA	B	1	0
3ZWM	F	3ZWV	D	1	0
3ZWU	B	4A9V	A	1	0
4GAL	B	5GAL	A	1	0
1OKO	C	2WYF	D	1	1
2ZTX	A	2ZTY	A	1	0
3ZZF	B	3ZZH	C	1	0
4AJ8	B	4V2N	A	1	0
43C9	E	43CA	A	1	0
4A02	A	4ALT	A	1	1
4A0T	B	4A0U	A	1	0
4A1T	B	4A1X	B	1	0
4A38	B	4A39	A	1	0
4A57	A	4A59	D	1	0
2GG5	A	2GGC	A	1	0
4A83	A	4MNS	A	1	0
4A9E	B	4UYF	C	1	0
4ADB	D	4ADD	C	1	0
2Q6W	A	4IS6	A	1	0
4AFG	D	4AFH	E	1	1
4AFQ	C	4AFU	D	1	0
4AHA	A	4D52	C	1	0
2XJQ	A	2XJR	A	1	0
4AJE	D	4AJK	A	1	0
1BS5	C	1BS8	A	1	0
4AM4	B	4AM5	B	1	0
1S83	A	1Z7K	A	1	0
2BAZ	B	2XX6	A	1	0
2WOR	A	3PSR	B	1	0
4AS2	D	4AS3	C	1	0
4AWJ	D	4W9F	J	1	0
4AY3	A	4B4K	E	1	0
4AYD	A	4AYI	E	1	0
4B04	A	4HRF	C	1	0
1ACB	I	2TEC	I	1	1
3GU8	A	3GUB	A	1	0
2NL9	B	3FDL	B	1	0
4B4U	A	4B4V	B	1	0
4B5T	A	4B5U	A	1	0
3O8C	A	3O8D	B	1	0
4B7F	C	4B7G	B	1	0
2BFX	A	2VGP	B	1	0
4B98	B	4BQ0	A	1	0

Table B.2 continued from previous page

pdb_id_1	chain_id_1	pdb_id_2	chain_id_2	analysis _complete	in_unique _CATH_subset
4BAZ	A	4BB0	A	1	0
4B9L	A	4B9M	A	1	0
4BBE	A	4BBF	D	1	0
4BC3	C	4BC5	A	1	0
4BFF	F	4BGL	D	1	0
4BLL	A	4BLN	A	1	0
4BLU	B	4BLV	A	1	0
4BM3	A	4BM4	A	1	0
4BQE	B	4BQF	A	1	0
4BT2	A	4BT5	A	1	0
4BV7	A	4BVC	A	1	0
4C3X	B	4C3Y	E	1	0
1NU4	B	1URN	B	1	0
4C5A	B	4C5C	B	1	0
2WGF	E	2WGG	A	1	0
4C8V	H	4C9A	B	1	0
4CAV	A	4CAX	A	1	0
4CC2	C	4CC7	K	1	0
1K3B	B	2DJF	B	1	0
1KBO	D	3JSX	H	1	0
2WJ4	D	2WJ6	B	1	0
4CK4	A	4NLJ	B	1	0
4CN5	A	4CN7	A	1	0
4CQW	C	4CQY	A	1	0
4CR7	I	4CR8	G	1	0
4CSH	B	4CT3	C	1	0
2QPJ	A	2YB9	A	1	0
4R8R	B	4R8S	A	1	0
4CZ6	D	4CZ7	B	1	0
4CZN	A	4CZP	A	1	0
4CYV	F	4CYZ	B	1	0
4D1I	A	4D1J	D	1	0
4D1L	F	4D1M	D	1	0
4D3I	A	4D3O	A	1	0
4D4B	A	4D4C	B	1	0
4D5R	A	4D5T	H	1	0
1L0L	D	2A06	D	1	0
1PPJ	S	2FYU	F	1	0
1PPJ	G	2FYU	G	1	0
4D8Y	D	4D98	B	1	0
3N4T	A	3N4U	A	1	0
4DI6	C	4DZ6	E	1	0
4CCJ	C	4CCK	A	1	0
4DLM	A	4DO7	B	1	0
4DO1	D	4EGO	B	1	0
4DP0	X	4DP4	X	1	0
4DPL	B	4DPM	D	1	0
4DQR	D	4E0D	A	1	0
4E44	D	4E45	I	1	0
4E44	C	4E45	M	1	0
4DS1	C	4HT6	A	1	0
1ZMH	A	1ZMI	B	1	0
1PX3	B	1PX4	C	1	0
4DW4	B	4DW5	A	1	0
1UC2	B	4DWQ	A	1	0
4DXN	A	4GOB	D	1	0
4E0R	D	4G42	A	1	0
4E0R	B	4G42	E	1	0
4E2B	A	4E2D	B	1	0
4E2K	A	4EE1	A	1	0
4K76	B	4NMR	A	1	0
4E3A	B	4JKU	B	1	0
4GY0	A	4GY1	B	1	0
4E5E	C	4E5J	B	1	0
4E97	B	4I7L	A	1	0
4EAC	A	4EAY	C	1	0
3JYH	C	3N0T	D	1	0
3UBE	D	3UBJ	L	1	0
4FDZ	C	4TME	B	1	0
4EDL	B	4EDM	A	1	0
4EE6	B	4EE8	A	1	0
4EGY	A	4EGZ	A	1	0
4EHY	A	4ITN	A	1	0
4L73	A	4L74	A	1	0
3WAE	B	3WAF	A	1	0
4KL7	D	4KPM	B	1	0
4EQ0	A	4EQJ	A	1	0
1KPC	A	1KPF	A	1	0
4ES8	A	4ES9	D	1	0
4EU5	B	4FAC	A	1	0
4EXL	A	4LAT	A	1	0
4F03	A	4G19	B	1	0
4F1W	A	4F3C	B	1	0
4F2O	A	4F31	B	1	0
1LVO	C	1P9U	B	1	0
4F9A	C	4F9C	A	1	0
4FAY	B	4I61	C	1	0
4FB9	C	4FBA	D	1	0
4GGO	D	4GGP	D	1	0
4FC6	B	4FC7	A	1	0

