Extracting Insights Into Protein Structures from Their Sequences

Etai Jacob

The Mina and Everard Goodman
Faculty of Life Sciences

Ph.D. Thesis

Submitted to the Senate of Bar-Ilan University

Ramat-Gan, Israel February 2016

This work was carried out under the supervision of

Prof. Ron Unger
Faculty of Life Sciences
Bar-Ilan University

and

Prof. Amnon Horovitz
Department of Structural Biology
Weizmann Institute of Science

To my father

Acknowledgments

Primarily, I would like to thank my advisors, Ron Unger and Amnon Horovitz, for their guidance, mentoring and invaluable support throughout my studies. Both Ron and Amnon had an enormous impact on my thesis, scientific and personal development. I am grateful that I had the opportunity to learn from their wisdom and experience. I thank Ron for the endless discussions about many ideas and his guidance and help to choose the right directions and make sense out of them. I thank Amnon for his strive for excellence and perfectionism, and teaching me how to communicate and present my thoughts and work both in writing and talks.

Last, I thank my wife Michal, for her unconditional love and endless support, for pushing me forward from the beginning and throughout the years.

Table of Contents

List of Figures

List of Tables

Abbreviations

3

3D - Three dimensional, 8, 57

A

ACO - Absolute Contact Order, 5, 6, 60, 61, 64, 65, 66, 67, 68, 69, 76
APC - Average Product Correction, 20, 21, 29, 32, 36, 37
AUC - Area Under the Curve, 32, 41, 42, 47

B

BLAST - Basic Local Alignment Tool, 7, 79, 80

C

CATH - (C)lass, (A)rchitecture, (T)opology or fold, and (H)omologous superfamily, 57, 58, 61, 64, 65, 66, 67, 68
CDS - Protein Coding Sequences, 17
CMA - Correlated Mutation Analysis, 10, 11, 12, 16, 17, 20, 30, 32, 54
CPU - Central Processing Unit, 22

D

DCA - Direct Coupling Analysis, 5, 12, 19, 22, 23, 24, 28, 32, 35, 36, 37, 38, 39, 41, 42, 43, 44, 45, 46
DI - Direct Information, 12, 19, 22
DNA - Deoxyribonucleic acid, 1, 2, 4, 9, 10, 13, 15, 35, 83, 84, 87

E

EBI - European Bioinformatics Institute, 4
EMBL - European Molecular Biology Laboratory, 4

G

GEO - Gene Expression Omnibus, 5

H

HMM - Hidden Markov Model, 15, 16, 17, 20, 57

M

McBASC - McLachlan-based Substitution Correlation, 11
MI - Mutual Information, 11, 19, 20, 21, 24, 29, 32, 35, 36, 37, 38, 41, 42, 43, 44
mRNA - messenger RNA, 2, 4, 8, 58
MSA - Multiple Sequence Alignment, 7, 8, 10, 11, 15, 16, 17, 20, 22, 23, 25, 27, 28, 29, 32, 35, 39, 41, 42, 44, 45, 46, 47, 50, 57

N

NGS - Next Generation Sequencing, 1, 2, 4
nsSNPs - nonsynonymous Single-nucleotide Polymorphisms, 33

O

OMES - observed-minus-expected-squared, 11, 20, 32, 35, 36, 37, 38, 41, 42, 43, 44

P

PDB - Protein Data Bank, 5, 9, 17, 30, 45, 57, 61, 78, 88, 2
PSICOV - Protein Sparse Inverse COVariance, 12, 22, 28, 29, 32, 35, 36, 37, 38, 41, 42, 43, 44, 83

R

RNA - Ribonucleic acid, 2, 4, 9, 10, 15, 35, 52, 53, 54, 84, 85, 88
ROC - Receiver Operating Characteristic, 47, 48, 49

S

SCA - Statistical Coupling Analysis, 11
SCOP - Structural Classification of Proteins, 57, 58, 61, 64, 65, 66, 67, 68, 77
SNP - Single-nucleotide Polymorphism, 5, 33, 34, 48, 49, 50, 51, 54
SRA - Sequence Read Archive, 2, 4

T

TCGA - The Cancer Genome Atlas, 4

W

WWW - World Wide Web, 4

Abstract

Amino acid and nucleotide sequences constitute a rich source of information that can be used to address a wide range of biological questions. The enormous amount of biological data that are rapidly accumulating from sequencing efforts, on the one hand, and from other types of experiments (e.g. three dimensional structure determination) on the other hand, are creating new opportunities to correlate protein sequences with their structure and function. Nevertheless, while the number of sequenced genomes continues to grow exponentially, other types of experiments have not kept pace. For instance, despite the great progress in experimental determination of protein three-dimensional structures, we know many more protein sequences than protein three-dimensional structures, and the gap is getting bigger. Thus, many bioinformatics applications which predict the properties of proteins or genes based on sequence data alone, were developed during the last three decades in order to bridge this gap. Nevertheless, the success of many of these prediction methods is limited, but their results are encouraging since they enable the discovery of knowledge that is difficult to obtain by experiments. Thus, despite many years of sequence analysis in biology, extracting biological insights from sequences alone is still a challenging task. In my thesis, I describe two studies in which I addressed this challenge.

First, I describe how codon-level information improves predictions of inter-residue contacts in proteins by correlated mutation analysis. Genomic sequences contain rich evolutionary information about functional and structural constraints on proteins. This information can be mined to detect correlated mutations in proteins and address the long-standing challenge of predicting protein three-dimensional structures from amino acid sequences. Methods for analysing correlated mutations in proteins are becoming an increasingly powerful tool for predicting contacts within and between proteins owing to the explosive growth in sequence data and significant theoretical progress. Nevertheless, limitations remain due to the requirement for large multiple sequence alignments (MSA) and the fact that, in general, only the relatively small number of top-ranking predictions are reliable. To date, methods for analysing correlated mutations have relied exclusively on amino acid MSAs as inputs. In my thesis, I describe a new approach for analysing correlated mutations that is based on combined analysis of amino acid and codon MSAs. I show that a direct contact is more likely to be present when the correlation between the positions is strong at the amino acid level but weak at the

codon level. The performance of different methods for analysing correlated mutations in predicting contacts is shown to be enhanced significantly when amino acid and codon data are combined.

In the second study, I revealed a strong tendency in all kingdoms of life for N-terminal domains in two-domain proteins to have shorter sequences than their neighboring C-terminal domains. Given that folding rates are affected by chain length, I asked whether the tendency for N-terminal domains to be shorter than their neighboring C-terminal domains reflects selection for faster folding N-terminal domains. Calculations of absolute contact order, another predictor of folding rate, provided additional evidence that N-terminal domains tend to fold faster than their C-terminal neighboring domains. A possible explanation for this bias is that faster folding of N-terminal domains reduces the risk of protein aggregation during folding by preventing formation of non-native interdomain interactions. This explanation is supported by protein expression analyses I performed which demonstrated that two-domain proteins with a shorter N-terminal domain are much more abundant than those with a shorter C-terminal domain. These findings, therefore, suggest a previously unrecognized mechanism for prevention of aggregation of neighboring domains in multi-domain proteins.

The first study of this thesis was published in *eLIFE*:

Jacob, E., Unger, R. and Horovitz, A. (2015). Codon-level information improves predictions of inter-residue contacts in proteins by correlated mutation analysis. eLife 2015;4:e08932.

The second study of this thesis was published in *Cell Reports*:

Jacob, E., Unger, R. and Horovitz, A. (2013). N-Terminal Domains in Two-Domain Proteins Are Biased to Be Shorter and Predicted to Fold Faster Than Their C-Terminal Counterparts. Cell Rep. 3 (4), 1051-1056.

Chapter 1

Introduction

1.1 Sequencing Methods

The determination of the first complete amino acid sequence of a protein by Sanger in 1955 showed that a protein has a unique amino acid sequence. Before that, it had been only known that different proteins had different amino acid compositions and the common assumption was that molecules of the same proteins are not identical to each other. Sequencing projects during the 1950's were a difficult manual process that consumed a lot of time. For example, the determination of the complete amino acid sequence of insulin (including 2 chains and disulfide bonds) by Sanger was an iterative process that lasted from 1945 to 1955 and led to approximately ten stand-alone publications describing each step separately (Stretton, 2002). In the following decade, manual sequencing processes were gradually improved and, consequently, the rate of sequence determination increased. By the mid 1960's, with the determination of the complete amino acid sequences of other proteins including ribonuclease by Anfinsen, there were a total of 65 known sequences (Table 1.1). By contrast with the advances in protein sequence determination technologies, sequencing nucleic acids had remained problematic mostly because of difficulties in purification and sequencing of long molecular fragments (less than ~500 bp). In 1977, however, Sanger introduced a DNA sequencing method (Sanger et al., 1977) that made it possible to sequence longer nucleotide fragments. His method became known as "Sanger sequencing" or "first-generation sequencing".

In the mid-1980's, a significant increase in productivity was made possible due to the automation of Sanger sequencing techniques. This advance led to a dramatic growth in the number of determined sequences (Figure 1.1) and laid the foundations for the sequencing of the first human genome. Eventually, in 2003, after almost 15 years of efforts around the world, the sequencing of the first human genome was declared completed (Table 1.1). Since then, much faster and cheaper sequencing methods have been developed that are known today as "Next Generation Sequencing" methods (e.g. The Illumina MiSeq and LifeTechnologies Ion Torrent Personal Genome Machine (PGM)) and have made sequencing accessible to more labs. These next-generation sequencing (NGS) methods are

based on massive parallel sequencing technology. In this technology, millions of fragments of nucleotides from a single sample are sequenced simultaneously, allowing an entire genome to be sequenced in less than one day. The easy accessibility and the short experimental time of these methods enabled the rapid increase in the amount of research being performed with nucleic acid sequencing. Consequently, the number of sequences deposited in public databases has been growing exponentially. Importantly, NGS has also become a platform to invent new research tools that are sequence-census based (Wold and Myers, 2007). For example, NGS is used to find transcription factor binding sites using ChIP-seq technology (Johnson et al., 2007), discover methylation patterns across the genome using Methyl-seq, measure mRNA expression using mRNA-seq, reveal folding principles of the human genome (Lieberman-Aiden et al., 2009) and much more.

Year	Protein	RNA	DNA	No. of residues
1935	Insulin			1
1945	Insulin			2
1947	Gramicidin S			5
1949	Insulin			9
1955	Insulin			51
1960	Ribonuclease			120
1965		tRNAAla		75
1967		5S RNA		120
1968			Bacteriophage λ	12
1977			Bacteriophage ϕX 174	174 5,375
1978			Bacteriophage ϕX 174	174 5,386
1981			Mitochondria	16,569
1982			Bacteriophage λ	48,502
1984			Epstein-Barr virus	172,282
2004*			*Homo sapiens*	2.85 billion
2009		Total base pairs in NCBI Sequence Read Archive (SRA)		1.35E+13
2015		Total base pairs in NCBI Sequence Read Archive (SRA)		2.75E+15

Table 1.1 Sequencing landmarks.
Table is based on the work of others (Attwood et al., 2011). * Completion of the human genome was already declared in 2003.

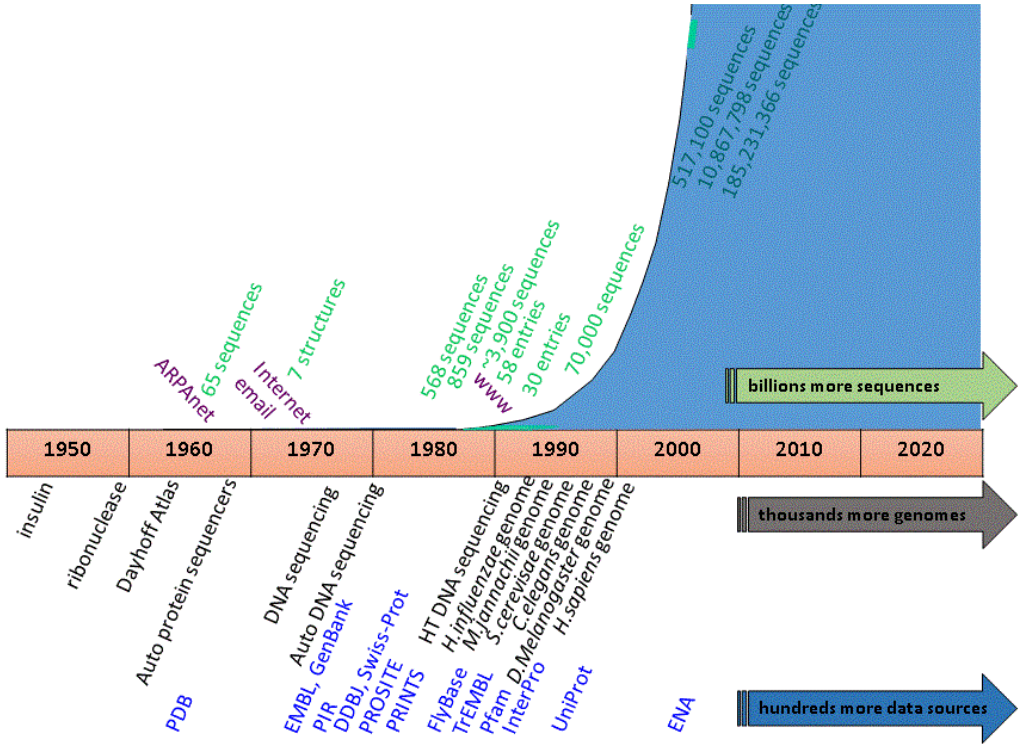


Figure 1.1. Historical milestones in bioinformatics.
Figure adopted from the work of others (Attwood et al., 2011) and a presentation by Teresa K. Attwood (with some modifications).

1.2 Databases

The increase in the number of known protein sequences prompted Margaret Dayhoff and co-workers (Dayhoff et. al, 1965) to organize the first computerized collection of protein sequences that initially comprised 65 sequences (the collection was called "Atlas of Protein Sequence and Structure"). Dayhoff and her colleagues understood that tremendous amounts of information about the evolutionary history and function are contained within each sequence.

Advances in nucleotide sequencing brought about a need for organization and analysis also of DNA and RNA sequences (Gingeras and Roberts, 1980). Several nucleotide databases were, therefore, established. In 1982, the European Molecular Biology Laboratory (EMBL) in Heidelberg released 568 sequences and GenBank, which was established in December that year, brought 606 sequences to the public domain (Figure 1.1; table 1.1). The world of protein sequences, which to a certain extent was overshadowed by the efforts to collect nucleotide sequences, continued to grow and reached a size of more than 1660 amino acid sequences (Figure 1.1). The late 1980's and early 1990's, just before the WWW emerged, were characterized by intense activity that gave rise to new sequences, databases (e.g. Swissprot in 1986), characterization of protein families (e.g. PROSITE), and data maintenance and support organizations (e.g. The European Bioinformatics Institute (EBI)). By 1995, sequencing technologies made whole genome sequencing feasible and the genomes of several organisms were sequenced including the first human genome . In the 2000's, Next Generation Sequencing (NGS) technologies led to enormous amounts of sequence data that were collected in several public archives such as NCBI Sequence Read Archive (SRA) and European Nucleotide Archive (ENA). As an example for the dramatic increase in the amount of data produced, SRA consisted in January 2009 of 1.35E+13 nucleotide bases and in November 2015 was already storing 2.75E+15 nucleotide bases. In addition, the significant changes in speed, costs and flexibility of NGS made it to become a central method that is used in international collaborations projects. An example of such a project is The Cancer Genome Atlas (TCGA), which was initiated in 2005 as an organized effort to accelerate our understanding of the molecular basis of cancer. Today, TCGA consists of many types of data (e.g. DNA and mRNA sequencing, protein expression, copy number and DNA methylation) for thousands of samples.

Although sequences provide a substantial source of information and have drawn a lot of attention over the years, valuable data of other types were also accumulating. One of

the main depositories that began to increase dramatically at the end of the 1980's was the protein data bank (PDB). The improvements in crystallization methods, the newly established molecular biology capabilities to clone genes and express proteins, and the technological advancements in computer software and X-ray detection methods, made it possible to make substantial progress in protein structural determination (Berman, 2008). Again, as in the case of biological sequences, the rapid development in all aspects of the experimental procedures resulted in the dramatic growth in the number of solved structures collected in the PDB. In addition to structure determination, numerous other experimental methods were invented and improved during the years (e.g. single cell analysis, Immunohistochemistry, methods to investigate protein-protein interactions). One example is the development of microarray technologies to measure gene expression during the 90's (Lenoir and Giannella, 2006). These technologies became widely used and have produced much information on gene expression in humans and other organisms under different conditions. Today, the gene expression omnibus at NCBI (GEO) maintains enormous amounts of microarray experiments data deposited by different laboratories across the world.

1.3 Sequence Analysis

The first sequence alignments, which were carried out for insulin, ribonuclease and a few other proteins, were based on a small number of homologous sequences from several species. Comparing two amino acid or nucleotide sequences was one of the first types of analysis that were required when sequencing data became available. In 1970, Gibbs and McIntyre described a simple method for comparing sequences that is called the dot matrix or diagram (Gibbs and Mcintyre, 1970). In this method, the two compared sequences are written along adjacent sides of a rectangular matrix with their N-terminal amino acids in the top left-hand corner of the diagram. Within the matrix, a dot is plotted whenever a row and a column share the same amino acid. Similarities of the two sequences are then indicated by a diagonal line of dots. (Figure 1.2). This method is also able to reveal insertions, deletions, and repeats when the same sequence is used in the horizontal and vertical axis of the dot matrix.

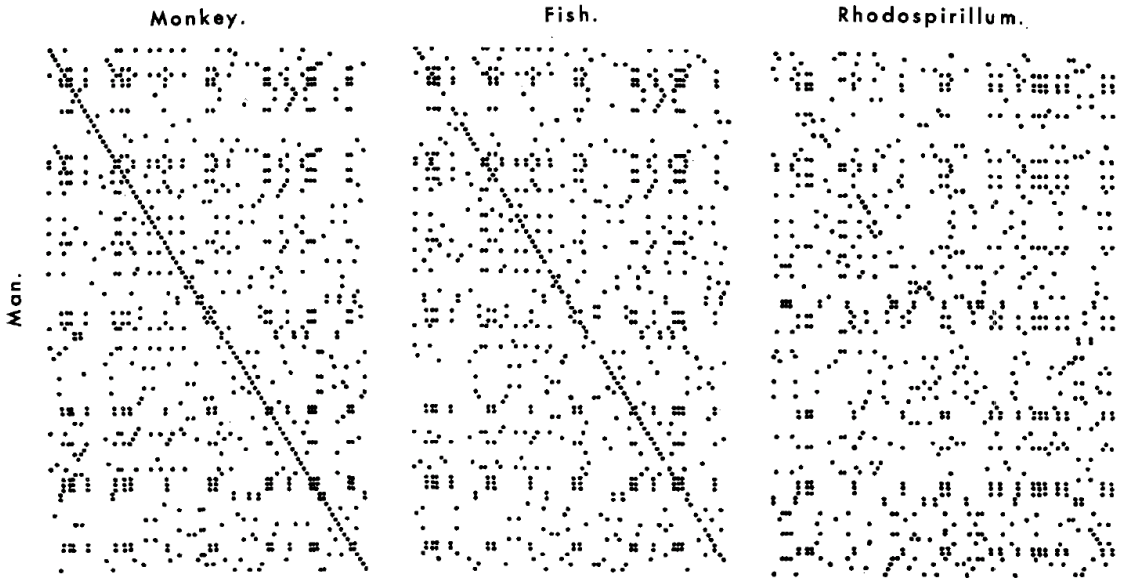


Figure 1.2. A dot matrix (diagram) obtained by comparing the human cytochrome c (Y-axis, N-terminal at the top) and the cytochrome c of monkey, fish and Rhodospirillum.
The figure is taken from previous work of others (Gibbs and Mcintyre, 1970).

Nevertheless, visual comparisons of sequences were tedious and involved subjective assessments and, thus, computer-based statistical approaches were required. Indeed, in the same year that the dot matrix was introduced, Needleman and Wunsch (Needleman and Wunsch, 1970) presented their dynamic programing approach for global sequence alignment. In their work, they described an iterative matrix procedure to find the maximum match between two sequences, that is, the largest number of amino acids or nucleotides in one sequence that can be matched with those of another sequence, while allowing for all possible deletions (Needleman and Wunsch, 1970). In their method, the problem is broken down into the smallest unit of comparison, a pair of amino acids. The alignment is built progressively by starting at the C-terminal end of each sequence and then moving ahead one amino acid pair at a time, allowing for various combinations of matched pairs, mismatched pairs, or insertion/deletion of amino acids in one sequence. This process results in every possible alignment between the two sequences and by using a scoring system which prioritizes a match over a mismatch and penalizes gaps, the alignment with the highest possible score was defined as the optimal alignment. An important modification to their algorithm was the local sequence alignment method introduced by Smith and Waterman (Smith and Waterman, 1981) in 1981. They recognized that the most biologically significant regions in sequences were the segments that aligned well and not the other less related regions that were not well aligned. Smith and Waterman extended Needleman's and Wunsch's idea to find a pair of sub-regions, one from each of two long sequences, such that there is no other pair of segments with greater homology (Smith and Waterman, 1981).

Given that the local and global sequence alignment methods required a great deal of time resources those days, a method that can perform a database scan for similarity in a short time was highly needed. In 1988, Pearson and Lipman developed a program called FASTA (Pearson and Lipman, 1988), which provided a rapid way to perform such similarity scans. Two years later, a faster method for similarity search was introduced by Altschul et al. (Altschul et al., 1990). basic local alignment tool, known as BLAST, has been and is still a widely used sequence analyses program.

Comparing more than two sequences simultaneously (i.e. multiple sequence alignment) required the design of new tools since dynamic programing implementations were too computationally demanding. Thus, several improvements were introduced (Johnson and Doolittle, 1986; Lipman et al., 1989), including the development of the commonly used multiple sequence alignment (MSA) software tool called CLUSTALW

(Thompson et al., 1994). Since its first implementation, MSA became an increasingly important tool in biology and has been used in molecular evolution to construct phylogenetic trees (Felsenstein, 1989; Hogeweg and Hesper, 1984), identify distantly related sequences of a protein family based on conserved regions (Gribskov et al., 1987), predict functionally or structurally important residues (e.g. Casari et al., 1995; Karlin and Brocchieri, 1996) and much more.

1.4 Recent Progress and Future Perspectives

As influx of biological data from different sources became a routine, efforts were made to improve the use of many biological resources (e.g. the 2015 Nucleic Acids Research (NAR) database summary paper reported over 1800 valid databases (Galperin et al., 2015)). In addition, attempts to link related bioinformatics databases together and enhance biological annotations, enabled an efficient retrieval of gene or protein related information from diverse resources (e.g. mRNA expression, structural and functional data). Increasingly, more nucleotide and amino acid sequences are linked to information from other sources, such as 3D structures, protein and mRNA expression. Nevertheless, while the number of sequenced genomes continues to grow exponentially, other types of experiments have not kept pace. For instance, despite the great progress in experimental determination of protein three-dimensional structures, we know many more protein sequences than protein three-dimensional structures, and the gap is getting bigger. Indeed, in order to bridge this gap, many bioinformatics applications which predict the properties of proteins or genes (as their three dimensional structures or functions) based on sequence data alone, were developed during the last three decades (Table 1.2). Many of these methods combine information from diverse sources in biology along with amino acid or nucleotide sequence information. Furthermore, although the success of many of these prediction methods is limited, their results are encouraging since they enable the discovery of knowledge that is difficult to obtain by experiments. Thus, with this theoretical progress, the exceptional advances of sequencing technologies, and the increase in the amount and availability of diverse data sources in biology it is clear that many insights remain to be obtained through analysis of protein sequences.

Computational method	Types of information used in analyses	The purpose of the tool	Examples of publication/Application name
Neural networks	MSAs and secondary structure data from the PDB	Predict protein secondary structure from sequence alone	(Rost and Sander, 1993) (Jones, 1999), PSIPRED
Neural network	Sequence and structural data of PDB complexes	Predict protein-protein interactions interfaces	(Ofran and Rost, 2003a, 2003b)
Bayesian framework	Combines protein-protein interactions data (e.g. Y2H) ,structural, functional, evolutionary and expression information	predicts whether a pair of proteins interact	(Zhang et al., 2012), PrePPI
Statistical model	ChIP-seq experiments and nucleotide sequences	Identifying transcription factor binding sites	(Wang et al., 2012a)
Statistical models	MSAs and structural data	Predict the three dimensional structure of proteins	(Balakrishnan et al., 2011; Hopf et al., 2012; Jones et al., 2015; Morcos et al., 2011)
Empirical Bayesian or ML algorithms	MSAs, derived phylogenetic tree and protein structure	estimating the evolutionary conservation of amino/nucleic acid positions in a protein/DNA/RNA molecule	(Glaser et al., 2003), ConSurf server

Table 1.2 Examples of applications that combine amino acid or nucleotide sequence and other types of information in their analysis.

Chapter 2

Incorporation of codon data in correlated mutation analysis

2.1 Introduction

The explosive growth in sequence data from current high-throughput techniques enables analysis of functional interaction patterns at the DNA or RNA level (e.g. RNA folding), cellular level (e.g. regulation and organization, interactions between proteins) and the amino acid residue level (e.g. protein contact prediction). In particular, genomic sequences contain rich evolutionary information about functional and structural constraints on proteins. For example, many computational methods for predicting protein three-dimensional structures were developed over the years for homology modeling, i.e. predicting structures using known three-dimensional structures with sequences that are similar to that of the protein of interest. Such structures are, however, not always available and evolutionary information found in patterns of correlated mutations in protein sequences can then play a major role in predicting the 3D structure of a protein (de Juan et al., 2013; Marks et al., 2012). Correlated mutations can arise since the effects of mutations which disrupt protein structure and/or function at one site are often suppressed by mutations that occur at another site (either in the same protein or in another protein). Such compensatory mutations can occur at positions that are distant from each other in space, thus, reflecting long-range interactions in proteins (Horovitz et al., 1994; Lee et al., 2008). It has often been assumed, however, that most compensatory mutations occur at positions that are close in space. This has motivated the development of computational methods for identifying co-evolving positions that can be used as distance constraints in protein structure prediction (Göbel et al., 1994) (Figure 2.1).

Methods of CMA consist of the following steps: first, a multiple sequence alignment (MSA) for the members of an evolutionary related family of proteins is created. Next, the frequencies of co-occurrence of all amino acids in all pairs of columns are calculated and compared to those expected assuming that the frequencies of occurrence at

one position are independent of those at the second position. Finally, the correlations are ranked according to the statistical and/or physical significance attached to them.

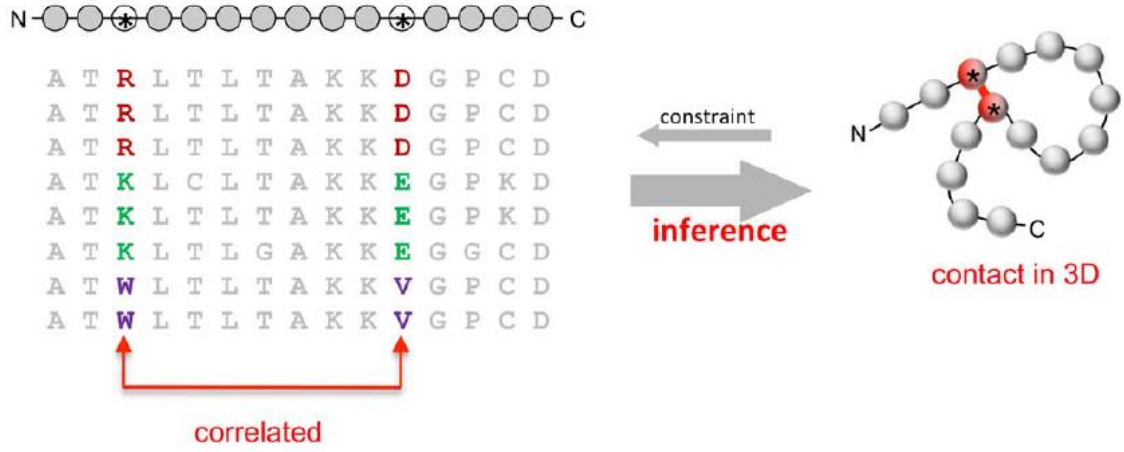


Figure 2.1. Identifying co-evolving positions as distance constraints in protein structure prediction.
The sequence of the protein (chain of gray circles) for which a 3D structure is to be predicted is a member of a family of evolutionary related sequences (the MSA in light gray letters). The evolutionary variation in the sequences (colored columns in the MSA) is restricted by a direct physical contact (red circles on the right). This figure is taken from previous work of others (Marks et al., 2011).

Various methods for CMA that have been developed in the past 15 years differ in the measures that they employ for attaching significance to the correlations (de Juan et al., 2013; Livesay et al., 2012; Mao et al., 2015). Early measures that were developed assume that pairs of residue positions are statistically independent of residues at other positions (Dekker et al., 2004; Göbel et al., 1994; Kass and Horovitz, 2002; Lockless et al., 1999; Martin et al., 2005). Such methods include, for example, mutual information (MI) from information theory (Gloor et al., 2005), observed-minus-expected-squared (OMES) in the chi-square test (Kass and Horovitz, 2002), statistical coupling analysis (SCA) (Lockless et al., 1999) and the McLachlan-based substitution correlation (McBASC) (Olmea et al., 1999)

Statistically significant correlations in MSAs that do not reflect interactions between residues in contact, i.e. false positives, can stem from (i) various indirect physical interactions and (ii) common ancestry. The extent of false positives due to the latter source is manifested in the large number of correlations between positions in non-interacting proteins that can be observed when the sequences of non-interacting proteins from the same organism are concatenated and subjected to CMA (Noivirt et al., 2005). Several approaches for removing false positives owing to common ancestry were developed (Dunn

et al., 2008; Noivirt et al., 2005; Pollock et al., 1999; Wollenberg and Atchley, 2000) on the basis of the early methods but their success in contact prediction remained limited. False positives due to the former source, i.e. indirect physical interactions, can occur when, for example, correlations corresponding to positions A and B that are in contact and positions B and C that are in contact lead to a correlation for positions A and C that are not in contact. Methods that remove such transitive correlations have been developed in recent years. These methods, in contrast with the earlier ones, consider correlated pairs of residues as being dependent on all other positions, thereby reducing the effect of noise due to transitivity. Examples of such methods include Direct Coupling Analysis (DCA or DI for Direct Information) (Morcos et al., 2011; Weigt et al., 2009), Protein Sparse Inverse COVariance (PSICOV) (Jones et al., 2012) and Gremlin's pseudo-likelihood method (Kamisetty et al., 2013). These methods have been found to be very successful in identifying contacting residues (Marks et al., 2012; Stein et al., 2015) and they outperform earlier methods (Mao et al., 2015). Nevertheless, their accuracy, which is ~80% for the correlations in the top 0.1% (ranked by their scores), drops to ~50% for the top 1% (Mao et al., 2015). Given that the number of contacts in a protein with N residues is ~N (Faure et al., 2008), it follows that for proteins with, for example, 100 residues (i.e. with 4,560 potential contacts between residues separated by at least 5 residues in the sequence) only about 25% of the contacts (i.e. 23 of the top 1% 46 predictions) will be identified by these CMA methods. In addition, these methods require large MSAs comprising thousands of sequences in order to perform well and such sequence data are not always available. Consequently, it is clear that much can be gained from further improvements in methods of CMA.

In this thesis, I describe a new approach for analysing correlated mutations that is based on combined analysis of amino acid and codon MSAs. I show that a direct contact is more likely to be present when the correlation between the positions is strong at the amino acid level but weak at the codon level (Figure 2.2). The performance of different methods for analysing correlated mutations in predicting contacts was found to be enhanced significantly when amino acid and codon data are combined.

