Predicting Multidomain Protein Structure And Function Via Co-evolved Amino Acids And Application To Polyketide Synthases.

by

Tuğçe Oruç

A thesis submitted to
The University of Birmingham
for the degree of
Doctor of Philosophy (PhD)

College of Life and Environmental Sciences
School of Biosciences
The University of Birmingham
March 2021

Abstract

Proteins are an important building block of life, and they are responsible for many processes in living organisms. Therefore, understanding their functions and working mechanisms has vital importance to answer many questions about diseases and is a basis for the development of novel drugs. Three dimensional (3D) structure of proteins determine their functions; therefore, the determination of the 3D structures of proteins has been studied widely. Although many experimental techniques have been developed to determine the structures of proteins, they have limitations, especially for large protein complexes. Protein structure can help understand protein function, as can looking at conserved residues, but typically time consuming mutagenesis experiments combined with protein function assays are needed. As an alternative to the experimental methods, researchers have been working on developing computational approaches. While it is relatively easy to predict structures when the structure of a homologous protein is known, as it can be used as a template, the prediction of protein structures in the absence of a template is more challenging. For template-free predictions, coevolved amino acid residue pairs, predicted from the alignment of the homologous sequences, provided promising improvements in the field. More recently, successful implementation of the artificial neural networks, fed by the predicted coevolved residue pairs, improved the accuracy of the predicted structures further. Although there are promising developments in the coevolution based approaches, especially for the structure prediction of small/medium-sized proteins, more developments are needed for predicting protein structure, particularly of large protein complexes. Here, we show that the prediction of distances between residue pairs, via deep neural networks fed by predictions of coevolved residue pairs, improves the accuracy of structure prediction in small/medium-sized proteins. The prediction of residue pair distances, using a similar approach, in two interacting domains also allows us to predict how two domains on the same chain interact with each other. Further, we show that prediction of coevolved residue groups, via statistical coupling analysis, allows us to determine functional boundaries of domains and diverged amino acid patterns in the sub-types of the domains in a multi-domain protein complex, a polyketide synthase. We found that using predicted distances, in addition to the predicted residue pairs in contact, allows us to generate structures closer to the experimental structures, and to select them as the final models in a straightforward approach. Additionally, we reveal that the distances of the residue pairs on interacting domain pairs can be predicted accurately leading to the successful prediction of the structural interface between two interacting proteins when the interface surface is large, and the sequence alignment is comprehensive enough. Finally, we found that functional domain boundaries, which are consistent with the experimental studies, can be determined. Also, some coevolved residue groups have distinct amino acid patterns in different domain sub-types including the positions that have already known as the fingerprint motifs of the different sub-types. These approaches can be applied to predict the structures of individual domains and to predict how two domains interact with each other, which can be used to predict the structure of multi-domain proteins. The work on polyketides here demonstrates how these developments might be applied, since identifying domain boundaries and residues important for substrate specificity should aid in the design of novel polyketide synthases and thus of novel polyketides. This in itself is an important development given the commercial and medicinal importance of polyketides, but also opens the way to similar analysis on other multidomain proteins.

To my mother, father and sister

ACKNOWLEDGEMENTS

First of all, I am extremely grateful to my supervisor Peter J. Winn for his supervision, for everything he taught me, for having an enormous contribution to my scientific perspective, and for being always supportive and understanding. I am also very thankful to my co-supervisor Christopher M. Thomas for his supervision and for always challenging me to think from another perspective. I am very thankful to my viva examiners Apoorva Bhatt and Maya Topf for their valuable comments and suggestions regarding my thesis, and thanks to them I could enjoy an exam for the first time in my unending - twenty-five years lasting - student life. I am also thankful to Jeff Cole for his time to read and comment on my introduction chapter. I am also extremely thankful to Deniz Sezer, my master's supervisor, for everything he taught me as they constitute the basis of my computational and biophysical knowledge.

I am extremely thankful to the Darwin Trust of Edinburgh for financially supporting my PhD studies. I am also very thankful to the Alan Turing Institute for the enrichment placement.

I am very thankful to Shuangxi Ji for helping me to adapt to a new research topic and patiently answering my endless questions. I am also very thankful to every member of the Winn group, it was very nice to meet and work with them.

Being in the same country with my friends from Turkey made my time in the UK much more enjoyable. I am very thankful to Cantuğ Bar for being such a great friend and being there for me whenever I need someone whom I can trust, and to Ahmet Sinan Yavuz for all the good times we have shared.

Being able to find good friends in my working environment has always been a motivation for me. I am very thankful Zuhal Eraslan, Eleni Christidi, Kim Summers, Edwin Aponte Angarita and all members of Centre for Computational Biology. I could spend just five months at the Turing Institute, but it was more than enough to have the greatest time. I am very thankful everyone I met at the Turing. I am especially thankful to the members of our “booth”, Victoria Volodina, Alessandro Ragano and Pedro Pinto da Silva, and beloved members of our social activity group Daniele Guariso, Ferran Gonzalez Hernandez, Sara Masarone, Beatriz Costa Gomes and Katriona Goldmann. I am also extemely thankful for my roommates, Ebru Şener and Mertkan Şener, for being like a family to me in London.

I am extremely thankful to my dearest friends Aslı Yenenler-Kutlu, Burak Büyüksakallı, Burcu Ekinci, Esra Sinoplu-Nalbat, Hande Aypek, Hande Kösek-Aluç, İlkcan Ercan-Orhan, Kader Özgür-Büyüksakallı, Mert Uygun, Nisan Erinç-Bayraktar, Tülin Ece for being in my life and their continous support even I was kilometers away from them.

Last, but not least, I am very thankful to all members of my great family. It would never be enough to thank to my mother Güldal Oruç, father Abdullah Oruç, sister Hande Oruç-Kırmızı and brother Burhan Kırmızı for their infinite and unconditional love and support. I am also very thankful to the all members of my family Ada Kırmızı, Aslı Mert, Buğra Mert, Burcu Mert, Erkay Derelli, Eşref Oruç, Faik Uyar, Fatma Uyar, Güler Uyar, Meltem Derelli, Nafi Ege Mert, Nafi Mert, Nil Mert, Nuran Mert, Onur Uyar, Özlem Derelli, Semahat Oruç, Sibel Uyar, Ümit Uyar, Yücel Uyar.

Tuğçe Oruç

Contents

List of Figures

List of Tables

Abbreviations

3DEM	3 Dimensional Electron Microscopy
ACP	Acyl Carrier Protein
AT	Acyltransferase
CAPRI	Critical Assesment of Predicted Interactions
CASP	Critical Assessment of Techniques for Protein Structure Prediction
DCA	Direct Coupling Analysis
DEBS	6-deoxyeryttomycin B synthase
DH	Dehydratase
ER	Enoyl Reductase
FAS	Fatty Acid Synthase
KAL	Ketosynthase - Acyltransferase Linker
KR	Ketoreductase
KRc	Ketoreductase catalytic region
KRs	Ketoreductase Structural Integrity Region
KS	Ketosynthase
mfDCA	Mean-Field Approximation
MI	Mutual Information
mpDCA	Message-Passing Algorithm
NMR	Nuclear Magnetic Resonance
PAL1	Post-AT Linker 1
PAL2c	Post-AT Linker 2 Conserved Region
PAL2nc	Post-AT Linker 2 Non-Conserved Region

PDB	Protein Data Bank
PKSs	Polyketide Synthases
plmDCA	Pseudolikelihood Maximization Direct Coupling Analysis
PSICOV	Protein Sparse Inverse Covariance
TE	Thioesterase

Chapter 1

Introduction

1.1 Proteins

Proteins are an extremely important building block of life and they are responsible for most of the functions in cells including catalytic activities, DNA replication, transcription, translation, signalling, and the structural integrity of the cell. Proteins are made of twenty types of amino acids. All of the amino acids have a backbone consisting of an amine (-NH_2) and a carboxyl (-COOH) group attached to a central carbon atom (C_α). Additionally, a side chain group (-R) is attached to (C_α) providing the diversity between the amino acid types (Fig. 1.1). An amino acid chain (i.e. polypeptide chain) should fold into a three dimensional (3D) structure in order to be functional. In the folding process, the polypeptide chain first forms a secondary structure that can be α- helix or β- sheet via hydrogen bonding between the backbone atoms (Fig. 1.2). Further interactions between the amino acids (via electrostatic interactions, salt bridges, disulfate bonds etc.) provide additional folding leading to the tertiary structure. Although some proteins can be functional in the tertiary state, most of the proteins interact with other folded amino acid chains to form a quarternary state (Fig. 1.2).

Proteins consist of domains that are separately evolved from the rest of the protein chain and independently fold into a 3D structure. Some proteins have only one domain whereas most of them have multiple domains. Predictions to determine the number of single and multi-domain proteins have been performed for at least two decades (Chothia 2003; Ekman *et al.*

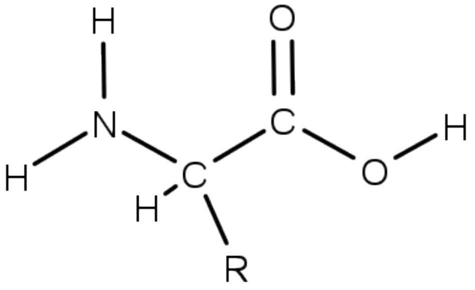


Figure 1.1: Structure of an amino acid. *There are 20 types of amino acids and all of them have a backbone consists of an amine (-NH_2) and a carboxyl (-COOH) group attached to a central carbon atom (C_α). Additionally, a side chain group (-R) is attached to (C_α) providing the diversity between the amino acid types.*

2005). Although estimates vary slightly, the calculated number of multi-domain proteins in prokaryotes is around 40%, which increases up to 65% in eukaryotes (Ekman *et al.* 2005).

Since the structure of a protein determines its function, it is critical to determine the structure of a protein. Many experimental techniques have been developed to determine the structure of the proteins. X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy are two commonly used methods. While X-ray crystallography provides high resolution (up to smaller than 1 Å), crystallization of a structure is experimentally challenging and costly, and it is particularly difficult for large protein complexes. On the other hand, while NMR spectroscopy is not as challenging as X-ray crystallography and determines the structure in solution (no need for crystalization), it is limited to smaller proteins. As of August 2020, 90% of all protein structures in the PDB are determined via X-ray crystallography and 7% via NMR spectroscopy. The remaining 3% are determined via electron microscopy and other methods (Berman 2000).

3D Electron Microscopy (3DEM) allows determination of protein structures with larger sizes. Although the method was not very successful in determination of the structures in better resolutions, recent developments improved the resolution quality (Malhotra *et al.* 2019), and the number of 3DEM structure releases in the PDB (Berman 2000) exceeded the number of NMR structure releases since 2016. Although very promising improvements have been achieved, structure determination for mega-Dalton sized proteins are still limited. For example, there are 969 structures in the PDB whose size is larger than 2 MDa; however only one of them has the

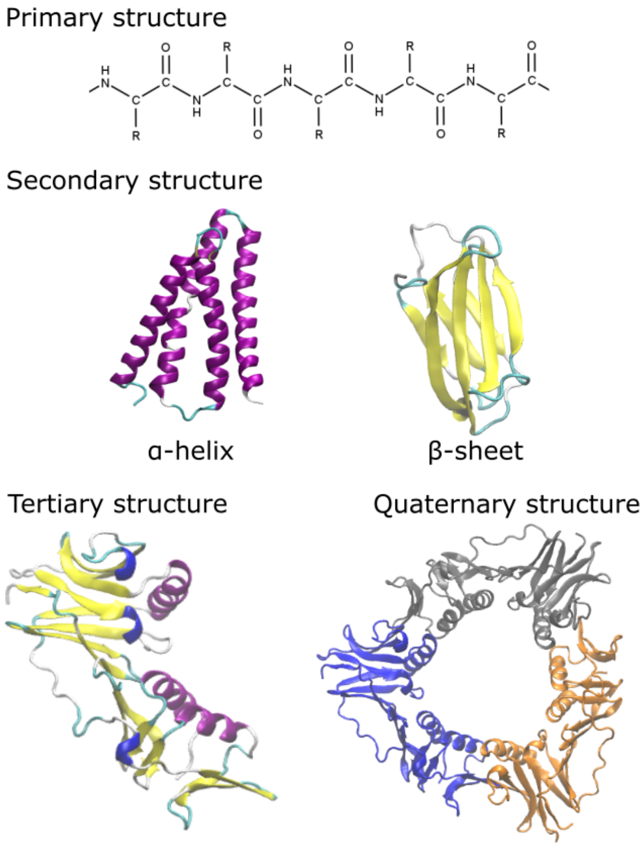


Figure 1.2: Folding of a polypeptide chain into the 3D structure. *The polypeptide chain first forms a secondary structure that can be α- helix or β- sheet via hydrogen bonding between the backbone atoms. Further interactions between the amino acids (via electrostatic interactions, salt bridges, disulfate bonds etc.) provide additional folding leading to the tertiary structure. Although some proteins can be functional in the tertiary state, most of the proteins interact with other folded amino acid chains to form a quarternary state.*

resolution smaller than 2 Å (PDB ID: 6e9d), indicating how challenging it is to determine the structures of large proteins with high quality of resolution.

Small-angle X-ray scattering (SAXS) is also used to determine structures of large proteins in solution; however, its resolution is low (Kikhney and Svergun 2015). The resolution of the structure can be improved by using structural information obtained from other methods (X-ray crystallography, NMR) similar to 3DEM.

These limiations in experimental methods directed researchers to develop computational approaches for protein structure prediction methods. Predictions of structures of individual domains are important for fast determination of the structures; however, due to the challenges of the current experimental techniques, it is critical for mega-sized protein complexes.

The computational methods can be divided roughly into two approaches. Template-based methods, which require known experimental structures, can result in very accurate predictions especially when the sequence similarity between the target and the template structures are high. For the proteins without any structural template, template-free approaches have been developed. The aim is to predict the structure of a protein from a solely amino acid sequence without using any template or fold similarity information.

Since it is a very important task to predict the structures accurately, the Critical Assessment of Techniques for Protein Structure Prediction (CASP) competition has been going on to evaluate the success of the predicted structures since 1994 (Moult *et al.* 1995). Many groups from all around the world participate in this competition for evaluation of their methods for protein structure prediction. It is a biennial, double-blind competition where the structures of the target proteins either have not been solved previously or have not been published; therefore, the predictors cannot know the structures of the target proteins. Hence, the success of the predictions only depends on the methodology that researchers developed. As all participants have been working on predicting the same set of target proteins, the competition ranks the methods revealing what kind of developments in the methodologies provide improvements in the area. These properties of the competition put CASP into a central role in the protein structure prediction challenge.

This chapter will begin with a summary of different computational methods for protein structure prediction including the milestone developments in the area with a specific emphasis on the coevolution-dependent approaches. Later, the methods for predicting the structure of multi-domain complex will be explained with details of developments and gaps in the field, highlighting the importance of the detection of coevolved residue pairs for the successful prediction of the domain-domain interactions. Then, detection of coevolved residue groups for identification of protein regions evolved for distinct functions will be explained, and their application area will be demonstrating the feasibility of its application on multi-domain proteins. Lastly, a multi-domain protein complex class, polyketide synthases, is explained in detailed, which will be used as a model system for the following chapters.

1.2 Computational Methodologies for Protein Structure Prediction

1.2.1 Template-Based Methods

1.2.1.1 Comparative (Homology) modelling

Comparative (homology) modelling is used to predict the 3D structure of a protein (target) based on a template structure that is homologous to the target protein. For the prediction of a structure, the first step is the detection of homologous structures from the PDB. This can be performed by similarity searching tool, like BLAST (Altschul *et al.* 1990), which can search the PDB for matching sequences or sequence fragments. If the identity is higher than 30 %, then the structure is selected to be used as a template (Xiang 2006). After alignment of the template sequence and the target sequence, a model is constructed. As the last step, the model is evaluated in order to check for errors (Eswar *et al.* 2007). The success of comparative modelling has increased with the development of better alignment algorithms that provided more accurate template selection and sequence alignment. Moreover, since the number of structures has been increasing, there are more templates available covering a greater range of patterns (KC 2016). MODELLER (Eswar *et al.* 2006), SWISS-MODEL (Guex and Peitsch 1997) are successful and widely used examples for homology modeling tools.

In the first decade of CASP experiments, predictions showed promising improvement due to the successful progress in alignment methods since the accuracy of comparative modelling is mainly based on the success of the alignment. On the other hand, in the second decade of CASP, this acceleration has disappeared. The misaligned regions of targets constitute no more than 15% of alignable regions so obtaining further improvements is more challenging via this approach. Although comparative modelling is in a stationary phase, the best predictions were obtained by homology modelling when there is a good template in recent CASP rounds (Moult *et al.* 2016; Croll *et al.* 2019).

The success of homology modelling depends on the identity of the target sequence with the homologous model. If the match between the sequences is higher than 40%, root-mean-square deviation (RMSD) of 90% of main-chain atoms is up to 1 Å. When the identity between the sequences is about 30-40%, alignment becomes challenging. While 80% of main-chain atoms RMSD can be less than 3.5 Å, it is higher for the rest. For the sequences with lower than 30% identity, finding homologous structures becomes difficult, which results in worse predictions (Xiang 2006).

1.2.1.2 Fold Recognition modelling (Threading)

When there is no high similarity between the target sequence and the sequences of the proteins in the PDB (sequence identity < 30%), alternative methods should be applied to those proteins to predict their structures (Khor *et al.* 2015). One approach to predict the structures of these proteins is *ab-initio* modelling, which is explained in detailed in the following section. The other approach is using similar folds or motifs of proteins as a structural template, rather than using the whole protein structure. This method is called as fold recognition modelling or threading.

The idea behind threading is placing (i.e. threading) the amino acids of a query sequence along the positions of amino acids on the structures of target proteins, and selecting template structures based on the quality of fit the by a scoring function. In other words, with this approach target proteins are searched to detect whether they have similar folds to the query sequence. That's why this approach is also called fold recognition. Using fold similarity to predict structures is based on the idea that there are a limited number of folds in nature (Chothia 1992; Zhang and Skolnick 2005; Chakraborty *et al.* 2017).

Originally the threading process was one where the sequence of the query protein was threaded through all proteins in the database and the ones with the best score, as judged by a statistical potential, were selected (Jones *et al.* 1992), although this idea is now often conflated with broader techniques for fold recognition, as described below. Similar to homology modelling, the critical step is to be able to find good templates, which requires successful alignments evaluated by accurate scoring functions (Khor *et al.* 2015). A variety of alignment scores including secondary structure match, sequence profile-profile alignment, sequence-structural profile alignments, hidden-Markov models (HMMs) as well as deep learning-based algorithms have been used to find precise query-template matches. For example, I-TASSER (Iterative Threading ASSEmbly Refinement) (Roy *et al.* 2010) is a widely used tool for fold recognition based protein structure prediction. To select the template structures, I-TASSER uses a program called LOMETS (Wu and Zhang 2007), which combines six different profile based threading algorithms and five different deep learning-based threading methods leading to better templates compared to any single method alone (Wu and Zhang 2007; Zheng *et al.* 2019). After selection of templates, I-TASSER generates full-length structures by iterative fragment assembly simulations (Yang and Zhang 2015).

1.2.2 Template-Free (*ab initio*, *de novo*) modelling

Template-free modelling aims to predict the 3D structure of a protein from only its amino acid sequence. Since template-based approaches need experimentally determined 3D structures, its application is limited to the presence of template structures. In the absence of a proper protein template, *ab initio* structure prediction is a unique way to predict the structure of a target sequence.

There are several approaches to predict protein structure from sequence. They can be classified as physics-based, fragment-based and covariance-based methods.

1.2.2.1 Physics based de novo structure prediction

All atom molecular dynamics (MD) simulations are used to understand the dynamics of biological molecules and their interactions at atomic resolution. Therefore, in theory, molecular dynamics simulations should properly provide native structures of proteins. However, there are some limitations to the application of MD simulations to predict protein structure. The first one is the computational cost of modelling. Since it takes up to 100 μs for the fast proteins to fold (Lindorff-Larsen *et al.* 2011), simulating folding processes is computationally costly. In order to be able to simulate these processes, the Shaw group developed a supercomputer - named Anton- with a chip architecture specially designed for fast MD simulations (Shaw *et al.* 2008; Shaw *et al.* 2014). The advanced version of Anton (Anton 2 (Shaw *et al.* 2014)) is capable of simulating a system with 10,000 atoms at a rate of 100 μs/day with 512 nodes; whereas the rate decreases to $\sim$ 15 μs/day when the system size increases up to 1,000,000 atoms (Shaw *et al.* 2014). The Shaw group demonstrated that simulation of the folding of ubiquitin, a 76 residue protein, which folds comparatively slow experimentally (Piana *et al.* 2013) with Anton took 3 ms of simulated time to fold, which is consistent with the experimental time scale. If this study were performed with Anton2, it would be expected to catch one folding event (3 ms) in slightly less than 30 days (with a system of around 9,000 atoms), using 512 nodes. Although it is not a very long time scale for the detailed study of protein folding, it is not feasible when the aim is to predict the 3D structures of proteins.

In a recent study from the same group, restraints that were determined from the first conformer of the NMR structure were applied to interacting residues to decrease simulation time. With this approach, they could obtain more than an order of magnitude faster folding of the same protein resulting in less than 1 Å RMSD from the NMR structure (Raval *et al.* 2015).

Apart from being costly, the success of all-atom MD simulations highly depends on the selected force field. Therefore, the determination of the best force field for a specific system may require additional studies and optimizations.

In order to decrease the cost of the calculation, coarse-grained modelling methods have been developed. Coarse-grained protein models use different representations of amino acids in a sequence. The resolution of the representations for a residue varies from several united atoms representing different parts of an amino acid up to a single united atom for the whole residue. These models can use physics-based, statistic-based and structure-based models of force fields. Since coarse-grained models allow the simulations of longer time scales and larger systems compared to all-atom MD, they promise a better understanding of biosystems (Kmiecik *et al.* 2016). However, similar to all-atom MD, selection of "best" force field is the challenging step of this approach.

1.2.2.2 Fragment based *de novo* structure prediction

The main idea of this approach is the assumption of peptide fragments having a limited number of conformations. Therefore, for the purpose of structure prediction of a target sequence, one should select the correct peptide fragments and assemble them. The fragment libraries are created via known structures (from the PDB). The programs select various fragments and replace them with one another to sample all possible conformations. After scanning the conformation space, the fragments are sampled and the potential energies of the conformations are calculated to compare the structures. The success of the prediction not only depends on the performance of the tool that performs these processes, but also the fragment library used. The fragments should be determined with detailed consideration. If long fragments are used, it would decrease the computational cost since the number of fragments used for a target would decrease. On the other hand, it would also reduce the chance of finding the correct fragment from the database. In order to obtain the balance between accuracy and speed, intermediate fragments (up to 21 amino acid length) are used from the libraries (Wang *et al.* 2016). One limitation of this approach is its dependency on the structures in the database. It may not to be possible to find candidate fragments (because of misalignments and missing templates), which results in a drastic decrease in the success of the prediction (Wang *et al.* 2016). For fragment library generation, several algorithms including NNmake (Gront *et al.* 2011), HHfrag (Kalev and Habeck 2011), Flib (Oliveira *et al.* 2015), LRFragLib (Wang *et al.* 2016) have been developed. Similarly, for fragment assembly, there are various algorithms including Rosetta AbinitioRelax (Simons *et al.* 1997; Rohl *et al.* 2004), Bilevel and ILS (Kandathil *et al.* 2018) protocols.

1.2.2.3 Covariance based *de novo* structure prediction

Interactions between amino acids provide proteins have their 3D structure and the maintenance of these interactions is critical for proteins to sustain their functions properly. During the evolutionary course, when a mutation occurs at one residue position, complementary mutations should occur at the interacting positions in order to preserve the 3D structure of the protein and the function. This evolutionary pressure results in the coevolution of residue pairs or groups. Detection of these coevolving residues promises to predict interacting amino acids in a protein structure that - in theory - should allow us to build the overall 3D structure.

1.2.2.3.1 Generation of Multiple Sequence Alignment

For the detection of coevolved residues from the sequence of the target protein, the first step is to find proteins with similar sequences, i.e. homologous proteins, via sequence search tools like BLAST (Altschul *et al.* 1990), PSI-BLAST (Altschul 1997), HMMER3 (Eddy 2011), HHsuite (Zimmermann *et al.* 2018). BLAST (Basic Local Alignment Search Tool) looks for sequences in a sequence database that have local similarities to the query (i.e. input) sequence and determines the statistical significance of the match (Altschul *et al.* 1990). PSI-BLAST (Position-Specific Iterative Basic Local Alignment Search Tool) is an iterative sequence searching tool. The first round is the same as a BLAST search. However, from the first run of BLAST, PSI-BLAST generates a position-specific scoring matrix (PSSM) (i.e. sequence profile) from the MSA based on the frequencies of the amino acids at each position in the sequence alignment. PSI-BLAST uses this PSSM to detect similar sequences in the following round. The PSSM is updated in every iteration and the updated PSSM is used to detect homologous sequences in the following iteration(Altschul 1997). As PSSM based sequence searching uses information not only from a unique sequence (i.e query sequence) but also information of an MSA consisting of similar sequences it provides a more sensitive search and allows the detection of distantly homologous sequences (Altschul 1997).

HMMs (Hidden Markov Models) are generally better at find remote sequence homologues than the BLAST family of methods. Profile HMMs include position specific penalties for insertions and deletions besides amino acid substitions, resulting in improvements in sensitivity. HMMER3 uses a profile HMM to search a sequence database to detect homologous sequences (Eddy 2011). A simple position-independent scoring system based profile is generated from the query sequence and homologous sequences are searched in a sequence database (by phmmer). If an iterative search is desired, the profile HMM is updated with detected sequences in each iteration and the updated profile HMM is used in the following iteration (by jackhmmer) (Eddy 2011).

HHsearch and HHblits from HHsuite also uses profile HMMs to find homologous sequences, the latter performing an iterative search (Zimmermann *et al.* 2018). On the contrary to PSI-

BLAST and HMMER3, HHblits searches a profile HMM database to detect homologous sequences rather than searching a raw sequence database. Using HMM profiles for both the query and target proteins makes HHblits a more sensitive tool for detection of distant sequence homologues (Zimmermann *et al.* 2018; Steinegger *et al.* 2019).

Due to their sensitivity, Jackhmmer and HHblits are commonly used tools to generate MSAs, which is the first step for predicting contacts in covariance based *de novo* structure prediction. The precision of predicted contacts vary depending on the prediction method (like plmDCA, PSICOV, described below), the parameters used in alignment generation or the databases used (Skwark *et al.* 2013; Tetchner 2015). However, the difference is not clear enough to claim one tool is better than the other for contact prediction studies.

1.2.2.3.2 Detection of Coevolutionary Information

Many approaches and methodologies have been developed to detect coevolutionary information from MSAs. The aim is to be able to detect contacting residue pairs (whose distance between C_β atoms (C_α for glycine) is less than 8 Å). On the other hand, the coevolutionary analysis gives not only directly coupled residues but also indirect couplings between the residues. Detection of indirect couplings is an inevitable artefact of coevolutionary analysis; since, for example, when residue A is contacting to residue B and when residue B is contacting to residue C, the only direct couplings between these three residues are between A-B and B-C. In this case, residues A-C have an indirect coupling through residue B, which is also detected in coevolutionary patterns. The problem is, when we cannot identify the real directly coupled residues, we introduce indirect couplings as direct couplings resulting in unsuccessful protein structure predictions.

Mutual information (MI) between the columns of the alignment has been used to determine the correlation. However, the MI approach could not eliminate indirect couplings. In order to overcome this problem global statistical models such as direct coupling analysis (DCA) (Weigt *et al.* 2008) and protein sparse inverse covariance (PSICOV) (Jones *et al.* 2011) approaches have been developed. DCA uses a maximum entropy method and the first reported version used a message-passing algorithm (mpDCA) (Weigt *et al.* 2008). mpDCA was computationally costly because it was using an iterative approach. As an alternative, a mean-field approximation was introduced to DCA (mfDCA) providing 103 -104 times faster calculations. Therefore it could be applied to longer sequences (Morcos *et al.* 2011). Afterwards, psuedolikelihood maximization was developed and applied to DCA (plmDCA) resulting in improvements in the accuracy of contact predictions (Ekeberg *et al.* 2013).

Apart from plmDCA, GREMLIN also uses a pseudolikelihood method and improves the predictions of plmDCA (Kamisetty *et al.* 2013). Comparison of the prediction with plmDCA, PSICOV and MI shows GREMLIN provides better predictions than all others while MI gives the worst. FreeContact implements EVfold-mfDCA and PSICOV that provides faster calculation but struggles to obtain accurate results as plmDCA and GREMLIN (Kaján *et al.* 2014; Seemayer *et al.* 2014). Another approach that uses pseudolikelihood maximization is CCMpred, which is faster than plmDCA, GREMLIN and PSICOV while as accurate as plmDCA and GREMLIN and better than PSICOV (Seemayer *et al.* 2014).

1.2.2.3.3 Progress in CASP competitions

Improvements in the contact prediction accuracies led to a leap in the success of *ab initio* structure predictions in CASP competitions (Schaarschmidt *et al.* 2018). Although the prediction of coevolved residue pairs to predict protein structures has been studied for more than two decades, the first encouraging results were obtained in CASP10 (2012). Removal of the indirect interactions was the reason for this improvement. Further success was obtained when the machine learning algorithms, using coevolutionary information as input, were implemented to predict the contacting pairs by MetaPSICOV (Jones *et al.* 2015) in CASP11 (2014) (Fig. 1.3). The average precision of top L/5 pairs (where L is the length of the sequence of a target protein) of long range contacts increased from 22% to 27% from CASP10 to CASP11, and it was acknowledged as a great success and oriented the attention to the implementation of the artificial neural networks for *ab-intio* structure prediction challenge. In CASP12, 26 of 32 groups performed better than the best performing group in CASP11 with reaching up to 47% average precision and among them, RaptorX (Wang *et al.* 2017) achieved the highest score. On the contrary of the results of CASP11, where there is a large difference between the top predictor (MetaPSICOV) and the second-best, in CASP12 top predictors have similar average precision on the target proteins.

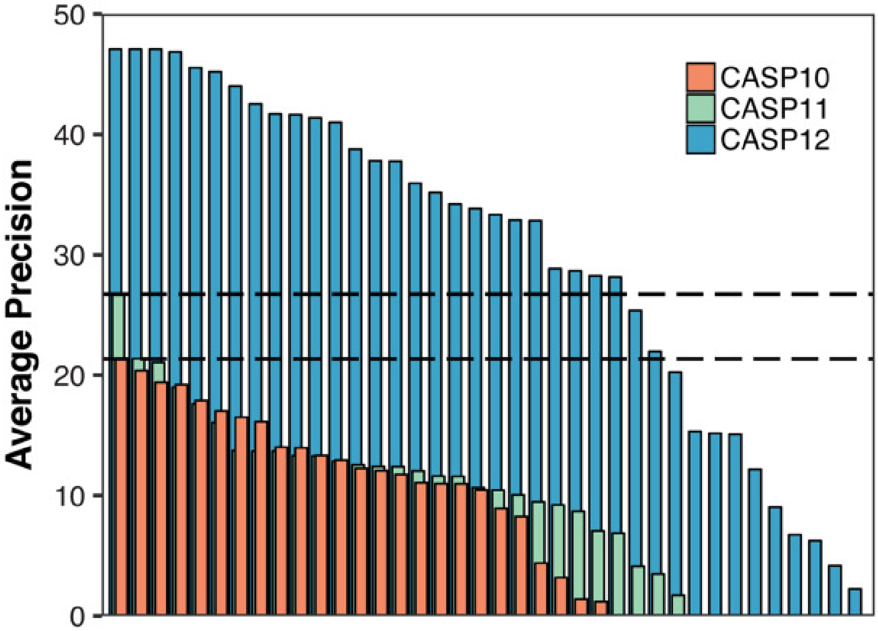


Figure 1.3: The precision of contact predictions in CASP12 jumped compared to predictions of CASP10 and CASP11 thanks to the developments in sequence databases and machine learning approaches. *Each bar represents the average precision score for the top L/5 residue pairs of long range contacts of a group, ranked along the x-axis from most successful to least successful group for each CASP competition (CASP10 (red), CASP11 (green) and CASP12 (blue)). Black dashed lines show the highest score for the corresponding competition. Figure is taken from ref. (Schaarschmidt* et al. *2018), for the permission to use, please see Appendix B.*

Improvements in the contact prediction has continued in CASP13 with another great jump in the average prediction (Shrestha *et al.* 2019; Kryshtafovych *et al.* 2019) (Fig. 1.4). Most of the participants performed well in contact prediction in this round of CASP where 16 of them outperformed the highest score of the previous round. Similar to the CASP12, Jinbo Xu's group achieved the highest contact prediction score with RaptorX, in CASP13 (Xu 2019).

Although great achievements have been obtained in the contact prediction accuracy, the big jump in structure prediction has been obtained by DeepMind's AlphaFold (Senior *et al.* 2019; Senior *et al.* 2020) in CASP13 (Fig. 1.5). The reason for this leap was obtained thanks to the successful predictions of the distance potentials between the residue pairs, rather than a successful prediction of contacting amino acids solely. Although AlphaFold showed a great success in CASP13 thanks to the prediction of distances between the residue pairs successfully, the idea of predicting distances and using them to predict the structure of proteins is not a novel

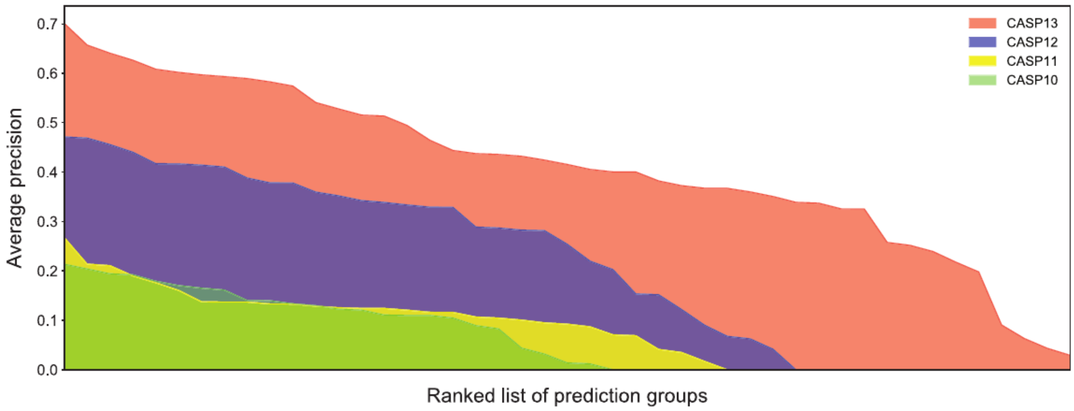


Figure 1.4: Contact predictions in CASP13 performed a further jump compared to predictions of previous CASP rounds thanks to impelementation of advanced neural networks. *Figure is taken from ref. (Shrestha* et al. *2019), for the permission to use, please see Appendix B.*

idea. Residue pairs were predicted previously by Pollastri et al., yet the implementation of these predictions to structure prediction was not that successful with a TM-score of 0.24 for the tested 9 CASP protein targets (Walsh *et al.* 2009; Kukic *et al.* 2014). The Jinbo Xu's group predicted inter-residue distance distribution for protein threading and showed that threading with distance information also helps to improve predictions on CASP12 free modelling targets (Zhu *et al.* 2018). To the best of my knowledge, we are one of the first groups who could predict the distances between the residues and obtained improvement in the predicted structure quality when distance predictions were implemented for *ab initio* structure prediction (Ji *et al.* 2019) (for details please see Chapter 2). Distances between residue pairs were later successfully predicted and implemented by other groups as well. For example, DMPfold was developed by David Jones and colleagues as an improved version on MetaPSICOV (Greener *et al.* 2019) and distance predictions were implemented to RaptorX by Jinbo Xu's group (Xu 2019).

1.2.2.3.4 Additional features used for *ab initio* structure prediction

In neural network architectures, in addition to the coevolutionary information, several other features, like predicted secondary structures, predicted solvent accessibility, statistical potentials, the effective number of sequences, amino acid composition, and Shannon entropy, are used to predict residue contacts (Jones *et al.* 2015; Wang *et al.* 2017) as well as in our protocols described in Chapters 2 and 3.

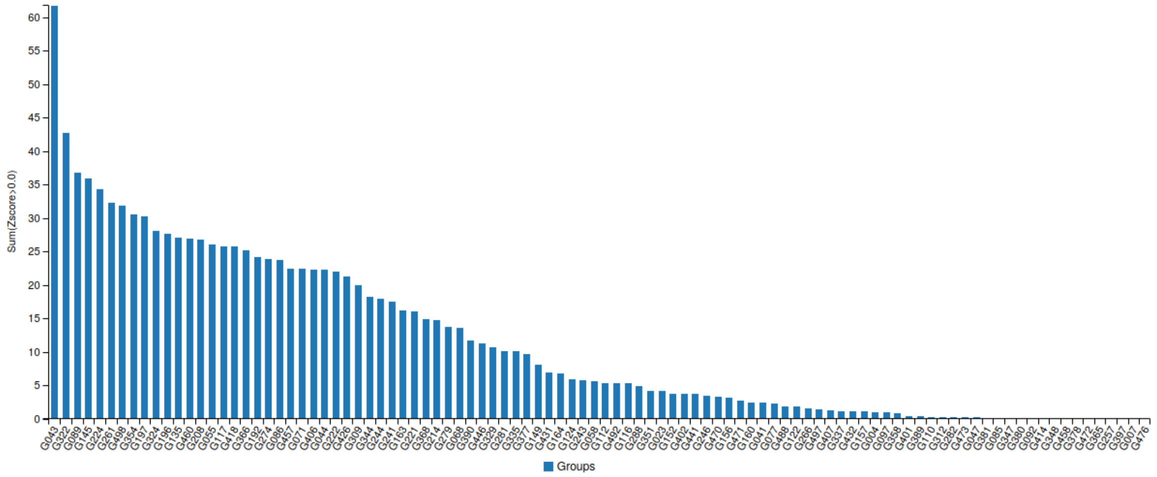


Figure 1.5: AlphaFold (G043) outperformed other predictors in free modeling catagory of CASP13 thanks to the successful predictions of the distances between the residue pairs and implementation of advanced neural networks. *Figure is taken from ref. (*Groups Analysis: zscores - CASP13*)*

Work to predict secondary structures has been ongoing for three decades as secondary structures provide information for functional annotation of proteins and also a step to predict the 3D structures (Zhu and Liu 2019). There are several successful secondary structure prediction tools like PSIPRED (Jones 1999), DeepCNF (Wang *et al.* 2016), SPIDER2 (Yang *et al.* 2016). For our predictions, we used SPIDER2, whose performance is ranked among top prediction tools on several benchmarks (Yang *et al.* 2016; Juan *et al.* 2020) with an high prediction speed (Juan *et al.* 2020).

Another feature that is fed to the ANNs are statistical potentials, which are also known as knowledge-based potentials. These are pairwise interaction scores for amino acid pairs generated from known 3D structures of proteins. In most of the approaches to determine a statistical potential, the residue pairs are assumed to be interacting and so taken into consideration if the distance between their side chains' center of mass is closer than a certain threshold. Generation of statistical potentials from known structures were first proposed by Tanaka and Scheraga (Tanaka and Scheraga 1976), and further expanded by Miyazawa and Jernigan (Miyazawa and Jernigan 1996). Many modificitions on the approach led to the generation of several statistical potentials. Betancourt and Thirumalai generated a potential matrix by rescaling Miyazawa and Jernigan's contact interaction giving hydrophobicities better aggreement with experiments (Betancourt and Thirumalai 2008). In MetaPSICOV, Jones et al. used the statistical potential of

Betancourt and Thirumalai, and for a given residue pair they calculated the weighted average (sequence weights) of the potentials based on the amino acid content of the corresponding MSA columns (Jones *et al.* 2015).

1.2.3 Implementation of machine learning algorithms for *ab initio* predictions

1.2.3.1 A brief introduction to machine learning

The working principle of machine learning algorithms is based on finding patterns in data and predicting the outcome with that detected patterns. Machine learning algorithms are used in many fields today; including finance, automatization, manufacturing as well as biology. In biology, the application of machine learning algorithms contains many fields as drug discovery, DNA sequencing, tumour detection.

Machine learning algorithms vary based on the type of data, availability of input and output and the approach to the task. For example, in supervised learning both input and output data are known and a model is trained to construct a pattern from input to output. Regression and classification are widely used types of supervised learning for discrete and continuous data, respectively. On the other hand, in unsupervised learning, there is no output data and the aim is to be able to find some patterns in the input data. Clustering, for example, is a type of unsupervised learning and the aim is to be able to cluster the input data points based on their "similarities". There are further types of learning algorithms exist in machine learning including semi-supervised learning, reinforcement learning; however, since I have used classification as a learning algorithm, details of the other types are beyond scope of this thesis.

One challenging point of using machine learning algorithms is to select the right approach for the specific task. While algorithm options can be narrowed based on data size and type, the most appropriate one can be selected with trial and error. Artificially neural networks (ANN) is one class of widely used machine learning algorithms. An ANN is composed of individual neurons and their interactions. A simple neuron is demonstrated in Fig. 1.6, left.

In a neuron, given information, input (p) is multiplied by a scalar weight (w) resulting a weighted input, wp (Fig. 1.6, left). When there is a bias, wp is added to the bias, b, resulting $wp + b$, which is called as net input, n. The net input, n, is acted on by a function called the transfer function, f, producing an output, a where $a = f(wp + b)$ (Fig. 1.6, right).

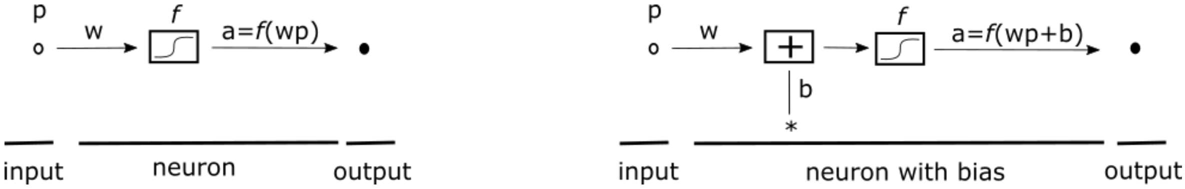


Figure 1.6: Structure of a single neuron with a scalar input without (left) and with (right) a bias. *An input (p) is multiplied by a scalar weight (w) resulting a weighted input, wp (left). When there is a bias, wp is added to the bias, b, resulting $wp + b$ (right).*

The overall point of ANNs is to optimize these w and b values based on the given data. In other words, the training process is a process of the adjustment of these parameters to obtain the given output. In the case where the input is not just a scalar but a vector, $\mathbf{p}$, then it is multiplied by a weight vector, $\mathbf{W}$, resulting $\mathbf{Wp} + b$ (Fig. 1.7).

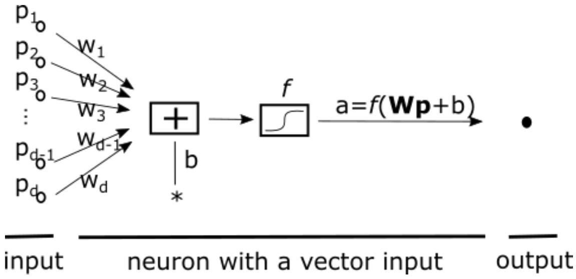


Figure 1.7: Structure of a single neuron with a vector input. *When an input is not just a scalar but a vector, $\mathbf{p}$, it is multipled by a weight vector, $\mathbf{W}$, resulting $\mathbf{Wp} + b$.*

Although these simplifications make it easier to understand the basic mechanism, neural networks used for training are much more complicated. A layer of neurons (also called as a hidden layer) contains multiple neurons as illustrated in Fig. 1.8. In the case where there is more than one hidden layer, the method is called a deep neural network (DNN). In Fig. 1.9, the structure of neural interactions with three hidden layers is shown. In this system, the first layer receives the input and after processing it produces an output, which is taken as input by the second layer and processed again to produce an output that is taken as an input for the third layer. The third layer processes it one more time to give the overall output. Deep learning is specifically used in image recognition applications including motion detection and face recognition (Demuth *et al.* 2000).

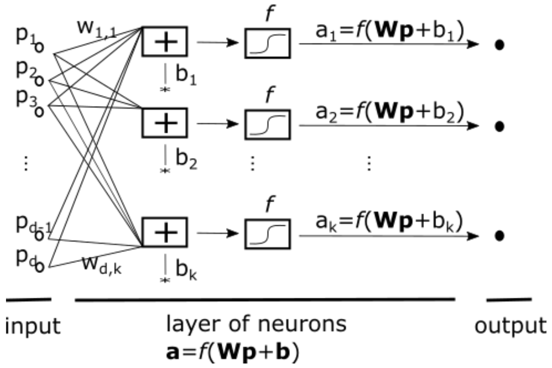


Figure 1.8: Neural network structure with multi-neurons in a network with one hidden layer.

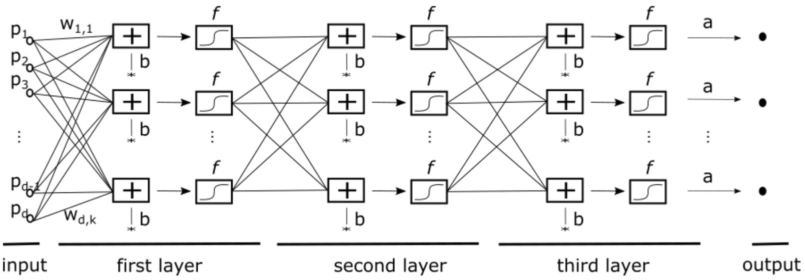


Figure 1.9: Neural network structure with multi-neurons in a network with three hidden layers.

As the output from one layer to the next layer is transferred via a transfer (or activation) function, one critical decision in a neural network is to be able to select a “proper” activation function for the specific network architecture and the task. An activation function decides whether to pass or not to pass the information from one layer to the following layer. Therefore an activation function generally just scales the output. There are different kind of activation functions, some of them are shown in Fig. 1.10. Each of them has some advantages and disadvantages. Therefore, finding the most “suitable” activation function for the specific settings of the neural network can be an additional task for training a neural network.

Whether it is a very simplistic network or a very complex DNN, the overall aim is to be able to find the best weights and biases as mentioned before. Based on the weights and biases an output is predicted and with a loss (error, cost) function the success of this prediction determined by comparing how close the prediction is to the target. For regression tasks mean squared

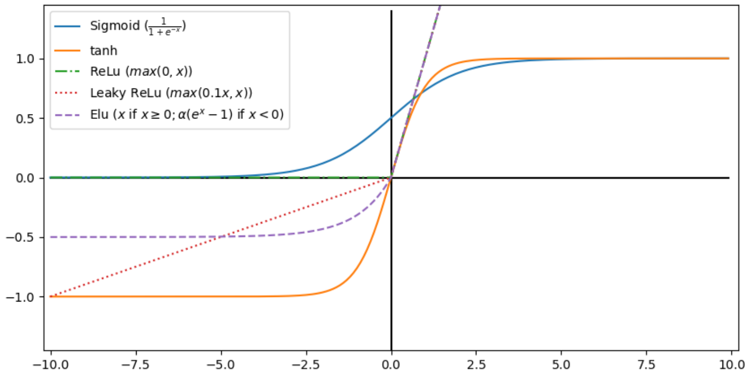


Figure 1.10: Examples of activation functions. *There are variety of activation functions, the most proper one should be selected based on the task.*

error, mean absolute error are widely used loss functions; for classification tasks cross-entropy is a widely used loss function. During the training of a model, weights and biases are aimed to be adjusted in a way to reduce the loss. During this process, the gradient of the loss function based on the weights is calculated with an algorithm called backpropagation. Backpropagation calculates how the loss function changes with respect to the weights; in other words, backpropagation calculates the gradients. Weights are modified based on this calculation in a way to find where the loss function gets the deepest step with an aim to find the local minimum. This is how neural networks learn. For this learning process - to optimize the weights and biases based on the backpropagation calculations - optimizers are used. The most common type of optimizer is gradient descent optimization and many variations of it have been applied to train a variety of neural networks.

For the application of these learning methods, there should be sets of training data, test data and validation data. Training set is used for the algorithms to adjust their parameters based on this set. In other words, it is the dataset network is trained on. Validation set is used to “try” the trained network on it to determine which adjusted parameters are better. The test set is used to apply this network on some independent data and determine the “success” of the trained network.

1.2.3.2 Neural networks used for protein structure prediction

The success of MetaPSICOV in CASP11 brought the attention to machine learning applications for protein structure predictions as mentioned in section 1.2.2.3. However, it is not the first study that uses a machine learning algorithm to predict protein structures. Elofsson and colleagues used random forest classifier to combine different coevolution based contact predictors and improve the prediction accuracies compared to their individual performances (Skwark *et al.* 2013). In that study, they showed that using both inverse covariance matrix estimation method (PSICOV) (Jones *et al.* 2012) and pseudolikelihood maximization with Potts model (plmDCA) (Ekeberg *et al.* 2013) increases the success of the prediction by 20% compared to the best individual method. In MetaPSICOV, Jones *et al.* used data from three different coevolution calculations (Jones *et al.* 2015). In addition to PSICOV and plmDCA (from CCMpred software (Seemayer *et al.* 2014)), they used the predictions of FreeContact (Kaján *et al.* 2014). Besides, they use additional properties including predicted secondary structure, solvent accessibility, Shannon entropy, mutual information and some additional features extracted from the alignments. As a machine learning algorithm, they used a feed-forward neural network with one hidden layer including only 55 neurons. Even with a shallow network test set's mean precision for the contact predictions were significantly higher than the contact predictions of individual methods whose results were used as input for the network.

Most of the top predicting tools in CASP12 used deep neural networks and coevolution data to predict contacts (Schaarschmidt *et al.* 2018). For example, RaptorX, the top-scoring predictor for contact prediction accuracies, uses more advanced deep neural networks (convolutional residual neural networks, explained in detail in Chapter 4) to predict contacting residue pairs rather than a 2-layer neural network as MetaPSICOV did in CASP11 (Wang *et al.* 2017). A subset of MetaPSICOV features was used in RaptorX; however, their predictions outperformed MetaPSICOV's predictions suggesting using a deeper network architecture improved the success of the predictions.

Although RaptorX became the winner in contact prediction in CASP13, in the structure prediction DeepMind's AlphaFold outperformed all other groups with an unignorable difference

(Fig. 1.5). Although the successful prediction of distance potentials, rather than predicting contacting pairs only, has a great role in this success, the part of very advanced neural networks cannot be underestimated (Senior *et al.* 2019; Senior *et al.* 2020).

Overall, it is seen that there are four main steps in improvements of *ab initio* protein structure prediction: (i) removal of indirect couplings, (ii) implementation of ANNs, (iii) implementation of deeper networks and (iv) distance predictions and further improvements in network architectures. Therefore it is clear that except for the first step, all milestones in the way of increasing the accuracy of protein structure prediction have been achieved with the successful implementation of the machine learning algorithms. This suggests that application of advanced network architectures and features extracted from the sequences and the sequence alignments not only can improve the prediction of average-sized proteins further but also promises to provide developments in other structural bioinformatics challenges like the prediction of the structures of large, multi-domain protein complexes.

