Radboud University Nijmegen

Master thesis information science

Improvements in structural contact prediction: opportunities in prediction difficulty and pairing preference of amino acids

Author:
Martijn Liebrand

Supervisor:
dr. Elena Marchiori

January, 2014

Abstract

Proteins are complicated molecules and their three-dimensional structure provides information about their functionality. These three-dimensional structures are determined by specific amino acid sequences. Hidden information in amino acid sequences, in particular homologous sequences, may offer possibilities in obtaining the three-dimensional structures of proteins. A common method to identify protein folding, is by using information about amino acid residue-residue contacts. These contact pairs can, in turn, be found by correlated mutations. If one residue changes by a mutation, its contacting counter partner will likely be mutated too, ensuring the native fold a protein.

Covariance in multiple sequence alignments offers a method in finding residue-residue contacts. PSICOV is a tool which tries to find residue-residue contacts by using a sparse inverse covariance estimation. This study investigates opportunities in further improvements in finding contact pairs using PSICOV, based on the assumption of improved predictions by using specific amino acid characteristics.

After an extensive study of the predictions made by PSICOV, it was possible to calculate higher mean precision values when prediction difficulty or pairing preferences of amino acids was used. An improvement in mean precision up to 0.03 can be made by a linear transformation of PSICOV predictions. It demonstrates that precise structural contact prediction can be further improved by a combination of machine learning algorithms and amino acid characteristics. In addition, this study shows the beginning of more hybrid residue-residue contact predictions tools. One of the conclusions of this thesis is that there is still a lot of profit to be made in this research field.

Preface

"Information science (or information studies) is an interdisciplinary field primarily concerned with the analysis, collection, classification, manipulation, storage, retrieval, movement, and dissemination of information."[1] Although this master thesis is about a topic within bioinformatics, it is closely related to the information science. I will discuss a research that deals with all the aforementioned (information science) concepts. Information science is widely applicable and in many cases the context changes, but the approaches or techniques are similar. Data mining or machine learning techniques are used in information science studies as well as in bioinformatics. This does not mean that biofinformatics is a redundant profession. As information scientists we certainly lack essential knowledge of (molecular) biology and therefore the focus in this thesis will be on the techniques and the improvements on the techniques.

[1] http://en.wikipedia.org/wiki/Information_science

Acknowledgments

I am grateful to everyone who helped and supported me in writing my master thesis. At first I would like to specially thank my supervisor dr. Elena Marchiori for the given opportunity to write my master thesis under her supervision. I found our discussions pleasant and very motivating. I would also like to thank prof. dr. Gert Vriend for taking time to answer my questions. Finally I would like to thank my girlfriend Erin, my parents and fellow students who each in their own way supported me.

Nijmegen, January 2014

Contents

List of Tables

List of Figures

Chapter 1

Introduction to structural contact prediction

1.1 Introduction

Nowadays, bioinformatics is an established research discipline. An increasing number of scientist are aware of the need for bioinformaticians. Like the fast growing amount of information on the internet, biological data is growing explosively. All the information of how we (our bodies/life in general) are build is described in our DNA. Consequently, the cause of serious diseases are frequently the consequence of harmful mutations in DNA. If people have serious diseases, it seems logical that we will look for an explanation in the main source of our body. We are able to obtain our main source by several sequencing techniques and this is becoming much cheaper and faster. The acceleration of DNA sequencing and lower sequencing costs in the past years can be compared with the exponential growth of the number of resistors on a computer chip (see table 1.1). In biology, the amount of data is growing at least as fast as computers did in the last few decades. Extracting the right

Date	Cost per Mb of DNA Sequence	Cost per Genome
September-2001	$5,292.39	$95,263,072
March-2002	$3,898.64	$70,175,437
September-2002	$3,413.80	$61,448,422
March-2003	$2,986.20	$53,751,684
October-2003	$2,230.98	$40,157,554
...	...	...
...	...	...
October-2010	$0.32	$29,092
January-2011	$0.23	$20,963
April-2011	$0.19	$16,712
July-2011	$0.12	$10,497
Jan-2012 (EST)	$0.09	$7,950

Table 1.1: A snippet of sequencing costs from the beginning of this century until 2012. (From: http://dnasequencing.org/history-of-dna)

information out of all the sequence data is a crucial and difficult task for a bioinformatician.

When we speak of sequence data, we do not necessarily mean DNA sequences. Sequence data exists at multiple levels such as DNA, RNA, (poly)peptides or proteins. In this thesis I will discuss an important piece of information that is hidden in protein sequences. An important topic within the field of bioinformatics is to make an accurate prediction of how proteins fold using nothing else than protein sequences and algorithms. How proteins fold and knowing the three-dimensional shape is essential for determining out how proteins function [3]. If we are able to make an accurate prediction of the three-dimensional shape of a protein using a computer, we may save a lot of time and money spending on experiments. To give an impression: as human beings we have over 20 thousands genes.[1] Automating the prediction of protein folding will be major breakthrough in biological research.

1.2 What is structural contact prediction?

A protein is a large molecule consisting of chains of individual amino acids. There are twenty different amino acids and they can occur more than once within a protein sequence. The reason why some amino acids appear in a particular order is not because of random chance. Each amino acid has his own special characteristics and all amino acids of a protein together define the function or functions of that protein [3].

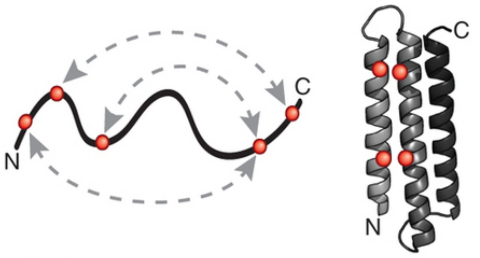


Figure 1.1: An image of residue-residue contacts (red dots), from sequence (left) to the three-dimensional structure (right). Figure from [17].

In literature it is established that we speak of residue-residue contacts if two amino acids of one protein sequence are connected together in the three-dimensional structure [17]. Within the three-dimensional structure the two residues are neighbours, but these residues may be far away from each other in the primary structure of a protein. As you can see in figure 1.1 the residue-residue contacts of the example are not located close to each other in the sequence strand (left), and one residue can make contact with one

[1] http://en.wikipedia.org/wiki/Human_genome

or more residues. In general there is a lot of variation in proteins and their structures.

Precise structural contact prediction is at this moment a problem within the bigger problem of the prediction of the actual three-dimensional structure of a protein. Residue-residue contacts is closely involved in guiding protein folding and maintaining the native fold of proteins [11, 13]. In addition to that, we also know that it is possible to elucidate the fold of the protein with sufficient correct information about a protein's residue-residue contacts [9, 13, 20]. Keeping these facts in mind, we can think of a method to predict the native fold of a protein using only the sequence. If we can extract the residue-residue contacts out of a sequence, we can consequently construct a complete protein.

1.3 Challenges in contact prediction

At this moment a key question is how residue-residue contacts can be found. The basis on which almost all contact prediction research leans is about the fact that proteins are rarely unique. Grishin states:"*[...] similar sequences typically yield similar three-dimensional structures, and experimentally determined structure for one family member offers reliable structure prediction for the rest of its members* [10]." In other words, important information about the three dimensional structure of proteins can be found within the information of homologous families. In this study we will not compare complete families with a specific experimentally determined protein. The idea is to extract residue-residue contacts using only the homologous families. Another critical piece of information protein families shares, is the evolutionary signals in sequences. How can we derive residue-residue contacts using the evolutionary background of a protein family?