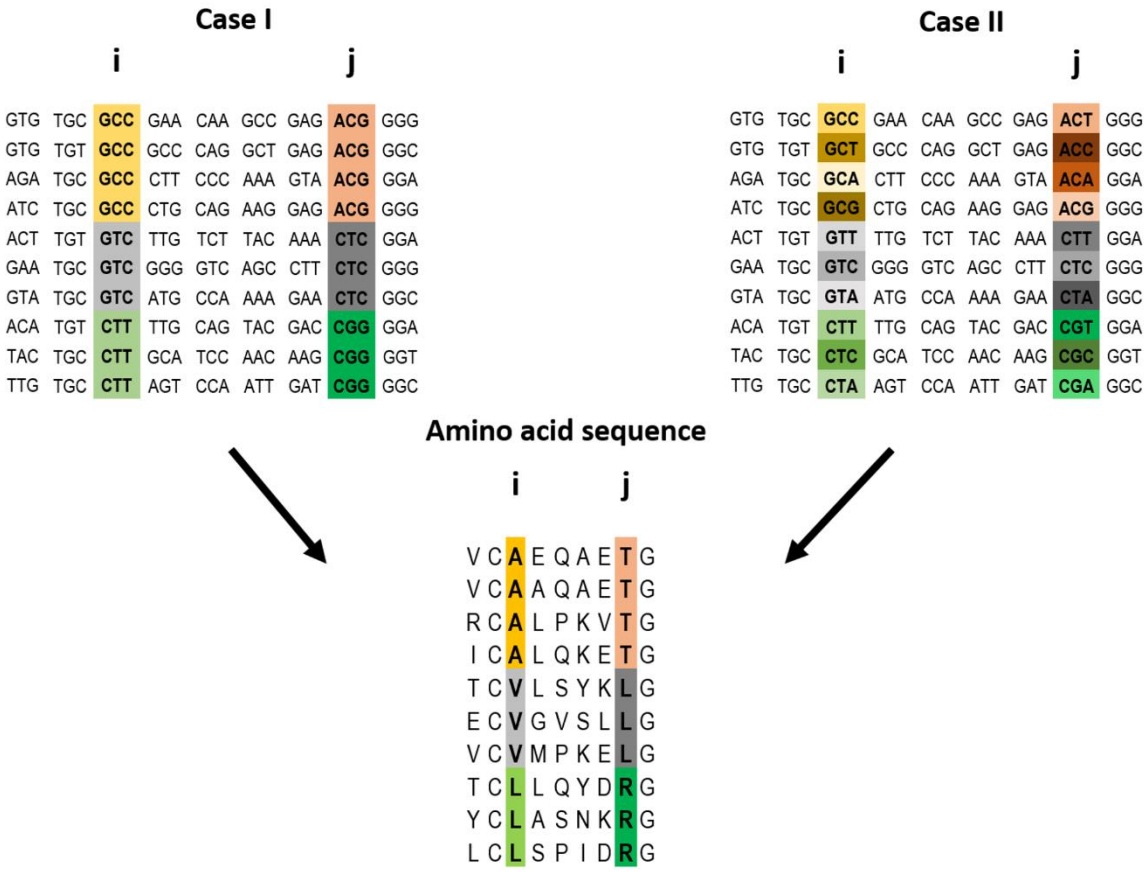


Figure 2.2 Example of a pairwise correlation in a multiple amino acid sequence alignment and two possible corresponding codon alignments.
A correlation at the amino acid level between two positions i and j may (top left) or may not (top right) be accompanied by a correlation at the codon level. The premise of the method introduced in my thesis is that a correlation at the amino acid level between two positions is more likely to reflect a direct interaction if the correlation at the codon level for these positions is weak (top right).

2.2 Methods

2.2.1 Collection of sequences

The growing availability of sequences of sufficient diversity as a result of advances in DNA sequencing technologies over the past decade (Figure 2.3) enabled the significant progress in protein structure prediction based on evolutionary information. Uniprot/TrEMBL currently consists of more than 571,000 species, with a strong bias towards several heavily sequenced species (6.6% of the whole database corresponds to 20 species that comprise

only 0.0035% of the total number of species). The most prominent source of sequences is bacteria that account for more than 60% of the sequences in Uniprot/TrEMBL (Figure 4).

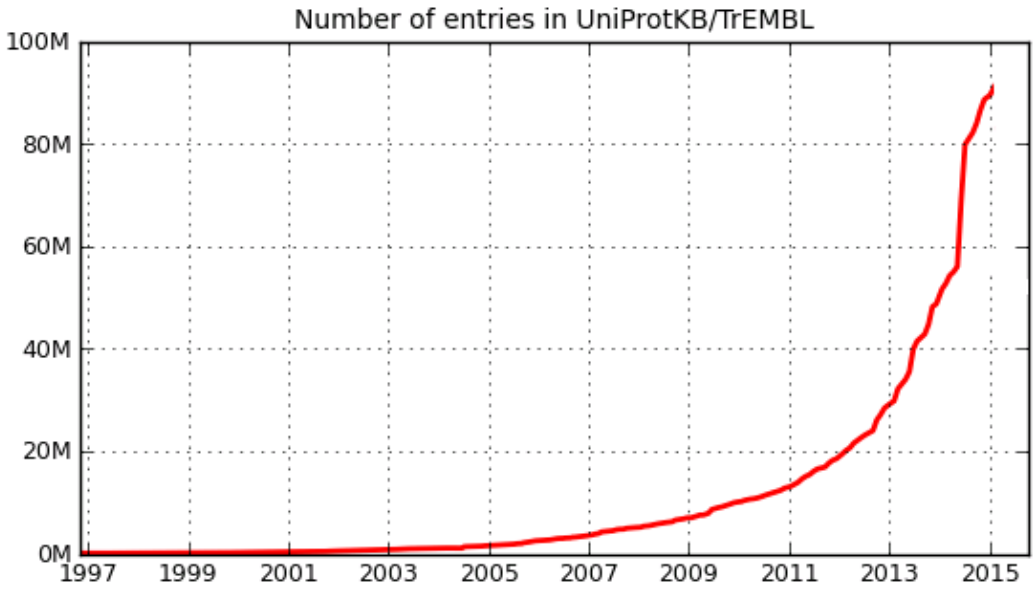


Figure 2.3 Growth of Uniprot/TrEMBL in the last ~20 years.
Based onRelease 2015_10 of 14-Oct-2015 of UniProtKB/TrEMBL. The figure is taken from Uniprot/TrEMBL statistics (http://www.ebi.ac.uk/uniprot/TrEMBLstats).

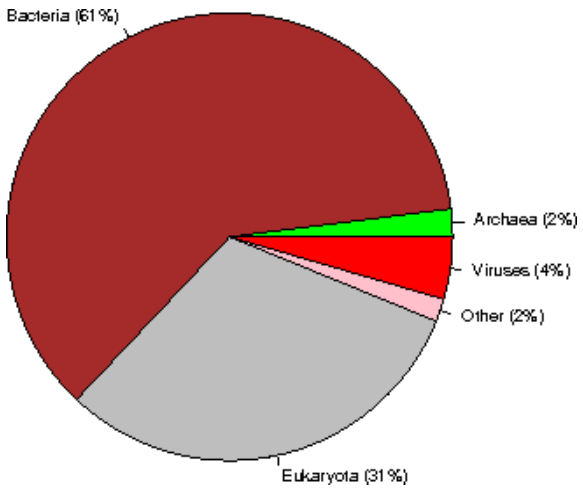


Figure 2.4 Taxonomic distribution of sequences.
Based onRelease 2015_10 of 14-Oct-2015 of UniProtKB/TrEMBL. The figure is taken from Uniprot/TrEMBL statistics (http://www.ebi.ac.uk/uniprot/TrEMBLstats).

2.2.2 Multiple sequence alignments

2.2.2.1 Overview of MSAs

Multiple sequence alignments (MSA) have been used in a wide range of bioinformatic applications. MSAs are used to infer functional sites (Glaser et al., 2003; Pupko et al., 2002), predict protein structure (de Juan et al., 2013; Marks et al., 2012), classify proteins into families using sequence phylogenetics (Casari et al., 1995; Pethica et al., 2012) and more. Although MSAs can be comprised of DNA, RNA or amino acid sequences, it is most common to align proteins using only their amino acid sequences. Such an analysis has advantages and disadvantages. For example, aligning nucleotide sequences instead of amino acid sequences introduces frame shifts but, on the other hand, closely related sequences can be more easily distinguished. The problem of frame shifts can be solved with the use of codons instead of nucleotides. Nevertheless, for more distantly related species, amino acid sequences may have the advantage of acting as a filter to reduce noise. Multiple amino acid sequence alignments have, to date, been the exclusive input for methods for analysing correlated mutations.

2.2.2.2 MSAs and correlated mutation analysis

MSAs consist of sequences that share an evolutionary relationship (i.e. homologs). These sequences are usually collected using a sequence-based search method (Altschul et al., 1990, 1997; Camacho et al., 2009), profile HMM (Eddy, 1998; Finn et al., 2011) or other methods and then aligned using a multiple sequence alignment tool (Edgar, 2004; Katoh et al., 2002; Notredame et al., 2000). The size of the MSA (i.e. the number of sequences that it includes) depends on the level of sequence similarity between its members (e.g. an alignment score), search method (e.g. blast or profile HMM as HMMER3) and types of homology considered (e.g. sequences of orthologs only or of both paralogs and orthologs). Other factors that influence the MSA quality and size are the databases that were used to search for homologous sequences, the level of redundancy that was used for filtering and more. The general rule of thumb is that larger MSAs contain more information. Nevertheless, the tradeoff is that including more sequences can result in adding bias or noise (e.g. bias towards a subset of related sequences in the MSA or certain species), low quality alignment and inclusion of proteins with unrelated function. The relevance of these factors needs to be considered in accordance with the application. In this thesis, in order

to correct the bias for certain species, a resampling technique is used before constructing an MSA. The sequences for each MSA of a protein family are collected from representative proteomes, i.e. proteomes each of which represents best a group of proteomes with similar sequences (Chen et al., 2011). In this way, over- and under-represented species contribute equally to the analysis. In addition, several methods for contact prediction described in this thesis include a reweighting procedure as a correction for this and other biases (see the section Regularization and reweighting of frequency counts).

In CMA, when one wishes to predict contacts within a protein family (i.e. intra-chain interactions) with high accuracy as done here, it is advantageous to use as much information as possible, i.e. sequences of distant homologs and both orthologs and paralogs, as long as the bias towards a subset of related sequences in the MSA or certain species is limited. On the other hand, when the goal is to predict contacts between two different proteins (i.e. inter-chain interactions), the MSAs usually comprise fewer sequences since only concatenated sequences of interacting proteins from the same organism can be included (Hopf et al., 2014). In this thesis, each MSA is built from sequences that are from the same protein family. A protein family is a group of proteins that are closely related with respect to their function, structure or evolution. Our analysis was done as before (Marks et al., 2011; Morcos et al., 2011) using MSAs from Pfam families. The Pfam database is a large collection of protein families (more than 14,000 different families) each of which is described by an MSA and a hidden Markov model profile (HMM) (Finn et al., 2014; Punta et al., 2012). HMM is a probabilistic model (Rabiner, 1989) used for the inference of a homology structure from a set of aligned family representative sequences (Eddy, 1998; Eddy and Wheeler, 2013; Krogh et al., 1994). A high-quality seed alignment is used to construct the profile HMM of a domain family with which it is then possible to search any large sequence database (e.g. UniprotKB) for all instances corresponding to a particular domain family. In this way, a large database of MSAs, representing domain families with diverse structures and lengths, was used for the analysis. Note that in most cases, each MSA contains many sequences with known structures, thereby helping to assess the reliability of contact prediction.

2.2.2.3 Generation of codon and amino acid MSAs

Protein sequence datasets were collected from Pfam version 27.0 (Finn et al., 2014) based on representative proteomes (Chen et al., 2011) at 75% co-membership threshold (RP75) in order to avoid overrepresentation of certain species. Protein coding sequences (CDS) of the collected proteins from Pfam were retrieved based on Uniprot cross reference annotations (for Refseq, Ensembl, EMBL and Ensembelgenomes databases in that order of priority) (Cunningham et al., 2014; Kanz et al., 2005; Pruitt et al., 2012) using the EMBL-EBI's WSDbfetch services (McWilliam et al., 2009) and Ensembl REST API (Beta version) (Yates et al., 2015). All collected CDSs were aligned in accordance to the Pfam HMM-based MSAs using tranalign tool from the EMBOSS package (Rice et al., 2000). Pfam domain families with more than 2,000 successfully retrieved coding sequences were used for further analysis (total of 551 MSA's). Only families with a known crystal structure at a resolution of 3 Å or better (more than 95% of the families have at least three such structures) and with an overlap of at least 80% of the domain sequence to the ATOMs sequence in the solved structure were included in the analysis (total of 460 MSA's). Our analysis was also restricted to proteins with more than 200 residues that have a large number of potential contacts for prediction (114 MSA's). PDB structures were assigned to Pfam families in accordance to the mapping in the files downloaded from http://www.rcsb.org/pdb/rest/hmmer?file=hmmer_pdb_all.txt and ftp://ftp.ebi.ac.uk/pub/databases/msd/sifts/text/pdb_chain_uniprot.lst. PDB structures were retrieved and their coordinates were extracted using the bio3D R package (Grant et al., 2006). Pairwise sequence alignments for mapping were performed using Biostrings [Pages H, Aboyoun P, Gentleman R and DebRoy S. Biostrings: String objects representing biological sequences, and matching algorithms. R package version 2.34.1.].

2.2.3 Methods for analysing correlated mutations

Early methods for CMA relied on the assumption that pairs of positions are statistically independent of other positions with respect to their amino acid frequencies (Göbel et al., 1994) (Dunn et al., 2008; Gloor et al., 2005; Kass and Horovitz, 2002). These methods do not take into account transitive correlations (chaining effect) and, to a certain extent, conserved positions (Fodor and Aldrich, 2004) and, consequently, result in many inaccurate predictions. By contrast, methods that were developed more recently (Baldassi

et al., 2014; Feizi et al., 2013; Jones et al., 2012, 2015; Kamisetty et al., 2013; Marks et al., 2011; Morcos et al., 2011; Weigt et al., 2009) consider the amino acid frequencies at a pair of positions to be dependent on the frequencies at all other positions, thereby reducing noise due to transitivity (Figure 2.5) and introducing a substantial improvement relative to the earlier methods (Figure 2.6). Nevertheless, recent methods require more extensive computations. In this thesis, both the early and more powerful recently developed contact prediction methods were examined.

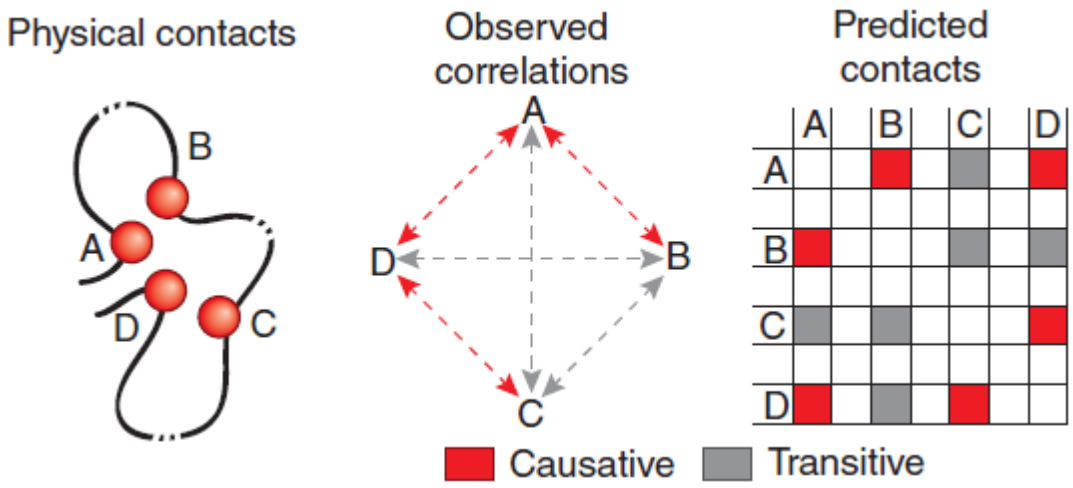


Figure 2.5 Transitivity (indirect) effects in protein contact prediction.
Transitivity occurs when correlations due to direct (causative) interactions between residues A and B, A and D, and residues D and C result in a transitive correlation between residues B and C. Transitive correlations can be stronger than causative correlations if, for example, two non-interacting residues have many common neighbors. This figure is taken from previous work of others (Marks et al., 2012).

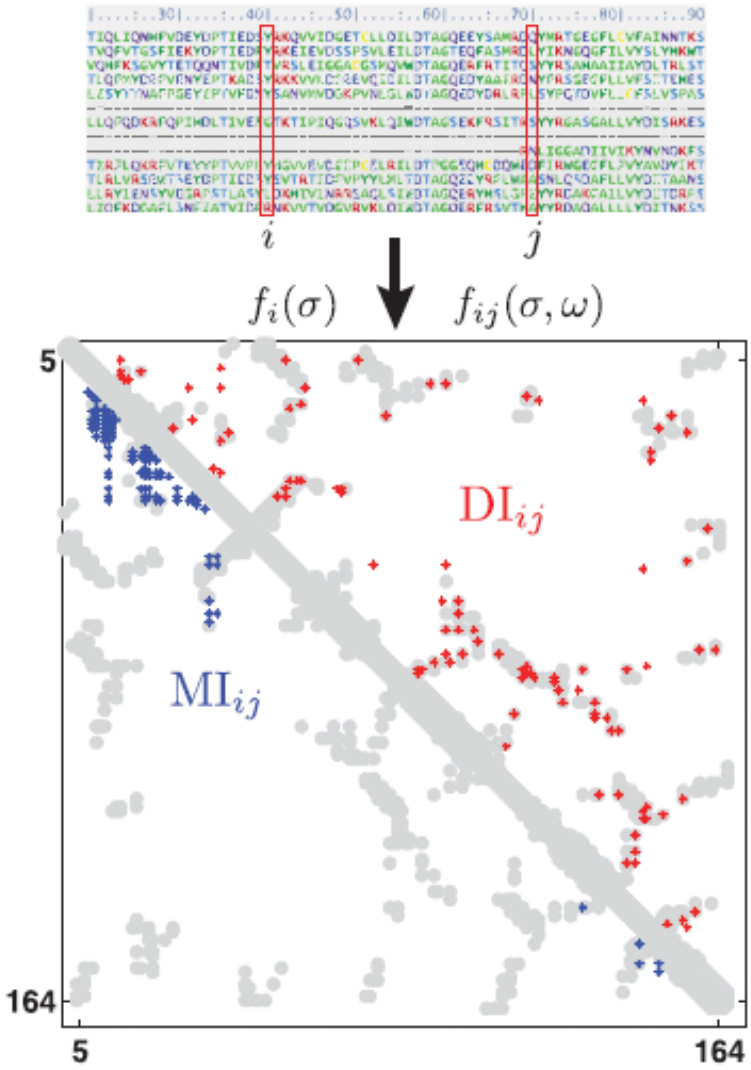


Figure 2.6 Protein contact prediction by representative early and recent methods. Protein contact prediction for the human Ras protein family using the early mutual information (MI) method and the more recent maximum entropy-based direct information (DI or DCA) methods (blue and red, respectively). The 150 predicted contacts with highest score obtained from both methods are shown in the protein contact map (in gray) derived from the experimentally determined structure of Ras. This figure is taken from previous work of others (Stein et al., 2015).

2.2.3.1 Early methods

2.2.3.1.1 The OMES method

The score for a pair of positions i and j, S (i,j), for the OMES (Observed Minus Expected Squared) method is calculated, as follows (Fodor and Aldrich, 2004; Kass and Horovitz, 2002):

$$S_{OMES}(i,j) = \sum_a \sum_b \frac{(OBS_{a_i b_j} - EXP_{a_i b_j})^2}{N_{valid}},$$

where $OBS_{a_i b_j}$ and $EXP_{a_i b_j}$ are the respective observed and expected number of sequences in the MSA with residue type a at position i and residue type b at position j. N_{valid} is the number of sequences in the alignment that have non-gapped residues at both i and j positions. Since gaps are excluded from the Pfam HMM based MSA before analysis (see details in the source code at https://github.com/etaijacob/CMA), $N_{valid} = 1$.

2.2.3.1.2 The MI method

Mutual information, MI, measures the reduction of uncertainty of one position given the information for the other (Cover and Thomas, 2005). MI can be viewed as the degree of correlation between two positions. The score for the MI method is calculated as follows (Gloor et al., 2005):

$$S_{MI}(i,j) = \sum_{a=1}^{21} \sum_{b=1}^{21} f_{(i,a;j,b)} \; log \frac{f_{(i,a;j,b)}}{f_{(i,a)} f_{(j,b)}}$$

where $f_{(i,a)}$ and $f_{(j,b)}$ denote the respective frequencies of occurrence of residue type a at position i and residue type b at position j and $f_{(i,a;\,j,b)}$ denotes the joint probability of occurrence of residue type *a* at position *i* and type *b* at position *j*.

2.2.3.1.3 Correction for phylogenetic background and entropic noise

Further improvements to MI, OMES and other methods can be done using correction methods that take into account the phylogenetic and entropic bias in the sequence family. Entropic noise originates from insufficient sequences in the MSA for adequate sampling of all residue types. Phylogenetic background refers to correlations due to the pattern of the underlying evolutionary tree. The correction method used in my analyses called average product correction (APC) is based on the assumption that each position in a MSA may have a propensity for a specific background signal M_b, which relates to its entropy and

phylogenetic history. The background M_b for any two positions can be approximated by the product of their propensities (Dunn et al., 2008). In the case of MI, an average product correction (APC) term is subtracted from the MI score for each pair of positions (the MI method including the APC correction is called MIp). The APC term, which is a measure of the background MI shared by positions i and j, is given by:

$$APC(i,j) = \frac{MI_{(i,\bar{x})}MI_{(j,\bar{x})}}{\overline{MI}},$$

where the terms in the nominator are the respective average MI values of positions i and j with all other positions in the alignment and the term in the denominator is the average background MI of all the positions in the alignment. The MIp score is given by:

$$S_{MI_p}(i,j) = S_{MI}(i,j) - APC(i,j)$$

2.2.3.2 Recent methods

2.2.3.2.1 Introduction to recent methods

Multivariate statistical methods (Balakrishnan et al., 2011; Ekeberg et al., 2013; Jones et al., 2012; Kamisetty et al., 2013; Morcos et al., 2011)and other recently developed methods (Feizi et al., 2013) are able to remove noise originated from transitivity and, thus, detect direct contacts more accurately than earlier methods. Employing these methods has become possible also because of the enormous increase in the number of sequences (i.e. a larger sample size) and the availability of more computing resources over the years. The first method developed, Direct Coupling Analysis (DCA or DI for direct information), was implemented using the message passing algorithm (Weigt et al., 2009), a computationally intense procedure that required a long time to complete a prediction for a very small number of pairs of positions (~4 days for 60 contacts on a single CPU). A significant breakthrough in the approximation method, which drastically reduces the computation time (Morcos et al., 2011), was introduced 2 years later. Other methods were also developed that include Protein Sparse Inverse COVariance (PSICOV) (Jones et al., 2012, 2015), a Bayesian network algorithm (Burger and van Nimwegen, 2010), Gremlin's pseudo-likelihood method (Kamisetty et al., 2013), the pseudo-likelihood maximization DCA (plmDCA) method (Ekeberg et al., 2013) and a network deconvolution approach based on spectral decomposition of the correlation matrix (Feizi et al., 2013). Here, I describe in more detail the two methods, DCA and PSICOV, which were chosen to assess our approach.

2.2.3.2.2 The Direct Coupling Analysis (DCA) method

2.2.3.2.2.1 Description of the DCA method

By contrast with the early methods, frequency counts in DCA are reweighted in order to avoid overrepresentation of similar sequences in the analysis. The weight of each sequence, a, is determined by its similarity to all the other sequences in the MSA. The weight $\frac{1}{m_a}$ of sequence a is given by:

$$m_a = \sum_{i=1}^{M} I(a, s_i),$$

where s_i is the sequence in row i in the MSA with M sequences and $I(a, b)$equals 1 if the sequence similarity between a and s_i is greateror equal to0.8 and 0 if otherwise. Note that

using a threshold of 1 instead of 0.8, would reweight each sequence by the number of times it appears in the MSA, thus removing simple sequence repeats. The effective number of independent sequences is defined here for later use by:

$$M_{eff} = \sum_{a=1}^{M} \frac{1}{m_a}.$$

In order to reweight the contribution of each sequence to the total frequency counts according to its similarity to other sequences, the marginal and joint frequency counts are calculated as follows:

$$f_i(A) = \frac{1}{\lambda + M_{eff}} \left(\frac{\lambda}{q} + \sum_{a=1}^{M} \frac{1}{m_a} \delta_{A,A_i^a} \right),$$

$$f_{ij}(A,B) = \frac{1}{\lambda + M_{eff}} \left(\frac{\lambda}{q^2} + \sum_{a=1}^{M} \frac{1}{m_a} \delta_{A,A_i^a} \, \delta_{B,A_j^a} \right).$$

Where in the amino acid analysis A and B designate amino acid types and q equals 21. A_i^a and A_j^a designate the amino acid type at specific position i or j in sequence a for $i, j = 1, \ldots, L$, where L is the sequence length. In the codon analysis case, A and B designate codon type and q equals 65, and A_i^a and A_j^a designate the codon types. δ denotes the Kronecker symbol, which equals one if the two indices agree and zero if otherwise. λ is the pseudo-count which is equal here to M_{eff} and will be discussed in the next paragraph. Thus, for example, if a group of 50 out of 100 sequences in a given MSA has a sequence similarity of more than 80% among all group members, each of these members will contribute 1/50 of its original count for the frequency calculations.

DCA uses the inverse covariance matrix (defined below) to predict direct coupling. In order to ensure that the covariance matrix is invertible (that is, the probability distribution is unique), a pseudo-count, λ, is used for regularization of the above frequency counts (Neher, 1994) for finite sample effect. In the extreme cases, the pseudo-count prevents counts from being equal to zero when there are not enough sequences in the MSA (i.e. finite sample effect) to sample all possible amino acids or codon pair combinations. For example, if in a given MSA there is not a single observation that accounts for the joint occurrence of arginine in column i and lysine in column j, instead of having a frequency count of zero, it will have a value of

$$\frac{1}{\lambda + M_{eff}} \frac{\lambda}{q^2} = \frac{1}{2M_{eff}} \frac{M_{eff}}{q^2} = \frac{1}{2q^2},$$

following the pseudo-count addition.

The covariance matrix is:

$$C_{ij}(A,B) = f_{ij}(A,B) - f_i(A) f_j(B),$$

and the coupling strength between amino acid types A and B, or codon types in the case of codon analysis, at positions i and j, respectively, denoted as $e_{ij}(A,B)$, is approximated by

$$e_{ij}(A,B) \simeq -\left(C^{-1}(A,B)\right)_{ij}.$$

Let us define H_{ij}, which can be interpreted as the Hamiltonian of positions i and j corresponding to the total coupling energy, $e_{ij}(A,B)$, and the local fields, $h_i(A)$ and $h_j(B)$.

$$H_{ij} = -\left(e_{ij}(A,B) + h_i(A) + h_j(B)\right).$$

The direct coupling is therefore,

$$P_{ij}^{dir}(A,B) = \frac{1}{Z_{ij}} exp\left(-H_{ij}\right),$$

whereZ_{ij} is the partition function (i.e. the normalizing constant). The local fields, $h_i(A)$ and $h_j(B)$, are determined for each pair A and B at positions i and j, respectively, by adjusting the marginal distributions of $P_{ij}^{dir}(A,B)$ to the reweighted frequency counts defined above, $f_i(A)$ and $f_j(B)$, as follows:

$$f_i(A) = \sum_B P_{ij}^{dir}(A,B) \text{and } f_j(B) = \sum_A P_{ij}^{dir}(A,B).$$

P_{ij}^{dir}, like the Boltzmann distribution, shows that coupling with lower energy will always have a higher probability than coupling with a higher energy.

Finally, the direct information formula is similar to the MI's, except for the joint frequency counts, f_{ij}, that were replaced by the direct coupling, P_{ij}^{dir}. Therefore,

$$DI_{ij} = \sum_{A,B} P_{ij}^{dir}(A,B)\, ln\left(\frac{P_{ij}^{dir}(A,B)}{f_i(A) f_j(B)}\right),$$

2.2.3.2.2.2 DCA Model formulation for continuous random variables

Recall that the model of the DCA method is defined by the following distribution function:

$$P(A_1, \ldots, A_N) = \frac{1}{Z} exp\left(\sum_{i<j} e_{ij}(A_i, A_j) + \sum_i h_i(A_i)\right)$$

Given that the aligned protein data are limited, many different probability distributions can be consistent with it and the choice is made by finding the probability distribution that both satisfies the constraints of marginal and joint frequencies and maximizes the entropy. For simplicity, I describe here how the model is derived by satisfying those constraints for the case of continuous random variable but it is equivalent to that used in the DCA procedure.

Let $X = (x_1, \dots, x_L)^T \in \mathbb{R}^L$ be a multivariate random variable. We require that the model will maximize the entropy

$$S = -\int_X P(X) ln P(X) dX,$$

and simultaneously meet the following constraints:

The first natural requirement for a probability distribution is that its integral equals one:

$$\int_X P(X) dX = 1$$

The second constraint is that the first moment of variable x_i, $\langle x_i \rangle$, should be equal to sample mean over M sequences in the MSA in each $i = 1, \dots, L$,

$$\langle x_i \rangle = \int_X P(X) x_i dX = \frac{1}{M} \sum_{m=1}^{M} x_i^m = \overline{x_i}$$

Equivalently, the third constraint requires that the second moment of the variables x_i and x_j, $\langle x_i\, x_j \rangle$, should be equal to its corresponding empirical expectation,

$$\langle x_i x_j \rangle = \int_X P(X) x_i x_j dX = \frac{1}{M} \sum_{m=1}^{M} x_i^m x_j^m = \overline{x_i x_j}$$

Finding the maximum of function S subject to the above three constraints is done using the method of Lagrange multipliers (Mead and Papanicolaou, 1984; Stein et al., 2015). With the Lagrange multipliers $\alpha, \beta = (\beta_i)_{i=1,\dots,L}$ and $\gamma = (\gamma_i)_{i=1,\dots,L}$ corresponding to the first, second and third constraints respectively, the Lagrangian $\mathcal{L} = \mathcal{L}(P(X); \alpha, \beta, \gamma)$ is defined as,

$$\mathcal{L} = S + \alpha(\langle 1 \rangle - 1) + \sum_{i=1}^{L} \beta_i (\langle x_i \rangle - \overline{x_i}) + \sum_{i=1}^{L} \gamma_{ij} (\langle x_i x_j \rangle - \overline{x_i x_j})$$

The maximum is then obtained by setting the derivative of $\mathcal{L}$ to zero with respect to the unknown density $P(X)$, given the definitions of the first and second moments above,

$$\frac{d\mathcal{L}}{dP(X)} = 0 \xrightarrow{yields} -ln P(x) - 1 + \alpha + \sum_{i=1}^{L} \beta_i\, x_i + \sum_{i=1}^{L} \gamma_{ij}\, x_i x_j = 0$$

The solution is therefore the Boltzmann distribution,

$$P(X; \beta, \gamma) = \exp\left(-1 + \alpha + \textstyle\sum_{i=1}^{L} \beta_i\, x_i + \sum_{i=1}^{L} \gamma_{ij}\, x_i x_j\right) =$$

$$\frac{1}{Z} exp\left\{-\left(\sum_{i=1}^{L} \beta_i\, x_i - \sum_{i=1}^{L} \gamma_{ij}\, x_i x_j\right)\right\}$$

With the normalization constant (also called the partition function) derived from the first constrained,

$$Z(\beta,\gamma) := \int_x \; exp\left(\sum_{i=1}^{L} \beta_i \, x_i + \sum_{i=1}^{L} \gamma_{ij} \, x_i x_j\right) dx \equiv \exp(1-\alpha)$$

The determining parameters of the probability distribution above, the Lagrange multipliers β and γ, that are equivalent to the coupling parameter, $e_{ij}(\sigma,\omega)$, and $\tilde{h}_i(\sigma)$ and $\tilde{h}_j(\omega)$ from the P_{ij}^{dir} formula described earlier, can be estimated from the closed-form solution in the case described here for the continuous variable model. In general, the Lagrange multipliers β and γ can be specified in terms of the empirical mean and the inverse covariance matrix, which is determined from the empirical correlation matrix,

$$C_{ij}^{(emp)}(x_i,x_j) = f_{ij}(x_i,x_j) - f_i(x_i)f_j(x_j).$$

Consequently, the maximum entropy distribution for the empirical first and second moments is found to be the multivariate Gaussian distribution. Further details can be found in (Morcos et al., 2011; Stein et al., 2015). The same numerical solution is obtained for the categorical variable using the mean-field approximation on the truncated Taylor series (Baldassi et al., 2014; Morcos et al., 2011).