1.3 Structure prediction of multi-domain protein complexes

While the approaches explained in the previous section have been used to predict the 3D structure of a single domain or single protein, the adjusted versions of them have been applied to predict multi-domain protein structures. This area is particularly important since experimental techniques struggle to determine the structure of large protein complexes while the number of multi-domain proteins in living organisms constitutes a large proportion (Ekman *et al.* 2005).

Domain-domain interaction problem to predict the overall structure of a multi-domain protein complex, or domain assembly problem, has been widely studied similar to the protein structure prediction problem. Two approaches have been developed to predict domain interfaces: (i) template-based methods and (ii) template-free methods. Template-based interface prediction methods, similar to template-based domain structure prediction methods, require to have structures of homologous protein complexes. The idea is if two protein complexes have similar sequence and structures, they would have similar domain interaction patterns and these interfaces can be used as a template to predict the unknown interfaces. PPI3D (Dapkunas *et al.* 2017), DEMO (Zhou *et al.* 2019), SWISS-MODEL (Waterhouse *et al.* 2018), InterPred (Mirabello and Wallner 2017) are successful examples of the tools that use template-based approaches for domain assembly.

Template-free (*ab initio, de novo*) approaches, on the other hand, aim to predict domain interfaces when no template structure is available. *Ab initio* methods generally approach to the problem as a docking problem; in which all possible interaction conformations are scanned and "the best" interface is chosen. However, the conformational space might be too broad and may require expensive computational sources. Therefore, reasonable restrictions to all possible interaction conformations have been applied. For that purpose, some researches focused on finding the correct domain interfaces by the searching the linker conformation space since for the domains on the same polypeptide chains, relative domain orientation is restricted by the conformation of the linker between two domains (Wollacott *et al.* 2006; Xu *et al.* 2015). Another method to restrict conformational space is to put restraints on the residue pairs, which are either determined with experimental methods or predicted with computational approaches. HADDOCK, for example, is one of the most successful docking tools, was developed to use experimental data to put restraints on residues for docking predictions to reduce the interaction conformational space. However, experimental data may be absent for most of the cases, where accurate computational predictions are needed for successful interface predictions. Many other tools allow users to put restrainst on the residues of two domains to reduce the space to span. One critical point in these approaches is to be able to implement correct restraints in order to get correct interfaces. Otherwise, these restraints would be forcing the domains to have wrong orientations with respect to each other. Therefore, many approaches have been developed to predict the interface surfaces or interacting residues pairs.

Interface residue predictions generally use conservation and/or coevolution information of the residues. Most recent tools that use conservation scores for interface prediction obtain conservation information from Shannon entropy or evolutionary rate and combine this information with additional features and implement to machine learning algorithms (Savojardo *et al.* 2017; Hou *et al.* 2017; Wang *et al.* 2019; Northey *et al.* 2018).

Similarly, coevolution information obtained from various approaches PSICOV, CCMpred are used as input to machine learning algorithms besides additional features extracted from the sequences (Hopf *et al.* 2014; Zeng *et al.* 2018).

Although many tools and approaches have been developed to predict interaction surfaces and interacting residues on a pair of domains or proteins, the success in *ab initio* prediction of single domain structures could not be achieved for the domain assembly challenge.

Similar to CASP, protein structure prediction competition, there is Critical Assessment of PRedicted Interactions (CAPRI) competition focusing on the successful predictions of protein complexes. CAPRI rounds are performed more frequently than CASP, and CASP rounds collaborate with CAPRI for multi-domain complex predictions since 2014. Although very encouraging improvements have been achieved in *ab initio* structure prediction, similar progress could not be observed for multimer predictions. In the last CASP/CAPRI (2018) competition, targets were classified into three: (i) easy targets for which template homologous models could be found for both the subunits and interfaces, (ii) medium targets for which template homologous structures could be found for the subunits but not for the interfaces, and (iii) difficult targets for which no template homologous information was available neither for the interface nor for the subunit structures. While the groups could make successful predictions for easy targets, the success of predictions for medium and hard targets was not very high. In other words, when a template for the interface available, successful prediction can be made, whereas for the targets without a template model, the interface cannot be predicted accurately. Another important point is among the participants, the ones with higher success rates are the ones with human intervention; whereas the ranking of the automated methods was similar. Overall, last CASP/CAPRI competition results suggest that predictions are not very successful when there is no template available and further improvements are needed in automation, suggesting further developments are needed for protein assembly/ domain interaction prediction area. (Dapkunas *et al.* 2017; Guzenko *et al.* 2019)

Although the progress in the domain assembly challenge is not as high as protein structure prediction area, the recent progress in the prediction of single-domain structures established motivation for further studies on the domain interface prediction problem. Implementation of advanced neural networks and prediction of distance potentials between the residues of domain pairs - which is studied in this thesis (chapter 3) - led to promising developments for the domain assembly problem.

1.4 Detection of coevolved residue groups in proteins

Another approach to detecting coevolved residues is using statistical coupling analysis (SCA) (Halabi *et al.* 2009; Rivoire *et al.* 2016). While with DCA and related methods the aim is to focus on minimal units of coevolution in order to detect interacting pairs, in the SCA the aim is to be able to detect global units of coevolution for a purpose to detect residue groups that are coevolved for a specific function (Cocco *et al.* 2013; Rivoire 2013; Rivoire *et al.* 2016). Therefore, while DCA and related methods are useful for predicting the structure of a protein/protein complex, for detection of functional residue networks, a more global perspective is needed, which can be detected with SCA.

In SCA analysis, a covariance matrix from the multiple sequence alignment is generated and this matrix is weighted with the conservation score of the amino acids. Top eigenmodes of this conservation-weighted covariance matrix give groups of residues that are coevolved together for a distinct function from the remaining coevolved residue groups. This approach has been applied to study a few protein families (Halabi *et al.* 2009; Rivoire *et al.* 2016; Narayanan *et al.* 2017).

In the first study by Ranganathan and colleagues, SCA was developed to see whether statistically coupled residues, obtained from evolutionary data, are related to each other thermodynamically (Lockless 1999). By mutational analysis, they showed that for the PDZ domain family, statistically coupled residues - regardless of their spatial distance to each other - are energetically coupled. Later, Halabi *et al.* detected multiple, independent residue networks, named sectors, that are evolved for a different function (Halabi *et al.* 2009). In serine protease protein family, they detected three sectors and with mutational studies, they showed that one of these sectors is functional for the thermal stability and the second one is responsible for the catalytic activity. In the following study, they overcame some methodological challenges and with a more straightforward protocol, they could detect different sectors on G-protein family, DHFR, β-lactamase as well as further investigation on serine protease enzyme family (Rivoire *et al.* 2016).

In order to avoid cross-coupling between the determined coevolved groups Rivorie et al. determined independent components, whose coupling with other residue groups are as less as possible (Rivoire *et al.* 2016).

In order to detect the independent components, covariation of amino acids along the sequence is calculated from the MSA as the first step:

$$C_{ij}^{ab} = f_{ij}^{ab} - f_i^a f_j^b \tag{1.1}$$

where f_i^a is the frequency of amino acid a at position i; similarly, f_j^b is the frequency of amino acid b at position j. f_{ij}^{ab} is their joint probability, which measures the probability of a and *b* being at positions *i* and *j* simultaneously.

Position-specific conservation is calculated by Kullback-Leibler relative entropy and covariance matrix (C_{ij}^{ab}) is weighted by the conservation score of each amino acid at each position with respect to the amino acid frequency, which is denoted by ϕ_i^a for amino acid *a* at position *i* and ϕ_j^b for amino acid *b* at position *j*.

$$\tilde{C}_{ij}^{ab} = \phi_i^a \phi_j^b C_{ij}^{ab} \tag{1.2}$$

Conservation weighted correlation matrix, $\tilde{C}_{ij}^{ab}$, is a four-dimensional matrix (LxLx20x20, where L is the length of the sequence) and it is compressed to a two-dimensional LxL matrix by taking the Frobenius norm. The resulting matrix $\tilde{C}_{ij}$ (or statistical coupling analysis (SCA) matrix) is used to determine the coevolved residue groups.

The matrix $\tilde{C}_{ij}$ gives the couplings between the positions of the amino acids but how can we learn which positions are more correlated to each other that can be useful to detect the residue groups evolved for a distinct function like catalytic activity, ligand binding etc.? In other words, can we group the amino acid positions which are coevolved together for a specific function?

To do that, we need to transform $\tilde{C}_{ij}$ in a way that can separate the residues as different groups and these residue groups should be as independent as possible from each other for proper detection of residue group-function relation. For this transformation, firstly, eigenvalue (spectral) decomposition is applied to the $\tilde{C}_{ij}$. This process decomposes $\tilde{C}_{ij}$ into three matrices:

$$\tilde{C} = \tilde{V}\tilde{\Delta}\tilde{V}^T \tag{1.3}$$

where $\tilde{V}$ is an eigenvector matrix and $\tilde{\Delta}$ is an eigenvalue matrix, which is a diagonal matrix bearing weight for each corresponding eigenvector. The matrix $\tilde{V}$ carries the information of the combination of the residue positions as scores. $\tilde{\Delta}$ contains the eigenvalues of $\tilde{V}$ that carries the information about the "importance" of the corresponding eigenvector. For example, low eigenvalue means an insignificant contribution of the corresponding eigenvector whereas high eigenvalue means a strong contribution of the corresponding eigenvector to the variance in $\tilde{C}_{ij}$. Therefore, the eigenvectors with high eigenvalues carry the information of a ”strong” relationship between the residue positions. This means these eigenvectors carry the information of residue positions that have a coevolutionary relationship. On the other hand, this separation of the residues into groups based on the eigenvalues and eigenvectors are unfortunately not successful enough since there can be strong coupling between the groups. To minimize this interaction between the groups, the significant eigenvectors (the eigenvectors with high weights, $\tilde{V}_{1 \cdots k^*}$) are transformed into new components, which are maximally independent from each other.

$$\tilde{V}^{p}_{1 \cdots k^*} = W\tilde{V}_{1 \cdots k^*} \tag{1.4}$$

where W is the transformation matrix. This mathematical transformation is known as independent component analysis (ICA) and resulting residue groups are called independent components (ICs) (Rivoire *et al.* 2016).

Detection of protein sectors by SCA on single-domain proteins have been used by other groups, too. More recently, Narayanan *et al.* applied this approach to study pancreatic-type ribonuclease superfamily and demonstrated one of the detected two sectors is responsible for thermal stability and the second sector residues are functional in catalytic activity. (Narayanan *et al.* 2017)

All these studies demonstrate that it is possible to detect the residue networks that are evolved separately for distinct functions from the sequence information alone. With this perspective, the application of this approach on multi-domain proteins can be a novel way to detect direct and indirect interactions within and between the domains of multi-domain proteins. Further, this analysis can detect functional residue groups in multi-domain proteins that can suggest new perspectives to experimental studies focusing on modification of protein complexes for the generation of novel products. With this motivation, I applied SCA analysis on a polyketide synthase, a multi-domain protein complex, to detect coevolved residue groups, which are explained further in Chapter 4.

1.5 Polyketide Synthases

Polyketides are secondary metabolites that are produced by many living organisms including bacteria, fungi, plant and some animal species. They are complex structures and have a huge role in pharmacology since they have many functions as an antibiotic, antifungal and antiparasitic drugs. They are produced by polyketide synthases (PKSs) that are extremely large proteins or protein complexes. Little is known about the higher-order structure of PKSs although many individual functional domains are known. Since PKS structures are massive, it is challenging to determine the structure of the whole PKSs with either experimentally or computationally. Despite all the challenges it is worth to put effort on the determination of the PKS structures and working mechanism since re-engineering the structures can overcome the limitations of the current drugs and provide production of new, target specific medicines.

1.5.1 Organization of Polyketide Synthases

PKSs are classified as type I, type II and type III PKSs (Shen 2003). Type I PKSs are large multifunctional proteins whose first reports were published in 1990. They can be either iterative or modular. Iterative type I PKSs produce aromatic and enediyne polyketides by using the same set of domains consecutively to elongate the extending unit. Modular type I PKSs consist of repeating unit of domain assembly and each module has a separate contribution to extending chain. In Figure 1.11a, 6-deoxyeryttomycin B synthase (DEBS) modular architecture is given as an example for type I PKSs. DEBS is one of the most commonly studied PKS and produces erythromycin A. One other PKS type known as type II PKS have multienzyme systems consisting of small distinct proteins for each catalytic activity and have an iterative function. These types of PKS generally produce aromatic polyketides. Tetracenomycin PKS is an example for type II PKS given in Figure 1.11b whose first report was published in 1999. Type III PKSs necessarily work iteratively and they do not use acyl carrier protein domains while PKS type I and type II does. In Figure 1.11c, RppA is given as an example of PKS type III. (Shen 2003) In this thesis, studies are focused on type I PKSs.

Each polypeptide chain in PKS type I is called a subunit and each subunit contains one or

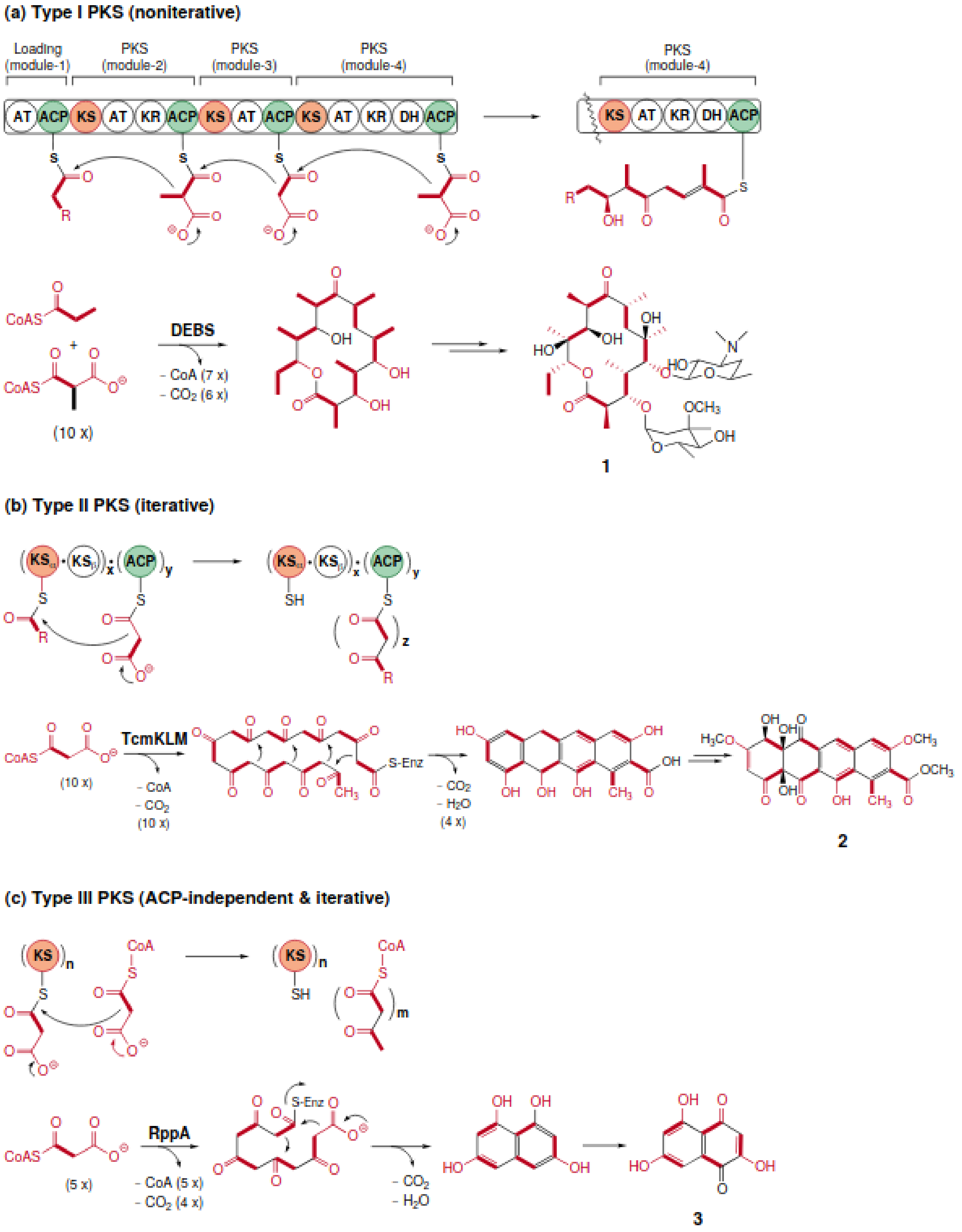


Figure 1.11: ***There are three types of PKSs.*** *DEBS is a type I PKS and a widely studied system producing erythromycin A. It has a modular structure consisting of one loading module and six elongation modules (a). Type II PKSs are multienzyme systems with small distinct proteins for each catalytic activity. Tetracenomycin PKS is an example of a type II PKSs (b). Type III PKSs are lack an acyl carrier protein domain and work iteratively. RppA is an example of a type III PKSs (c). Figure is taken from (Shen 2003), for permission to use, please see Appendix B.*

more modules. These modules are located in a way that the “tail” of the previous module interacts with the “head” of the following module. These interactions are provided by unstructured peptide linkers or docking domains. In each module intermediate product is elongated with or without modifications. Each elongation module contains a minimum of a ketosynthase (KS) domain, an acyltransferase (AT) domain and an acyl carrier protein (ACP) domain. The KS domain gets the intermediate product from the previous module, the AT domain chooses the proper extender unit from the cytoplasm and transfers it to the ACP. Chain elongation is performed by the KS domain via a Claisen condensation reaction between the intermediate product and the extender unit remaining on the ACP. A module may also include one or more modification domains like ketoreductase (KR), dehydratase (DH) and enoyl reductase (ER). At the C-terminus of the PKSs, there is a thioesterase (TE) domain that releases the final product and in some systems, it also performs macrocyclization (Robbins *et al.* 2016).

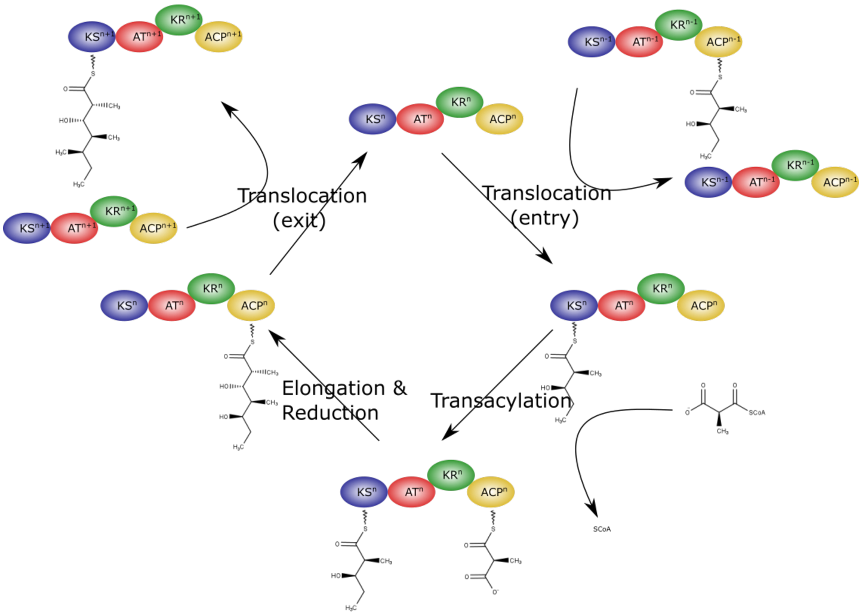


Figure 1.12: Schematic representation of turnstile mechanism. *Figure is plotted based on ref. (Bayly and Yadav 2017).*

While there is an extending intermediate in the module, a KS does not accept a new polyketide intermediate from the previous module. In Fig. 1.12, the so-called turnstile mechanism is shown. Although it is not clear how a KS ”senses” there is an extending chain in the module, interdomain linkers are the potential candidate in mediating the conformational changes leads

KS to ”sense” the information (Bayly and Yadav 2017). For a module with only KS, AT and ACP domains, the resulting chain has a β-keto group. In the case when a module has all three modification domains KR, DH and ER, elongating chain ends up with fully-reduced methylene as KR reduces β-keto group into hydroxyl group, DH reduces the hydroxyl group into a double bond and finally, ER reduces the double bond into a single bond. In Fig. 1.13, it is illustrated how each domain modifies the extending unit.

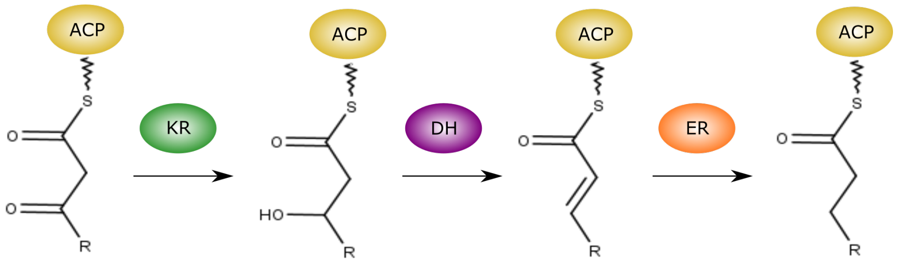
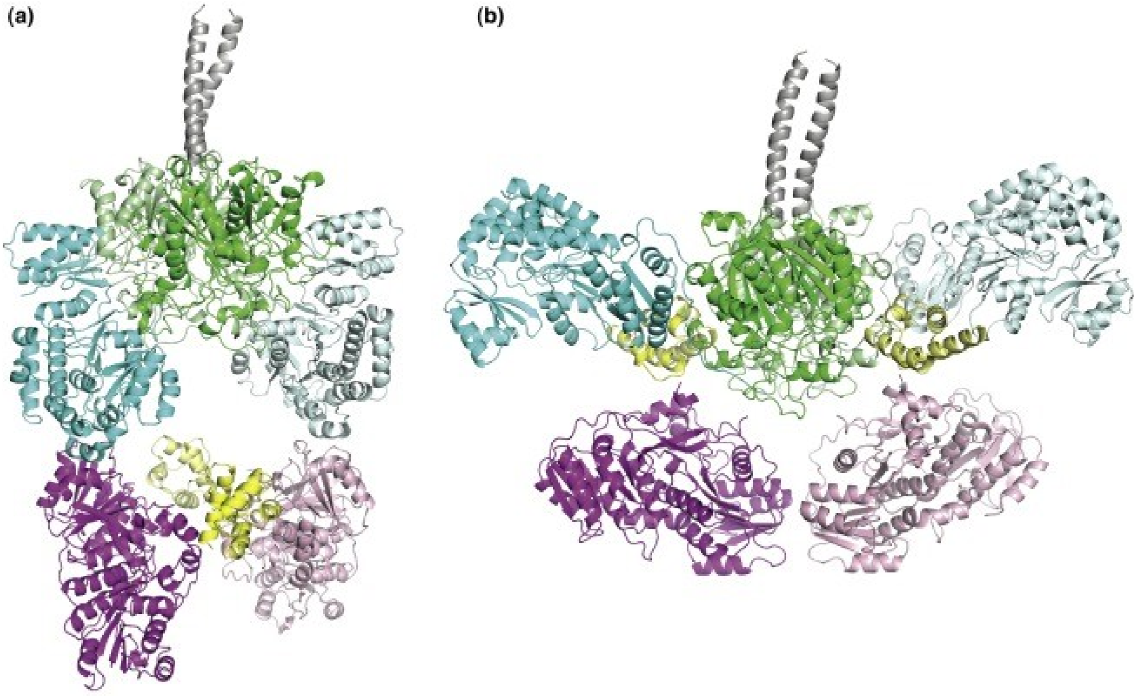


Figure 1.13: Modification on β-keto acyl-ACP by KR, DH and ER.

The AT domain of PKSs can either be within the module (*cis*-AT) or separate from the module (*trans*-AT). For *cis*-AT PKSs, the structure of the polyketide is predictable from the organization of the modules. In other words, there is a strong correlation between the sequence of the modules and the structure of the polyketide. On the other hand, this correlation is drastically low in *trans*-AT systems. *Trans*-AT systems include additional catalytic functions like β-branching and C-methyl transfer activities and the ER domains typically act in *trans*. They may also have nonribosomal peptide synthase (NRPS) modules that introduce amino acids into the intermediate product (Weissman 2015). Despite many challenges, there are plenty of studies to understand the structure and the mechanism of the PKS systems. It is hoped that solving the structure and the dynamics will provide new insights into protein engineering allowing modification of the current products or the production of novel ones.

1.5.2 Structural studies of PKSs

Since PKSs are extremely large, most structural studies focus on the determination of the domains individually. The largest complex determined experimentally is just one module of a PKS (Dutta *et al.* 2014; Edwards *et al.* 2014). Therefore, there is little information about the overall structure of the whole complex and how domains interact with each other. One of the widely studied PKSs is 6-Deoxyerythronolide B Synthase (DEBS). In Fig. 1.14b, the structure of DEBS module 3 is presented, which was determined via Small Angle X-ray Scattering (SAXS). During rigid-body modelling, KS-AT was treated as a unit e.g. the dimer retained its linear structure during the process. The KS domain (green and light green) forms a homodimer structure and KS-AT (AT domains are presented with blue and light blue color) didomain construct a nearly linear structure. This KS-AT structure is highly similar to vertebrate fatty acid synthase (FAS) (Edwards *et al.* 2014; Robbins *et al.* 2016).



Current Opinion in Structural Biology

Figure 1.14: Modular structure of polyketide synthases. *(a) Structure of module five of pikromycin synthase (determined via Cyrogenic Electron Microscopy, Cyro-EM), (b) structure of module three of DEBS (determined via Small Angle X-ray Scattering, SAXS). Although both modules have the same domain composition, the structures of pikromycin synthase module five and DEBS module three have drastic differences. KS domains are shown with dark and light green, AT domains are shown with dark and light blue, KR domains are shown with purple and pink, ACP domains are shown with dark and light yellow. Figure is taken from ref. (Robbins* et al. *2016), for permission to use, please see Appendix B.*

The structure of pikromycin synthase module 5 (PikAIII) was determined via Cryogenic Electron Microscopy (Cyro-EM) (Dutta *et al.* 2014) in 2014. The structure is different from the module of DEBS (Fig. 1.14). AT domains are bent downwards via rotation about 120° along KS-AT linker. The following KR domains complete a circle resulting in a hole (reaction chamber) at the center of the structure that provides a space for the ACP to move and reach other domains. Until this study, all proposed structures of PKSs and FASs included two reaction chambers (Fig. 1.14b). In contrast, there is only one reaction chamber in the PikAIII structure and the active sites of the domains face into the chamber. Additionally, ACPs were observed in two different locations (i) near the KR domain, (ii) near the AT domain. Since the dimerization element in the module locates at the C terminus, the two ACPs move together in the chamber. Even though the structures in Fig. 1.14 seem different, the reason can be using KS and AT not individually but as a dimer during rigid body modelling as discussed in ref. (Weissman 2015). Since SAXS data provided approximately 40 Å resolution, it is not known whether using these domains individually could give a structure similar to PikAIII structure. Although recent developments in electron microscopy have allowed higher resolution structures for larger proteins, no atomic resolution structure has been published for a module of a PKS at the time of writing this thesis.

1.5.3 Experimental studies on AT and KR domains to generate novel polyketides

Apart from structural approaches to understand the structure, dynamics and function of the PKS, many other experimental studies have been conducted to figure out the working mechanisms in order to be able to successfully manipulate the protein complexes for generation new polyketides as novel drug candidates.

AT and KR domains are frequently manipulated to produce new polyketides. Many experimental studies have been performed to modify the extender unit specificity of the AT domain or switch the KR domain sub-type (Fig. 1.15) (Weissman 2016; Barajas *et al.* 2017; Musiol-Kroll and Wohlleben 2018; Kornfuehrer and Eustáquio 2019). One common approach to change the specificity of that domain is to swap the native domain with an external corresponding domain from another PKS system or module. As another approach, site-specific mutations have been applied on certain residues that are thought critical for the specific function of the domain.

The challenging point of the domain swapping experiments is to sustain the interaction be-

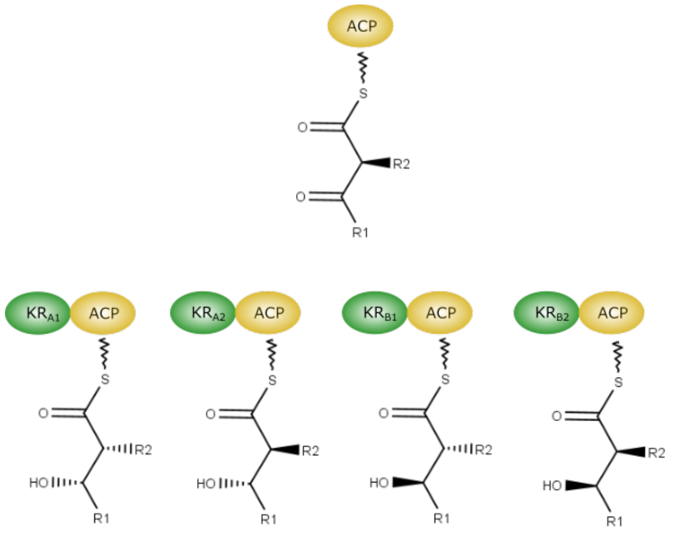


Figure 1.15: Different sub-types of KR domain produce intermediate products with different configurations.

tween the inserted domain and the remaining domains on the host system, to maintain structural and functional integrity. Early attempts for the AT domain swapping experiments were performed between erythromycin (DEBS) and rapamycin (RAPS) systems. The AT domain of the first module in a system where the TE domain was attached to the end of the first two modules of DEBS (DEBS1-TE) was replaced with the AT domain of the RAPS system second module. This swap resulted in successful switch from malonyl specificity to methylmalonyl specificity with almost no change in the product amount (Oliynyk *et al.* 1996). This result made domain swapping experiments a promising approach to generate new polyketides. On the other hand, not all AT swapping experiments were successful. When the full PKS system is used instead of DEBS1-TE, the product yield decreased significantly (Ruan *et al.* 1997). Confusing results from several experiments led researchers to think that when the overall PKS system is used, the position of the modified module in the system is important for the success of the experiment. When the modifications were performed on initial modules, the product yield decreased significantly (Ruan *et al.* 1997; Lau *et al.* 1999), yet swapping the last AT domain did not make any change in the product yield (Liu *et al.* 1997). However, it was later shown that swapping the AT domain of the fourth module provided a similar amount of erythromycin analog contrary to similar studies (Petkovic *et al.* 2003). The important point of that study was the fact that they used different domain boundaries for swapping. This suggested that for domain swapping experiments, domain boundaries should be optimized carefully to obtain similar amounts of product as the wild type. Recently, a comprehensive kinetic study with different domain boundaries on DEBS module 6 showed that optimized AT domain boundary works successfully in a variety of PKS systems (Yuzawa *et al.* 2017).

Another approach to change the extender unit specificity is to mutate certain residues on the AT domain via site-directed mutagenesis. Sequence analysis revealed that methylmalonyl specific AT domains bear YASH motif whereas malonyl specific AT domains have HAFH motif at the corresponding position in the sequence (Haydock *et al.* 1995). The targets for the modifications were selected on either the fingerprint motifs (YASH, HAFH), or other residues close to the catalytic site. Failure in switching extender unit specificity by mutating YASH/HAFH motifs led researchers to try to characterise additional residues likely to be funtional in extender unit binding. Koryakina *et al.* applied single and double mutations on DEBS-AT6 to convert methylmalonyl specificity to propargylmalonyl specificity (Koryakina *et al.* 2017). For that study, they worked both on only DEBS module 6, and bimodule of DEBS module 5 and 6. In the former experiment, conversion of the specificity was highly achieved by mutating YASH motif to RASH by a single mutation. In the latter experiment, the same mutation provided incorporation of the propargylmalonyl-CoA, as well; however, the amount of the product was significantly lower compared to the methylmalonyl-CoA incorporation with the same mutation. Kalkreuter *et al.* applied a similar approach to modules 5 and 6 of pikromycin polyketide synthase whose AT domains are specific for malonyl-CoA with low promiscuity to other extender units (Kalkreuter *et al.* 2019). With mutating only one module in bimodule systems, they were able to increase the incorporation of the propargylmalonyl-CoA up to 50% with a slight decrease in the relative activity. More recently, Zhang *et al.* studied on switching malonly/methlymalonyl/ethylmalonyl specificities on salinomycin PKS by gradually mutating four residues in addition to the fingerpront motif residues (Zhang *et al.* 2019). By these experiments they could incorporate targeted extender unit without losing the efficiency in catalytic activity only when they manupalted the residues at all six positions. On the other hand, the success of switching the extender unit decreased drastically when cognate holo-ACP was incubated with the AT domains. These experiments showed that switching the extender unit specificity from one type to another is possible only with successful characterization of the critical residues additional to the YASH/HAFH motif.

Similar to AT domain modifications, one widely applied approach on KR domains is to swap the reductive loop with the corresponding position from another module or PKS system. Kellenberger *et al.* swapped KR domain by using different domain boundaries in DEBS1-TE system where the host KR is A1 type (module 2) ("A Polylinker Approach to Reductive Loop Swaps in Modular Polyketide Synthases"). Although swapping the domain with A1 and B1 type KRs gave good results, swapping with A2 and B2 KRs did not produce detectable amounts of products. Additionally, despite different domain boundaries were used, more than one boundary were not tested on the same host-donor system; hence, it is hard to conclude on the optimized domain boundaries. On the other hand, one year later Valenzano *et al.* were able to swap a non-epimerazing KR to an epimerazing KR domain in DEBS module 6 (Valenzano *et al.* 2009). Another study swapped module 2 KR (type A1) with a range of type A2 and B2 KRs in order to test the incorporation of epimerization activity further ("Evaluating Ketoreductase Exchanges as a Means of Rationally Altering Polyketide Stereochemistry"). For A2 swapped systems, epimerized structures were able to be observed besides non-epimerized or non-reduced structures. Although their amount was not as high as the wild type product, it was significantly higher compared to the previous researches. On the other hand, type B2 KR swapping experiments were not that successful as only two out of six swaps yielded traces of the epimerized structure. Consistent with these results, another experiment performed on lipomycin PKS module 1 showed that swapping the A2 type host domain with A type donor KRs produced the expected product; yet when B-type KRs were used as a donor, expected products were unable to be detected (Eng *et al.* 2016).

Sequence analysis on KR domains revealed that LDD motif approximately 57 residues before the catalytic tyrosine is conserved (specifically the latter aspartic acid residue is strictly conserved) in type-B KR domains and this motif is absent in the type-A KR domains. Instead, type-A KR domains bear tryptophan eight residues before the catalytic tyrosine. It was also detected that three residues before the catalytic tyrosine, leucine, histidine, and glutamine residues are conserved in B2 type, A2 type and A1-B1 types of KRs, respectively. Site-directed mutagenesis studies on these motifs to switch the KR type on isolated DEBS modules 1 and 2, KR domains yielded a decreased amount of the expected product (Baerga-Ortiz *et al.* 2006). Nevertheless, the same approach applied to DEBS1-TE system did not provide any switch from the natural product towards expected product (Kwan *et al.* 2011). Later, Zheng *et al.*, successfully converted A1 type KR to A2 type KR on the second module of amphotericin PKS via mutating two residues detected by structural analysis (Zheng *et al.* 2013). The same group later showed that, with single or double mutations, they were successfully converted type B2 KR of DEBS module 1 into type A2 KR (Bailey *et al.* 2016). Mutations on the leucine (three residues before the catalytic tyrosine) to histidine and glutamine, which are conserved in A2 type and A1-B1 types respectively, did not lead any conversion. However, mutating this residue to alanine provided to obtain A2 type product. Moreover, with mutations in the same positions, they converted amphotericin KR2 (type A1) and tylosin KR1 (type B1) to type A2 KR. However, corresponding mutations on rifamycin KR7 (type A2) did not change the KR type. This suggested that for the reduction process, A2 type KR follows the most energetically favored pathway among these four types.

Although there is an intense effort to determine the structure of PKSs, even the most comprehensive studies include no more than one module. Even though there are improvements in experimental systems, it seems there is a long way (if it is not impossible) to determine the whole PKS structure experimentally. In this thesis, domain interactions are predicted via an approach that I developed by using coevolutionary information and deep neural networks to have furhter insights about relative orientations of the domains with respect to each other (Chapter 3) and statistical coupling analysis have been applied on the first module of DEBS to detect coevolved residue groups (Chapter 4).

1.6 Scope of the thesis

Understanding the structure and the function of multi-domain complexes has critical importance since they constitute a large proportion of all proteins. As their structural determination with experimental methods is highly time-consuming and expensive, when possible to figure out, computational approaches for determining the structures of multi-domain complexes and for understanding their working mechanism could provide a cheaper method for addressing the problem.

Analysis of coevolutionary patterns in multiple sequence alignments arise due to natural selection, to maintain important structure-function relationships. The hypothesis presented here is that coevolutionary information, extracted from sequence alignment, can be used to obtain information about the structure and function of multi-domain proteins, including information about the distance between residues within individual domains, and between domains, the boundaries of individual domains, modes of functional cooperation between residues in different domains and residues that confer critical functionality to the protein e.g. by determining substrate specificity. This thesis shows developments in all these areas, improving and extending existing methods in the field and breaking ground by specifically applying them to multi-domain proteins.

This thesis includes four additional chapters. The second chapter of the thesis focuses on our contributions to *ab initio* protein structure prediction problem, where deep neural networks - with coevolution data - were used to predict the distances between the residue pairs resulting in improvements in the quality of the predicted structures. In the third chapter, deep neural networks fed by the features extracted from sequence information, including predicted coevolved residue pairs, are used to predict distance potentials between the residue pairs of the domain pairs. It will be demonstrated that the successful predictions of distance potentials provide successful predictions of the interface between two interacting domains. In the fourth chapter, coevolved residue groups in a multi-domain protein complex are detected via statistical coupling analysis. It will be shown that, by this analysis, domain boundaries and residue groups that are functional in sub-type specification can be detected. In the last chapter, the results and the contribution to the literature will be summarized.

Chapter 2

Ab-initio prediction of protein domain structure

2.1 Overview of the Chapter

Protein structure prediction has been one of the most important and difficult challenges of biology. Knowing the structure of a protein is essential for better understanding its function, and performing successful experiments for protein engineering. Therefore, many groups have been working on the prediction of protein structures for decades as explained in Section 1.2. Here, in this study, we worked on the *ab-initio* protein structure prediction problem. In the covariance-based *ab-initio* protein structure prediction problem, the main approach is to be able to predict the contacting residue pairs (distance between the C_β, C_α for glycine, atoms of two residues is less than 8 Å) to predict the experimental structure of a protein (Jones *et al.* 2015; Wang *et al.* 2017; Schaarschmidt *et al.* 2018). In this study, in addition to predicting contacting residues, we aimed to predict residue pairs whose distance is larger than 8 Å.

At the point that the work in this chapter was started, there were few people in the field who had tried to predict distances and none had shown significant benefits (Walsh *et al.* 2009; Kukic *et al.* 2014). There are more recent studies demonstrating the successful distance predictions and their contributions to the structure prediction (Xu 2019; Senior *et al.* 2019; Senior *et al.* 2020), and we are one of the first groups to show that the successful prediction of distances between residue pairs improves structure prediction accuracy for *ab initio* structure prediction. Besides, we show that implementation of distance predictions leads to select better models without using any additional model assessment tool.

In the following sections, I describe the methodology we followed for neural network training and testing, and generating the structures by using predicted contacts and distances. Then, I present the results demonstrating the effect of distance predictions on structure prediction quality, analysis on training and test protein structures belonging to the same topology class, the impact of increasing the structure pool size on the selected final model, and amino acid pair type distributions on the successful predictions. Lastly, I discuss the importance of our results and the contribution of this study to the literature.

This study was published in PLoS One in January 2019 (Ji *et al.* 2019). I have an equal contribution in this study with Shuangxi Ji, and contributions of the other authors are cited in the following sections.

2.2 Methods

2.2.1 Neural network architecture and parameters

As the input for the neural network, a feature vector was generated for all residue pairs of the training and the test set proteins with information extracted and predicted from the target sequence and the alignment containing homologous sequences (multiple sequence alignment, MSA). From the MSA, mutual information with the average product correlation (APC) (Dunn *et al.* 2007), normalized mutual information with APC, CCMPred (Seemayer *et al.* 2014), QUIC (Hsieh *et al.* 2014) and mfDCA (Morcos *et al.* 2011) were used to detect coevolved residue pairs. SPIDER2 (Yang *et al.* 2016) was used to predict the secondary structure and solvent accessibility. Statistical potential (Betancourt and Thirumalai 2008), the effective number of sequences, amino acid composition, and Shannon entropy were calculated by a script from the source code of MetaPSICOV (Jones *et al.* 2015).

Training and test sets included 1701 and 108 proteins, respectively. For each residue pair in a protein, a feature vector with 733 elements was generated, with features very similar to those used in MetaPSICOV feature vector (Jones *et al.* 2015). Details of the training and test set preparation as well as feature vector generation are explained in Appendix A.

For all networks, a simple nine-layer neural network architecture was used with one input layer, eight hidden layers and one output layer (Fig. A.2). The first hidden layer consist of 120, the second hidden layer consist of 50 and the remaning hidden layers consist of 30 neurons (Fig. A.2). Keras (Chollet 2015) with Tensorflow (Abadi *et al.* 2015) backend was used for training the models, stochastic gradient descent (SGD) was used as the optimizer, binary cross-entropy was used as loss function, SELU was used as the activation function between the layers except for the last hidden layer and the output layer where the sigmoid function was used. In order to avoid overfitting, the early-stopping algorithm was chosen with 20% of the data randomly assigned as the validation set. The maximum number of epochs was set to 300, patience was set to 40. The batch size was selected as 32.

Seven different neural networks with the same network architecture were trained. For the contact predictions (0 - 8 Å), four different networks with varying upper boundary, (0-7.9] Å, (0-8.0] Å, (0-8.1] Å and (0-8.2] Å, were trained. For each residue pair, the average score of four networks was calculated and used as the final network score. For each distance bin, one neural network was trained for the distance intervals of (8-13] Å, (13-18] Å, and (18-23] Å.

Feature vector generation, determination of the training and the test sets, initial optimization of the neural network architecture and the parameters of the network, implementation protocol of the contact and distance predictions into Rosetta for *ab initio* structure prediction were studied by Shuangxi Ji (Ji 2019) and the details are explained in Appendix A. Further optimization of the network architecture and the parameters were studied by Liam Mead (Mead 2018). Contact bin predictions were made by Liam Meads and distance bin predictions were made by me.

2.2.2 Reducing the effective number of sequences of test proteins

In order to study the effect of having fewer sequences in the alignment, some sequences were removed from the alignment of the test set proteins. To reduce the Nf value of the alignment (Nf: the number of sequences in the alignment, having maximum 80% identity to each other, divided by the square root of the protein length) of alignment, subsampled MSAs from the initial MSA were generated. For that, firstly the MSA was filtered with 80% identity threshold by HHfilter (Zimmermann *et al.* 2018). For a target Nf, Nf* $\sqrt{L}$ sequences, keeping the target sequence in the alignment, were randomly selected to generate a subsampled MSA, where L is the length of the sequence.

2.2.3 Generation of an additional test set

Although using 25% identity threshold allows us to ensure that we are not using similar sequences in the training and the test sets, from a structural perspective, it does not guarantee we are not using similar structures. The possibility of using similar structures raises the question, would having proteins with similar structures in the training set and the test set cause the neural network to learn the structural patterns and lead to a bias in the prediction accuracy.

To investigate this question, we analysed the training and the test set proteins based on their CATH database (Dawson *et al.* 2016) classification. The CATH database classifies protein structures in a hierarchical scheme. In the highest order, Class (C of CATH), protein structures are grouped based on their secondary structure: (i) α-helix only, (ii) β-sheet only, (iii) both $alpha$-helix and $beta$-sheet and (iv) little secondary structure. At the Architecture (A of CATH) level, proteins are further grouped based on the arrangements of their secondary structures. At the Topology/Folding (T of CATH) level, the further grouping is made based on how secondary structures are connected. And at the Homologous Superfamily (H of CATH) level, proteins are classified based on their evolutionary relation, i.e. grouped if there is any evidence that they are homologous (Dawson *et al.* 2016).

For analysing the training and test sets based on their structural similarity we used topology and homologous superfamily level similarities based on the CATH database downloaded in January 2018.

For further investigation of the effect of having test set proteins belonging to the same topology/homologous superfamily classes as the training set proteins, an additional test set whose proteins do not share any common topology with the training sets of DeepCDpred, MetaPSI-

COV and RaptorX, was generated. In CATH v4.0.0 database, there are 1391 topology classes and only 327 of them do not have a common topology with the proteins of the training sets. There are 12234 protein chains in the 327 topology classes. The PISCES server was used to cull the proteins to select the ones with maximum 25% sequence identity, maximum 2.5 Å resolution, and sequence length between 40 - 400 amino acids. From the remaining protein set, we removed the ones with missing residues or atoms in the structure, leaving us with 50 proteins. PDB IDs of this additional, non-topologous, test set proteins are given in Table 2.1.

Table 2.1: *PDB ID list of the test set with 50 proteins.*

1b12A	1ckmA	1d0qA	1dd9A	1dmgA
1e1hA	1e1hB	1g2rA	1hufA	1i71A
1inpA	1io1A	1j3aA	1o9iA	1okcA
1r7lA	1rajA	1sknP	1svbA	1tgrA
1w2yA	1whiA	1wjxA	1yrtA	1yu5X
1ywmA	2j7aC	2p84A	2rhkC	2vnlA
2wqiA	3bl9B	3bqwA	3girA	3hrdB
3o79A	3pn3A	3rioA	3rlfG	3ts2A
3vtoQ	3x02A	3x34A	4x8yA	4xb4A
4ymuC	4z6mA	5b66O	5hobA	5hocA

2.2.4 Expansion of the structure pool

For each test protein, initial 100 structures were generated by Rosetta as explained in Appendix A. An additional 100 structures, with the same set of parameters, were generated to see the effect of model selection and quality between 100 and 200 structure pools. The model with the lowest Rosetta score was selected as the final model from the structure pool. To evaluate the quality of the predicted structures, i.e. how close is the predicted structure to the experimental structure, we used TM-score (Xu and Zhang 2010). TM-score measures the similarity between two structures, score of > 0.5 means the predicted structure the experimental structure is very likely to have the same fold; whereas, a score of < 0.5 indicates the two structures do not have the same fold. To determine the “best” structure in a structure pool, we calculated the TM-score of all structures (by comparing them against the experimental structures of the proteins) and selected the one with the highest TM-score for analysis.

2.2.5 Determination of the proportion of amino acid types in correctly predicted residue pairs.

Amino acid types of the successfully predicted contacts and distances were analyzed to determine whether there is any bias in the types of residues that are successfully predicted in each of the four distance bins. Residue pairs with high network score (0.7) were analysed. The observed distribution (O) and expected distribution (E) as

$$E = \frac{\sum AB \in d}{\sum all\ residue\ pairs \in d} \tag{2.1}$$

$$O = \frac{\sum AB\ correctly\ predicted \in d}{\sum all\ residue\ pairs\ correctly\ predicted \in d}. \tag{2.2}$$

Here, AB is a given amino acid pair type and d is the distance bin where AB belongs, based on their distance in the 3D structure. For each pair type, AB, the mean O/E was determined over the 108 test set proteins.

We calculated the precision as the number of correctly predicted contacts/distances for that pair of amino acid types divided by the all contact/distance predictions for that pair.

$$precision = \frac{\sum_{correct} AB\ \in d}{\sum_{correct} AB\ \in d + \sum_{incorrect} AB\ \in d}. \tag{2.3}$$

Calculations were made on the residue pairs with high network score (0.7).

All scripts were written in Python (Van Rossum and Drake Jr 1995).

2.3 Results

2.3.1 Implementation of distance restraints to improve protein structure prediction accuracy.

For the trained neural network (nine-layered), the accuracies for the 108 protein test set were determined for the top L/10, L/5, L/4, L/3, L/2, L and 1.5L residue pairs (where L is the length of the protein sequence) (Fig. A.3). Although the prediction accuracies of DeepCDpred are better than MetaPSICOV and NeBcon, RaptorX contact prediction accuracies are better than DeepCDpred's.

Distance prediction accuracies of DeepCDpred show variation between the distance bins (Fig. 2.1). Bin 8-13 Å predictions have a very high accuracy, the top 1.5L prediction accuracy is even higher than the top 1.5L contact prediction accuracy of all methods (Fig. A.3). Bin 13-18 Å drops slightly yet the accuracy for top 1.5L residue pairs is similar to top 1.5L contact predictions of DeepCDpred. On the other hand, bin 18-23 Å accuracies drop drastically leading to ~ 60 % accuracy for top 1.5L residue pairs.

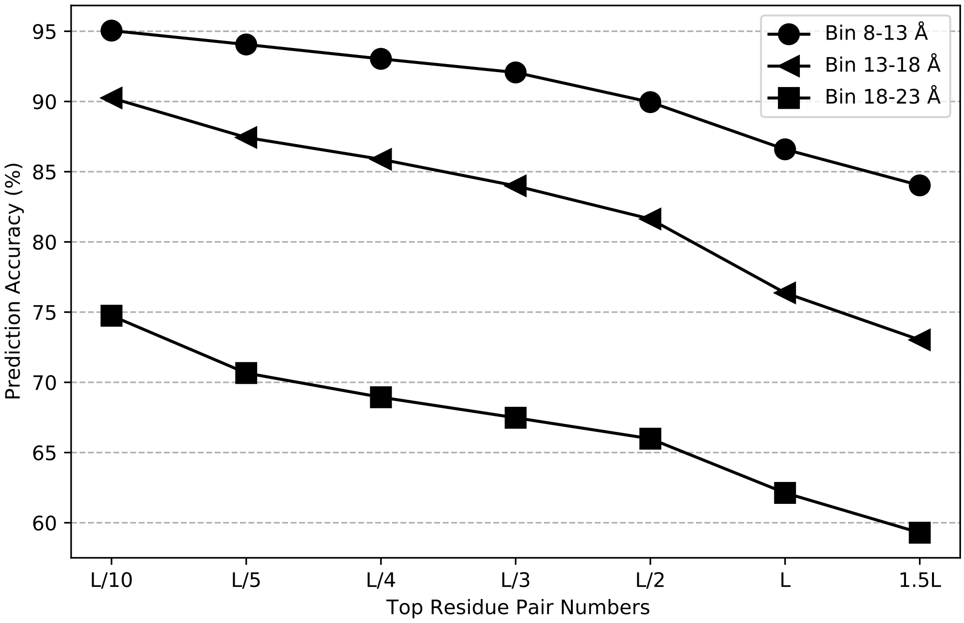


Figure 2.1: Distances were predicted with high accuracy by DeepCDpred on 108 test proteins.