Successful approaches attempt to extract contact information from multiple sequence alignments [13]. This information can be found by investigating correlated mutations. Several research groups demonstrated that extracting covariation information from sequences is sufficient to predict a protein fold to reasonable accuracy [17]. Table 1.2 shows a simple example of how you can recognize correlated mutations. For a more biological background such as the idea behind multiple sequence alignments I refer to chapter two.

"The underlying rationale rests on the fact that any given contact critical for maintaining the fold of a protein will constrain the physicochemical properties of the amino acids involved. Should a given contacting residue mutate and potentially perturb the properties of the contact, then its contacting partner will be more likely to mutate to a physicochemically complementary amino acid residue, to ensure the native fold of the protein remains stabilized [13]." As shown in table 1.2 the residues of column two, five and seven

	Residue Nr.							
	1	2	3	4	5	6	7	8
Sequence 1	A	A	A	A	A	A	C	A
Sequence 2	A	E	A	A	D	A	A	A
Sequence 3	A	A	A	A	A	A	A	A
Sequence 4	A	E	A	A	D	A	C	A
Sequence 5	A	E	A	A	D	A	C	A

Table 1.2: Simple correlated mutation example. The colored columns indicates possible correlated mutations. When the second residue varies in the protein family, other residues (eg. 5) may change as well.

may have some covariance. If one residue is mutated, other residues may change because the protein may need to ensure its stability. The biggest challenge within the prediction of the right residue-residue contacts is handling false positives based on phylogenetic effects or indirect coupling [13]. False positives that are observed, for example, when two residues contact the same third residue but do not actually contact each other. This is called the chaining effect [15, 17]. The chaining effect explained with reference to table 1.2, it means that the two blue residues (column 2 and 7) both contacts with the red residue (5) and therefore a transitive influence can be observed between the blue residues.

1.4 Problem Statement

In this thesis I will go deeper into PSICOV, a specific method to predict co-evolution. PSICOV uses a sparse inverse covariance estimation for amino acid residue-residue contact prediction. I will explain the functionalities of PSICOV and perform experiments considering the data provided by the authors of PSICOV [13]. For benchmarking purposes we have a golden truth for each target protein. After running experiments with PSICOV, it might be interesting to look at outliers of residue contact prediction. Are there any amino acids or amino acid pairs which PSICOV predict exceptionally good or bad? Can we explain these results and are we able to optimize PSICOV in such a way that its predictions are more accurate? One possibility is weighting of specific output values.

Chapter 2

Biological sequence data and residue-residue contacts

In this chapter I will discuss about biological processes and the meaning of biological data like sequence data. The purpose of this chapter is to give basic background information in such way that non-biologists (like informatics and information science students) obtain a better understanding of the research topic.

Sequence data exists at multiple levels. In order to emphasize the distinction between different types of sequence data, I show very briefly the central dogma of biology. The translation of the DNA code into a protein. After the gene-protein introduction I will explain how multiple sequence alignments (MSA) function. MSA is a common bioinformatics method for analyzing sequence data. Finally I discuss how we can extract from a MSA interesting information like residue-residue contacts.

2.1 From gene to protein

The genetic instructions for a polypeptide chain are written in the DNA as series of non overlapping three-nucleotide codons (combinations of nucleotides) [4]. 64 codons function as a base for 20 amino acids where several codons can lead to the same amino acid (or residue). Figure 2.1 shows that, for instance, the three-nucleotide code ACC results in the amino acid tryptophan. The translation from gene to protein has several steps and each step results in another type of sequence. Transcription of a gene, by so-called RNA polymerase, leads to mRNA (messenger RNA) which is the messenger of genetic information. After the translation (by ribosomes) of mRNA into amino acid chains we have a polypeptide chain which is not necessarily a protein, although the terms “protein” and “polypeptide” are sometimes used interchangeably [19]. Proteins can consist of long polypeptide chains.

Figure 2.1 suggests that there are three types of sequences, i.e. DNA,

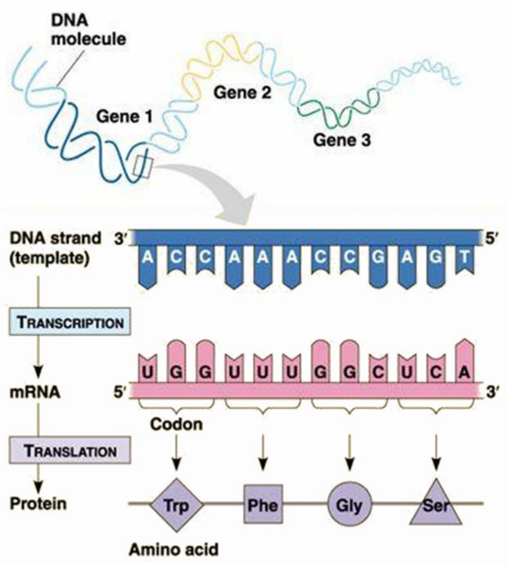
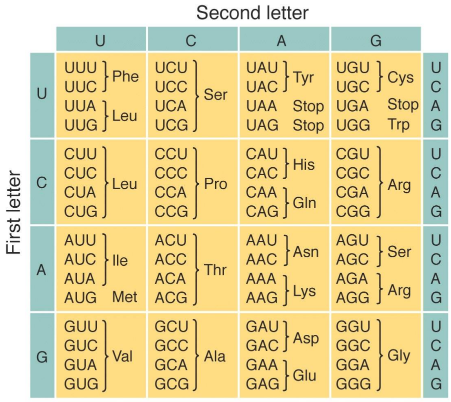


Figure 2.1: Central dogma in (molecular) biology for generating of proteins out of genes. Figure from [4].

mRNA and protein. The function of each protein molecule depends on its three-dimensional structure. The three-dimensional structure is determined by its amino acid sequence, which is in turn is determined by the nucleotide sequence of the structural gene [3]. For this thesis the focus lies on protein sequences, because finding residue-residue contacts in protein sequence may be a way to extract the three-dimensional structure.

2.1.1 The amino acid dictionary

Figure 2.2 shows the composition of nucleotides for specific amino acids. Some codons lead to the beginning mRNA translation (by the amino acid methionine), and some combinations of nucleotides are stop codons [19]. There are two common abbreviations for amino acids. A notation by one or three letters. For investigation of protein sequences we use the one-letter code of amino acids.