2.2.3.2.2.3 Extension of categorical variables to binary representation

Furthermore, the categorical variables as represented in the MSA, can be extended using a binary representation (see figure 2.7) to a continuous one, with the advantage of the analytical framework (Baldassi et al., 2014; Stein et al., 2015).

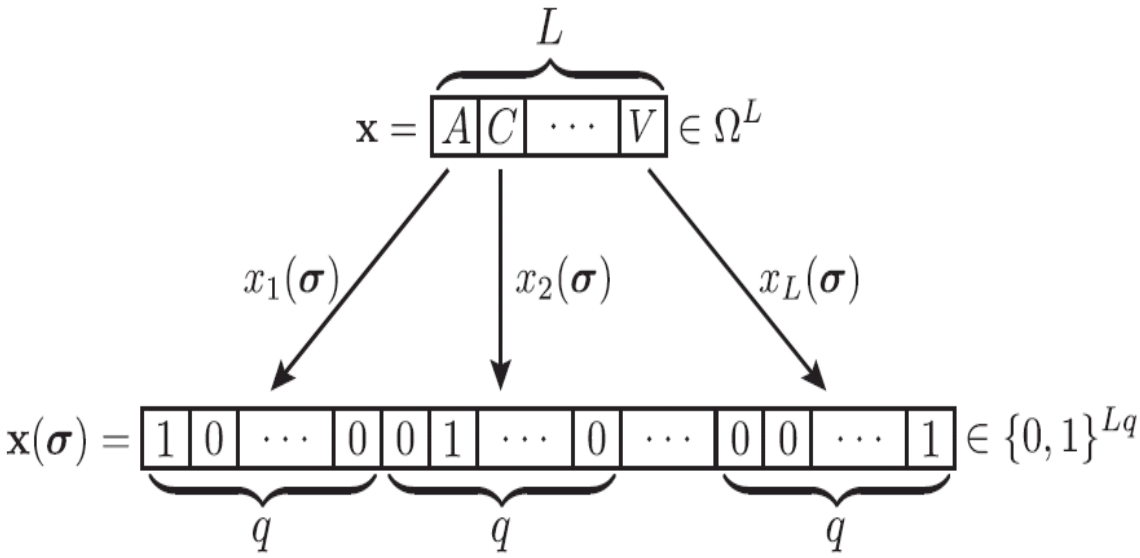


Figure 2.7 Illustration of binary translation of a categorical representation of amino acids.

The binary translation, $\Omega \rightarrow \{0,1\}^{Lq}$, maps each vector ofcategorical random variables, $X \in \Omega^L$, here represented by a sequence of amino acids from the amino acid alphabet, Ω = {A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R,S, T, V,W, Y,−}, onto a unique binary representation, $X(\sigma) \in \{0,1\}^{Lq}$. This figure is taken from previous work of others (Stein et al., 2015).

2.2.3.2.3 The PSICOV method

PSICOV is based on the sparse inverse covariance technique (Meinshausen and Bühlmann, 2006) and estimates the coupling effect between two positions in the protein based on the MSA of its related family members. As the DCA method, PSICOV method aims to correct for transitivity and also uses in its recipe the inverse covariance matrix (see description in the global model introduction above). As described above, given the observed marginal frequencies,

$$f_i(A) = \frac{1}{M}\sum_{a=1}^{M} \delta_{A,A_i^a},$$

$$f_{ij}(A,B) = \frac{1}{M}\sum_{a=1}^{M} \delta_{A,A_i^a}\,\delta_{B,A_j^a},$$

with $1 \le i,j \le L, 1 \le A,B \le q$ and δ denoting the Kronecker symbol, which equals one if the two indices agree and zero otherwise. The empirical covariance matrix is then,

$$C_{ij}(x_i,x_j) = f_{ij}(x_i,x_j) - f_i(x_i)f_j(x_j).$$

Assuming that the underlying distribution of the data is multivariate Gaussian, in the inverse covariance matrix, C^{-1}, the element ${C^{-1}}_{ij}$ represents the covariance between the residuals resulting from a regression of i with all other positions and the residuals resulting from the regression of j with all other positions. Thus, the matrix of partial correlation coefficients for all pairs of positions can be obtained using the Pearson correlation coefficient as follows:

$$\rho_{ij} = -\frac{{C^{-1}}_{ij}}{\sqrt{{C^{-1}}_{ii}{C^{-1}}_{jj}}}$$

As stated earlier, the empirical covariance matrices of MSAs are singular because the number of observed variables is often smaller than the dimensionality of the problem. Since the matrix cannot be directly inverted, PSICOV method uses the sparse inverse covariance estimation. In general, protein contact maps are sparse since only about 3% of all residue pairs in a protein structure tend to have a direct contact. This method uses this expected sparsity (i.e. low number non-zero terms in the matrix) of the covariance matrix as a constraint on the obtained solution. The PSICOV method used in this thesis, is based on the graphical Lasso technique (Banerjee et al., 2008; Friedman et al., 2008). This method estimates the inverse covariance matrix, given S, the empirical covariance matrix with $d \times d$ dimensions, by minimizing the objective function:

$$trace(S\Theta) - log(det\Theta) + \rho \sum_{i,j \in d} |\Theta_{ij}|$$

The third term is the regularization part, a type of a penalty (which is also called ℓ_1 norm) that favors sparse solutions in the sense that many of the positive values in Θ will become zero during the minimization process.

The norm of contacting residues i and j is the sum of the 20×20 absolute values in Θ, corresponding to the 20 amino acid types observed in the alignment columns i and j:

$$S_{ij}^{contact} = \sum_{ab} |\Theta_{ij}^{ab}|,$$

for a and b amino acid types.

The score used for prediction is corrected for entropic and phylogenetic noise using the average product correction (APC), exactly as described above concerning the MI method:

$$\left(S_{ij}^{contact}\right)_{APC} = S_{ij}^{contact} - \frac{S_{(i,\bar{x})}^{contact} S_{(j,\bar{x})}^{contact}}{\overline{S^{contact}}},$$

where $S_{(i,\bar{x})}^{contact}$ and $S_{(j,\bar{x})}^{contact}$ are the mean norm between column i and all other columns or column j and all other columns, respectively. $\overline{S^{contact}}$ is the mean norm across whole MSA. In the implementation of PSICOV, as used in this thesis, additional standardization is done to this score for the final prediction output.

2.2.4 Contact definitions and performance evaluation

2.2.4.1 Choosing contact definitions for the evaluation

In order to assess the accuracy of contact predictions we must first decide which atom types and distance cutoffs will be used to define a contact (Yuan et al., 2012). In CASP (Moult et al., 1995, 2011) and many other applications (Jones et al., 2012; Kamisetty et al., 2013) residues are defined as being in contact if the distance between their C_β atoms is ≤ 8 Å. In several other applications (Baldassi et al., 2014; Morcos et al., 2011), a contact is assumed to exist if at least one inter-atomic distance between the residues is ≤ 8 Å. The former is referred here as the C_β-based definition and the latter as the "all" definition. A direct physical contact occurs when two heavy atoms of the respective amino acids are at a distance < 3.5 Å. We, therefore, examined which of the above two definitions is better at identifying such direct contacts. We then used the more accurate definition to assess the performance of CMA methods in contact prediction. The examination was performed on a compilation of a non-redundant set of thousands of proteins with an available crystal structure (next section describes the technical details of this analysis). We determined the fraction of the amino acid pairs defined as contacts using the C_β-based definition or the "all" definition that are actually in direct physical contact (using the definition given above). This calculation was done separately for each protein. Direct physical contacts were found to comprise 30% of the interactions identified using the C_β-based definition and only 10% of the interactions identified using the "all" definition. We, therefore, considered the C_β-based definition to be better for our analysis (Figure 2.8).

2.2.4.2 Procedure for determining physical contacts

A non-redundant set of 2,481 PDB entries was downloaded from the CullPDB website (Wang and Dunbrack, 2003, 2005) at http://dunbrack.fccc.edu/Guoli/pisces_download.php on Feb. 25th, 2015. The downloaded set was compiled based on the following properties: (i) a protein sequence identity cutoff of 20%; (ii) structures with an X-ray resolution higher than 1.6 Å; and (iii) an R-factor cutoff of 0.25. For each protein, I identified the residue pairs in contact according to the "all" and C_β-based definitions. The fraction of residue pairs that are in a true physical contact (i.e. if they have at least one pair of atoms with a distance < 3.5 Å) was then calculated for each of these sets. Only pairs of residues that are separated by at least five amino acids along the protein sequence were considered.

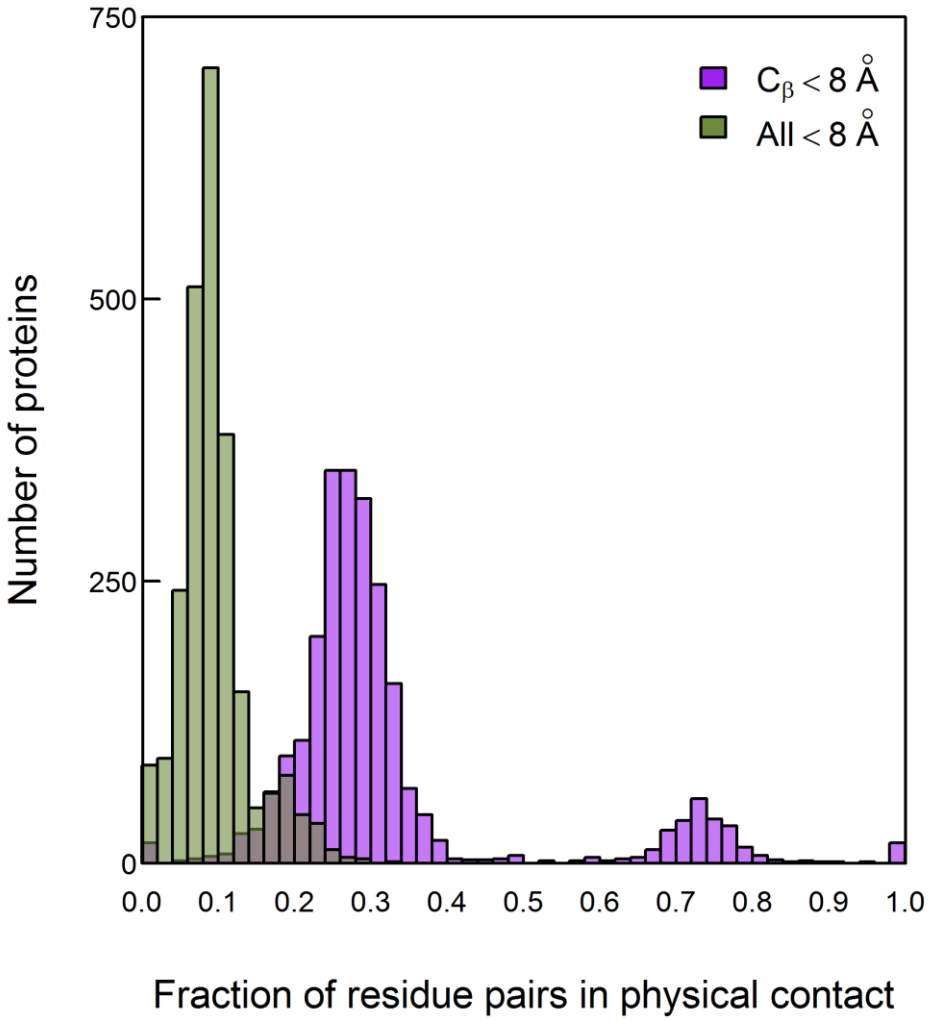


Figure 2.8 Histogram of the fractions of residue pairs in physical contact out of those considered to be in contact according to two widely used definitions.
Residue pairs defined to be in contact if at least one inter-atomic distance between them is ≤ 8 Å (designated 'All') or if the distance between their C_β atoms is ≤ 8 Å were identified in 2,481 proteins with high-resolution structures. The fraction of these residue pairs that are in direct contact, i.e. with a distance < 3.5 Å between two of their respective heavy atoms, was then determined for each protein. Only pairs of residues that are separated by at least five amino acids along the protein sequence were considered.

2.2.4.3 Evaluation of prediction accuracy

The evaluation was based on the all structures with the highest resolution (at least 3 Å) but, in cases where families have more than 30 known structures with unique sequences, only the 30 with the best resolution were used (in cases of structures with the same resolution we arbitrarily chose one). The average accuracy of contact predictions for all the crystal structures of each domain family was then calculated so that domain families with many

crystal structures would not be over-represented. Accuracy was calculated as the proportion of true contacts from the N pairs with the highest score in that set. We evaluated the improvement of our method using the difference in the area under the curve (AUC) of the accuracy vs. number of predicted pairs of our method relative to the results of the original OMES, MI, MIp, PSICOV and DCA methods. AUC was calculated using the auc function in MESS package in R with the default parameters.

2.2.5 Contact prediction implementation

The Direct Coupling Analysis (DCA) method (Morcos et al., 2011) was implemented and optimized in R and C for amino acid and codon MSAs based on a Matlab source code provided by Weigt et al. (http://dca.rice.edu/portal/dca/download). The PSICOV code was downloaded from http://bioinfadmin.cs.ucl.ac.uk/downloads/PSICOV/ and used for the predictions based on amino acid MSAs with the default parameters for faster options as recommended by the authors (-p -r 0.001 and with the -l option in order to avoid using the APC term). The PSICOV code was modified in order to carry out the same analysis for codon MSAs and a python script was implemented to perform the whole analysis as done for the other methods using Pfam MSA files in Stockholm format and fasta MSA files as inputs. PSICOV was used here either with the APC for amino acid MSAs or without the APC for the predictions based on both amino acid and codon MSAs.

2.2.5.1 Available software for CMA analysis

The R and Python source codes for the contact prediction by all methods, C source code modifications to PSICOV V2.1b3, R source code for structure-domain sequence mapping and python scripts for generating codon MSAs are available at https://etaijacob.github.io/. Details on the relevant R packages that will be available on CRAN will also be provided at: https://etaijacob.github.io/.

2.2.6 Other applications using codon information - Deleterious SNPs prediction

2.2.6.1 SNPs Datasets collection

Several methods for predicting damaging SNPs were successfully developed over the years. One of the best performing methods is PolyPhen-2, which is based on a classification method that uses two different datasets for training and testing

(http://genetics.bwh.harvard.edu/pph2/dokuwiki/_media/nmeth0410-248.pdf). Here, I used these two datasets to train and test a new classifier for this purpose. In order to assess if combining codon information with other features improves the performance of deleterious SNP prediction, I built a new classifier which uses both amino acid and codon information. One dataset used was HumDiv, which is compiled from all 3,155 damaging alleles annotated in the UniProt database as causing human Mendelian diseases and affecting protein stability or function. This dataset also includes 6321 non-damaging SNPs defined here as positions at which there is a difference between the human protein and its closely related mammalian homologs. The second dataset, HumVar, consists of all the 13,032 human disease-causing mutations from UniProt. This dataset also includes 8,946 human nonsynonymous single-nucleotide polymorphisms (nsSNPs) without annotated involvement in disease, which were treated as non-damaging.

2.2.6.2 Generation of independent variables for the prediction

The evolutionary conservation at the position of a SNP in a protein was used as the independent variable for predicting whether it is deleterious or not. I estimated such conservation by calculating the entropy at the SNP's location in the multiple sequence alignment of that protein and its homologs. Calculations were restricted to SNPs that are located within domain regions of Pfam families in proteins. Therefore, families from the Pfam database (RP75 redundancy level and Pfam version 27) that included a protein member with an indicated SNP in its domain region, were those considered in the analysis. This resulted in 1005 MSAs of different domains with mapped SNPs. The transcripts for the codon based MSAs were collected based on Uniprot cross reference annotations (for Refseq, Ensemble, EMBL and Ensembelgenomes databases) and aligned in accordance to the proteins MSAs using tranalign software tool. The entropy for each mapped SNP for the proteins MSAs and the transcripts MSAs was calculated for all SNPs positions with less than 50% gaps in the alignment. In order to include in the same analysis the entropy measurements of SNPs from different domains with MSAs of different size and compositions, each entropy value was standardized (that is, centered by the mean and scaled by the standard deviation of the entropy values).

2.2.6.3 Prediction model

I used multivariate logistic regression to estimate the probability of an SNP to be damaging or non-damaging (i.e. a binary response) from the entropy calculations (i.e.

independent variables) of the amino acid and codon MSAs. The regression model is defined as follows:

Let Y be the probability for a SNP to be damaging. The multiple logistic regression is defined as follows:

$$\ln\left(\frac{Y}{(1-Y)}\right) = a + b_1 Zscore(H(AA)) + b_2 Zscore(H(C))$$

where $\frac{Y}{(1-Y)}$ is the odds ratio of a SNP to be damaging compared to non-damaging, $H(AA)$ and $H(C)$ are the conservation scores based on the amino acid and codon sequences, respectively, and $Zscore$ indicates the standardization function.

The logistic regression coefficients, b_1 and b_2 can be used to interpret the relation between the codon and the amino acid based independent variables. Performance evaluation consists of data divided into two equally sized sets: test and learning (i.e. two-fold cross validation).

2.3 Results

2.3.1 The rationale of the method

The key premise underlying the method introduced in this thesis is that a correlation at the amino acid level between two positions is more likely to reflect a direct interaction if the correlation at the codon level for these positions is weak (Figure 2.2). In other words, it is assumed that cases of strong correlations at both the amino acid and codon levels for a pair of positions are less likely to reflect selection to conserve protein contacts and more likely to reflect selection to conserve interactions involving DNA or RNA and/or common ancestry. Given this rationale in mind, we decided to test whether contact identification improves when all the pairs of positions are ranked using a score that increases with (i) increasing strength of the correlation at the amino acid level and (ii) decreasing strength of the correlation at the codon level. Such a score, S_i, is given, for example, by:

$$S_i = S_i^{\alpha}(aa)/S_i(c),$$

where $S_i(aa)$ and $S_i(c)$ are the scores generated by method i (e.g. MI) for the amino acid and codon alignments, respectively, and the value of the power α is determined empirically depending on the method (see below).

2.3.2 Performance analysis and comparison

Our approach was tested for the OMES (Kass and Horovitz, 2002), MI (Gloor et al., 2005), MIp (Dunn et al., 2008) and DCA (Marks et al., 2011; Morcos et al., 2011) methods using 114 MSAs each comprising at least 2000 sequences of length between 200 and 500 residues. In the case of the PSICOV method (Jones et al., 2012), only 86 MSAs out of the 114 MSAs were used since the others didn't pass this method's threshold for amino acid sequence diversity. Each MSA also included at least one sequence with a known crystal structure at a resolution < 3 Å in which at least 80% of all the residues are resolved. The mean accuracy of contact identification was plotted as a function of the top ranked number of predicted pairwise contacts (Figure 2.10) or as a function of the top ranked fraction of protein length, L (Figure 2.9). Residues were considered as being in contact if the distance between their C_β atoms is $\leq$ 8 Å following the definition used in CASP experiments (Ezkurdia et al., 2009) and other studies (Kamisetty et al., 2013; Skwark et al., 2014) (see also Figure 9). The results show that the PSICOV and DCA methods outperform the

OMES, MI and MIp methods (Figure 2.9, 2.10) as established before (Mao et al., 2015). They also show that combining amino acid and codon data leads to an improvement in the predictions by OMES, MI, DCA and PSICOV. In the case of MIp, however, no improvement was observed despite the fact that this method performs worse than DCA and PSICOV. In MIp, a term called average product correction (APC) is subtracted from the MI score for each pair of positions in order to reduce false positives. Removing this correction from PSICOV where it also exists and including the codon data yielded the best method (Figure 2.9, 2.10). Hence, we can conclude that there is an overlap between the background noise reduced upon including the APC term and codon data and that including the latter can be more advantageous as we observe for PSICOV.

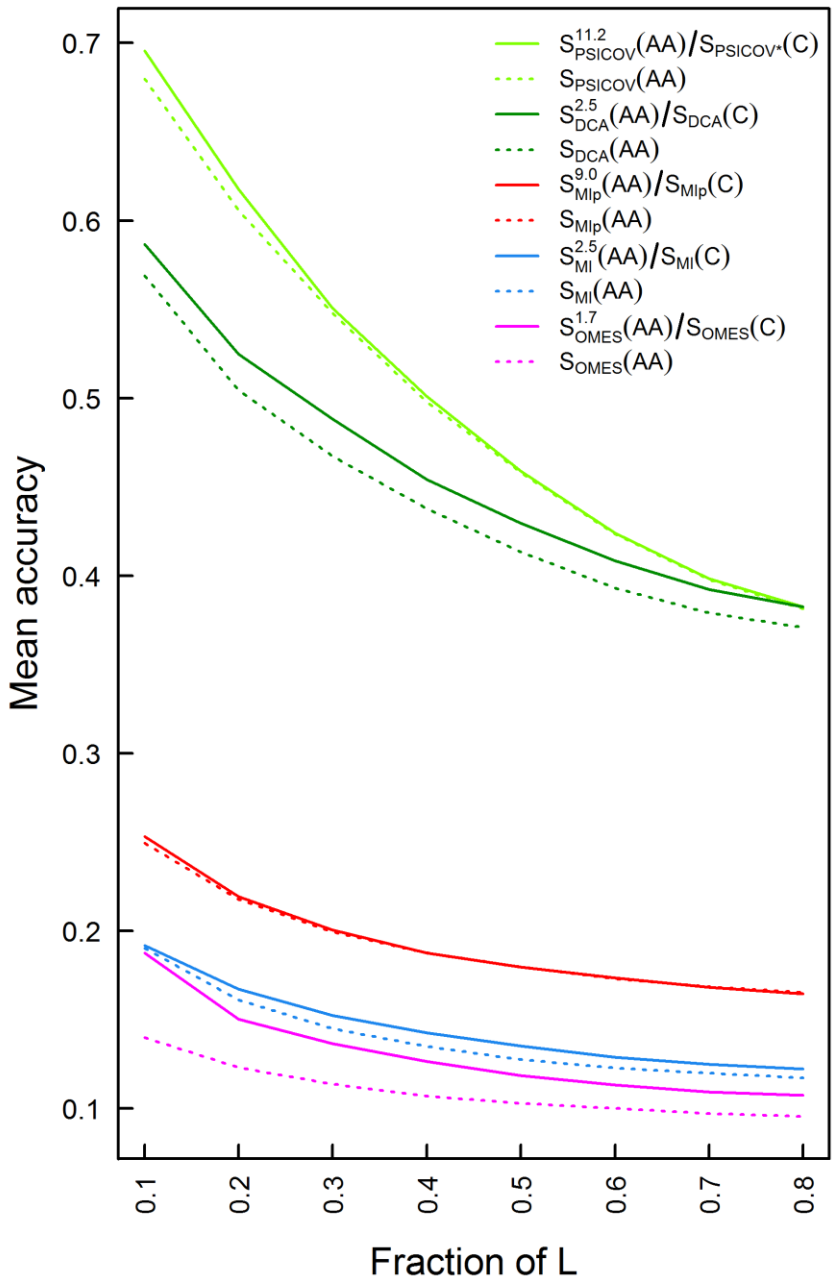


Figure 2.9 Plots of the mean accuracy of contact identification by various methods of correlated mutation analysis as a function of the top ranked fraction of protein length, L, number of predicted pairwise contacts.

The mean accuracies of contact identification by the OMES, MI, MIp, DCA and PSICOV methods are shown either with or without incorporating codon data. Residues were defined as being in contact if the distance between their C_β atoms is ≤ 8 Å. PSICOV* indicates that it was carried out without the APC.

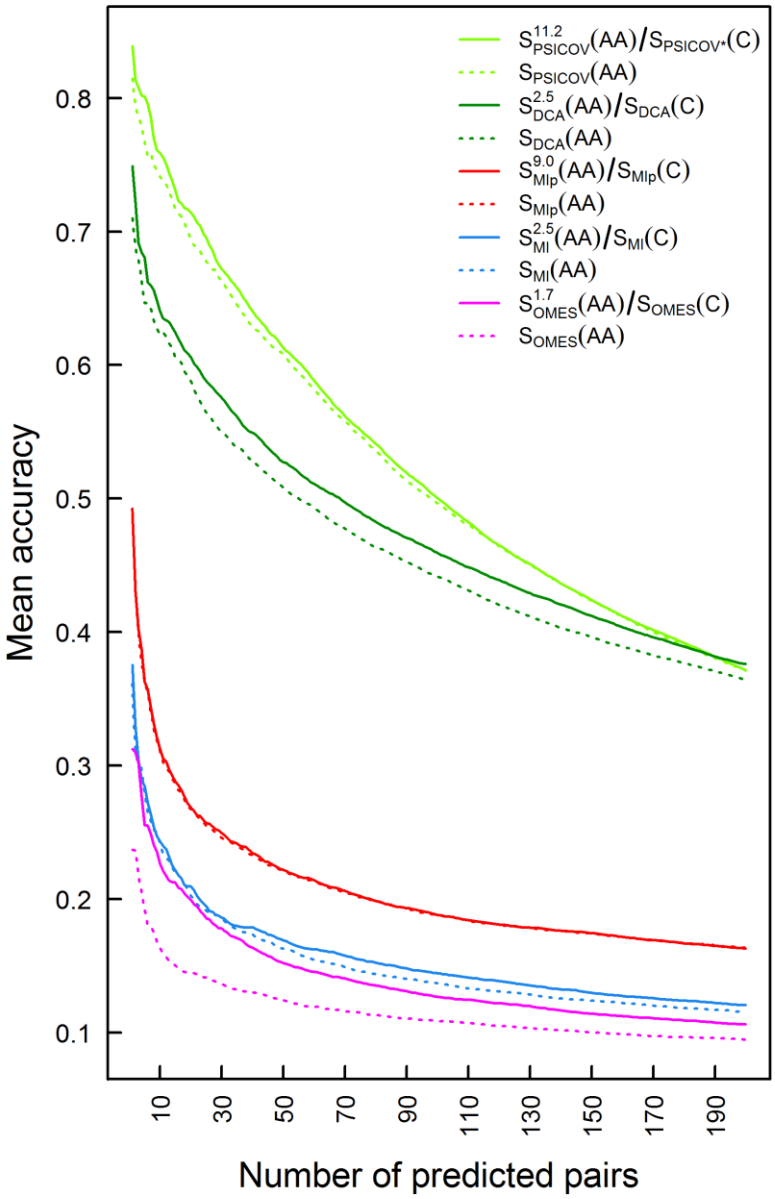
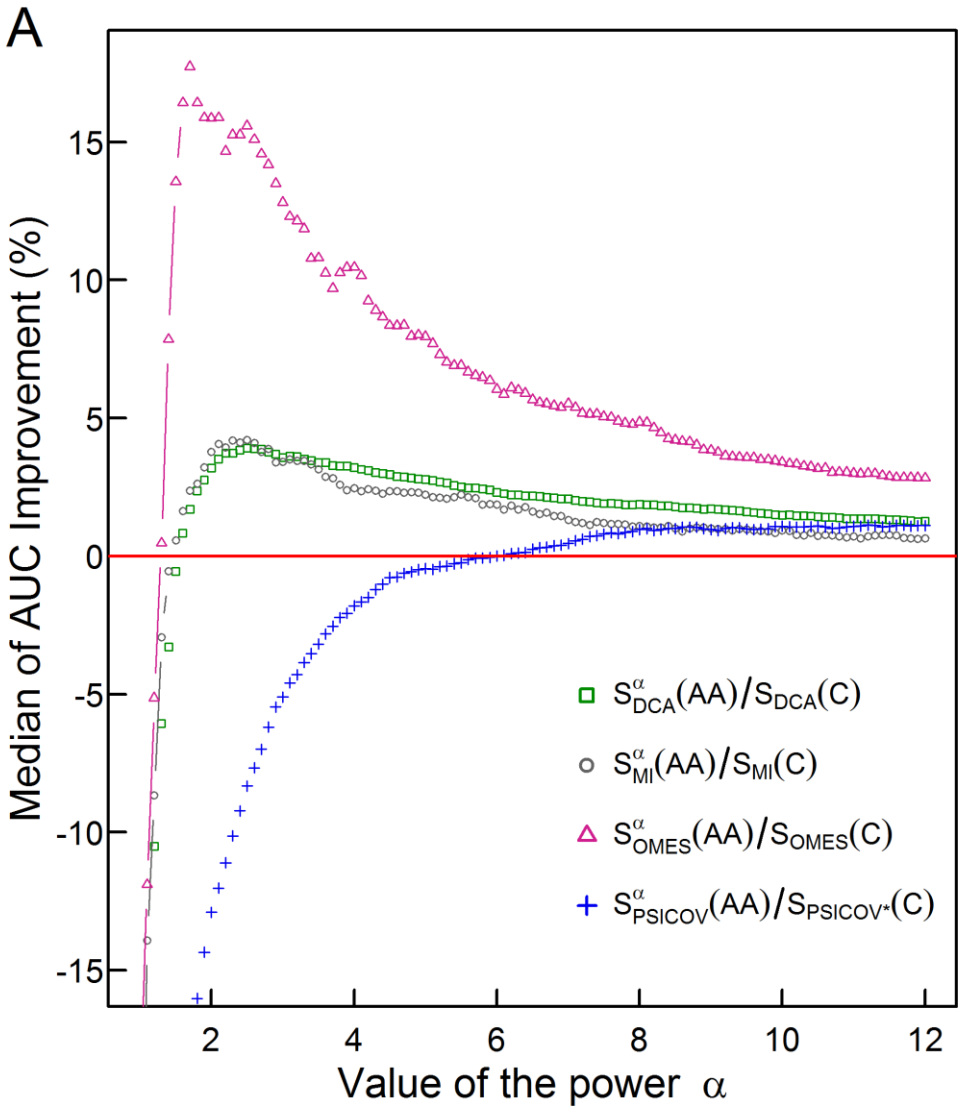


Figure 2.10 Plots of the mean accuracy of contact identification by various methods of correlated mutation analysis as a function of the top ranked number of predicted pairwise contacts.
The mean accuracies of contact identification by PSICOV, DCA MIp, MI and OMES are shown either with or without incorporating codon data. Residues were defined as being in contact if the distance between their C_β atoms is ≤ 8 Å.

2.3.3 Method optimization

The extent of improvement increases with increasing values of the power α until a maximum is reached (Figure 2.11A) at a value of α that depends on the method used and different values of α_{max} were, therefore, chosen accordingly. Cross-validation by dividing

the MSA data into training and test sets showed that the values of α_{max} are stable, i.e. they do not vary depending on the set of MSAs (Figure 2.12). Given these values of α_{max}, the significance of the extent of improvement was assessed by comparing for each MSA the accuracy of the contact predictions using the different methods with and without incorporating codon data. Significance levels were determined using two non-parametric tests: (i) the Wilcoxon signed-rank test, which takes into account both the number of MSAs for which the accuracy of the contact predictions increases upon incorporating codon data (e.g. 81 in the case of DCA) and the magnitude of the improvement; and (ii) the sign test, which only considers the number of MSAs with improved accuracy. The extent of improvement achieved by incorporating codon data was found to be highly significant as indicated by the P-values obtained using both tests (Figure 2.11B).