To see the effect of distance predictions on the quality of predicted structures, distance predictions were implemented as constraints besides contact predictions for the 108 test set proteins. Since the contact prediction accuracy of RaptorX is better than DeepCDpred's, we introduced the distance predictions alongside DeepCDpred contact predictions or RaptorX predictions. For each test protein, 100 structures were generated by Rosetta and the structure with the lowest Rosetta score was selected as the final model. Comparison of DeepCDpred, RaptorX and RaptorX+DeepCDpred distance predictions are given in Fig. 2.2.

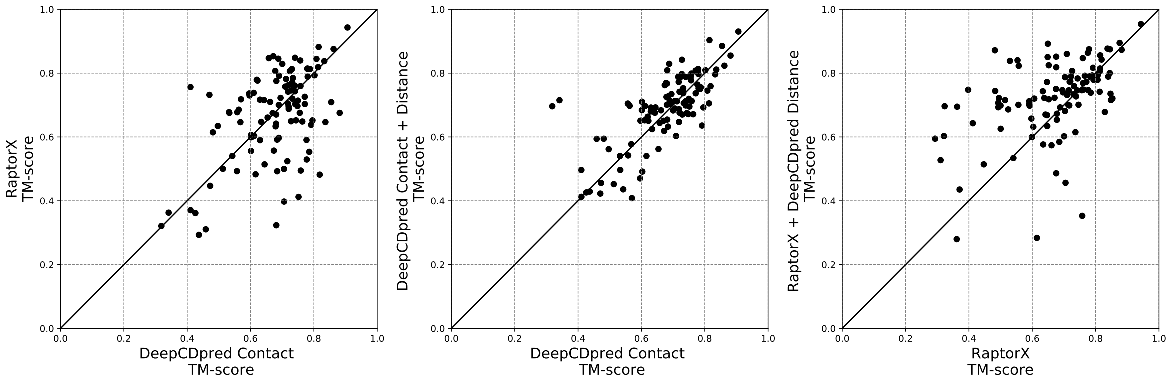


Figure 2.2: Implementation of distance predictions allows the selection of a better model when the Rosetta score is used as selection criteria. *TM-scores for DeepCDpred contact vs. RaptorX (left), DeepCDpred contact vs. DeepCDpred contact + distance (middle), and RaptorX vs. RaptorX + DeepCDpred distance (right) are plotted. There is no significant difference between RaptorX and DeepCDpred contact (p-value: 0.4325). While RaptorX + DeepCDpred distance is significantly better than RaptorX predictions (p-value: 9.24e-05), DeepCDpred contact + distance is not significantly better than DeepCDpred contact predictions (p-value:0.2121). Contact predictions were made by Liam Mead.*

For the prediction of the structures, using contact predictions from DeepCDpred or RaptorX did not lead to any significant difference between the average TM-scores of the selected models (selected by Rosetta score) of the test proteins (paired t test p-value=0.4325) (Fig. 2.2, left), although contact prediction accuracies of RaptorX was higher than DeepCDpred's (Fig. A.3). While implementation of DeepCDpred distance predictions did not lead any significant improvement on the TM-score of the predicted structures (paired t-test p-value=0.2121) (Fig. 2.2, middle), implementation of the distance predictions to RaptorX predictions provided a significant improvement in the final structure selected by Rosetta energy (paired t-test p-value= 9.24e-05) (Fig. 2.2, right). This result suggests that implementation of DeepCDpred distance predictions increases the accuracy of the predicted structure.

Although using Rosetta energy is a reasonable and easy way to to select the final structure from the structure pool, it fails to find the structure closest to the experimental structure (the structure with the highest TM-score). Selecting the structure with the highest TM-score as the final structure reveals that predictions with RaptorX contact predictions produce significantly better results than the structures generated with DeepCDpred contact predictions (paired t test p-value: 1.9e-12) (Fig. 2.3, left). Furthermore, implementation of DeepCDpred distance predictions improves the predicted structure quality when they are implemented together with DeepCDpred contact or RaptorX (paired t test p-value: 0.0108 and 0.0370, respectively) (Fig. 2.3, middle, right).

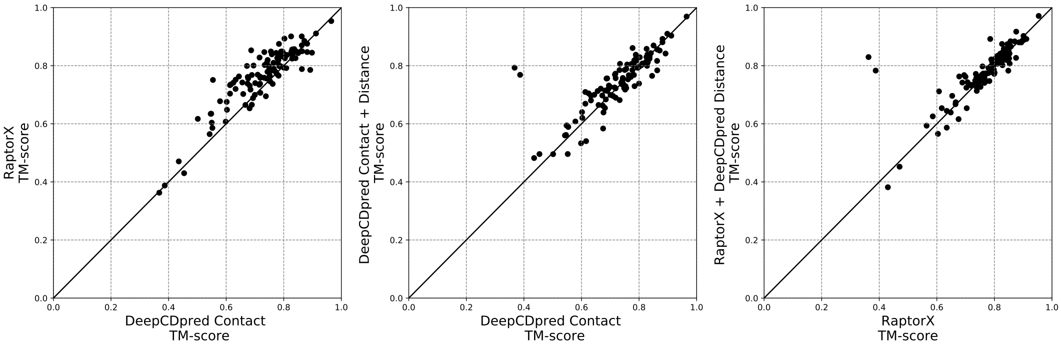


Figure 2.3: Implementation of distance predictions allows the selection of a better model when the best model is selected from the structure pool. *TM-scores (structure with the best TM-score is selected) for DeepCDpred contact vs. RaptorX (left), DeepCDpred contact vs DeepCDpred contact + distance (middle), and RaptorX vs. RaptorX + DeepCDpred distance (right) are plotted. RaptorX is significantly better than DeepCDpred contact (p-value: 1.9e-12). Distance predictions have a significant effect in both DeepCDpred contact and RaptorX predictions (p-values: 0.0108 and 0.0370, respectively). Contact predictions were made by Liam Mead.*

2.3.2 The effect on the prediction accuracy of sharing common homology/topology

Training and test set proteins were originally selected based on having sequence similarity less than 25 % identity as mentioned in section Section A.0.2. While this approach ensures that we do not use homologous sequences, the approach ignores the fact that structures of proteins may be similar despite their sequence similarities being low (Orengo *et al.* 1994). Therefore, we wanted to analyse whether having proteins in the training set with the same topology and/or homology class as the test proteins increased the prediction accuracy by causing a bias, possibly due to a bias arising during training..

For this purpose, we first calculated how many of the test proteins share common topology and homology with the training set proteins. Among 108 test set proteins, 80 of them share the same homology classes as the training set proteins (homologous proteins); whereas 28 of them do not belong to any mutual homologous superfamily with the training proteins (unique). In terms of sharing the same topology classes, 90 of the test proteins belong to the same topology classes as the training set proteins (topologous proteins); whereas, 18 of them do not have any common topology with the training set proteins (unique proteins).

Since the number of unique proteins and homologous/topologous protein groups is not balanced, we decided to train the neural network with different subsets of the training set. The training set was divided into three groups and the network was trained on the proteins from individual groups separately. The groups were determined by Tugce Oruc, but the training was performed by Liam Mead (Mead 2018).

Summary of the new training subsets are as following:

1. homology training set (includes proteins that belong to the same homologous superfamily as the test set proteins) - 532 proteins
2. topology training set (includes proteins that belong to the same topology class but not to the same homologous superfamily as the test set proteins) - 446 proteins
3. unique training set (does not include any protein that belongs to the same homologous superfamily or topology class as the test set proteins) - 723 proteins

As the number of the proteins in the subsets is not the same, ten different training sets - from all three sub-training protein sets - with 200 proteins were constructed by a random selection of the proteins, ending up with 30 training subsets in total. Each subset was trained with the nine-layer neural network architecture. The accuracy for the 108 test proteins was calculated and the average accuracy of the ten sets was used for comparisons. The mean contact prediction accuracy (for top 1.5L residue pairs) of the test set for the models trained with topologous training sets is 63.8% (σ=0.41), and the mean accuracy of the test set for the models trained with homologous sets is 65.6% (σ=0.83). Both of the accuracies are significantly higher than the mean accuracy of the test set for the models trained with set of unique proteins, which is 62.6% (σ=0.90) (two-tailed paired t-test p-values are 7.92 x 10^{-7} and 6.28 x 10^{-19}, respectively).

When we first focus on the accuracy difference between the homologous and unique training sets, the results suggest that training on the homologous proteins provides more successful predictions for the test set proteins. One possible reason for that can be thought as the neural network learns patterns from the homologous structures leading to better predictions for the test set proteins. One potential other reason is differences in the depth of the MSA and length of the sequences. When the Nf is low, prediction accuracy drops (Ovchinnikov *et al.* 2015; Ovchinnikov *et al.* 2017), and the average Nf value of the homologous training set is higher compared to both the unique set and the topologous set (Fig. 2.4).

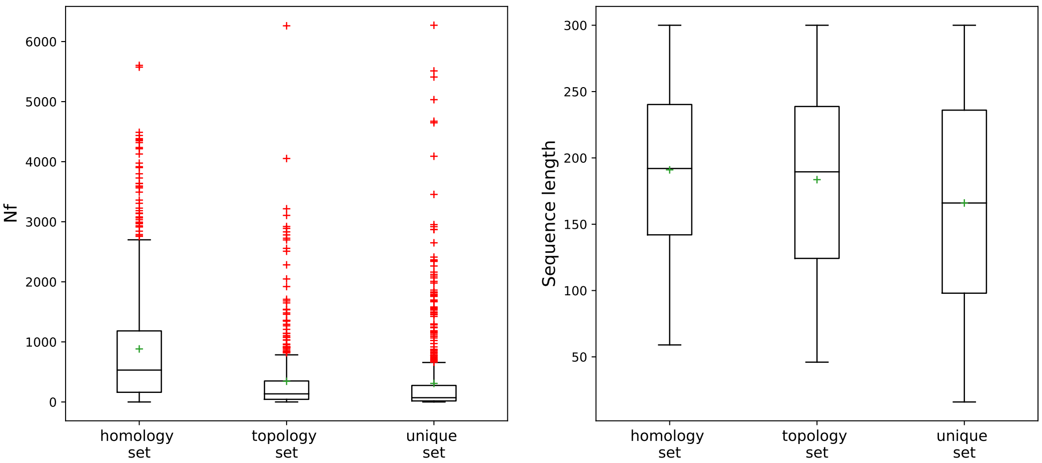


Figure 2.4: Nf values and sequence lengths of homology, topology and unique training set proteins. *While Nf values of topology and unique protein sets have a similar distribution, Nf values of homology set proteins are higher than topology and unique protein sets. Sequence lengths of three sets have a similar distribution. Nf: number of sequences in the alignment, having maximum 80% identity to each other, divided by the square root of the protein length. The boundareis of the box range from the lower quartile of the data to the upper quartile, the line shows the median, the whiskers demonstrate the range of the data. Outliers are shown with red plus signs. The mean values are shown with green plus signs.*

To eliminate the effect of the Nf on the success of training, the number of sequences in the homologous training proteins were decreased to give a similar mean to the unique set proteins. Comparison of the Nf value distribution of ten unique sets, ten homologous sets and ten homologous sets with reduced Nfs is shown in Fig. 2.5. The average of 1.5L contact prediction accuracy with trimmed MSAs of the training proteins is 61.74; whereas, the average accuracy of the trainings with the unique test set is 62.61. As the average accuracy of the network models trained on unique proteins are significantly higher than the average accuracy of the models trained on homologous proteins with reduced Nfs (paired t-test p-value = 0.00157), it can be suggested that having almost three times higher average Nf score is the primary reason why training on the homologous sets resulted in a significantly higher contact prediction accuracy compared to training with unique sets.

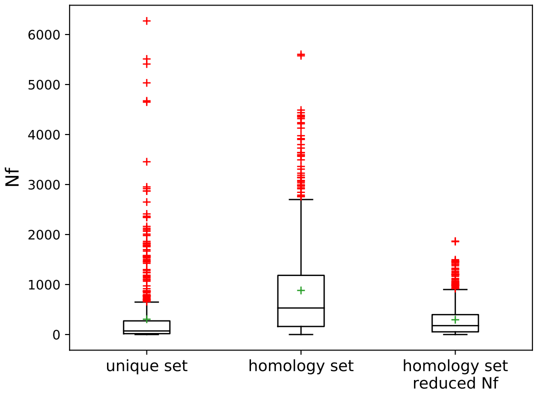


Figure 2.5: Nf values of the proteins in unique set, homology set and homology set with reduced Nf. *Nf values of homology set proteins were reduced in a way to have a similar mean Nf similar to the unique set proteins. Nf: number of sequences in the alignment, having maximum 80% identity to each other, divided by the square root of the protein length.The boundareis of the box range from the lower quartile of the data to the upper quartile, the line shows the median, the whiskers demonstrate the range of the data. Outliers are shown with red plus signs. The mean values are shown with green plus signs.*

Individual analysis of the accuracy of test set proteins between the homologous and the unique sets, and homologous set with reduced Nfs and the unique sets are shown in Fig. 2.6 and Fig. 2.8. The percentage change of the accuracy (i.e. the difference in the accuracy) of test set proteins from unique sets to homologous sets (the homologous set accuracy minus the unique set accuracy, divided by the unique set accuracy), unpaired t-test p-values between the accuracy calculations of ten sets (ten predictions for the unique trainings and ten for homologous trainings), and the number of homologous proteins in the homologous training sets for the test proteins are shown in Fig. 2.6. Not surprisingly, for the proteins with the high difference between unique and homologous trainings, the p-value is lower. If training on the homologous set allows the network to learn the patterns of the homologous proteins, it can be expected that test set proteins which have a greater number of homologous structures in the homologous training set would have higher predictions compared to the trainings on the unique training sets. However, for the proteins, which have a greater number of homologous proteins in the homologous training sets (for example the one with 52 homologous proteins) the accuracy difference is not significant. Similarly, for the proteins, even the ones that do not have any homologous proteins in the homologous training set, there can be a significant difference between the trainings. The correlation coefficient between the number of homologous proteins in the training set and the ratio of the difference is 0.37 (p-value of 0.00007), indicating a significant weak correlation (Fig 2.7). This suggests that there is not any strong evindence to state that having homologous proteins in the training set cause a bias towards homologous proteins for our test set proteins.

When the Nf numbers were reduced in the homologous training sets, there are fewer proteins whose percentage difference are significant (Fig. 2.8). Further, it is clear that the difference (the homologous set accuracy minus the unique set accuracy, divided by the unique set accuracy), is not only positive, but it is negative for most of the proteins, indicating there are more proteins with higher accuracy on the unique set. Moreover, it is seen that for four of the six proteins with the highest number of homologous proteins in the training set, there is no significant difference between the training on homologous sets and the training on the unique sets. The correlation coefficient between the number of homologous proteins in the training set and the ratio of the difference is 0.32 (p-value of 0.0007), again indicating a significant weak correlation (Fig. 2.9).

Analyses of the test set contact prediction accuracy of the models trained on the topologous sets and the unique sets are shown in Fig. 2.10. Although the average prediction accuracy of models trained on topologous sets is significantly higher than the trained models with unique set protein (paired t-test p-value= $7.92 \text{x} 10^{-7}$), this difference does not seem to be sourced by

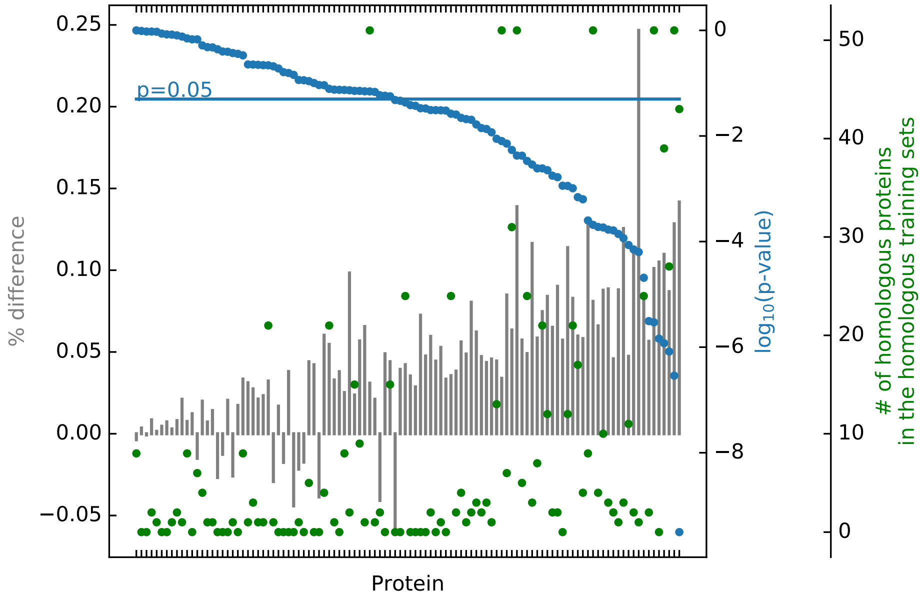


Figure 2.6: Contact prediction accuracy comparison of the test proteins on the models trained with unique set proteins and homology set proteins. *The difference between the accuracy of unique and homology sets (the homologous set accuracy minus the unique set accuracy, divided by the unique set accuracy) (gray bars, left hand axis), the statistical significance of that difference measured as an unpaired p-value between the ten predictions of unique and homology sets (blue dots) and the number homologous proteins in the homology training sets (green dots) are shown for the test proteins. The training was performed by Liam Mead, the groups were determined and the analysis were performed by me.*

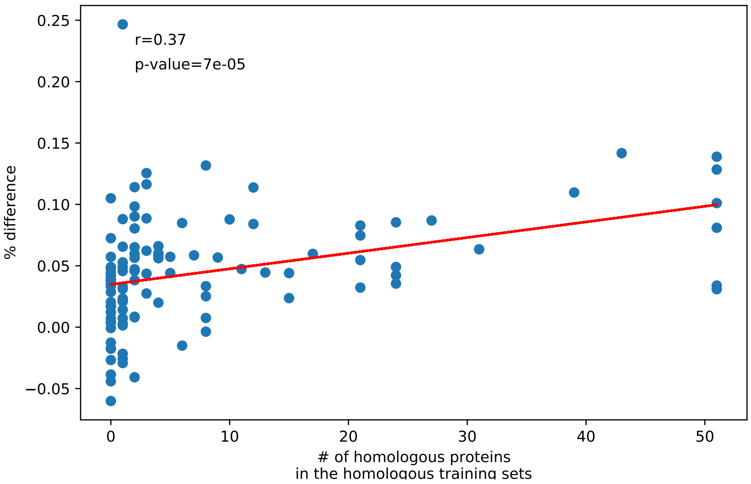


Figure 2.7: Correlation between the number of homologous proteins in the homologous training set and the percentage difference in accuracy is significant but weak. *The correlation coefficient, r, is 0.37 with a p-value of 0.000007. R^2=0.139.*

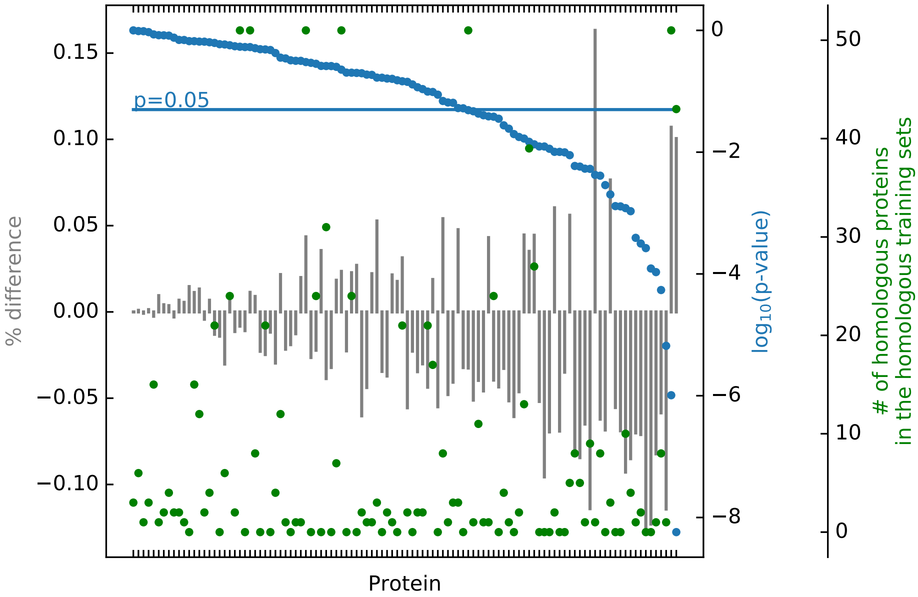


Figure 2.8: Contact prediction accuracy comparison of the test proteins on the models trained with unique set proteins and homology set proteins with reduced Nfs. *The difference between the accuracy of unique and reduced Nf-homology sets (the homologous set accuracy minus the unique set accuracy, divided by the unique set accuracy) (gray bars, left hand axis), the statistical significance of that difference measured as an unpaired p-value between the ten predictions of unique and reduced Nf-homology sets (blue dots) and the number homologous proteins in the reduced Nf-homology training sets (green dots) are shown for the test proteins.*

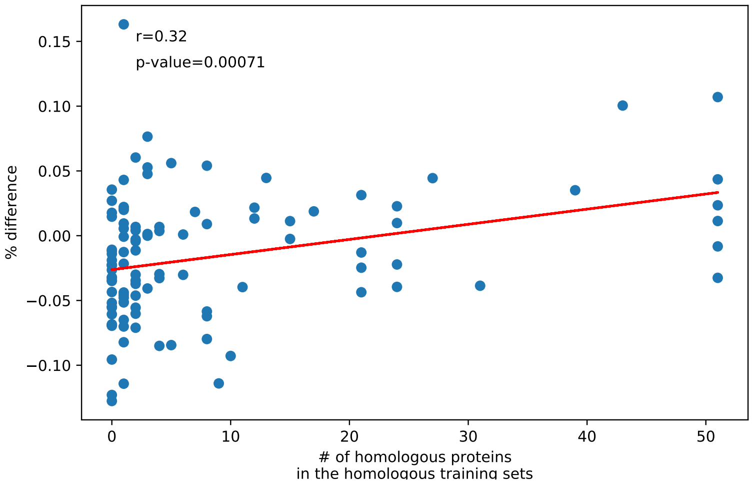


Figure 2.9: Correlation between the number of homologous proteins in the homologous training set with reduced Nf values and the percentage difference in accuracy is significant but weak. *The correlation coefficient, r, is 0.32 with a p-value of 0.00007. R^2=0.103.*

having topologous proteins in the training set. In that case, it would be expected that the test set proteins, which have more topologous proteins in the topologous training set, should have a significant increase, which is not the case (Fig. 2.10). The contact prediction accuracy for the proteins that have the maximum number of topologous proteins (as many as 116), changed both significantly and insignificantly. Similarly, for the proteins that do not have any topologous proteins in the topologous training set, a significant change in the contact prediction accuracy was observed. There is no correlation between the number of topologous proteins in the topologous training set and the percentage difference (Fig. 2.11). This suggests that having topologous proteins in the training set does not cause a bias in the predictions.

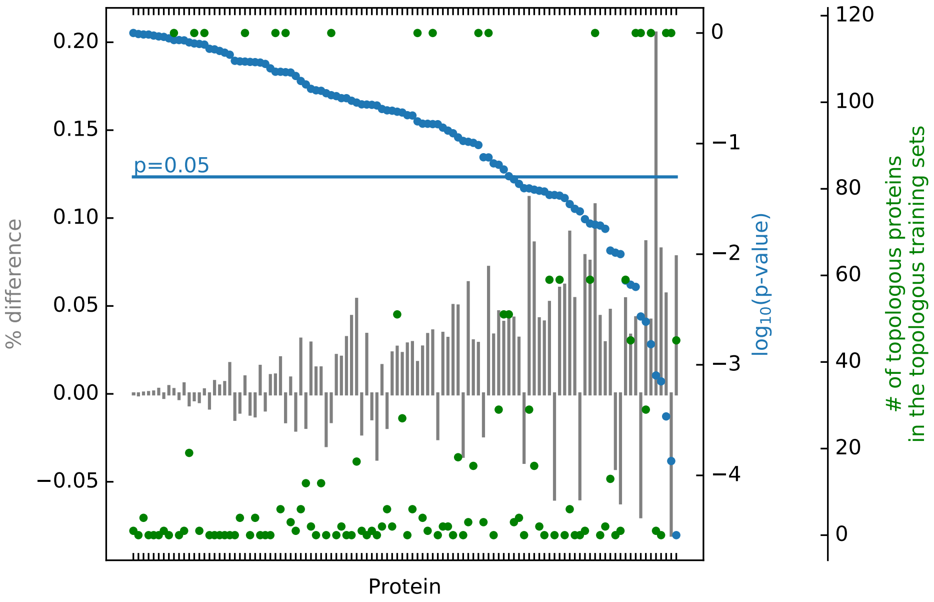


Figure 2.10: Contact prediction accuracy comparison of the test proteins on the models trained with unique set proteins and topology set proteins. *The difference between the accuracy of unique and topology sets (the topologous set accuracy minus the unique set accuracy, divided by the unique set accuracy) (gray bars, left hand axis), the statistical significance of that difference measured as an unpaired p-value between the ten predictions of unique and topology sets (blue dots) and the number topologous proteins in the topology training sets (green dots) are shown for the test proteins. The training was performed by Liam Mead, the groups were determined and the analysis were performed by me.*

Although a significant difference was detected in the comparison of the models trained on the unique training set versus the homologous training set with reduced Nf values, and the unique training set versus topologous sets, a detailed analysis suggested that having homolo-

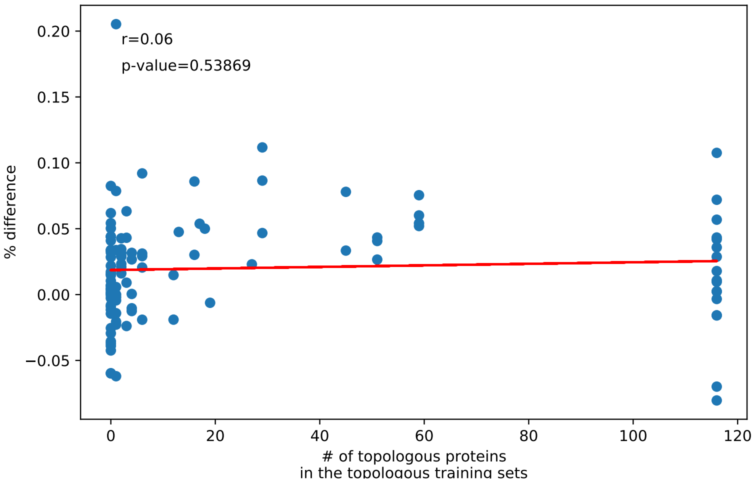


Figure 2.11: There is no correlation between the number of topologous proteins in the topologous training set and the percentage difference. *The correlation coefficient, r, is 0.06 with a p-value of 0.53869. R^2=0.004.*

gous/topologous proteins in the training and the test sets are not the source of this difference. The fluctuation in the prediction accuracy can be caused by a poor generalization of the models since only 200 proteins were used in a training set. Overall, these analyses suggest that our trained models do not learn specific patterns belonging to some topology class or homologous superfamily class resulting in a bias in the prediction accuracy.

2.3.3 Structure prediction on an additional non-topologous test set

To further investigate whether having topologous/homologous structures in the training set causes a bias in the prediction accuracy, another test set whose structures are not topologically related to the training set proteins was generated. The new test set includes 50 proteins that do not belong to the same topology class as the training set proteins of DeepCDpred, MetaPSICOV or RaptorX (for further details please see Section 2.2.3).

Comparison of the Nf values and the sequence lengths of the test sets with 108 proteins and 50 non-topologous proteins is shown in Fig. 2.12. The average of the Nf values for the 50 proteins ($\overline{\text{Nf}}_{50} = 75$) is almost an order of magnitude lower than the average of the Nf values of the 108 protein test set ($\overline{\text{Nf}}_{108} = 794$). The maximum sequence length of the 50 test proteins is 400; whereas, it is 242 for the 108 proteins. Having lower Nf values makes contact and structure prediction of the 50 proteins a harder challenge as the success of the coevolved pair detection is related to the depth of the alignment and the length of the sequence (Ovchinnikov *et al.* 2015; Ovchinnikov *et al.* 2017).

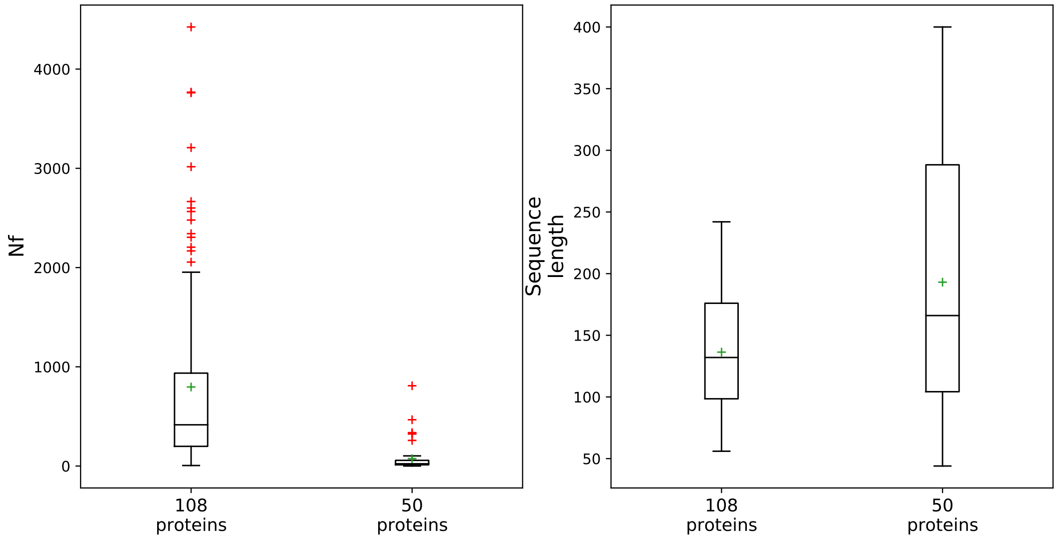


Figure 2.12: Nf values and sequence lengths of the test sets with 108 and 50 proteins. *The average Nf value of the test set with 108 proteins is almost an order of magnitude larger than the average Nf value of the test set with 50 proteins. Nf: number of sequences in the alignment, having maximum 80% identity to each other, divided by the square root of the protein length.*

Contact prediction accuracies for the 50 additional test set proteins are drastically low compared to the accuracy of the 108 proteins for all of the contact prediction tools that we use to compare (Fig. 2.13). Among the three tools we used, RaptorX provided the highest accuracy similar to the test set with 108 proteins, where DeepCDpred and MetaPSICOV performed with similar accuracy. The prediction of distance accuracy is low for the 50 protein test set compared toe 108 protein set. (Fig. 2.14).

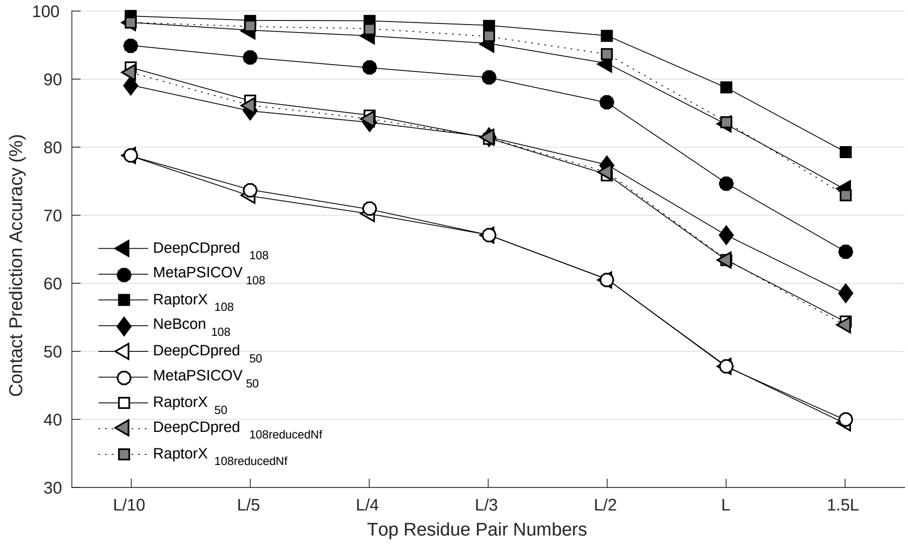


Figure 2.13: Contact prediction accuracy of both 108 protein and 50 protein test sets; besides, the accuracy of 108 test set proteins when their Nf values were reduced. *Prediction accuracy for all DeepCDpred, RaptorX and MetaPSICOV is lower for the test set with 50 proteins compared to the test set with 108 proteins. When the Nf values were reduced of the 108 proteins, performances of both DeepCDpred and RaptorX decreased. Nf: number of sequences in the alignment, having maximum 80% identity to each other, divided by the square root of the protein length.*

This data suggests that contact and distance predictions of the 50 proteins are less successful compared to the 108 test proteins. As the Nf values of the test set with 108 proteins are an order of magnitude greater than the test set with 50 proteins, we investigated the accuracy of predictions on the 108 protein test set but with lower Nf values. To test that, their alignments were trimmed in a way to obtain a similar average Nf value as the 50 test set proteins. The analysis showed that both the accuracy of DeepCDpred and RaptorX contact predictions decreased compared to the full alignment predictions (Fig. 2.13). On the other hand, the decrease for DeepCDpred is larger compared to the decrease of RaptorX predictions (Fig. 2.13). Although a decrease could be observed due to the dropped Nf values, their contact prediction accuracy is

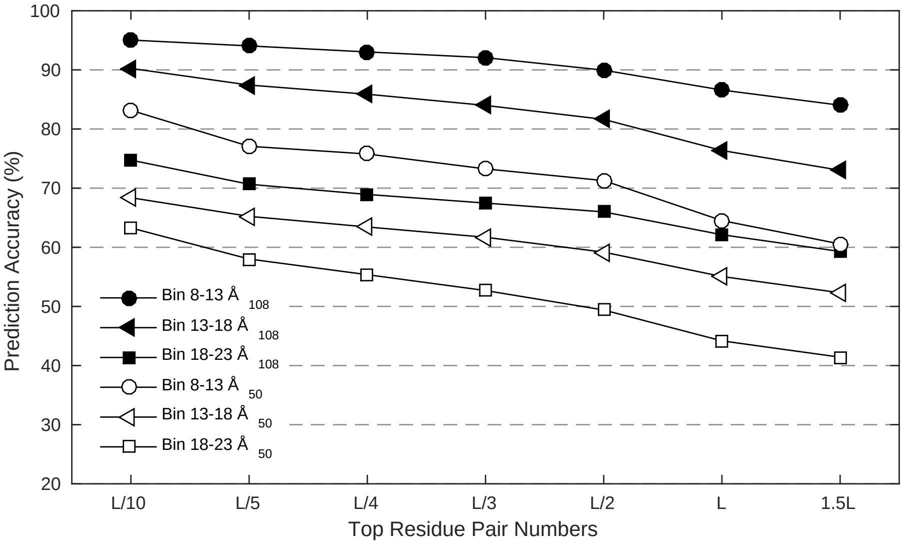


Figure 2.14: Distance prediction accuracy for both 108 protein and 50 protein test sets. *Prediction accuracy for all three bins is lower for the test set with 50 proteins compared to the accuracy for the test set of 108 proteins.*

not as low as the contact prediction accuracy of the 50 test proteins.

Structure prediction comparison with DeepCDpred and RaptorX predictions for the 50 non-topologous proteins are given in Fig. 2.15. Results suggest that for the 50 proteins, RaptorX contact predictions provide better structure predictions compared to DeepCDpred contact predictions (paired t-test p-value:0.0008), which is not surprising as the contact prediction accuracy is higher than DeepCDpred contact prediction accuracy (Fig. 2.13). While the implementation of distance predictions to the DeepCDpred contact predictions did not lead any improvements in the selected models (paired t-test p-value:0.0621), implementation of distance predictions to RaptorX contact predictions led a significant increase in the TM-scores of the predicted structures (paired t-test p-value:0.0013).

On the other hand, the increase in the TM-score for selected structure with implementation of the distance predictions could not be observed for the comparisons of the best structures (with the highest TM-score) in the pool (Fig. 2.16). TM-scores of the best-predicted structures by RaptorX contact predictions are significantly higher than the TM-scores of the best-predicted structures by DeepCDpred (paired t-test p-value: 2.869e-05) (Fig. 2.16, left). Implementation of distance predictions to DeepCDpred contact predictions surprisingly reduces the average TM-score (paired t-test p-value=0.0397) (Fig. 2.16, middle). Implementation of DeepCDpred

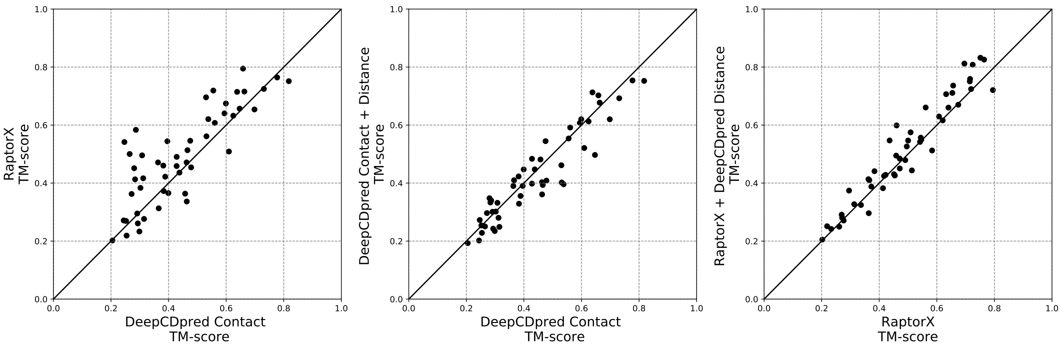


Figure 2.15: Implementation of distance predictions allows the selection of a better model when the Rosetta score is used as selection criteria for addtional 50 test proteins. *TM-scores for DeepCDpred contact vs. RaptorX (left), DeepCDpred contact vs. DeepCDpred contact + distance (middle), and RaptorX vs. RaptorX + DeepCDpred distance (right) are plotted. Paired t-test p-values are 0.0008, 0.0621 and 0.0013, respectively.*

distance predictions to RaptorX contact predictions lead a slight but insignificant increase in the average TM-score of the best predictions (paired t-test p-value=0.4533) (Fig. 2.16, right).

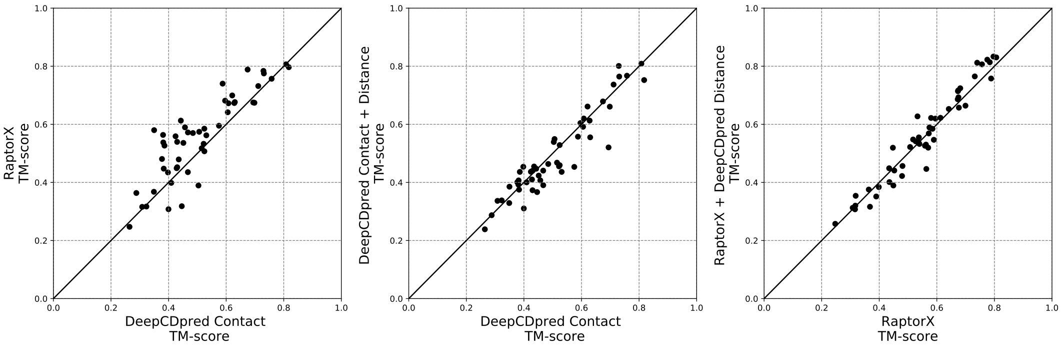


Figure 2.16: Implementation of distance predictions provides small but insignificant improvement on model generation when RaptorX contact predictions are used for 50 test proteins. *(TM-scores for DeepCDpred contact vs. RaptorX (left), DeepCDpred contact vs. DeepCDpred contact + distance (middle), and RaptorX vs. RaptorX + DeepCDpred distance (right) are plotted. Paired t-test p-values: 2.869e-05, 0.0397 and 0.4533, respectively.*

2.3.4 Effect of structure pool size in the prediction accuracy.

To see how increasing the size of the structure pool affects the prediction accuracy, 100 additional structures were generated for both test sets (with 108 proteins and 50 non-topologous proteins) for RaptorX contact predictions and RaptorX contact predictions with DeepCDpred distance predictions. Comparison of the structures from the pool with 100 structures and the pool with 200 structures selecting one per pool using the minimum Rosetta energy or the true best structure, determined by the TM score with respect to the experimental structure are shown in Table 2.2. For the test set with 108 proteins, increasing the number of structures in the pool from 100 to 200 produced structures closer to the experimental structure when both distance predictions were used and were not used (p-values are 7 x 10^{-5} and 0.04, respectively). When Rosetta score was used as the selection criteria, a significant improvement was detected in the average TM-score when distance predictions were implemented (p-value = 0.008); however, a significant improvement could not be detected when only RaptorX contact predictions were used to generate structures (p-value = 0.320). It suggests that increasing the pool size provides to generate structures closer to the experimental structure.

Table 2.2: Comparison of average TM-scores of the structure pools with 100 vs. 200 models.

	Lowest Rosetta Energy			**Model with highest TM**		
	Average TM from 100	**Average TM from 200**	**p-value**[1]	**Average TM from 100**	**Average TM from 200**	**p-value**[1]
RaptorX contact only (108)	0.667	0.665	0.320	0.766	0.775	0.04
RaptorX + Distance (108)	0.720	0.737	0.008	0.780	0.786	7x10^{-5}
RaptorX contact only (50)	0.493	0.494	0.831	0.557	0.568	1x10^{-5}
RaptorX + Distance (50)	0.516	0.515	0.850	0.561	0.576	8x10^{-5}

[1] Paired t-test to determine if there is a significant improvement when the pool size is increased.

For the test set with 50 non-topologous proteins, increasing the pool size from 100 to 200 increases the average TM-score when the structure with the highest TM-score is selected. In other words, generating more structures increases the number of structures in the pool that are close to the experimental structure. On the other hand, this increase could not be detected when minimum Rosetta score was used to select the final model for this test set. Overall, using another model assesment method may increase the accuracy of the final model; however, without a better scoring method there’s not much point in increasing the pool size.

2.3.5 Comparison of the effect of distance predictions in proteins from two test sets with 200 generated structures

Comparisons of selected models from the pool with 200 structures are given in Fig. 2.17. For the 108 protein test set, implementing the DeepCDpred distance predictions significantly improved the average TM-score with an average increase of ~ 0.07 (paired t-test p-value = 9x10−9) compared to using the RaptorX contact predictions only (Fig. 2.17A). Similarly, for the 50 non-topologous test set, addition of the distance predictions increased the average TM-score by ~ 0.03 (paired t-test p-value = 0.004) (Fig. 2.17C). While using distance predictions improves the average TM-score of the best models significantly by ~ 0.01 (paired t-test p-value = 0.001) for the test set with 108 proteins (Fig. 2.17B), a significant improvement could not be observed for the 50 non-topologous test set (paired t-test p-value = 0.158) where the average TM-score of the best models increased only by ~ 0.008 (Fig. 2.17D).

Although using the lowest Rosetta score to select the final model from a structure pool is an easy and time-saving approach, it rarely selects the best model from the pool. Selecting the model closest to the experimental structure, model assessment, is another challenge of structural bioinformatics which has been studied by many groups (Cheng *et al.* 2019). However, we see that when we use the distance predictions as additional constraints for the structure predictions, using the lowest Rosetta scores led us to select the models closer to the experimental structures (Fig. 2.18). For the 108 protein test set, the correlation between the TM-scores of selected models by the lowest Rosetta energy and the models closest to experimental structures is 0.62 when only contact predictions were used as restraints (Fig. 2.18A). When distance predictions were implemented as additional restraints, the correlation increased to 0.89 (Fig. 2.18B), which indicates the addition of distance restraints allows the selection of better models from the structure pool. A similar, but not that distinct pattern is seen for the 50 non-topologous test proteins. The correlation between the models selected by the Rosetta score and the models closest to the experimental structure improved from 0.95 to 0.97 by addition of the distance restraints (Fig. 2.18C,D).

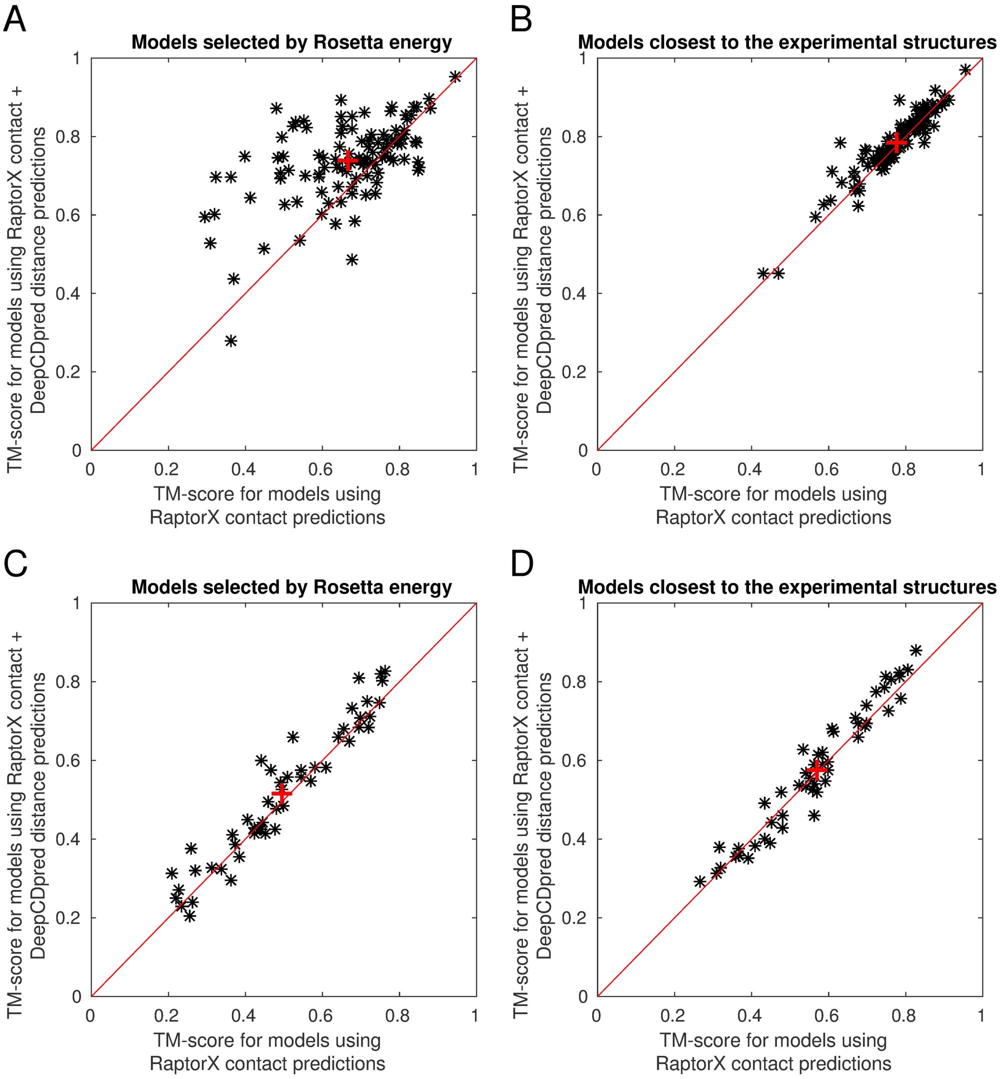


Figure 2.17: Distance predictions improve the accuracy of predicted structures. *Implementation of distance predictions improves the accuracy of the final structure when Rosetta score is used as selection criteria for both test sets (A, C) as well as improving the quality of predicted structures significantly for the test set with 108 proteins (B). For the test set with 50 proteins, the addition of distance predictions leads a small but insignificant increase in the quality of predicted structures (D). The red '+' sign indicates the average TM-score.*

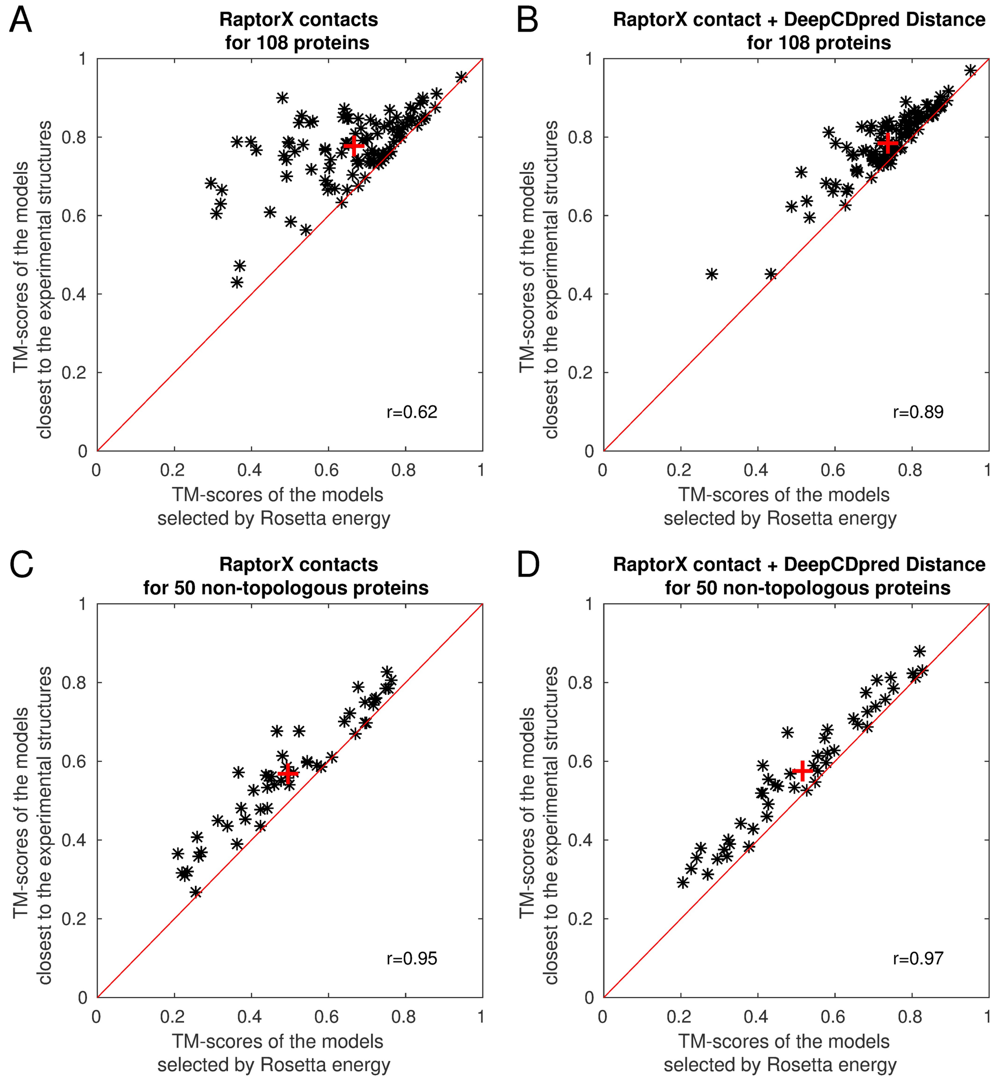


Figure 2.18: The addition of distance predictions to contact predictions provides to select models closer to the experimental structure with Rosetta energy.*The red ‘+’ sign indicates the average TM-score.*

2.3.6 Amino acid type distribution in successfully predicted residue pairs

The propensity of amino acid types in correctly predicted pairs was analysed to see which amino acid types are observed more or less frequently than the expected in the successful predictions. It is seen that in the contact bin (0-8 Å bin), the propotion of hydrophobic amino acid interactions are higher than expected, whereas a similar abundancy cannot be observed for hydrophobic-hydrophilic or hydrophilic-hydrophilic amino acid pairs 2.19A). For 8-13 Å bin, while hydrophobic pair types are still over-represented, hydrophilic - hydrophobic amino acid pairs become more abundant, as well (Fig. 2.19B). When the distance between the residue pairs increases, hydrophobic interactions lose their over-representation and hydrophilic residue pairs become more abundant (Fig. 2.19C, D). The precision of amino acid pair types are high for all amino acid types in the prediction bins, indicating there is no bias in predicting some residue pair types better than others (Fig. 2.20).