Table B.2 continued from previous page

pdb_id_1	chain_id_1	pdb_id_2	chain_id_2	analysis _complete	in_unique _CATH_subset
4FEE	A	4KGD	B	1	0
4FFH	D	4FFI	A	1	0
4FI1	A	4JQE	A	1	0
4IX2	D	4QQ3	A	1	0
4FO7	C	4FO8	B	1	0
4B0J	P	4CL6	E	1	0
4HFG	A	4Q96	A	1	0
4FNO	B	4QBK	A	1	0
4G8B	B	4G8D	A	1	0
1SWF	C	1SWG	C	1	0
4MYD	B	4MYS	C	1	0
4GE1	D	4GET	A	1	0
3ROI	B	3SLH	D	1	0
1BFD	A	2FWN	A	1	0
4GGF	V	4XJK	J	1	0
4GGT	A	4GGZ	A	1	0
4GH9	A	4GHA	C	1	0
4GI9	B	4GIA	A	1	0
4GIJ	C	4GIK	B	1	1
4GK9	A	4GU8	B	1	0
4HAX	B	4HB2	B	1	0
4GN7	A	4GN9	B	1	0
3G4E	B	3G4H	B	1	0
4GND	C	4GNG	D	1	0
4GOE	A	4GOF	B	1	0
4JST	B	4JT2	A	1	0
4GQG	A	4JII	X	1	0
4GVP	B	4IF4	B	1	0
4GVR	B	4GVS	B	1	1
4GWA	A	4IPM	A	1	0
1GIQ	A	1GIR	A	1	0
4GZ0	A	4PUQ	A	1	0
4GZW	B	4GZX	C	1	0
4H00	A	4H01	A	1	0
4H15	D	4H16	A	1	0
2OVX	B	2OW1	A	1	0
4K1H	B	4K1K	A	1	0
4H5F	B	4H5G	B	1	0
4H73	D	4NMJ	E	1	0
4H7V	B	4H8U	B	1	0
4H9N	C	4H9P	C	1	0
4HCF	B	4HCI	B	1	0
4HFK	B	4JUR	I	1	0
4HIO	A	4HJ7	A	1	0
4MZC	A	4N11	A	1	0
3T36	B	4HJV	A	1	0
4HLK	A	4HLM	B	1	0
4HLK	D	4HLM	C	1	0
4HMC	A	4HME	A	1	0
4HMS	A	4HMT	B	1	0
4HO4	B	4HO9	A	1	0
2POX	A	2Z6Z	C	1	0
4HQL	B	4HQO	B	1	0
4HRR	E	4HRT	E	1	0
4HRR	F	4HRT	H	1	0
4HRX	A	4HTA	A	1	0
4HS5	A	4LK8	B	1	1
4HSN	C	4UMB	A	1	0
4HSW	A	4HSX	B	1	0
4HU5	A	4HU6	A	1	0
4HUR	A	4HUS	A	1	0
4HWK	A	4J7U	D	1	0
1EVK	A	1EVL	C	1	0
4HY4	A	4MTI	A	1	0
3LDC	A	3LDE	A	1	0
4HYV	A	4KCT	B	1	0
3RYP	A	4N9H	A	1	0
4HZY	A	4I00	A	1	0
4I1W	B	4I2R	A	1	0
4I2N	F	4I46	B	1	0
3P8O	A	4JMY	A	1	0
4I3V	C	4I3X	A	1	0
4I4P	B	4I4R	C	1	0
4I7U	D	4I7W	B	1	0
4ICM	E	4NI8	G	1	0
4ICU	A	4ICV	A	1	0
4IDA	A	4IDF	A	1	0
4IGX	B	4IH2	A	1	0
4IIU	A	4IIV	B	1	0
4IME	G	4IMG	B	1	0
2GMR	H	2J8D	H	1	1
1RG5	L	2GNU	L	1	0
4QS5	B	4QS6	B	1	0
4IOD	C	4IRF	A	1	0
2R0L	L	3GRW	L	1	0
4IOJ	B	4IOK	B	1	0
4ION	B	4IYB	A	1	0
4IP2	A	4IWS	C	1	0
1J7D	B	4DHJ	G	1	0
4IPU	B	4IPV	A	1	0

Table B.2 continued from previous page

pdb_id_1	chain_id_1	pdb_id_2	chain_id_2	analysis _complete	in_unique _CATH_subset
4IR1	F	4IR9	A	1	0
4IU4	A	4IU5	A	1	0
4IX4	A	4IX6	B	1	0
4IXT	B	4IXW	A	1	0
4IY8	A	4IY9	B	1	0
4J10	B	4J11	A	1	1
4J0N	A	4M8D	B	1	0
4JGN	A	4JNX	D	1	0
4IPF	A	4LWU	A	1	0
4J3G	B	4JWP	A	1	0
4J3K	A	4LM7	A	1	0
4J3Y	C	4J48	A	1	0
4J4R	A	4J4V	C	1	0
4J56	H	4J57	F	1	0
4J6B	B	4J6D	A	1	0
4J9F	A	4J9H	A	1	0
4JAO	D	4JAR	D	1	0
4JAY	B	4JB1	A	1	0
3TX6	A	4FB4	A	1	0
1SDQ	A	2B11	A	1	0
4JBG	A	4JBI	J	1	0
4JDR	B	4JQ9	F	1	0
4JEK	A	4NXL	C	1	0
4JF9	B	4JGE	B	1	0
4JES	A	4JET	B	1	0
4JF5	A	4JF6	A	1	0
4JFX	A	4JG1	L	1	0
4JH2	A	4JH9	B	1	0
4JIC	C	4JIP	A	1	0
4JNU	A	4JO9	C	1	0
1XJB	A	4L1W	A	1	0
4JQS	C	4PXY	A	1	0
3DX8	A	3KPQ	A	1	0
4JRH	B	4JRM	C	1	0
1R8M	E	1R8Q	F	1	1
1XLN	A	1XLO	B	1	0
4LOC	B	4M6V	A	1	0
4MSJ	B	4MSQ	C	1	0
4JZX	B	4K10	B	1	0
2I5Y	M	2I60	S	1	0
4K0Z	A	4K1Q	A	1	0
4K1Y	D	4K1Z	B	1	0
4K2E	A	4OOI	C	1	0
4K3G	B	4K3H	H	1	0
4K3U	A	4K7J	B	1	0
4K4P	A	4K4Q	B	1	0
3DDK	A	4WFZ	A	1	0
4K5U	A	4K79	B	1	0
4K6C	B	4K6F	A	1	0
4K84	A	4KBS	B	1	0
4K9Z	A	4KA0	B	1	0
4KEN	B	4KG6	C	1	0
1KX5	G	4J8U	G	1	0
4KIC	B	4KIG	A	1	0
4LJ3	B	4LYK	A	1	0
4KL9	A	4KM2	B	1	0
4KNK	B	4KNL	B	1	0
4KNS	F	4KNU	C	1	0
4KRX	B	4KRY	E	1	0
2W9N	A	3ZNZ	B	1	0
1DY9	B	1W3C	A	1	0
4BH1	F	4CQY	B	1	0
4KXA	A	4KXD	A	1	0
4KXX	A	4KXY	A	1	0
4L63	B	4L6T	D	1	0
4L7B	A	4L7D	C	1	0
4L9X	B	4LH8	A	1	0
4LBH	A	4LBI	A	1	0
4LE3	A	4LE4	B	1	0
4LF1	D	4LF2	F	1	0
4LFL	C	4LFN	A	1	0
4LFK	D	4LFL	B	1	0
3QBT	C	3TNF	A	1	0
4LJK	B	4LJL	B	1	0
4LN6	E	4LN8	K	1	0
4LO1	A	4QD2	H	1	0
4LO2	C	4LO8	B	1	0
3TX0	A	3UO0	C	1	0
4LS7	B	4LS8	A	1	0
4LXZ	C	4LY1	A	1	0
3LER	B	3M5V	B	1	0
3NN1	B	3NN2	A	1	0
2WYE	B	3SRA	B	1	0
4M3N	B	4MAR	A	1	0
2XCY	B	2XZK	A	1	0
4M5R	A	4M5U	A	1	0
1QF1	A	3T2J	E	1	0
4M93	C	4MA1	L	1	0
4M8A	A	4Q0F	C	1	0
4M8G	B	4M90	A	1	0