A
Median of AUC Improvement (%)
15
10
5
0
-5
-10
-15
2
4
6
8
10
12
Value of the power α
$S^{\alpha}_{DCA}(AA)/S_{DCA}(C)$
$S^{\alpha}_{MI}(AA)/S_{MI}(C)$
$S^{\alpha}_{OMES}(AA)/S_{OMES}(C)$
$S^{\alpha}_{PSICOV}(AA)/S_{PSICOV*}(C)$

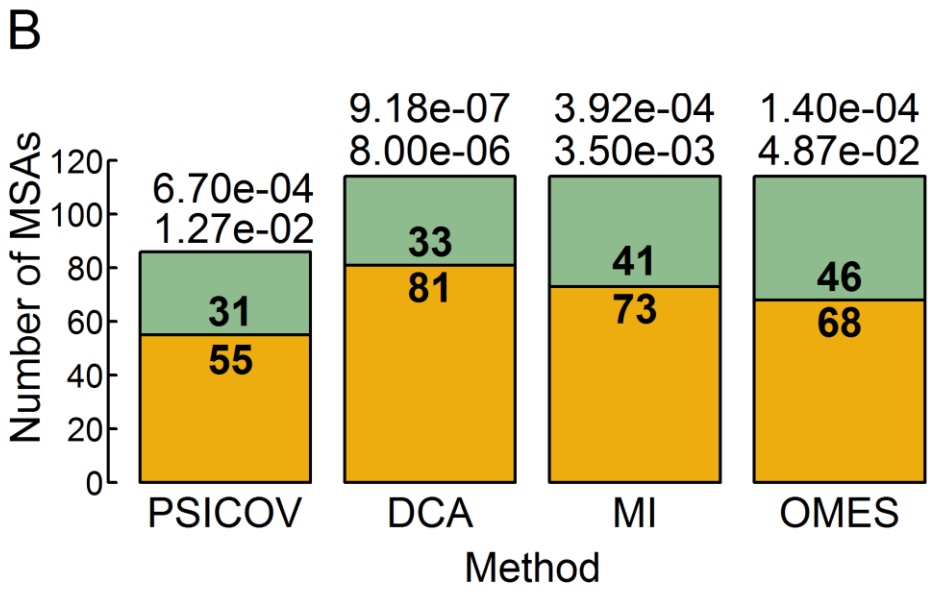

B
Number of MSAs
120
100
80
60
40
20
0
6.70e-04
1.27e-02
9.18e-07
8.00e-06
3.92e-04
3.50e-03
1.40e-04
4.87e-02
31
55
33
81
41
73
46
68
PSICOV
DCA
MI
OMES
Method

Figure 2.11 The effect of the relative weights of amino acid and codon information on contact prediction improvement and its statistical significance.
(A) The median of the extent of improvement in contact prediction for 114 MSAs (86 in the case of PSICOV) is plotted as a function of the value of the power α which determines the relative weights of the amino acid and codon correlations in the score, S_i ($S_i = S_i^{\alpha}(aa)/S_i(c)$, where $S_i(aa)$ and $S_i(c)$ are the respective amino acid and codon scores generated by method i). The extent of improvement was determined by calculating the difference in the areas under the curves (AUC) of prediction accuracy vs. number of predictions for each method i with and without incorporation of the codon data normalized by the area under the curve generated without codon data. The analysis was done for domains of length between 200 and 500 residues and at least 2000 coding sequences in their MSA. The value of α which maximizes the median improvement was used for predictions. Maximal respective improvements of 3.9% and 4.2% were found for DCA and MI when α is 2.5, 17.6% for OMES when α is 1.7 and 1.13% for PSICOV when α is 11.2. (B) Stacked bar plots showing the number of MSAs for which including codon data improved the contact predictions using the different methods (orange) and the number of those for which it was otherwise (green). The statistical significance of the improvement achieved by incorporating codon data is indicated by the top and bottom P-values obtained using the Wilcoxon signed-rank and sign tests, respectively.

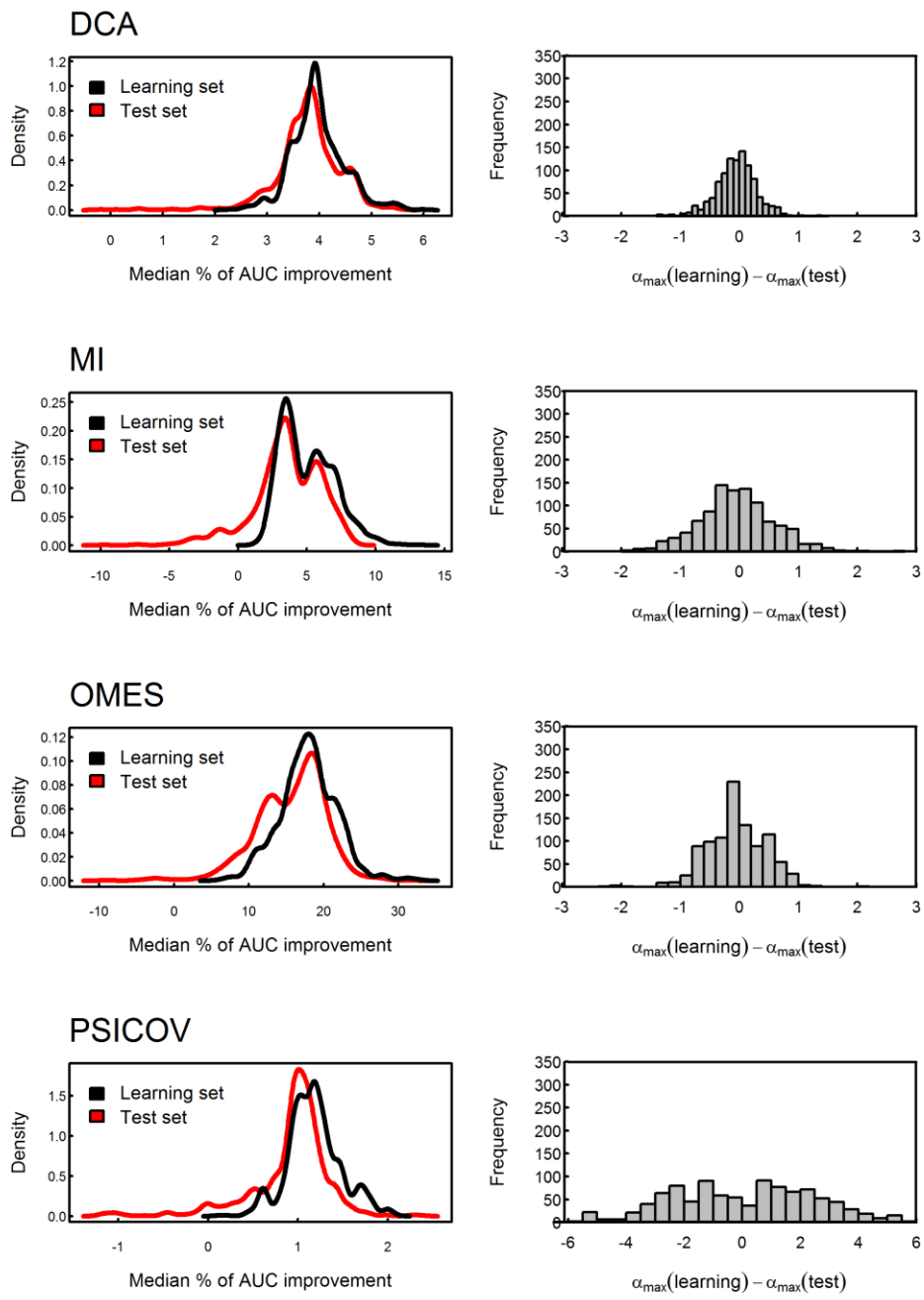


Figure 2.12 Testing the stability of the value of α by cross-validation.
The MSA data set was divided into 10,000 different learning and test sets of equal size. The value of αwhich produces the maximal median percent of AUC improvement in accuracy of contact prediction was obtained for each learning set and then used to assess the median percent of AUC improvement for the corresponding test set. The distributions of the median percent of improvement obtained for the test sets are shown for OMES, MI, DCA and PSICOV. The mean values of these test sets distributions obtained using the different methods are similar to those of their learning sets, thus, showing that the improvement is not due to over-fitting. In the case of OMES, MI and DCA, the mean difference between the values of α which maximizes the median of the percent of AUC improvement for the learning and test sets equals zero, thus, reflecting the stability of the values of α. In the case of PSICOV, the variance of that difference is high due to the asymptotic nature of the median percent of AUC improvement as a function of α.

2.3.4 Performance analysis for different contact definitions

The better success of DCA and other methods in identifying contacts according to the C_β-based definition when amino acid and codon data are combined is an important result since as stated earlier, more pairs that are in true physical contact are identified in this way. Nevertheless, my finding that the C_β-based definition of contacts is better than the 'All' definition but still poor (only 30% of the pairs defined as being in contact are in physical contact) prompted me to test the performance of our method for additional contact definitions. The mean of the extent of improvement in contact prediction for 114 domains (or 86 in the case of PSICOV) was, therefore, determined as a function of the distance that must exist between at least two C_β atoms in different residues in order for them to be defined as being in contact. It may be seen that, in the cases of PSICOV, OMES and DCA, the maximum improvements in contact prediction upon combining amino acid and codon data are when these distances are about 5.5, 7 and 5.5 Å, respectively, and that, in the cases of DCA and OMES, the improvement decreases dramatically when this distance is >~10 Å (Figure 2.13). In the case of MI, the extent of improvement upon combining amino acid and codon data is found to be relatively insensitive to the distance used to define a contact and is maximal when it is ~4.5 Å (Figure 2.13). These data, therefore, show again that the improvement in contact prediction upon combining amino acid and codon data is greatest when the distance used for contact definition does not lead to many pairs being defined in contact when in fact they are not in direct physical contact.

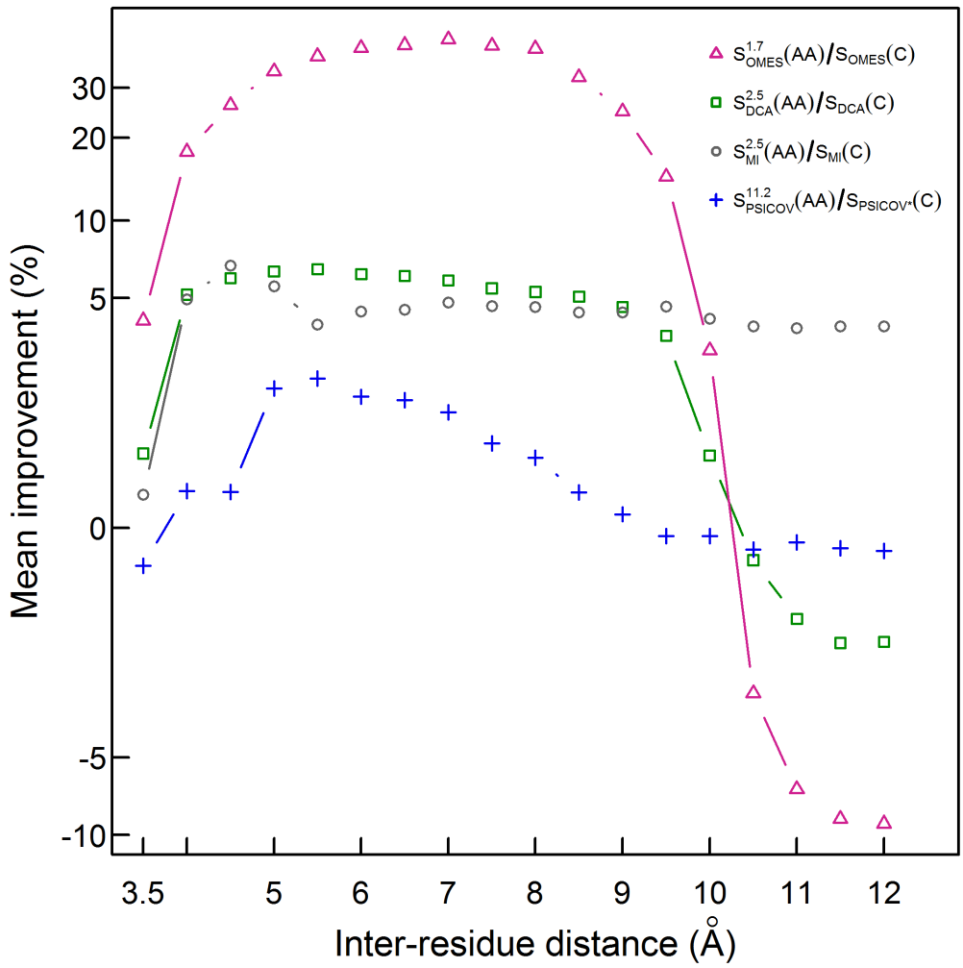
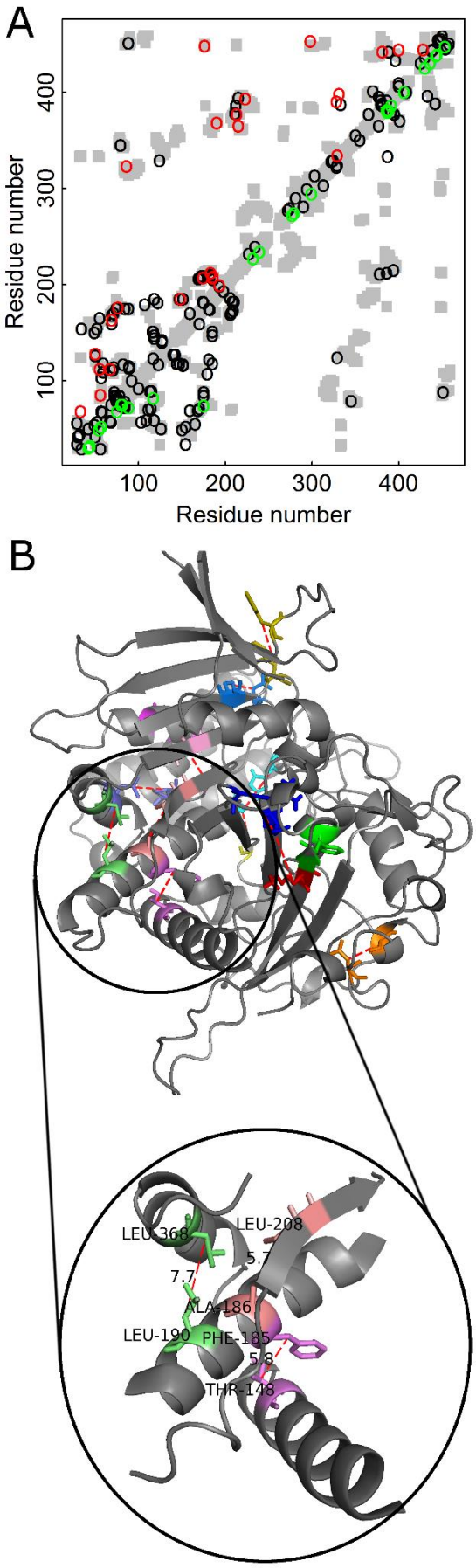


Figure 2.13 Improvement in contact prediction as a function of the distance used to define a physical contact.
The mean of the extent of improvement in contact prediction for 114 domains (or 86 in the case of PSICOV) is plotted as a function of the distance that must exist between two C_β atoms in different residues in order for them to be defined as being in contact. The extent of improvement was determined by calculating the difference in the areas under the curves of prediction accuracy vs. number of predictions by OMES, MI, DCA and PSICOV with and without incorporation of the codon data normalized by the area under the curve generated without codon data. The analysis was done for domains of length between 200 and 500 residues and at least 2000 coding sequences in their MSA. The contact predictions were made for the seven sequences with available crystal structures that have the highest resolution and that in all cases is < 3 Å.

2.3.5 Illustrative examples

The added value in combining amino acid and codon data can be illustrated for contact prediction by DCA in the case of Kex1Δp, a prohormone-processing carboxypeptidase from *Saccharomyces cerevisiae* that lacks the acidic domain and membrane-spanning portion of Kex1p. The crystal structure of Kex1Δp was solved at a resolution of 2.4 Å



(Shilton et al., 1997)and its MSA consists of 1,877 sequences. The predictions by DCA with or without incorporating codon data are shown in the respective top and bottom halves of the Kex1Δp contact map (Figure 2.14A). A comparison of the predictions by the two approaches shows that those made with incorporation of codon data are more long-range (in sequence) and more spread throughout the protein structure than those made without incorporation of codon data. Examples for such long-range contacts between different secondary structure elements in Kex1Δp that are predicted only when also the codon data is used include the interactions between Thr148 with Phe185, Ala186 with Leu208 and Leu190 with Leu368 (Figure 2.14B). This and other examples (Figure 2.15) show that incorporation of codon data can yield predictions of contacts between residues that are distant in sequence and are, thus, of more value for structure prediction.

Figure 2.14 Added value of combining amino acid and codon data in contact prediction by DCA illustrated for Kex1Δp.
A prohormone-processing carboxypeptidase from *S. cerevisiae*. (A) Contact map of the structure of Kex1Δp(PDB ID: 1AC5) in which all the contacts are shown as gray rectangles. Residues were defined as being in contact if at least one inter-atomic distance between their C_β atoms (C_α for glycine) is $\leq$ 8 Å. The top 100 predicted contacts made with or without incorporating

codon data are highlighted above (in red) and below (in green) the diagonal, respectively, and those predicted by both methods by black circles. (B) The crystal structure of Kex1Δp with predicted contacts highlighted. Only true predicted contacts that were not predicted by the original method are highlighted. Each contacting pair has a different color. The contacts were predicted using an MSA with 1,877 coding sequences with a length of 415 codons.

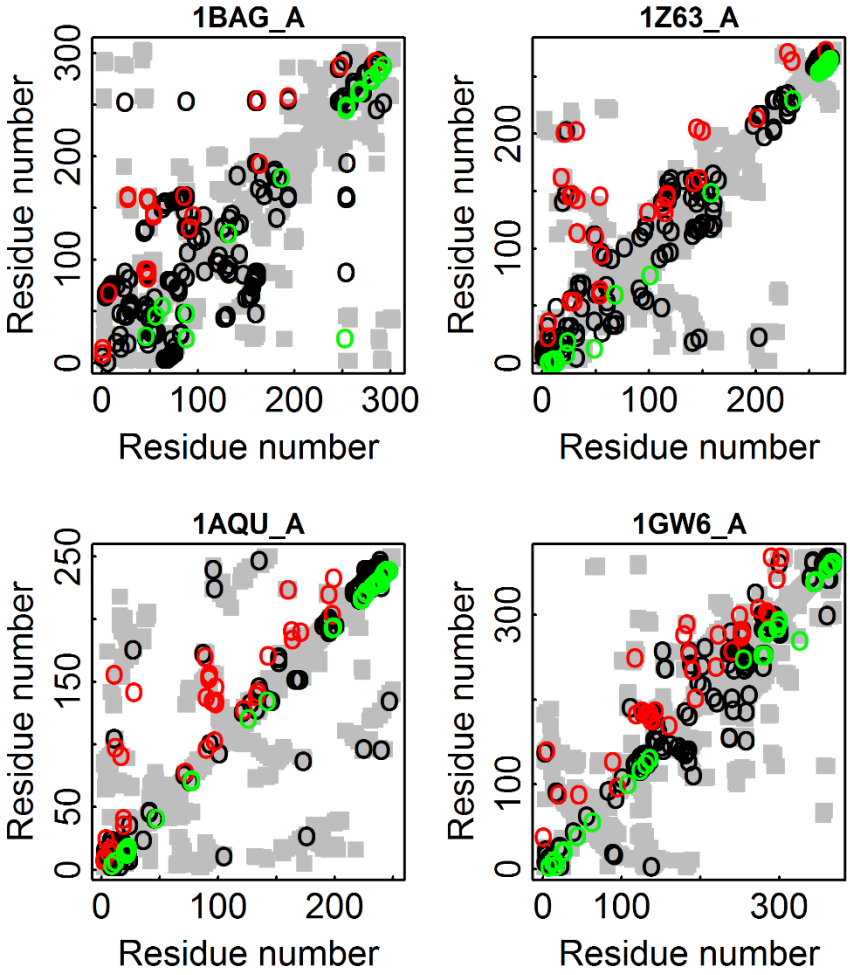


Figure 2.15 Illustration for four proteins of added value of combining amino acid and codon data in contact prediction by DCA.
All the contacts are shown as gray rectangles. Residues were defined as being in contact if at least one inter-atomic distance between their C_β atoms (C_α for glycine) is ≤ 8 Å. The top 100 predicted contacts made with or without incorporating codon data are highlighted above (in red) and below (in green) the diagonal, respectively, and those predicted by both methods by black circles. 1BAG_A - contact map of the structure of alpha-amylase from *Bacillus subtilis* (Pfam id: PF00128).1Z63_A - contact map of the structure of *Sulfolobussolfataricus* SWI2/SNF2 ATPase core (Pfam id: PF00176). 1AQU_A - contact map of the structure of mouse estrogen sulphotransferase (Pfam id: PF00685). 1GW6_A - contact map of the structure of human leukotriene A4 hydrolase (Pfam id: PF01433).

2.3.6 The potential value of codon information in other applications

The improvement in contact prediction when incorporating codon information in the analysis raised the question whether other applications in bioinformatics can benefit from it. An interesting application that also uses amino acid information in the form of an MSA is predicting whether a single nucleotide polymorphism (SNPs) is deleterious or non-deleterious. One example for such a method that was reported to perform well is Polyphen-2 (Adzhubei et al., 2013, 2010). This method is based on a machine learning approach that includes in its model a large number of features comprising phylogenetic and structural information characterizing the substitution., Here, I tested whether a much simpler model, based only on amino acid and codon MSA data, would be adequate. Thus, by using the same learning and test sets used in the development of the Polyphen-2 method (HumDiv and HumVar, see details in the methods section), we were able to demonstrate that incorporation of codon information in the prediction model significantly improves the performance of the predictions (Figures 2.16 and 2.17). The performance on the HumDiv test set was improved from an AUC of the receiver operating characteristic (ROC) curve with a value of 0.69 (based on amino acid MSAs only) to an AUC value of 0.75 (based on a model which combines both amino acids and codon information). The same trend is observed for the HumVar database (Figures 2.16 and 2.17).

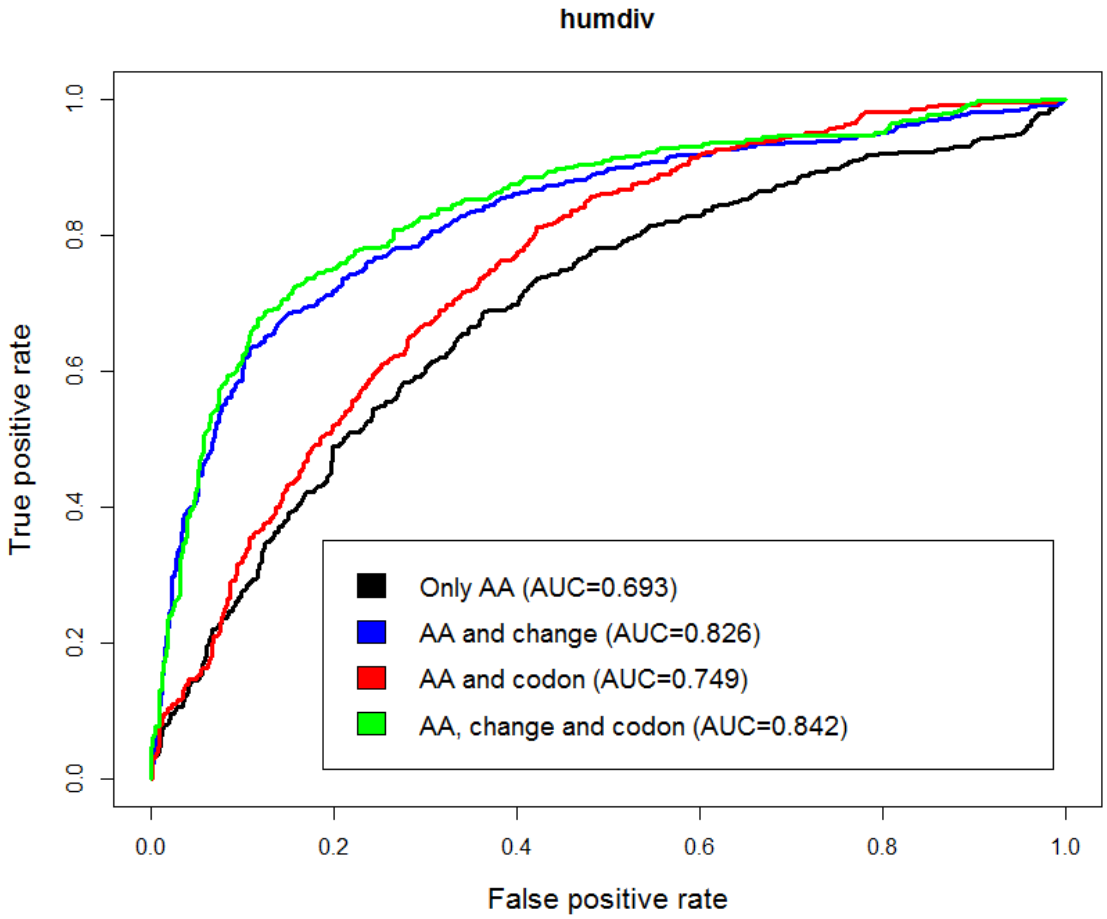


Figure 2.16 Performance evaluations of four different deleterious SNP predictors based on the HumDiv dataset.
ROC curves are plotted based on 2-fold cross validation. See methods below for details of the features. An additional independent variable incorporated into the prediction model was the type of mutation designated in the figure legend as "change".

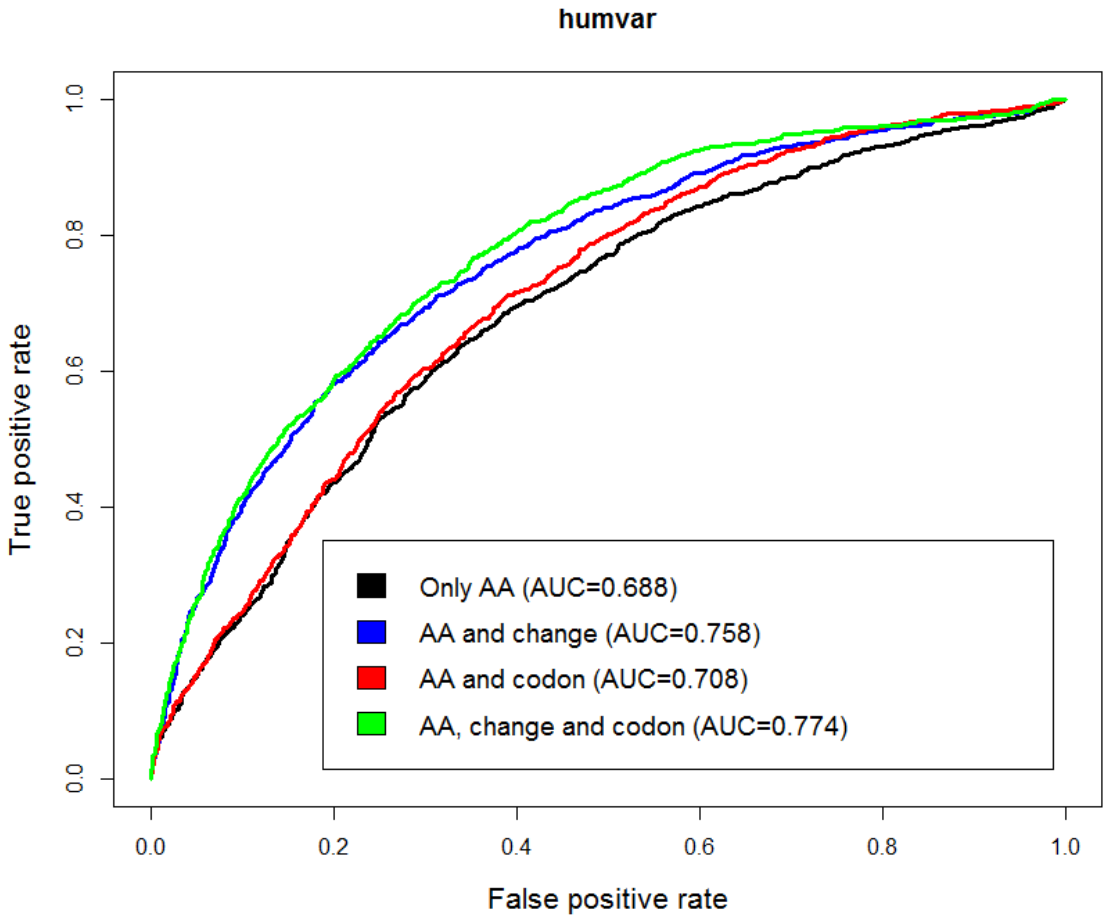


Figure 2.17 Performance evaluations of four different deleterious SNP predictors based on the HumVar dataset.
ROC curves are plotted based on 2-fold cross validation. See methods below for details of the features.

In the correlated mutations analysis, the amino acid and codon based scores had an opposite effects on the likelihood that a pair will have a physical contact. The same effect is observed when fitting a logistic regression model to the SNPs' data sets described above. Here, the coefficients of the codon and amino acid entropies have opposite signs as follows:
The fitted model based on HumDiv:

$$F(x) = \frac{1}{1 + e^{-(-0.56+1.84H(AA)-\mathbf{0.85}\boldsymbol{H}(\boldsymbol{Codon})+0.17H(AA)H(Codon)}}$$

The fitted model based on HumVar:

$$F(x) = \frac{1}{1 + e^{-(-1.06+1.35H(AA)-\mathbf{0.51}\boldsymbol{H}(\boldsymbol{Codon})+0.15H(AA)H(Codon)}}$$

where H is the entropy calculation for a SNP site in the MSA.

2.4 Summary

I have shown that improved contact prediction can be achieved by analysing both amino acid and codon MSAs together. The premise of my approach is that direct contacts are more likely to be present if the correlation at the amino acid level is high but at the codon level is low. Of particular importance, I find many cases where contacts between residues that are distant in sequence and, thus, of greatest value for structure prediction, are predicted only by using the combined method. It will be interesting to see whether our method succeeds better than other methods in contact prediction for a specific group of cases such as long proteins, long range interactions, inter-chain interactions and more.

With regards to other applications with which codon and amino acid information can be combined, I have shown that the prediction of deleterious SNPs can be improved using both codon and amino acid information. Interestingly, the opposite relations between amino acid and codon data also appeared in the regression model for the damaging SNP detection; a result that emphasizes the generality of my approach.

2.5 Discussion

2.5.1 False signals from phylogenetic bias and mRNA structures

High correlation at the codon level can be a consequence of a phylogenetic bias, a signal that results from functionally related clusters of residues that co-evolve according to the structure of the underlying tree. In such cases, the driving force for co-variations are at the codon level and the resulting correlations at the amino acid level do not reflect a true physical contact. Alternatively, when the main factors are direct physical contacts, the effect will be at the amino acid level, and as long as the variations at the codon level remain synonymous, their impact will be negligible. This will result in a high correlation value at the amino acid level and low correlation at the codon level.

Another factor that can influence the correlation values at the codon level, is the formation of mRNA secondary and tertiary structures (Katz and Burge, 2003). mRNA structures are widespread around coding regions (Mortimer et al., 2014; Wan et al., 2014) and in some cases are linked to translational regulation (Katz and Burge, 2003). In these cases, synonymous mutations can have a direct effect on the mRNA structure. One example for such a scenario is when a stable secondary structure protects the mRNA sequence from degradation. In this situation base pairing preservation in stem regions requires the

selection of nucleotides at synonymous sites (Katz and Burge, 2003). As a consequence, the correlation values at the codon level will be high and the correlation at the amino acid level will not reflect a true inter-residue contact. Therefore, a combined score as described here will detect that it is not a true contact. Another example concerns ribosome pausing and translation efficiency. The rate of translation in many proteins, which can greatly vary across transcripts, influences the protein folding pathway (Komar, 2009; Shah et al., 2013; Wolin and Walter, 1988). RNA structures can have a profound effect on translation rates, since a highly structured RNA region can cause ribosome pausing, which may facilitate the folding of individual domains (Figure 2.18) (Dana and Tuller, 2012; Meyer and Miklós, 2005; Wen et al., 2008). The above two factors, phylogenetic bias and mRNA secondary and tertiary structure formation, are important sources of false signals that might be detected at the codon level, and when explicitly combined with the amino acid information, can significantly reduce false positive contact predictions.