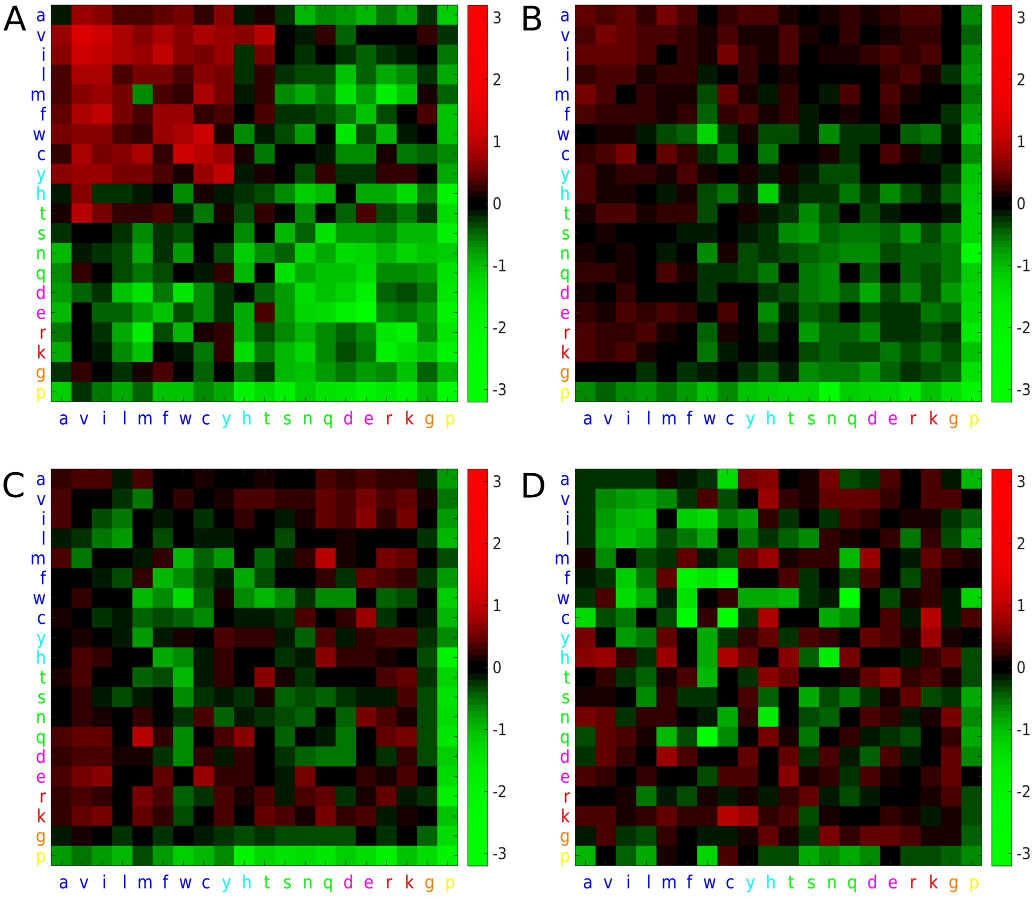


Figure 2.19: Amino acid type distribution in successfully predicted residue pairs for contact and distance bins. *A: 0-8 Å, B: 8-13 Å, C: 13-18 Å, D: 18-23 Å. Squares represent* $\log_2(<O/E>)$ *value for each amino acid pair type. The scale is shown on the right hand side of each plot. Amino acids are colored based on ClustalX coloring scheme: hydrophobic as dark blue; tyrosine and histidine in cyan; non-charged polar amino acids in green; acidic residues as magenta; basic aliphatic residues as red; glycine: orange; proline: yellow.*

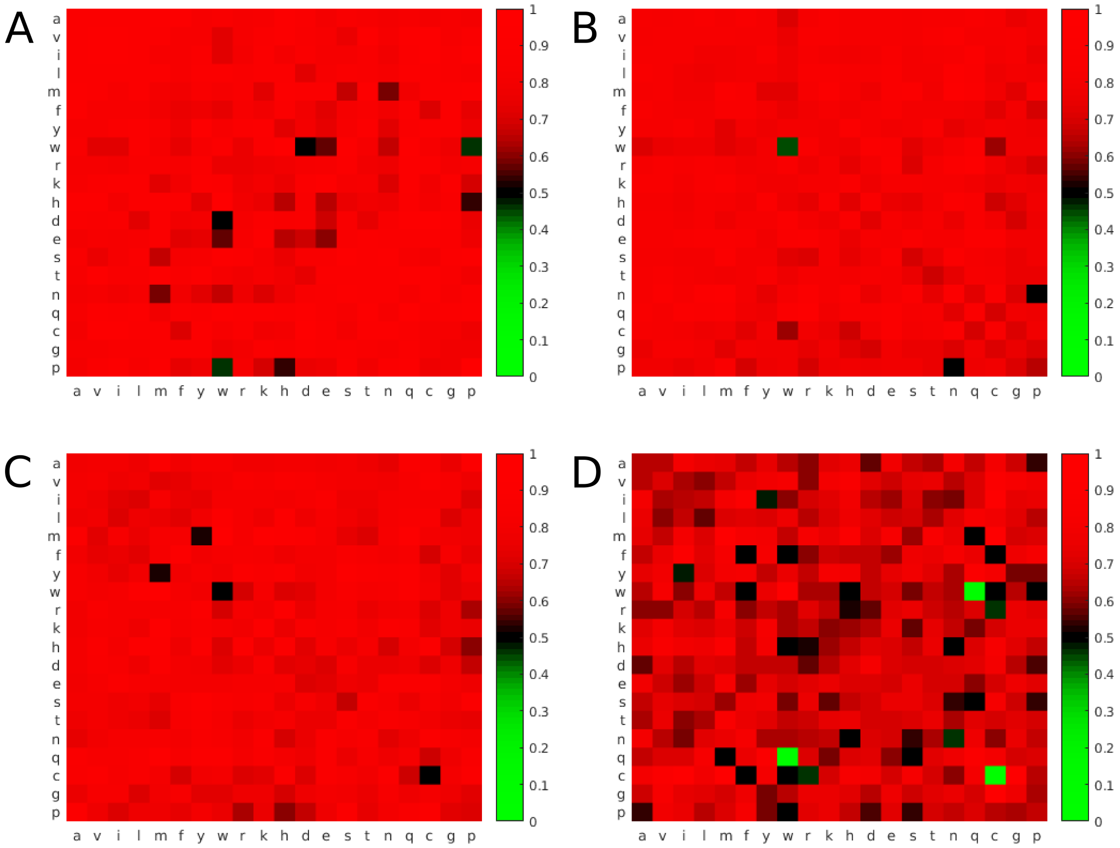


Figure 2.20: The precision of contact and distance bin predictions of amino acid pair types. *A: 0-8 Å, B: 8-13 Å, C: 13-18 Å, D: 18-23 Å. The scale is shown on the right hand side of each plot. Amino acids are colored based on ClustalX coloring scheme: hydrophobic as dark blue; tyrosine and histidine in cyan; non-charged polar amino acids in green; acidic residues as magenta; basic aliphatic residues as red; glycine: orange; proline: yellow.*

2.4 Discussion

The prediction of structures from amino acid sequences is an appealing and challenging area. In the process of predicting high-quality structures, there are critical points for generation of accurate models. One of them is accurate contact predictions, and distance predictions as in this study. Using neural networks is a useful tool for accurate predictions, as it has been shown by many studies (Jones *et al.* 2015; Wang *et al.* 2017; Senior *et al.* 2019; Senior *et al.* 2020). However, for successful predictions, there are important points that should be taken into careful consideration if one is to obtain a good network model. One of the possible problems is overfitting of the model. That means weights are trained in a way to fit on the training data successfully without generalization, leading to the model failing on the test set. To avoid such incidents, we used the early stopping algorithm. With this approach, the training stops earlier than the total number of predefined training cycles (epochs) when the no decrease is detected in validation loss for a predefined epochs (patience). For example, in our study, we trained the network with patience 40 epochs for a total of 300 epochs. This setting led the network to stop training when there is no decrease in the validation loss for 40 epochs. By this way, the training of the models was terminated even though the training loss might still be decreasing, avoiding overfitting of the model on the training data.

To be sure that our neural network does not learn some specific patterns from the training set related to structural features of some proteins, we trained new models considering the structural classification of the proteins and generated an additional test set with distinct structural organization. We could not find any results suggesting that the network learns patterns specific to some topology/homology classes leading a bias in the prediction accuracy (Sections 2.3.2 and 2.3.3). Although we could not find any data to support that possibility, the maximum number of proteins belonging to the same topology and homologous superfamily classes in the 108 test set proteins were 116 and 52, respectively, which might be too low to cause any bias. Therefore, it is not clear whether having more topologous/homologous proteins in the training set would result in a bias in the trained network. Since this question is beyond the focus of this study, the effect of having topologous/homologous proteins in the training set was not investigated further.

Another critical point for structure prediction is to use a successful tool to implement contact and distance predictions for generation of accurate 3D models. We used Rosetta to predict structures by implementing predicted contacts and distances. However, there are other approaches to generate structures. One of the alternative approaches is using a faster tool like CNS (Crystallography & NMR System) suite (Brünger *et al.* 1998), as Jinbo Xu and colleagues used in the RaptorX study (Wang *et al.* 2017). Although CNS is faster compared to Rosetta, the prediction quality is lower (Wang *et al.* 2017; Ji 2019).

Another alternative is using MD simulations to predict structures (explained in Section 1.2.2.1). Predicted contact and distances can be applied as restraints on the protein chain to generate a 3D structure. To test how effective this approach is, I tested the performance of three force fields (Amber ff99SB-ILDN(Lindorff-Larsen *et al.* 2010), Amber ff14SB(Maier *et al.* 2015) and YAMBER(Krieger *et al.* 2004)), two environmental conditions (vacuum and explicit water) with different restraint sets (with contact restraints only, with both contact and distance restraints) on one of the test proteins from the test set with 108 proteins. A simulated annealing protocol (which is basically simulation at high temperature and then slowly cooling down) was applied with restraints to scan the conformational space and find the global minimum. Besides contact and distance predictions, secondary structure predictions from SPIDER were applied as restraints. Among the three force fields, YAMBER provided the best structure. This is not quite surprising because YAMBER was developed from the Amber ff99SB force field via adjusting the parameters in a way to get closer structures to ones obtained from X-ray crystallography (Krieger *et al.* 2004). Among three conditions ((i)contact predictions in vacuum, (ii) contact predictions in explicit water and (iii) contact + distance predictions in explicit water), maximum TM-score was obtained when only contact predictions were applied in a vacuum (TM-score=0.77, RMSD=2.99). Although this protocol worked on one protein very well, its performance failed on an additional four proteins with an RMSD of 13.42 for the worst prediction. Although further optimizations could be performed to find a protocol that works on more proteins, the time cost of the simulations makes further studies inefficient. The protein I worked on has only 145 amino acids, and it took approximately six-core hours to run just one simulation in a vacuum. When explicit solvent simulations were performed, the time needed to complete a simulation increased to approximately 11 core hours. Increasing the simulation number up to 100 structures, which is the minimum number of structures generated with Rosetta, would require 1100 core hours which makes the approach computationally costly. Therefore, no further study was performed to generate structures with this approach and we generated structures with Rosetta.

We showed that increasing the pool size provides the generation of structures closer to the experimental structure. Although it can be anticipated that there can be a stationary point where a further generation of structures does not lead any further improvements in the structure quality, more structures should be generated to test it. Since this question is beyond the scope of this study, we did not generate more structures to see a pattern.

For the selection of the final model from a structure pool, we used Rosetta score and selected the model with the lowest Rosetta energy as the final model. Addition of distance predictions improved the quality of the final structure in both test sets compared to using contact predictions alone; whereas, the increase for the 50 non-topologous proteins was not as high as the increase for the 108 test proteins (by 11% and 8% for the 108 test proteins and 50 non-topologous test proteins, respectively) (Fig. 2.17). This result is not quite surprising since the Nf values of the 50 non-topologous proteins are lower compared to 108 test proteins (Fig. 2.12), and we showed that decreasing the Nf values decreases the prediction accuracy (Fig. 2.13).

Selection of the best models from a structure pool is another critical point for predicting protein structures. Many groups have been working on the model assessment challenge and develop separate tools to determine the quality of individual structures and pick the one closest to the experimental structure (Cheng *et al.* 2019). Although, they can be successful in selecting better models, using an additional tool for model assessment can be time-consuming. Implementation of distance predictions also allowed us to select better models by simply using Rosetta energy (Fig. 2.18), which is a handy and fast model selection method, especially if one needs to predict structures of tens of different proteins.

The overrepresentation of hydrophobic residue pairs in the contact bin suggests that the the network is more sensitive to predict hydrophobic pairs at the shorter distance range. On the other hand, hydrophilic interactions can be predicted more accurately for longer distances. This observation can be caused by the relative occurance of residue pair types in different bins. For example, for the contact bin, it is expected to see higher propotion of hyrophobic interactions and the loss function can be optimized by the network in a way to predict accurate hydrophobic contacts easily due to the high abundance of hydrophobic interactions at 0-8Å distance range.

Chapter 3

Interaction surface prediction of domain pairs

3.1 Overview

Prediction of multi-domain protein structure is another challenging area of computational biology. However, the developments in the domain assembly area are not as promising as the structure prediction of small or medium-sized proteins. In the last CASP/CAPRI competition, it was revealed that the template free prediction accuracies are not high, and further improvements are needed for automation (Guzenko *et al.* 2019).

Here, in this chapter, a method is developed to predict how two domains on the same chain interact with each other, without using any structural template for the interaction. Initially, I tested different feature vectors by looking at their accuracy at predicting different distances, and then I moved to predicting distances potentials, with my best model. Predicted distance potentials between the residue pairs were applied as constraints on the domain pairs for interface prediction. For almost half of the predicted domain pairs, correct interfaces could be obtained. The model was applied on a multi-domain protein complex, fatty acid synthase as a test system. To the best of our knowledge, this is the first study that predicts the distance potentials of the residue pairs on domain pairs for structure prediction of multi-domain complexes.

I first explain how I generated the training, validation and test sets, followed by introducing features I tested, and how I generated the feature matrix. After the details of the neural network architecture, I explain how I predicted the structure of domain interfaces with the predicted distances between the residue pairs. In the Results section, the results of different feature sets on the validation data are compared, and the predictions on the test data with the selected feature set are made. After presenting how the network output is converted to distance potentials, I present the interface prediction quality with and without distance potential constraints. Lastly, the effect of domain-domain interface size the Nf value and the number of predicted residue pairs on the predicted interface surface quality is analysed.

3.2 Methods

3.2.1 Training, validation and test set generation

The 3did database was used to generate training, validation and test sets. The 3did database classifies the domain-domain interactions based on Pfam families. For each domain-domain interaction (either homodimeric and heterodimeric interactions) it lists all proteins (with PDB IDs) belonging to the interactions of the Pfam domain(s). It also indicates whether the domains are on the same or different chains, the residue numbers of interacting amino acids and whether main or side chains are interacting. In the database, two domains are accepted as interacting if there are at least five estimated contacts (via van der Waals interactions, electrostatic interactions or hydogen bonds between the atoms of amino acids) between them. The version of the database that was used included PDB 2017_06 and Pfam30 data with 11200 domain-domain interactions.

I extracted the domain pairs that belong to the different Pfams (for heterodimeric interaction prediction). Among them, 1607 domain pairs are found on the same chain. Based on the PDB ID, proteins were culled to give identity lower than 25%, length between 40 and 400 amino acids and resolution 3 Å or better, via the PISCES server (Wang and Dunbrack 2003). The remaining proteins were divided into training, validation and test sets with 813, 161 and 161 proteins, respectively, such that they had a similar average Nf value (460, 405, 485, respectively). As we calculated the statistical coupling matrix (SCA) as a feature for the network (described in detail below), sequences whose effective number of sequences were insufficient to generate an SCA matrix were removed from the sets. From the test set, domain pairs with fewer than 10 contacting residue pairs were removed (as judged by C_β to C_β distance of 8Å or less, C_α for glycine) were also removed. This processes left 804, 156 and 133 proteins in training, validation and test sets. As a final process, sequence identities of individiual domains were compared. Not only should whole sequences be less than 25% identical to each other, domains within the proteins under consideration should not match between the test, training and validation sets. Therefore, individual domain sequence between the validation and training set, and the test and training set were compared. As a result, 16 of the validation proteins and 13 of the test proteins have sequence identity more than 25 % for both domains with the training set domain pairs; hence, they were discarded from the predictions.

3.2.2 Feature generation

For the generation of features, first of all, multiple sequence alignments (MSAs) were generated. Sequences of two domains were concatenated with respect to their order in the chain. Homologous sequences were searched by HHblits in uniprot20_2016_02 databases with setting e-value to 0.001, coverage to 60, minimum sequence identity to 0, maximum sequence identity to 90. 4 iterations were performed and a maximum of 500,000 sequences were allowed to pass to the next iteration.

The following features were calculated from the generated MSAs and fed into the neural network: CCMpred results as a prediction of coevolving residues (DCA) (Seemayer *et al.* 2014), mutual information (Dunn *et al.* 2007; Jones *et al.* 2015), normalized mutual information(Jones *et al.* 2015), statistical potential (Betancourt and Thirumalai 2008; Jones *et al.* 2015), secondary structure predictions (Yang *et al.* 2016), predicted accessible surface area (Yang *et al.* 2016), and the statistical coupling analysis matrix (Rivoire *et al.* 2016) were used. Additionally, the dot product of the DCA matrix (DCAdot) was also used as a potential enhancer of the coevolution signals from the DCA matrix since the dot product of any two residues would be large if their coevolution patterns with the all other residues in the dimer is similar; would be low, otherwise. Finally, a binary matrix was used to indicate whether a position carries information about the residue pairs, which is not always the case. Since the lengths of two domains are different, in the final input matrix (described below) there are cells in the matrix which do not correspond to a real residue pair. A position was assigned in the binary information matrix as 1 if the position corresponds to a real pair; as 0 otherwise. The weights of the neural networks are trained on these features.

3.2.3 Neural network architecture

A two-dimensional convolutional neural network was used, with the feature vector containing 2D data. DCA, DCAdot, mutual information, normalized mutual information, statistical potential are already 2D data; therefore, there was no need for further processing. In order to convert secondary structure information into 2D data, predicted values of a residue (from the first domain) for helix, strand and coil probabilities were multiplied by the values of the second residue (from the second domain) leading to nine values for each residue pair. Therefore, for secondary structure prediction, L x L x 9 matrices were generated where L is the total length of the sequence. Similarly, predicted accessible surface area of positions were multiplied to obtain a 2D matrix.

All resulting matrices have L x L x 1 or L x L x 9 dimensions; however, for training only d1 x d2 region of L x L matrix were extracted where d1 is the length of the first domain and d2 is the length of the second domain (Fig. 3.1). As training data size is small, the matrix with d1 x d2 size was chopped into smaller squares whose size are 32 amino acids by 32 amino acids, leading to obtaining more than one input from one domain pair. For example, for a domain pair whose sequence lenghts are, let's say, 100 and 50 amino acids, chopping the matrix into 32 amino acids-sized matrices provides to have eight different input matrices rather than one input matrix. Moreover, in order to increase the training data size further and avoid edge effects, three additional boundaries, 16 residue-size away from the boundaries of the original d1 x d2 matrix, were selected for cropping (Fig. 3.2). For the final score of a residue pair, the average of four predictions were taken. This setting for the training with four different boundaries will be called as T4 setting in the remaining of the text.

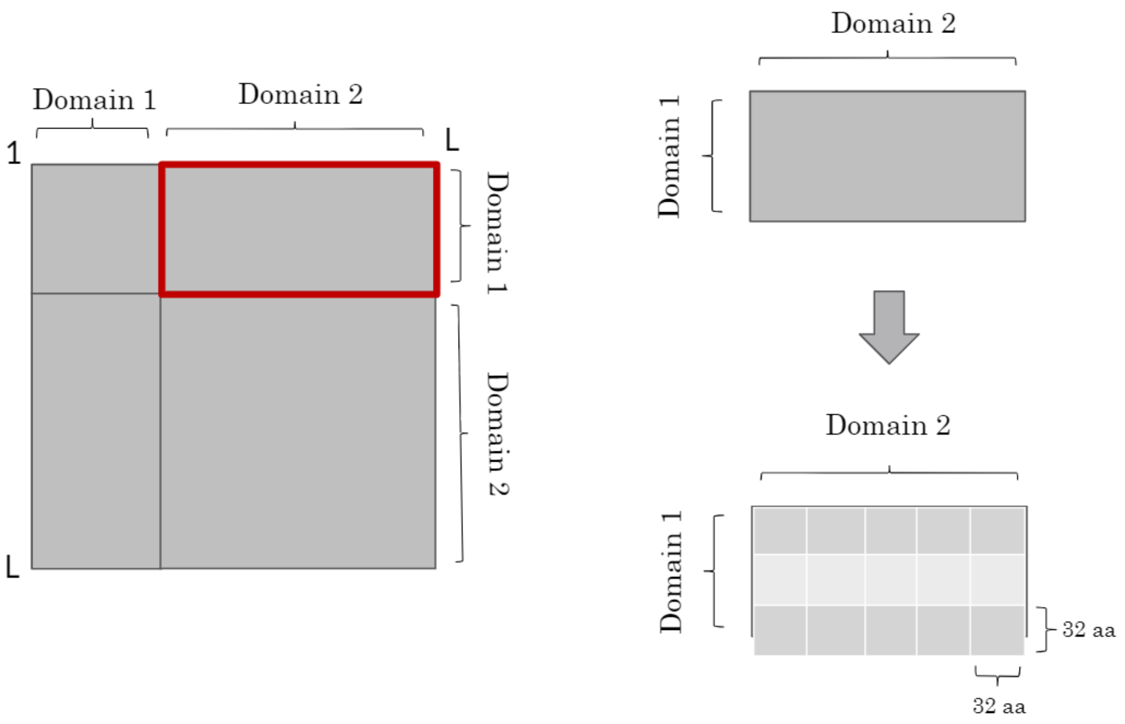


Figure 3.1: Feature matrix generation for neural networks. *From the overall matrix, the intersection region of the first domain and the second domain is extracted and further divided into square matrixes with a length of 32 amino acids (aa).*

As the output matrix, 1 Å interval bins were used ([0, 3.5), [3.5, 4.5), [4.5, 5.5) .... [19.5, 20.5), [20.5, ∞)), where for a residue pair, the real distance bin is marked with 1 and the other layers are marked as 0 for corresponding residue pair (Fig. 3.3). The design of the network arhitecture was inspired by the architecture of AlphaFold (Senior *et al.* 2019; Senior *et al.* 2020). 24 residual blocks were used consisting of three layers with a 1 x 1 x 64 projection layer, a 3 x 3 x 64 dilated convolution layer, and a 1 x 1 x 128 projection layer. Elu was used as the activation function, batch normalization was applied before every layer, dropout (0.3 rate) was used to avoid overfitting, 'he_normal' was used as the kernel initializer, Adam optimizer was used with its default settings in Keras. Dilated convolutional layers were used in order to increase receptive field with a rate of 1 (no dilation), 2, 4 and 8 in a cyclic manner. Keras (Chollet 2015) with Tensorflow (Abadi *et al.* 2015) backend was used.

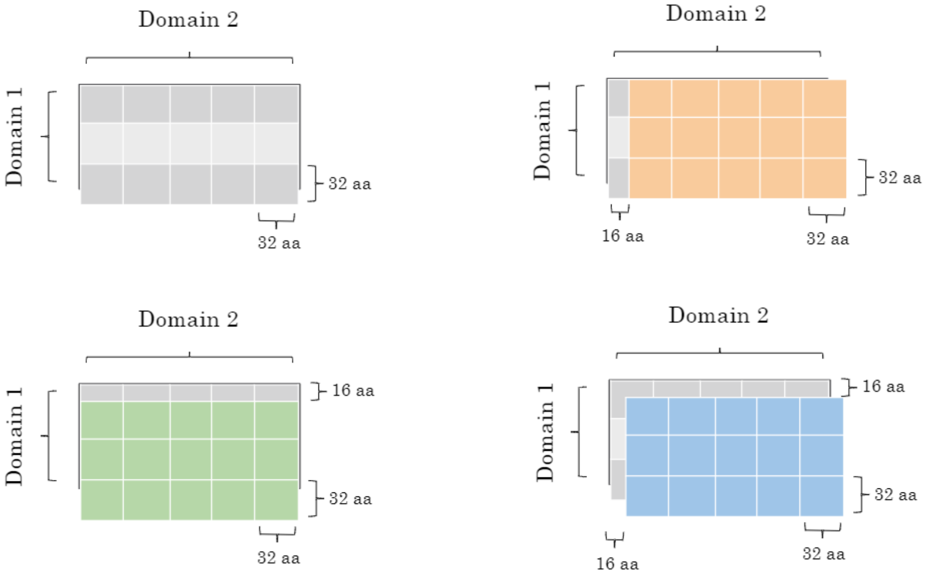


Figure 3.2: Increasing the input size via chopping the input matrix from three additional boundaries (T4 setting). *In order to increase the input data size d1 x d2 matrix was chopped from four different boundaries. Additional three matrices (colored with green, orange and blue) whose boundaries start from 16 amino acid (aa) away from the boundary of d1 x d2 matrix. As the final network score for a residue pair, the average of four settings was calculated.*

3.2.4 Measuring the accuracy of the the residue-residue interaction predictions

In order to determine how accurately residue pair distances were predicted, the accuracy of pairs in 0 - 8 Å, 8 - 13 Å, 13 - 18 Å distances were calculated. For 0 - 8 Å, 8 - 13 Å, 13 - 18 Å distances, accuracy was estimated via

$$accuracy = \frac{\#\, of\ residue\ pairs\ correctly\ predicted\ in\ a\ bin}{\#\, of\ residue\ pairs\ predicted\ in\ a\ bin.} \tag{3.1}$$

A measure we refer to as the accuracy for ±2 Å was also calculated. This is the proportion of residue pair predictions where the highest scoring bin is within 2Å of that observed in the experimental structure as compared to the total number of predictions.

3.2.5 Distance potential prediction between two residues

To create a distance potential, the bin sizes were set to one angstrom and the network was trained using different sets of features. The probabilities for each bin were then converted to a

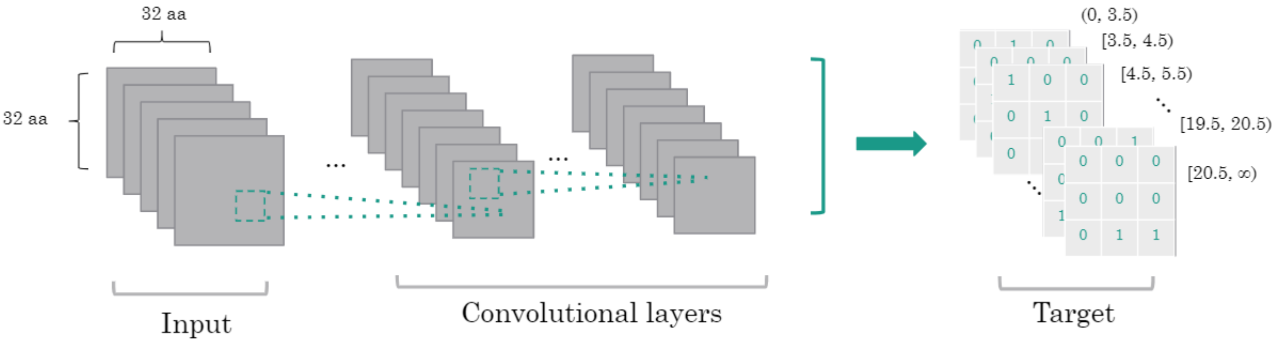


Figure 3.3: Neural network architecture for residue pair distance prediction. *Convolutional neural networks were used to train models. As output matrix, 1 Å interval bins were used. For all residue pairs in the target matrix, the real distance bin is marked with 1 and the other layers are marked as 0.*

distance potential by taking their negative log. In more detail, the network was trained intially six times, but later increased to ten for some feature sets. As a result, a score distribution for each residue pair was obtained for all residue pairs of the domain pairs. The pairs were removed if the highest score was detected in the last bin ([20.5, ∞)) since the range is too large. For a domain pair, all remaining residue pair score distributions from the six (or ten) trained models were pooled to use for interface prediction. If a residue pair had acceptable predictions from two or more networks then the final score for each distance bin was their average. The score distributions were converted to distance potentials by taking negative logarithm of the network scores. Further, these scores were subtracted from the highest score of the distribution to fix the maximum potential score to 0.

3.2.6 Predicting the structure of domain-domain interactions.

The rosetta docking application was used to predict the structure of domain-domain interactions. Predicted potentials were implemented as spline constraints on the C_β (C_α for glycine) atoms of residue pairs. For implementation of spline constraints, the predicted potentials are given in a histogram file including an *x_axis* row, which has the distance values between the residues; and a *y_axis* row, which has the corresponding predicted potential values. For constraining the distance between an atom pair, Rosetta generates a cubic spline over the data given in the histogram file using the Rosetta SplineGenerator. Experimental structures of individual domains were used for docking. The orientations of both docking partners were randomized (by the -randomize1, -randomize2 flags), the second docking partner was allowed to spin around the centre of mass of the first docking partner (by -spin flag), random perturbation of the input structure was allowed (by -dock_pert flag with recommended usage), and extra side-chain rotamers were added as recommended (by -ex1, -ex2aro flags). 4500 interfaces were generated for each domain pair and the structure with the minimum Rosetta energy (Rosetta score) was selected as a final model from the structure pool.

3.2.7 Interface evaluation

In the CAPRI docking competition, the success of the prediction of a protein complex is assessed by categorizing the quality of the interface. Four groups are defined (Table 3.1) based on f_{nat}, L-RMSD and I-RMSD values where f_{nat} is the fraction of the successfully predicted native contacts (native contacts are defined as residue pairs in domain pairs whose any heavy atom distance is ≤ 5 Å to each other). Ligand RMSD, L-RMSD, is the backbone (N, C_α, C, O) RMSD between the ligands (smaller domains) when the receptors (larger domains) are superposed. Interface RMSD, I-RMSD, is the backbone (N, C_α, C, O) RMSD of the interface residues when the interface residues are superposed. A residue is defined as an interface residue if one of its heavy atoms is close to ≤ 10 Å to the any heavy atom of the binding domain.

Table 3.1: *CAPRI competition accuracy cutoffs.*

Quality	f_{nat}	L-RMSD		I-RMSD
High	≥ 0.5	≤ 1.0	or	≤ 1.0
Medium	≥ 0.5	> 1.0	or	> 1.0
	≥ 0.3	1.0 < x ≤ 5.0	or	1.0 < x ≤ 2.0
Acceptable	≥ 0.3	> 5.0	or	> 2.0
	≥ 0.1	5.0 < x ≤ 10.0	or	2.0 < x ≤ 4.0
Incorrect	≥ 0.1	> 10.0	or	> 4.0
	< 0.1			

All scripts were written in Python (Van Rossum and Drake Jr 1995).

3.3 Results

3.3.1 Investigation of the effect of different features on the prediction accuracy.

Different feature combinations were tested to find the optimum input vector. Keeping secondary structure predictions, accessible surface area predictions, binary information, mutual information and normalized mutual information always in the input vector, the effect of DCA, DCAdot and SCA matrices were investigated. Further, the effect of chopping the input matrix from additional three different boundaries (T4 setting) was investigated. Comparisons were made on the validation set (Table 3.2).

Table 3.2: Accuracies of predictions on the 140 validation domain pairs with different feature sets.

	bin 0 - 8			bin 8 - 13			bin 13 - 18			±2 Å range		
features	accuracy[1]	# of non-zero predictions[2]	# of all predictions[3]	accuracy[1]	# of non-zero predictions[2]	# of all predictions[3]	accuracy[1]	# of non-zero predictions[2]	# of all predictions[3]	accuracy[1]	# of non-zero predictions[2]	# of all predictions[3]
DCA	0.316	61	72	0.665	75	80	0.585	76	85	0.341	84	106
SCA	0.039	2	31	0.672	41	48	0.610	48	61	0.145	55	92
DCA + SCA	0.243	62	71	0.606	78	82	0.589	86	91	0.351	85	116
DCA + DCAdot	0.300	67	77	0.614	82	86	0.566	83	94	0.297	91	125
DCA + DCAdot + SCA	0.335	68	80	0.598	89	95	0.572	97	105	0.287	100	127
DCA + DCAdot + SCA T4	0.308	72	96	0.625	101	105	0.550	103	112	0.276	108	133

[1] Mean accuracy of only non-zero predictions.

[2] Number of domain pairs which has at least one correct residue-residue prediction (non-zero prediction).

[3] Number of total domain pairs including zero and non-zero accuracy predictions.

Table 3.3: Accuracies of predictions on the 140 validation domain pairs for different feature sets and requiring a prediction to have been made by at least two of the trained networks.

	bin 0 - 8			bin 8 - 13			bin 13 - 18			±2 Å range		
features	accuracy[1]	# of non-zero predictions[2]	# of all predictions[3]	accuracy[1]	# of non-zero predictions[2]	# of all predictions[3]	accuracy[1]	# of non-zero predictions[2]	# of all predictions[3]	accuracy[1]	# of non-zero predictions[2]	# of all predictions[3]
DCA	0.369	53	61	0.677	61	64	0.588	59	65	0.440	64	77
DCA + SCA	0.301	57	67	0.641	67	67	0.598	61	68	0.455	69	86
DCA + DCAdot	0.330	61	66	0.690	67	72	0.523	64	67	0.431	73	85
DCA + DCAdot + SCA	0.344	59	68	0.681	70	72	0.530	71	76	0.419	76	89
DCA + DCAdot + SCA T4	0.339	67	76	0.669	78	78	0.544	75	81	0.433	81	102
DCA + DCAdot + SCA T4 10tr[4]	0.324	71	84	0.638	85	86	0.565	80	90	0.380	89	113

[1] Mean accuracy of only non-zero predictions.

[2] Number of domain pairs which has at least one correct residue-residue prediction (non-zero prediction).

[3] Number of total domain pairs including zero and non-zero accuracy predictions.

[4] 10tr: Ten trained networks.

Since the residue pairs whose maximum network score was detected in the last bin ([20.5, ∞)) were not taken into consideration, for some proteins no residue pair prediction could be made, resulting in no interface prediction. Accuracies were calculated for three distance bins (0 - 8 Å, 8 - 13 Å and 13 - 18 Å) as well as for ±2 Å range. Average accuracies were calculated over the domain pairs whose accuracy is greater than zero. The number of domain pairs whose accuracy is greater than zero and the number of total domain pairs for which at least one residue interaction prediction were made are given in Table 3.2.

The average accuracies of 8 - 13 Å and 13 - 18 Å bin are higher than those in the 0 - 8 Å bin for all feature vectors. Although it is not clear why, it emphasizes the importance of predicting not only contacting pairs but also the residue pairs at longer distances. Although 0 - 8 Å, 8 - 13 Å and 13 - 18 Å bin accuracies gives us insight about the prediction success in different intervals, for the network comparisons we will focus on ±2 Å range accuracies as it gives information about the success of the prediction in a continuous metric rather than a discrete metric (as in 0 - 8 Å, 8 - 13 Å and 13 - 18 Å accuracies).

Accuracies presented in Table 3.2 indicate that DCA predictions form the backbone of the calculations, as removal of it (and using SCA only) reduces the number of domain pairs for which predictions could be made, i.e. for which there were predictions of less than 20.5 Å separation, and reduces the accuracy of any predictions. While there are minimal variations in 0 - 8 Å, 8 - 13 Å and 13 - 18 Å accuracies for DCA, DCA + SCA, DCA + DCAdot and DCA + SCA + DCAdot, the maximum number of predictions was obtained training with DCA + SCA + DCAdot, suggesting a more generalised model. Although the ±2 Å range accuracy of DCA + SCA is ~ 0.06 higher than DCA + SCA + DCAdot, predictions could be made for 100 domain pairs, i.e. 100 pairs had at least one residue pair predicted to be less than 20.5 Å A apart, for the latter, whereas there are only 85 predictions for the former. These results suggested that inclusion of all features (DCA, SCA and DCAdot) provides better generalization without any loss in prediction accuracy.

As mentioned above, the input matrix was further chopped from additional three boundaries to increase the input data size (T4 setting). This step (DCA + SCA + DCAdot T4) increased the number of domain pairs for which predictions could be made, with a slight change in the accuracy.

To reduce the number of incorrect predictions, for a residue pair to be considered as a predicted distance it needed to be predicted as such by at least two of the six (or ten) trained networks. This modification in selection criteria increased accuracies but reduced the number of domain pair that predictions were made for, as shown in Table 3.3. Increasing the number of trainings from six to ten, increasing the number of domain pairs predictions that were made but caused a slight decrease in ±2 Å range predictions.

Although it is clear that accepting a residue pair if it was predicted at least two times leads to better predictions, which feature set from Table 3.3 is the best one is challenging to decide. Average ±2 Å accuracy of DCA, DCA + SCA, DCA + DCAdot + SCA and DCA + DCAdot + SCA T4 settings varies with the lowest ±2 Å accuracy of 0.419 while the number of domain predictions whose accuracy is larger than zero increases, suggesting that the addition of DCAdot and SCA, and chopping the matrix from three additional boundaries improve the generalization of the network. Although increasing the number of the trained network from six to ten decreases the average ±2 Å accuracy, since more non-zero accuracy predictions were made with DCA + DCAdot + SCA T4 10tr feature set, predictions of this feature set were selected for further analysis.

Accuracy of residue pair predictions on the test set based on the models trained with DCA + DCAdot + SCA T4 10tr feature set are given in Table 3.4. Among 120 test domain pairs, predictions were able to be performed for 91 of them. Average accuracies for test domain pairs are similar to the validation domain pairs (Table 3.3).

Table 3.4: *Accuracies of predictions on the 120 test domain pairs with DCA + DCAdot +SCA + B T4 10tr feature set. At least two occurancse among all trainings was used as selection criteria for a residue pair to be taken into account.*

	bin 0 - 8			bin 8 - 13			bin 13 - 18			±2 Å range		
features	accuracy[1]	# of non-zero predictions[2]	# of all predictions[3]	accuracy[1]	# of non-zero predictions[2]	# of all predictions[3]	accuracy[1]	# of non-zero predictions[2]	# of all predictions[3]	accuracy[1]	# of non-zero predictions[2]	# of all predictions[3]
DCA + DCAdot +SCA + B T4 10tr[4]	0.347	67	73	0.629	74	76	0.561	76	77	0.397	78	91

[1] Mean accuracy of only non-zero predictions.

[2] Number of domain pairs which has at least one correct residue-residue prediction (non-zero prediction).

[3] Number of total domain pairs including zero and non-zero accuracy predictions.

[4] 10tr: Ten trained networks.

3.3.2 Distance potentials were generated from network score distribution.

Based on the results of the previous section the feature vector was chosen as DCA + DCAdot + SCA T4 10tr, and the network was trained six times with distance bins of 1 Å width (Fig. 3.4), from which a distance potential was calculated. The network score distribution was converted into a distance potential by calculation negative log of the score (Fig. 3.4). As an example, the predicted distance potentials of one of the test proteins are shown in Fig. 3.5. Distance distributions for each pair are shown in separate subplots and real distance for that residue pair is shown with a red line and 8 Å contact threshold is shown with a green line. Successful predictions include both contacting and non-contacting residue pairs.

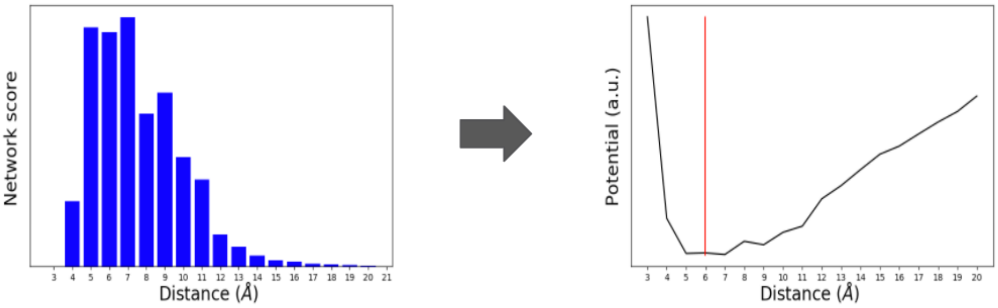


Figure 3.4: Distance potentials were calculated from network score distribution. *For a chosen residue pair the distribution of network scores was converted into a distance potential by calculation of negative logarithm.*

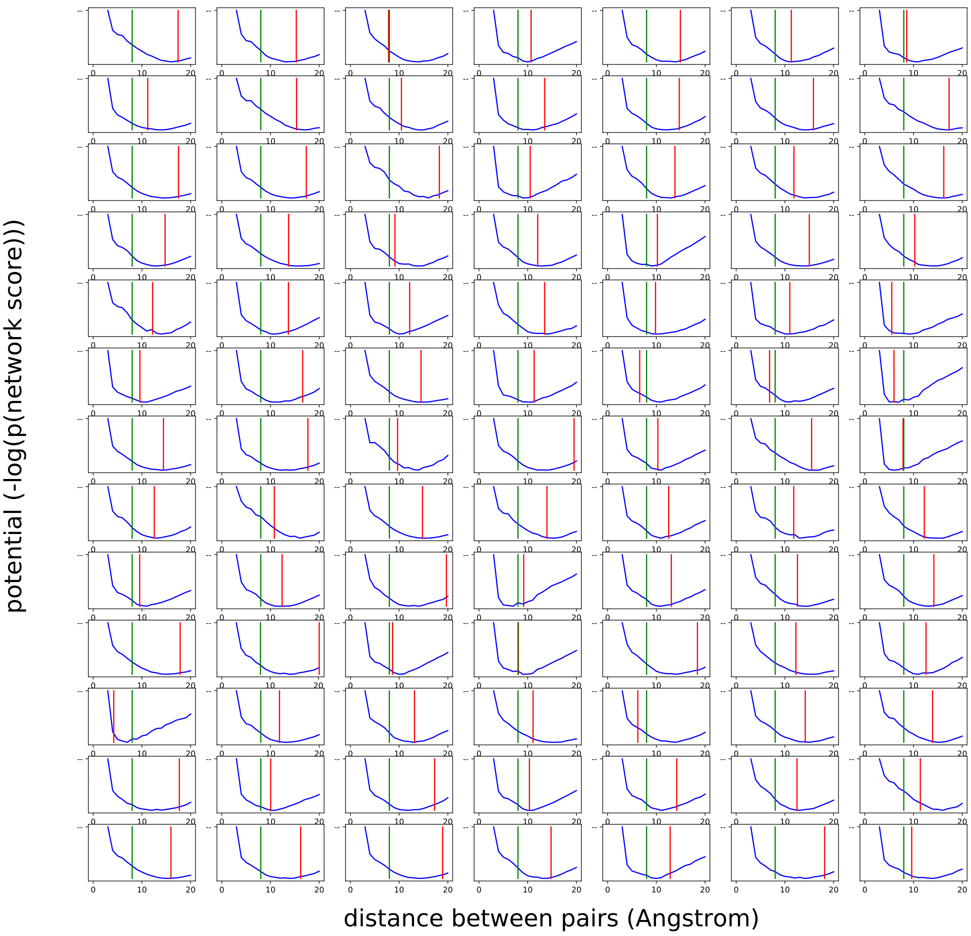


Figure 3.5: Predicted distance distributions of residue pairs of a test set domain pair. *Distance potentials were determined by calculating negative log likelihood of neural network score for that protein. The potential values were substracted from the highest score in the distribution (blue lines). Real distance between the residue pairs are shown with red lines, 8 Å contact threshold is shown with green lines.*

3.3.3 The structures of domain domain interactions were predicted from the distance potentials.

The distance potentials, described in the previous section, were applied as constraints within Rosetta Dock to predict the structure of domain interactions. The success of a prediction was evaluated using the criteria of the CAPRI docking competition, which assesses the quality of the interface between the two docked proteins (Table 3.1). Besides applying the constraints from DCA + DCAdot +SCA T4 10w tr feature set, structures were predicted without applying predicted distance potential constraints, to compare the results with constraints and without constraints. Out of 91 domain pair predictions, 17 of the interface structures have high accuracy, 27 have medium accuracy and 6 have acceptable accuracy; whereas the predicted interfaces for 41 domain pairs are incorrect based on CAPRI evaluation metric (Table 3.5, Fig. 3.6). Examples for high, medium, acceptable and incorrect interface predictions are given in Fig. 3.7. When predicted distance distributions were not applied as constraints, only 7 domain interfaces were predicted with high accuracy, 3 domain interfaces were predicted with medium accuracy, and incorrect predictions were made for 81 domain pairs. Interestingly for six domain pairs, prediction quality is better without restraints, probably because the introduction of the unsuccessfully predicted residue pair constraints forces the domain pairs have wrong orientations. However, for 45 domain pairs, having constraints improves the quality of the predicted structure. This suggests that implementation of the distance potentials provides successful domain interface predictions.

The domain pairs with higher ±2 Å range accuracy correlated with more successful interface predictions (Fig. 3.6). When the ±2 Å range accuracy is lower than ~ 0.3, incorrect interfaces were predicted; whereas, when ±2 Å range accuracy is higher than ~ 0.3, acceptable, medium and high predictions were made. This result is not surprising as a more accurate distance potential is expected to provide more accurate interface predictions.

As the Rosetta score does not always select the best model from the structure pool, the quality of the structures with the lowest I-RMSD was recorded for all domain types. Predic-

Table 3.5: The number of high, medium, acceptable and incorrect domain pairs. The structure with the minimum Rosetta energy was selected as the final model.

	high	medium	acceptable	incorrect
No constraints	7	3	0	81
DCA + DCAdot +SCA T4 10w tr	17	27	6	41

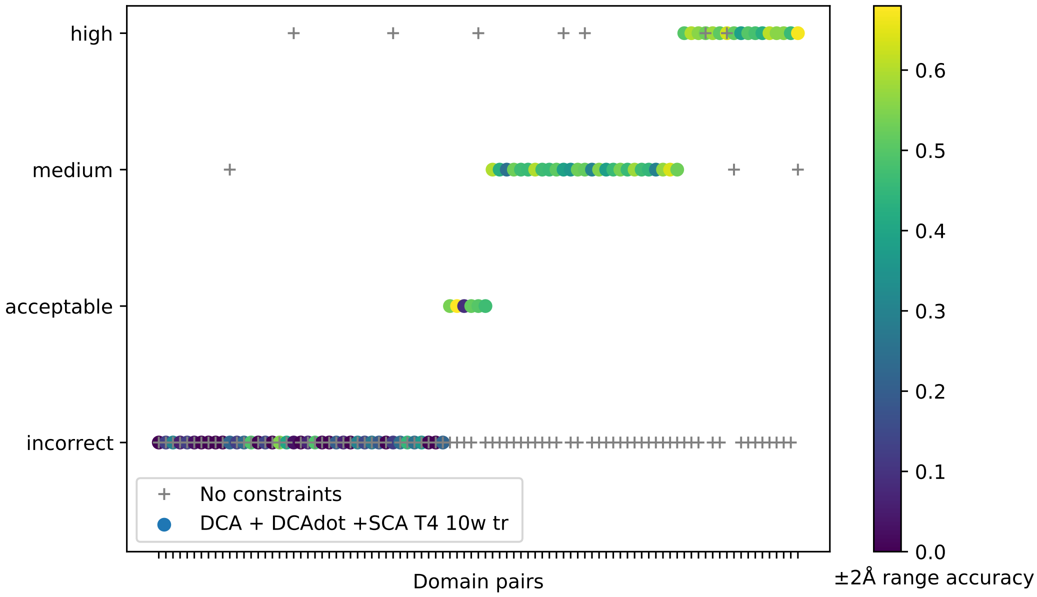


Figure 3.6: Quality evaluation of predicted domain interfaces with and without constraints. *When constraints were not applied only ten interfaces could be predicted correctly; whereas, 50 interfaces could be predicted correctly when predicted distance poteintials were applied. For the domain pairs with ±2 Å range accuracies, interfaces could be predicted with at least acceptable quality.*

tions without inter-residue distance constraints the choosing the model with lowest I-RMSD gives 15 high, 25 average, and 31 medium accuracy predictions and 20 incorrect predictions. Application of predicted distance potentials predicted 34 interfaces with high accuracy, 31 interfaces with medium accuracy and 7 interfaces with acceptable accuracy, whereas no correct predictions were made for 19 domain interfaces (Table 3.6, Fig. 3.8). The models of domain interactions with the highest accuracy interfaces that were not selected because they didn't have the lowest Rosetta energy were those with low ±2 Å accuracy. This is not surprising as Rosetta energy includes constraint energy and unsuccessful constraints contribute to the Rosetta score and misleads the correct structure selection.