Table B.2 continued from previous page

pdb_id_1	chain_id_1	pdb_id_2	chain_id_2	analysis _complete	in_unique _CATH_subset
4M9B	A	4MAP	B	1	0
4MA3	B	4O51	F	1	0
4MAL	B	4MBQ	C	1	0
1YVM	A	4A6W	A	1	0
4MBX	G	4MBY	D	1	0
2ZNT	A	3FV2	A	1	0
4MFP	A	4MFQ	A	1	0
4MHD	B	4MI4	A	1	0
4MIE	A	4MIN	D	1	0
4MIR	B	4ML7	D	1	0
4HCD	A	4HCL	B	1	0
3PS2	A	4IS9	A	1	0
4MT8	B	4MTX	A	1	0
4MTD	B	4MTE	D	1	0
4MTU	A	4MZQ	K	1	0
4MUS	B	4OAK	A	1	0
1DC4	A	1DC6	A	1	0
4MWQ	A	4MWU	A	1	0
4MWL	A	4MX0	A	1	0
4MY8	C	4MYA	A	1	0
4MZM	D	4MZT	B	1	0
4N21	F	4N23	A	1	0
1RF5	C	1RF6	D	1	0
4N4N	F	4N4O	D	1	0
4BSA	B	4BSD	B	1	0
4N5L	B	4N5M	B	1	0
4KDN	C	4KDO	A	1	0
4N8N	A	4N8O	B	1	0
4N9K	B	4N9L	B	1	0
3SOR	A	4X6P	A	1	0
4NDT	A	4NDV	A	1	0
4NE3	A	4NE5	G	1	0
4NE3	B	4NE5	H	1	0
3IX8	D	3JPU	C	1	1
2FX8	N	2FX9	L	1	0
4NI3	B	4PSR	A	1	0
4NJI	A	4NJK	B	1	0
4NJS	C	4NJU	D	1	0
2WZZ	A	4GZB	A	1	0
4NM0	B	4NM7	B	1	0
3BO8	A	4NQV	A	1	0
3V2V	A	3VM9	B	1	0
4NT1	B	4NT2	A	1	0
4NXK	B	4NZF	G	1	0
4NXT	A	4NXU	B	1	0
4NZ1	B	4NZ3	A	1	0
4O0Q	B	4O1F	A	1	0
1BX4	A	2I6B	B	1	0
4O2C	A	4O2F	D	1	0
4O59	O	4O63	O	1	0
4O6Y	B	4O79	A	1	0
4O7M	D	4OA4	B	1	0
4OA5	A	4OA8	B	1	0
4OAF	D	4OAG	A	1	0
4OBC	A	4WTL	A	1	0
4OBD	B	4OBG	A	1	0
4OBW	B	4OBX	A	1	0
1LYA	A	1LYW	C	1	0
1LYB	D	1LYW	H	1	0
4OCF	D	4OD7	C	1	0
4WH1	A	4WH2	A	1	0
4OCM	C	4OCN	C	1	0
4OGN	A	4WT2	A	1	0
4CDE	F	4CDF	F	1	0
4OET	A	4OEV	B	1	0
4OF6	B	4OF7	D	1	0
4OIT	C	4OKC	B	1	0
4OKE	B	4OKJ	A	1	0
4OMW	B	4TLJ	A	1	0
4X0C	A	4X7D	A	1	0
4OOV	A	4X05	A	1	0
4X07	B	4X7C	A	1	0
1Y7L	A	4LHG	X	1	0
4OT6	A	4RG0	A	1	0
4OWJ	D	4OWK	B	1	0
4OXM	A	4P67	F	1	0
4PFN	A	4TN4	C	1	0
3S6E	B	4J5O	A	1	0
4P0I	A	4POX	B	1	0
4P0Y	B	4PM3	B	1	0
2HHN	A	2HXZ	A	1	0
4PCA	C	4PCL	B	1	0
4PEX	A	4PF0	A	1	0
2QUL	D	2QUM	C	1	0
1WBZ	A	3ROO	C	1	0
4PGN	C	4PGP	B	1	0
2RD6	A	3GJW	A	1	0
4PKG	G	4PKH	B	1	0
4PLC	D	4PLG	A	1	0
4POR	J	4POT	H	1	0

Table B.2 continued from previous page

pdb_id_1	chain_id_1	pdb_id_2	chain_id_2	analysis _complete	in_unique _CATH_subset
4PTZ	B	4PU0	D	1	0
4PU6	B	4PV3	D	1	0
3M9J	A	3M9K	B	1	0
4Q0U	A	4Q0V	A	1	0
2AAI	A	4HV7	X	1	0
4Q6F	B	4QF2	B	1	0
4QH7	E	4QH8	G	1	0
2GTQ	A	4PU2	A	1	0
4QJ7	D	4QJ8	D	1	0
4QKD	C	4QKF	C	1	0
4QLX	B	4QLY	C	1	0
4QOP	A	4QOQ	C	1	0
4QPD	A	4TS4	A	1	0
4QQH	A	4QQL	I	1	0
4QR0	A	4QR2	B	1	0
4QSE	A	4RJZ	A	1	0
4QX8	A	4QXB	C	1	0
4QX8	D	4QXH	B	1	0
4QY0	K	4QY2	I	1	0
1A1L	A	1AAY	A	1	1
3Q43	A	3Q44	A	1	0
4R7W	A	4R88	D	1	0
4R83	D	4R84	C	1	0
4RC1	I	4U9P	C	1	0
4RDK	A	4RDL	B	1	0
4RDN	A	4RDO	D	1	0
4RDZ	B	4RE0	A	1	0
4RHR	C	4RHS	E	1	0
4RJJ	E	4RJK	B	1	0
4RKP	B	4RKS	A	1	0
4RPN	B	4RPO	C	1	0
4ROF	B	4RRE	E	1	0
4RP4	A	4RP5	B	1	0
4RUQ	B	4RUS	A	1	0
3PD2	A	3PD5	B	1	0
4TLX	C	4TM3	B	1	0
4TMV	A	4TMW	B	1	0
2XMB	A	2XMC	A	1	0
4TRB	B	4TRC	A	1	0
4TZO	E	4TZQ	C	1	0
1B4V	A	2GEW	A	1	0
4U5G	A	4U5H	D	1	0
4U64	A	4U65	D	1	0
4U9J	B	4U9K	A	1	0
4UAR	A	4UAS	A	1	0
4UD2	Q	4UD6	S	1	0
1I40	A	2AU9	A	1	0
4UME	A	4UMF	C	1	0
4UOX	A	4UOY	A	1	0
4UQL	B	4URH	C	1	0
4UQL	R	4UQP	Q	1	0
4URZ	R	4US0	R	1	0
4UTT	B	4UTW	B	1	0
1LE6	A	1LE7	A	1	0
4QYD	A	4QYL	A	1	0
4W53	A	4W58	A	1	0
4TYU	A	4U7S	B	1	0
4W7Y	A	4W7Z	B	1	0
4W85	A	4W89	B	1	0
4WJ9	A	4X4L	A	1	0
1YMK	A	1YS0	A	1	0
2XIB	B	2XII	A	1	0
4XKD	A	4XKE	C	1	0
4XKD	F	4XKE	D	1	0
4WWJ	B	4WWZ	B	1	0
4WX6	C	4WX7	A	1	0
4X1W	A	4X1Z	B	1	0
4X3T	E	4X3U	A	1	0
4X7C	D	4X7E	C	1	0
4X7H	A	4X7J	A	1	0
4X8B	B	4X8D	A	1	0
4XCK	D	4XDA	A	1	0
4X90	B	4X91	A	1	0
1BLS	A	1XX2	B	1	0
4Y0G	B	4YEE	H	1	0
4Y6P	A	4Y6S	B	1	0
1M1D	A	1PU9	A	1	0
4NIA	K	4OQ9	A	0	0
1B0P	B	2C42	B	0	0
1DM0	F	2XSC	B	0	0
4AR4	A	4K9F	A	0	0
1C48	E	1CQF	C	0	0
1GI7	A	1GJA	A	0	0
2JGO	A	3LJM	B	0	0
1CZW	J	1D1I	E	0	0
1HDX	B	1U3U	A	0	0
1XLC	A	1XLL	B	0	0
4OIJ	D	4OIK	B	0	0
1UGI	E	2UGI	A	0	0
1FFK	G	4V9F	J	0	0