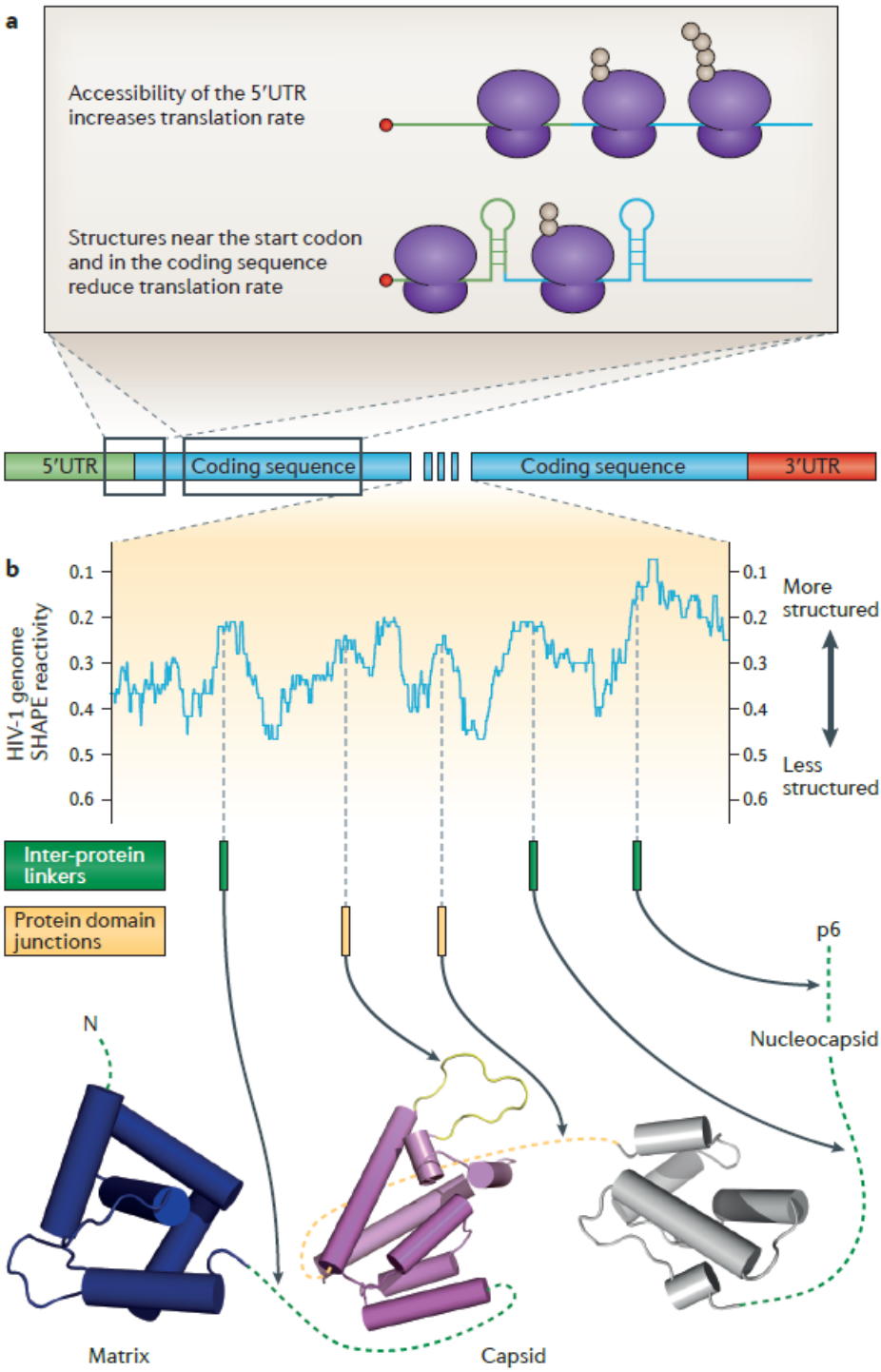


Figure 2.18 Structure around start codons and translational efficiency.
(a) Accessibility of the 5′ untranslated region (UTR) increases translation rate due to decreased structure in this region that allows efficient ribosome binding and start codon scanning. (b) The RNA folding energy of different segments of the coding region is associated with protein structure. The increased structure of these regions promotes ribosome pausing and assists in protein domain folding. The protein domains shown are Protein Data Bank identifiers 2GOL and 1A43. SHAPE, selective 2′-hydroxyl acylation analysed by primer extension. The figure and caption is taken from previous work of others (Mortimer et al., 2014).

2.5.2 Extensions and future work

The score I propose can be used in conjunction with different methods of CMA; however, other possible scores or model refinements should be examined in future work. For example, a refinement of the combined score in this thesis could be done with additional data sources such as ribosome profiling, RNA secondary structure predictions, and more.

As discussed earlier, the incorporation of codon information in contact prediction involves dividing the amino acid based score (generated by the CMA) by the codon based score. Therefore, two amino acids are more likely to be predicted in contact if their amino acid based score is high and their codon based score is low. Interestingly, in the regression model fitted for the deleterious SNP prediction, the predictors, i.e. the conservation scores based on the amino acid and the codon information also have opposite effects on the odds that a SNP will be damaging. This similarity in the relationship between codon and amino acid based scores in these two examples suggests that incorporation of codon information may have a wide range of applications.

With regards to the regularization of a codon covariance matrix (pseudo-count weights of codons), it might be that additional optimizations could be done, as with regards to the underlying dependencies in the codon table.

The potential of other applications which combine amino acid and codon MSAs as part of their analysis should be tested, such as predicting protein-protein interactions and, more generally, in feature selection in machine learning.

Chapter 3

A Mechanism for Prevention of Aggregation of Neighboring Domains

3.1 Introduction

Protein domains play a major role in evolution and are occasionally referred to as the building blocks of proteins. The term "domain" has been employed in different ways over the years. Here, it is defined as an evolutionary unit whose coding sequence (typically corresponding to 100-250 residues) can be duplicated or undergo recombination (Chothia et al., 2003). In addition, domains usually have an independent function either alone or with other domains (Vogel et al., 2004) and a compact structure that folds independently (Figure 3.1). Nearly half of all proteomes and more than 70% of all eukaryotic proteins are multi-domain proteins. About 95% of multidomain proteins contain 2–5 domains and their combinations follow a power law distribution, i.e, a small number of domains recombine with many different partners whereas most domains are found only in combination with a few other partner domains (Han et al., 2007).

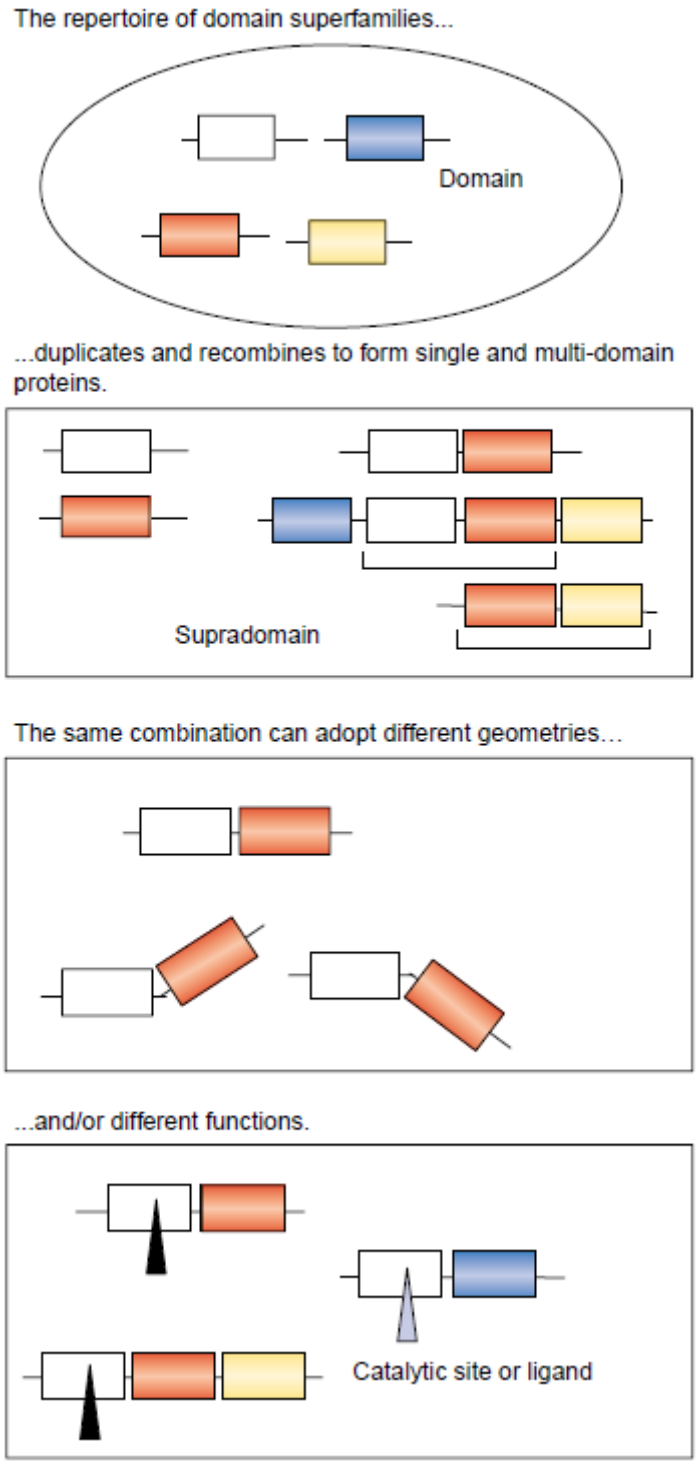


Figure 3.1 The role of domains as building blocks of proteins.
Domains form different multi-domain proteins by duplication and recombination. Domains belonging to the same superfamily are represented as rectangles in the same color. Various domain combinations in a certain order (that is, supradomains) can form functional units that are reused in different protein contexts. This figure is taken from previous work of others (Vogel et al., 2004).

In general, domains are classified into families based on sequence, structure or function. The expansion of the PDB in the mid-90's, inspired the development of several protein domain classifications (Mizuguchi et al., 1998; Murzin et al., 1995; Orengo et al., 1997; Siddiqui et al., 2001) that are hierarchical. For instance, in a protein domain classification database called CATH, structures are first divided into their constituent domains and then classified at four major levels: (C)lass, (A)rchitecture, (T)opology or fold, and (H)omologous superfamily. SCOP, another domain classification database, employs similar categories (fold, superfamily, family and domain) with some differences in the classification process (Cuff et al., 2009) (Figure 3.2).

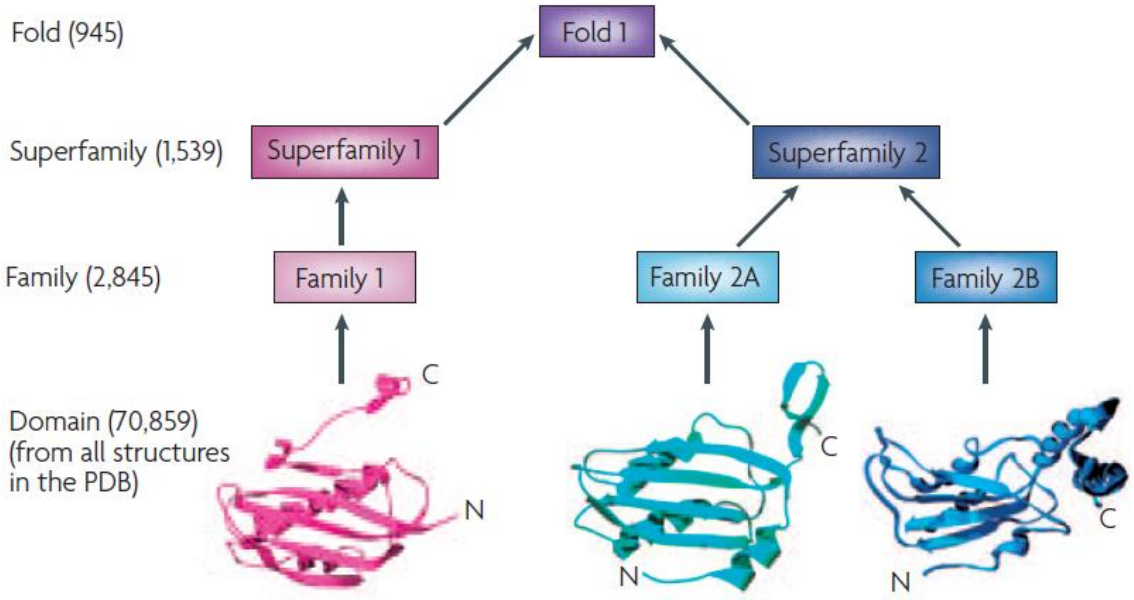


Figure 3.2 Hierarchal classification of protein domain families.
This figure is taken from previous work of others (Han et al., 2007).

The 3D structures of most proteins are not known and, therefore, using the structure-based assignment of a domain as described above (SCOP or CATH) is not possible and an alternative sequence-based domain definition is required. Pfam, the widely used database of protein families, (Finn et al., 2014; Punta et al., 2012) is a comprehensive source for domain families which is based on sequence alone. In this database, families are sets of regions in proteins that share a significant degree of sequence similarity (i.e. homologous sequences). A multiple sequence alignment (MSA) of each family of homologous sequences can be formed and turned into a position-specific scoring system based on a profile hidden Markov model (HMM). Then, the profile HMM, one for each protein family, can be used for searching sequence databases (e.g. UniprotKB) for remotely homologous sequences (Eddy,

1998). Similarities between sequences in Pfam are detected using HMMER3 software tool. Despite the fundamental difference between sequence-based and structure-based domain family databases, many of the former (that is, Pfam families) can be related to the latter (that is, SCOP families) (Pandit et al., 2002). Nevertheless, a significant difference in domain definitions between Pfam and SCOP and CATH is due to discontinuous domains (a discontinuous domain is one where the linear sequence of the domain is interrupted by another inserted domain) (Bateman et al., 2004). In this thesis, multidomain proteins are considered using both the sequence-based Pfam definition described above and the SCOP or CATH structural definitions (only consecutive domains are taken into account), depending on the type of analysis.

Multi-domain proteins are potentially more aggregation-prone owing to the high effective protein concentration near each domain (Han et al., 2007). Aggregation of misfolded proteins is associated with many diseases such as Alzheimer's disease and type II diabetes (Luheshi and Dobson, 2009; Selkoe, 2003). Protein misfolding is also harmful to cells owing to the energetic costs involved in the synthesis and degradation of non-functional proteins and the lack of folded protein molecules that may have essential functional roles (Figure 3.3). Hence, it is not surprising that evidence for strong selection against misfolding has been found in all kingdoms of life (Dill et al., 2011; Drummond and Wilke, 2008). Previous work has shown that selection against mis-folding is reflected in various correlations between measures of protein abundance and the probability of generating mis-folded proteins upon translation and folding (Drummond and Wilke, 2008; Tartaglia and Vendruscolo, 2009). For example, it has been suggested that the observed correlation between protein abundance and optimal codon usage reflects selection against mis-folding (Drummond and Wilke, 2008). A similar explanation has been offered for the inverse correlation between mRNA expression levels and predicted protein aggregation propensities (Tartaglia and Vendruscolo, 2009). Preventing aggregation is particularly significant in the case of multi-domain proteins since as stated earlier, they account for nearly half of all proteomes (and more than 70% of all eukaryotic proteins) and are potentially more aggregation-prone owing to the high effective protein concentration near each domain (Han et al., 2007). Mechanisms for preventing multidomain proteins from mis-folding are not yet fully understood (Borgia et al., 2011). One likely mechanism is co-translational folding that can help to prevent aggregation of a newly synthesized domain with other domains in the same polypeptide chain that were synthesized first (Netzer and Hartl, 1997). Domain-by-domain folding can also be facilitated by controlled domain-by-

domain release from chaperones (Jacob et al., 2007; Rivenzon-Segal et al., 2005). Aggregation (not just of multi-domain proteins) can also be prevented by increasing folding rates and decreasing unfolding rates (Batey et al., 2006; Oberhauser et al., 1999) . Finally, it has been suggested that aggregation of multi-domain proteins is also minimized by selection for neighboring domains with low sequence identity (Wright et al., 2005).. In this thesis, I suggest a previously unrecognized mechanism for preventing aggregation which is based on protein length.

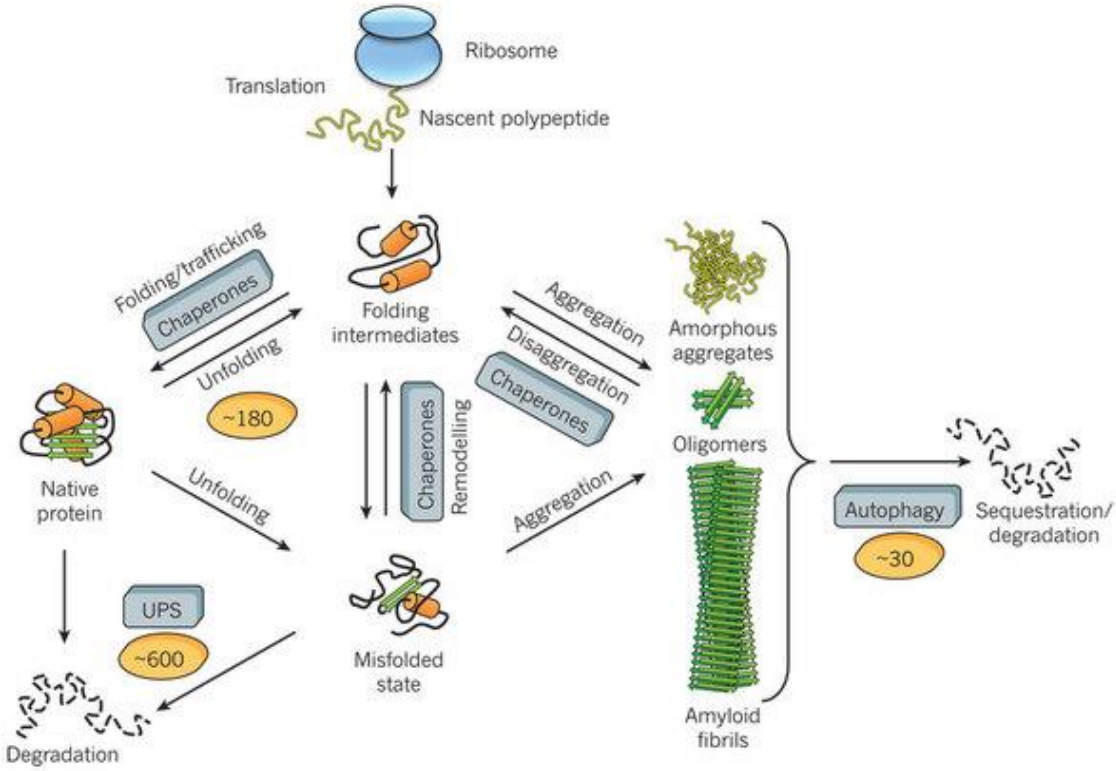


Figure 3.3 Protein life time from synthesis to degradation.
The figure describes the proteostasis network which integrates pathways for the folding of newly synthesized proteins, refolding of misfolded states and disaggregation with protein degradation mediated by the ubiquitin-proteasome system and the autophagy system. This figure is taken from previous work of others (Hartl et al., 2011).

Many properties of proteins depend on their chain length (Dill et al., 2011; Thirumalai et al., 2010). Protein folding rates, for example, are known to be inversely correlated with chain length. I found that there is a very significant tendency for N-terminal domains in double-domain proteins to be shorter than their neighboring C-terminal domains. A possible explanation for this observation, given the negative correlation between folding rates and protein length (Galzitskaya et al., 2003; Thirumalai et al., 2010), is that there is selection for N-terminal domains to fold faster than their C-terminal

counterparts. In addition to protein length, folding rates have also been found to be inversely correlated with absolute contact order (ACO), i.e. the average separation in sequence between residues that are in contact in the folded structure (Galzitskaya et al., 2003; Plaxco et al., 1998). Independent support for the existence of selection for faster folding N-terminal domains is, therefore, provided here by showing that the ACO values of N-terminal domains in two-domain proteins with available three-dimensional structures tend to be lower than those of their respective C-terminal neighbors. I, therefore, reasoned that if the bias for two-domain proteins with a faster folding N-terminal domain is due to selection against protein mis-folding then proteins with a shorter N-terminal domain should be more abundant than those with a shorter C-terminal domain as indeed I found to be the case. Taken together, the findings presented in this thesis suggest the existence of a previously unrecognized mechanism for prevention of aggregation of neighboring domains.

3.2 Methods

3.2.1 Construction of datasets of two-domain proteins

Pfam (Finn et al., 2014) domains sequence assignments (release 26.0) from Swissprot were downloaded from ftp://ftp.sanger.ac.uk/pub/databases/Pfam/current_release/. Only protein sequences that were assigned two consecutive Pfam domains (each formed by a continuous sequence of 50 to 200 amino acids and connected by a linker that is shorter than 30 amino acids) with no additional nested or overlapping domain assignments were included in the database. This dataset comprises 32,567 proteins from 3,995 different organisms. Evidence for existence at the protein level was taken from SwissProt annotations (UniProt Consortium, 2012) in the file downloaded from ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/complete/uniprot_sprot.fasta.gz and is available for 2,964 of these proteins. A non-redundant set of 6,739 two-domain proteins was created by intersecting our dataset of 32,567 proteins with that of UniRef (Suzek et al., 2007) that was downloaded from ftp://ftp.uniprot.org/pub/databases/uniprot/uniref/uniref50/uniref50.fasta.gz using a redundancy cutoff of 50%.

3.2.2 Contact order analysis

Two databases of structural classification of proteins were used in the analysis: (i) version 1.75A of SCOP (Andreeva et al., 2008; Murzin et al., 1995) that was downloaded from http://scop.berkeley.edu/astral; and (ii) version 3.4 of CATH (Orengo et al., 1997; Pearl, 2003) that was downloaded from http://release.cathdb.info. Only proteins that contain two domains belonging to the same family (the lowest level in the structural hierarchy as defined by CATH and SCOP) were included in the analysis. In addition, I considered only proteins in which the length of each domain is between 50 and 300 residues and where the combined lengths of the two domains is > 80% of the length of the PDB entry and that the length of the linker is less than 30 amino acids. In cases where different two-domain proteins contain the same domain, I required in order to avoid redundancy that the non-shared domains differ in sequence by at least 5% (using other cutoffs did not alter the results). This process yielded 454 entries for 174 domain families in SCOP and 1247 entries (808 of which belong to the immunoglobulins) for 92 domain families in CATH. Data for families with more than one member were included in the analysis using their average so that large families (e.g. the immunoglobulins) would not be overrepresented.

ACO was calculated as described (Galzitskaya et al., 2003; Plaxco et al., 1998) using the script downloaded from http://depts.washington.edu/bakerpg/contact_order/contactOrder.pl (written by Erik Alm). ACO is the average sequence separation between contacting residues in the native structure and is given by:

$$ACO = \frac{1}{N}\sum_{i,j\in N, i>j} \Delta S_{i,j},$$

where N is the number of contacts in the native structure and $\Delta S_{i,j}$ is the number of amino acids between residues i and j that are in contact. The relative contact order (RCO) is equal to ACO/L where L is the length of the protein.

3.2.3 Protein abundance analysis

Protein abundance analysis was carried out using data downloaded from http://pax-db.org (Wang et al., 2012b), release 2.1, for 1,699 two-domain proteins in the Pfam database from 12 different organisms.. The abundance data is expressed as parts per million (ppm), *i.e.* the abundance of each protein is quantified relative to those of all other protein

molecules in the sample. Data for samples from different sources can, therefore, be compared.

3.3 Results

3.3.1 N-terminal domains in two-domain proteins tend to be shorter than C-terminal domains

Density plots generated for the chain lengths of N- and C-terminal domains in 2964 two-domain proteins in the SwissProt database show a clear bias for N-terminal domains to be shorter than C-terminal domains (Figure 3.4A). It is not possible, however, to determine from these plots to what extent, if any, the bias is influenced by domain pairing, i.e. the tendency of the N-terminal domain in two-domain proteins to be shorter than its neighboring C-terminal domain. We, therefore, decided to compare the bias in the 2964 real two-domain proteins with the biases in 10,000 sets of 2964 randomly chosen domain pairs generated by shuffling the N-terminal domains of the real proteins while keeping the C-terminal domains in place. A histogram of these biases shows that the bias in the real two-domain proteins is significantly larger than in any of the sets of randomly generated domain pairs (Figure 3.4B). This analysis shows that the domain pairing in real proteins increases the bias much beyond what is expected given that, in general, N-terminal domains tend to be shorter than C-terminal domains (Figure 3.4A). In other words, there appears to be selective pressure for N-terminal domains in double-domain proteins to be shorter than their C-terminal counterparts.

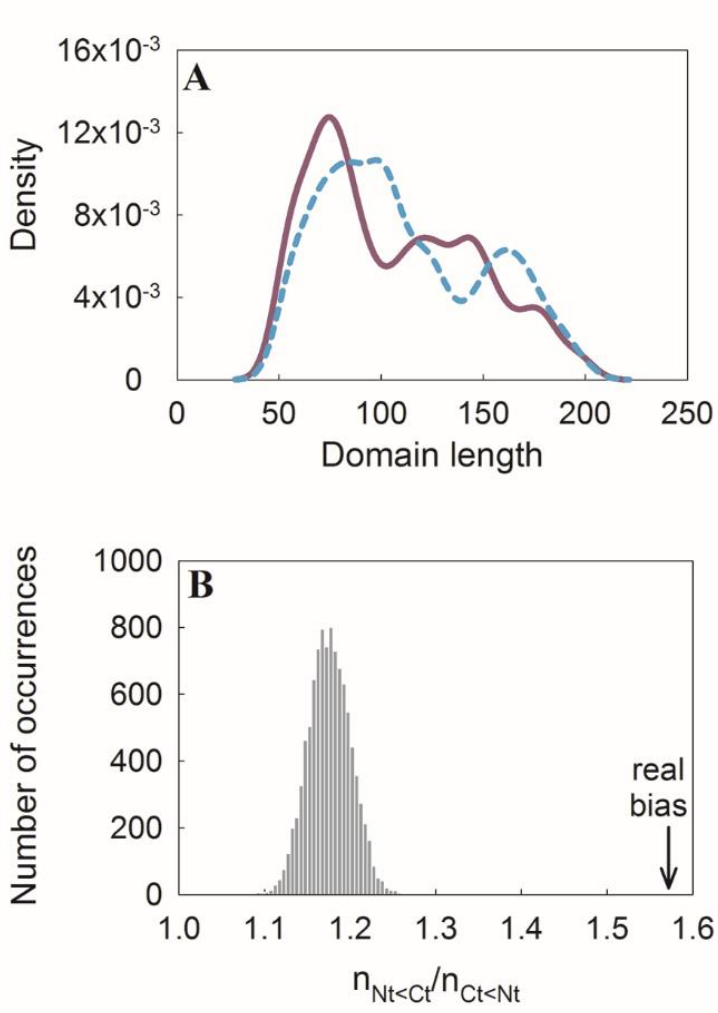


Figure 3.4 Distribution of chain lengths of N- and C-terminal domains in two-domain proteins.
(A) Density plots of the chain length distributions of N- (purple) and C-terminal (turquoise) domains in double-domain proteins shows that the tendency for C-terminal domains to be longer is significant with a Wilcoxon rank sum test (two sided) P-value of 8.3×10^{-8}. The analysis is based on 2964 two-domain proteins in the SwissProt database that (i) comprise domains of length between 50 to 200 amino acids connected by a linker that is shorter than 30 amino acids and (ii) for which there is evidence at the protein level. (B) Histogram showing the bias for shorter N-rerminal domains in real two-domain proteins and in two-domain proteins comprising randomly chosen domain pairs generated by shuffling the N-terminal domains of the real proteins while keeping the C-terminal domains in place. The bias, which corresponds to the number of proteins with shorter N-terminal domains, $n_{Nt<Ct}$, divided by the the number of proteins with shorter C-terminal domains, $n_{Ct<Nt}$, was calculated for the 2964 real two-domain proteins (arrow) and for 10,000 sets of 2964 randomly chosen domain pairs (gray bars).

Next, we asked whether the bias seen in Figure 3.4A is general or limited to certain datasets. The data in Table 3.1 show that the bias is found in both eukaryotic and prokaryotic proteins. It is also seen when the analysis was carried out for all the Pfam entries in SwissProt and for the Uniref50 nonredundant prokaryotic, eukaryotic and

combined datasets. Finally, the bias was also observed in datasets of proteins from very different organisms such as *E. coli* and fly. The bias for N-terminal domains in two-domain proteins to be shorter than their C-terminal counterparts is, therefore, found to be ubiquitous but is stronger in prokaryotic proteins.

Protein Data Set	$n_{Nt<Ct}$	$n_{Ct<Nt}$	$n_{Nt<Ct}/n_{Ct<Nt}$
All proteins for which there is evidence at protein level	1,757	1,116	1.57
All eukaryotic proteins for which there is evidence at protein level	1,139	739	1.54
All prokaryotic proteins for which there is evidence at protein level	521	311	1.68
All Pfam entries in Swiss-Prot	18,763	13,198	1.42
UniRef50 nonredundant set of proteins	3,747	2,871	1.31
UniRef50 nonredundant set of eukaryotic proteins	1,220	1,016	1.20
UniRef50 nonredundant set of prokaryotic proteins	2,148	1,599	1.34
Representative Organisms			
Human (*H. sapiens*)	339	309	1.10
Mouse (*M. musculus*)	298	280	1.06
Arabidopsis thaliana	327	193	1.69
Yeast (*S. cerevisiae*)	111	98	1.13
Escherichia coli	166	100	1.66
Fly (*D. melanogaster*)	54	34	1.59
Worm (*C. elegans*)	78	59	1.32
Methanocaldococcus jannaschii	59	37	1.59

Table 3.1 Number of Two-Domain Proteins with Shorter N- or C-Terminal Domains in Different Protein Data Sets.
The table is based on proteins in Swiss-Prot with two domains of length between 50 and 200 amino acids connected by a linker that is less than 30 amino acids. Proteins with additional nested or overlapping Pfam domain indications were excluded.

3.3.2 N-terminal domains in two-domain proteins are predicted to fold faster than C-terminal domains

The inverse dependence between folding rate and chain length (Galzitskaya et al., 2003; Thirumalai et al., 2010) suggests that N-terminal domains in double-domain proteins are selected to be shorter than their C-terminal counterparts so that they fold faster. Given that the detailed folding kinetics of most proteins are not known, we decided to test this idea using ACO as a predictor of the relative folding rates of the individual domains in double-domain proteins. We restricted our analysis to two-domain proteins with known structure in which both domains belong to the same family (as defined by the CATH (Orengo et al., 1997) and SCOP (Murzin et al., 1995) databases) so that the strong dependence of ACO and chain length on topology would not mask a signal that arises from the domain order. The analysis was carried out using both CATH and SCOP in order to ensure that the ACO

values that are calculated separately for each domain do not depend on the choice of domain boundaries that may differ in the two databases. We also required that each domain is formed by a continuous sequence of 50 to 300 residues and is, thus, in the range where ACO and chain length were shown to have predictive value. Finally, we only considered two-domain proteins in which the combined lengths of the two domains is >80% of the length of the full protein and the linker connecting the two domains is less than 30 amino acids. Data for families with more than one member were included in the analysis using their average so that large families would not be overrepresented.

A significant tendency is observed for the ACO values of the N-terminal domains in two-domain proteins (satisfying the criteria described above) to be smaller than those of their neighboring C-terminal domains (Figure 3.5). The values of the ratio between the number of all two-domain proteins in SCOP and CATH with a predicted faster folding N-terminal domain and the number of all those with a predicted faster folding C-terminal domain ($n_{ACO(Nt)<ACO(Ct)}/n_{ACO(Ct)<ACO(Nt)}$) are 1.4 and 1.7, respectively, with respective binomial test P-values of 0.04 and 0.016. This tendency is observed for all domain classes (α, β, α/β and $\alpha+\beta$) in both SCOP and CATH. The values of $n_{ACO(Nt)<ACO(Ct)}/n_{ACO(Ct)<ACO(Nt)}$ are 1.7, 1.5 and 1.8 for the 19, 28 and 44 respective members of the α, β, α/β and $\alpha+\beta$ classes in CATH and 1.8, 1.3 and 1.4 for the 31, 46 and 88 members of these classes in SCOP. Importantly, the bias for ACO values of the N-terminal domains in two-domain proteins to be smaller than those of their neighboring C-terminal domains is not due to differences in domain lengths as it is observed also for proteins with domains of similar size (Table 3.2). For example, the values of the ratio $n_{ACO(Nt)<ACO(Ct)}/n_{ACO(Ct)<ACO(Nt)}$ for all the two-domain proteins in CATH and SCOP, when those with a difference of more than ten amino acids in their domain lengths were excluded from the analysis, are 1.6 and 1.4, respectively. The corresponding *P*-values of 0.053 and 0.078 are, however, somewhat higher owing to the smaller sizes of the datasets when only two-domain proteins comprising domains with similar lengths are considered. We also calculated Fisher exact test of independence P-values to determine to what extent ACO values contain information beyond that which is provided by domain length. The respective Fisher exact test P-values of 0.32 and 0.798 for the case above indicate that the bias in ACO values is not due to differences in domain lengths (Table 3.2).