	Second letter								
First letter	U		C		A		G		Third letter
U	UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys	U
	UUC		UCC		UAC		UGC		C
	UUA	Leu	UCA		UAA	Stop	UGA	Stop	A
	UUG		UCG		UAG	Stop	UGG	Trp	G
C	CUU	Leu	CCU	Pro	CAU	His	CGU	Arg	U
	CUC		CCC		CAC		CGC		C
	CUA		CCA		CAA	Gln	CGA		A
	CUG		CCG		CAG		CGG		G
A	AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser	U
	AUC		ACC		AAC		AGC		C
	AUA		ACA		AAA	Lys	AGA	Arg	A
	AUG	Met	ACG		AAG		AGG		G
G	GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly	U
	GUC		GCC		GAC		GGC		C
	GUA		GCA		GAA	Glu	GGA		A
	GUG		GCG		GAG		GGG		G

Amino acid	3-letter	1-letter
Alanine	Ala	A
Arginine	Arg	R
Asparagine	Asn	N
Aspartic acid	Asp	D
Cysteine	Cys	C
Glutamic acid	Glu	E
Glutamine	Gln	Q
Glycine	Gly	G
Histidine	His	H
Isoleucine	Ile	I
Leucine	Leu	L
Lysine	Lys	K
Methionine	Met	M
Phenylalanine	Phe	F
Proline	Pro	P
Serine	Ser	S
Threonine	Thr	T
Tryptophan	Trp	W
Tyrosine	Tyr	Y
Valine	Val	V

Figure 2.2: The amino acid dictionary. All codons resulting in amino acids. (From: http://www.lucasbrouwers.nl/blog/2010/11/the-algaes-accent/ & wikipedia.org/wiki/Amino_acid)

2.2 Multiple sequence alignment (MSA)

When we consider a protein, one of the most fundamental questions is which other proteins are related to it, because biological sequences often occur in families. By introducing sequences into a multiple sequence alignment, we can define members of a gene or protein family. *"The function of most proteins is assigned on the basis of homology to other known proteins rather than on the basis of results from biochemical or cell biological assays* [22]." How does a MSA work?

2.2.1 Principles of a sequence alignment

DNA and protein sequences change during evolution. Nucleotides and the amino acids they encode can change as result of point mutations, and sequence lengths can be quite different as a result of insertions and deletions. To find underlying similarities, an alignment is needed to maximize their similarities [25]. The following example shows a simple alignment and how to handle evolutionary changes.

T	H	I	S	S	E	Q	U	E	N	C	E
|	|			|	|	|	|	|	|	|	|
T	H	A	T	S	E	Q	U	E	N	C	E

Table 2.1: Simple alignment with two hypothetical amino acid sequences. Example from [25].

Table 2.1 shows a strong similarity between two sequences. The identical letters are highlighted in red after the compared sequences are lined up in the best possible way. This is a very simple example, what happens if two sequences differ more from each other? Look at the following sequences, THATSEQUENCE and THISISASEQUENCE.

T	H	A	T	S	E	Q	U	E	N	C	E			
|	|							|						
T	H	I	S	I	S	A	S	E	Q	U	E	N	C	E

After introducing gaps:

T	H	–	–	–	–	A	T	S	E	Q	U	E	N	C	E
|	|					|		|		|	|	|	|	|	|
T	H	I	S	I	S	A	–	S	E	Q	U	E	N	C	E

Table 2.2: Introducing gaps into alignments. Example from [25].

You can immediately see that there are a lot of similarities between both sequences, but you are not able to find them by simply sliding one sequence over another. Introducing gaps into alignments is a way to handle false matches due to different sequence lengths. In table 2.2 you can see how introducing gaps improves an alignment.

A multiple sequence alignment is essentially series of pairwise alignments between, for instance, a group of proteins. There are several approaches and programs available. For more information I refer to Bioinformatics and Functional Genomics and Understanding Bioinformatics [22, 25].

2.2.2 MSA and residue-residue contacts

What are correlated mutations, residue-residue contacts and how are these two concepts related? Before explaining this relation, you need some knowledge of the basic building principles of proteins. It is not necessary to go deeply into protein structures, but it is good to know why we are interested in the relation between correlated mutations and residue-residue contacts.

Basic structural principles

When we discuss protein structures, it is not always clear what we mean. We make a distinction between different levels of structures when discussing proteins. For a better understanding of this thesis there is no need to know all the details about amino acids or proteins, but to avoid confusion it is required to know about the four levels of protein structures.

After reading the first part of this chapter you should be familiar with the primary structure of a protein. It is the amino acid sequence of a protein's polypeptide chain. Such a polypeptide chain (or different regions of it), may result into units of secondary structures such as helices or strands (see figure 2.3). Combinations of secondary structures may, in turn, form (parts of) a tertiary structure. The final protein may contain several subunits (tertiary structures). This is the quaternary structure. Amino acids located far apart in the sequence are brought close together in the three-dimensional structure, and may form a functional region [3, 19].

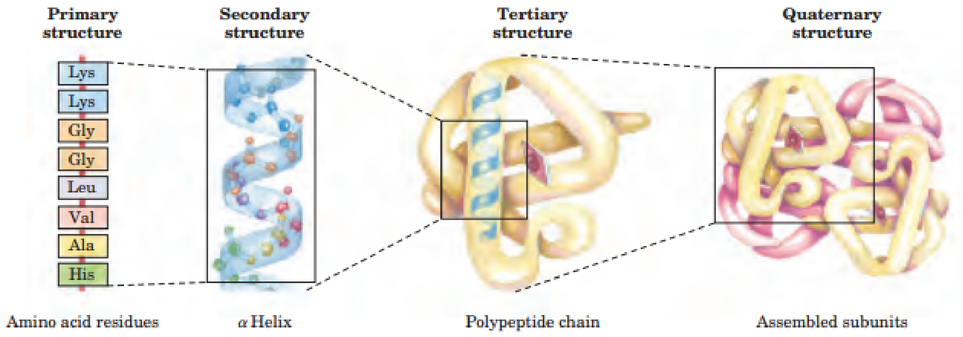
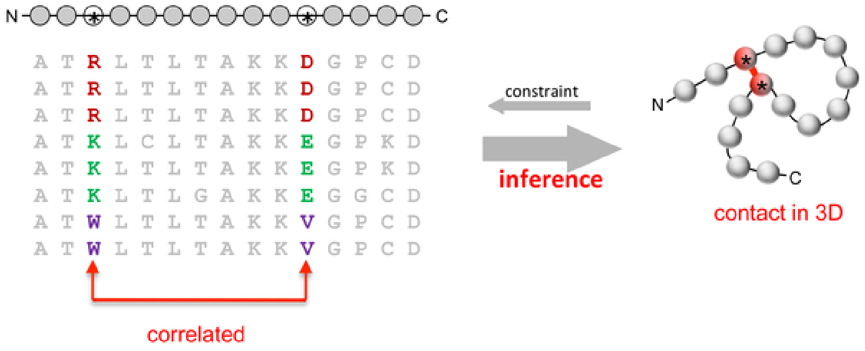


Figure 2.3: Protein structures. Figure from [19].

MSA and correlated mutations

The main problem is, how can we make an accurate prediction of tertiary or quaternary protein structures out of protein sequences? There is much potential in this research field if you imagine using great protein (family) databases like PFAM. PFAM contains nearly 12.000 protein families with a growing number of families over 100.000 sequences [6]. We can extract (evolutionary) information out of multiple sequence alignments using protein sequences and its family.

Correlated mutations are based on the idea that most proteins want to maintain its stability. For instance, amino acids that are brought close together to form an active site, do not mutate in a single way. If one residue changes, then its contacting partner will be more likely to mutate to a physicochemically complementary amino acid residue, to ensure the native fold of the protein [13]. Are we able to find the so called correlated mutations in multiple sequence alignments? We have a large amount of sequence data of protein families, and we may be able to see these evolutionary changes by means of correlated mutations in MSA's.



```
A T R L T L T A K K D G P C D
A T R L T L T A K K D G P C D
A T R L T L T A K K D G P C D
A T K L C L T A K K E G P K D
A T K L T L T A K K E G P K D
A T K L T L G A K K E G G C D
A T W L T L T A K K V G P C D
A T W L T L T A K K V G P C D
```

Figure 2.4: Using a multiple sequence alignment to find correlated mutations. Figure From [16].