Table B.2 continued from previous page

pdb_id_1	chain_id_1	pdb_id_2	chain_id_2	analysis _complete	in_unique _CATH_subset
1FFK	H	4V9F	K	0	0
1FFK	T	1VQ9	W	0	0
1VQ9	3	4V9F	3	0	0
1GW9	A	9XIA	A	0	0
2BE5	M	2O5I	C	0	0
1KD8	D	1KDD	B	0	0
2HOR	A	2HOX	A	0	0
1MKF	B	2NZ1	X	0	0
3CWH	A	4A8I	A	0	0
1OXS	C	1OXU	C	0	0
1P1Q	B	1P1U	B	0	0
1PX8	B	1UHV	B	0	0
1Q2V	D	1Q3Q	C	0	0
1Q43	B	1Q5O	A	0	0
1Q9L	C	3T77	A	0	0
1QMG	C	1YVE	I	0	0
1QOH	Q	2BOS	A	0	0
1R14	A	3PAQ	A	0	0
1XC1	D	3TYH	H	0	0
2XAP	E	2XO4	F	0	0
1R1U	B	1R1V	B	0	0
1R4P	C	4M1U	E	0	0
1R8W	B	4MTJ	A	0	0
1R9Z	A	1RAK	A	0	0
1IVW	B	1WMP	A	0	0
1RJR	A	1RK5	A	0	0
1RVX	G	1RVZ	E	0	0
1RWA	A	1RWF	A	0	0
1FFK	X	1JJ2	Z	0	0
1VQ9	2	4V9F	2	0	0
1VQ9	A	4V9F	A	0	0
1VQ9	B	4V9F	B	0	0
1VQ7	E	4V9F	E	0	0
1VQO	F	4V9F	F	0	0
1VQN	N	4V9F	N	0	0
1FFK	L	1JJ2	N	0	0
1FFK	O	1JJ2	Q	0	0
1FFK	P	1JJ2	R	0	0
1VQ9	T	4V9F	T	0	0
1VQ7	V	4V9F	V	0	0
1FFK	U	1JJ2	W	0	0
1SF2	D	1SFF	A	0	0
1SOI	A	1SU2	B	0	0
1SLI	A	1SLL	A	0	0
1NTM	B	1PPJ	B	0	0
1T0R	A	1T0S	A	0	0
1NDF	B	1NDI	A	0	0
1TJV	D	1TJW	B	0	0
1U0F	B	1U0G	A	0	0
1U15	B	1U16	A	0	0
3DP8	C	4DCY	A	0	0
2HU5	B	3O4G	B	0	0
1W4X	A	2YLR	A	0	0
6XIM	A	7XIM	B	0	0
1XYC	A	2GYI	B	0	0
1YQ5	B	1YQ6	A	0	0
2BUK	A	4V4M	l	0	0
2BWM	A	2BWR	B	0	0
2GEL	G	3ZEU	D	0	0
2GHA	B	2GHB	A	0	0
2IHS	D	3F2O	C	0	0
2UW8	A	2VNW	A	0	0
1HI8	B	1UVI	C	0	0
1WZB	B	3B2C	I	0	0
3INS	A	3ZS2	C	0	0
3UC7	F	3UC8	C	0	0
2O7I	A	3I5O	B	0	0
3BOG	D	3BOI	A	0	0
1FFK	R	1YHQ	U	0	0
2QUD	A	2QUX	A	0	0
3PUX	F	3RLF	F	0	0
3CAY	F	3CBA	E	0	0
2UZT	A	2UZU	E	0	0
3F86	H	3F87	D	0	0
3FXK	A	3FXM	A	0	0
3U5N	C	4MZG	A	0	0
3JTM	A	3N7U	D	0	0
3PON	B	4LOR	D	0	0
3C2P	A	3C46	B	0	0
4XKN	A	4XKR	A	0	0
4GL4	A	4JJI	B	0	0
3WLE	C	3WLF	B	0	0
4B4P	B	4W6X	A	0	0
4C33	A	4C35	A	0	0
4DLA	B	4DLB	A	0	0
4GLN	D	4GLS	C	0	0
4LOO	B	4LOQ	K	0	0
4MHX	B	4MIV	C	0	0
4MPB	B	4NEA	D	0	0
4MRN	A	4MRP	B	0	0

Table B.2 continued from previous page

pdb_id_1	chain_id_1	pdb_id_2	chain_id_2	analysis _complete	in_unique _CATH_subset
4PEH	B	4PEI	E	0	0
4UR7	C	4UR8	A	0	0
3EYD	D	3KF2	D	0	0
1CAG	B	1CGD	A	0	0
3V2P	C	3V2Q	E	0	1
1GCL	D	1UO2	B	0	0
1WS4	H	1WS5	F	0	0
1RB4	C	3K7Z	C	0	0
2B1F	C	3CRP	C	0	0
2B22	A	3CK4	G	0	0
2EFR	C	2EFS	D	0	0
2IPZ	A	3CK4	L	0	0
2NRN	B	3CRP	D	0	0
2R5B	A	2R5D	C	0	0
2Z2T	A	3AHA	C	0	0
3F4Y	C	3VTQ	E	0	0
3W92	B	3W93	B	0	0
4AK4	P	4AKC	D	0	0
4HZ1	D	4JKN	D	0	0
4BXP	A	4BXR	B	0	0
3ITM	A	3ITU	D	0	0
4C2Y	B	4C2Z	A	0	0
4C3T	B	4UOV	B	0	0
4JZL	B	4JZP	B	0	0
2X4W	B	2X4X	B	0	0

C

Supporting Information for Transmembrane Protein Folding Pathway Prediction

Table C.1: Analysed OPM structures

tm_count	famid	pdbid	chain_id
3	364	6zkd	A
3	364	3rko	K
3	364	6ztq	A
3	714	4rng	A
3	714	4qnd	B
3	714	4x5m	A
3	714	4qnc	B
3	9	1kf6	C
3	9	1nek	C
3	9	6lum	C
3	9	2h88	C
3	9	1zoy	C
3	9	4ysx	C
3	2	6a2w	A
3	2	3pl9	A
3	2	1rwt	A
3	2	2bhw	A
4	22	6y5a	A
4	22	6w1j	A
4	22	7ekp	A
4	22	6x40	A
4	22	4cof	E
4	22	5osa	A
4	22	6pxd	E

Table C.1 continued from previous page

tm_count	famid	pdbid	chain_id
4	567	6k4j	A
4	567	5tcx	A
4	567	6wvg	A
4	279	7jkc	G
4	279	7jjp	A
4	279	6l3t	A
4	254	6vgc	A
4	254	4ntf	A
4	254	2uuh	A
4	254	4yl3	A
4	265	6fl9	E
4	265	3tlw	A
4	265	3rqw	B
4	265	6hjx	A
5	3	2j8c	L
5	3	1dxr	L
5	3	1eys	M
5	532	7c9r	L
5	532	5y5s	L
5	532	6et5	L
5	532	6z5s	M
6	701	4mnd	A
6	701	4o6m	A
6	701	6wm5	A
6	701	6h59	A
6	682	6pqo	A
6	682	7b0j	A
6	682	6aei	A
6	682	6uz8	A
6	99	3qf4	A
6	99	4pl0	A
6	99	3wmg	A
6	99	6bpl	A
6	99	6bl6	A
6	99	4ayt	A
6	99	2hyd	A
6	99	5c78	A
6	5	6wj6	B
6	5	3wu2	B
6	5	4yuu	B
6	5	6dhe	B
6	5	6kac	B
6	237	2nr9	A
6	237	2xow	A
6	237	4qo2	A
6	15	3c02	A
6	15	1ymg	A
6	15	1j4n	A
6	15	6f7h	A