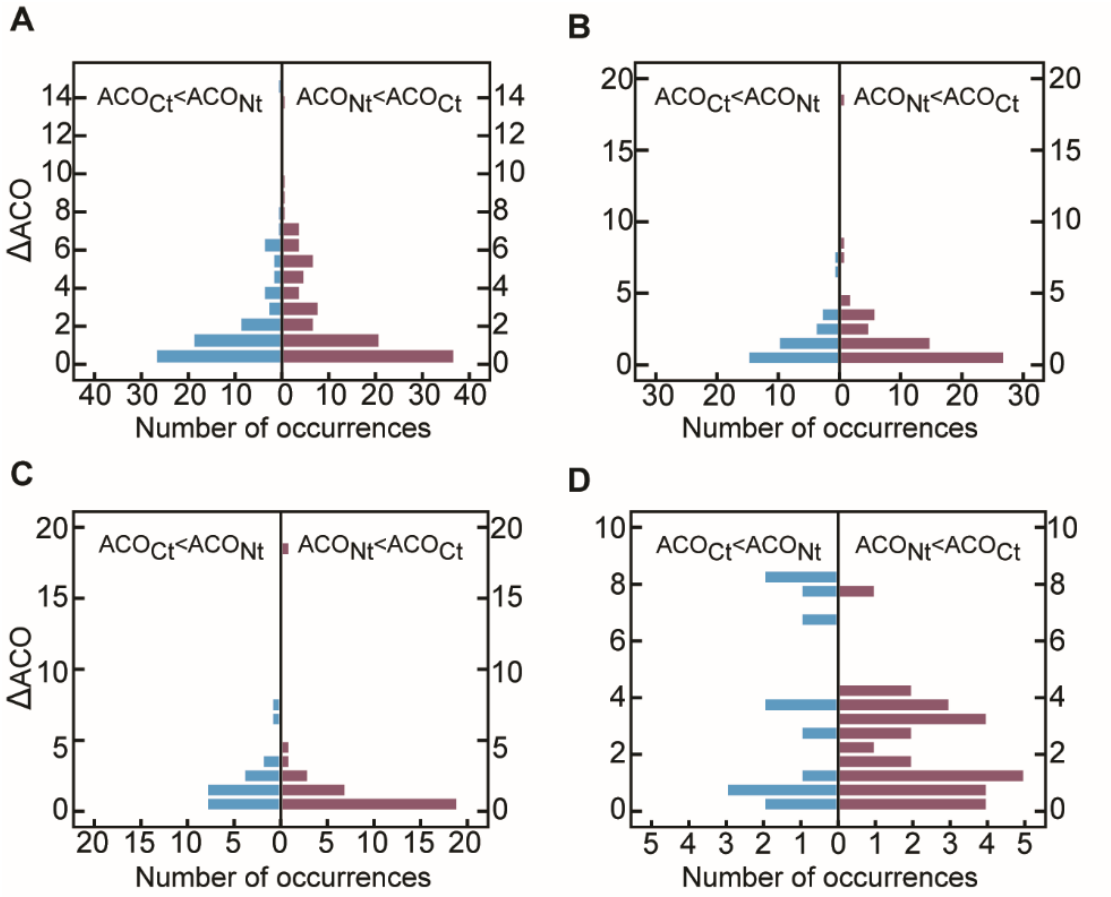


Figure 3.5 Distribution of the differences in absolute contact order (ACO) values of the N- and C-terminal domains in proteins with two domains that belong to the same family.

(A) Two-domain proteins in SCOP, (B) two-domain proteins in CATH, (C) two-domain eukaryotic proteins in CATH, and (D) two-domain prokaryotic proteins in CATH. The proteins were binned so that the frequency of two-domain proteins with a positive value of ΔACO (i.e. $ACO_{Nt} < ACO_{Ct}$) in a certain range can be compared with the frequency of two-domain proteins with a negative value of ΔACO (i.e. $ACO_{Nt} > ACO_{Ct}$) in the same range of absolute values. The frequency of two-domain proteins with a positive value of ΔACO in a certain range is nearly always found to be greater than the frequency of two-domain proteins with a negative value of ΔACO in the same range of absolute values. In the SCOP database (A), there are 101 and 73 proteins for which $ACO_{Nt} < ACO_{Ct}$ and $ACO_{Nt} > ACO_{Ct}$, respectively (binomial two-sided test *P*-value of 0.04). In the CATH database (B), there are 58 such proteins for which the ACO value of the N-terminal domain is smaller than that of its neighbouring C-terminal domain ($ACO_{Nt} < ACO_{Ct}$) and 34 proteins for which $ACO_{Nt} > ACO_{Ct}$ (binomial two-sided test *P*-value of 0.02). Such a tendency is observed for all thresholds of non-redundancy analyzed in SCOP and CATH.

In the case of relative contact order (RCO) calculations (see Methods) for two-domain proteins with a difference of less than ten amino acids in their domain lengths, the values of the ratio $n_{RCO(Nt)<RCO(Ct)}/n_{RCO(Ct)<RCO(Nt)}$ for the two-domain proteins in CATH and SCOP are, as expected, similar to the corresponding values of $n_{ACO(Nt)<ACO(Ct)}/n_{ACO(Ct)<ACO(Nt)}$ but

the dependence of the bias on domain length is greater as reflected in the respective Fisher exact test P-values of 0.1 and 0.077 (Table 3.2). In summary, therefore, two predictors of folding rate, domain length and ACO, indicate independently of each other that N-terminal domains in two-domain protein tend to fold faster than their neighboring C-terminal domains. The bias for positive values of ΔACO is found to be significantly greater in the two-domain proteins in CATH from prokaryotes (D) than in those from eukaryotes (C). See Figure 6 for additional data.

SCOP

	ACO			**RCO**		
Maximum difference between domain lengths	$n_{ACO(Nt)<ACO(Ct)}/n_{ACO(Ct)<ACO(Nt)}$	Fisher exact test P-value	Binomial test P-value	$n_{RCO(Nt)<RCO(Ct)}/n_{RCO(Ct)<RCO(Nt)}$	Fisher exact test P-value	Binomial test P-value
10	62/43	0.32	0.078	63/42	0.1	0.05
15	72/47	0.13	0.027	75/44	0.12	0.005
20	75/53	0.046	0.063	77/51	0.13	0.026
30	86/59	0.005	0.030	82/63	0.16	0.134
40	90/65	0.005	0.053	88/67	0.046	0.107
50	90/69	0.009	0.112	88/71	0.03	0.204

CATH

	ACO			**RCO**		
Maximum difference between domain lengths	$n_{ACO(Nt)<ACO(Ct)}/n_{ACO(Ct)<ACO(Nt)}$	Fisher exact test P-value	Binomial test P-value	$n_{RCO(Nt)<RCO(Ct)}/n_{RCO(Ct)<RCO(Nt)}$	Fisher exact test P-value	Binomial test P-value
10	43/26	0.798	0.053	41/28	0.077	0.148
15	41/33	0.228	0.41	46/28	0.08	0.047
20	45/34	0.360	0.26	48/31	0.03	0.071
30	46/35	0.103	0.266	44/37	0.011	0.505
40	54/35	0.069	0.055	48/41	0.001	0.525
50	56/33	0.039	0.019	48/41	0.004	0.525

Table 3.2 Summary of statistics for the relative (RCO) and absolute (ACO) contact order values for two-domain proteins in which the difference in the lengths of the N- and C-terminal domains is restricted[a].

[a]Fisher exact test is used to calculate the likelihood that a bias in ACO or RCO values is independent of domain length (i.e. $P > 0.05$ indicates that the two measurements are independent). The probability of the indicated bias to happen by chance is calculated using the binomial two-sided test.

3.3.3 Bias for faster folding N-terminal domains is greater in prokaryotes than in eukaryotes

The tendency for N-terminal domains to be predicted as faster folders than their C-terminal neighboring domains is found to be greater in prokaryotes than in eukaryotes (Figures 3.5C,D and 3.6). In 28 families in CATH that comprise prokaryotic proteins, the ACO values of the N-terminal domains are smaller than those of their respective C-terminal neighboring domains whereas only in 13 families the opposite is found (binomial two-sided test P-value of 0.03). In eukaryotic families, the ACO values of the N-terminal terminals are smaller than those of the neighboring C-terminal domains in 32 families in CATH whereas in 24 families the opposite is seen and while the trend is, therefore, maintained it is not significant statistically (binomial two-sided test P-value of 0.35). This difference between prokaryotes and eukaryotes is also seen in the two-domain proteins in SCOP (Figure 3.6). The bias for N-terminal domains in two-domain proteins to be shorter than their C-terminal counterparts was also found to be stronger in prokaryotic proteins (Table 3.1). Taken together, therefore, the data indicate that selection for N-terminal domains to fold faster than their C-terminal neighboring domains is much stronger in prokaryotes than in eukaryotes.

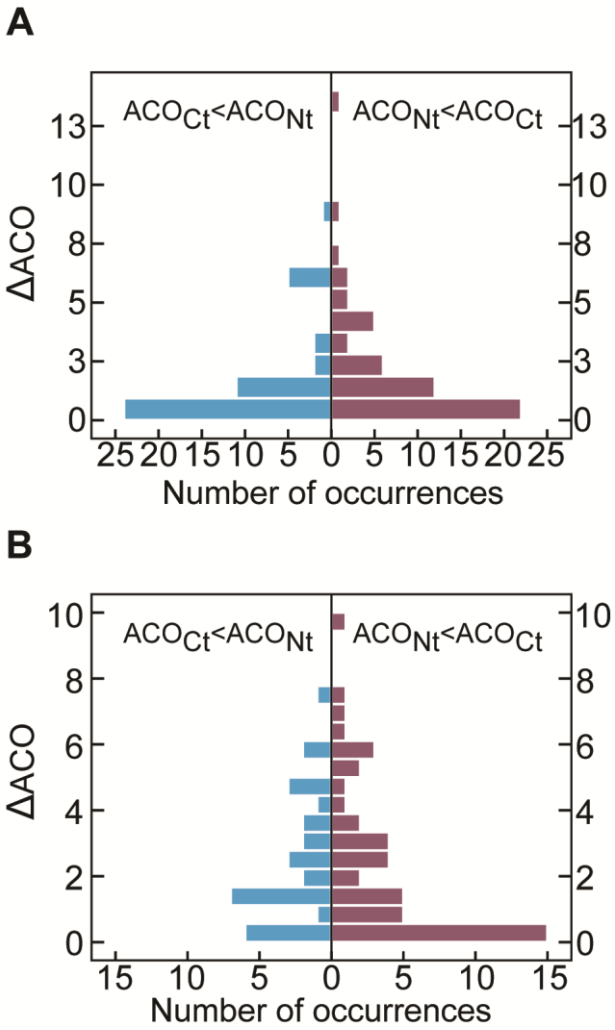


Figure 3.6 Distribution of the differences in absolute contact order (ACO) values of the N- and C-terminal domains in two-domain proteins in SCOP that belong to the same family in prokaryotes (A) versus eukaryotes (B).

The frequency of two-domain proteins with a positive value of ΔACO (i.e. $ACO_{Nt} < ACO_{Ct}$) in a certain range is compared with the frequency of two-domain proteins with a negative value of ΔACO (i.e. $ACO_{Nt} > ACO_{Ct}$) in the same range of absolute values. The frequency of two-domain proteins with a positive value of ΔACO is found to be significantly greater than the frequency of two-domain proteins with a negative value of ΔACO in prokaryotes than in eukaryotes.

3.3.4 Two-domain proteins with an N-terminal domain that is shorter than its neighboring C-terminal domain are more abundant

Given that proteins with a propensity to mis-fold tend to be less abundant (Drummond and Wilke, 2008; Tartaglia and Vendruscolo, 2009), we reasoned that two-domain proteins with a shorter N-terminal domain should be more abundant than those with a shorter C-terminal domain if this bias reflects selection against mis-folding. Surprisingly, the overall lengths of two-domain proteins with a shorter N-terminal domain tend to be less than those of two-domain proteins with a shorter C-terminal domain (Figure 3.7). We decided, therefore, to compare the abundances of two-domain proteins with different domain lengths but with a similar overall chain length. Strikingly, we find that two-domain proteins with an N-terminal domain that is shorter than the C-terminal domain are more abundant than two-domain proteins with similar overall chain length but with shorter C-terminal domains (Figure 3.8). This tendency is also seen for each of the individual species in the dataset (see Methods), when the data for the different species, which include both prokaryotic and eukaryotic model organisms such as *E. coli, S. cerevisiae and H. sapiens*, are analyzed separately (not shown). Furthermore, when the analysis is restricted to two-domain proteins with a linker that is ten or less amino acids long (Figure 3.9), this trend becomes more pronounced as might be expected since a short linker length can increase the probability of mis-folding or aggregation (e.g. (Arndt et al., 1998). Finally, we also found that three-domain proteins in which the N-terminal domain is appreciably shorter (> ten amino acids) than the middle domain which, in turn, is appreciably shorter than the C-terminal domain are more abundant compared to triple-domain proteins with the other five possible rank orders of domain sizes (data not shown). We do not find, however, that triple-domains with any particular rank order of domain lengths are more abundant but the analysis of proteins with more than two domains is restricted by less available data on the one hand and more potential rank orders on the other hand.

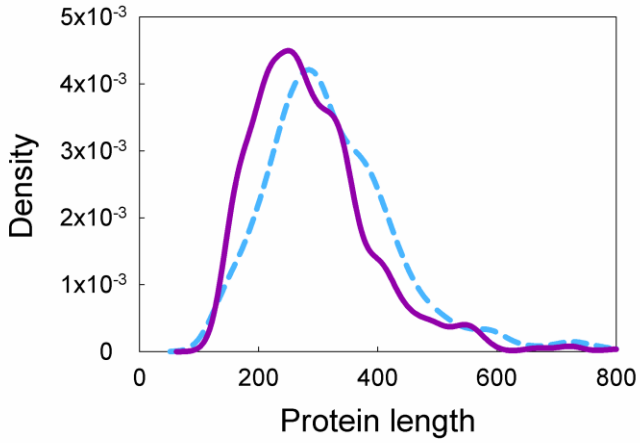


Figure 3.7 Chain length distributions of two-domain proteins with shorter N- or C-terminal domains.
Density plots of the overall length of two-domain proteins with shorter N- (purple) or C-terminal (turquoise) domains show that the tendency for those with shorter N-terminal domains to have a shorter overall length is significant with a Wilcoxon rank sum test (two sided) P-value of 2.5×10^{-11}. The mean values of the overall length of the two-domain proteins with shorter N- or C-terminal domains are 298 and 329 residues, respectively. The analysis is based on 1,699 two-domain proteins.

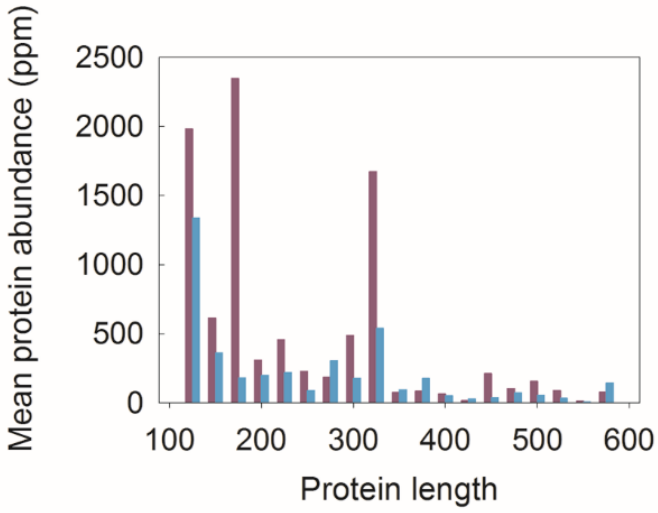


Figure 3.8 Comparison between the mean abundances of two-domain proteins of similar overall chain length with either a shorter N-terminal domain or with a shorter C-terminal domain.
The two-domain proteins were binned according to their overall length (bin width is set here to 25 amino acids) (see Figure 3.7) and the average abundances of all the proteins with either a shorter N-terminal domain (purple) or with shorter C-terminal domain (turquoise) in each bin were calculated separately. The analysis was carried out using abundance data downloaded from http://pax-db.org for 1,699 two-domain proteins (with length between 50-200 amino acids and a linker shorter than 30 amino acids) in the Pfam database from 12 different organisms. The mean abundances of all the two-domain proteins with a shorter N-terminal domain or with a shorter C-terminal domain are 590.53 and 229.75 ppm, respectively. The difference between the binned data for the two groups was found to be significant (P-value = 0.01) using the Wilcoxon rank sum paired test (two sided). For additional data and analyses, see Figures 3.9 and 3.10.

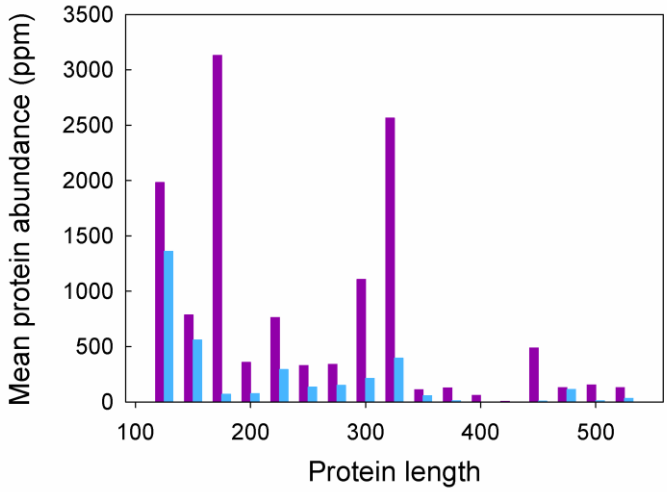


Figure 3.9 Comparison between the mean abundances of two-domain proteins of similar overall chain length with shorter N- (purple) or C-terminal (turquoise) domains that are connected by a linker of ten residues or less.
The analysis was carried out for 825 two-domain proteins as described in the legend to Figure. 3.8. The mean abundances of all the two-domain proteins with a shorter N-terminal domain or with a shorter C-terminal domain are 995.30 and 224.30 ppm, respectively. The difference between the binned data for the two groups was found to be significant (P-value $<3x10^{-5}$) using the Wilcoxon rank sum paired test (two sided).

3.3.5 Higher abundance of proteins with shorter N-terminal domains is much more pronounced for longer proteins

The risk of protein mis-folding as a result of formation of non-native inter-domain interactions is likely to increase as folding times approach translation times which can be the case for proteins longer than 150 amino acids (Naganathan and Muñoz, 2005). Hence, the benefit reflected in protein abundance that is associated with having a shorter N-terminal domain is expected to increase with protein size. We, therefore, compared the mean abundances of two-domain proteins with shorter N- or C-terminal domains (Figure 3.10) but with a similar overall chain length that is either less than 150 residues, between 150 and 300 residues or between 300 and 450 residues. In the case of two-domain proteins shorter than 150 residues, the mean abundance of two-domain proteins with a shorter N-terminal domain is only marginally higher than that of those with a shorter C-terminal domain (Figure 3.10). However, in the case of two-domain proteins that are longer than

150 residues, the mean abundance of those with a shorter N-terminal domain is significantly higher than that of those with a shorter C-terminal domain (Figure 3.10). These findings, therefore, suggest that selection for shorter N-terminal domains increases when the folding and translation times are in the same range.

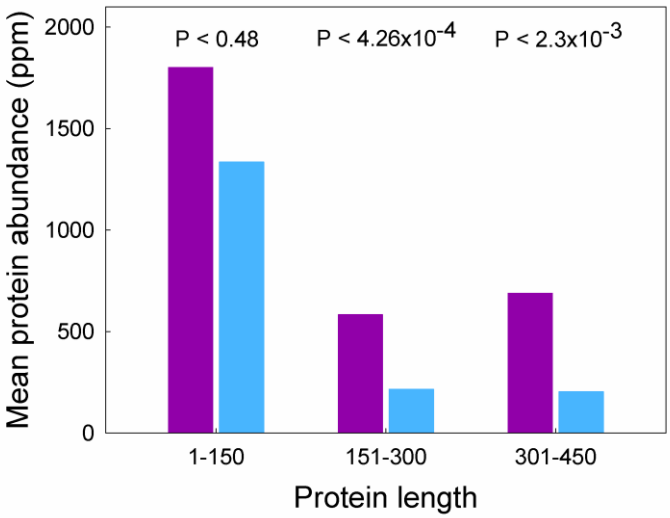


Figure 3.10 Comparison between the mean abundances of two-domain proteins with shorter N-terminal (purple) or C-terminal (turquoise) domains for three ranges of protein size.
The analysis was carried out for 33, 882 and 574 two-domain proteins with similar overall chain lengths that are either less than 150 residues, between 151 and 300 residues or between 301 and 450 residues, respectively. The ratios between the mean abundances of the proteins with a shorter N-terminal domain and those with a shorter C-terminal domain are 1.34 (1801.62/1335.67), 2.69 (583.20/216.20) and 3.36 (689.60/204.63) for the three protein size ranges (in order of increasing size), respectively. The P-values for the difference in the abundances of the two-domain proteins with shorter N-terminal or C-terminal domains in each of the three bins were calculated using the Wilcoxon rank sum paired test (two sided) and are indicated above the respective abundances.

3.3.6 Bias in proteins with more than two domains

Expanding the analysis for proteins with more than two domains is restricted by less available data on the one hand and more potential rank orders (6 for triple-domains) on the other hand. We do not find that triple-domains with any particular rank order of domain sizes are more common. However, we do see that triple-domains in which the first domain is appreciably shorter (by 10 amino acids) than the second which, in turn, is

appreciably shorter than the third (designated by 1<2<3) are more abundant compared to triple-domains with all other five rank orders (see Figure 3.11) with a P-value of 0.00038. This result supports our conclusions and is mentioned in the revised text.

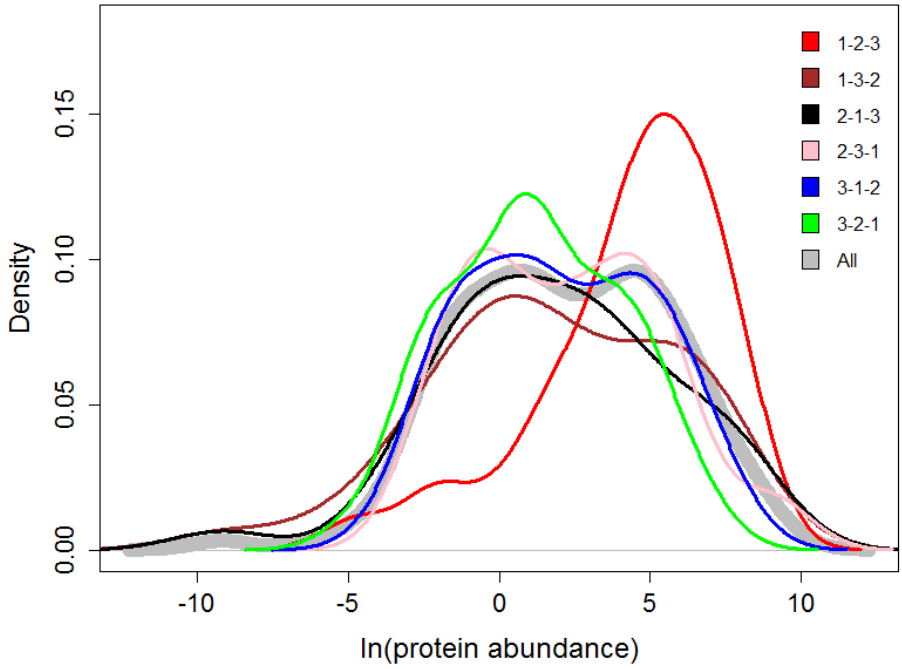


Figure 3.11 Abundance distributions for triple-domain proteins in all triple configurations.
Abundance values are taken from PaxDb.

3.4 Conclusions

This study shows a significant tendency for the N-terminal domains of two-domain proteins to be shorter than their neighboring C-terminal domains (Figure 3.4 and Table 3.1). I have also found that the ACO values of N-terminal domains tend to be smaller than those of their neighboring C-terminal domains (Figure 3.5). Given that both chain length and ACO are inversely correlated with folding rate, our results suggest that there is a bias for two-domain proteins in which the N-terminal domain folds faster than its C-terminal counterpart. In such two-domain proteins, folding of the N-terminal domain is predicted by the exponential dependence of folding rate on chain length and ACO (Ivankov et al., 2003; Plaxco et al., 1998) to be 10-14 times faster, on average, than that of the C-terminal domain. Such a bias in folding rates may reflect selection against mis-folding since domain-by-domain folding can minimize formation of non-native interdomain interactions. In addition, folding of an N-terminal domain can catalyse the folding of its neighboring C-terminal domain as shown for spectrin domains (Batey and Clarke, 2008), thereby reducing the risk of aggregation. Support for the suggestion that the bias reflects selection against mis-folding is provided by the observation that two-domain proteins with a shorter N-terminal domain are more abundant than those with a shorter C-terminal domain (Figure 3.8) since proteins with a tendency to mis-fold are, in general, less abundant (Drummond and Wilke, 2008; Tartaglia and Vendruscolo, 2009).

There is increasing evidence that folding of multi-domain proteins takes place co-translationally in both eukaryotes (Netzer and Hartl, 1997) and prokaryotes (Cabrita et al., 2010; Nicola et al., 1999). Co-translational folding is potentially more efficient than post-translational folding as it can facilitate domain-by-domain folding, thereby minimizing mis-folding owing to formation of non-native interdomain interactions. It has been suggested that translational pausing owing to the presence of rare codons might allow one domain to fold before synthesis of the other is completed (Komar, 2009). The data presented here, based on a survey of a large number of domain families, indicate that increased efficiency of multi-domain protein folding is often achieved by another mechanism, i.e. selection for faster folding of N-terminal domains relative to their C-terminal neighboring domains *via* fine tuning of their respective structural properties. The observation that this bias is greater in prokaryotes than in eukaryotes (Figures 3.5C,D and 3.6) is intriguing and may reflect compensation for the absence in prokaryotes of an extensive chaperone network that interacts with nascent chains (Albanèse et al., 2006).

The bias is also expected to be greater when the time-scales of translation and folding are closer and may, therefore, be more pronounced in prokaryotes since their translation rates are 5- to 10-fold faster than those of eukaryotes (Liang et al., 2000; Mathews et al., 2007). This expectation is also consistent with our observation that the tendency for proteins with shorter N-terminal domains to be more abundant than those with shorter C-terminal domains is more pronounced for longer proteins.

3.5 Future work

Another possible function for the bias for shorter N-terminal domains vs. their C-terminal counterparts described here, is structural templating, a process in which a folded domain facilitates the folding of its C-terminal neighboring domain that is translated subsequently,(Figure 3.12). A preliminary analysis of a non-redundant set (NR95) from SCOP (almost 2000 two-domain proteins) reveals more than 100 two-domain proteins with an observed template-like interface. Future work can extend this data set and continue in studying this direction in relation to folding rates, folding characteristics (e.g. intrinsically unfolded proteins and nanny proteins (Tsvetkov et al., 2009)) and structural density in multi-domain proteins.

In order to prevent aggregation and promote efficient folding under stress and normal conditions, many cellular resources are invested in a complex network of chaperones (Figure 3.3) (Albanèse et al., 2006; Hartl et al., 2011). The analysis of domain length bias that was based on more than 1300 complete proteomes from bacteria to human, showed that the higher the organism the lower is the bias (table 1.1). It is possible that the development of the chaperone network in higher organisms enabled the evolution of more multi-domain proteins in those genomes (Albanèse et al., 2006), while maintaining adequately low aggregation rates. This raises the question of identifying the part of that network that is involved in preventing multi-domain protein aggregation. Another interesting question is whether two-domain proteins with a strong bias towards shorter N-terminal residues potential substrates for chaperonins more than other multi-domain proteins (assuming they can enter the chaperonin cavity)?

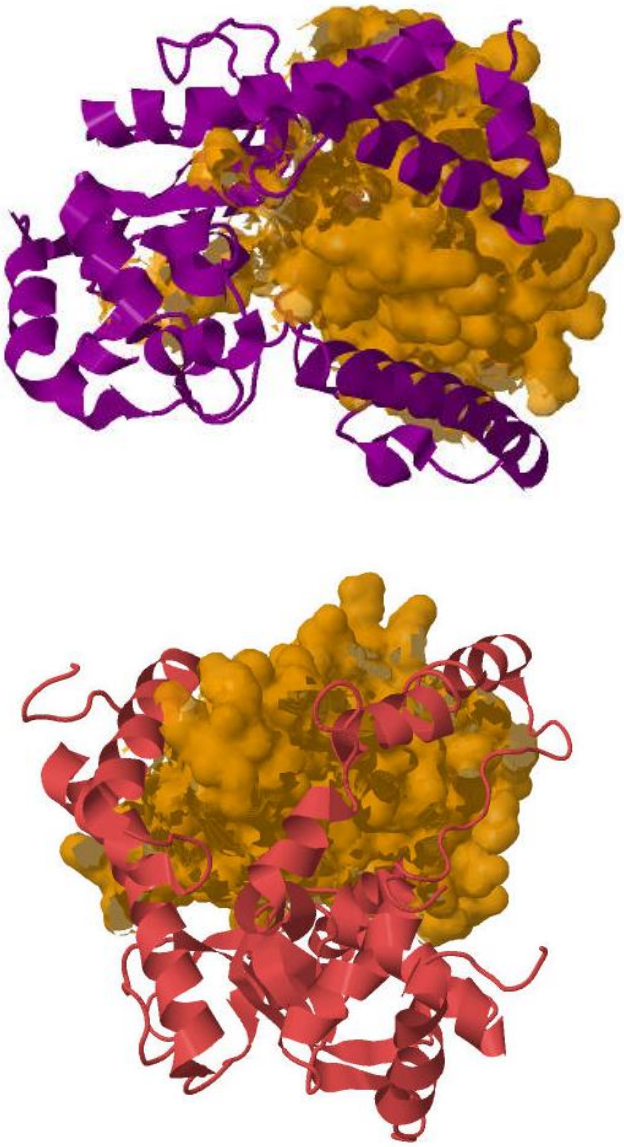


Figure 3.12 A possible templating mechanism of the N-terminal domain on its C-terminal counterpart.
Densely structured N-terminal (light brown) domain serves as a template for its neighboring C-terminal domain (purple or peach color). Top – PDB code 1vjtA, represents a hetero two-domain protein with a large interface between its domains. Bottom – PDB code 2k49A, represents a homo two-domain protein with a large interface between its domains.

Chapter 4

References

Adzhubei, I., Jordan, D.M., and Sunyaev, S.R. (2013). Predicting Functional Effect of Human Missense Mutations Using PolyPhen-2. Curr Protoc Hum Genet. Chapter 7:Unit7.20

Adzhubei, I. a, Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S., and Sunyaev, S.R. (2010). A method and server for predicting damaging missense mutations. Nat. Methods *7*, 248–249.

Albanèse, V., Yam, A.Y.-W., Baughman, J., Parnot, C., and Frydman, J. (2006). Systems analyses reveal two chaperone networks with distinct functions in eukaryotic cells. Cell *124*, 75–88.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J. Mol. Biol. *215*, 403–410.

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Res. *25*, 3389–3402.