Finding co-evolution in multiple sequence alignments is shown in figure 2.4. Marks et al. states: "*The sequence of the protein for which the 3D structure is to be predicted [...] is part of an evolutionarily related family of sequences that are presumed to have essentially the same fold* [16]." The colors red, green and purple are used to indicate the two residues that seems to have correlated mutations. If the one residue mutates, it seems to affects another residue. This could mean that, because the two residues covary, the residues are connected in their three-dimensional structure.

Recap

There are several types of sequence data. Multiple sequence alignments are useful for collecting more information about sequences or proteins thanks to homologous families. From sequence variation in MSA's we want to extract correlated mutations, because the involved residues may have a connection in the three-dimensional structure of a protein. A sufficient amount of residue-residue contacts offers a possible way to predict a protein's tertiary or quaternary structure from his primary structure.

Chapter 3

Protein structure prediction from sequence variation

3.1 Local versus global statistics

There are several statistical models for predicting co-evolution between protein residues. We can roughly divide all the available approaches into two groups: local and global statistical models. The devision of approaches of co-evolution prediction depends on the fundamental principles of the approaches.

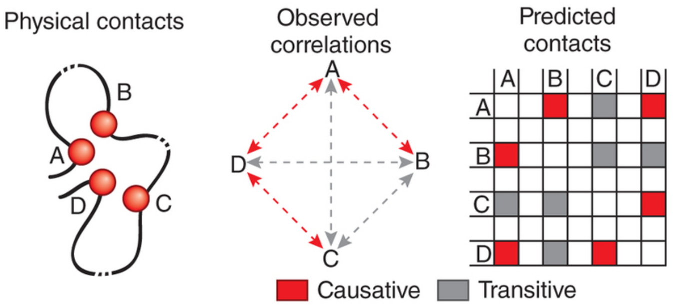


Figure 3.1: Principles of confounding effects. There are causative correlations between residues A-B, A-D and D-C, because of direct interactions. Transitive correlations can be found between, for instance, residues B-D, because of their direct interactions with residue A. Figure from [17].

In chapter 1 I mentioned the chaining effect and figure 3.1 shows the underlying confounding problem. True evolutionary covariation can be masked by transitive correlations and sometimes transitive correlations are even stronger than causative correlations [17]. Presence of common neighbors of two noninteracting residues could be a reason for transitive correlations. The assumption of local statistical models is that pairs of residue posititions are statistically independent of other pairs of residues. This makes local approaches less successful in contact identification for the prediction of three-dimensional structures of proteins, because they do not take into account that transitive correlations causes noise [17].

In contrast to local statistical models, global models assume that correlated residue pairs are dependent on each other. Residue pairs that are high globally correlated are more likely to be true residue couples. Marks et al. states: "*[...] predicted contacts based on the global probability models provide a base for the computation of three-dimensional folds* [17]." In table 3.1 an overview of local and global statistical models used for prediction of three-dimensional structures or residue-residue contacts can be found.

	Method	Statistics
Global (3d folds)	EVfold, EC's	Maximum entropy
	EVfold-transmembrane	Maximum entropy
	DCA-fold	Maximum entropy
	FILM3	Partial correlations
Global (contacts)	Boltzmann network model	Maximum entropy
	Bayesian network model	Conditional ratio of spanning trees
	PSICOV	Sparse inverse covariance estimation
	DCA-BP	Maximum entropy, belief propagation
	DCA-mean field	Maximum entropy
Local	Correlated mutation analyses	Correlations
	MI, SCA, McBasc, OMES	(Weighted) mutational information, substitution correlations observed minus expected
	MIp	Phylogeny-corrected mutational information
	SCA	Weighted mutual Information

Table 3.1: Statistical models that are used for prediction of co-evolution. We can distinguish global models in two groups. Prediction of three-dimensional folds and prediction of residue-residue contacts. Local models focus mainly on residue-residue contacts using mutational information. Table from [17].

3.2 PSICOV

PSICOV is a tool that is based on a global statistical approach and tries to tackle the problem of transitive correlations. Using a sparse inverse covariance estimation PSICOV takes transitive coupling into account [13]. The source code is freely available and supplementary data is available at Bioinformatics online.[1] For many reasons is PSICOV the most logical choice when it comes to a master thesis research. The reasons are the availability and possibilities to install the software anywhere and the alternative statistical model which is not too complex, but tries to handle transitive couplings. In addition, we may be able to adapt the predictions after running experiments.

[1] http://bioinfadmin.cs.ucl.ac.uk/downloads/PSICOV/

Chapter 4

Inferring directly coupled sites using covariance

4.1 Mutational information

Calculating the mutational information between two specific sites can provide interesting information. It is a common method for identifying correlated mutations [13].

$$MI = \sum_{ab} f(A_iB_j) log \frac{f(A_iB_j)}{f(A_i)f(B_j)} \tag{4.1}$$

The idea is to calculate with the frequency of the observed amino acid combination $f(A_iB_j)$ and the frequencies of the amino acids independently. By using the following example (a very small MSA) I will explain a major problem when calculating the correlated mutations by mutational information.

1	2	3
A	A	A
F	C	A
A	D	E

Assume that 1-2 and 2-3 have some direct interaction, but the combination of 1-3 has not. By calculating the mutational information between 1 and 3 we might observe a correlation:

$$MI^{13} = \frac{1}{3} log(\frac{\frac{1}{3}}{\frac{2}{3}\frac{2}{3}}) + \frac{1}{3} log(\frac{\frac{1}{3}}{\frac{1}{3}\frac{2}{3}}) + \frac{1}{3} log(\frac{\frac{1}{3}}{\frac{2}{3}\frac{1}{3}}) \approx 0.17 \tag{4.2}$$

Even when some sites are not connected in the three-dimensional structure of their protein you can find some mutational information and think there is some correlation.

4.2 The covariance matrix

To avoid the false positive observation of indirect couples, PSICOV uses a sparse inverse covariance estimation. First I explain the covariance estimation and the covariance matrix. Unfortunately, even by a really small MSA the covariance matrix will explode in size, therefore I am not able to prove that the covariance matrix handles indirect couples much better. The best way to explain this method is by a small MSA and a reduced set of amino acids.

1	2
A	A
D	C

Suppose we have a MSA of two sequences of two residues each and a set of three amino acids types {A,C,D}. We can compute the covariance between specific amino acid combinations at two given alignment sites.

$$S_{ij}^{ab} = \frac{1}{n}\sum_{k=1}^{n}(x_i^{ak} - \bar{x}_i^{a})(x_j^{bk} - \bar{x}_j^{b}) \tag{4.3}$$

The result (S) is the matrix representing the covariance of any amino acid combination at any combination of two sites. n is the number of sequences or rows in the MSA. x_i^{ak} and x_j^{bk} have a value of 0 or 1, depending on the presence of amino acid type a/b at column i/j at row k. If we want to know the covariance between amino acid A and C for column one and two from our example, we can calculate it in the following way:

$$\begin{aligned} S_{12}^{AC} &= \frac{1}{2}\sum_{k=1}^{2}(x_1^{Ak} - \bar{x}_1^{A})(x_2^{Ck} - \bar{x}_2^{C}) \\ S_{12}^{AC} &= \frac{1}{2}((1-\frac{1}{2})(0-\frac{1}{2}) + (0-\frac{1}{2})(1-\frac{1}{2})) = -\frac{1}{4} \end{aligned}$$