Comparison of Table 3.5 and Table 3.6 results suggest that even though there are more

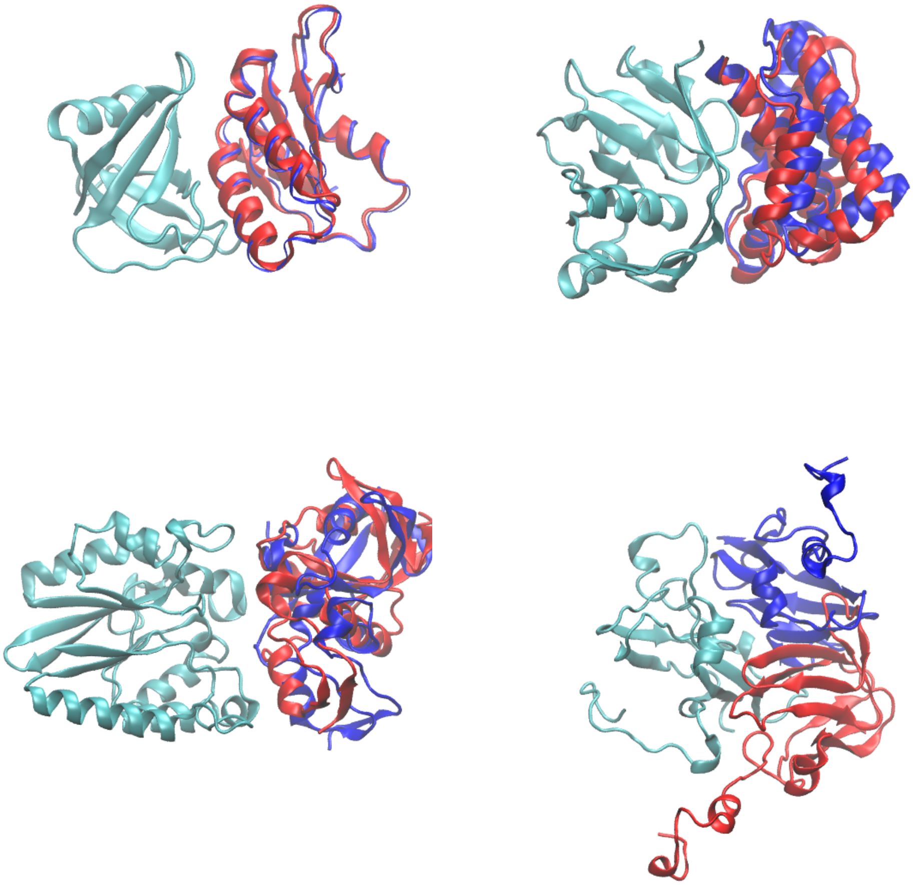


Figure 3.7: Example structures for high (upper left, PDB id:1qjf), medium (upper right, PDB id:5ayv), acceptable (lower left, PDB id:1oi8), and incorrect (lower right, PDB id:1kcw) interface predictions. *The superposed domain of the pair is coloured cyan, the experimental orientation is coloured blue and the predicted orientation is coloured red.*

successful predictions in the structure pool, we cannot pick them as the final model. Although the number of different domain pairs with at least ”acceptable” predictions is almost the same with and without constraints (72 and 71, respectively), the number of high and medium quality interfaces is greater for the predictions using constraints. When no constraints were applied, only the structures of 10 domain pairs, selected as having the lowest Rosetta energy from a pool of 4500 putative structures per domain pair, had ”acceptable” or better interfaces, whereas the structure pools actually contained ”acceptable” or better predictions for 71 different domain pairs. On the other hand, when constraints were applied, classifying the best model using the Rosetta energy, including the energies associated with the constraints gave 50 different domain pairs with ”acceptable” or better interfaces and in total 72 of them have correct predictions in the structure pools. Therefore, these results indicate that implementation of distance potentials improves the quality of the correct interfaces, as well as providing to select better models from the structure pool.

Table 3.6: *Number of high, medium, acceptable and incorrect domain pairs. The structure with the lowest I-RMSD value was selected as final structure.*

feature	high	medium	acceptable	incorrect
No constraints	15	25	31	20
DCA + DCAdot +SCA T4 10w tr	34	31	7	19

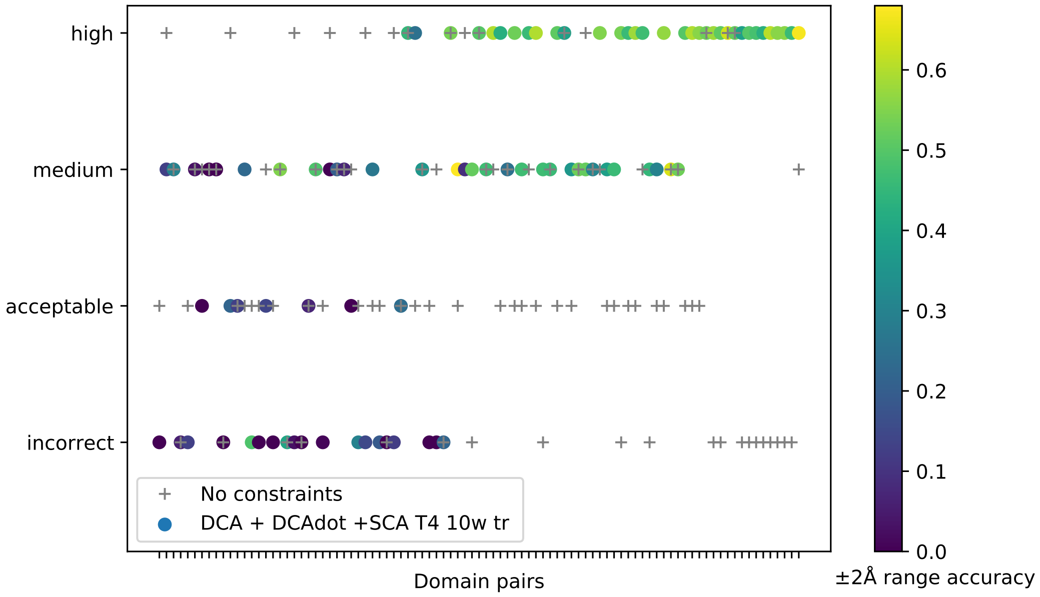


Figure 3.8: Quality evaluation of predicted domain interfaces with and without constraints when the best structure from the pool was selected. *When constraints were applied more interfaces were generated with high and medium accuracy.*

3.3.4 Effect of Nf and real contacting pair number on the prediction quality.

Since successful predictions could not be made for all domain pairs in the test set, we investigated the possible reasons for that. The Nf value (i.e. the number of sequences in the alignment with maximum 80% pairwise sequence identity, divided by the square root of the sequence length) is critical for the successful prediction of residue pairs as discussed in Chapter 2. When the alignment is not comprehensive enough, coevolution of residue pairs cannot be detected successfully leading to inaccurately predicted protein structures. The surface area of the interface, on the other hand, can also be expected to have an importance for the prediction of the correct interface. Since the larger surface area means more interacting residue pairs between the domains, for the proteins with large surface area, more residue pairs can be predicted correctly, which would assist accurate prediction of the interfaces. Therefore, we investigated how the ±2 Å range accuracy, f_{nat}, I-RMSD and L-RMSD scores change with the Nf values and the number of real contacts (within 0 - 8 Å distance) in the experimental structures (Fig. 3.9).

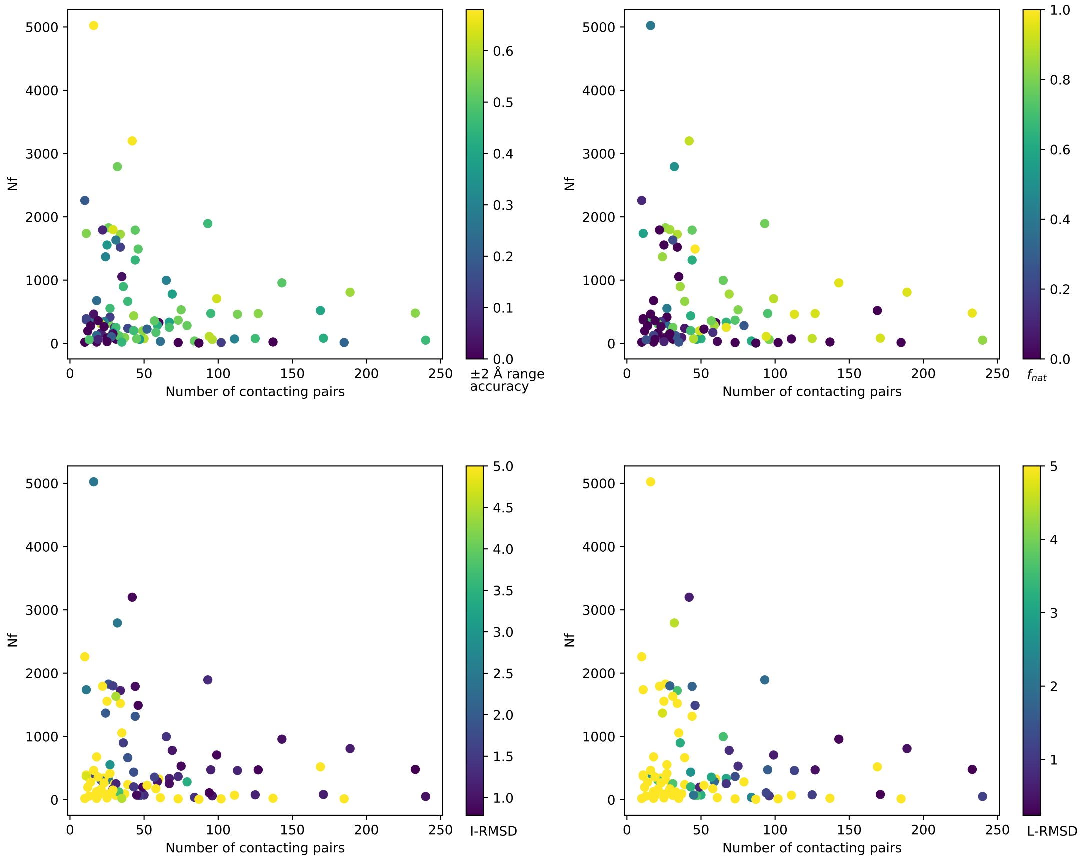


Figure 3.9: The success of predictions using distance constraints increases as the number of real residue contacts and Nf values increase. *The interface of pairs of domains is predicted best where they have a high number of contacts in the experimental structure and a high Nf value in the sequence alignment. The number of real residue contacts are determined based on the C_{β} (C_{α} for glycine) atom distance between two residues being less than 8 Å.*

For all four metrics, the prediction success increases as the Nf value and the number of real contacts increase (Fig. 3.9). The accuracy of the ±2 Å range increases as the number of contacting pairs increase excepting a few domain pairs with low Nf values. Similarly, the fraction of successfully predicted native contacts, f_{nat}, increases as there are more interacting residue pairs in the domain pairs. The residue pairs that have a high number of contacting pairs but have low f_{nat} have mostly low Nf values. For the domain pairs with a high number of contacting pairs, I-RMSD and L-RMSD are low except for the domain pairs with low Nf values.

As monotonic relationships are observed between the number of real contacts vs. the interface evaluation metrics (±2 Å range accuracy,f_{nat}, I-RMSD and L-RMSD), and Nf values vs. the interface evaluation metrics, Spearman's rank correlation coefficient was calculated to determine whether there is any significant monotonic correlation (Table 3.7). For all comparisons weak but significant correlations are detected. Since both Nf values and real contact numbers seem to contribute to the accuracy of the predicted structure of the interfaces, Spearman's correlation coefficient correlations were calculated for subsets of test domain pairs with higher real contact numbers (>40) (Table 3.8) and higher Nf values (>300) (Table 3.9) to analyze their effect individually. Correlation coefficients (ρ) and corresponding p-values reveal a clearer relationship between the Nf values and the interface quality, when the domain pairs with lower real contact pair number are discarded, as higher significant correlations can be detected (Table 3.8). Similarly, correlation coefficients (ρ) and corresponding p-values reveal a clearer relationship between the real residue contact number and the interface quality, when the domain pairs with the lower Nf values are discarded, since higher significant correlations can be detected (Table 3.9). Overall, it is seen that although there are a few exceptions, the common trend is that when the surface area is larger, and when the sequence alignment is comprehensive enough, the quality of the predicted interfaces is better.

3.3.5 The effect of the number of predicted residue pairs on the prediction quality.

The relation between the number of predicted residue pairs and the prediction quality was investigated. Analysis suggests that when there are more constriants predicted for a domain pair,

Table 3.7: *Spearman's rank correlation coefficient (ρ) and corresponding p-values between the Nf value and the interface evaluation metrics (±2 Å range accuracy,f_{nat}, I-RMSD and L-RMSD), and the number of real contacts and the evalution metrics.*

	±2 Å	f_{nat}	I-RMSD	L-RMSD
Nf	ρ= 0.34	ρ=0.31	ρ=-0.35	ρ=-0.31
	p=8.90E-04	p=2.56E-03	p=5.67E-04	p=2.71E-03
real contacts	ρ= 0.35	ρ=0.46	ρ=-0.40	ρ=-0.50
	p=7.55E-04	p=4.62E-06	p=7.78E-05	p=5.59E-07

Table 3.8: *Spearman's rank correlation coefficient (ρ) and corresponding p-values between the Nf value and the interface evaluation metrics (±2 Å range accuracy,f_{nat}, I-RMSD and L-RMSD), and the number of real contacts and the evalution metrics for the subset of test domain pairs (44 domain pairs) whose real contact pairs are greater than 40.*

	±2 Å	f_{nat}	I-RMSD	L-RMSD
Nf	ρ= 0.39	ρ=0.48	ρ=-0.47	ρ=-0.47
	p=9.68E-03	p=9.54E-04	p=1.19E-03	p=1.17E-03
real contacts	ρ= -0.07	ρ=0.12	ρ=0.00	ρ=-0.19
	p=6.33E-01	p=4.57E-01	p=9.82E-01	p=2.22E-01

Table 3.9: *Spearman's rank correlation coefficient (ρ) and corresponding p-values between the Nf value and the interface evaluation metrics(±2 Å range accuracy,f_{nat}, I-RMSD and L-RMSD), and the number of real contacts and the evalution metrics for the subset of test domain pairs (47 domain pairs) whose Nf values are greater than 300.*

	±2 Å	f_{nat}	I-RMSD	L-RMSD
Nf	ρ= 0.20	ρ=0.18	ρ=-0.11	ρ=-0.08
	p=1.67E-01	p=2.28E-01	p=4.45E-01	p=5.96E-01
real contacts	ρ= 0.35	ρ=0.59	ρ=-0.62	ρ=-0.65
	p=1.46E-02	p=1.09E-05	p=3.71E-06	p=7.77E-07

it is more likely to have a better interface (Fig. 3.10). On the other hand, it is not clear why for some of the domain pairs there are drastically more predictions (i.e outliers).

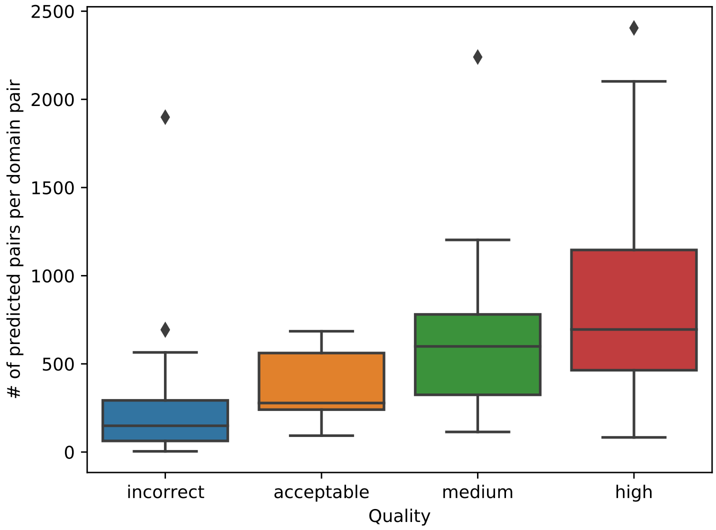


Figure 3.10: Comparison of the number of predicted residue pairs per domain pair between the quality groups. *The number of predictions per domain pair is higher in the domain pairs with higher quality interface predictions. The boundaries of the box range from the lower quartile of the data to the upper quartile, the line shows the median, the whiskers demonstrate the range of the data. Outliers are shown with black diamond signs.*

3.3.6 Domain-domain interaction prediction on a multidomain protein.

Fatty acid synthases (FAS) is a multi-domain protein complex whose domain organization is very similar to polyketide synthases (PKS). Since the crystal structure of wild boar, *Sus scrofa*, FAS was determined at 3.2 Å (PDB ID:2vz9), we used this FAS as a model multi-domain protein complex to test the success of our domain-domain prediction methodology. The domain organization of the FAS is given in Fig. 3.11.

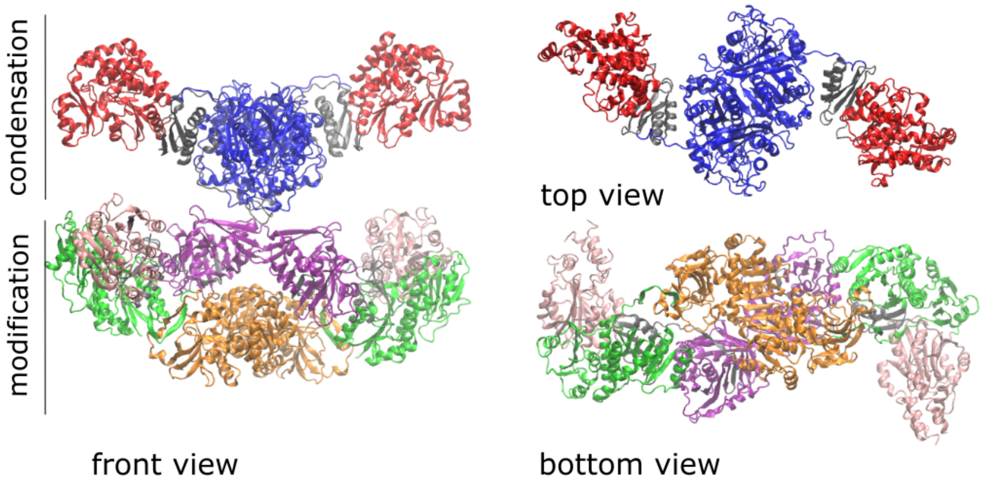


Figure 3.11: Experimentally determined domain organization of a fatty acid synthase structure. *. KS:ketosynthase, LD: ketosynthase - acyltransferase linker domain, MAT: acyltransferase (malonyl transferase), KR: ketoreductase, ψKR: pseudo-ketoreductase, DH: dehydratese, ER: enoyl reductase. PDB id: 2vz9.*

To analyze the interactions between FAS domains, HHblits was run to find homologous sequences (the whole FAS sequence including ACP and TE domain sequences was used). There are 1999 sequences in the alignment including sequences from PKSs. Only the sequences that are labelled as an FAS system or as uncharacterized were kept in the alignment resulting in 715 sequences in the MSA, 136 of which were labelled as FASs.

From the MSA, the positions of the selected domain pairs were extracted. The selected domain pairs are (i) KS - LD and MAT, (ii) KS and DH, (iii) DH and ER, (iv) DH and KR, (v) ψKR and KR, (vi) KS and LD and (vii) LD and MAT. The extracted alignments were filtered to remove the sequences with a high number of gaps (minimum ≥ 80% coverage with the query sequence). The number of residue pairs in the contact, 8-13 Å bin and 13-18 Å bin, as determined from the experimental structure for these domain pairs are given in Table 3.10 as an indicator of the interface area.

Table 3.10: *Real contact, bin8-13 and bin13-18 numbers of interactions between two domains.*

domain pairs	contact	bin 8-13	bin 13-18
KS-LD and MAT	35	250	699
KS and DH	0	46	247
DH and ER	3	68	300
DH and KR	30	212	60
ψKR and KR	5	40	176
KS and LD	245	1291	2821
LD and MAT	35	243	659

For the first four domain pairs listed in the Table3.10, residue distance potentials were predicted. Although predictions could be done for all domain pairs, accuracies for all bins were 0. Therefore, further predictions were performed by keeping all sequences (1999 sequences, including the ones from PKSs) in the alignment. From this alignment, again the positions of the selected domain pairs were extracted and the alignment was filtered to remove the highly gapped sequences. Accuracies are given in Table 3.11.

Table 3.11: *FAS domain pair predictions. The ratio of the number of the correct predictions to the number of all predictions is given for each bin.*

domain pairs	contact	bin 8-13 Å	bin 13-18 Å	± 2 Å	# of seqs in MSA	length	Nf
KS-LD and MAT	0	0	0	0	1701	792	49.6
KS and DH	0	0	0	0	1715	592	58.9
DH and ER	0	0	0	0	1736	548	66.0
DH and KR	0	3/7	4/20	7/458	1899	414	81.8
ψKR and KR	0	0	2/3	1/580	1283	367	59.8
KS and LD	26/140	295/472	261/470	567/1456	1663	473	62.6
LD and MAT	1/8	11/14	2/7	8/323	1891	431	79.4

For all tested FAS domain pairs, Nf values are low; therefore, the expected accuracies would be low based on Fig. 3.9. For KS-LD and MAT, KS and DH, and DH and ER domain pairs, no successful predictions could be made; whereas, for DH and KR, ψKR and KR, KS and LD, and LD and MAT pairs at least one successful predictions could be obtained. When the surface is area is very large (as in KS and LD domain pair), it was possible to obtain successful predictions. On the other hand, when the surface area is very low (as in KS and DH domain pair and DH and ER domain pair), no successful predictions were obtained. Although the surface area of ψKR and KR is low, two out of three predictions in 13-18 Å bin were correct, whereas only one predicted distance out of 580 predictions was within 2 Å away from the experimental distance.

Another interesting observation is the prediction of the interface between the LD and MAT domains. KS-LD and MAT pair and LD and MAT pair have the same interface surface (i.e. there is no direct interaction between the KS and the MAT domains). When the structure of the interface between the LD and the MAT domains was aimed to be predicted by including the KS domain, no successful prediction could be obtained. However, when KS domain was discarded, successful residue-residue interaction predictions were obtained. This difference in the success of predicting the same interface structure can be caused by two reasons. First one is, for the KS-LD and MAT system, the Nf value is smaller (as the sequence length is larger) and that may cause worse predictions. The second possible reason is having another domain, KS, which also interacts with the LD domain, can introduce noise.

Overall, the interaction prediction between the FAS domain pairs gives consistent results with the previous analysis. For successful residue pair predictions on a domain pair, large interface area and comprehensive sequence alignment are necessary.

3.4 Discussion

This chapter aimed to investigate the prediction of multi-domain protein structures without needing a template for the domain-domain interactions. Convolutional neural networks were used to predict distance potentials between pairs of residues on the two domains that are interacting. The predicted distances were applied as constraints on the domain pairs to predict domain interfaces correctly. There are studies in the literature aiming to predict how two domains interact without using any template for the interface (Ovchinnikov *et al.* 2014; Zeng *et al.* 2018), where they aim to predict contacting pairs only, not the distances between the residue pairs. Moreover, they generate the alignments of two domains separately and then match the sequences for further analysis on the contrary to concatenating the sequences and finding the homologous sequences. To the best of my knowledge, this is the first study predicting the distance potentials between the residue pairs of two same chain domains to predict correct interfaces.

In addition to DCA matrix, SCA and DCAdot matrices were used in the feature vector, and their contribution was tested. In Table 3.3, it is seen that the implementation of SCA matrix slightly increases both the number of domain pairs for which predictions were made, and mean accuracy. Implementation of DCAdot also increases the number of domain pairs for which predictions were made; whereas, the average accuracy slightly decreases. And using both of them increases the number of domain pairs for which predictions were made, further with again a slight decrease in the average accuracy. Overall, implementation of SCA and DCAdot matrices allow us to make non-zero accuracy predictions for additional 12 domain pairs, yet result in $\sim$ 0.02 decrease in the average accuracy. To the best of my knowledge, we are the first group using SCA and DCAdot matrices in the feature vector and demonstrate their contribution to the prediction accuracy.

Although successful distance potentials and domain interfaces were predicted, correct interfaces could not be obtained for all of the domain pairs in the test set. One of the reasons for that result is almost certainly an insufficient number of domain pairs in the training set. Usually, the training set includes thousands of samples for successful training and generalization.

For example, the AlphaFold training set included almost 30,000 training proteins, whereas our training set included only 804 domain pairs. Chopping the input data into smaller matrices and choosing four additional chopping boundaries improved the success of the network (Table 3.2, 3.3), yet more data would be helpful.

As the success of correct detection of residue pairs depends on the alignment depth, the predicted interface accuracy also depends on the Nf values of the domain pairs. Therefore, not surprisingly, when there are fewer sequences in the alignment, the accuracy of the predictions decreases. Moreover, the quality of the prediction also depends on the interface surface area. When the interface surface area is larger, the correct interface can be predicted (Fig. 3.9). We also demonstrated that, if the number of predicted residue pairs is high for a domain pair, it is an indication that an interface prediction with at least acceptable quality can be obtained (Fig. 3.10).

The ranking of models in both the protein structure prediction area and the protein (or domain) docking area is an important challenge. We might be able to generate a correct structure in a pool of model structures but if we have no good ranking method then we cannot identify it as the correct structure. Using Rosetta energy to pick the final model is a very time efficient and straightforward approach to pick the final model. However, unfortunately, the best prediction cannot be selected most of the time. Implementation of distance potentials for docking of two domains gives more successful predictions in the pool and helps to select better models as the Rosetta score includes the constraint energy.

The success of the method was tested on a multi-domain protein complex, fatty acid synthase which has a similar domain organization as polyketide synthases. Similar to the test set proteins, successful residue pair predictions obtained as long as there were enough sequences in the alignment (high Nf value), and the surface area was large.

Overall, in this study, successful predictions of domain domain interactions were achieved without using an experimental model of the complex. Although limitations are causing unsuccessful predictions for some cases, our method can be used to predict domain interfaces correctly, which should allow the prediction of multi-domain protein structures.

Chapter 4

Determination of coevolved residue groups on DEBS module 1

4.1 Overview

Understanding the structure and working mechanism of multi-domain proteins and protein complexes is important to be able to modify them for many protein engineering purposes including the development of novel drug candidates. Here, we focus on polyketide synthases (PKSs), which are multi-domain proteins or protein complexes, producing polyketides that have various functions including antibiotic, antitumor, antifungal effects. Protein engineering applications have been widely applied to PKSs to generate novel polyketides. However, most of the experiments fail since the working mechanism of PKSs are not known in detail yet Weissman 2016. Domain swapping and site-directed mutagenesis studies performed to manipulate polyketide production resulted in either low yield product or no product at all. Experimental evidence suggests that domain swapping studies without thorough optimization of the functional boundaries of the domains and applying point mutations on only a few residues are not adequate to shift the specificity of the domains successfully Weissman 2016; Barajas *et al.* 2017; Musiol-Kroll and Wohlleben 2018; Kornfuehrer and Eustáquio 2019. In this chapter, it will be demonstrated that co-evolved networks of residues (independent components and sectors) can be detected by statistical coupling analysis (SCA) and that these residue networks have specific functions within a

PKS, notably defining domain boundaries consistent with experimental data. Further, we detect residue groups that might make important contributions to the domain subtype functionality. Detection of functional domain boundaries and residue groups specific to the domain sub-types can enhance the success of the experimental studies to generate new drug candidates.

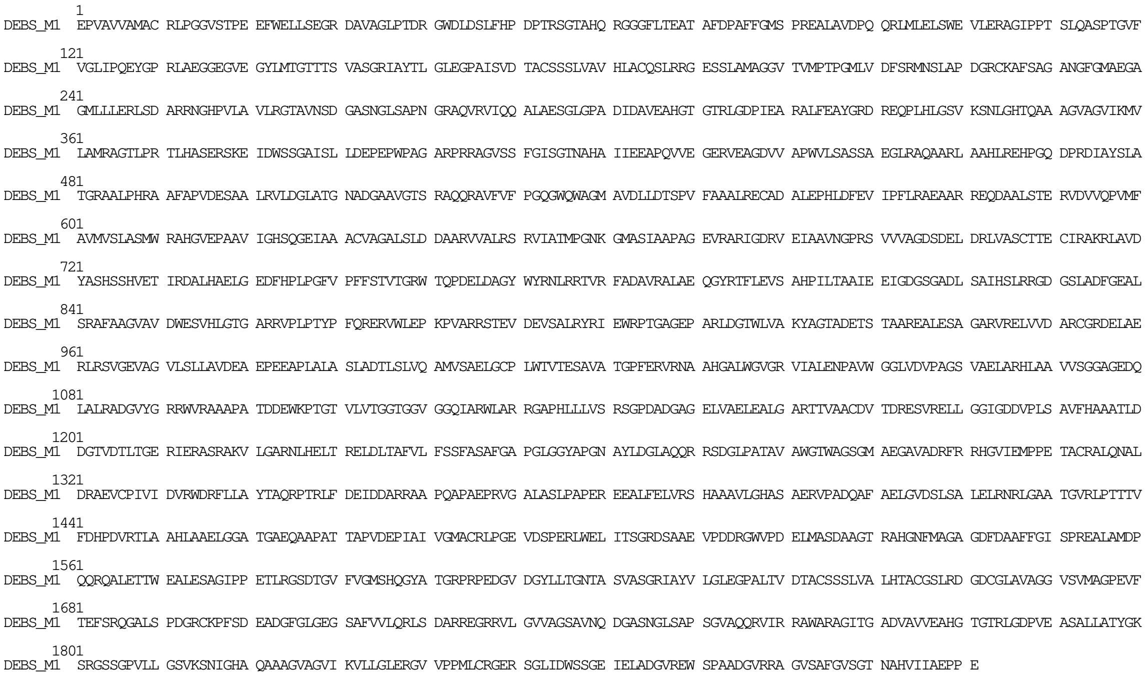
In this chapter, the methodology will be explained. On several model systems, I tested the number of sequences needed for a convergence in the analysis, which has not been addressed before. Then, I continue with the detailed steps of determination of coevolved amino acids in the target system, which is the first module of DEBS, explained further in Chapter 1.5. Further, I applied sequence-position mapping on the alignment including the determination of residues that might make important contributions to subtype functionality. This section will be followed by the results and discussion section demonstrating the detected residue networks and their relation with experimental evidence. Lastly, concluding marks will be given summarizing the outcomes of the study.

4.2 Methods

4.2.1 Multiple Sequence Alignment Generation

The sequence of DEBS1 module 1 together with the KS of module 2 was selected as the target sequence (UniprotID:Q03131) (Fig 4.1). Homologous sequences were detected with HHblits Remmert *et al.* 2011 and from which a multiple sequence alignment was generated. For each sequence in the alignment, pfam domains El-Gebali *et al.* 2018 were determined via hmmscan Eddy 2011 where e-value was set to 0.001. Sequences whose original module includes domains other than KS, AT, KR, DH, ER, TE and ACP were removed from the alignment. After preprocessing the MSA to improve its quality, as described in ref. Rivoire *et al.* 2016, e.g. to remove highly gapped columns and sequence fragments, the final alignment included 2303 sequences.

Preproccesing on the MSA was applied with the scaProcessMSA.py script in SCA analysis tool Rivoire *et al.* 2016 and the reference sequence was set to the target sequence by –refindex 0 command. Second and third steps of the analysis, which applies the SCA method and determines the independent components, were performed by scaCore.py and scaSector.py scripts from the same tool package.



```
           1
DEBS_M1  EPVAVVAMAC RLPGGVSTPE EFWELLSEGR DAVAGLPTDR GWDLDSLFHP DPTRSGTAHQ RGGGFLTEAT AFDPAFFGMS PREALAVDPQ QRLMLELSWE VLERAGIPPT SLQASPTGVF
         121
DEBS_M1  VGLIPQEYGP RLAEGGEGVE GYLMTGTTTS VASGRIAYTL GLEGPAISVD TACSSSLVAV HLACQSLRRG ESSLAMAGGV TVMPTPGMLV DFSRMNSLAP DGRCKAFSAG ANGFGMAEGA
         241
DEBS_M1  GMLLLERLSD ARRNGHPVLA VLRGTAVNSD GASNGLSAPN GRAQVRVIQQ ALAESGLGPA DIDAVEAHGT GTRLGDPIEA RALFEAYGRD REQPLHLGSV KSNLGHTQAA AGVAGVIKMV
         361
DEBS_M1  LAMRAGTLPR TLHASERSKE IDWSSGAISL LDEPEPWPAG ARPRRAGVSS FGISGTNAHA IIEEAPQVVE GERVEAGDVV APWVLSASSA EGLRAQAARL AAHLREHPGQ DPRDIAYSLA
         481
DEBS_M1  TGRAALPHRA AFAPVDESAA LRVLDGLATG NADGAAVGTS RAQQRAVFVF PGQGWQWAGM AVDLLDTSPV FAAALRECAD ALEPHLDFEV IPFLRAEAAR REQDAALSTE RVDVVQPVMF
         601
DEBS_M1  AVMVSLASMW RAHGVEPAAV IGHSQGEIAA ACVAGALSLD DAARVVALRS RVIATMPGNK GMASIAAPAG EVRARIGDRV EIAAVNGPRS VVVAGDSDEL DRLVASCTTE CIRAKRLAVD
         721
DEBS_M1  YASHSSHVET IRDALHAELG EDFHPLPGFV PFFSTVTGRW TQPDELDAGY WYRNLRRTVR FADAVRALAE QGYRTFLEVS AHPILTAAIE EIGDGSGADL SAIHSLRRGD GSLADFGEAL
         841
DEBS_M1  SRAFAAGVAV DWESVHLGTG ARRVPLPTYP FQRERVWLEP KPVARRSTEV DEVSALRYRI EWRPTGAGEP ARLDGTWLVA KYAGTADETS TAAREALESA GARVRELVVD ARCGRDELAE
         961
DEBS_M1  RLRSVGEVAG VLSLLAVDEA EPEEAPLALA SLADTLSLVQ AMVSAELGCP LWTVTESAVA TGPFERVRNA AHGALWGVGR VIALENPAVW GGLVDVPAGS VAELARHLAA VVSGGAGEDQ
        1081
DEBS_M1  LALRADGVYG RRWVRAAAPA TDDEWKPTGT VLVTGGTGGV GGQIARWLAR RGAPHLLLVS RSGPDADGAG ELVAELEALG ARTTVAACDV TDRESVRELL GGIGDDVPLS AVFHAAATLD
        1201
DEBS_M1  DGTVDTLTGE RIERASRAKV LGARNLHELT RELDLTAFVL FSSFASAFGA PGLGGYAPGN AYLDGLAQQR RSDGLPATAV AWGTWAGSGM AEGAVADRFR RHGVIEMPPE TACRALQNAL
        1321
DEBS_M1  DRAEVCPIVI DVRWDRFLLA YTAQRPTRLF DEIDDARRAA PQAPAEPRVG ALASLPAPER EEALFELVRS HAAAVLGHAS AERVPADQAF AELGVDSLSA LELRNRLGAA TGVRLPTTTV
        1441
DEBS_M1  FDHPDVRTLA AHLAAELGGA TGAEQAAPAT TAPVDEPIAI VGMACRLPGE VDSPERLWEL ITSGRDSAAE VPDDRGWVPD ELMASDAAGT RAHGNFMAGA GDFDAAFFGI SPREALAMDP
        1561
DEBS_M1  QQRQALETTW EALESAGIPP ETLRGSDTGV FVGMSHQGYA TGRPRPEDGV DGYLLTGNTA SVASGRIAYV LGLEGPALTV DTACSSSLVA LHTACGSLRD GDCGLAVAGG VSVMAGPEVF
        1681
DEBS_M1  TEFSRQGALS PDGRCKPFSD EADGFGLGEG SAFVVLQRLS DARREGRRVL GVVAGSAVNQ DGASNGLSAP SGVAQQRVIR RAWARAGITG ADVAVVEAHG TGTRLGDPVE ASALLATYGK
        1801
DEBS_M1  SRGSSGPVLL GSVKSNIGHA QAAAGVAGVI KVLLGLERGV VPPMLCRGER SGLIDWSSGE IELADGVREW SPAADGVRRA GVSAFGVSGT NAHVIIAEPP E
```

Figure 4.1: Input sequence used for DEBS module 1.

4.2.2 AT Domain Classification

Sequences that did not give a hit for the AT domain with hmmscan were labelled as *trans*-AT systems; whereas, the ones that gave a hit for the AT domain were labelled as *cis*-AT sequences. The *cis*-AT sequences in the alignment were classified as malonyl-CoA, methylmalonyl-CoA and ethylmalonyl-CoA specific based on their fingerprint motifs HAFH, YASH and (T/F/V/H)AGH, respectively Haydock *et al.* 1995. The ones do not bear any of these motifs were label as 'Unclassified'.

4.2.3 KR Domain Classification

Sequences were classified based on the subtype motifs of the KR domains. The ones with LDD motifs were labelled as type B KRs, the ones with a tryptophan eight residues before the catalytic tyrosine were labelled as type A. Further classification was made based on the three residues before the catalytic tyrosine. Since leucine, histidine, and glutamine residues are conserved in B2 type, A2 type and A1-B1 types of KRs, respectively Zheng and Keatinge-Clay

2013, a further classification was made based on these specifications. KRs that did not bear any of those sequence motifs were labelled as ’Unclassified’ and the ones that have both patterns of A and B types are labelled as ’KRbothmotifs’. These groups are not shown in the sequence-position mapping plots. Among all the sequences only 81 do not have KR domains. The most abundant type in the alignment is type-B1, with 763 sequences. There are also 6 type-B2, 239 type-A1, 73 type-A2 and 67 type-C KR present. Unfortunately, 996 sequences could not be classified and 20 of them bear motifs of both type A and type B KRs.

4.2.4 Determination of coevolved residue groups

Weighted correlation matrix and independent component analysis were applied as described in Section 1.4 via scripts from py-SCA tool Rivoire *et al.* 2016.

In the process of eigenvalue decomposition, when the first eigenvalue is much larger than the remaining significant eigenvalues, it indicates that the first eigenvector is the ”coherent” mode and it describes the contribution to all positions to the total correlation Halabi *et al.* 2009. In our system, the first eigenvalue is 822, whereas the second highest is only 143, indicating the dominant first mode. As the first mode has the contributions from all positions, it was removed for the SCA matrix visualization and hierarchical clustering for a clearer detection of the covariation of the residue groups.

4.2.5 Hierarchical clustering of ICs

We performed hierarchical clustering based on average coupling scores of inter-ICs. In the coupling matrix $\tilde{C}_{ij}$, the average coupling score of each square (scores of intra-ICs and inter-ICs) is calculated ending up with 34 x 34 size matrix (for the analysis after removal of some ICs, the size of the average-scored coupling matrix is 22 x 22). Hierarchical clustering was applied on the average-coupling score matrix by scipy.cluster hierarchical clustering package where complete linkage calculations were performed on the average-coupling matrix as distance matrix and the clusters were generated from the calculated linkage matrix.

4.2.6 Sequence-position mapping

Covariation of the MSA matrix columns reveals the information of amino acids changing patterns of positions resulting in the detection of the coevolved residue groups (i.e ICs). Similarly, covariation of the MSA matrix rows is expected to give information about amino acid changing patterns of the sequences.

The sequences, which have a high correlation in the row analysis of the MSA matrix, are the sequences with high similarity in the pattern of the amino acid order i.e. amino acid sequence. Grouping the sequences based on their correlation in the amino acid order means grouping them based on their sequence similarity, which is similar to phylogenetic analysis. Therefore, in other words, we can say that the correlation analysis of sequences gives information about the closeness of the sequences. Although this approach is not the best way to learn more about the similarity order of a given set of whole sequences, it is useful when we want to figure out the sequence patterns of the coevolved residue groups.

Since both sequence correlations and position correlations are obtained from the same matrix (MSA), they are related to each other by a mathematical approach known as singular value decomposition (SVD).

Basically, SVD is used to decompose any matrix into three matrices:

$$X = U\lambda V^T. \tag{4.1}$$

Here, U carries the information about sequence similarities (rows-based analysis) and V carries information about position similarities (columns-based analysis). λ matrix is a diagonal matrix that carries information about which parts of the U and V matrices provide “more important” contribution to the X matrix.

With a slight modification, the matrix U, which carries the sequence similarity pattern, can be obtained from the alignment matrix, X, and position correlation matrix, V:

$$U = XV\lambda^{-1} \tag{4.2}$$

This means we can obtain sequence similarity patterns specific to coevolved residue groups i.e. independent components.

However, since we applied modifications on the positional correlations of the X matrix to obtain the coupling matrix ($\tilde{C}_{ij}$), we need to follow the same transformations.

$$\tilde{U} = \tilde{x}\tilde{V}\tilde{\Delta}^{-1/2} \tag{4.3}$$

where $\tilde{x}$ is a compressed alignment matrix (from MxLx20 to MxL), $\tilde{V}$ and $\tilde{\Delta}$ are eigenvectors and eigenvalues of $\tilde{C}_{ij}$, respectively. And as the final step, the ICA transformation is applied:

$$\tilde{U}^p = W\tilde{U} \tag{4.4}$$

where W is the same transformation matrix that was used to transform the top eigenvectors to independent components.

The final $\tilde{U}^p$ matrix carries the information of sequence divergence patterns of the coevolved residue groups (independent components). Therefore, this analysis allows us to map positional correlations on sequences and thus this protocol is referred to as sequence-position mapping in the following part of the chapter. For a more detailed explanation of the method please see Rivoire *et al.* 2016.

In order to identify the importance of the residues for specific sub-types, we calculated the projection of the separate $\tilde{C}_{ij}$ matrices onto the ICs for sub-type specific sequences only. For the AT domain sub-types, we first generated two separate MSAs including malonyl- or methylmalonyl- specific sequences and calculated the coupling matrices $\tilde{C}^m_{ij}$ and $\tilde{C}^{mm}_{ij}$ for malonyl- and methylmalonyl- specific MSAs, respectively. The projections of $\tilde{C}^m_{ij}$ and $\tilde{C}^m_{ij}m$ matrices on the $\tilde{V}^p_{1\cdots k^*}$ matrix were calculated via

$$\tilde{V}^{p_m}_{1\cdots k^*} = \tilde{C}^m_{ij}\tilde{V}^p_{1\cdots k^*} \tag{4.5}$$

$$\tilde{V}^{p_{mm}}_{1\cdots k^*} = \tilde{C}^{mm}_{ij}\tilde{V}^p_{1\cdots k^*}. \tag{4.6}$$

To determine the difference between the malonyl- and methylmalonic sub-types, we substracted the scores of projected ICs.

$$\Delta \tilde{V}^{p}_{1\cdots k^*} = \tilde{V}^{p_m}_{1\cdots k^*} - \tilde{V}^{p_{mm}}_{1\cdots k^*} \tag{4.7}$$

A similar approach was also applied for the analysis of KR sub-types.

All scripts were written in Python Van Rossum and Drake Jr 1995.

4.3 Results and Discussion

4.3.1 Determining the number of sequences needed for a sectors analysis.

The literature suggests that 100 sequences are sufficient for a sectors analysis Rivoire *et al.* 2016. In contrast, other covariance methods such as direct coupling analysis require thousands of sequences to give reliable results. The DEBS module that was selected to be analysed has 1901 amino acids, which is considerably larger than any sequence previously analysed by sectors analysis. Hence, we anticipated a larger number of sequences in the alignment might be needed to obtain robust results. We, therefore, analysed several uni-domain proteins to see whether we can detect a trend between the length of a protein and the number of sequences adequate for the analysis of the sector.

To test how the number of sequences in an MSA affected the ICs, varying numbers of sequences, from 100 to the full complement of the sequence alignment, in increments of 100 sequences, were randomly were randomly selected from a source MSA, without replacement, and ICs were determined. This was repeated three times for each different number of sequences (n=3). To see the similarity between the ICs from subsampled MSAs and the ICs from the source MSA a similarity score (ss) was calculated between the sets of ICs.

$$ss = \frac{1}{M}\sum_{m=1}^{M}\sum_{n=1}^{N}\delta(IC_m, IC_n) \tag{4.8}$$

where $\delta(IC_m, IC_n)$ is equal to one if the ratio of the number of mutual residues between two ICs to the number of residues in the IC with the fewer sequences (min(len(IC_m),len(IC_n))) is greater than a threshold (set at 0.7) and either IC_m or IC_n is not matched with another IC; zero otherwise.

For most of the proteins analysed, 100 sequences in the MSA is not sufficient to get ICs similar to the ICs obtained using the MSA with the full complement of sequences available, irrespective of the length of the protein (Fig. 4.2). On the other hand, the similarity score of the all tested proteins converged to one. However, no trend was detected between the length of the proteins and the ratios of the minimum number of sequences adequate for achieving the same IC sets as the source MSA ICs (saturation point, M_s, shown in Fig. 4.2) to the number of sequences (M), defined as saturation ratio (Fig. 4.3).

Even though subsampled MSAs with fewer sequences failed to provide the same IC sets for most of the analysed proteins, some ICs were consistently detected even in the MSAs with few sequences (Fig. 4.4). This suggests that although the number of sequences in the MSA is critical for the sector analysis, some ICs can be successfully detected even when the MSA is not comprehensive enough.

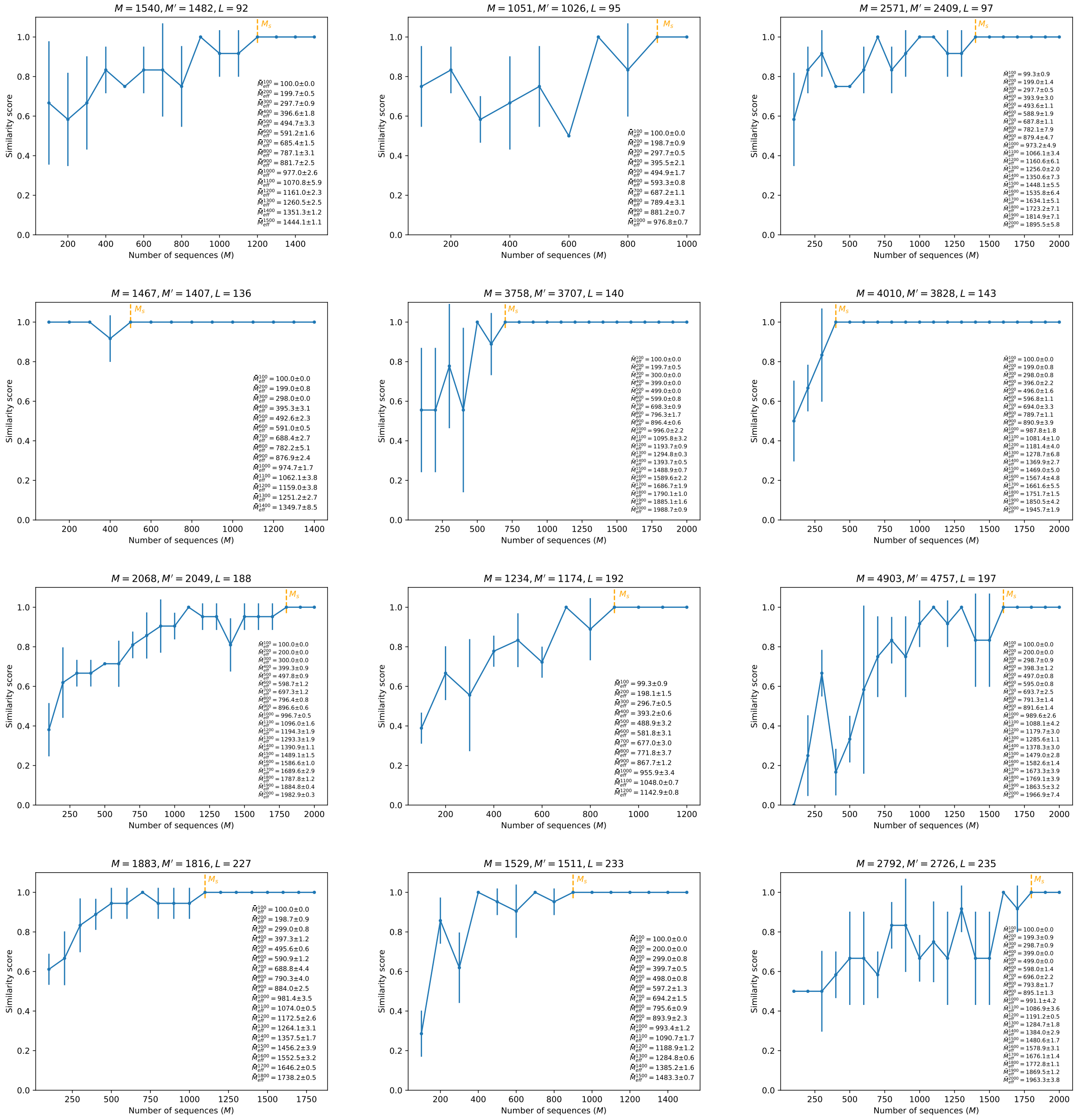


Figure 4.2: As the number of sequences in the subsampled MSAs increase, similarity score converges to 1. *100 sequences in the subsampled MSAs are generally insufficient to achieve the same set of ICs are the source MSA. Increasing the number of sequences in the subsampled MSAs eventually gives a saturation point (Ms) that varies among the proteins.*

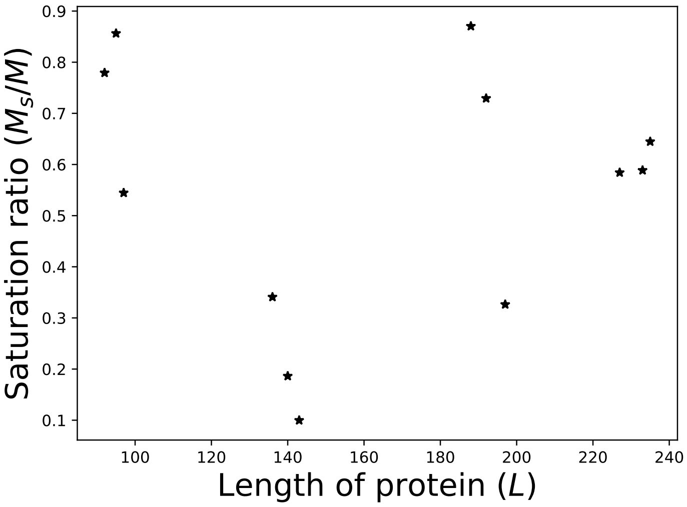


Figure 4.3: A trend could not be detected between the length of the proteins** (L) **and the saturation ratios** (M_s/M)**. *M_s values are defined as the minimum number of sequences adequate to achieve the same set of ICs as the source MSA (with M number of sequences). Selected M_s values for each protein are shown in Fig. 4.2.*

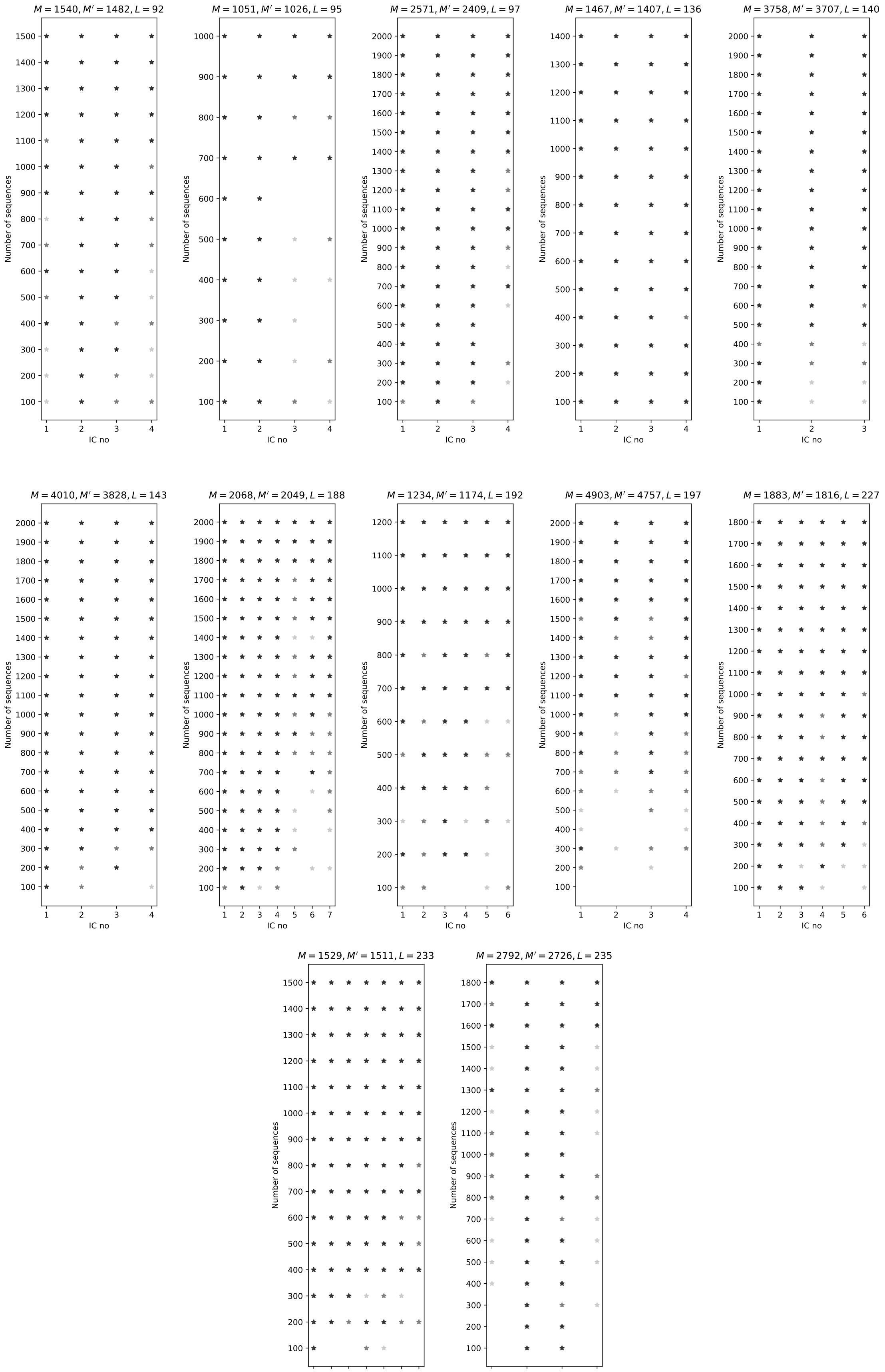


Figure 4.4: Some ICs can be detected successfully even though few sequences are used in the alignment. *A light grey star indicates one occurrence in three replicates, a grey star indicates two occurrences in three replicates and a black star indicates three occurrences in three replicates.*

4.3.2 Construction of a multiple sequence alignment for DEBS module one.

An initial multiple sequence alignment (MSA) was generated by using the sequence of the first module plus the KS domain of the second module (KS1_AT_KR_ACP_KS2) of the DEBS system (Fig. 4.1). After processing the MSA to remove highly gapped positions and sequences, and highly identical ones, as described in Methods, 2303 sequences remained in the alignment. Only 682 of the sequences have the exact domain composition as the input sequence. The remaining sequences have either a missing domain (like AT and KR) or an additional domain (like DH and ER). Having 600 and fewer sequences in the alignment were sufficient for consensus sector analysis for some of the proteins we analysed in the previous section, but here the input sequence is longer than these proteins. Thus, we investigated the effect of the number of sequences in the alignment on the ICs.