Table C.1 continued from previous page

tm_count	famid	pdbid	chain_id
6	15	4nef	A
6	15	3gd8	A
6	15	3d9s	A
6	15	6qzi	A
6	15	3ne2	A
6	15	2f2b	A
6	15	2w2e	A
6	15	1z98	A
6	15	5i32	A
6	15	2o9g	A
6	15	3llq	A
6	15	1ldf	A
6	301	3tds	E
6	301	6vqr	A
6	301	3q7k	A
6	301	3kcu	A
6	301	3kly	A
6	301	4fc4	A
7	823	5h35	C
7	823	5wud	A
7	823	5eik	A
7	823	6iyx	A
7	667	6bd4	A
7	667	6o3c	A
7	667	4jkv	A
7	14	4iar	A
7	14	6a93	A
7	14	4ib4	A
7	14	6bqg	A
7	14	2ydv	A
7	14	6kux	A
7	14	6k41	R
7	14	6kuw	A
7	14	5vbl	B
7	14	6h7j	A
7	14	7bu6	A
7	14	4lde	A
7	691	7c7q	A
7	691	4or2	A
7	691	6ffi	A
7	242	6uun	R
7	242	4k5y	B
7	242	7d68	R
7	242	5xez	A
7	242	6fj3	A
7	242	7d3s	R
7	13	1vgo	A
7	13	6gux	A
7	13	5jsi	A

Table C.1 continued from previous page

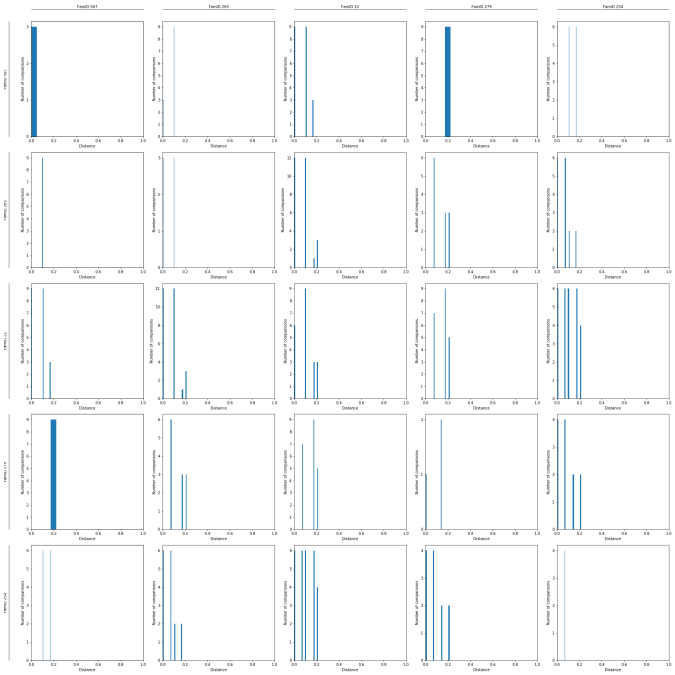
tm_count	famid	pdbid	chain_id
7	13	5azd	A
7	13	1m0l	A
7	13	6eyu	A
7	13	4pxk	A
7	13	4wav	A
7	13	4qi1	A
7	13	4knf	A
7	13	6k6i	A
7	13	5b2n	A
7	13	4jr8	A
7	13	6lm0	A
7	13	6lm1	A
7	13	4fbz	A
7	13	3a7k	A
7	13	1e12	A
7	13	3vvk	A
7	13	6gyh	A
7	13	6csm	A
7	13	4jq6	A
7	13	4hyj	A
7	13	6nwd	A
7	13	5ax0	A
7	13	5zih	A
7	13	4tl3	A
7	13	1h68	A
10	302	3k3f	A
10	302	4ezc	A
10	302	6qd5	A
10	30	6zhh	A
10	30	3ar2	A
10	30	5mrw	A
10	27	6z3t	A
10	27	3mp7	A
10	27	5aww	Y
10	675	6bw6	A
10	675	5jnq	A
10	675	5ckr	B
10	25	4r9u	A
10	25	2nq2	A
10	25	5b57	A
11	21	1u7g	A
11	21	6eu6	A
11	21	2b2f	A
11	21	5af1	A
11	21	3b9y	A
11	4	7coy	A
11	4	6kig	A
11	4	6sl5	A
11	4	1jb0	A

Table C.1 continued from previous page

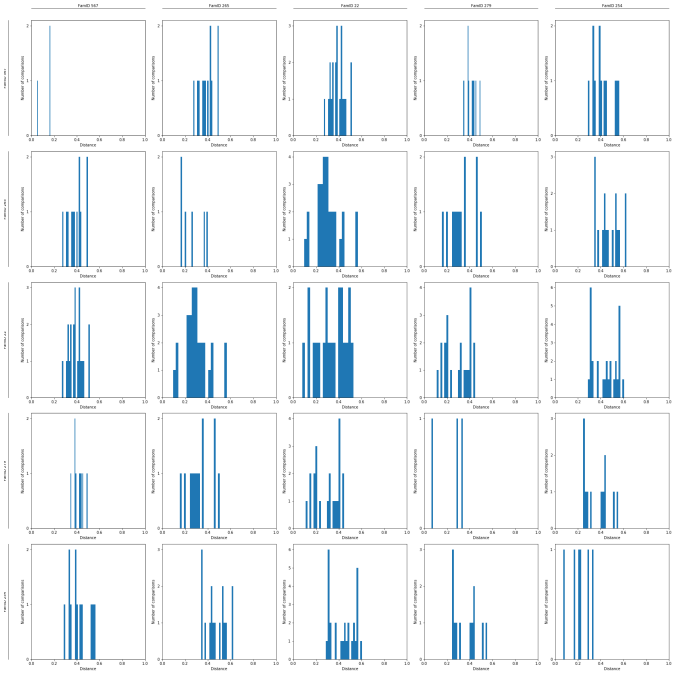
tm_count	famid	pdbid	chain_id
11	4	4kt0	A
11	4	6kmx	A
11	4	5l8r	A
11	4	6l4u	A
11	4	6ijj	A
11	4	6k61	A
12	11	3hb3	A
12	11	1m56	A
12	11	7coh	A
12	11	4xyd	A
12	11	3wfd	B
12	1010	6npl	A
12	1010	6kkt	A
12	1010	7d99	A
12	1010	6m22	A
12	78	6m18	A
12	78	6m0z	A
12	78	7dii	A
12	23	6e9n	A
12	23	6g9x	A
12	23	6t1z	A
12	23	6kkl	A
12	24	6zoe	A
12	24	6vkt	A
12	24	6ows	A
12	24	3w9i	A
12	281	3gia	A
12	281	5oqt	A
12	281	5j4i	A
12	327	6fv8	A
12	327	6fhz	A
12	327	5y50	A
12	327	6idp	A
12	327	5xjj	A
12	327	4z3n	A
12	327	5c6p	A
12	656	4gc0	A
12	656	5eqg	A
12	656	4zwc	A
12	656	6m20	A
12	656	4j05	C
12	656	6h7d	A
14	18	7jm6	A
14	18	1ots	A
14	18	1kpl	A
14	18	6d0j	A
14	333	6gs4	A
14	333	4uvm	A
14	333	6exs	A