For $k = 1$ and $k = 2$, we subtract the frequency of the specific amino acid of the number that represents the presence of the same amino acid. If we calculate the covariance between A and C for all positions we can fill in a matrix:

$$S^{AC} = \begin{pmatrix} s_{1,1} & s_{1,2} \\ s_{2,1} & s_{2,2} \end{pmatrix}$$

We are not only interested in one amino acid combination at two sites, but in all the possible variations. The matrix grows enormously, even by an unrealistic MSA of sequences consisting of only two residues out of a subset of three AA-types.

$$\begin{pmatrix} s_{1,1}^{AA} & s_{1,2}^{AA} & s_{1,1}^{AD} & s_{1,2}^{AD} & s_{1,1}^{AC} & s_{1,2}^{AC} \\ s_{2,1}^{AA} & s_{2,2}^{AA} & s_{2,1}^{AD} & s_{2,2}^{AD} & s_{2,1}^{AC} & s_{2,2}^{AC} \\ s_{1,1}^{DA} & s_{1,2}^{DA} & s_{1,1}^{DD} & s_{1,2}^{DD} & s_{1,1}^{DC} & s_{1,2}^{DC} \\ s_{2,1}^{DA} & s_{2,2}^{DA} & s_{2,1}^{DD} & s_{2,2}^{DD} & s_{2,1}^{DC} & s_{2,2}^{DC} \\ s_{1,1}^{CA} & s_{1,2}^{CA} & s_{1,1}^{CD} & s_{1,2}^{CD} & s_{1,1}^{CC} & s_{1,2}^{CC} \\ s_{2,1}^{CA} & s_{2,2}^{CA} & s_{2,1}^{CD} & s_{2,2}^{CD} & s_{2,1}^{CC} & s_{2,2}^{CC} \end{pmatrix} = \begin{pmatrix} \frac{1}{4} & \frac{1}{4} & -\frac{1}{4} & 0 & 0 & -\frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} & -\frac{1}{4} & 0 & 0 & -\frac{1}{4} \\ -\frac{1}{4} & -\frac{1}{4} & \frac{1}{4} & 0 & 0 & \frac{1}{4} \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ -\frac{1}{4} & -\frac{1}{4} & \frac{1}{4} & 0 & 0 & \frac{1}{4} \end{pmatrix} \tag{4.4}$$

Matrix S (4.4) contains all the covariances of the three types of amino acids at any two sites of the simple MSA. Imagine how the matrix will look like if you want to calculate covariances for a MSA consisting of thousand sequences, over one-hundred residues each and twenty-one possible residues (including gaps).

If the covariance matrix is invertible, you can calculate the concentration matrix. The matrix above is not invertible (because the determinant is zero), but the concentration matrix (Θ) should look like this:

$$\Theta = \begin{pmatrix} \theta_{1,1} & \theta_{1,2} & \theta_{1,3} & \theta_{1,4} & \theta_{1,5} & \theta_{1,6} \\ \theta_{2,1} & \theta_{2,2} & \theta_{2,3} & \theta_{2,4} & \theta_{2,5} & \theta_{2,6} \\ \theta_{3,1} & \theta_{3,2} & \theta_{3,3} & \theta_{3,4} & \theta_{3,5} & \theta_{3,6} \\ \theta_{4,1} & \theta_{4,2} & \theta_{4,3} & \theta_{4,4} & \theta_{4,5} & \theta_{4,6} \\ \theta_{5,1} & \theta_{5,2} & \theta_{5,3} & \theta_{5,4} & \theta_{5,5} & \theta_{5,6} \\ \theta_{6,1} & \theta_{6,2} & \theta_{6,3} & \theta_{6,4} & \theta_{6,5} & \theta_{6,6} \end{pmatrix}$$

By using the inverse of the covariance matrix you can obtain the partial correlation matrix (4.5). The partial correlation matrix gives the correlation between any pair of amino acids at any two sites, conditional on the frequencies of amino acids at all other sites [13].

$$\rho_{ij} = -\frac{\Theta_{ij}}{\sqrt{\Theta_{ii}\Theta_{jj}}} \tag{4.5}$$

Jones et al. states that if the covariance matrix can be inverted, the inverse covariance matrix provides information on the degree of coupling between pairs of sites in the given MSA [13]. Excluding diagonal elements, numbers significantly different from zero may indicate pairs of sites which have strong direct coupling [13].

4.3 Sparse inverse covariance estimation

The size of a covariance matrix is $21m$ by $21m$, where m is the number of columns or residues (to keep the matrix manageable the example of 4.4 the matrix is 6 by 6 and not 42 by 42). Besides the size of the matrix, a second problem is the fact that the covariance matrix contains more variables than observations.

KVFGRCELAAAMKR-HGLDNYRGYSLGNWV-CAAKFESNFNTQATN
KVFGRCELAAAMKR-HGLDNYRGYSLGNWV-CAAKFESNFNTQATN
KVFGRCELAAAMKR-HGLDNYRGYSLGNWV-CAAKFESNFNTQATN
KVFGRCELAAAMKR-HGLDNYRGYSLGNWV-CAAKFESNFNTQATN
KVFGRCELAAAMKR-HGLDNYRGYSLGNWV-CAAKFESNFNTQATN
KVFGRCELAAAMKR-HGLDNYRGYSLGNWV-CAAKFESNFNTQATN
KVFGRCELAAAMKR-HGLDNYRGYSLGNWV-CVAKFESNFNTQATN
KVFGRCELAAAMKR-HGLDNYRGYSLGNWV-CVAKFESNFNTQATN
KVFGRCELAAAMKR-HGLDNYRGYSLGNWV-CAAKFESNFNSQATN

Table 4.1: Variables vs. observations. In a MSA, some sites have more variation than others. If some residues do not appear at specific sites, than you have more variables than observations (blue vs. red).

Perhaps you noticed many zero's in the covariance matrix of 4.4. The explanation for this can be found in table 4.1. Not every amino acid will be observed at every site, even in very large families, and thus there will be more variables than observations [13]. For instance, if residue K never appears at column 17 and residue L never appears at column 33, $S_{17,33}^{KL}$ will be zero. Thanks to the variables vs. observation problem, the covariance matrix cannot be directly inverted.

The problem is solved by using the graphical lasso method of the sparse inverse covariance estimation. The algorithm is implemented in PSICOV, but a software package written in R is also freely available.[1] According to Jones et al., "*where an inverse covariance estimate is constrained to be sparse, the non-zero terms tend to more accurately relate to correct positive correlations in the true inverse covariance matrix* [13]." The user is able to choose his own sparsity level. On average only around 3% of all residue pairs are observed to be in direct contact [13].

$$\sum_{ij=1}^{d} S_{ij}\Theta_{ij} - logdet\Theta + \rho \sum_{ij=1}^{d} |\Theta_{ij}| \tag{4.6}$$

In the glasso function (4.6), the ρ is the parameter chosen by the user. If the user chooses a higher ρ, the number of zero elements increases until there are only zero elements [13]. The final solution is a good estimation of the true inverse covariance matrix. For more information see the references [1, 7, 13, 18, 24].