Similarity scores calculated between different alignments with a varying number of sequences showed that as the number of sequences increases in the sub-sampled alignment, similarity to the original alignment (with 682 sequences) increases, too; but it does not converge to a point where the addition of new sequences has no further effect (Fig. 4.5). This indicates that to achieve the best analysis, we should include as many sequences as possible. Therefore, we included sequences with different domain compositions. Keeping the two KS domains and the ACP domain present in the sequences, we included sequences without AT and/or KR domains or domain compositions with DH and/or ER domains. This allowed us to work on alignment with 2303 sequences.

Although the inclusion of additional sequences was not adequate to obtain the convergence of the results, the maximum similarity score is higher than the one where only sequences with KS1_AT_KR_ACP_KS2 domain composition were included in the MSA (Fig. 4.6).

34 ICs were detected in the studied system (Fig. 4.7). Eighteen sequence alignments were sampled from the full alignment with 100, 500, 900, 1300, 1700 and 2100 sequences drawn randomly three times over. Analysing the presence of the ICs in different subsets revealed

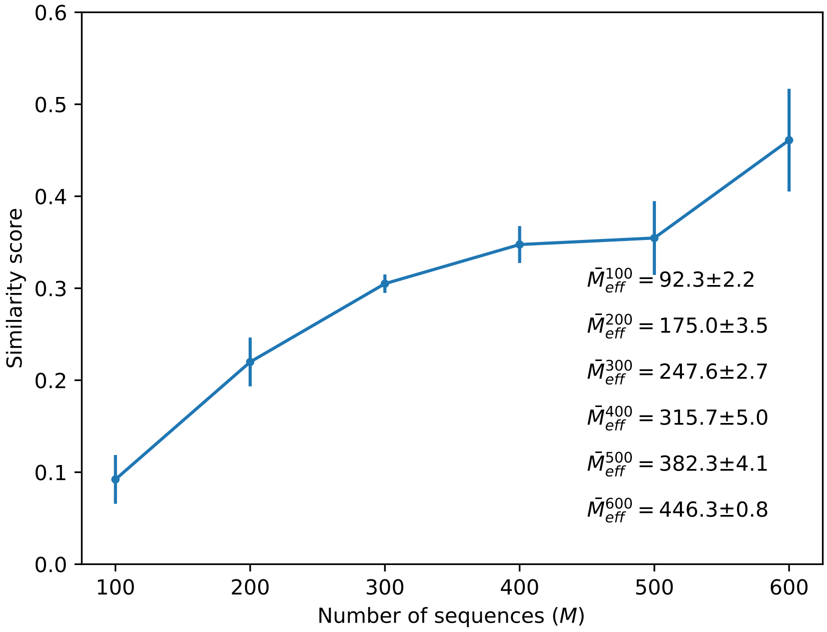


Figure 4.5: The similarity of the detected ICs to the source MSA analysis. *The MSA contains only sequences of domain composition KS1_AT_KR_ACP_KS2. The lack of convergence to a plateau indicates that more sequences are needed for the analysis.*

that some ICs were able to be detected even there are few sequences in the alignments (Fig. 4.8). Removal of the ICs that occurred less than or equal to three times in whole bootstrapping analysis (ICs 9, 19, 20, 23, 24, 28, 29, 31, 32, 33, 34) and one IC with only two residues (IC_25) finally gave 22 ICs for further analysis (Fig. 4.10).

4.3.3 Functional domain boundaries can be detected.

Some independent components consist of residues that are predominantly from only one domain while most ICs have a signal from multiple domains and linkers (Fig. 4.10). Although an IC is a grouping of alignment positions that have coevolved together, predominantly independent of other ICs, there can be couplings between residues in different ICs (Fig. 4.9). Therefore, previous work has grouped together ICs with high inter-IC coupling, each grouping termed a sector Rivoire *et al.* 2016. However, there is no standard way to group ICs into sectors. Therefore, in this study, we do not definitively define sectors but instead make hierarchical clustering of ICs based on the average interaction score between residues in each pair of ICs

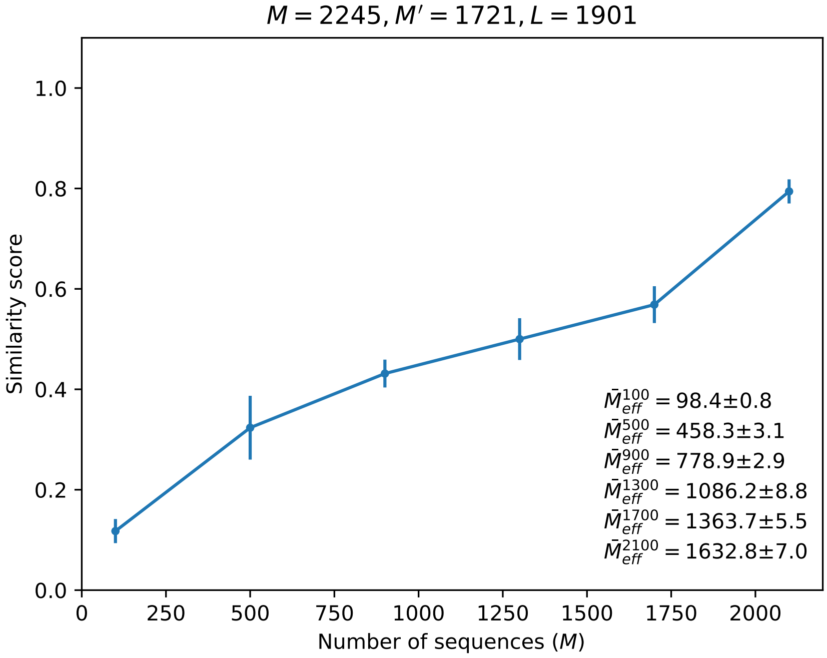


Figure 4.6: The similarity of the detected ICs to the source MSA analysis (with varied domain compositons in the MSA) increased as the number of sequences in the sub-sampled MSA increases. *Although convergence could again not be detected, the similarity score is higher compared to the KS1_AT_KR_ACP_KS2-only MSA analysis.*

as explained further in Methods section 4.2.5 (Fig. 4.11). Such hierarchical clustering shows that AT and KR domains are characterised by several coupled ICs (Fig. 4.10) and they are predominantly from the same one domain, thus allowing us to define domain boundaries.

The whole KR domain is defined by ICs 10, 17 and 21, which form a cluster of ICs that are more cross-coupled to each other than to other ICs (Fig. 4.10). KR domain contains two parts: a catalytic region (KRc) and a structural, non-catalytic, region (KRs) Keatinge-Clay and Stroud 2006b. KRc has strong inter-residue coevolutionary couplings that result in IC3 with only weak couplings outside the KRc; on the other hand, ICs 10, 17, 21 have strong coupling across the whole KR domain, resulting in defining the domain boundaries. Although there are other ICs that have residues in the KR domain (ICs 1, 16, 18, 26, 27, 30), their coupling strength is not as strong as the ICs 3, 10, 17 and 21 (Fig. 4.10 upper panel).

As residues of ICs 10, 17 and 21 have high coupling only within the KR domain, we were concerned that these results may be an artefact of some sequences in the MSA not having a

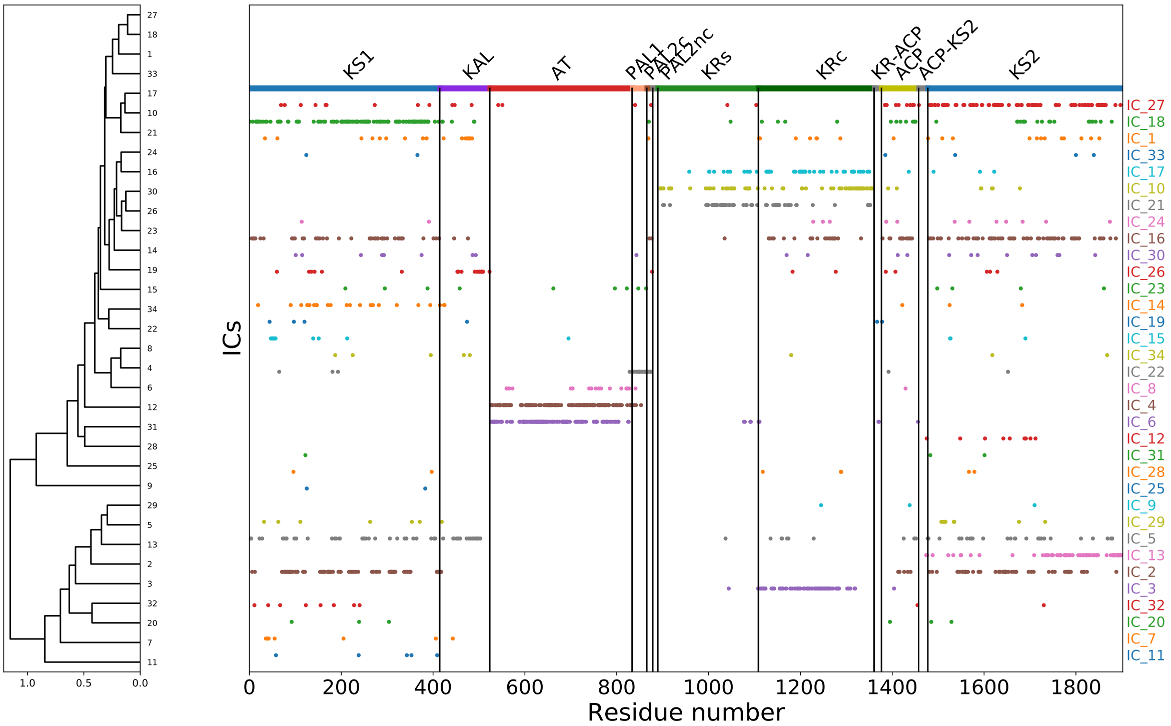


Figure 4.7: SCA analysis revealed 34 coevolved residue groups spanning either one domain or multiple domains and linkers, and clustering analysis based on the average coupling score between the ICs grouped highly coupled ICs. *The distributions of residue positions contributing to each co-evolving group (i.e. IC) are shown. KS: Ketosynthase, AT: Acyltransferase, KAL: KS-AT linker, PAL1: post-AT linker 1, PAL2c: post-AT linker 2 conserved region, PAL2nc: post-AT linker 2 non-conserved region, KRs: Ketoreductase structural integrity region, KRc: Ketoreductase catalytic region , ACP: acyl carrier protein, KS2: ketosynthase of module 2.*

KR domain. To investigate that, sequences without the KR domains were removed from the alignment and the same analysis was applied. Highly similar domain boundaries were detected when only sequences with a domain composition that includes a KR were analysed (Fig. 4.12 IC10). In that analysis KRc also still comes out as defined by one IC (IC_5, Fig. 4.12) but interestingly the KRs domain is detected as a separate entity in one IC (IC_20 Fig. 4.12), but there are only 33 residues from a KRs subdomain of 219 residues, and their coupling scores are low (Fig. 4.12 lower panel). It is not possible to tell if these differences are due to a change in the number of sequences analysed or are an artefact arising from the presence/absence of the KR domain in some sequences, but the results of the two analyses are similar. The analysis of the MSA with all non-KR containing sequences removed indicates that the boundaries for the full KR are expected to be R886 to A1356 and for the KRc domain to be P1107 to L1320 (ICs 10 and 5 of that analysis, respectively). For the full MSA analysis, then ICs 10, 17 and 21 together

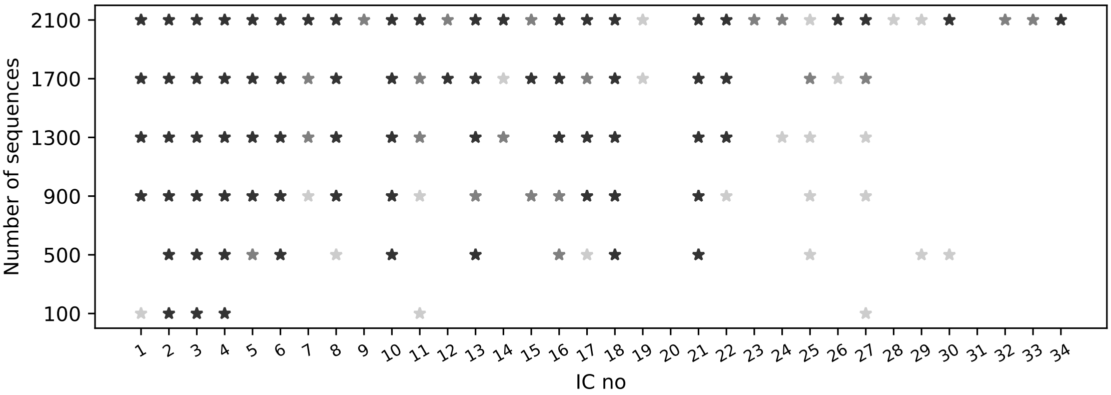


Figure 4.8: As the number sequences in the alignment increases, the number of the detected ICs increases. *22 out of 34 ICs were selected for further analysis as they are detected at least three times over all bootstrapping analyses. A light grey star indicates one occurrence in three replicates, a grey star indicates two occurrences in three replicates and a black star indicates three occurrences in three replicates.*

define a boundary for the whole KR of L896 to R1357 for DEBS1 and IC3 defines a boundary to the KRc of G1109 to L1320 compared to the boundaries in the literature Keatinge-Clay and Stroud 2006b of V890 to A1360 where KRc starts at position T1110.

IC_3 includes highly conserved residue positions including the catalytic triad of the KR domain, K1219, S1243 and Y1256 and the NADPH binding site TGGTGxLG (T1114, G1115, G1116, G1118, L1120, G1121) Zheng and Keatinge-Clay 2013. Keatinge-Clay and Stroud identified that the adenine ring of NADPH stacks with R1141 and forms hydrogen bonds with D1169 and V1170 Keatinge-Clay and Stroud 2006a, which are also detected in IC_3. Additionally, they showed that the phosphate group of adenine ribose forms a salt bridge with R1141 and hydrogen bonds with S1142, which are detected in IC_3 and IC_21, respectively. For determination of the stereochemistry of KR products, an alpha helix proceeded by a loop close to the active site, referred as the lid region, play role in cooperation with LDD motifs in B-type KRs and conserved tryptophan residue in A-type KRs Keatinge-Clay 2007. The highly-conserved second aspartic acid in the LDD motif (D1201) and W1282, T1284, W1285 and G1303 residues of the lid region are detected in IC_3; whereas the LD of LDD motif, conserved tryptophan and the rest of lid residues are detected in five different ICs. Position G1283 is detected in IC_16, A1286, G1287, A1291, F1299, R1300 and H1302 are in IC_10, S1288 is in IC_1, G1289 and

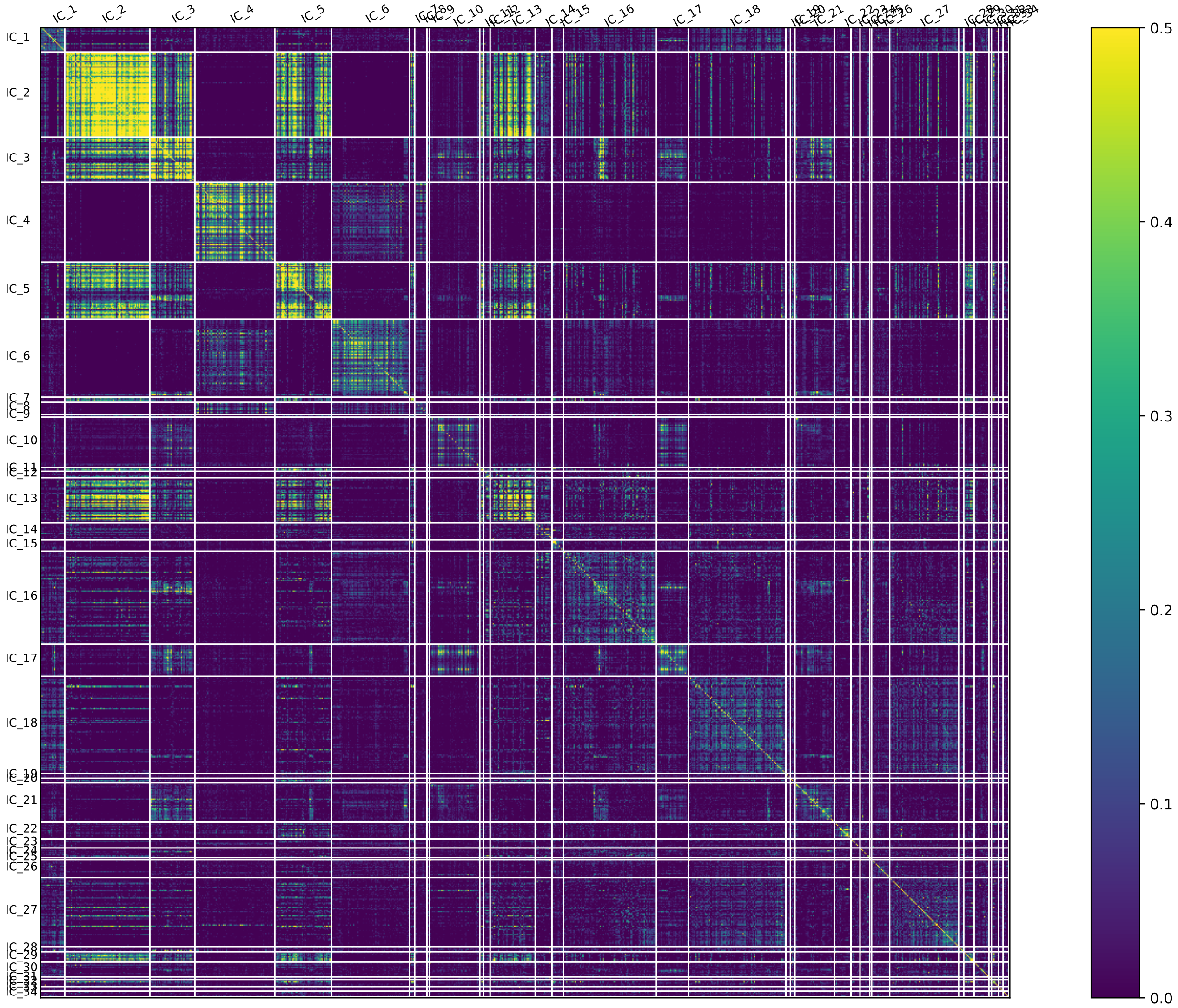


Figure 4.9: The coupling matrix sorted by the detected ICs. *ICs are clustered further based on the average inter-coupling score of the independent components*

M1290 are in IC_28, V1295 and R1298 are in IC_17.

The ICs coupled to IC_3 (ICs 2, 5, 7, 11 and 13) bear highly conserved residues, as well (Fig. 4.13). IC_2 contains the GXDS motif that is highly conserved in ACPs. Catalytic triad residues of the KS1 domain (C173, H308 and H346) are also detected in IC_2. For the KS2 domain, while the catalytic residues C1644 and H1819 detected in IC_2, H1779 is detected in IC_13, which contains highly conserved, only-KS2 domain residues. One intrigue here is how such highly conserved residues can have a coevolutionary signal. Although these positions are highly conserved, there are sequences where these residues are different and this can be seen in the residues associated with IC_2 and IC_3 when the MSA consist of the positions of only IC_2 or IC_3 filtered by HHfilter[ref] to have only sequences with <80% sequence ID with respect to each other, is displayed as a logos plot (Fig. 4.14).

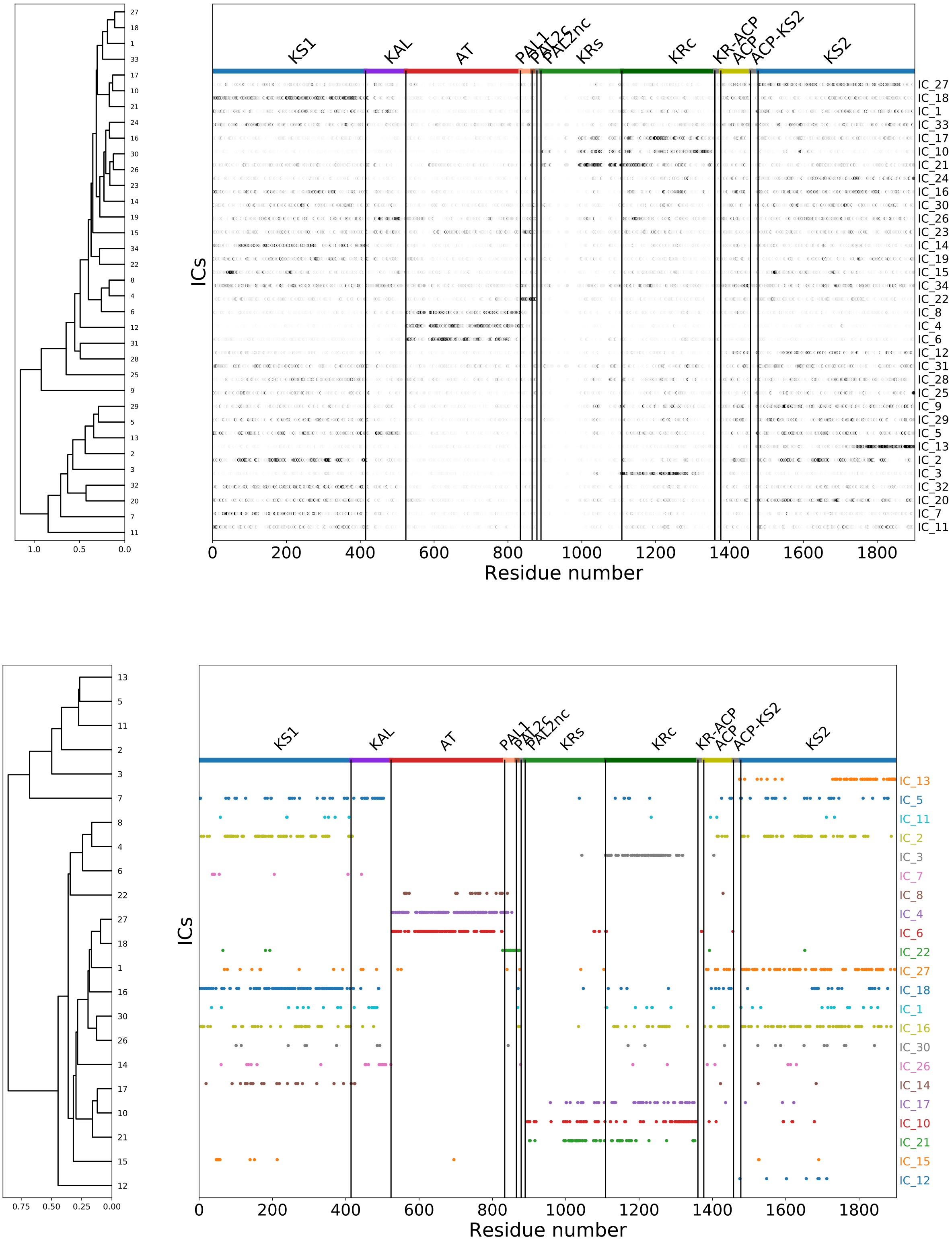


Figure 4.10: SCA analysis revealed 34 coevolved residue groups of which 22 were consistent in subsamples of sequences. *The residues in the upper panel are shaded according to their contribution to the IC, dark indicating a strong contribution. Hierarchical clustering of the ICs, which were consistent in subsamples of sequences and the distribution of selected residues of these ICs (based on p=0.95 threshold), are shown in the lower panel. KS1: Ketosynthase of module 1, AT: Acyltransferase, KAL: KS-AT linker, PAL1: post-AT linker 1, PAL2c: post-AT linker 2 conserved region, PAL2nc: post-AT linker 2 non-conserved region, KRs: Ketoreductase structural integrity region, KRc: Ketoreductase catalytic region, ACP: acyl carrier protein, KS2: Ketosynthase of module 2.*

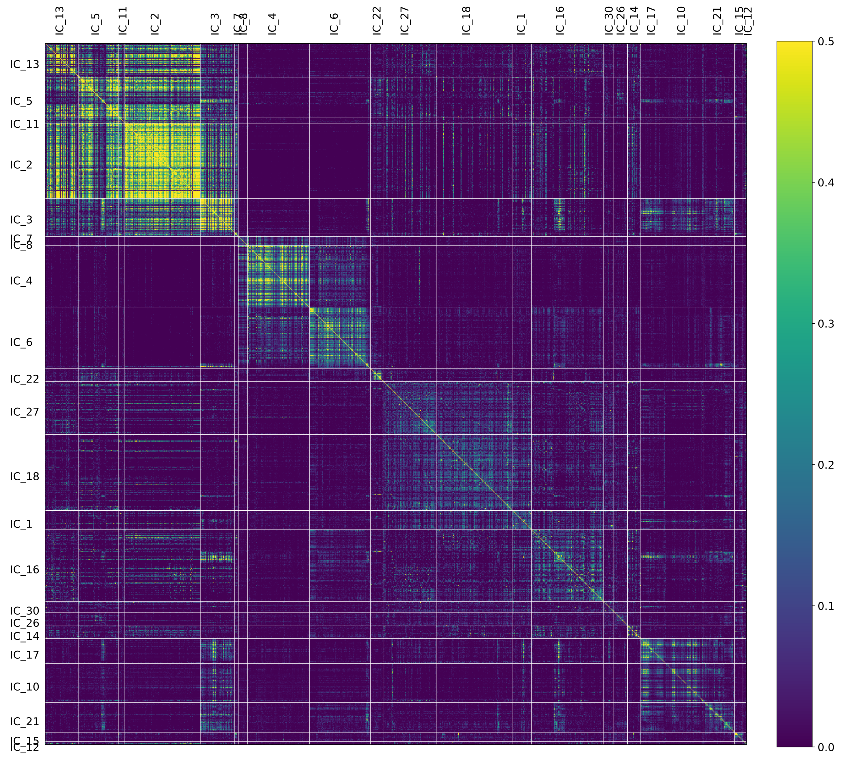


Figure 4.11: Grouping of highly coupled ICs. *Groupings were determined by cluster analysis as explained in Methods.*

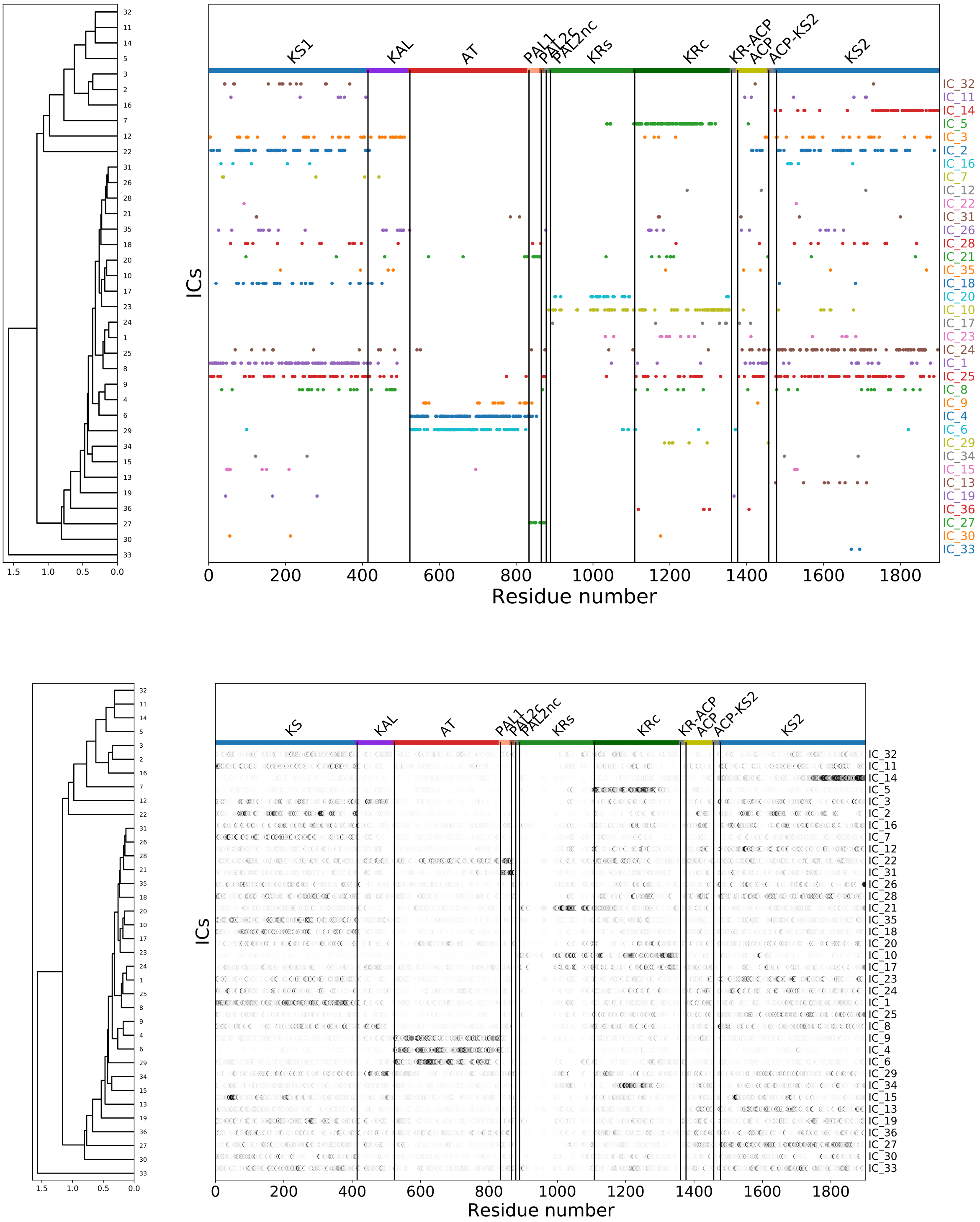


Figure 4.12: Using an MSA with sequences composed of a strict KS1_AT_KR_ACP_KS2 domain composition provided domain boundaries consistent with the analysis of the MSA with all homologous sequences. *Hierarchical clustering of the ICs are shown in the upper panel. The residues in the lower panel are shaded according to their contribution to the IC, dark indicating a strong contribution. KS1:Ketosynthase of module 1, AT: Acyltransferase, KAL: KS-AT linker, PAL1: post-AT linker 1, PAL2c: post-AT linker 2 conserved region, PAL2nc: post-AT linker 2 non-conserved region, KRs: Ketoreductase structural region, KRc: Ketoreductase catalytic region, ACP: Acyl carrier protein, KS2: Ketosynthase of module 2.*

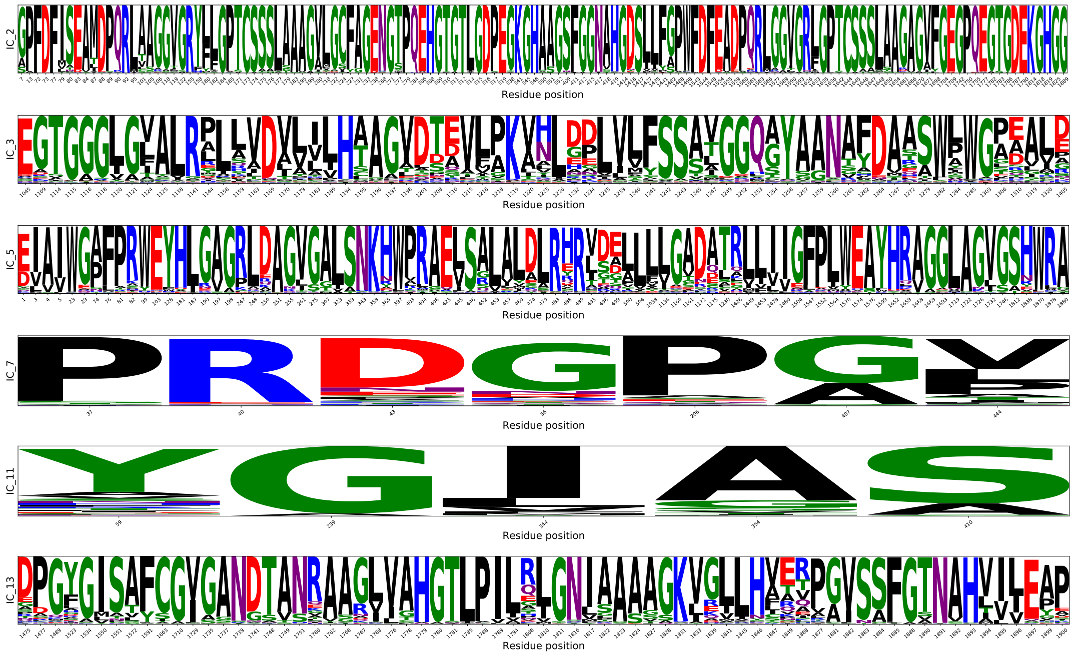


Figure 4.13: Highly conserved positions were detected in one branch of the cluster. *Sequence logos are shown for IC_2, IC_3, IC_5, IC_7, IC_11, IC_13.*

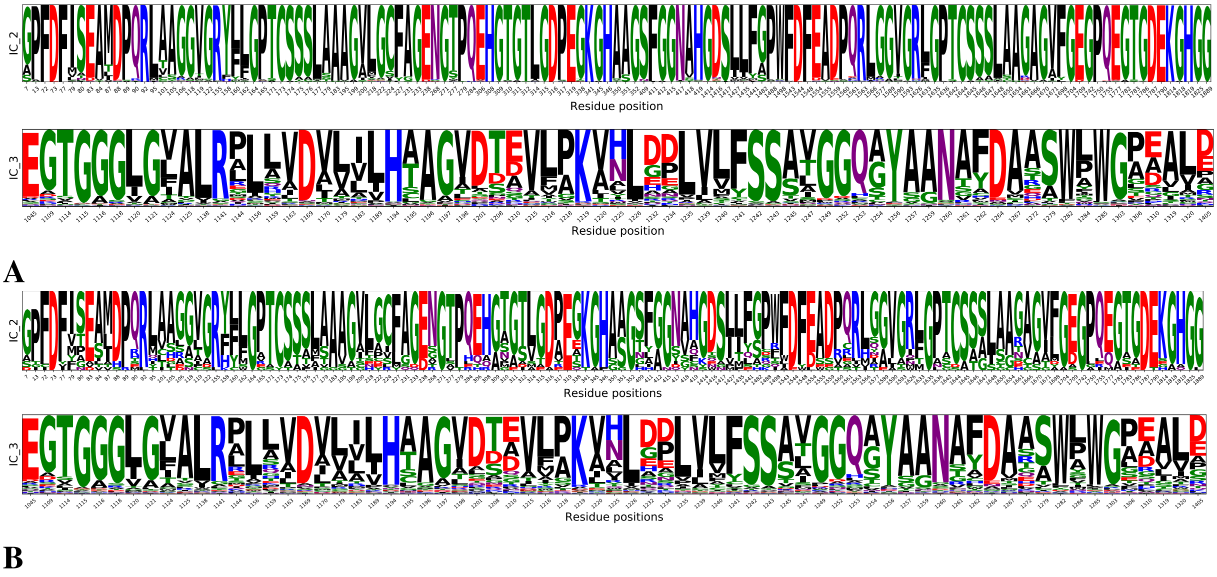


Figure 4.14: Filtering the IC_2 and IC_3 sequences based on the sequence identity provided a clear observation for the variation of the amino acids at highly conserved positions. *Filtering IC_2 and IC_3 sequences (A) based on maximum 80% pairwise sequence identitiy demonstrates that although the positions in ICs 2 and 3 are highly conserved, there are vairations in the alignment for the corresponding positions (B).*

The AT seems to be an independently evolving unit consisting of three ICs (4, 6 and 8). These ICs have residues that are coupled to each other, and therefore they are clustered together (Fig. 4.10). On the other hand, they do not have strong coupling with other ICs, except IC_20, which has residues from the linker region at the C-terminus of the AT domain (Fig. 4.10). ICs 6 and 8 are coupled to residues from the ACP, and IC_6 also couples with five residues from KRs,and one residue from KRc; however, these couplings to the residues outside of the AT_PAL region are very weak compared to those between residues within the AT domain. IC_4 contains no residues from outside of the AT_PAL1 boundary.

Removing *trans*-AT sequences from the MSA, to test if they influence the boundaries defined here, leads to almost identical boundaries of coevolving residue clusters, although inevitably not identical since the MSA differs (Fig. 4.15). In this *cis*-AT only analysis, the AT is now represented by two ICs, one consisting of AT and PAL1 (Fig. 4.15). Although there are a few residues from other domains, their coupling strength is low on the contrary of the residues of the AT domain (Fig. 4.15, lower panel). The second IC that has a signal from the AT domain is IC_5 and similar to IC_2, coupling scores of the positions outside of the AT domain is weak (Fig. 4.15, lower panel). IC_12 that has the residues from the post AT linker region (PAL1 and PAL2c) is clustered with ICs 2, 5 with a similar pattern as the full MSA analysis.

Taking ICs 8, 4, 6 of the full MSA together, they define an AT_PAL1 unit of residues V527 to S854, whereas IC_2 from the MSA without *trans*-AT sequences define the coevolving unit as residues V527 to R863. Recent experimental work has demonstrated the need for the PAL for the successful replacement of the AT domain of DEBS module 6 with that of the equivalent residues from EPOS module 4 Yuzawa *et al.* 2017, equivalent to residues Q524 to P865 in DEBS1. This is consistent with the results here; however, they obtained better kinetic parameters (a lower K_M value) when they included residues from the KAL region, which is not evident in our ICs, but which they demonstrated led to functional swaps in other systems (the AT_PAL unit was only tested in one construct).

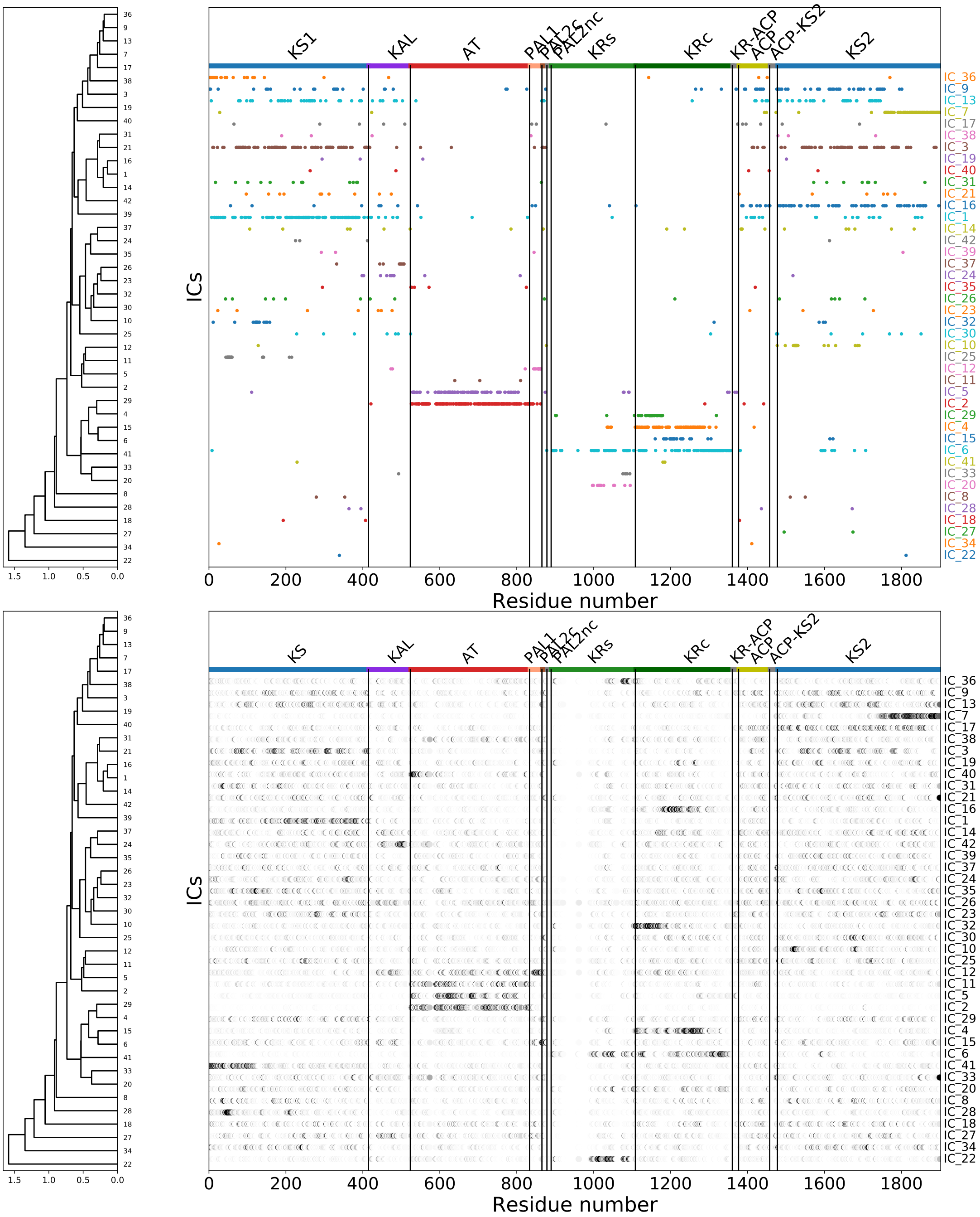


Figure 4.15: Using only cis-AT sequences in the MSA provides similar AT domain boundaries to the whole MSA analysis. *Hierarchical clustering of the ICs when only cis-AT sequences were kept in the alignment are shown in the upper panel. The residues in the lower panel are shaded according to their contribution to the IC, dark indicating a strong contribution. KS1:Ketosynthase of module 1, AT: Acyltransferase, KAL: KS-AT linker, PAL1: post-AT linker 1, PAL2c: post-AT linker 2 conserved region, PAL2nc: post-AT linker 2 non-conserved region, KRs: Ketoreductase structural region, KRc: Ketoreductase catalytic region, ACP: Acyl carrier protein, KS2: Ketosynthase of module 2.*

4.3.4 Amino acid patterns specific to different domain compositions are detected.

We further investigated whether groups of sequences in the MSA show different amino acid patterns at the residue positions detected in the ICs. To detect the sequence divergence patterns of the amino acids within the ICs, we applied sequence-position mapping via singular value decomposition of the weighted covariance matrix Rivoire *et al.* 2016. For this analysis, firstly, sequences were classified based on their domain compositions, and sub-types of the AT and KR domains.

Although the DEBS sequence has KS1_AT_KR_ACP_KS2 domain composition, not all sequences in the MSA have the same domain composition. Some sequences do not have a KR or AT domain; whereas, there are sequences whose original modular context include additional reducing domains (DH and ER). Although these domains are removed from the final alignment before the analysis, the rest of their sequence is kept and thus traces of these domains are likely to be seen in the co-evolution networks. Within the MSA, the most abundant domain compositions detected are KS1_AT_KR_DH_ACP_KS2 (with 725 sequences), KS1_AT_KR_ACP_KS2 (673), KS1_AT_KR_DH_ER_ACP_KS2 (225), KS1_KR_ACP_KS2 (161), KS1_KR_KS2 (132), KS1_KR_DH_ACP_KS2 (126).

Sequence-position mapping analysis based on domain composition classification reveals distinct patterns on four ICs (IC_4, IC_6, IC_8, and IC_17, Fig. 4.16). The *cis*-AT and *trans*-AT systems are distinguished by IC_4 and IC_6. Interestingly, IC_8 has two peaks: the first one has sequences from all domain compositions while the second peak consists of the sequences only from the *cis*-AT systems.

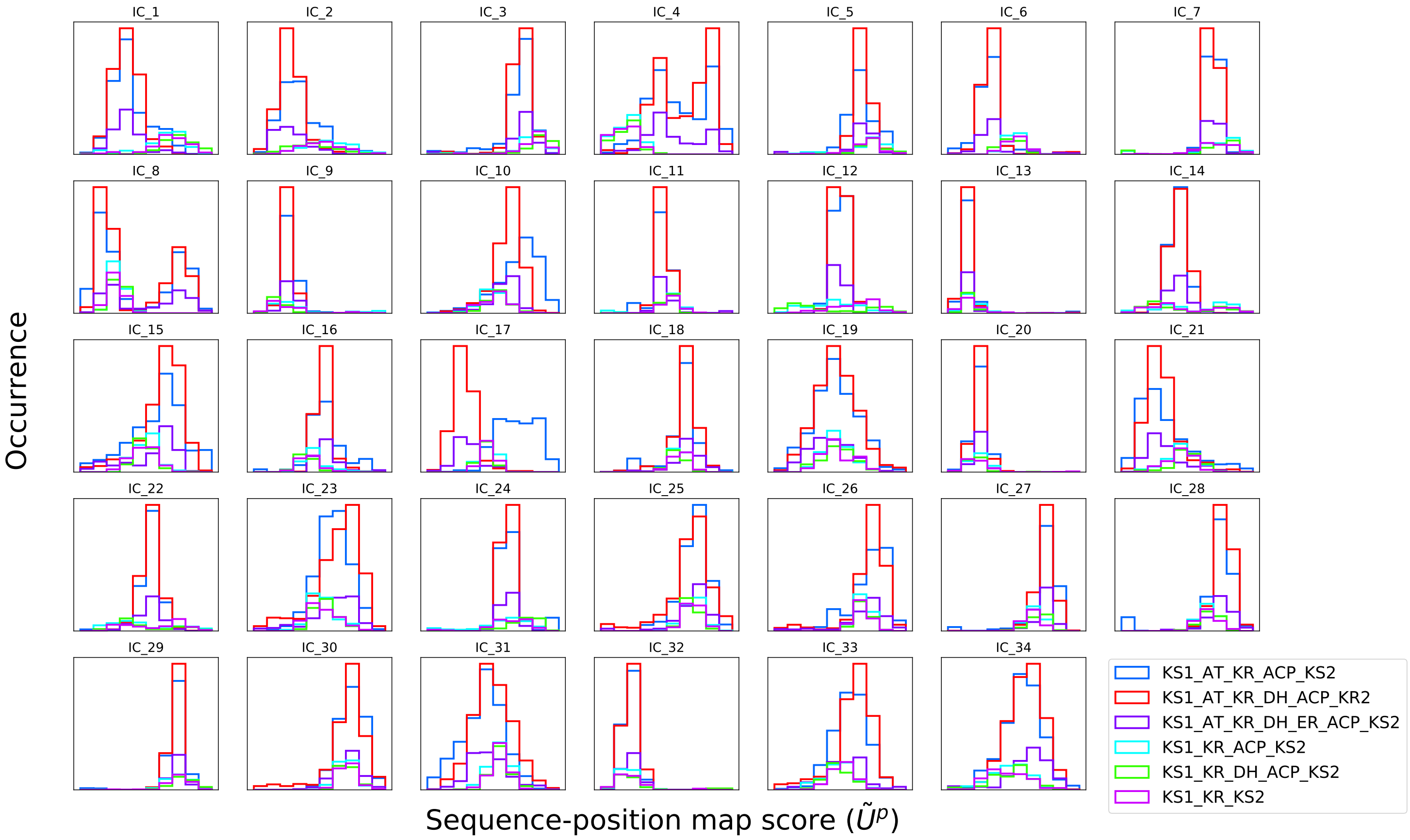


Figure 4.16: Sequence-position mapping based on domain compositions of the sequences in the MSA. *In these graphs, if any two classes of domain composition are highly overlapped (i.e. showing a similar pattern) in an IC, it means the amino acid patterns of these positions detected in that IC are similar. On the other hand, if any two class of domain compositions show diverged patterns, for these positions detected in that particular IC, amino acid patterns are not the same. Although for most of the ICs, there is no clear distinction in the sequence patterns of the different domain compositions, IC_4, IC_6, IC_8, and IC_17 show distinction between some domain compositions.*

On the other hand, the KS1_AT_KR_ACP_KS2 domain composition is distinguished in IC_17 from the KS1_AT_KR_**DH_ER**_ACP_KS2, with additional DH and ER domains, and KS1_AT_KR_**DH**_ACP_KS2, with an additional DH domain. Residue distributions of the positions in these four ICs along the sequence show that the residues of the ICs 4, 6 and 8 are predominantly from the AT domain, whereas the residues of IC_17 are mainly accumulated in the KR domain (Fig. 4.10). Since the signals originate from the AT and the KR domains, we further classified the sequences based on the AT extender unit specificity and the KR domain subtypes.