Andreeva, A., Howorth, D., Chandonia, J.-M., Brenner, S.E., Hubbard, T.J.P., Chothia, C., and Murzin, A.G. (2008). Data growth and its impact on the SCOP database: new developments. Nucleic Acids Res. *36*, D419–D425.

Arndt, K.M., Müller, K.M., and Plückthun, A. (1998). Factors influencing the dimer to monomer transition of an antibody single-chain Fv fragment. Biochemistry *37*, 12918–12926.

Attwood, T.K., Gisel, A., Eriksson, N.-E., and Bongcam-Rudloff, E. (2011). Concepts, historical milestones and the central place of bioinformatics in modern Biology: a european perspective, Bioinformatics - Trends and Methodologies (InTech).

Balakrishnan, S., Kamisetty, H., Carbonell, J.G., Lee, S.-I., and Langmead, C.J. (2011). Learning generative models for protein fold families. Proteins *79*, 1061–1078.

Baldassi, C., Zamparo, M., Feinauer, C., Procaccini, A., Zecchina, R., Weigt, M., and Pagnani, A. (2014). Fast and accurate multivariate Gaussian modeling of protein families: Predicting residue contacts and protein-interaction partners. PLoS One *9*, 1–12.

Banerjee, O., El Ghaoui, L., d'Aspremont, A., and Ghaoui, L. El (2008). Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. J. Mach. Learn. Res. *9*, 485–516.

Bateman, A., Coin, L., Durbin, R., Finn, R.D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E.L.L., et al. (2004). The Pfam protein families database. Nucleic Acids Res. *32*, D138–D141.

Batey, S., and Clarke, J. (2008). The folding pathway of a single domain in a multidomain protein is not affected by its neighbouring domain. J. Mol. Biol. *378*, 297–301.

Batey, S., Scott, K.A., and Clarke, J. (2006). Complex folding kinetics of a multidomain protein. Biophys. J. *90*, 2120–2130.

Berman, H.M. (2008). The Protein Data Bank: a historical perspective. Acta Crystallogr. *64*, 88–95.

Borgia, M.B., Borgia, A., Best, R.B., Steward, A., Nettels, D., Wunderlich, B., Schuler, B., and Clarke, J. (2011). Single-molecule fluorescence reveals sequence-specific misfolding in multidomain proteins. Nature *474*, 662–665.

Burger, L., and van Nimwegen, E. (2010). Disentangling Direct from Indirect Co-Evolution of Residues in Protein Alignments. PLoS Comput. Biol. *6*, e1000633.

Cabrita, L.D., Dobson, C.M., and Christodoulou, J. (2010). Protein folding on the ribosome. Curr. Opin. Struct. Biol. *20*, 33–45.

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

Casari, G., Sander, C., and Valencia, A. (1995). A method to predict functional residues in proteins. Nat. Struct. Biol. *2*, 171–178.

Chen, C., Natale, D. a, Finn, R.D., Huang, H., Zhang, J., Wu, C.H., and Mazumder, R. (2011). Representative proteomes: a stable, scalable and unbiased proteome set for sequence analysis and functional annotation. PLoS One *6*, e18910.

Chothia, C., Gough, J., Vogel, C., and Teichmann, S.A. (2003). Evolution of the protein repertoire. Science *300*, 1701–1703.

Cover, T.M., and Thomas, J.A. (2005). Elements of Information Theory.

Cuff, A., Redfern, O.C., Greene, L., Sillitoe, I., Lewis, T., Dibley, M., Reid, A., Pearl, F., Dallman, T., Todd, A., et al. (2009). The CATH hierarchy revisited-structural divergence in domain superfamilies and the continuity of fold space. Structure *17*, 1051–1062.

Cunningham, F., Amode, M.R., Barrell, D., Beal, K., Billis, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fitzgerald, S., et al. (2014). Ensembl 2015. Nucleic Acids Res. *43*, D662–D669.

Dana, A., and Tuller, T. (2012). Determinants of translation elongation speed and ribosomal profiling biases in mouse embryonic stem cells. PLoS Comput. Biol. *8*, e1002755.

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

Dill, K.A., Ghosh, K., and Schmit, J.D. (2011). Physical limits of cells and proteomes. Proc. Natl. Acad. Sci. U. S. A. *108*, 17876–17882.

Drummond, D.A., and Wilke, C.O. (2008). Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. Cell *134*, 341–352.

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

Bioinformatics *24*, 333–340.

Eddy, S.R. (1998). Profile hidden Markov models. Bioinformatics *14*, 755–763.

Eddy, S.R., and Wheeler, T.J. (2013). HMMER User ' s Guide.

Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. *32*, 1792–1797.

Ekeberg, M., Lövkvist, C., Lan, Y., Weigt, M., and Aurell, E. (2013). Improved contact prediction in proteins: Using pseudolikelihoods to infer Potts models. Phys. Rev. E - Stat. Nonlinear, Soft Matter Phys. *87*, 1–21.

Ezkurdia, L., Grana, O., Izarzugaza, J.M.G., and Tress, M.L. (2009). Assessment of domain boundary predictions and the prediction of intramolecular contacts in CASP8. Proteins Struct. Funct. Bioinforma. *77*, 196–209.

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

Feizi, S., Marbach, D., Médard, M., and Kellis, M. (2013). Network deconvolution as a general method to distinguish direct dependencies in networks. Nat. Biotechnol. *31*, 726–733.

Felsenstein, J. (1989). Phylip: phylogeny inference package (version 3.2). Cladistics *5*, 164–166.

Finn, R.D., Clements, J., and Eddy, S.R. (2011). HMMER web server: interactive sequence similarity searching. Nucleic Acids Res. *39*, W29–W37.

Finn, R.D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Heger, A., Hetherington, K., Holm, L., Mistry, J., et al. (2014). Pfam: the protein families database. Nucleic Acids Res. *42*, D222–D230.

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

Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. Biostatistics *9*, 432–441.

Galperin, M.Y., Rigden, D.J., and Fernández-Suárez, X.M. (2015). The 2015 Nucleic Acids Research Database Issue and Molecular Biology Database Collection. Nucleic Acids Res. *43*, D1–D5.

Galzitskaya, O. V., Garbuzynskiy, S.O., Ivankov, D.N., and Finkelstein, A. V. (2003). Chain length is the main determinant of the folding rate for proteins with three-state folding kinetics. Proteins *51*, 162–166.

Gibbs, A., and Mcintyre, G. (1970). The diagram, a method for comparing sequences. Eur. J. Biochem. *16*, 1–11.

Gingeras, T.R., and Roberts, R.J. (1980). Steps toward computer analysis of nucleotide sequences. Science (80-.). *209*, 1322–1328.

Glaser, F., Pupko, T., Paz, I., Bell, R.E., Bechor-Shental, D., Martz, E., and Ben-Tal, N. (2003). ConSurf: identification of functional regions in proteins by surface-mapping of phylogenetic information. Bioinformatics *19*, 163–164.

Gloor, G.B., Martin, L.C., Wahl, L.M., and Dunn, S.D. (2005). Mutual information in protein multiple sequence alignments reveals two classes of coevolving positions. Biochemistry *44*, 7156–7165.

Göbel, U., Sander, C., Schneider, R., and Valencia, a (1994). Correlated mutations and residue contacts in proteins. Proteins *18*, 309–317.

Grant, B.J., Rodrigues, A.P.C., ElSawy, K.M., McCammon, J.A., and Caves, L.S.D. (2006). Bio3d: an R package for the comparative analysis of protein structures. Bioinformatics *22*, 2695–2696.

Gribskov, M., Mclachlant, A.D., and Eisenberg, D. (1987). Profile analysis : Detection of distantly related proteins. *84*, 4355–4358.

Han, J., Batey, S., Nickson, A.A., Teichmann, S.A., and Clarke, J. (2007). The folding and evolution of multidomain proteins. Nat. Rev. Mol. Cell Biol. *8*, 319–330.

Hartl, F.U., Bracher, A., and Hayer-Hartl, M. (2011). Molecular chaperones in protein folding and proteostasis. Nature *475*, 324–332.

Hogeweg, P., and Hesper, B. (1984). The alignment of sets of sequences and the construction of phyletic trees: an integrated method. J. Mol. Evol. *20*, 175–186.

Hopf, T. a, Schärfe, C.P.I., Rodrigues, J.P.G.L.M., Green, A.G., Kohlbacher, O., Sander, C., Bonvin, A.M.J.J., and Marks, D.S. (2014). Sequence co-evolution gives 3D contacts and structures of protein complexes. Elife *3*, 1–45.

Hopf, T.A., Colwell, L.J., Sheridan, R., Rost, B., Sander, C., and Marks, D.S. (2012). Three-Dimensional Structures of Membrane Proteins from Genomic Sequencing. Cell 1–15.

Horovitz, A., Bochkareva, E.S., Yifrach, O., and Girshovich, A.S. (1994). Prediction of an inter-residue interaction in the chaperonin GroEL from multiple sequence alignment is confirmed by double-mutant cycle analysis. J. Mol. Biol. *238*, 133–138.

Ivankov, D.N., Garbuzynskiy, S.O., Alm, E., Plaxco, K.W., Baker, D., and Finkelstein, A. V. (2003). Contact order revisited : Influence of protein size on the folding rate. 2057–2062.

Jacob, E., Horovitz, A., and Unger, R. (2007). Different mechanistic requirements for prokaryotic and eukaryotic chaperonins: a lattice study. Bioinformatics *23*, i240–i248.

Johnson, M.S., and Doolittle, R.F. (1986). A method for the simultaneous alignment of three or more amino acid sequences. J. Mol. Evol. *23*, 267–278.

Johnson, D.S., Mortazavi, A., Myers, R.M., and Wold, B. (2007). Genome-Wide Mapping of in Vivo Protein-DNA Interactions. Science (80-.). *316*, 1497–1502.

Jones, D.T. (1999). Protein secondary structure prediction based on position-specific scoring matrices. J. Mol. Biol. *292*, 195–202.

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

Jones, D.T., Singh, T., Kosciolek, T., and Tetchner, S. (2015). MetaPSICOV: combining coevolution methods for accurate prediction of contacts and long range hydrogen

bonding in proteins. Bioinformatics *31*, 999–1006.

de Juan, D., Pazos, F., and Valencia, A. (2013). Emerging methods in protein co-evolution. Nat. Rev. Genet. *14*, 249–261.

Kamisetty, H., Ovchinnikov, S., and Baker, D. (2013). Correction for Kamisetty et al., Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era. Proc. Natl. Acad. Sci. *110*, 18734–18734.

Kanz, C., Aldebert, P., Althorpe, N., Baker, W., Baldwin, A., Bates, K., Browne, P., van den Broek, A., Castro, M., Cochrane, G., et al. (2005). The EMBL nucleotide sequence database. Nucleic Acids Res. *33*, 29–33.

Karlin, S., and Brocchieri, L. (1996). Evolutionary conservation of RecA genes in relation to protein structure and function. J. Bacteriol. *178*, 1881–1894.

Kass, I., and Horovitz, A. (2002). Mapping pathways of allosteric communication in GroEL by analysis of correlated mutations. Proteins *48*, 611–617.

Katoh, K., Misawa, K., Kuma, K., and Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. *30*, 3059–3066.

Katz, L., and Burge, C.B. (2003). Widespread Selection for Local RNA Secondary Structure in Coding Regions of Bacterial Genes. Genome Res. 2042–2051.

Komar, A.A.. (2009). A pause for thought along the co-translational folding pathway. Trends Biochem. Sci. *34*, 16–24.

Krogh, A., Brown, M., Mian, I.S., Sjolander, K., and Haussler, D. (1994). Hidden Markov Models in Computational Biology: Applications to Protein Modeling. J. Mol. Biol. *235*, 1501–1531.

Lee, J., Natarajan, M., Nashine, V.C., Socolich, M., Vo, T., Russ, W.P., Benkovic, S.J., and Ranganathan, R. (2008). Surface sites for engineering allosteric control in proteins. Science *322*, 438–442.

Lenoir, T., and Giannella, E. (2006). The emergence and diffusion of DNA microarray technology. J. Biomed. Discov. Collab. *1*, 11.

Liang, S.T., Xu, Y.C., Dennis, P., and Bremer, H. (2000). mRNA composition and control of bacterial gene expression. J. Bacteriol. *182*, 3037–3044.

Lieberman-Aiden, E., Berkum, N.L. Van, Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O., et al. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science (80-.). *326*, 289–293.

Lipman, D.J., Altschul, S.F., and Kececioglu, J.D. (1989). A tool for multiple sequence alignment. Proc Natl Acad Sci USA *86*, 4412–4415.

Livesay, D.R., Kreth, K.E., and Fodor, A.A. (2012). A critical evaluation of correlated mutation algorithms and coevolution within allosteric mechanisms. Methods Mol. Biol. *796*, 385–398.

Lockless, S.W., Lockless, W.S., and Ranganathan, R. (1999). Evolutionarily Conserved Pathways of Energetic Connectivity in Protein Families. Science (80-.). *286*, 295–299.

Luheshi, L.M., and Dobson, C.M. (2009). Bridging the gap: From protein misfolding to protein misfolding diseases. FEBS Lett. *583*, 2581–2586.

Mao, W., Kaya, C., Dutta, A., Horovitz, A., and Bahar, I. (2015). Comparative Study of the Effectiveness and Limitations of Cur- rent Methods for Detecting Sequence Coevolution. 1–8.

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

Marks, D.S., Hopf, T. a, and Sander, C. (2012). Protein structure prediction from sequence variation. Nat. Biotechnol. *30*, 1072–1080.

Martin, L.C., Gloor, G.B., Dunn, S.D., and Wahl, L.M. (2005). Using information theory to search for co-evolving residues in proteins. Bioinformatics *21*, 4116–4124.

Mathews, M.B., Sonenberg, N., and Hershey, J.W.B. (2007). 1 Origins and Principles of Translational Control. Cold Spring Harb. Monogr. Arch. *48*, 1–40.

McWilliam, H., Valentin, F., Goujon, M., Li, W., Narayanasamy, M., Martin, J., Miyar, T., and Lopez, R. (2009). Web services at the European Bioinformatics Institute-2009. Nucleic Acids Res. *37*, W6–W10.

Mead, L.R., and Papanicolaou, N. (1984). Maximum entropy in the problem of moments. J. Math. Phys. *25*, 2404–2417.

Meinshausen, N., and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the Lasso. Ann. Stat. *34*, 1436–1462.

Meyer, I.M., and Miklós, I. (2005). Statistical evidence for conserved, local secondary structure in the coding regions of eukaryotic mRNAs and pre-mRNAs. Nucleic Acids Res. *33*, 6338–6348.

Mizuguchi, K., Deane, C.M., Blundell, T.L., and Overington, J.P. (1998). HOMSTRAD: a database of protein structure alignments for homologous families. Protein Sci. *7*, 2469–2471.

Morcos, F., Pagnani, A., Lunt, B., Bertolino, A., Marks, D.S., Sander, C., Zecchina, R., Onuchic, J.N., Hwa, T., and Weigt, M. (2011). Direct-coupling analysis of residue coevolution captures native contacts across many protein families. Proc. Natl. Acad. Sci. *108*, E1293–E1301.

Mortimer, S.A., Kidwell, M.A., and Doudna, J.A. (2014). Insights into RNA structure and function from genome-wide studies. Nat. Rev. Genet. *15*, 469–479.

Moult, J., Pedersen, J.T., Judson, R., and Fidelis, K. (1995). A large-scale experiment to assess protein structure prediction methods. Proteins Struct. Funct. Bioinforma. *23*, ii – iv.

Moult, J., Fidelis, K., Kryshtafovych, A., and Tramontano, A. (2011). Critical assessment of methods of protein structure prediction (CASP)-round IX. Proteins Struct. Funct. Bioinforma. *79*, 1–5.

Murzin, A.G., Brenner, S.E., Hubbard, T., and Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. J.

Mol. Biol. *247*, 536–540.

Naganathan, A.N., and Muñoz, V. (2005). Scaling of folding times with protein size. J. Am. Chem. Soc. *127*, 480–481.

Needleman, S.B., and Wunsch, C.D. (1970). A general method applicable to search for similarities in amino acid sequence of two proteins. J. Mol. Biol. *48*, 443–453.

Neher, E. (1994). How frequent are correlated changes in families of protein sequences? Proc. Natl. Acad. Sci. U. S. A. *91*, 98–102.

Netzer, W.J., and Hartl, F.U. (1997). Recombination of protein domains facilitated by co-translational folding in eukaryotes. Nature *388*, 343–349.

Nicola, A. V., Chen, W., and Helenius, A. (1999). Co-translational folding of an alphavirus capsid protein in the cytosol of living cells. Nat. Cell Biol. *1*, 341–345.

Noivirt, O., Eisenstein, M., and Horovitz, A. (2005). Detection and reduction of evolutionary noise in correlated mutation analysis. Protein Eng. Des. Sel. *18*, 247–253.

Notredame, C., Higgins, D.G., and Heringa, J. (2000). T-Coffee: A novel method for fast and accurate multiple sequence alignment. J. Mol. Biol. *302*, 205–217.

Oberhauser, A.F., Marszalek, P.E., Carrion-Vazquez, M., and Fernandez, J.M. (1999). Single protein misfolding events captured by atomic force microscopy. Nat. Struct. Biol. *6*, 1025–1028.

Ofran, Y., and Rost, B. (2003a). Analysing six types of protein-protein interfaces. J. Mol. Biol. *325*, 377–387.

Ofran, Y., and Rost, B. (2003b). Predicted protein-protein interaction sites from local sequence information. FEBS Lett. *544*, 236–239.

Olmea, O., Rost, B., and Valencia, a (1999). Effective use of sequence correlation and conservation in fold recognition. J. Mol. Biol. *293*, 1221–1239.

Orengo, C.A., Michie, A.D., Jones, S., Jones, D.T., Swindells, M.B., and Thornton, J.M. (1997). CATH - a hierarchic classification of protein domain structures. Structure *5*, 1093–1108.

Pandit, S.B., Gosar, D., Abhiman, S., Sujatha, S., Dixit, S.S., Mhatre, N.S., Sowdhamini, R., and Srinivasan, N. (2002). SUPFAM-a database of potential protein superfamily relationships derived by comparing sequence-based and structure-based families: implications for structural genomics and function annotation in genomes. Nucleic Acids Res. *30*, 289–293.

Pearl, F.M.G. (2003). The CATH database: an extended protein family resource for structural and functional genomics. Nucleic Acids Res. *31*, 452–455.

Pearson, W.R., and Lipman, D.J. (1988). Improved tools for biological sequence comparison. Proc Natl Acad Sci USA *85*, 2444–2448.

Pethica, R.B., Levitt, M., and Gough, J. (2012). Evolutionarily consistent families in SCOP: sequence, structure and function. BMC Struct. Biol. *12*, 27.

Plaxco, K.W., Simons, K.T., and Baker, D. (1998). Contact order, transition state placement and the refolding rates of single domain proteins. J. Mol. Biol. *277*, 985–994.

Pollock, D.D., Taylor, W.R., and Goldman, N. (1999). Coevolving protein residues: maximum likelihood identification and relationship to structure. J. Mol. Biol. *287*, 187–198.

Pruitt, K.D., Tatusova, T., Brown, G.R., and Maglott, D.R. (2012). NCBI Reference Sequences (RefSeq): Current status, new features and genome annotation policy. Nucleic Acids Res. *40*, 130–135.

Punta, M., Coggill, P.C., Eberhardt, R.Y., Mistry, J., Tate, J., Boursnell, C., Pang, N., Forslund, K., Ceric, G., Clements, J., et al. (2012). The Pfam protein families database. Nucleic Acids Res. *40*, D290–D301.

Pupko, T., Bell, R.E., Mayrose, I., Glaser, F., and Ben-Tal, N. (2002). Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. Bioinformatics *18 Suppl 1*, S71–S77.

Rabiner, L.R. (1989). Tutorial on Hmm and Applications.Pdf. Proc. IEEE *77*, 257–286.

Rice, P., Longden, I., and Bleasby, A. (2000). EMBOSS: The European Molecular Biology Open Software Suite. Trends Genet. *16*, 276–277.

Rivenzon-Segal, D., Wolf, S.G., Shimon, L., Willison, K.R., and Horovitz, A. (2005). Sequential ATP-induced allosteric transitions of the cytoplasmic chaperonin containing TCP-1 revealed by EM analysis. Nat. Struct. Mol. Biol. *12*, 233–237.

Rost, B., and Sander, C. (1993). Prediction of protein secondary structure at better than 70% accuracy. J. Mol. Biol. *232*, 584–599.

Sanger, F., Nicklen, S., and Coulson, A.R. (1977). DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci USA *74*, 5463–5467.

Selkoe, D.J. (2003). Folding proteins in fatal ways. Nature *426*, 900–905.

Shah, P., Ding, Y., Niemczyk, M., Kudla, G., and Plotkin, J.B. (2013). Rate-limiting steps in yeast protein translation. Cell *153*, 1589–1601.

Shilton, B.H., Thomas, D.Y., and Cygler, M. (1997). Crystal structure of Kex1deltap, a prohormone-processing carboxypeptidase from Saccharomyces cerevisiae. Biochemistry *36*, 9002–9012.

Siddiqui, A.S., Dengler, U., and Barton, G.J. (2001). 3Dee: a database of protein structural domains. Bioinformatics *17*, 200–201.

Skwark, M.J., Raimondi, D., Michel, M., and Elofsson, A. (2014). Improved contact predictions using the recognition of protein like contact patterns. PLoS Comput. Biol. *10*, e1003889.

Smith, T.F., and Waterman, M.S. (1981). Identification of Common Molecular Subsequences. J. Mol. Biol. *147*, 195–197.

Stein, R.R., Marks, D.S., and Sander, C. (2015). Inferring Pairwise Interactions from Biological Data Using Maximum-Entropy Probability Models. PLOS Comput. Biol. *11*, e1004182.

Stretton, A.O.W. (2002). Anecdotal , Historical and Critical Commentaries on Genetics The First Sequence : Fred Sanger and Insulin. Genetics *162*, 527–532.

Suzek, B.E., Huang, H., McGarvey, P., Mazumder, R., and Wu, C.H. (2007). UniRef: comprehensive and non-redundant UniProt reference clusters. Bioinformatics *23*, 1282–1288.

Tartaglia, G.G., and Vendruscolo, M. (2009). Correlation between mRNA expression levels and protein aggregation propensities in subcellular localisations. Mol. Biosyst. *5*, 1873–1876.

Thirumalai, D., O'Brien, E.P., Morrison, G., and Hyeon, C. (2010). Theoretical perspectives on protein folding. Annu. Rev. Biophys. *39*, 159–183.

Thompson, J.D., Higgins, D.G., and Gibson, T.J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. *22*, 4673–4680.

Tsvetkov, P., Reuven, N., and Shaul, Y. (2009). The nanny model for IDPs. Nat. Chem. Biol. *5*, 778–781.

Vogel, C., Bashton, M., Kerrison, N.D., Chothia, C., and Teichmann, S.A. (2004). Structure, function and evolution of multidomain proteins. Curr. Opin. Struct. Biol. *14*, 208–216.

Wan, Y., Qu, K., Zhang, Q.C., Flynn, R. a, Manor, O., Ouyang, Z., Zhang, J., Spitale, R.C., Snyder, M.P., Segal, E., et al. (2014). Landscape and variation of RNA secondary structure across the human transcriptome. Nature *505*, 706–709.

Wang, G., and Dunbrack, R.L. (2003). PISCES: A protein sequence culling server. Bioinformatics *19*, 1589–1591.

Wang, G., and Dunbrack, R.L. (2005). PISCES: Recent improvements to a PDB sequence culling server. Nucleic Acids Res. *33*, 94–98.

Wang, J., Zhuang, J., Iyer, S., Lin, X., Whitfield, T.W., Greven, M.C., Pierce, B.G., Dong, X., Kundaje, A., Cheng, Y., et al. (2012a). Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. Genome Res. *22*, 1798–1812.

Wang, M., Weiss, M., Simonovic, M., Haertinger, G., Schrimpf, S.P., Hengartner, M.O., and von Mering, C. (2012b). PaxDb, a Database of Protein Abundance Averages Across All Three Domains of Life. Mol. Cell. Proteomics *11*, 492–500.

Weigt, M., White, R.A., Szurmant, H., Hoch, J.A., and Hwa, T. (2009). Identification of direct residue contacts in protein-protein interaction by message passing. Proc. Natl. Acad. Sci. U. S. A. *106*, 67–72.

Wen, J.-D., Lancaster, L., Hodges, C., Zeri, A.-C., Yoshimura, S.H., Noller, H.F., Bustamante, C., and Tinoco, I. (2008). Following translation by single ribosomes one codon at a time. Nature *452*, 598–603.

Wold, B., and Myers, R.M. (2007). Sequence census methods for functional genomics. Nat. Methods *5*, 19–21.

Wolin, S.L., and Walter, P. (1988). Ribosome pausing and stacking during translation of a eukaryotic mRNA. EMBO J. *7*, 3559–3569.

Wollenberg, K.R., and Atchley, W.R. (2000). Separation of phylogenetic and functional associations in biological sequences by using the parametric bootstrap. Proc. Natl. Acad. Sci. U. S. A. *97*, 3288–3291.

Wright, C.F., Teichmann, S.A., Clarke, J., and Dobson, C.M. (2005). The importance of sequence diversity in the aggregation and evolution of proteins. Nature *438*, 878–881.

Yates, A., Beal, K., Keenan, S., McLaren, W., Pignatelli, M., Ritchie, G.R.S., Ruffier, M., Taylor, K., Vullo, A., and Flicek, P. (2015). The Ensembl REST API: Ensembl Data for Any Language. Bioinformatics *31*, 143–145.

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

תקציר - Abstract in Hebrew

מבוא

רצף חומצות האמינו הראשון שפוצח במלואו היה של חלבון האינסולין בשנת 1955. תהליך הפיצוח ארך כעשר שנים ע״י סנגר ועמיתיו ובמהלכו פותחו שיטות ניסוי חדשות. חשיבותם של ממצאים אלו, מעבר לשיטות הניסוי אשר פותחו, הייתה השינוי בתפיסה המדעית של אותה תקופה. במהלך השנים שופרו שיטות ריצוף החלבונים ותהליכי הפיצוח התקצרו באופן משמעותי – ולראייה, בשנת 1965 כבר היו בנמצא 65 רצפים של חלבונים שונים. בניגוד להצלחות בריצוף חלבונים בתקופה זו, ריצוף הדנ״א נתקל במכשולים רבים, ובפרט בקושי לרצף מולקולות דנ״א ארוכות. המהפך המשמעותי קרה בשנת 1977 כאשר סנגר ועמיתיו, פיתחו שיטה המסוגלת להתמודד בהצלחה עם ריצוף של מולקולות דנ״א ארוכות. שיטה זו, הקרויה ״שיטת ריצוף הדור-הראשון״, שימשה כבסיס לטכנולוגית הריצוף ב- 25 שנים הבאות, ונעשה בה אף שימוש לפיצוחם המלא של רצפי גנומים של מספר לא מבוטל של אורגניזמים, לרבות הגנום האנושי.

למן ההכרזה על פיצוח הגנום האנושי בשנת 2003, עבר עולם ריצוף הנוקלאוטידים (רנ״א ודנ״א) מהפך משמעותי המבוסס על טכנולוגית ריצוף הדור החדש. טכנולוגיה זו מבוססת על ריצוף מקביל של מיליוני חתיכות קצרות של דנ״א או רנ״א. תהליך הריצוף באמצעות שיטה זו הוא מהיר ביותר ומאפשר ריצוף של גנום אנושי בשלמותו בפחות מיום בודד. בנוסף, העלויות הנמוכות ופשטות השימוש והתפעול של הטכנולוגיה הנ״ל הביאו לתפוצה רחבה של טכנולוגיות אלו ברחבי העולם, וכיוצא בזה גם ליצירה של כמויות רחבות היקף של מידע רצפי. טכנולוגיות ריצוף הדור החדש משמשות גם לאפיון תופעות ביולוגיות. לדוגמא, ע״מ לחקור אינטראקציות של חלבונים עם דנ״א, נעזרים בשיטה המשלבת את טכנולוגית ריצוף הדור החדש עם מיצוי נוגדני של כרומטין (ChIP-seq). בשיטה זו, גדילי הדנ״א נחתכים, והמקטעים המכילים אזורים שעוברים אינטראקציה עם חלבונים – כגון פקטורי שעתוק – מושקעים ע״י חלבונים ספציפיים שמזהים אותם ומנופים משאר המקטעים. תלכידי החלבונים והדנ״א שמוצו עוברים הפרדה, ומקטעי הדנ״א מרוצפים בשיטת ריצוף הדור החדש. תוצר הרצפים ממופה לגנום, ובכך ניתן להעריך באילו אתרים גנומים קיימת אינטראקציה של החלבונים שמוצו עם הדנ״א, לרבות אינטראקציה של פקטורי שעתוק עם פרומוטורים.

בד בבד עם התפתחות שיטות הריצוף, נעשתה התקדמות משמעותית גם בשיטות ניסוי שתוצרתן אינה רצפים ביולוגים. לדוגמא, בסוף שנות השמונים של המאה הקודמת, התפתחויות בשיטות הקריסטלוגרפיה, שיבוט גנים, ביטוי חלבונים ויכולות החישוב הביאו להצלחות בתחום פיתרון המבנים השלישוניים של חלבונים רבים, וכתוצאה מכך הביאו להרחבה משמעותית של מסד הנתונים של בנק מבני החלבונים (Protein Data Bank). דוגמא משמעותית נוספת לטכנולוגיה שפותחה בשנות התשעים של המאה הקודמת היא שיטת מדידה של ביטוי גנים הנקראת הצ׳יפ הגנטי. שיטה זו, אשר הלכה ונהפכה למקובלת בעולם המחקר במשך שנים רבות, מאפשרת מדידה במקביל של רמות הביטוי של אלפי גנים עבור דגימה ביולוגית באופן פשוט יחסית. כתוצאה מכך, הלכו והצטברו במאגרי הנתונים הציבוריים כמויות רבות ביותר של מדידות ביצורים שונים ממעבדות ברחבי העולם.

הצטברות המידע הרב במהלך השנים מכל אותם ניסויים יצר הכרח לארגנו במסדי נתונים ע״מ לאפשר לקהילת המחקר גישה נוחה ויעילה אליו, לצורך ביצוע מגוון ניתוחים וחישובים. דוגמא מובהקת לארגון מסד נתונים כאמור הנה יצירתו של מסד הנתונים הראשון עוד בשנות ה-60׳ של המאה הקודמת, אשר מנה 65 רצפי חלבונים. בהמשך, בשנות השמונים של המאה הקודמת, מסדי הנתונים הלכו וגדלו בעקבות ההתפתחויות הטכנולוגיות, והגיעו לכדי מאות רצפים. הגידול בכמות הרצפים התבטא ביתר שאת לגבי רצפי

נוקלאוטידים (דנ״א ורנ״א). בשנת 1995 כבר כללו מסדי הנתונים רצפי גנומים שלמים שהצטברו, ובהם הגנום האנושי אשר פוצח כאמור בשנת 2003. במהלך שנות האלפיים, טכנולוגיית ריצוף הדור החדש, עליה עמדנו לעיל, הביאה לגידול מעריכי של מספר הרצפים. לשם המחשה, בשנת 2009 הכיל אחד ממאגרי המידע הנוקלאוטידי המרכזי בעולם 1.35E-13 בסיסי נוקלאוטידים, ואילו בשנת 2015 הכיל מאגר זה 2.75E-15 בסיסים. לצד הגידול חסר התקדים במאגרי הרצפים, ניכר גידול משמעותי ביותר גם במאגרים של נתונים אחרים – כגון מאגר המידע של מבני החלבון השלישוני (PDB), מאגר המידע של ביטוי גנים (GEO) ורבים אחרים.

השוואות בין רצפים וניתוחם, אשר מטרתם לזהות אזורים ברצף בעלי חשיבות פונקציונלית או מבנית, התבצעו כבר בשנות החמישים של המאה הקודמת – עם השלמת ריצופם של חלבוני האינסולין מיצורים שונים. שיטות אלו – כגון עימוד רצפים (sequence alignment) ועימוד מרובה רצפים (Multiple sequence alignment) – שוכללו ושופרו ע״מ להתמודד עם העלייה במספר הרצפים שהצטברו במאגרי המידע במהלך השנים.