[1] `http://www-stat.stanford.edu/~tibs/glasso/`

4.4 Final prediction

The final processing of the prediction is a bit complicated, also because of the size of the inverse covariance matrix. For the following example, I pulled apart the original inverse covariance matrix into sub-matrices. Keep in mind that the numbers are unrealistic, extremely small matrices and I left sub-matrices out (eg. AD). We want to make a prediction for contacts between two specific residues. So, a score $S^{contact}$ for alignment columns i and j. Assume we have the following sub-matrices.

$$\Theta^{AC} = \begin{pmatrix} -0.4 & -0.8 & 0.2 \\ -0.8 & -0.6 & 0.4 \\ 0.2 & 0.4 & 0.4 \end{pmatrix}, \text{ and } \Theta^{DC} = \begin{pmatrix} -0.4 & 0.8 & -0.2 \\ 0.8 & -0.6 & 0.4 \\ -0.2 & 0.4 & 0.4 \end{pmatrix}$$

We do not want to know all the scores for any specific amino acid combination, but for every combination of alignment columns.

$$S_{ij}^{contact} = \sum_{ab} |\Theta_{ij}^{ab}| \tag{4.7}$$

If we apply formula 4.7 on the matrices above, we can fill in the formula, for $i = 1$ and $j = 3$ as follows:

$$S_{13}^{contact} = \sum_{ab} |\Theta_{13}^{ab}| = |\theta_{13}^{AC}| + |\theta_{13}^{DC}| = |0.2| + |-0.2| = 0.4$$

Values of the inverse covariance contain information about the degree of direct coupling. The scores of the different amino acids are combined on basis of alignment columns. Finally we have a correction of the final prediction versus all results.

$$PC_{ij} = S_{ij}^{contact} - \frac{\bar{S}_{(i-)}^{contact} \bar{S}_{(-j)}^{contact}}{\bar{S}^{contact}} \tag{4.8}$$

PC_{ij} represents the final prediction score (see formula 4.8). We subtract a background correction from the calculated $S_{ij}^{contact}$. Mean prediction norms of columns i and j versus all other columns divided by the mean precision norm of all columns. The output of PSICOV may consist of negative prediction values. Using parameter -p, it is possible to convert the scores into an estimated positive prediction value. PSICOV will use a logistic function to the observed distribution of scores [13].

Chapter 5

Experiments using PSICOV

Can we find interesting information from the contact prediction results of PSICOV, related to the difficulty to predict contacts for specific types of amino acids and amino acid pairs? If we do find unexpected or interesting results, we can try to adapt PSICOV and improve its predictions. One possibility is weighting of specific output values. But, at first we take a look into the contact predictions of PSICOV.

5.1 Data

5.1.1 Input data

We consider the data provided by the authors of the PSICOV software package. The data has been carefully selected, because protein contact prediction tools like PSICOV need highly varied alignments for accurate predictions. According to the guidelines of Jones et al., Pfam families must contain ≥1000 sequences with a highly resolved X-ray crystallographic structure available, resulting 150 target proteins. Sequences of the targets must be between 50 and 275 residues, duplicate rows in alignments are deleted and so are the columns containing gaps in the target sequence [13].

In order to increase the scope of our analysis, we could also use the HSSP database as input data [14]. If we do so, we must extract the alignments from the HSSP files and stick to the mentioned guidelines.

5.1.2 PSICOV output and parameters

First we tried to replicate the results in the PSICOV paper [13]. The program offers a set of parameters in order to manage the computation time or the score of the results. How can we reproduce the predictions?

```
./psicov -d 0.03 demo.aln > output
```

The parameter setting of -d 0.03 represents the target density 3% ("i.e. 3% non-zero terms in the final precision matrix"[13]). These settings leads to iteratively adjusting ρ in the glasso function (equation 4.6). We are able to achieve similar results, but a 100% match can not be found because the software of PSICOV and the glasso library have changed over time. David T. Jones wrote to us the following.

"The results supplementary material was for versions of PSICOV and the glasso library code that existed when the paper was published (i.e. back in mid-2011). Since then the glasso code has been improved by its authors (Robert Tibshirani's group). PSICOV itself probably hasn't changed so much - apart from fixing a few bugs here and there - but some of these changes may have slightly changed the output. Assuming the differences you see are minor, I would just go with whatever the current version of the code calculates."

Since PSICOV did not change significantly, in our analysis we use the results from the supplementary material of [13].

5.1.3 Top-L/l and sequence separation

The output of PSICOV is a list of predicted contacts and a score for each contact. In contact prediction it is common to look at the top-L/l scores, where L is the length of the target sequence [5, 13]. If the sequence is 150 residues long, then we have 150 predicted contacts. The top-$L/2$, -$L/5$ and -$L/10$ are also considered, which have been shown to yield more accurate predictions [13].

Another criteria to filter the results is based on sequence separation between two residues. By how many residues are the contacts separated in the sequence? In the experiments we look at the same cutoffs as used in the PSICOV paper. We assume that for each contact (with residue pair (i,j)), the absolute distance between i and j in the sequence must be greater than 4, 8, 11 and 23 residues. These cutoffs are not chosen randomly, but they represent characteristics of protein structures. The distance between contact pairs in an alpha helix, separated by one rotation is about four residues [3]. Eight represents two rotations. In other studies sequence separations of >12 and >23 are used, mainly because long distance contacts are more valuable as structure constraints in 3D structure predictions [5].

5.2 Experimental setup

5.2.1 Defining a contact pair

What is the definition of a contact pair in a fully folded protein? The following definition is used in the PSICOV paper and CASP8 experiment: a pair of residues whose C-beta to C-beta distance is <8Å [5, 13]. Å stands

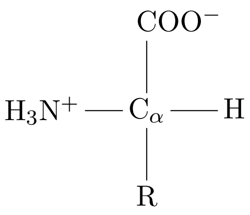


Figure 5.1: Single amino acid molecule where R stands for the R-group or side chain [19].

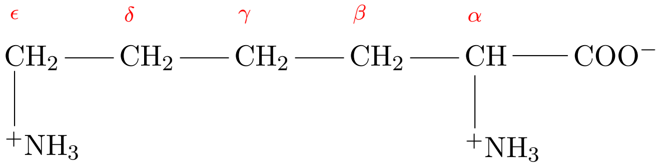


Figure 5.2: The amino acid lysine. Additional carbons in an R-group are named (started from the alpha carbon in the middle) C-beta, C-gamma, C-delta and so on. Figure and theory from [19].

for the Angstrom unit. One Angstrom is equal to 0.1nm (or 10^{-10}m) and called after physicist Anders J. Angstrom [19].[1] Now we will briefly explain what C-beta, and C-beta to C-beta distances are.

Figure 5.1 shows a single amino acid. The R-group is different for any of the twenty amino acids and it is attached to the alpha carbon in the middle. Figure 5.2 shows the amino acid lysine and how additional alpha carbons are called. All amino acids, except glycine, have an C-beta atom in their R-group. Glycine only contains a hydrogen atom in its side chain. We look into the distances between the C-beta atoms of the considered amino acids. In case of glycine we use the C-alpha atoms. For more information see chapter 3.1 of Lehningner, Principles of Biochemistry [19].

In literature you can find more definitions of residue-residue contacts [23]. For instance another method used by the authors of PSICOV considers minimum distances between any two heavy atoms as definition of residue-residue contacts. Results of PSICOV are slightly better for this more sensitive definition. Nevertheless, the first experiments of this research have been done with the C-beta to C-beta contact definition. The C-beta definition is still a common method and differences are expected to be minor.