Table C.1 continued from previous page

tm_count	famid	pdbid	chain_id
14	333	6ei3	A
14	333	4ikv	A

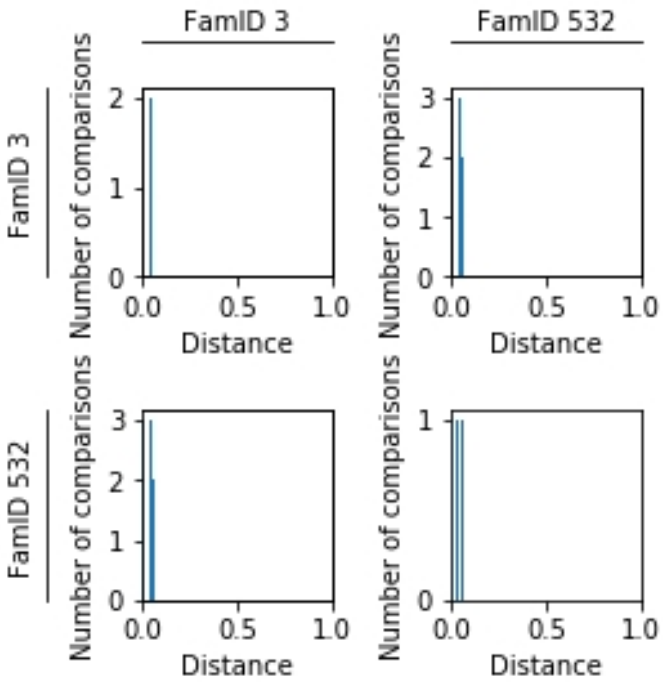


(a) Pathway matrix comparisons (Jaccard distance)

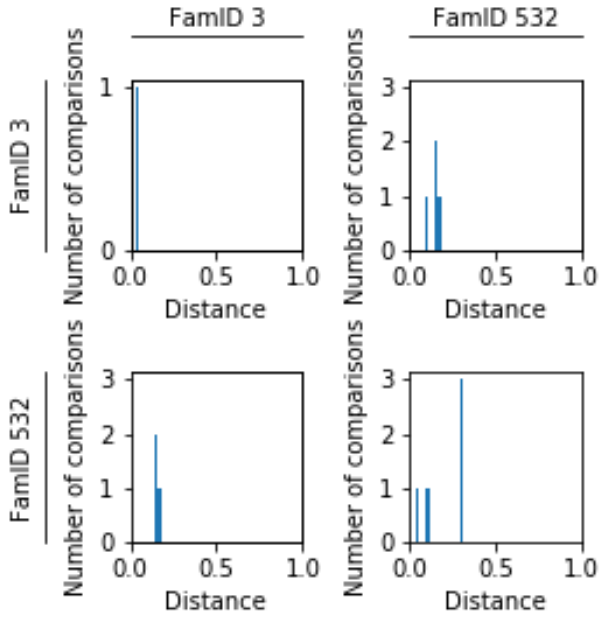


(b) Energy matrix comparisons (Sum of absolute differences)

Figure C.1: Pathway and energy comparisons of families with 4 TM helices. Normalised distances (x-axes) are computed between pathway (**a**)) or energy (**b**)) matrices of all proteins of one family and all proteins of another family (diagonal has within family comparisons). Order of OPM family IDs from left to right and top to bottom: 567, 265, 22, 279, 254.

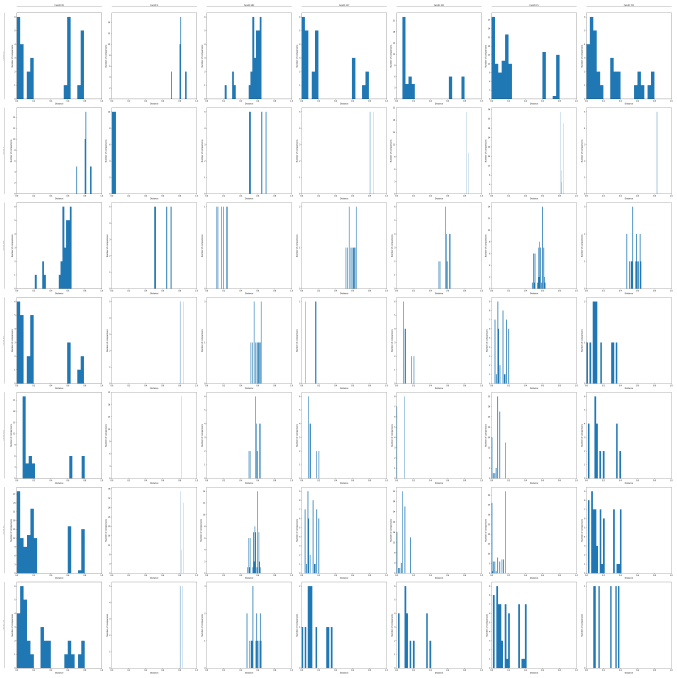


(a) Pathway matrix comparisons (Jaccard distance)

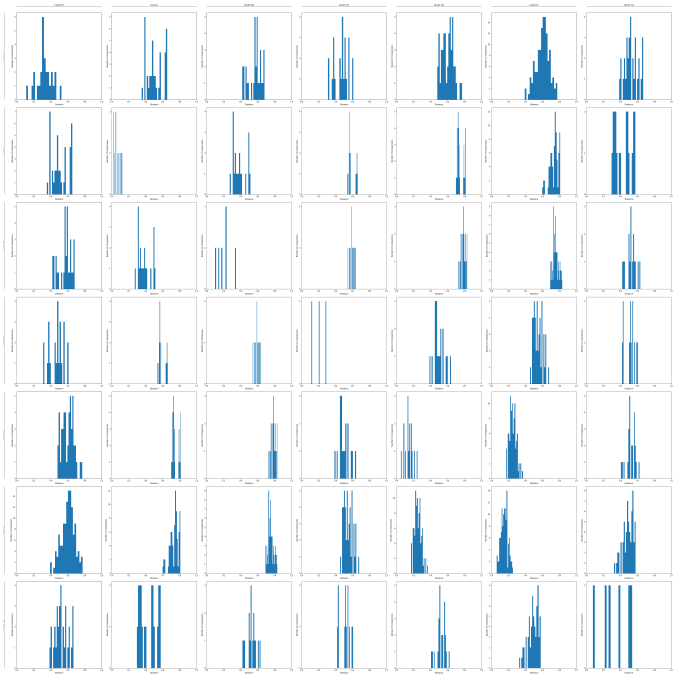


(b) Energy matrix comparisons (Sum of absolute differences)

Figure C.2: Pathway and energy comparisons of families with 5 TM helices. Normalised distances (x-axes) are computed between pathway (**a**)) or energy (**b**)) matrices of all proteins of one family and all proteins of another family (diagonal has within family comparisons). Order of OPM family IDs from left to right and top to bottom: 3, 532.

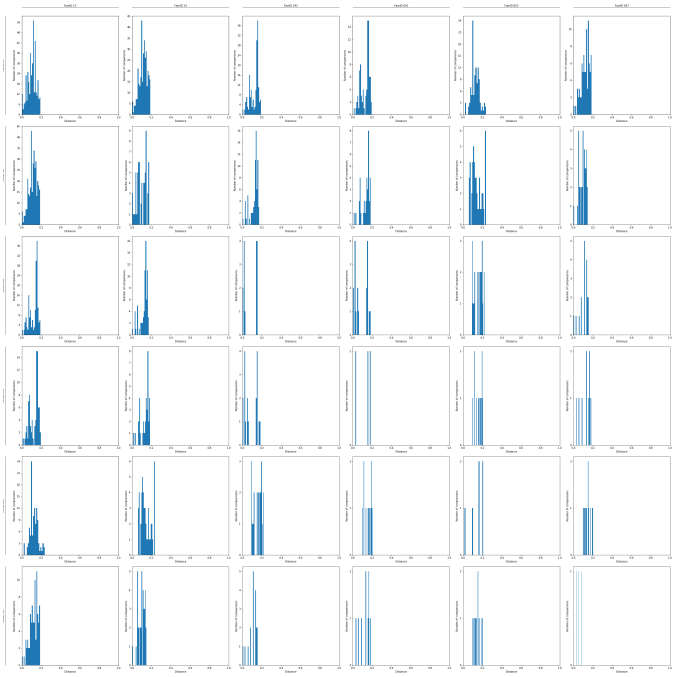


(a) Pathway matrix comparisons (Jaccard distance)