4.3.5 AT domains can be distinguished based on their extender unit specificity.

The AT domains in the sequence alignment were classified based on their specificity for their malonyl-CoA, methylmalonyl-CoA, or ethylmalonyl-CoA extender unit, or unclassified, as described in Methods. The sequence score distribution of the different AT types are distinguished in ICs 1, 4, 6, and 8 (Figs. 4.18, 4.20, 4.21, 4.22, 4.17).

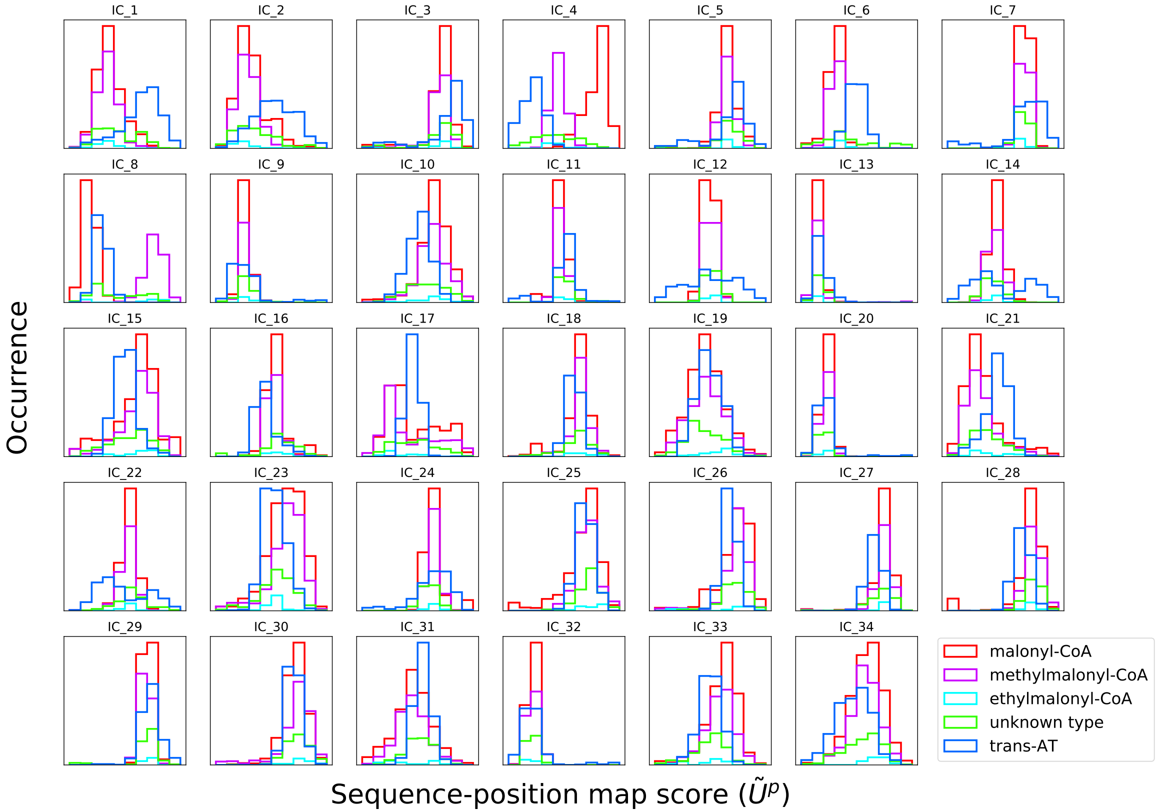


Figure 4.17: Sequence-position mapping based on extender unit specificity of the AT domains. *Although for most of the ICs, there is no clear distinction in the sequence patterns of the different domain compositions, IC_1, IC_4, IC_6 and IC_8 show distinction between some AT domains with different extender unit specificity.*

Methylmalonyl-CoA and malonyl-CoA specific AT domains show distinct patterns in IC_4 suggesting that these residues should be functional in the extender unit specificity of the AT domain (Fig. 4.18A). The separation that the position-sequence map suggests can also be seen by phylogenetic analysis of the IC_4 residue positions (Fig. 4.18B). From the phylogenetic tree, a clear distinction can be detected between malonyl-, methylmalonyl- specific ATs and *trans*-AT sequences. This pattern of the clustering is quite similar to the phylogenetic analysis applied on the whole AT domain, with boundaries defined by IC_4 (Fig. 4.18C). These results suggests that IC_4 might be important for the AT domain sub-type specification.

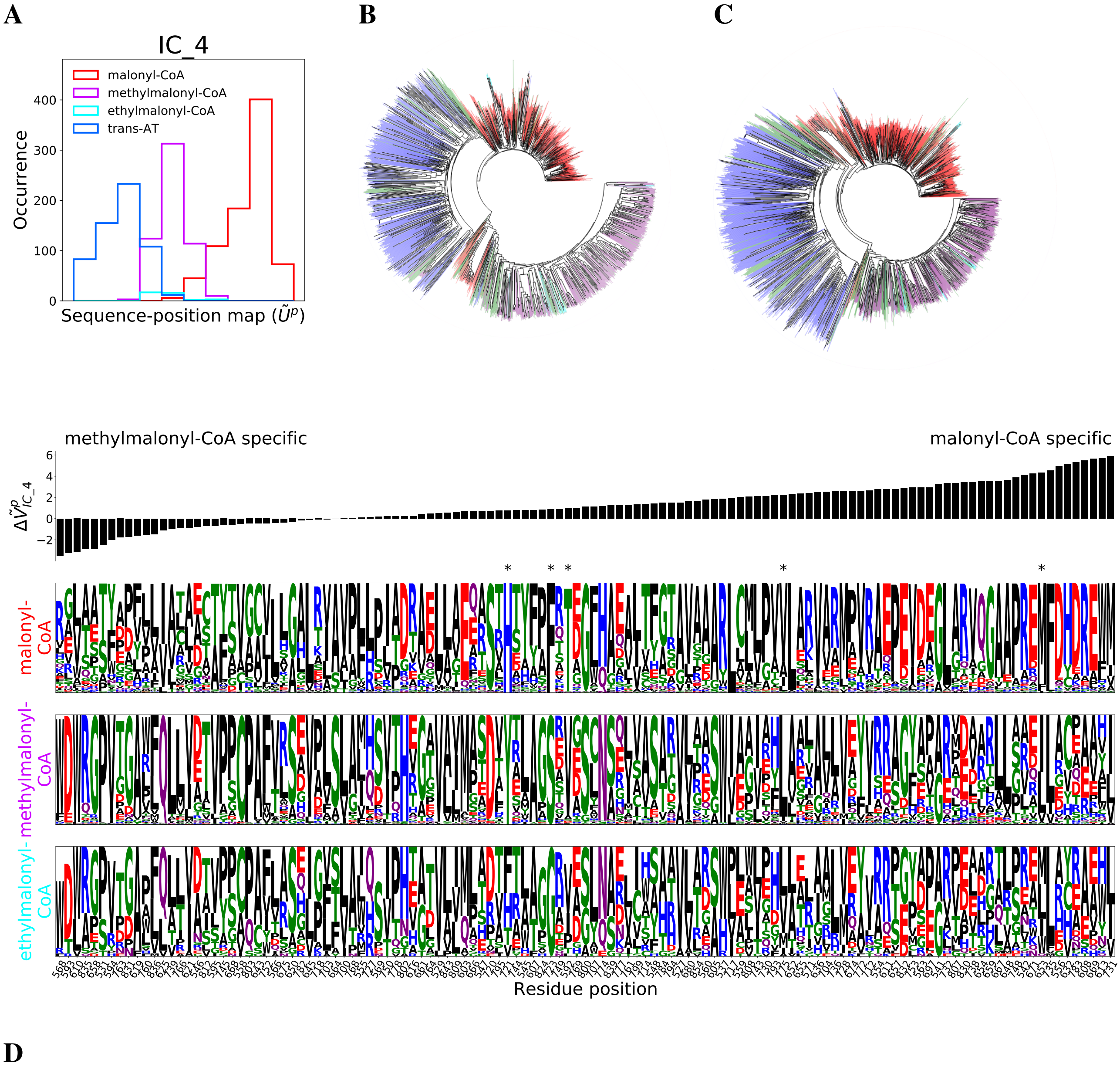


Figure 4.18: Different sequence patterns for methylmalonyl-CoA, and malonyl-CoA specific cis-AT systems were detected in IC_4. *Sequence-position map separates the AT domains based on their extender unit specifity (A). Phylogenetic analysis of the sequences of IC_4 residues distinguishes malonly-, methylmalonyl- specific and trans-AT systems supporting the sequence-position mapping patterns of the sequences (B). Pyhlogenetic analysis of whole AT domain reveals very similar clustering pattern as IC_4, which suggests residues of IC_4 might be important for AT domain subtype specification (C). Sequence logos show that sub-types of the AT domains have different amino acid patterns at IC_4 residue positions (D). Y and H residues of YASH motif (methylmalonyl-CoA specificity) and H and F residues of HAFH motif (malonly-CoA specificity) that provides the distinction between the fingerprint motifs are detected in this IC (positions at 721 and 723, respectively). Amino acids are coloured based on their chemical properties. Positions are sorted based on $\Delta\tilde{V}^p_{IC\_4}$ scores where $\Delta\tilde{V}^p_{IC\_4} = \tilde{V}^{p_m}_{IC\_4} - \tilde{V}^{p_{mm}}_{IC\_4}$.*

Analysis of sequence logos of malonyl-CoA and methylmalonyl-CoA specific AT domains shows the divergence of the amino acids detected in IC_4 (Fig. 4.18D). Sorting based on $\Delta\tilde{V}^{p}_{IC\_4}$ scores, as described in Methods, gives the positions that are more critical in methylmalonyl- (left end) and malonyl- (right end) specific ATs. When the sequence logos are analysed, it is seen that high-scored positions are mostly the ones that are conserved for that sub-type, whereas diverged for the other sub-types. There is an exception at position 760 that has a score in favour of methylmalonyl- specificity yet the conservation is not high. And a similar, but more conserved pattern is seen for position 763. In order to investigate these positions further, $\tilde{C}^{mm}_{ij}$ was analysed revealing that residue 760 has the highest coupling with residue 753. The spatial positions of 753, 760 and 763 on the 3D structure are close to each other (Fig. 4.19B). The most abundant pairs for positions 760 and 753 in methylmalonyl- specific sequences are Trp-Phe, Leu-Tyr and Arg-Tyr suggesting that the interaction between these two positions may be important and conserved (Fig. 4.19C); whereas, a similar pattern is not observed for malonyl-specific sequences (Fig. 4.19D).

Sequence logo of the trans-AT system includes amino acids although only gaps would be expected to be detected in the AT domain positions. It should be noted that a sequence is classified as a *trans*-AT sequence when no domain hit was detected by hmmscan. Therefore, the signals from the trans-AT sequences are either by the presence of remnants of the AT domain or parts of other domains that were misaligned.

It also is important to note that, the Y and H residues of the YASH motif and H and F residues of the HAFH motif, which provides the distinction between the fingerprint motifs, are detected in this IC (positions at 721 for Y/H and 723 for S/F, marked with stars). Site-directed mutagenesis studies on YASH and HAFH motifs aiming to switch the extender unit specificity generally result in promiscuous domains that can accept both methylmalonyl-CoA or malonyl-CoA as an extender unit. Further, the kinetic analysis showed that the cause of the promiscuity is not an increase in the tendency to bind non-native extender units, but because of a decrease in the capability of accepting the native one. In a recent study by Zhang et al., they performed experiments on salinomycin polyketide synthase targeting residues on malonyl, methylmalonyl and ethylmalonyl specific ATs beyond the residues of the YASH/HAFH motif residues Zhang *et al.* 2019. After performing structural analysis and molecular dynamic simulations, they determined that hydrophobic residues at positions 592, 653, 662 and 775 (DEBS1 module 1 numbering) are critical for substrate specificity. Mutations at these four positions in addition to YASH/HASH mutations switch specificity when mutating YASH/HASH residues did not. Consistent with these results, positions 592, 653 and 775 are all in IC_4. The exception, residue 662, is discussed below. This suggests that we can detect the residues that are critical for the AT domain extender unit specificity by sequence-position mapping analysis.

Sequence-position mapping based on the AT extender unit specificity also clarifies the ambiguity of the double peak in IC_8 of domain composition based mapping (Fig. 4.20A). AT extender unit type specificity based mapping reveals that the second peak is composed of only methylmalonyl specific ATs while the first peak bears the mixture of the rest. The distinction between the methylmalonyl- (and ethylmalonyl-) specific ATs and the others is also clear by the phylogenetic tree of the IC_8 residues' sequences (Fig. 4.20B). Detection of highly conserved positions in the methylmalonyl specific AT domain within IC_8 suggests those residues may have a role for methylmalonyl selection (Fig. 4.20C).

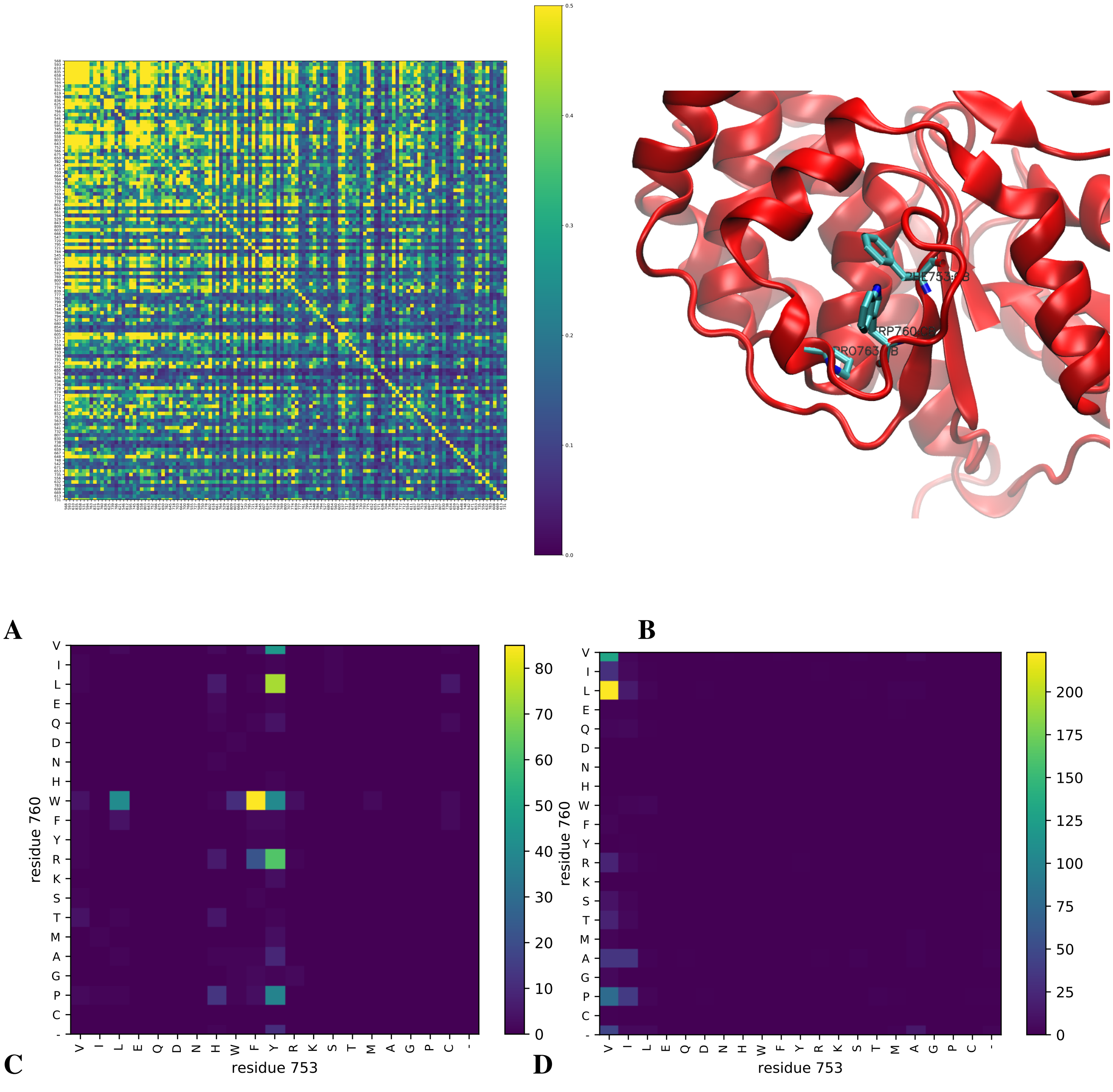


Figure 4.19: Residues 760 and 763 that are favoured for methlymalonyl specificity yet not highly conserved have high coupling with residue 753. *The $\tilde{C}_{ij}^{mm}$ matrix reveals that positions 760 and 763 have strong coupling with position 753 (A). On a 3D model of the structure residues 753, 760 and 763 on the 3D structure model are close to each other (the structural model was obtained by a former master student, Ruairi O'Brien, in the group via homology modeling) (B). Highly abundant pairs for positions 760 and 753 in methylmalonyl- specific sequences are Trp-Phe, Leu-Tyr and Arg-Tyr (B); whereas a similar pattern is not observed for malonyl- specific sequences (C).*

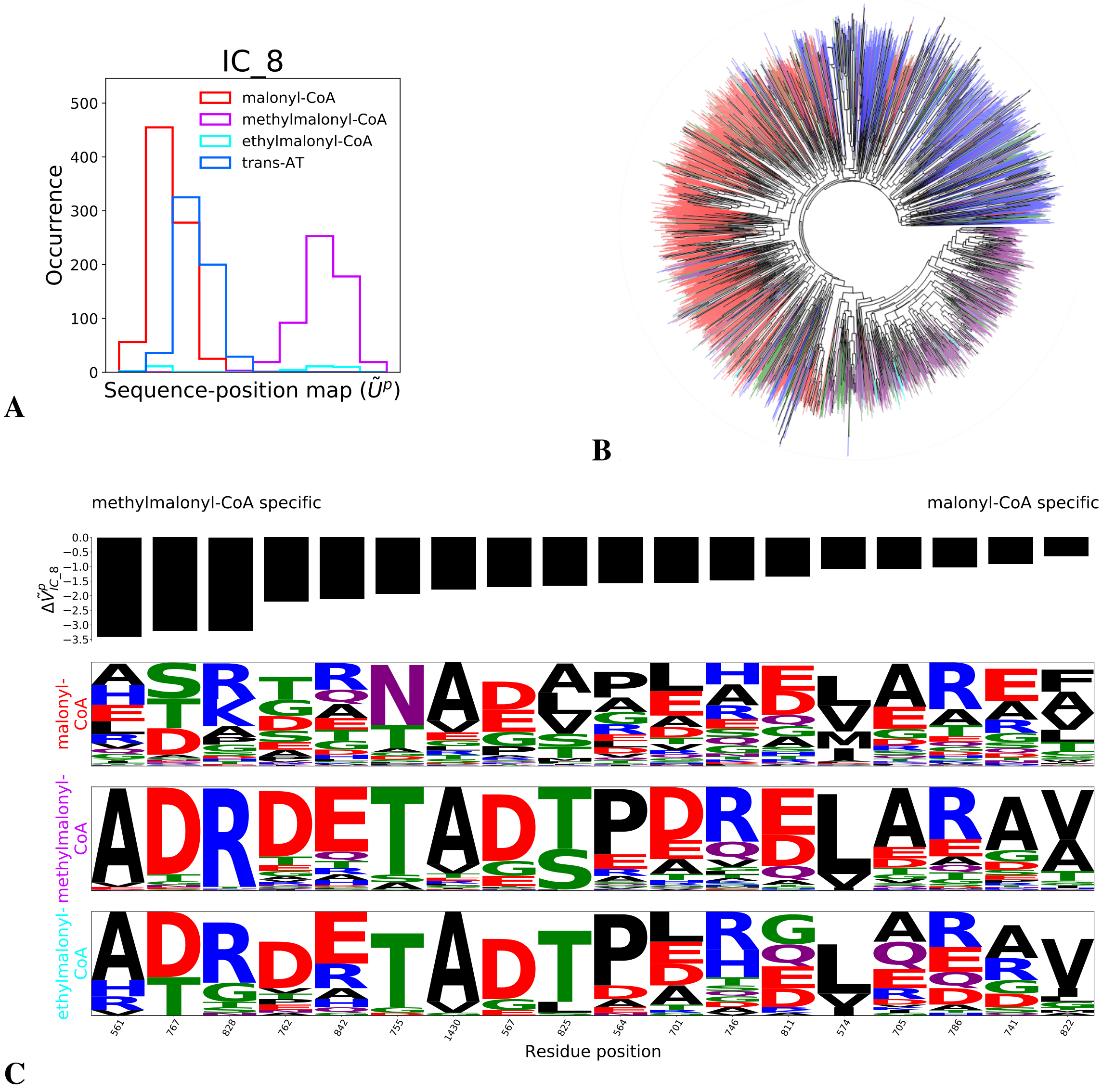


Figure 4.20: IC_8 reveals more conserved positions in methylmalonyl- and ethylmalonyl- specific ATs compared to malonyl- specific ATs. *Sequence-position map of IC_8 separates methylmalonyl- (and ethylmalonyl-) specific ATs from malonyl- specific ATs (A). Phylogenetic analysis of sequences of IC_8 residues clusters methylmalonyl- specific ATs in the same branch suggesting the closer sequence patters for the residues of IC_8 for this group (B). Sequence logos of IC_8 residues reveal that those positions are more conserved in methylmalonyl- and ethylmalonyl- specific AT domains, contrary to malonyl-CoA specific ATs (C). Positions are sorted based on* $\Delta\tilde{V}^{p}_{IC\_8}$ *scores where* $\Delta\tilde{V}^{p}_{IC\_8} = \tilde{V}^{p_m}_{IC\_8} - \tilde{V}^{p_{mm}}_{IC\_8}$.

Highly conserved residues in both malonyl-CoA and methylmalonyl-CoA are detected in IC_6 including catalytic residues (S624, H724) (Fig 4.21). These highly conserved residues, irrespective of extender unit specificity, that the residues detected in IC_6 are critical for the proper function of the AT domain.

Residue 662, which was detected as an important residue in the study of the salinomycin PKS Zhang *et al.* 2019, is detected in IC_6 as a highly conserved Met in all malonyl/methylmalonyl/ethylmalonyl specificity contrary to the system they studied as the residue at the corresponding position is Val in the ethylmalonyl specific AT.

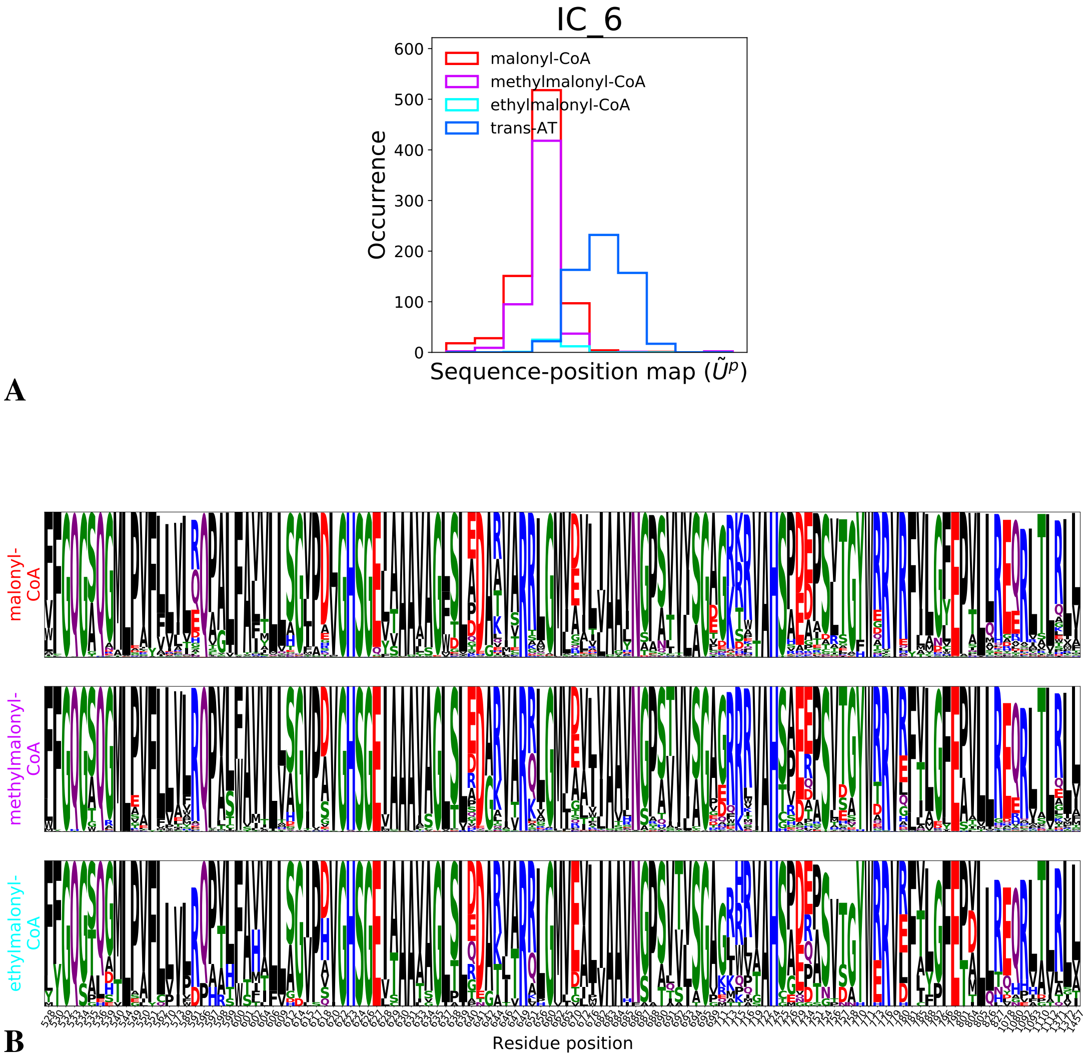


Figure 4.21: Highly conserved positions of the AT domain are detected in IC_6, including the catalytic residues. *Sequence-position mapping distinguishes accumulates cis-AT sequences in IC_6. Sequence logos reveal that highly conserved positions, including catalytic residues (S624, H724), are detected in in all methylmalonyl-CoA, malonyl-CoA and ethylmalonyl-CoA specific AT domains (B).*

4.3.6 Residues that distinguish *trans*-AT and *cis*-AT systems.

Trans-AT and *cis*-AT systems are separated in IC_1 (Fig. 4.22A), which was not detected as having different patterns in domain based sequence-position mapping. In contrast to the IC_4, 6 and 8, residues of IC_1 are dispersed along all the domains and the linkers except the AT domain. Phylogenetic analysis also shows that *trans*-AT systems have a distinct clade (Fig. 4.22B). Having different conservation patterns between *cis*- and *trans*-AT systems suggest that these positions (Fig. 4.22C) may have distinct roles in *cis*- and *trans*-AT systems.

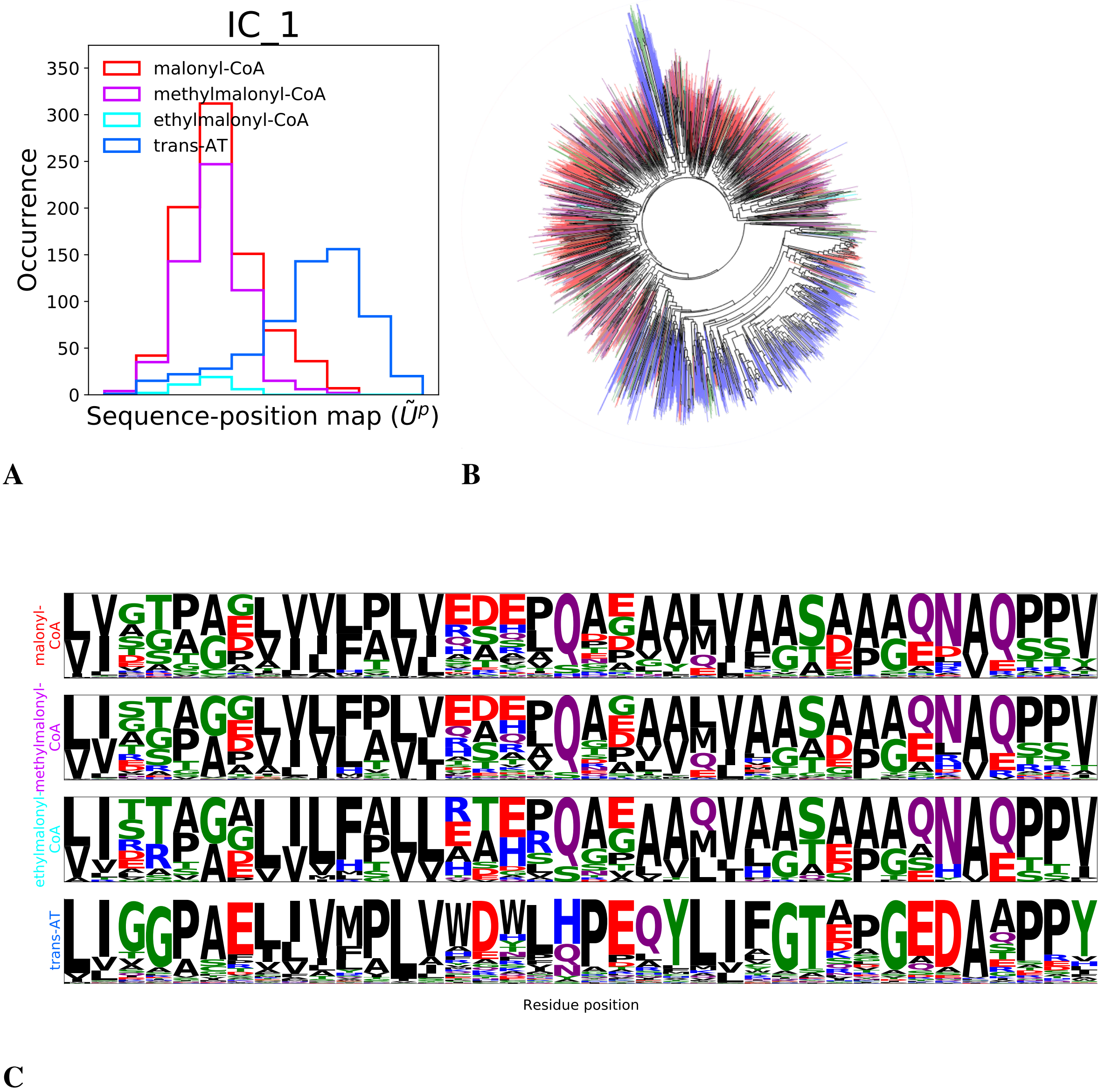


Figure 4.22: Different sequence patterns outside of the AT domain in cis- and trans- AT systems are detected in IC_1. *Sequence-position map of IC_1 separates cis- and trans- AT systems (A). Phylogenetic tree of sequences of IC_1 residues clusters trans- AT sequences suggesting these position have similar sequence patterns (B). Sequence logos demonstrates that sequences of trans-AT and cis-AT systems have distinct amino acid patterns (C).*

4.3.7 KR domain types can be distinguished by the independent components.

For most ICs, the different KR types show a very similar distribution (Fig. 4.23). In IC_3, reducing and non-reducing type KRs are separated, which is not surprising as IC_3 includes the catalytic core. Additionally, there is a slight but distinguishable separation between the reducing and non-reducing KR domains in IC_10 and IC_21, which are the ICs containing residues from both KRc and KRs parts.

The results suggest that A and B type KR domains can also be separated along with the ICs, although there are too few examples of B2 type for any conclusions to be drawn for these. IC_17 shows a distinction for A-type and B1 type sequences (Fig. 4.24A). This suggests that IC_17 residues might be functional in the determination of whether the beta-hydroxyl is L or D configured. B type KR domains have a sequence fingerprint of an LDD motif, of which the KR1 of the DEBS pathway, the one we are studying here, is an example. The LD of this motif is present in IC_17. The final D of the motif has been shown to be important and is hypothesised to control the direction of entry of the substrate into the active site, controlling the chirality of the resulting beta-hydroxyl Keatinge-Clay and Stroud 2006b, and this residue is detected in IC_3, which has the highly conserved KRc positions. A type KR domains have no LDD domain but rather have a conserved tryptophan and these two changes, as compared to B type, are thought lead to substrate entering with the opposite face leading to the L stereo-configuration of the resulting beta-hydroxyl group. This conserved tryptophan is detected in IC_10.

Clustering the sequences of KR domains is shown in Fig. 4.24B is similar to the clustering of sequences based solely on the residues in IC_17. Although the trees are similar, KR sub-types cluster more closely when all domain positions are used for analysis.

With a similar approach as we applied on the AT domain to identify the residues more critical for different subtypes, we identified the residues of IC_17 that are important in the type-B1 specification. To determine the difference between type-B1 and type-A1 KRs, we substracted the scores of projected ICs. Since the number of sequences of type-B2 and type-A2 is low, we applied this approach only on types B1 and A1.

$$\Delta\tilde{V}^{p}_{1\cdots k^*} = \tilde{V}^{p_{B1}}_{1\cdots k^*} - \tilde{V}^{p_{A1}}_{1\cdots k^*} \tag{4.9}$$

The positions of the IC was sorted based on $\Delta\tilde{V}^{p}_{IC\_17}$ scores (Fig. 4.24D). We can detect LD residues of LDD motif close to the right end consistent with the importance of these residues for type B specification (Fig. 4.24D, marked with stars). Interestingly, we detect 12 more residues having higher scores than L1199 and D1200 residues, suggesting these 12 positions should be considered for experimental studies attempting to switch KR domains to type B2.

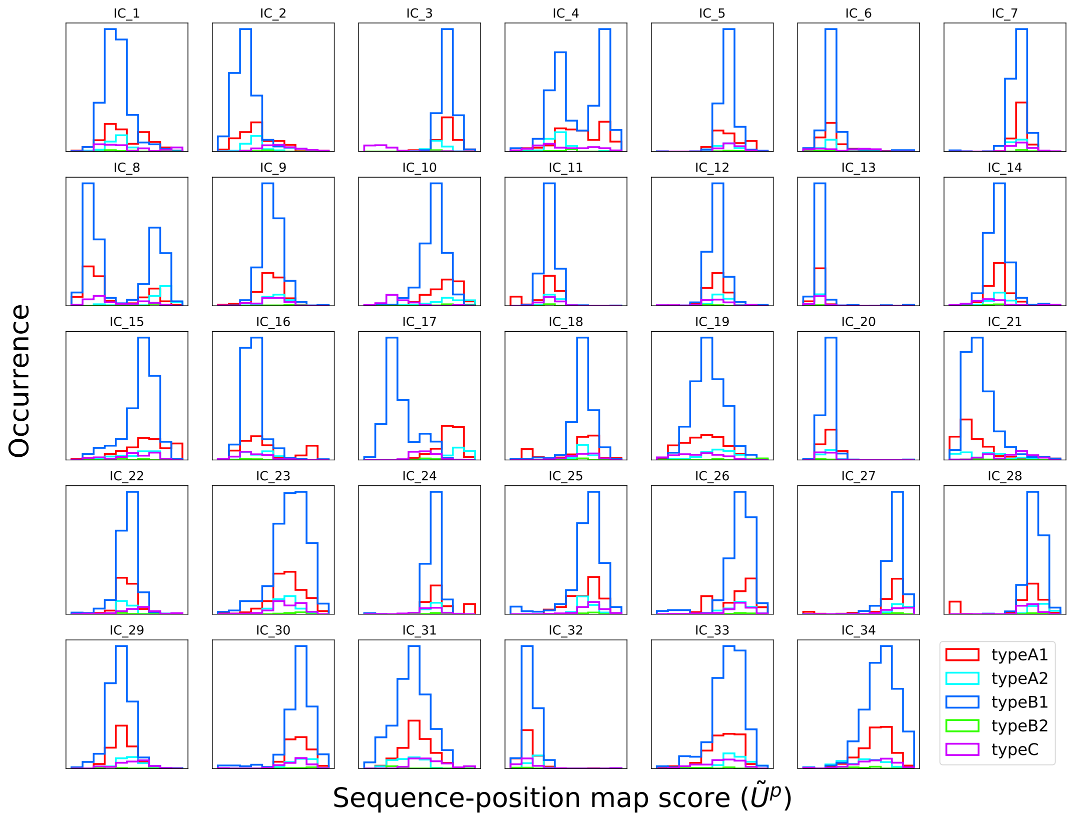


***Figure 4.23: Sequence-position mapping based on subtypes of the KR domain.** Although for most of the ICs, there is no clear distinction in the sequence patterns of the different domain compositions, IC_17 shows distinction between some subtypes of the KR domain.*

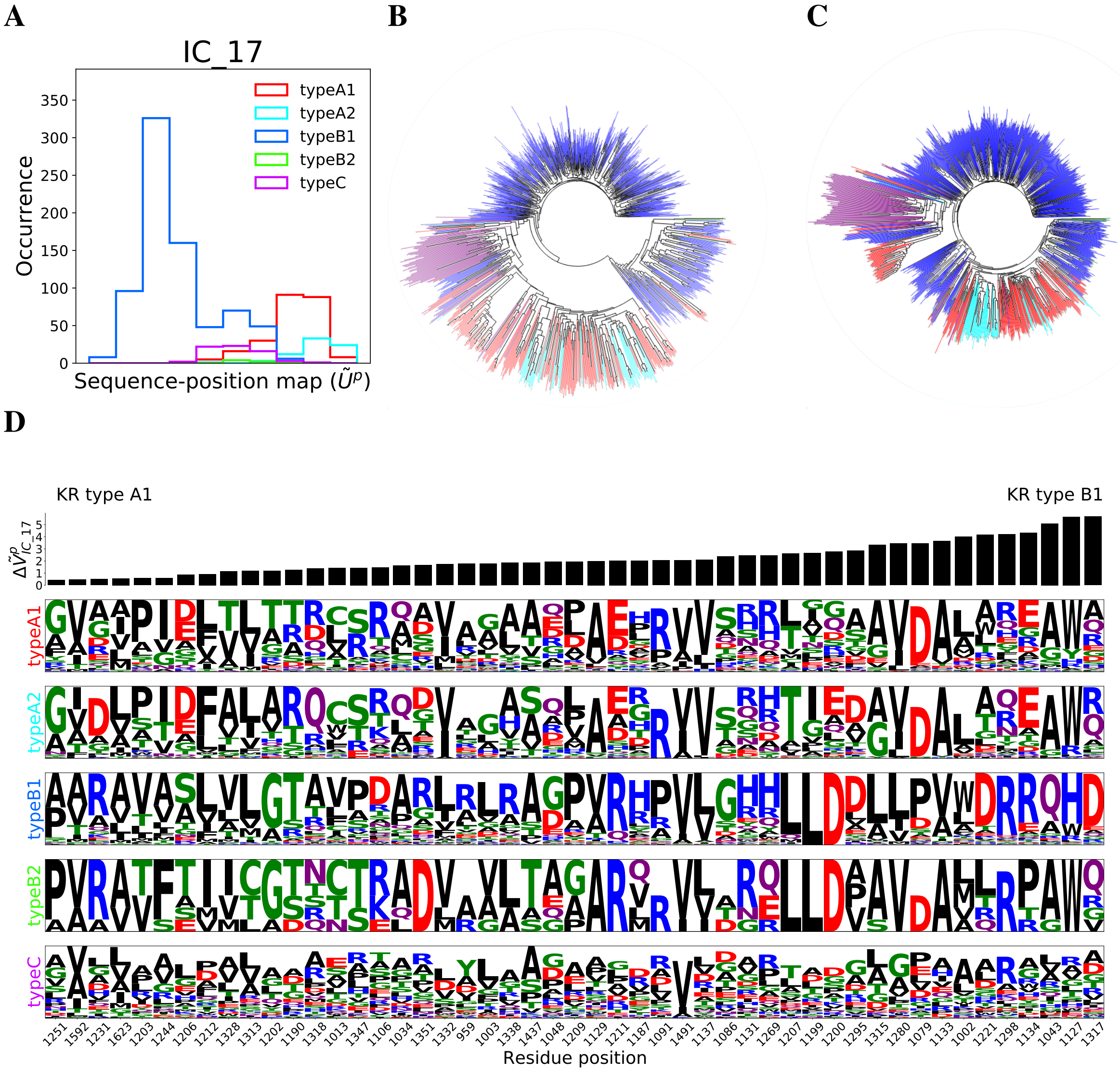


Figure 4.24: Different sequence patterns for type-B and type-A KR domains are detected in IC_17. *Sequence-position map separates the KR domains based on their sub-types (A). Phylogenetic analysis of the sequences of IC_17 residues distinguishes type A and type B KRs supporting the sequence-position mapping patterns of the sequences (B). Pyhlogenetic analysis of whole KR domain reveals very similar clustering pattern as IC_17 (C). Sequence logos shows that positions detected in IC_17 have more conserved pattern in type B KRs compared to type A (D). Positions are sorted based on* $\Delta\tilde{V}^p_{IC\_17}$ *scores, where* $\Delta\tilde{V}^p_{IC\_17} = \tilde{V}^{p_{B1}}_{IC\_17} - \tilde{V}^{p_{A1}}_{IC\_17}$.

4.4 Conclusion

With the SCA, we can detect functional domain boundaries consistent with experimental studies and that explains the importance of domain boundary optimisation for successful domain swap experiments Yuzawa *et al.* 2017, since residues within the domain boundaries have clearly been under selective pressure to function together. Furthermore, sequence-position mapping analysis shows that there are groups of residues co-evolved specifically within different domain subtypes. This is consistent with the experimental results since mutating only a few residues has not switched the specificity of a domain from one subtype to another particular one.

In this study, we made extensions on the existing methodology to apply it on multi-domain proteins and to determine key residues potentially important for the sub-type specificity.

Chapter 5

Conclusion

The structures of proteins and their working mechanisms have been a widely studied area of biology. Knowing protein structure and function helps answer many questions about diseases and is a basis for discovering new drugs. As the function of a protein is determined by its three dimensional (3D) structure, various experimental approaches have been developed to determine the structure. Although experimental approaches are very successful in certain conditions, they have limitations. For example, crystallization of proteins for X-ray crystallography is a very challenging and time-consuming step, and NMR is limited to only for small-sized proteins. Additionally, both of them are insufficient to determine the structures of mega-Dalton sized proteins and protein complexes like polyketide synthases (PKS). Electron microscopy, on the other hand, is a promising approach to determine larger structures. As of September 2020, all of 969 structures whose size is larger than 2 MDa - the size of DEBS, which is one of the smallest PKSs - were determined via electron microscopy. The resolution of just one of these structures, whose size is larger than 2MDa, is lower than 2 Å (PDB ID: 6e9d), indicating how challenging it is to determine the structures of mega-Dalton sized proteins and protein complexes. While researchers have been working to solve these experimental limitations, recent developments in computational approaches to determine the structure of proteins are in the progress of being a fast and reliable alternative.

Here in this thesis, firstly, I worked on *ab initio* protein structure prediction problem. Later, I applied the successful approaches in *ab initio* protein structure prediction area to the challenge of structure prediction of multi-domain proteins. Lastly, I used another approach to study multi-domain proteins and as a model system, I used a DEBS module.

In the second chapter, I explained how we contributed to *ab initio* protein structure prediction. Besides predicting contacting residue pairs in a 3D structure, we successfully predicted distances between the residue pairs via deep neural networks, resulting in an improvement in the quality of predicted structures. In this project, I mostly focused on the analyses of (i) whether having structures in the test set sharing the same topology or homologous superfamily class with the training set cause a bias in the prediction, (ii) the effect of predicting more structures in the accuracy of the final structure, and (iii) the proportion of amino acid types in correctly predicted pairs, as well as analysing the effect of distance predictions on the structure prediction accuracy with Shuangxi Ji.

Overall, we showed that not only contacting pairs but also residue pairs in longer distances can be predicted. Implementation of distance constraints into Rosetta for *ab initio* prediction improves the structure quality as well as allowing the selection of better models than using the Rosetta score without the restraint energy. It was also shown that having structurally similar proteins (based on the CATH classification) in the training set and the test set did not cause any bias in our predictions. Increasing the structure pool size from 100 to 200 generated better models (higher TM-score); however, a better model could not be always selected with Rosetta energy. Lastly, it was demonstrated that the proportion of amino acid types in correctly predicted pairs changes depending on the distance between the pairs. Contribution of distance predictions to *ab initio* protein structure prediction problem was further demonstrared by many other groups (Zhu *et al.* 2018; Greener *et al.* 2019; Xu 2019; Senior *et al.* 2019; Senior *et al.* 2020).

In the third chapter, I applied a similar approach, which was shown as successful for prediction of small/medium proteins, to predict how two domains on the same chain interact with each other. For this purpose, different feature sets were tested for their ability to predict the distance potentials between inter-domain residue pairs. Successful domain-domain interactions could be predicted for almost half of the test set; whereas, for the domain pairs with fewer sequences in the alignment and the ones with smaller interface area, successful predictions could not be obtained. The trained network is used to predict interactions between the domain pairs of a fatty acid synthase, which has a similar domain organization as PKSs. Consistent with the results of the test proteins, successful predictions could only be obtained when the sequence alignment is comprehensive enough and the domain interface is large. To sum up, it is shown that the evolutionary information in an MSA can be mined for structural information that improves the prediction of the structures of domain-domain interactions. Moreover, criteria were identified for the likelihood of a successful prediction, namely the Nf value and the interface size.

For the prediction of how two domains interact, the experimental structures of the domains were used. As a further study, the same trained models can be tested using monomeric structures that were generated by structure prediction algorithms (rather than experimental structures), to see how the quality of prediction changes for the interaction prediction when there is no structural information available.

In the fourth chapter, I worked on understanding the multi-domain proteins better with an alternative approach. Here, I aimed to detect the coevolved residue groups in a multi-domain protein. DEBS was selected as a model PKS since it is one of the most studied PKS systems. Initially, the method was benchmarked using a number of small domains, to test the number of sequences needed for a converged result. Contrary to claims in the existing literature this was found to vary between systems from 100 to 1800 sequences. With this approach, the coevolved residue groups within the DEBS1 module 1 were detected. The domain boundaries of the AT and the KR domains were determined consistent with the experimental results in the literature. Further, sequence-position mapping allowed us to determine residue groups that are potentially important in domain sub-type specification.

In this study, the method was extended by (i) using a bootstrapping technique to look for convergence in the results of the analysis, and (ii) projecting subtype covariance data onto the ICs explaining the whole data set to determine key residues defining specificity, which still needs to be tested experimentally but seems to agree with existing experimental data and provides a means for identifying critical residues for determining specificity, and thus allowing rational reengineering.

List of References

Abadi, Martín, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng (2015). *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. Software available from tensorflow.org (cit. on pp. 42, 76).

Altschul, S. (1997). "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs". *Nucleic Acids Research* 25.17, pp. 3389–3402 (cit. on p. 10).

Altschul, Stephen F., Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman (1990). "Basic local alignment search tool". *Journal of Molecular Biology* 215.3, pp. 403–410 (cit. on pp. 5, 10).

Annaval, Thibault, Cédric Paris, Peter F. Leadlay, Christophe Jacob, and Kira J. Weissman. "Evaluating Ketoreductase Exchanges as a Means of Rationally Altering Polyketide Stereochemistry". *ChemBioChem* 16.9 (), pp. 1357–1364 (cit. on p. 36).

Baerga-Ortiz, Abel, Bojana Popovic, Alexandros P. Siskos, Helen M. O'Hare, Dieter Spiteller, Mark G. Williams, Nuria Campillo, Jonathan B. Spencer, and Peter F. Leadlay (2006). "Directed Mutagenesis Alters the Stereochemistry of Catalysis by Isolated Ketoreductase Domains from the Erythromycin Polyketide Synthase". *Chemistry & Biology* 13.3, pp. 277–285 (cit. on p. 37).

Bailey, Constance B., Marjolein E. Pasman, and Adrian T. Keatinge-Clay (2016). "Substrate structure–activity relationships guide rational engineering of modular polyketide synthase ketoreductases". *Chem. Commun.* 52 (4), pp. 792–795 (cit. on p. 37).

Barajas, Jesus F., Jacquelyn M. Blake-Hedges, Constance B. Bailey, Samuel Curran, and Jay. D. Keasling (2017). "Engineered polyketides: Synergy between protein and host level engineering". *Synthetic and Systems Biotechnology* 2.3, pp. 147 –166 (cit. on pp. 33, 100).

Bayly, Carmen L. and Vikramaditya G. Yadav (2017). *Towards precision engineering of canonical polyketide synthase domains: Recent advances and future prospects* (cit. on pp. 30, 31).

Berman, H. M. (2000). "The Protein Data Bank". *Nucleic Acids Research* 28.1, pp. 235–242 (cit. on p. 2).

Betancourt, Marcos R. and D. Thirumalai (2008). “Pair potentials for protein folding: Choice of reference states and sensitivity of predicted native states to variations in the interaction schemes”. *Protein Science* 8.2, pp. 361–369 (cit. on pp. 15, 41, 74, 159).

Brünger, A. T., P. D. Adams, G. M. Clore, W. L. DeLano, P. Gros, R. W. Grosse-Kunstleve, J. S. Jiang, J. Kuszewski, M. Nilges, N. S. Pannu, R. J. Read, L. M. Rice, T. Simonson, and G. L. Warren (1998). “Crystallography & NMR System: A New Software Suite for Macromolecular Structure Determination”. *Acta Crystallographica Section D Biological Crystallography* 54.5, pp. 905–921 (cit. on p. 69).

Chakraborty, Hirak Jyoti, Aditi Gangopadhyay, Sayak Ganguli, and Abhijit Datta (2017). “Protein structure prediction”. *Applying Big Data Analytics in Bioinformatics and Medicine* (cit. on p. 6).

Cheng, Jianlin, Myong-Ho Choe, Arne Elofsson, Kun-Sop Han, Jie Hou, Ali H. A. Maghrabi, Liam J. McGuffin, David Menéndez-Hurtado, Kliment Olechnovič, Torsten Schwede, Gabriel Studer, Karolis Uziela, Česlovas Venclovas, and Björn Wallner (2019). “Estimation of model accuracy in CASP13”. *Proteins: Structure, Function, and Bioinformatics* 87.12, pp. 1361–1377 (cit. on pp. 62, 70).

Chollet, François *et al.* (2015). *Keras*. `https://keras.io` (cit. on pp. 42, 76).

Chothia, C. (2003). “Evolution of the Protein Repertoire”. *Science* 300.5626, pp. 1701–1703 (cit. on p. 1).

Chothia, Cyrus (1992). *One thousand families for the molecular biologist* (cit. on p. 6).

Cocco, Simona, Remi Monasson, and Martin Weigt (2013). “From Principal Component to Direct Coupling Analysis of Coevolution in Proteins: Low-Eigenvalue Modes are Needed for Structure Prediction”. *PLoS Computational Biology* 9.8. Ed. by Björn Wallner, e1003176 (cit. on p. 24).