[1] http://en.wikipedia.org/wiki/Angstrom

5.2.2 Golden truth: experimentally determined contact pairs

In order to assess the results of PSICOV and contact prediction tools in general, we need the real three-dimensional structure of proteins. We use the crystallographic structures which are experimentally determined protein 3D structures. X-ray structures can be determined at different levels of resolution. "*At low resolution only the shape of the molecule is obtained, whereas at high resolution most atomic positions can be determined to a high degree of accuracy. The quality of the final three-dimensional structure depends on the resolution of the X-ray data and the degree of refinement* [3]." Brandon and Tooze state that a highly refined structure at a resolution around 2.0Å, the amino acid sequence is known. Since Jones et al. use a resolution threshold of ≤1.9Å, we assume that we have a reliable benchmark.

Atomic coordinates

Information about a protein 3D structure can be found in so-called pdb (format) files. "*A typical pdb format file contains atomic coordinates for a diverse collection of proteins, small molecules, ions and water. Each atom is entered as a line of information that starts with a keyword: either ATOM or HETATM* [2, 21]." These atomic coordinates are used to calculate the absolute distance between any two atoms. Therefore pdb files can be directly used to calculate the 'true' residue-residue contacts, for various types of contact pair definitions.

Keyword	Nr.in file	Name	AA	Chain-id	Res nr	X	Y	Z	
ATOM	1	N	LEU	A	4	-3.883	41.780	40.071	N
ATOM	2	CA	LEU	A	4	-4.394	42.817	39.132	C
ATOM	3	C	LEU	A	4	-5.413	42.211	38.178	C
ATOM	4	O	LEU	A	4	-5.799	42.876	37.208	O
ATOM	5	CB	LEU	A	4	-3.265	43.498	38.355	C
ATOM	6	CG	LEU	A	4	-3.604	44.793	37.605	C
ATOM	7	CD1	LEU	A	4	-4.290	45.783	38.535	C
ATOM	8	CD2	LEU	A	4	-2.372	45.423	36.968	C
ATOM	9	N	PHE	A	5	-5.860	40.976	38.465	N
ATOM	10	CA	PHE	A	5	-6.835	40.387	37.540	C

Table 5.1: A snippet of the atomic contacts and coordinates in a pdb file (pdb-id 1a3a). The abbreviation AA represent the three letter code of amino acids. Chain-id specifies the chain if the protein is oligomeric (a molecular complex).

Table 5.1 shows part of the description of atomic coordinates. Each line consists of an atom and the information related to its residue (number) and coordinates.

5.2.3 Comparison of predicted and experimentally determined contacts

Interatomic contacts vs. PSICOV output

The WHAT IF servers of the CMBI provide many different functions around protein data calculations [12]. One is the calculation of C-beta pairs <10Å out of atomic contact information of pdb files. C-beta pairs within a smaller threshold can be filtered afterwards.

seq nr.	AA	pdb nr.		seq nr.	AA	pdb nr.	Å
1	LEU	(4)A	-	2	PHE	(5)A	5.358
1	LEU	(4)A	-	3	LYS	(6)A	8.838
1	LEU	(4)A	-	73	LYS	(76)A	9.874
1	LEU	(4)A	-	74	THR	(77)A	5.247
1	LEU	(4)A	-	75	GLY	(78)A	7.254
1	LEU	(4)A	-	76	VAL	(79)A	9.641
1	LEU	(4)A	-	103	ALA	(106)A	8.314
1	LEU	(4)A	-	107	GLU	(110)A	7.507
1	LEU	(4)A	-	110	GLN	(113)A	6.981
1	LEU	(4)A	-	111	VAL	(114)A	6.741
1	LEU	(4)A	-	114	SER	(117)A	7.432
1	LEU	(4)A	-	144	ARG	(147)A	9.819

Table 5.2: All results of interatomic contacts <10Å for the first amino acid leucine in pdb file 1a3a. Each row represents a contact pair with associated distance in Angstrom.

Pos i	Pos j			Score
96	129	0	8	17.366371
25	100	0	8	12.980494
43	100	0	8	12.909767
12	80	0	8	11.863224
23	40	0	8	11.837438
46	52	0	8	10.869245
26	40	0	8	10.746915
25	79	0	8	10.317855
75	102	0	8	10.081199
84	96	0	8	9.303587

Table 5.3: Top ten results of a PSICOV prediction. Each row stands for a predicted contact pair between positions i and j. All results are sorted by score. Column three and four contain dummy values.

As can be seen in table 5.2, the sequence number does not correspond with the pdb number. The supplementary data of PSICOV includes fixed pdb files, where sequence number and pdb number are matched correctly. As illustrated in table 5.2 and 5.3 one can relate the predictions of PSICOV to the golden truth, and compute the true positive and false positive predictions of amino acids (or pairs) used in our analysis.

5.2.4 True positives and false positives

The computation of the set of true positives and false positives depends on two parameters, here called selected number of ranked predictions (denoted by l) and distance between the residues of a pair in the protein sequence (denoted by d). We used the parameter settings considered in [13], with $l \in \{L, L/2, L/5, L/10\}$ and $d \in \{4, 8, 11, 23\}$. So, for a given target protein p, the list O consisting of the contact pairs output by PSICOV and sorted according to their score, is processed as follows to compute true and false positives.

Algorithm 1 Define true and false positives

Input: PSICOV predictions and interatomic contacts (golden truth)

Output: Determined true and false positive PSICOV predictions

```
true positive = [ ];
false positive = [ ];
set the values of parameters l and d;
k = 1;
while k < l do
   (i,j) = k-th pair in 0;
   if sequence distance between i and j > d then
      if prediction$_{(i,j)}$ in interatomic contacts then
         true positive = [ true positive ; (i,j) ];
      else
         false positive = [ false positive ; (i,j) ];
      end if
   end if
   k = k + 1;
end while
```

With interatomic contacts is meant the list of interatomic contacts calculated by the WHAT IF webservers. For the considered 150 target proteins, and for each pair of values of the parameters l and d we calculate the true and false positives, for a total of 4800 experiments.

Chapter 6

Analysis of PSICOV predictions

We want to analyze the results of PSICOV in order to determine whether sequence properties of an amino acid and/or of an amino acid pair can be linked to the difficulty of their prediction. The findings could be used to improve the prediction of PSICOV by incorporating the discovered properties as bias in the method (predictions).

We investigated correlations between vectors using Pearson and Spearman correlations. A Spearman correlation coefficient is a rank ordered based test. Two datasets are ranked by value and both rankings are tested. We also use a Pearson correlation coefficient to test the linear relationship between two datasets. For both tests, a Python library called Scipy provided the functionality.[1]

6.1 Analysis of true positives

We first investigated true positives. For each amino acid, we correlate the distribution of the number of true positives and of the true contact pairs containing that amino acid, over the 150 target proteins. Amino acids that do not yield a positive correlation are considered as outliers and could have peculiar biological characteristics. Specifically, for each target protein p, sequence separation $d \in \{4, 8, 11, 23\}$, and selected number of predictions $l \in \{L, L/2, L/5, L/10\}$, we construct a 20x20 matrix $M^p_{d,l}$ such that $M^p_{d,l}(i,j)$ contains the number of true positive pairs which are equal to the pair of amino acids (i,j) (or (j,i)).