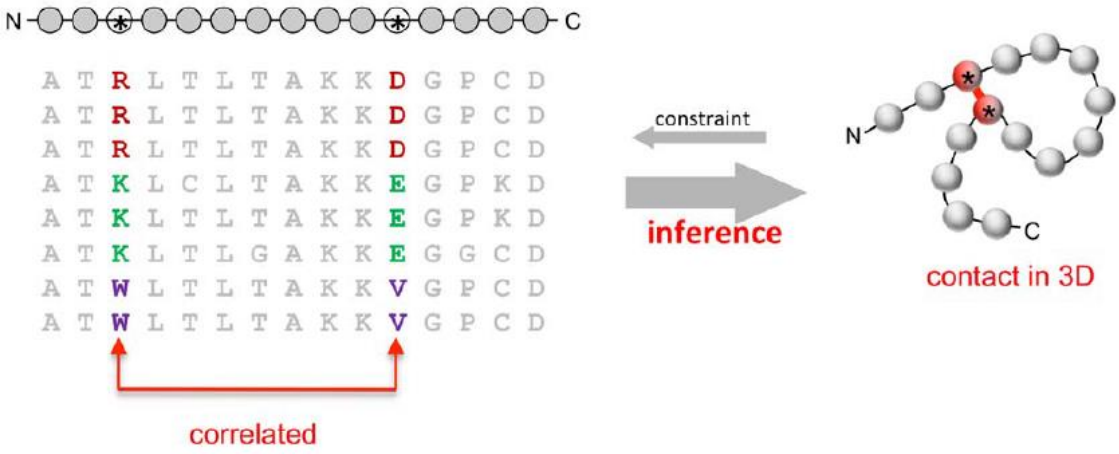
לצד הגידול חסר התקדים במאגרי המידע של הרצפים, חל גידול משמעותי גם במאגרי מידע ביולוגיים אחרים (שהמידע בהן אינו רצף חומצות אמינו או נוקלאוטידים), כגון מאגר המבנים השלישוניים של החלבונים (PDB) – ואולם הגידול במאגרי מידע אלה היה איטי לאין שיעור מהגידול במאגרי המידע של הרצפים. ביטוי מובהק לכך מצוי בעובדה כי מספר המבנים השלישוניים הידוע כיום נמוך בהרבה ממספר הרצפים הידועים, ופער זה רק הולך וגדל. במטרה לגשר על הפער האמור, החלו לפתח שיטות ביואינפורמטיות, שמנבאות תכונות של חלבונים על בסיס הרצף לבדו – ובכלל זה גם ניבוי מבנה שניוני ושלישוני של חלבונים.

הצלחתן של שיטוי ניבוי אלה (אשר הולכות ומשתכללות אף הן בחלוף הזמן) עודנה מוגבלת בחלק לא מבוטל מהמקרים, ואולם עצם האפשרות למצוא מידע שהיה קשה לגלותו באמצעות ניסויים במעבדה בלבד מדגיש את הפוטנציאל הרב הטמון בהן.תזה זו מתמקדת בגילוי תכונות של חלבונים על בסיס רצפם, באמצעות שיטות חישוביות ועל בסיס מידע ממקורות ביולוגיים שונים. בפרק הראשון תובא ותנותח גישה חדשה למיצוי מידע על רצפים לצורך הגעה להבנות חדשות בנוגע למבנה השלישוני שלהם ; ובפרק השני יוצע מנגנון אפשרי למניעת אגרגציה בחלבונים מרובי דומיינים.

פרק ראשון

שיטות רבות לניבוי מבנה שלישוני של חלבון אשר פותחו במהלך השנים, מתבססות על דמיון בין הרצף שאת מבנהו מעוניינים לנבא לבין רצף חלבונים אחרים שהמבנה השלישוני שלהם כבר ידוע – אלא שבמקרים רבים אין חלבונים בעלי מבנה שלישוני ידוע, אשר רצפם דומה לרצף החלבון אותו מעוניינים לנבא. במקרים אלה, שיטות ניתוח המבוססות על הרצף בלבד ומשתמשות בניתוחים סטטיסטיים על מידע אבולוציוני – יכולות לתת את המענה. הרעיון בבסיס שיטות ניתוח אלו, המכונות "ניתוח מוטציות בקורלציה", הוא שכאשר מתרחשת מוטציה באתר מסוים בחלבון, המשפיעה באופן כלשהו על מבנהו או יציבותו, היא גוררת בעקבותיה מוטציה באתר אחר שיש לו איזושהי אינטראקציה עם אותו אתר, על מנת לפצות על השינוי שנגרם. ההערכה היא כי במקרים רבים מוטציות מפצות אחת על השנייה באתרי חלבון הנמצאים קרוב אחד לשני ויוצרים לרוב קשר פיזיקלי. לכן, במהלך השנים, פותחו שיטות חישוביות שמטרתן לזהות אתרי חלבון בקורלציה גבוהה, כדרך לנבא קירבה מרחבית בין אותם אתרים. בכך, ניתן להעריך את המבנה השלישוני של החלבון כולו על בסיס אותם אתרים שנובאו כקרובים מרחבית.

במהלך עשרים השנים האחרונות, פותחו שיטות רבות לניבוי קשרים בין חומצות אמינו בחלבונים על בסיס ניתוח מוטציות בקורלציה. בחלק גורף מהמקרים, השיטות השונות כוללות את השלבים כדלהלן (ראה שרטוט 1). ראשית, מתבצע עימוד מרובה רצפים של חלבונים בעלי קירבה אבולוציונית (הומולוגיה רצפית) לרצף החלבון הנדרש לניבוי. שנית, ערכי המתאם בין כל זוגות האתרים בחלבון נמדדים באופן הבא: תדירות המופעים המשותפים של חומצות האמינו בין כל שני אתרים בחלבון נמדדת ומשווה לצפי התדירות המשותפת בהנחה של אי תלות בין שני האתרים. שלישית, ערכי מתאם אלה מדורגים ע"פ סדר חשיבות סטטיסטית או פיזיקלית שמשוערכת עבור כל ערך. לזוגות עם ערך מתאם גבוה שעבר סף מסוים אשר נקבע מראש (לדוגמא, סף של 200 הזוגות בעלי ערך המתאם הגבוה ביותר), מיוחסת קירבה פיזית שיכולה לשמש לניבוי המבנה השלישוני של החלבון.



שרטוט 1. זיהוי אתרים בעלי זיקה אחד לשני ברצף החלבון היכולה לרמוז על קירבה פיזית במבנהו השלישוני. רצף החלבון עבורו נבקש לנבא את המבנה השלישוני (שרשרת העיגולים האפורים), שייך למשפחת חלבונים בעלי קירבה אבולוציונית (העימוד מרובה הרצפים באותיות לטיניות אפורות). השונות האבולוציונית בין שני הטורים בעימוד הרצפים היא תוצר של אילוץ פיזיקלי הנובע מקשר ישיר בין שני האתרים. תמונה זו נלקחה ממחקר שנעשה בעבר ופורסם ע"י אחרים (Marks et al., 2015).

ברם, ערכי מתאם גבוהים לעיתים אינם משקפים קירבה פיזית בין שני אתרים בחלבון ויכולים לנבוע משתי סיבות מרכזיות נוספות. האחת, שנקראת מוצא אבולוציוני משותף ("רעש" אבולוציוני), היא תוצאה של מבנה העץ הפילוגנטי שמכתיב את הקורלציות בין האתרים ללא קשר לאילוצים הנובעים ממבנה החלבון. השנייה, שנקראת קשר לא ישיר בין שני האתרים, היא תוצאה של מצב בו קשרים ישירים (קירבה פיזית) בין אתר א' לאתר ב' בחלבון ובין אתר ב' לאתר ג' גוררים קורלציה לכאורה בין אתר א' לאתר ג', אף על פי שאין ביניהם קשר ישיר. שתי סיבות מרכזיות אלו מביאות לשגיאה בזיהוי, כלומר למסקנה כי הקורלציה בין שני אתרים מעידה על קשר ישיר ביניהם אף על פי שלא כך הוא הדבר בפועל. יוצא אפוא, כי הרעש האבולוציוני והקשר הלא ישיר הנ"ל מהווים את המכשול המרכזי ביכולת הניבוי של מבנים שלישונים. לאור זאת, נעשו במהלך השנים ניסיונות לנפות מקרים אלו. ואכן, כבר עם פיתוחן של השיטות המוקדמות של ניתוח מוטציות בקורלציה, הומצאו גם דרכים שנועדו להתמודד עם שגיאות שנובעות ממוצא אבולוציוני משותף והביאו לשיפור בהצלחות הניבוי. אף על פי כן, מכיוון שבשיטות המוקדמות של ניתוח מוטציות בקורלציה הייתה הנחת מוצא, לפיה אין תלות של זוגות אתרים באתרים נוספים בחלבון, שגיאות בזיהוי הנובעות מהסיבה השנייה שהוזכרה לעיל – קשר לא ישיר בין שני אתרים – המשיכו להוות מכשול מרכזי.

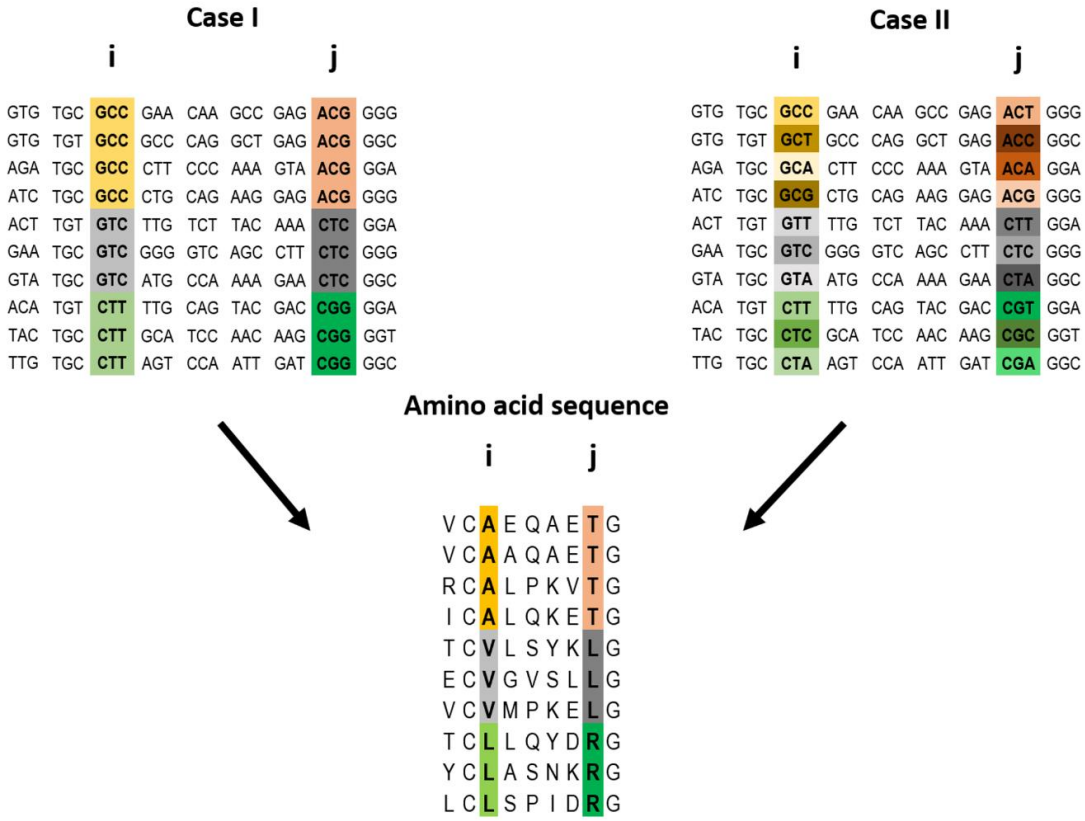
בשנים האחרונות, עם העלייה במספר הרצפים, אשר הביאה בהתאמה לגידול בכמות המידע שהצטבר בנוגע למשפחות חלבונים רבות, פותחו שיטות סטטיסטיות שלוקחות בחשבון את קיום התלות בין זוגות אתרים לבין כל שאר האתרים בחלבון, ובכך מצליחות להתמודד עם מכשול הרעש, אשר נובע כאמור מקשר לא ישיר בין שני אתרים. הסרת מכשול הרעש שיפרה אמנם, באופן משמעותי, את הביצועים של השיטות החדשות לעומת המוקדמות – בכל הנוגע לניבוי מבנה שלישוני של חלבונים על בסיס הרצף,ואולם רמת הדיוק של שיטות אלה עודנה מוגבלת. חיסרון נוסף של שיטות חדשות

אלה מתבטא בכך שהן דורשות מספר רב של רצפים על מנת להגיע לרמת ניבוי טובה יותר מהשיטות המוקדמות.

עד עתה, ניתוח מוטציות בקורלציה התבסס באופן בלעדי על רצפי חלבונים. אולם, מידע נוסף ובעל ערך נמצא ברמת רצפי הנוקלאוטידים המקודדים לחלבונים אלו. מקורו של מידע זה מגיע מיתירות הקוד הגנטי המתבטאת בכך שיותר מסוג אחד של קודון מתורגם לאותה חומצה אמינית.

בעבודת מחקר זו, מתוארת גישה חדשה לניתוח מוטציות בקורלציה המבוססת על מידע ברמת רצף חומצות האמינו בשילוב עם מידע ברמת רצף הנוקלאוטידים המקודדים לאותם חלבונים. העיקרון מאחורי גישה זו הנו פשוט: הסבירות לקשר ישיר בין שני אתרים בחלבון היא גבוהה יותר כאשר הקורלציה ברמת חומצות האמינו בין אתרים אלה חזקה ובו בעת הקורלציה ברמת הקודונים המתאימים היא חלשה. על ידי השימוש הנוסף במידע ברמת הקודונים בשיטות שונות של ניתוח מוטציות בקורלציה, נראה כיצד יש שיפור בניבוי קשרים בין אתרים בחלבון – והדברים יובאו בתמצית להלן:

באמצעות מספר שיטות של ניתוח מוטציות בקורלציה, הציון המנבא את פוטנציאל הקשר עבור כל זוגות האתרים בחלבון מחושב פעמיים, האחת ברמת חומצות האמינו; והשנייה ברמת הקודונים. ציון זה שונה אמנם בכל שיטה, אך במהותו הוא מבוסס בכולן על ערך מתאם הקורלציה בין זוג אתרים. הציון המשולב מחושב ע״י חלוקה של הציון ברמת חומצות האמינו בציון ברמת הקודונים, בהתבסס על הנחת המוצא הברורה מאליה, לפיה חלוקה זו תתעדף מקרים בהם יש קורלציה חזקה ברמת חומצות האמינו וקורלציה חלשה ברמת הקודונים, ובכך יעלה הסיכוי להציף מקרים של קשר ישיר בין שני אתרים בחלבון. בהשוואה שמתבצעת במסגרת עבודת מחקר זו בנוגע לקבוצה גדולה של משפחות חלבונים, ניתן לראות כי הציון המשולב מחזיק בצדו יתרון משמעותי אל מול הציון שמבוסס רק על רמת חומצות האמינו במספר שיטות שונות של ניתוח מוטציות בקורלציה. בנוסף, שילוב זה מנבא בהצלחה מספר רב של קשרים ישירים בין אתרים בחלבונים הנמצאים רחוק אחד מהשני ברצף – קשרים, אשר אינם עולים בשיטות המשתמשות במידע ברמת חומצות האמינו בלבד ומחזיקים בצדם חשיבות יתרה לניבוי מבנה שלישוני של חלבונים.



שרטוט 2. דוגמא לקורלציה בין זוג אתרים בעימוד מרובה רצפי חומצות אמינו ועימוד הקודונים המתאים לו. קורלציה ברמת חומצות האמינו באתר i ו- j יכולה להתאים למצב שבו הקודונים המתאימים הם בקורלציה (צד שמאל למעלה) או לאו (צד ימין למעלה). הנחת היסוד המוצגת בעבודת מחקר זו היא שקורלציה ברמת חומצות האמינו יכולה לשקף קשר ישיר בין שני אתרים בחלבון בסיכוי יותר גבוה כאשר ברמת הקודונים הקורלציה באתרים אלה היא חלשה.

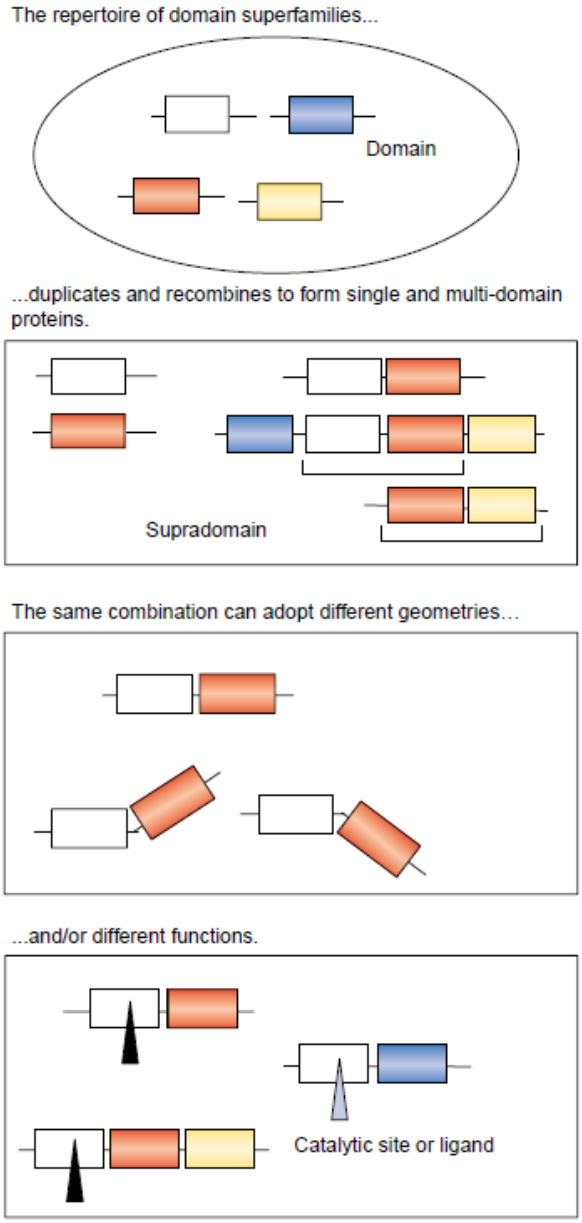
פרק שני

התפתחותם של אורגניזמים מורכבים במהלך האבולוציה, לוותה בגידול במבחר החלבונים היכולים למלא פונקציות שונות. מקום חשוב בתהליך ההתפתחות מילאו הדומיינים (protein domain), אשר לעיתים קרובות מכונים גם אבני הבניין של החלבונים. דומיין הוא יחידה אבולוציונית של רצף מקודד (בדר״כ באורך של 100-250), אשר עוברת שכפול ורקומבינציה (שרטוט 3). לעיתים קרובות דומיין ממלא תפקיד ביולוגי באופן עצמאי או במשותף עם דומיין או דומיינים אחרים. בנוסף, בחלק לא מבוטל מהמקרים, הדומיין מתקפל באופן עצמאי למבנה קומפקטי העומד בפני עצמו. שכיחותם של הדומיינים בטבע גבוהה, כמחצית מהפרוטאומים של יצורים רבים מורכבים מחלבונים בעלי שני דומיינים לפחות. כחמישה אחוז מאותם חלבונים מרובי דומיינים אף מורכבים משישה דומיינים ויותר.

חלבונים מרובי דומיינים נוטים יותר לאגרגציה מחלבונים בעלי דומיין יחיד. הסיבה לכך היא כי הריכוז החלבוני בסביבתו הקרובה של כל דומיין הינו גבוה ומעלה את הסבירות למגעים לא רצויים בין דומיינים שכנים, אשר גורמים לקיפול לקוי של החלבון. באופן כללי, פוטנציאל הנזק של קיפול לקוי של חלבונים הנו גבוה: הוא גורם לאגרגציה שמיוחסת בין השאר למחלות רבות, ביניהם אלצהיימר וסוכרת מסוג 2; והוא גורם למחסור בחלבונים הנדרשים לתפקוד התא ולהשקעת אנרגיה לשווא בייצורם, פינויים ופירוקם של החלבונים הפגומים שנוצרו.

בנסיבות אלה, לא בכדי קיימות ראיות כי במהלך האבולוציה התרחשה סלקציה חזקה למניעת קיפול לקוי של חלבונים בכל שלוש ממלכות החיים. לדוגמא, נמצא כי יש קורלציה שלילית בין רמות הרנ״א שליח לבין הנטייה לאגרגציה שככל הנראה נובעת מהצורך למזער נזקים הנובעים בעקבות קיפול לקוי. הימנעות מאגרגציה חשובה בפרט כאשר מדובר בחלבונים מרובי דומיינים, מכיוון שכפי שהוזכר לעיל – הם מהווים חלק לא מבוטל מהחלבונים בתא וגם בעלי נטייה מוגברת לאגרגציה.

במהלך השנים נעשו ניסיונות להבין באילו מנגנונים משתמש התא ע״מ להימנע מאגרגציה של חלבונים מרובי דומיינים. מנגנון אחד שעלה כאפשרי הוא קיפול במהלך התרגום שתפקידו למנוע אגרגציה בין דומיין שזה הרגע תורגם עם דומיינים שכנים שכבר תורגמו באותה שרשרת פפטידית. צ׳פרונים (חלבונים מלווים) גם הם יכולים לסייע לקיפול של דומיין אחרי דומיין, ואכן בתא (בפרט תאים איוקריוטים) יש מערך צ׳פרונים משמעותי המסייע בין השאר בתהליכי קיפול חלבונים ומניעת אגרגציה. מנגנון נוסף שהוצע למניעת אגרגציה של חלבונים מרובי דומיינים הוא סלקציה של דומיינים שכנים עם דמיון רצפי נמוך ביניהם. הסיבה לכך היא שדמיון רצפי גבוה בין דומיינים סמוכים מעלה את הסיכויים לאינטראקציות ביניהם במקום האינטראקציות הנדרשות בתוך הדומיין לקיום מבנה יציב ועצמאי. חלבונים מרובי דומיינים ללא ספק מוסיפים רמה נוספת של מורכבות למערכת שלא הייתה קיימת בחלבונים בעלי דומיין בודד והבנתנו כיום כיצד חלבונים מרובי דומיינים נמנעים מקיפול לקוי היא חלקית. במסגרת עבודת מחקר זו נשתמש בידע על רצף, מבנה וביטוי של חלבונים ע״מ לחפש דרכים נוספות שבעזרתן נמנע קיפול לקוי. להלן יוצע מנגנון למניעת אגרגציה שלא הוכר בעבר ומבוסס על אורך החלבון.

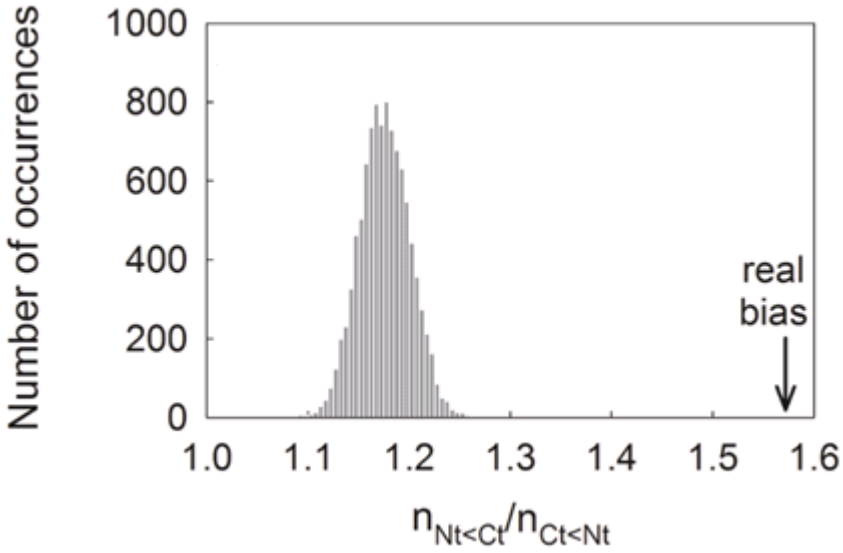


שרטוט 3. תפקיד הדומיינים כאבני הבניין של חלבונים. חלבונים מרובי דומיינים נוצרים ע״י שכפול ורקומבינציה. דומיינים השייכים לאותה משפחה מיוצגים כמלבנים באותו צבע. קומבינציה של דומיינים בסדר מסוים יכולה לשמש לפונקציה מסוימת ולהופיע בחלבונים שונים (supradomain). השרטוט לקוח מעבודה שנעשתה בעבר ע״י אחרים (Vogel et al., 2004).

אורך שרשרת החלבון הוא נתון בסיסי שדורש מידע מינימלי בלבד ע״מ להעריכו – מידע כדוגמת משקלו של החלבון או רצפו. למרות פשטותו, אורך החלבון יכול לספק לנו הרבה ידע על תכונותיו – ובכלל זה: אורך החלבון מספק הערכה בנוגע למהירות בה החלבון מתקפל והערכה בנוגע לרמות הביטוי היחסי שלו בתא (הערכה המבוססת על ניתוח נתונים שנעשה על מדידות של ביטוי חלבונים רבים והשוואה לאורכם). לכן, שימוש בנתון האורך יכול לספק לנו מידע על מספר עצום של חלבונים.

במסגרת עבודת מחקר זו, נמצא כי בחלבונים בעלי שני דומיינים קיימת נטייה חזקה לדומיינים בקצה האמיני להיות קצרים משכניהם בקצה הקרבוקסילי (שרטוט 4). הסבר אפשרי לנטייה זו מגיע מהקשר ההופכי בין אורך החלבון לקצב הקיפול שלו – קיימת סלקציה בטבע, כך שדומיינים בקצה האמיני יתקפלו מהר יותר משכניהם בקצה הקרבוקסילי. על מנת לוודא כי נטייה זו אינה ייחודית לאורגניזמים ספציפיים או למשפחה מסוימת של חלבונים, הורחבה הבדיקה של אורכי הדומיינים לאלפי פרוטאומים שלמים של אורגניזמים ומשפחות החלבונים הידועות. אכן, נטייה זו נמצאה ברוב הגורף של הפרוטאומים ובכל קבוצות החלבונים שנבדקו, ואף התגלה כי היא משמעותית יותר ביצורים פרוקריוטים מאשר איוקריוטים.

דרך נוספת להעריך את קצב הקיפול של חלבונים, היא מדד המכונה ״הערך המוחלט של סדר הקשרים״ (absolute contact order), אשר הנו ממוצע מספר חומצות האמינו ברצף המפריד בין שיירים הנמצאים בקשר פיזיקלי במבנה החלבון המקופל. במחקרים שנעשו בעניין זה נמצא, כי מדד זה נמצא בקשר הופכי לקצב הקיפול של החלבון. מניתוח שבוצע במסגרת עבודת מחקר זו בנוגע לחלבונים בעלי שני דומיינים עם מבנה שלישוני פתור – נמצא כי יש נטייה לדומיינים בקצה האמיני להיות עם ערך מוחלט של סדר הקשרים נמוך מהדומיינים בקצה הקרבוקסילי הסמוך להם. ממצאים אלה תומכים, באופן בלתי תלוי, בהנחה שיש סלקציה בטבע, כך שדומיינים בקצה האמיני יתקפלו מהר יותר מהדומיינים השכנים בקצה הקרבוקסילי. סיבה אפשרית לסלקציה כה משמעותית יכולה להיות מניעת קיפול לקוי של חלבונים בעלי שני דומיינים – תהליך שיכול להביא לאגרגציה. כפי שהוזכר לעיל, על מנת שהתא ימזער ככל הניתן נזקים הנובעים מאגרגציה, נוצר לחץ אבולוציוני שהביא למצב בו לחלבונים עם רמות ביטוי גבוהות תהיה נטייה פחותה יותר לאגרגציה מחלבונים עם רמות ביטוי נמוכות. אכן, בניתוח רחב היקף שבוצע במסגרת עבודת מחקר זאת, נמצא כי חלבונים בעלי שני דומיינים עם דומיין בקצה האמיני הקצר מהדומיין השכן בקצה הקרבוקסילי, הם עם רמות ביטוי יותר גבוהות. ממצאים אלה תומכים בהשערה, לפיה הנטייה של דומיינים בקצה האמיני להיות קצרים מהדומיינים בקצה הקרבוקסילי בחלבונים בעלי שני דומיינים – נובעת מסלקציה נגד קיפול לקוי. הממצאים לפיהם נטייה זו חזקה יותר בפרוקריוטים מאשר איוקריוטים יכולים לרמוז על דרך ההתמודדות של התא הפרוקריוטי עם בעיית האגרגציה בהעדר מערך הצ׳פרונים המקיף שקיים בתא האיוקריוטי. לסיכום, ממצאים אלו מציעים מנגנון שטרם הוכר בעבר למניעת אגרגציה בין דומיינים סמוכים בחלבונים מרובי דומיינים.



שרטוט 4. הנטייה של חלבונים בעלי שני דומיינים בטבע לדומיין קצר בקצה האמיני לעומת מדגם אקראי. הציר האופקי מייצג את היחס בין מספר החלבונים בעלי דומיין קצר בקצה האמיני לעומת מספר החלבונים בעלי דומיין קצר בקצה הקרבוקסלי. המדגם האקראי נועד לשקף מצב בו לא הייתה סלקציה על קבוצת החלבונים הנמדדת. המדגם יוצר ע״י מדידת היחס כפי שבוצע בקבוצת החלבונים המקורית אך לאחר החלפות אקראיות בין הדומיינים בקצה האמיני תוך קיבוע הדומיינים בקצה הקרבוקסילי. פעולה זו בוצעה 10,000 פעמים ע״מ ליצור את ההיסטוגרמה בשרטוט. מספר החלבונים בקבוצה הוא כמעט 3000.

שני הנושאים בעבודת מחקר זו מבוססים על ניתוח רצפים של חלבונים, ברמת חומצות האמינו או הנוקלאוטידים. באמצעות שילוב בין הרצף ומקורות מידע אחרים, כמבנים שלישוניים ורמות ביטוי של חלבון התגלו מספר תובנות חדשות. כמויות הרצפים העצומות במספר רב של אורגניזמים מצד אחד, וחיבורם בתהליך הניתוח למקורות מידע אחרים שאפשרו לתמוך בהשערות המחקר מצד שני, הביאו לביסוסן של הנחות היסוד בעבודה זו שלא היה מתאפשר בדרך של ניסויי מעבדה בטכנולוגיה של ימינו. אין ספק שמרחב האפשרויות בניתוח מידע ביולוגי על בסיס הרצף טומן בחובו עוד הרבה אפשרויות שיביאו להבנה יותר טובה של הביולוגיה.

הפרק הראשון של העבודה התפרסם בכתב העת המדעי eLIFE:

Jacob, E., Unger, R. and Horovitz, A. (2015). Codon-level information improves predictions of inter-residue contacts in proteins by correlated mutation analysis. eLife 2015;4:e08932.

והפרק השני של עבודת המחקר התפרסם בכתב העת המדעי Cell Reports:

Jacob, E., Unger, R. and Horovitz, A. (2013). N-Terminal Domains in Two-Domain Proteins Are Biased to Be Shorter and Predicted to Fold Faster Than Their C-Terminal Counterparts. Cell Rep. 3 (4), 1051-1056.

תוכן עניינים

עבודה זו נעשתה בהדרכתם של

פרופ׳ רון אונגר

מהפקולטה למדעי החיים של אוניברסיטת בר-אילן.

ופרופ׳ אמנון הורוויץ

מהמחלקה לביולוגיה מבנית של מכון וויצמן למדע.

מרצף למבנה – ניתוח רצפים להבנת מבנה וקיפול של חלבונים

חיבור לשם קבלת התואר "דוקטור לפילוסופיה"

מאת:

איתי יעקב

הפקולטה למדעי החיים ע"ש מינה ואברד גודמן

הוגש לסנט של אוניברסיטת בר-אילן

רמת גן שבט, תשע"ו