Croll, Tristan I., Massimo D. Sammito, Andriy Kryshtafovych, and Randy J. Read (2019). “Evaluation of template-based modeling in CASP13”. *Proteins: Structure, Function, and Bioinformatics* 87.12, pp. 1113–1127 (cit. on p. 5).

Dapkunas, Justas, Albertas Timinskas, Kliment Olechnovič, Mindaugas Margelevičius, Rytis Dičiunas, Česlovas Venclovas, and Anna Tramontano (2017). “The PPI3D web server for searching, analyzing and modeling protein-protein interactions in the context of 3D structures”. *Bioinformatics* (cit. on pp. 22, 24).

Dawson, Natalie L., Tony E. Lewis, Sayoni Das, Jonathan G. Lees, David Lee, Paul Ashford, Christine A. Orengo, and Ian Sillitoe (2016). “CATH: an expanded resource to predict protein function through structure and sequence”. *Nucleic Acids Research* 45.D1, pp. D289–D295 (cit. on p. 43).

Demuth, Howard, M Beale, and M Hagan (2000). “Neural network toolbox User’s guide: For Use with MATLAB”. *The Mathworks* (cit. on p. 18).

Dunn, S.D., L.M. Wahl, and G.B. Gloor (2007). “Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction”. *Bioinformatics* 24.3, pp. 333–340 (cit. on pp. 41, 74, 159).

Dutta, Somnath, Jonathan R Whicher, Douglas A Hansen, Wendi A Hale, Joseph A Chemler, Grady R Congdon, Alison R H Narayan, Kristina Håkansson, David H Sherman, Janet L Smith, and Georgios Skiniotis (2014). "Structure of a modular polyketide synthase". *Nature* 510.7506, 512—517 (cit. on pp. 31, 32).

Eddy, Sean R. (2011). "Accelerated Profile HMM Searches". *PLoS Computational Biology* 7.10. Ed. by William R. Pearson, e1002195 (cit. on pp. 10, 101).

Edwards, Andrea L., Tsutomu Matsui, Thomas M. Weiss, and Chaitan Khosla (2014). "Architectures of Whole-Module and Bimodular Proteins from the 6-Deoxyerythronolide B Synthase". *Journal of Molecular Biology* 426.11, pp. 2229–2245 (cit. on pp. 31, 32).

Ekeberg, Magnus, Cecilia Lövkvist, Yueheng Lan, Martin Weigt, and Erik Aurell (2013). "Improved contact prediction in proteins: Using pseudolikelihoods to infer Potts models". *Phys. Rev. E* 87 (1), p. 012707 (cit. on pp. 12, 20).

Ekman, Diana, Åsa K. Björklund, Johannes Frey-Skött, and Arne Elofsson (2005). "Multidomain Proteins in the Three Kingdoms of Life: Orphan Domains and Other Unassigned Regions". *Journal of Molecular Biology* 348.1, pp. 231–243 (cit. on pp. 1, 2, 22).

El-Gebali, Sara, Jaina Mistry, Alex Bateman, Sean R Eddy, Aurélien Luciani, Simon C Potter, Matloob Qureshi, Lorna J Richardson, Gustavo A Salazar, Alfredo Smart, Erik L L Sonnhammer, Layla Hirsh, Lisanna Paladin, Damiano Piovesan, Silvio C E Tosatto, and Robert D Finn (2018). "The Pfam protein families database in 2019". *Nucleic Acids Research* 47.D1, pp. D427–D432 (cit. on p. 101).

Eng, Clara H., Satoshi Yuzawa, George Wang, Edward E. K. Baidoo, Leonard Katz, and Jay D. Keasling (2016). "Alteration of Polyketide Stereochemistry from anti to syn by a Ketoreductase Domain Exchange in a Type I Modular Polyketide Synthase Subunit". *Biochemistry* 55.12. PMID: 26976746, pp. 1677–1680 (cit. on p. 36).

Eswar, Narayanan, Ben Webb, Marc A. Marti-Renom, M.S. Madhusudhan, David Eramian, Min yi Shen, Ursula Pieper, and Andrej Sali (2006). "Comparative Protein Structure Modeling Using Modeller". *Current Protocols in Bioinformatics* 15.1 (cit. on p. 5).

Eswar, Narayanan, Ben Webb, Marc A. Marti-Renom, M.S. Madhusudhan, David Eramian, Min-yi Shen, Ursula Pieper, and Andrej Sali (2007). "Comparative Protein Structure Modeling Using MODELLER". *Current Protocols in Protein Science* (cit. on p. 5).

Greener, Joe G., Shaun M. Kandathil, and David T. Jones (2019). "Deep learning extends de novo protein modelling coverage of genomes using iteratively predicted structural constraints". *Nature Communications* (cit. on pp. 14, 143).

Gront, Dominik, Daniel W. Kulp, Robert M. Vernon, Charlie E. M. Strauss, and David Baker (2011). "Generalized Fragment Picking in Rosetta: Design, Protocols and Applications". *PLoS ONE* 6.8. Ed. by Vladimir N. Uversky, e23294 (cit. on p. 9).

Groups Analysis: zscores - CASP13. `https://www.predictioncenter.org/casp13/zscores_final.cgi`. (Accessed on 08/04/2020) (cit. on p. 15).

Guex, Nicolas and Manuel C. Peitsch (1997). “SWISS-MODEL and the Swiss-Pdb Viewer: An environment for comparative protein modeling”. *Electrophoresis* 18.15, pp. 2714–2723 (cit. on p. 5).

Guzenko, Dmytro, Aleix Lafita, Bohdan Monastyrskyy, Andriy Kryshtafovych, and Jose M. Duarte (2019). “Assessment of protein assembly prediction in CASP13”. *Proteins: Structure, Function, and Bioinformatics* 87.12, pp. 1190–1199 (cit. on pp. 24, 72).

Halabi, Najeeb, Olivier Rivoire, Stanislas Leibler, and Rama Ranganathan (2009). “Protein Sectors: Evolutionary Units of Three-Dimensional Structure”. *Cell* 138.4, pp. 774–786 (cit. on pp. 24, 25, 103).

Haydock, Stephen F., Jesús F. Aparicio, István Molnár, Torsten Schwecke, Lake Ee Khaw, Ariane König, Andrew F.A. Marsden, Ian S. Galloway, James Staunton, and Peter F. Leadlay (1995). “Divergent sequence motifs correlated with the substrate specificity of (methyl)malonyl-CoA:acyl carrier protein transacylase domains in modular polyketide synthases”. *FEBS Letters* 374.2, pp. 246–248 (cit. on pp. 35, 102).

He, Baoji, S M Mortuza, Yanting Wang, Hong-Bin Shen, and Yang Zhang (2017). “NeBcon: protein contact map prediction using neural network training coupled with naïve Bayes classifiers”. *Bioinformatics* 33.15. Ed. by Alfonso Valencia, pp. 2296–2306 (cit. on p. 162).

Hopf, Thomas A., Charlotta P.I. Schärfe, João P.G.L.M. Rodrigues, Anna G. Green, Oliver Kohlbacher, Chris Sander, Alexandre M.J.J. Bonvin, and Debora S. Marks (2014). “Sequence co-evolution gives 3D contacts and structures of protein complexes”. *eLife* (cit. on p. 23).

Hou, Qingzhen, Paul F.G. De Geest, Wim F. Vranken, Jaap Heringa, and K. Anton Feenstra (2017). “Seeing the trees through the forest: Sequencebased homo- and heteromeric protein-protein interaction sites prediction using random forest”. *Bioinformatics* (cit. on p. 23).

Hsieh, Cho-Jui, Mátyás A. Sustik, Inderjit S. Dhillon, and Pradeep Ravikumar (2014). “QUIC: Quadratic Approximation for Sparse Inverse Covariance Estimation”. *Journal of Machine Learning Research* 15.83, pp. 2911–2947 (cit. on pp. 41, 159).

Ji, Shuangxi (2019). “Improving Protein Structure Prediction Using Amino Acid Contact & Distance Prediction.” PhD thesis. University of Birmingham (cit. on pp. 42, 69, 160, 162).

Ji, Shuangxi, Tugce Oruc, Liam Mead, Muhammad Fayyaz Rehman, Christopher Morton Thomas, Sam Butterworth, and Peter James Winn (2019). “DeepCDpred: Inter-residue distance and contact prediction for improved prediction of protein structure”. *PLOS ONE* 14.1, pp. 1–15 (cit. on pp. 14, 41).

Jones, D. T., W. R. Taylort, and J. M. Thornton (1992). “A new approach to protein fold recognition”. *Nature* 358.6381, pp. 86–89 (cit. on p. 6).

Jones, David T (1999). “Protein secondary structure prediction based on position-specific scoring matrices 1 1Edited by G. Von Heijne”. *Journal of Molecular Biology* 292.2, pp. 195–202 (cit. on p. 15).

Jones, David T., Daniel W. A. Buchan, Domenico Cozzetto, and Massimiliano Pontil (2011). "PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments". *Bioinformatics* 28.2, pp. 184–190 (cit. on p. 11).

Jones, David T., Daniel W.A. Buchan, Domenico Cozzetto, and Massimiliano Pontil (2012). "PSICOV: Precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments". *Bioinformatics* (cit. on p. 20).

Jones, David T., Tanya Singh, Tomasz Kosciolek, and Stuart Tetchner (2015). "MetaPSICOV: Combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins". *Bioinformatics* (cit. on pp. 12, 14, 16, 20, 40–42, 68, 74, 159, 160, 162).

Juan, Sheng-Hung, Teng-Ruei Chen, and Wei-Cheng Lo (2020). "A simple strategy to enhance the speed of protein secondary structure prediction without sacrificing accuracy". *PLOS ONE* 15.6. Ed. by M. Sohel Rahman, e0235153 (cit. on p. 15).

Kaján, László, Thomas A. Hopf, Matúš Kalaš, Debora S. Marks, and Burkhard Rost (2014). "FreeContact: Fast and free software for protein contact prediction from residue co-evolution". *BMC Bioinformatics* (cit. on pp. 12, 20).

Kalev, Ivan and Michael Habeck (2011). "HHfrag: HMM-based fragment detection using HH-pred". *Bioinformatics* 27.22, pp. 3110–3116 (cit. on p. 9).

Kalkreuter, Edward, Jared M. CroweTipton, Andrew N. Lowell, David H. Sherman, and Gavin J. Williams (2019). "Engineering the Substrate Specificity of a Modular Polyketide Synthase for Installation of Consecutive Non-Natural Extender Units". *Journal of the American Chemical Society* 141.5, pp. 1961–1969 (cit. on p. 35).

Kamisetty, H., S. Ovchinnikov, and D. Baker (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* 110.39, pp. 15674–15679 (cit. on p. 12).

Kandathil, Shaun M., Mario Garza-Fabre, Julia Handl, and Simon C. Lovell (2018). "Improved fragment-based protein structure prediction by redesign of search heuristics". *Scientific Reports* 8.1 (cit. on p. 9).

KC, Dukka B (2016). "Recent advances in sequence-based protein structure prediction: Table 1". *Briefings in Bioinformatics*, bbw070 (cit. on p. 5).

Keatinge-Clay, Adrian T. (2007). "A Tylosin Ketoreductase Reveals How Chirality Is Determined in Polyketides". *Chemistry & Biology* 14.8, pp. 898–908 (cit. on p. 115).

Keatinge-Clay, Adrian T. and Robert M. Stroud (2006a). "The Structure of a Ketoreductase Determines the Organization of the -Carbon Processing Enzymes of Modular Polyketide Synthases". *Structure* 14.4, pp. 737–748 (cit. on p. 115).

— (2006b). "The Structure of a Ketoreductase Determines the Organization of the -Carbon Processing Enzymes of Modular Polyketide Synthases". *Structure* 14.4, pp. 737 –748 (cit. on pp. 113, 115, 137).

Kellenberger, Laurenz, Ian S. Galloway, Guido Sauter, Günter Böhm, Ulf Hanefeld, Jesús Cortés, James Staunton, and Peter F. Leadlay. "A Polylinker Approach to Reductive Loop Swaps in Modular Polyketide Synthases". *ChemBioChem* 9.16 (), pp. 2740–2749 (cit. on p. 36).

Khor, Bee Yin, Gee Jun Tye, Theam Soon Lim, and Yee Siew Choong (2015). "General overview on structure prediction of twilight-zone proteins". *Theoretical Biology and Medical Modelling* 12.1 (cit. on p. 6).

Kikhney, Alexey G. and Dmitri I. Svergun (2015). "A practical guide to small angle X-ray scattering (SAXS) of flexible and intrinsically disordered proteins". *FEBS Letters* 589.19PartA, pp. 2570–2577 (cit. on p. 3).

Kmiecik, Sebastian, Dominik Gront, Michal Kolinski, Lukasz Wieteska, Aleksandra Elzbieta Dawid, and Andrzej Kolinski (2016). "Coarse-Grained Protein Models and Their Applications". *Chemical Reviews* 116.14, pp. 7898–7936 (cit. on p. 8).

Kornfuehrer, Taylor and Alessandra S. Eustáquio (2019). "Diversification of polyketide structures via synthase engineering". *Med. Chem. Commun.* Pp. – (cit. on pp. 33, 100).

Koryakina, Irina, Christian Kasey, John B. McArthur, Andrew N. Lowell, Joseph A. Chemler, Shasha Li, Douglas A. Hansen, David H. Sherman, and Gavin J. Williams (2017). "Inversion of Extender Unit Selectivity in the Erythromycin Polyketide Synthase by Acyltransferase Domain Engineering". *ACS Chemical Biology* 12.1, pp. 114–123 (cit. on p. 35).

Krieger, Elmar, Tom Darden, Sander B. Nabuurs, Alexei Finkelstein, and Gert Vriend (2004). "Making optimal use of empirical energy functions: Force-field parameterization in crystal space". *Proteins: Structure, Function, and Bioinformatics* 57.4, pp. 678–683 (cit. on p. 69).

Kryshtafovych, Andriy, Torsten Schwede, Maya Topf, Krzysztof Fidelis, and John Moult (2019). "Critical assessment of methods of protein structure prediction (CASP)—Round XIII". *Proteins: Structure, Function, and Bioinformatics* 87.12, pp. 1011–1020 (cit. on p. 13).

Kukic, Predrag, Claudio Mirabello, Giuseppe Tradigo, Ian Walsh, Pierangelo Veltri, and Gianluca Pollastri (2014). "Toward an accurate prediction of inter-residue distances in proteins using 2D recursive neural networks". *BMC Bioinformatics* (cit. on pp. 14, 40).

Kwan, David H., Manuela Tosin, Nadin Schläger, Frank Schulz, and Peter F. Leadlay (2011). "Insights into the stereospecificity of ketoreduction in a modular polyketide synthase". *Org. Biomol. Chem.* 9 (7), pp. 2053–2056 (cit. on p. 37).

Lau, Janice, Hong Fu, David E. Cane, and Chaitan Khosla (1999). "Dissecting the Role of Acyltransferase Domains of Modular Polyketide Synthases in the Choice and Stereochemical Fate of Extender Units". *Biochemistry* 38.5. PMID: 9931032, pp. 1643–1651 (cit. on p. 34).

Lindorff-Larsen, K., S. Piana, R. O. Dror, and D. E. Shaw (2011). "How Fast-Folding Proteins Fold". *Science* 334.6055, pp. 517–520 (cit. on p. 7).

Lindorff-Larsen, Kresten, Stefano Piana, Kim Palmo, Paul Maragakis, John L. Klepeis, Ron O. Dror, and David E. Shaw (2010). "Improved side-chain torsion potentials for the Amber ff99SB protein force field". *Proteins: Structure, Function, and Bioinformatics* 78.8, pp. 1950–1958 (cit. on p. 69).

Liu, Lu, Arinthip Thamchaipenet, Hong Fu, Mary Betlach, and Gary Ashley (1997). "Biosynthesis of 2-Nor-6-deoxyerythronolide B by Rationally Designed Domain Substitution". *Journal of the American Chemical Society* 119.43, pp. 10553–10554 (cit. on p. 34).

Lockless, S. W. (1999). "Evolutionarily Conserved Pathways of Energetic Connectivity in Protein Families". *Science* 286.5438, pp. 295–299 (cit. on p. 25).

Maier, James A., Carmenza Martinez, Koushik Kasavajhala, Lauren Wickstrom, Kevin E. Hauser, and Carlos Simmerling (2015). "ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB". *Journal of Chemical Theory and Computation* 11.8, pp. 3696–3713 (cit. on p. 69).

Malhotra, Sony, Sylvain Träger, Matteo Dal Peraro, and Maya Topf (2019). "Modelling structures in cryo-EM maps". *Current Opinion in Structural Biology* 58, pp. 105–114 (cit. on p. 2).

Mead, Liam (2018). "'The optimisation of neural networks to increase the accuracy of amino acid contact predictions' and 'Structural studies of proteins required for gliding motility in the predatory bacteria Bdellovibrio bacteriovorus'". MA thesis. University of Birmingham (cit. on pp. 42, 49, 160).

Mirabello, Claudio and Björn Wallner (2017). "InterPred: A pipeline to identify and model protein–protein interactions". *Proteins: Structure, Function and Bioinformatics* (cit. on p. 22).

Miyazawa, Sanzo and Robert L. Jernigan (1996). "Residue – Residue Potentials with a Favorable Contact Pair Term and an Unfavorable High Packing Density Term, for Simulation and Threading". *Journal of Molecular Biology* 256.3, pp. 623–644 (cit. on p. 15).

Morcos, F., A. Pagnani, B. Lunt, A. Bertolino, D. S. Marks, C. Sander, R. Zecchina, J. N. Onuchic, T. Hwa, and M. Weigt (2011). "Direct-coupling analysis of residue coevolution captures native contacts across many protein families". *Proceedings of the National Academy of Sciences* 108.49, E1293–E1301 (cit. on pp. 12, 41, 159).

Moult, John, Jan T. Pedersen, Richard Judson, and Krzysztof Fidelis (1995). "A large-scale experiment to assess protein structure prediction methods". *Proteins: Structure, Function, and Genetics* 23.3, pp. ii–iv (cit. on p. 4).

Moult, John, Krzysztof Fidelis, Andriy Kryshtafovych, Torsten Schwede, and Anna Tramontano (2016). "Critical assessment of methods of protein structure prediction: Progress and new directions in round XI". *Proteins: Structure, Function, and Bioinformatics* 84, pp. 4–14 (cit. on p. 5).

Musiol-Kroll, Ewa Maria and Wolfgang Wohlleben (2018). "Acyltransferases as Tools for Polyketide Synthase Engineering". *Antibiotics* 7.3 (cit. on pp. 33, 100).

Narayanan, Chitra, Donald Gagné, Kimberly A. Reynolds, and Nicolas Doucet (2017). "Conserved amino acid networks modulate discrete functional properties in an enzyme superfamily". *Scientific Reports* 7.1, p. 3207 (cit. on pp. 25, 27).

Northey, Thomas C., Anja BarešiÄ ‡, and Andrew C.R. Martin (2018). “IntPred: A structure-based predictor of protein-protein interaction sites”. *Bioinformatics* (cit. on p. 23).

Oliveira, Saulo H. P. de, Jiye Shi, and Charlotte M. Deane (2015). “Building a Better Fragment Library for De Novo Protein Structure Prediction”. *PLOS ONE* 10.4. Ed. by Yang Zhang, e0123998 (cit. on p. 9).

Oliynyk, Markiyan, Murray J.B. Brown, Jesús Cortés, James Staunton, and Peter F. Leadlay (1996). “A hybrid modular polyketide synthase obtained by domain swapping”. *Chemistry & Biology* 3.10, pp. 833 –839 (cit. on p. 34).

Orengo, Christine A., David T. Jones, and Janet M. Thornton (1994). “Protein superfamilles and domain superfolds”. *Nature* (cit. on p. 48).

Ovchinnikov, Sergey, Hetunandan Kamisetty, and David Baker (2014). “Robust and accurate prediction of residue–residue interactions across protein interfaces using evolutionary information”. *eLife* 3 (cit. on p. 98).

Ovchinnikov, Sergey, Lisa Kinch, Hahnbeom Park, Yuxing Liao, Jimin Pei, David E Kim, Hetunandan Kamisetty, Nick V Grishin, and David Baker (2015). “Large-scale determination of previously unsolved protein structures using evolutionary information”. *eLife* 4 (cit. on pp. 50, 57).

Ovchinnikov, Sergey, Hahnbeom Park, Neha Varghese, Po-Ssu Huang, Georgios A. Pavlopoulos, David E. Kim, Hetunandan Kamisetty, Nikos C. Kyrpides, and David Baker (2017). “Protein structure determination using metagenome sequence data”. *Science* 355.6322, pp. 294–298 (cit. on pp. 50, 57).

Petkovic, Hrvoje, Rachel E. Lill, Rose M. Sheridan, Barrie Wilkinson, Ellen L. McCormick, Hamish A. I. McArthur, James Staunton, Peter F. Leadlay, and Steven G. Kendrew (2003). “A Novel Erythromycin, 6-Desmethyl Erythromycin D, Made by Substituting an Acyltransferase Domain of the Erythromycin Polyketide Synthase”. *The Journal of Antibiotics* 56.6, pp. 543–551 (cit. on p. 34).

Piana, S., K. Lindorff-Larsen, and D. E. Shaw (2013). “Atomic-level description of ubiquitin folding”. *Proceedings of the National Academy of Sciences* 110.15, pp. 5915–5920 (cit. on p. 8).

Raval, Alpan, Stefano Piana, Michael P. Eastwood, and David E. Shaw (2015). “Assessment of the utility of contact-based restraints in accelerating the prediction of protein structure using molecular dynamics simulations”. *Protein Science* 25.1, pp. 19–29 (cit. on p. 8).

Remmert, Michael, Andreas Biegert, Andreas Hauser, and Johannes Söding (2011). “HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment”. *Nature Methods* 9, 173 EP – (cit. on pp. 101, 159).

Rivoire, Olivier (2013). “Elements of Coevolution in Biological Sequences”. *Physical Review Letters* 110.17 (cit. on p. 24).

Rivoire, Olivier, Kimberly A. Reynolds, and Rama Ranganathan (2016). “Evolution-Based Functional Decomposition of Proteins”. *PLOS Computational Biology* 12.6, pp. 1–26 (cit. on pp. 24, 25, 27, 74, 101, 103, 105, 106, 112, 124).

Robbins, Thomas, Yu-Chen Liu, David E Cane, and Chaitan Khosla (2016). “Structure and mechanism of assembly line polyketide synthases”. *Current Opinion in Structural Biology* 41, pp. 10–18 (cit. on pp. 30, 32).

Rohl, Carol A., Charlie E.M. Strauss, Kira M.S. Misura, and David Baker (2004). “Protein Structure Prediction Using Rosetta”. *Methods in Enzymology*. Elsevier, pp. 66–93 (cit. on p. 9).

Roy, Ambrish, Alper Kucukural, and Yang Zhang (2010). “I-TASSER: a unified platform for automated protein structure and function prediction”. *Nature Protocols* 5.4, pp. 725–738 (cit. on p. 7).

Ruan, X, A Pereda, D L Stassi, D Zeidner, R G Summers, M Jackson, A Shivakumar, S Kakavas, M J Staver, S Donadio, and L Katz (1997). “Acyltransferase domain substitutions in erythromycin polyketide synthase yield novel erythromycin derivatives.” *Journal of Bacteriology* 179.20, pp. 6416–6425 (cit. on p. 34).

Savojardo, Castrense, Piero Fariselli, Pier Luigi Martelli, and Rita Casadio (2017). “ISPRED4: Interaction sites PREDiction in protein structures with a refining grammar model”. *Bioinformatics* (cit. on p. 23).

Schaarschmidt, Joerg, Bohdan Monastyrskyy, Andriy Kryshtafovych, and Alexandre M.J.J. Bonvin (2018). “Assessment of contact predictions in CASP12: Co-evolution and deep learning coming of age”. *Proteins: Structure, Function and Bioinformatics* (cit. on pp. 12, 13, 20, 40).

Seemayer, Stefan, Markus Gruber, and Johannes Söding (2014). “CCMpred - Fast and precise prediction of protein residue-residue contacts from correlated mutations”. *Bioinformatics* (cit. on pp. 12, 20, 41, 74, 159).

Senior, Andrew W., Richard Evans, John Jumper, James Kirkpatrick, Laurent Sifre, Tim Green, Chongli Qin, Augustin Žídek, Alexander W.R. Nelson, Alex Bridgland, Hugo Penedones, Stig Petersen, Karen Simonyan, Steve Crossan, Pushmeet Kohli, David T. Jones, David Silver, Koray Kavukcuoglu, and Demis Hassabis (2019). “Protein structure prediction using multiple deep neural networks in the 13th Critical Assessment of Protein Structure Prediction (CASP13)”. *Proteins: Structure, Function and Bioinformatics* (cit. on pp. 13, 21, 40, 68, 76, 143).

— (2020). “Improved protein structure prediction using potentials from deep learning”. *Nature* (cit. on pp. 13, 21, 40, 68, 76, 143).

Shaw, David E., Martin M. Deneroff, Ron O. Dror, Jeffrey S. Kuskin, Richard H. Larson, John K. Salmon, Cliff Young, Brannon Batson, Kevin J. Bowers, Jack C. Chao, Michael P. Eastwood, Joseph Gagliardo, J. P. Grossman, C. Richard Ho, Douglas J. Ierardi, István Kolossváry, John L. Klepeis, Timothy Layman, Christine McLeavey, Mark A. Moraes, Rolf Mueller, Edward C. Priest, Yibing Shan, Jochen Spengler, Michael Theobald, Brian Towles, and Stanley C. Wang (2008). “Anton, a special-purpose machine for molecular dynamics simulation”. *Communications of the ACM* 51.7, pp. 91–97 (cit. on p. 7).

Shaw, David E., J.P. Grossman, Joseph A. Bank, Brannon Batson, J. Adam Butts, Jack C. Chao, Martin M. Deneroff, Ron O. Dror, Amos Even, Christopher H. Fenton, Anthony

Forte, Joseph Gagliardo, Gennette Gill, Brian Greskamp, C. Richard Ho, Douglas J. Ierardi, Lev Iserovich, Jeffrey S. Kuskin, Richard H. Larson, Timothy Layman, Li-Siang Lee, Adam K. Lerer, Chester Li, Daniel Killebrew, Kenneth M. Mackenzie, Shark Yeuk-Hai Mok, Mark A. Moraes, Rolf Mueller, Lawrence J. Nociolo, Jon L. Peticolas, Terry Quan, Daniel Ramot, John K. Salmon, Daniele P. Scarpazza, U. Ben Schafer, Naseer Siddique, Christopher W. Snyder, Jochen Spengler, Ping Tak Peter Tang, Michael Theobald, Horia Toma, Brian Towles, Benjamin Vitale, Stanley C. Wang, and Cliff Young (2014). "Anton 2: Raising the Bar for Performance and Programmability in a Special-Purpose Molecular Dynamics Supercomputer". *SC14: International Conference for High Performance Computing, Networking, Storage and Analysis*. IEEE (cit. on pp. 7, 8).

Shen, Ben (2003). *Polyketide biosynthesis beyond the type I, II and III polyketide synthase paradigms* (cit. on pp. 28, 29).

Shrestha, Rojan, Eduardo Fajardo, Nelson Gil, Krzysztof Fidelis, Andriy Kryshtafovych, Bohdan Monastyrskyy, and Andras Fiser (2019). "Assessing the accuracy of contact predictions in CASP13". *Proteins: Structure, Function and Bioinformatics* (cit. on pp. 13, 14).

Simons, Kim T., Charles Kooperberg, Enoch Huang, and David Baker (1997). "Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and bayesian scoring functions". *Journal of Molecular Biology* 268.1, pp. 209–225 (cit. on p. 9).

Simons, Kim T., Rich Bonneau, Ingo Ruczinski, and David Baker (1999). "Ab initio protein structure prediction of CASP III targets using ROSETTA". *Proteins: Structure, Function, and Bioinformatics* 37.S3, pp. 171–176 (cit. on p. 162).

Skwark, Marcin J., Abbi Abdel-Rehim, and Arne Elofsson (2013). "PconsC: Combination of direct information methods and alignments improves contact prediction". *Bioinformatics* (cit. on pp. 11, 20).

Steinegger, Martin, Markus Meier, Milot Mirdita, Harald Vöhringer, Stephan J. Haunsberger, and Johannes Söding (2019). "HH-suite3 for fast remote homology detection and deep protein annotation". *BMC Bioinformatics* 20.1 (cit. on p. 11).

Tanaka, Seiji and Harold A. Scheraga (1976). "Medium- and Long-Range Interaction Parameters between Amino Acids for Predicting Three-Dimensional Structures of Proteins". *Macromolecules* 9.6, pp. 945–950 (cit. on p. 15).

Tetchner, Stuart (2015). "Computational Modelling of Multidomain Proteins with Covarying Residue Pairs." PhD thesis. University College London (cit. on p. 11).

Valenzano, Chiara R., Rachel J. Lawson, Alice Y. Chen, Chaitan Khosla, and David E. Cane (2009). "The Biochemical Basis for Stereochemical Control in Polyketide Biosynthesis". *Journal of the American Chemical Society* 131.51. PMID: 19928853, pp. 18501–18511 (cit. on p. 36).

Van Rossum, Guido and Fred L Drake Jr (1995). *Python reference manual*. Centrum voor Wiskunde en Informatica Amsterdam (cit. on pp. 45, 79, 106).

Walsh, Ian, Davide Baù, Alberto Jm Martin, Catherine Mooney, Alessandro Vullo, and Gianluca Pollastri (2009). *Ab initio and template-based prediction of multi-class distance maps by two-dimensional recursive neural networks* (cit. on pp. 14, 40).

Wang, G. and R. L. Dunbrack (2003). “PISCES: a protein sequence culling server”. *Bioinformatics* 19.12, pp. 1589–1591 (cit. on pp. 73, 160).

Wang, Sheng, Siqi Sun, Zhen Li, Renyu Zhang, and Jinbo Xu (2017). “Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model”. *PLoS Computational Biology* (cit. on pp. 12, 14, 20, 40, 68, 69, 160, 162).

Wang, Tong, Yuedong Yang, Yaoqi Zhou, and Haipeng Gong (2016). “LRFragLib: an effective algorithm to identify fragments for de novo protein structure prediction”. *Bioinformatics*, btw668 (cit. on pp. 9, 15).

Wang, Xiaoying, Bin Yu, Anjun Ma, Cheng Chen, Bingqiang Liu, and Qin Ma (2019). “Protein-protein interaction sites prediction by ensemble random forests with synthetic minority oversampling technique”. *Bioinformatics* (cit. on p. 23).

Waterhouse, Andrew, Martino Bertoni, Stefan Bienert, Gabriel Studer, Gerardo Tauriello, Rafal Gumienny, Florian T. Heer, Tjaart A.P. De Beer, Christine Rempfer, Lorenza Bordoli, Rosalba Lepore, and Torsten Schwede (2018). “SWISS-MODEL: Homology modelling of protein structures and complexes”. *Nucleic Acids Research* (cit. on p. 22).

Weigt, M., R. A. White, H. Szurmant, J. A. Hoch, and T. Hwa (2008). “Identification of direct residue contacts in protein-protein interaction by message passing”. *Proceedings of the National Academy of Sciences* 106.1, pp. 67–72 (cit. on p. 11).

Weissman, Kira J. (2015). “Uncovering the structures of modular polyketide synthases”. *Natural Product Reports* 32.3, pp. 436–453 (cit. on pp. 31, 33).

— (2016). “Genetic engineering of modular PKSs: from combinatorial biosynthesis to synthetic biology”. *Nat. Prod. Rep.* 33 (2), pp. 203–230 (cit. on pp. 33, 100).

Wollacott, A. M., A. Zanghellini, P. Murphy, and D. Baker (2006). “Prediction of structures of multidomain proteins from structures of the individual domains”. *Protein Science* (cit. on p. 22).

Wu, Sitao and Yang Zhang (2007). “LOMETS: A local meta-threading-server for protein structure prediction”. *Nucleic Acids Research* 35.10, pp. 3375–3382 (cit. on p. 7).

Xiang, Zhexin (2006). “Advances in Homology Protein Structure Modeling”. *Current Protein & Peptide Science* 7.3, pp. 217–227 (cit. on pp. 5, 6).

Xu, Dong, Lukasz Jaroszewski, Zhanwen Li, and Adam Godzik (2015). “AIDA: Ab initio domain assembly for automated multi-domain protein structure prediction and domain-domain interaction prediction”. *Bioinformatics* (cit. on p. 22).

Xu, Jinbo (2019). “Distance-based protein folding powered by deep learning”. *Proceedings of the National Academy of Sciences of the United States of America* (cit. on pp. 13, 14, 40, 143).

Xu, Jinrui and Yang Zhang (2010). "How significant is a protein structure similarity with TM-score = 0.5?" *Bioinformatics* 26.7, pp. 889–895 (cit. on p. 44).

Yang, Jianyi and Yang Zhang (2015). "I-TASSER server: new development for protein structure and function predictions". *Nucleic Acids Research* 43.W1, W174–W181 (cit. on p. 7).

Yang, Yuedong, Rhys Heffernan, Kuldip Paliwal, James Lyons, Abdollah Dehzangi, Alok Sharma, Jihua Wang, Abdul Sattar, and Yaoqi Zhou (2016). "SPIDER2: A Package to Predict Secondary Structure, Accessible Surface Area, and Main-Chain Torsional Angles by Deep Neural Networks". *Methods in Molecular Biology*. Springer New York, pp. 55–63 (cit. on pp. 15, 41, 74, 159, 160).

Yuzawa, Satoshi, Kai Deng, George Wang, Edward E. K. Baidoo, Trent R. Northen, Paul D. Adams, Leonard Katz, and Jay D. Keasling (2017). "Comprehensive in Vitro Analysis of Acyltransferase Domain Exchanges in Modular Polyketide Synthases and Its Application for Short-Chain Ketone Production". *ACS Synthetic Biology* 6.1. PMID: 27548700, pp. 139–147 (cit. on pp. 35, 122, 141).

Zeng, Hong, Sheng Wang, Tianming Zhou, Feifeng Zhao, Xiufeng Li, Qing Wu, and Jinbo Xu (2018). "ComplexContact: A web server for inter-protein contact prediction using deep learning". *Nucleic Acids Research* (cit. on pp. 23, 98).

Zhang, Fa, Ting Shi, Huining Ji, Imtiaz Ali, Shuxin Huang, Zixin Deng, Qing Min, Linquan Bai, Yilei Zhao, and Jianting Zheng (2019). "Structural Insights into the Substrate Specificity of Acyltransferases from Salinomycin Polyketide Synthase". *Biochemistry* 58.27, pp. 2978–2986 (cit. on pp. 35, 131, 134).

Zhang, Yang and Jeffrey Skolnick (2005). "TM-align: A protein structure alignment algorithm based on the TM-score". *Nucleic Acids Research* (cit. on p. 6).

Zheng, Jianting and Adrian T. Keatinge-Clay (2013). "The status of type I polyketide synthase ketoreductases". *MedChemComm* 4.1, pp. 34–40 (cit. on pp. 102, 115).

Zheng, Jianting, Shawn K. Piasecki, and Adrian T. Keatinge-Clay (2013). "Structural Studies of an A2-Type Modular Polyketide Synthase Ketoreductase Reveal Features Controlling - Substituent Stereochemistry". *ACS Chemical Biology* 8.9. PMID: 23755878, pp. 1964–1971 (cit. on p. 37).

Zheng, Wei, Chengxin Zhang, Qiqige Wuyun, Robin Pearce, Yang Li, and Yang Zhang (2019). "LOMETS2: improved meta-threading server for fold-recognition and structure-based function annotation for distant-homology proteins". *Nucleic Acids Research* 47.W1, W429–W436 (cit. on p. 7).

Zhou, Xiaogen, Jun Hu, Chengxin Zhang, Guijun Zhang, and Yang Zhang (2019). "Assembling multidomain protein structures through analogous global structural alignments". *Proceedings of the National Academy of Sciences of the United States of America* (cit. on p. 22).

Zhu, Jianwei, Sheng Wang, Dongbo Bu, and Jinbo Xu (2018). "Protein threading using residue co-variation and deep learning". *Bioinformatics* 34.13, pp. i263–i273 (cit. on pp. 14, 143).

Zhu, Shuping and Yihui Liu (2019). "Protein Secondary Structure Online Server Predictive Evaluation". *Journal of Physics: Conference Series* 1237, p. 052005 (cit. on p. 15).

Zimmermann, Lukas, Andrew Stephens, Seung-Zin Nam, David Rau, Jonas Kübler, Marko Lozajic, Felix Gabler, Johannes Söding, Andrei N. Lupas, and Vikram Alva (2018). “A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at its Core”. *Journal of Molecular Biology* 430.15, pp. 2237–2243 (cit. on pp. 10, 11, 43, 159).

Appendix A

Additional information on the methods of "*ab initio* prediction of protein domain structure" study

A.0.1 Feature Vector Generation

As the input for the neural network, a feature vector was generated for all residue pairs of the training and the test set proteins with information extracted and predicted from the target sequence and the alignment containing homologous sequences. Multiple sequence alignments (MSAs) were generated by using HHblits (Remmert *et al.* 2011; Zimmermann *et al.* 2018). Homologous sequences were searched by setting e-value to 0.001, coverage to 60, minimum sequence identity to 0, maximum sequence identity to 90. 4 iterations were performed and maximum 500,000 sequences were allowed to pass to the next iteration. As the database, uniprot20_2016_02 was used.

For the feature vector generation, mutual information with the average product correlation (APC) (Dunn *et al.* 2007), normalized mutual information with APC, CCMPred (Seemayer *et al.* 2014), QUIC (Hsieh *et al.* 2014) and mfDCA (Morcos *et al.* 2011) were used to detect coevolved residue pairs. SPIDER2 (Yang *et al.* 2016) was used to predict the secondary structure and solvent accessibility. Statistical potential (Betancourt and Thirumalai 2008), the effective number of sequences, amino acid composition, and Shannon entropy were calculated by a script from the source code of MetaPSICOV (Jones *et al.* 2015).

For each residue pair in a protein, a feature vector with 733 elements was generated, with features very similar to those used in MetaPSICOV feature vector (Jones *et al.* 2015). For a residue pair (i, j), 13-residue length windows were used for both residues i and j (Fig. A.1). Additionally, a central window with a length of five residues was used centred at the residue position of $(i + j)/2$ (Fig. A.1). Therefore, for each residue pair (i, j), we used information from $(2 * 13 + 5 = 31)$ positions. For these 31 residues, we used secondary structure predictions (predicted likelihood of α-helix, β-sheet and coil), predicted solvent accessibility, entropy and a binary value to indicate whether the corresponding position is within the sequence as for the windows at the beginning and the end of a sequence, non-residue positions were taken into account. In addition to $31 * (3 + 1 + 1 + 1) = 186$ features, coevolution calculations from CCMPred, mfDCA and QUIC were used. These calculations were used as 13 by 13 window where the residue pair (i, j) is located at the centre of the $13 * 13$ window. Therefore, in addition to 186 features, $3 * (13 * 13) = 507$ elements were added to the feature vector. Finally, for each residue (i, j) pair, 40 additional features from mutual information, normalized mutual information, statistical potential, sequence length, number of sequences in the alignment, the effective number of sequences in the alignment, amino acids and gap position frequency in the MSA, an average of predicted likelihood of α-helix, β-sheet, coil and solvent accessibility, site entropies and sequence separation (which is given as eight binary inputs for different separation intervals) were included resulting in 733 elements in the feature vector.

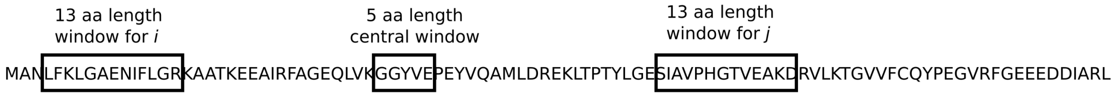

13 aa length window for *i*
5 aa length central window
13 aa length window for *j*
MANLFKLGAENIFLGRKAATKEEAIRFAGEQLVKGGYVEPEYVQAMLDREKLTPTYLGESIAVPHGTVEAKDRVLKTGVVFCQYPEGVRFGEEEDDIARL

***Figure A.1: Residue windows used in feature generation for neural network training.** For each target sequence, two 13-amino acid (aa) lenght windows were used for a residue pair (i,j) centred at positions i and j. One central five-amino acid (aa) length window was used centred at position (i + j)/2.*

First optimization of the feature vector was studied by Shuangxi Ji(Ji 2019), and further improved by Liam Mead(Mead 2018).

A.0.2 Training and Test Sets

As a test set, the test set proteins of MetaPSICOV was selected (Jones *et al.* 2015). From the initial 150 test proteins, the ones having >25% sequence identity with the training set proteins of DeepCDpred, RaptorX(Wang *et al.* 2017) and SPIDER2 (Yang *et al.* 2016) were removed, resulting in 108 test proteins. The sequence length of the 108 proteins varies with a minimum of 56 and a maximum of 242 and their sequence similarity to each other is less than 25%. PDB IDs of 108 proteins are given in Table B.1. Selection of the test set proteins was performed by Shuangxi Ji (Ji 2019).

Proteins used for the training set were selected from precompiled culledPDB lists of the PISCES database (Wang and Dunbrack 2003), downloaded in November 2016. Protein list with less than 25% sequence identity, a maximum resolution of 2 Å, a maximum R value of 0.25 was used to select training proteins. Among them, the ones with maximum 25% sequence identity with the test set and the ones with maximum 400 amino acid length were selected. 1701 of them were arbitrarily selected as the training set. Training set proteins were selected by Shuangxi Ji (Ji 2019).

Table A.1: PDB ID list of the test set with 108 proteins.

1a3aA	1cc8A	1dsxA	1gzcA	1im5A	1ku3A	1p90A	1vjkA
1aapA	1chdA	1eazA	1h2eA	1j3aA	1kw4A	1pchA	1vmbA
1abaA	1cjwA	1ej8A	1h4xA	1jfuA	1lm4A	1qf9A	1vp6A
1ag6A	1ckeA	1f6bA	1hdoA	1jl1A	1lo7A	1qjpA	1w0hA
1aoeA	1ctfA	1fcyA	1hfcA	1jo0A	1m4jA	1r26A	1whiA
1atzA	1cxyA	1fk5A	1hh8A	1jo8A	1m8aA	1roaA	1wjxA
1avsA	1cznA	1fl0A	1htwA	1josA	1mk0A	1rw1A	1wkcA
1bdoA	1d0qA	1fvgA	1hxnA	1jwqA	1mugA	1smxA	1xffA
1bebA	1d1qA	1fx2A	1i1jA	1jyhA	1nb9A	1svyA	2cuaA
1behA	1d4oA	1g2rA	1i1nA	1k6kA	1ne2A	1t8kA	2phyA
1bkrA	1dixA	1g9oA	1i4jA	1k7jA	1npsA	1tifA	1c44A
1dlwA	1gmiA	1i58A	1kq6A	1nrvA	1tqgA	1c52A	1dmgA
1gmxA	1i71A	1kqrA	1ny1A	1tqhA	1c9oA	1dqgA	1gz2A
1iibA	1ktgA	1o1zA	1vfyA				

A.0.3 Neural Network Architecture and Hyperparameters

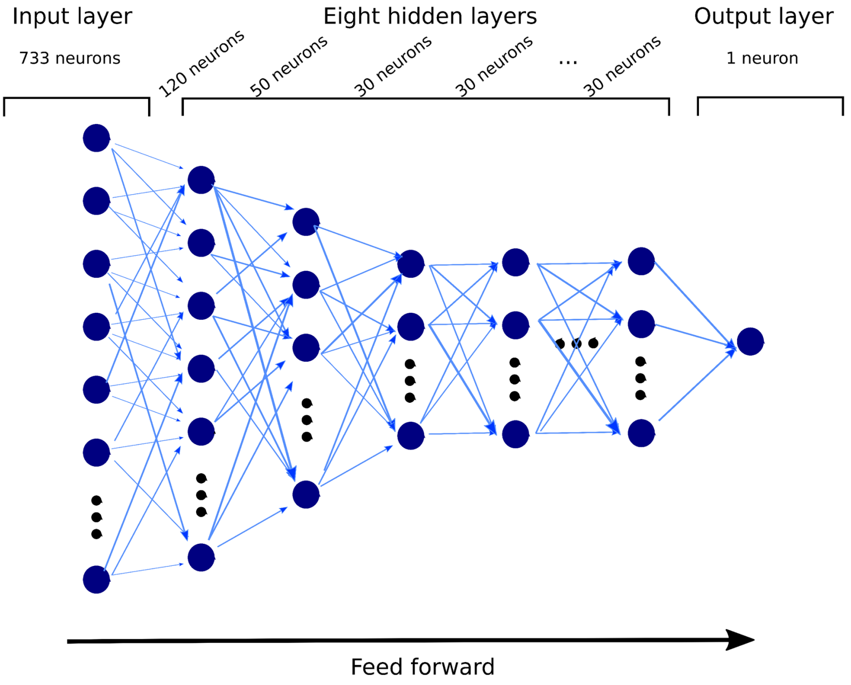


Figure A.2: Neural network architecture used to train models. *The nine-layered neural network consists of one input layer, eight hidden layers and one output layer with varying number of neurons.*

A.0.4 Contact and distance prediction accuracy determination and comparison

Accuracy for top L/10, L/5, L/4, L/3, L/2, L and 1.5L number of residue pairs (where L is the length of the protein sequence) was calculated for the predicted pairs in each bin. It was determined as

$$Accuracy = \frac{Correct\ Predictions}{All\ Predictions} = \frac{True\ Pozitives + False\ Pozitives}{All\ Predictions}. \tag{A.1}$$

Since we did not make true negative and false negative predictions, we calculated accuracy simply by dividing the number of true positives to all predictions.

We compared our contact prediction accuracies with results of MetaPSICOV (Jones *et al.* 2015), RaptorX (Wang *et al.* 2017) and NeBcon (He *et al.* 2017). Predictions were performed by Shuangxi Ji, Liam Mead, Muhammad Fayyaz Rehman and me.

A.0.5 Structure prediction

Rosetta AbInitioRelax (Simons *et al.* 1999) was used for structure prediction. Three-residue and nine-residue fragments were generated by using make_fragments.pl script of Rosetta with -nohoms option in order to exclude homologous structures. As secondary structure predictions, we used the results from SPIDER2. Top 1.5L contact and distance predictions were implemented as bounded function constraints. The parameters for the constraints are shown in Table A.2, which were optimized by Shuangxi Ji (Ji 2019). For each bin, the boundaries of the bounded function were determined based on the regression of the real distance against the neural network score of the predicted pairs over 435 training proteins. For the pairs with lower network score, the bounded function ranges were kept large with lower weights allowing having values outside of the distance bin range. For the implementation of the predictions from RaptorX, top 1.5L contact predictions were used with the same boundary conditions shown in Table A.2.

Since we trained different neural networks for different distance ranges, for some residue pairs there were high scoring predictions in more than one bin. In this case, network scores were compared to decide to select which bin prediction to use. When there are predictions in both contact bin and 8 - 13 Å bin (or 13 - 18 Å bin), if the 8 - 13 Å bin (or 13 - 18 Å bin) score is 0.3 higher than the contact bin score, then 8 - 13 Å bin (or 13 - 18 Å bin) is used; otherwise, the constraint for the residue pair was implemented as a contact bin prediction. If a residue pair was predicted in both 8 - 13 Å bin and 13 - 18 Å bin, then the one with the higher score was selected. 18 - 23 Å bin predictions were only selected if the pair score is 0.5 higher than the predictions in the other bins. This priority order was determined based on the prediction accuracy of the bins, it was not systematically optimized.

Table A.2: *Parameters of the contact and distance constraints.*

Range /Å	DeepCDpred score (s)	Upper boundary	Lower boundary	Standard deviation	Weight
bin 0 - 8	>= 0.9	$-10.8 * s + 16.7$	3.2	0.5	2.5
	>= 0.8 & <0.9			0.7	1.5
	<0.8			1.0	1.0
bin 8 - 13	>= 0.8	$-12 * s + 23.5$	7.5	1	1.5
	<0.8			1.5	0.5
bin 13 - 18	>= 0.8	$-8.6 * s + 25.17$	$8.6 * s + 4.84$	1.5	0.8
	<0.8			1.0	0.3
bin 18 - 23	>= 0.8	$-7.2 * s + 29.2$	$7.2 * s + 11.2$	1.5	0.6
	<0.8			1.0	0.3

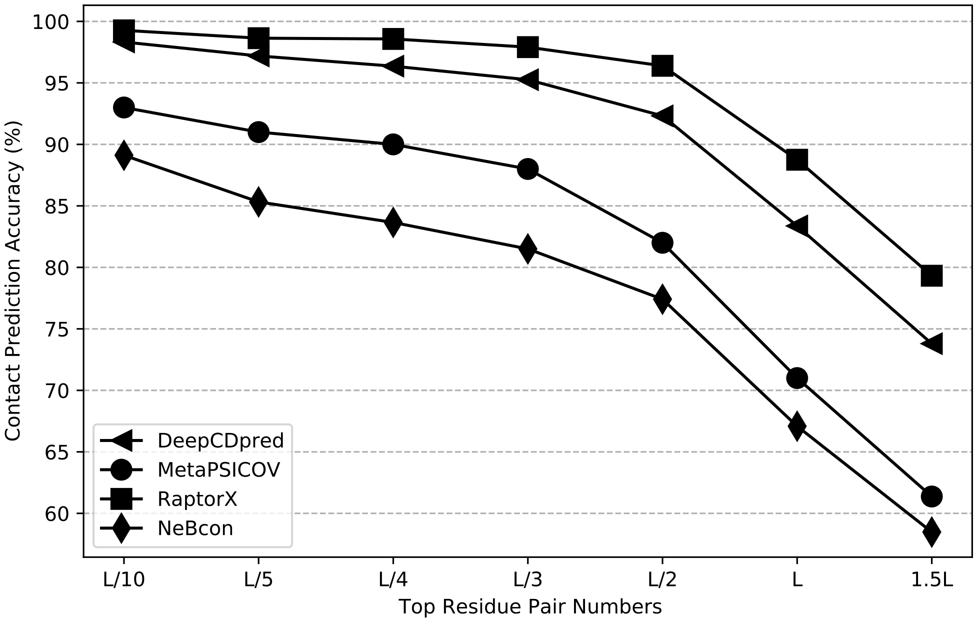


Figure A.3: Contact prediction accuracy of DeepCDpred on 108 test proteins and their comparison with MetaPSICOV, RaptorX and NeBcon results. *DeepCDpred contact predictions outperform predictions of MetaPSICOV and NeBcon for all top L/10, L/5, L/4, L/3, L/2, L and 1.5L residue pairs; however, RaptorX predictions outperform DeepCDpred contact predictions. Predictions were made by Liam Mead and Muhammad Fayyaz Rehman.*

APPENDIX B

LICENCE REFERENCE NUMBERS OF FIGURES FOR PERMISSIONS TO USE

Table B.1: *Licence reference numbers of figures for permissions to use.*

Fig no	Reference no
Fig. 1.3	5025770325530
Fig. 1.4	5025770080424
Fig. 1.11	5025770539654
Fig. 1.14	5016591236267