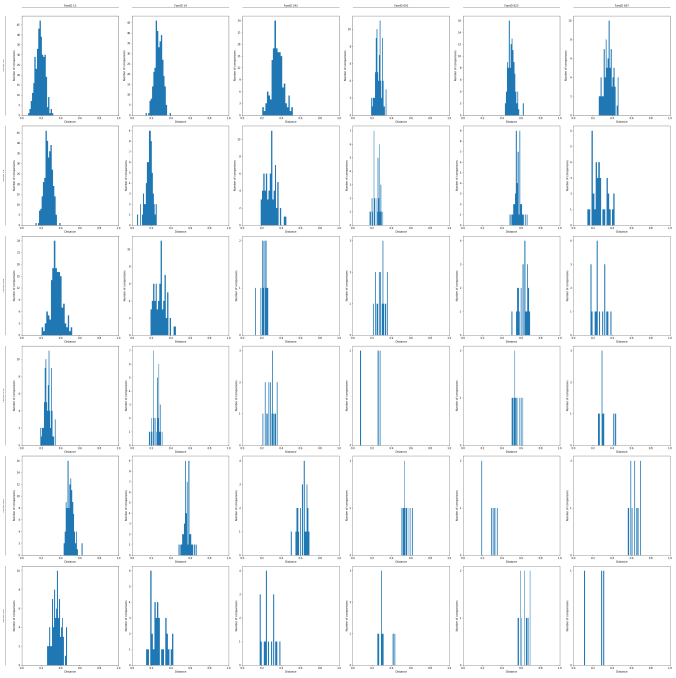


(b) Energy matrix comparisons (Sum of absolute differences)

Figure C.3: Pathway and energy comparisons of families with 6 TM helices. Normalised distances (x-axes) are computed between pathway (**a**)) or energy (**b**)) matrices of all proteins of one family and all proteins of another family (diagonal has within family comparisons). Order of OPM family IDs from left to right and top to bottom: 99, 5, 682, 237, 301, 15, 701.

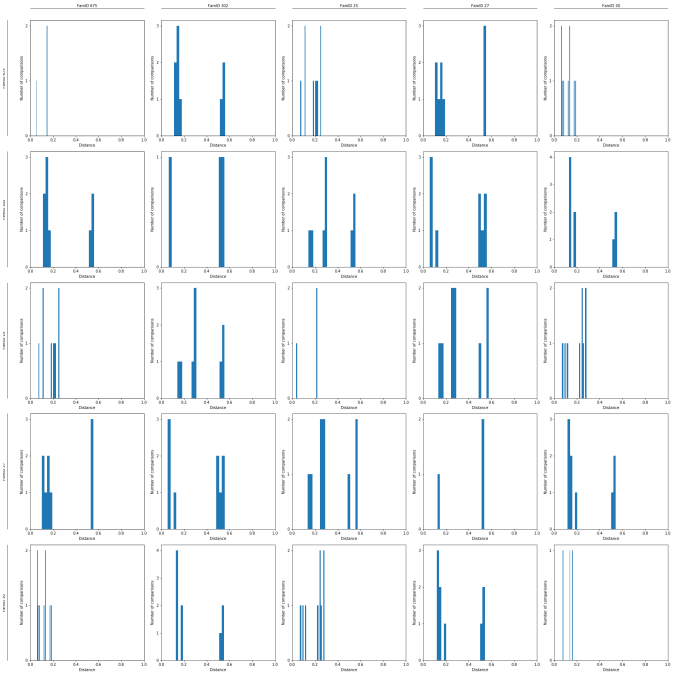


(a) Pathway matrix comparisons (Jaccard distance)

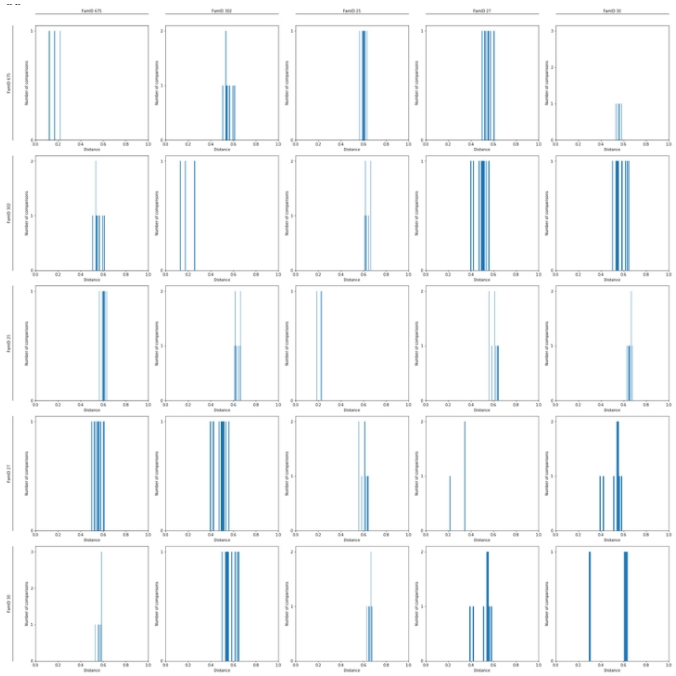


(b) Energy matrix comparisons (Sum of absolute differences)

Figure C.4: Pathway and energy comparisons of families with 7 TM helices. Normalised distances (x-axes) are computed between pathway (**a**)) or energy (**b**)) matrices of all proteins of one family and all proteins of another family (diagonal has within family comparisons). Order of OPM family IDs from left to right and top to bottom: 13, 14, 242, 691, 823, 667.

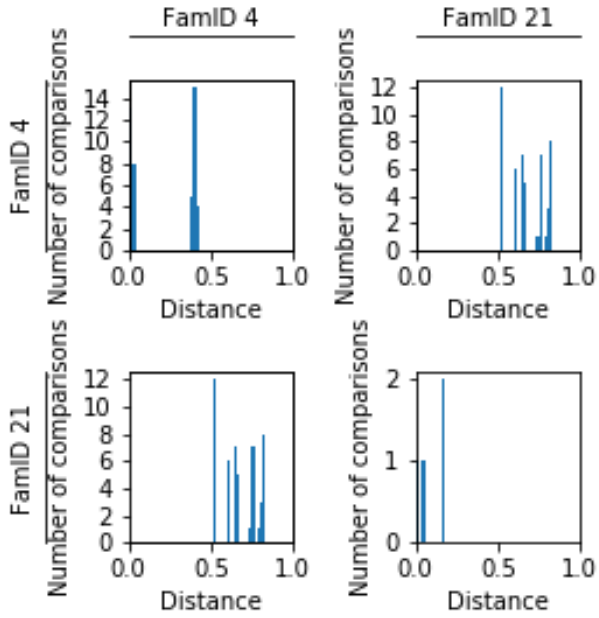


(a) Pathway matrix comparisons (Jaccard distance)

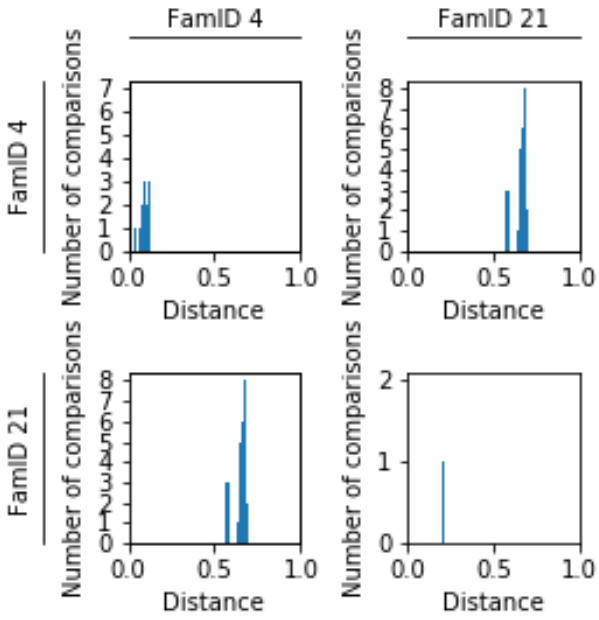


(b) Energy matrix comparisons (Sum of absolute differences)

Figure C.5: Pathway and energy comparisons of families with 10 TM helices. Normalised distances (x-axes) are computed between pathway (**a**)) or energy (**b**)) matrices of all proteins of one family and all proteins of another family (diagonal has within family comparisons). Order of OPM family IDs from left to right and top to bottom: 675, 302, 25, 27, 30.

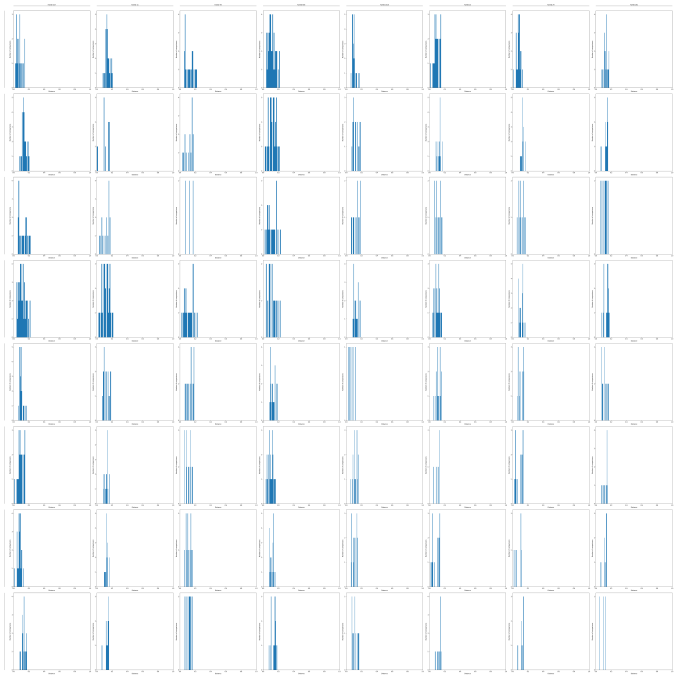


(a) Pathway matrix comparisons (Jaccard distance)

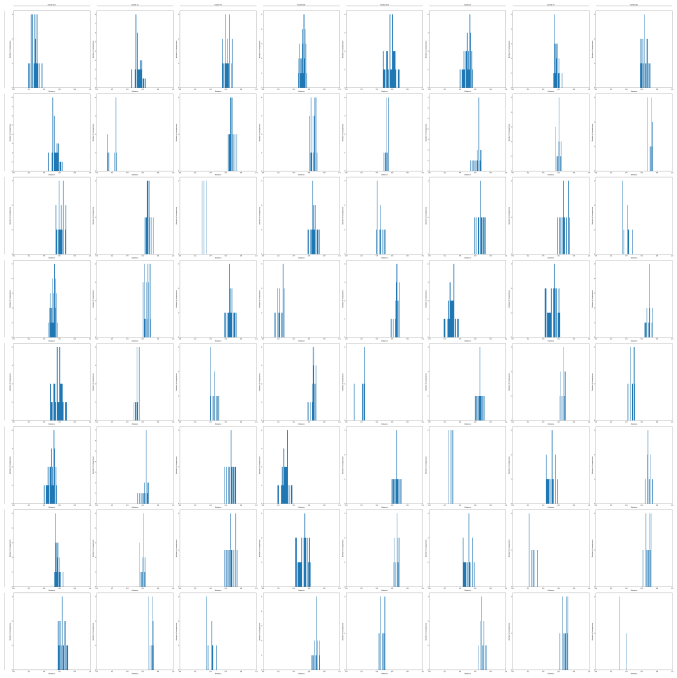


(b) Energy matrix comparisons (Sum of absolute differences)

Figure C.6: Pathway and energy comparisons of families with 11 TM helices. Normalised distances (x-axes) are computed between pathway (**a**)) or energy (**b**)) matrices of all proteins of one family and all proteins of another family (diagonal has within family comparisons). Order of OPM family IDs from left to right and top to bottom: 4, 21.

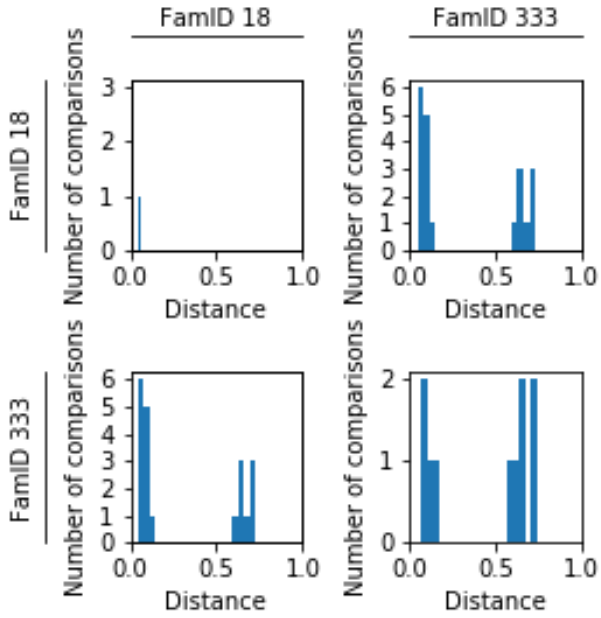


(a) Pathway matrix comparisons (Jaccard distance)

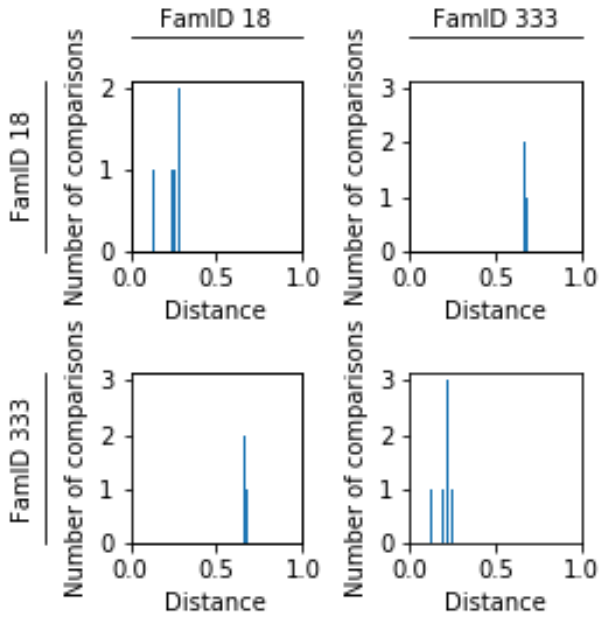


(b) Energy matrix comparisons (Sum of absolute differences)

Figure C.7: Pathway and energy comparisons of families with 12 TM helices. Normalised distances (x-axes) are computed between pathway (**a**)) or energy (**b**)) matrices of all proteins of one family and all proteins of another family (diagonal has within family comparisons). Order of OPM family IDs from left to right and top to bottom: 327, 11, 78, 656, 1010, 23, 24, 281.



(a) Pathway matrix comparisons (Jaccard distance)



(b) Energy matrix comparisons (Sum of absolute differences)

Figure C.8: Pathway and energy comparisons of families with 14 TM helices. Normalised distances (x-axes) are computed between pathway (**a**)) or energy (**b**)) matrices of all proteins of one family and all proteins of another family (diagonal has within family comparisons). Order of OPM family IDs from left to right and top to bottom: 18, 333.