Consider the vector $V^j_{d,l}$ such that $V^j_{d,l}(k)$ is $\sum\limits_{i=1}^{20} M^k_{d,l}(i,j)$, that is, the number of amino acid pairs occurring in the set of true positives and containing amino acid j for the k-th target protein PSICOV experiment (see 6.1).

[1] http://www.scipy.org/

$$V_{d,l}^{j} = (\sum_{i=1}^{20} M_{d,l}^{1}(i,j), \sum_{i=1}^{20} M_{d,l}^{2}(i,j), \sum_{i=1}^{20} M_{d,l}^{3}(i,j), ..., \sum_{i=1}^{20} M_{d,l}^{150}(i,j)) \quad (6.1)$$

Analogously we consider the vector C_d^j (see equation 6.2), where I denotes the matrix containing the number of true interatomic contacts for each pair of amino acids. $C_d^j(k)$ is the number of amino acid pairs occurring in the set of true interatomic contacts of the k-th target protein and containing amino acid j. We compute a Pearson and Spearman correlation between $V_{d,l}^j$ and C_d^j.

$$C_{d}^{j} = (\sum_{i=1}^{20} I_{d}^{1}(i,j), \sum_{i=1}^{20} I_{d}^{2}(i,j), \sum_{i=1}^{20} I_{d}^{3}(i,j), ..., \sum_{i=1}^{20} I_{d}^{150}(i,j)) \quad (6.2)$$

Example of true positive analysis

Consider the following matrices, for sequence separation 4 and number of predictions L/1, we have fictional true positive predictions (M) and a fictional golden truth (I). In all examples we pretend that there are only three types of amino acids ($\in \{A, C, D\}$). Sorted by name, $1a3a$ is the first pdb-id and $5ptp$ the hundred fiftieth.

$$M_{4,L/1}^{1a3a} = \begin{array}{cccc} & A & C & D \\ A & 3 & 0 & 2 \\ C & 0 & 1 & 2 \\ D & 2 & 2 & 1 \end{array}, ..., M_{4,L/1}^{5ptp} = \begin{array}{cccc} & A & C & D \\ A & 1 & 1 & 2 \\ C & 1 & 3 & 1 \\ D & 2 & 1 & 1 \end{array}$$

$$I_{4}^{1a3a} = \begin{array}{cccc} & A & C & D \\ A & 3 & 2 & 2 \\ C & 2 & 4 & 2 \\ D & 2 & 2 & 3 \end{array}, ..., I_{4}^{5ptp} = \begin{array}{cccc} & A & C & D \\ A & 1 & 3 & 2 \\ C & 3 & 4 & 1 \\ D & 2 & 1 & 1 \end{array}$$

In this example, each amino acid can form three types of pairs instead of twenty pairs in the original experiment. Therefore, the sum consists of three elements.

$$\begin{aligned} V_{4,L/1}^{A} &= (\sum_{i=1}^{3} M_{4,L/1}^{1a3a}(i,A), ..., \sum_{i=1}^{3} M_{4,L/1}^{5ptp}(i,A)) \\ &= ((M_{4,L/1}^{1a3a}(A,A)) + M_{4,L/1}^{1a3a}(C,A)) + M_{4,L/1}^{1a3a}(D,A)), \\ & \quad ..., \\ & \quad (M_{4,L/1}^{5ptp}(A,A)) + M_{4,L/1}^{5ptp}(C,A)) + M_{4,L/1}^{5ptp}(D,A)) \\ &= ((3+0+2), ..., (1+1+2)) = (5, ..., 4) \end{aligned}$$

$$C_4^A = (\sum_{i=1}^{3} I_4^{1a3a}(i,A), ..., \sum_{i=1}^{3} I_4^{5ptp}(i,A)) = ((3+2+2), ..., (1+3+2)) = (7, ..., 6)$$

Finally, we compute a Pearson and Spearman correlation between vectors $V_{4,L/1}^A$ and C_4^A. For amino acids C and D we can do the same.

6.2 Analysis of false positive analysis

Besides true positives, we can also try to find information in the results of the false positive predictions. Our hypothesis is that the difficulty of predicting a contact containing a given amino acid is related to how often it occurs in the target protein. When amino acids occur fewer times, they are more difficult to predict.

We test this hypothesis by means of the following two methods.

1. Amino acid false positive vs frequency analysis. For each amino acid, we correlate the distribution of its false positives and its frequency over the target sequences. We would expect this analysis to yield positive correlations.

2. Amino acid prediction difficulty vs frequency analysis. We quantify the difficulty to predict a contact pair containing a given amino acid. Specifically, for a given target protein, we subtract the number of true positives from the number of false positives of pairs containing that amino acid. Then we correlate the vector obtained by considering all target proteins, with the amino acid frequency vector. We would expect this analysis to yield negative correlations.

In the sequel we describe these methods in detail.

6.2.1 Amino acid false positive vs frequency analysis

For each target protein sequence p, sequence separation $d \in \{4, 8, 11, 23\}$, and selected number of predictions $l \in \{L, L/2, L/5, L/10\}$, construct a 20x20 matrix $M'^{k}_{d,l}$ such that $M'^{k}_{d,l}(i,j)$ contains the number of false positive pairs which are equal to the pair of amino acids (i,j) (or (j,i)).

Consider the vector $V'^{j}_{d,l}$ such that $V'^{j}_{d,l}(k)$ is $\sum_{i=1}^{20} M'^{k}_{d,l}(i,j)$, that is, the number of amino acid pairs occurring in the set of false positives and containing amino acid j for the k-th target protein PSICOV experiment. We want to correlate this vector with the vector F^j of the amino acid frequencies in the 150 target protein sequences, such that $F^j(p)$ is the number of times amino acid j occurs in protein sequence p divided by the length of p. We compute Pearson and Spearman correlation between $V'^{j}_{d,l}$ and F^j.

Example of the frequency vector

Vector $V'^{j}_{d,l}$ is created in the same way as shown in the example of the true positive analysis, but for $V'^{j}_{d,l}$ we use matrices with false positive predictions.

```
Fictional targets:
1a3a: CCDAAADDACADDCA, ...,5ptp: ADCADDDACADDACCD
```
F^A = (F^A(1a3a), ...,F^A(5ptp)) = (6/15, ..., 5/16)

We compute Pearson and Spearman correlation between $V'^{A}_{4,L/1}$ and F^A (F^A as illustrated above, all target sequence frequencies for amino acid A).

6.2.2 Amino acid prediction difficulty vs frequency analysis

$$D^j_{d,l}(k) = \sum_{i=1}^{20}(M'^k_{d,l}(i,j) - M^k_{d,l}(i,j)) \tag{6.3}$$

Consider the vector $D^j_{d,l}$ (see equation 6.3), where $D^j_{d,l}(k)$ is the relative difference of amino acid occurrence j in the false positive and true positive predictions of the k-th target protein. We compute Pearson and Spearman correlation between $D^j_{d,l}$ and frequency vector F^j (for F^j see section 6.2.1).

Example of amino acid prediction difficulty vs frequency analysis

In amino acid prediction difficulty we look at both true positive and false positive matrices. We calculate the vector element for pdb-id $1a3a$ ($k = 1a3a$). The final vector $D^j_{d,l}$ consist of 150 elements or targets (sorted alphabetically).

$$M^{1a3a}_{4,L/1} = \begin{array}{cccc} & A & C & D \\ A & 3 & 0 & 2 \\ C & 0 & 1 & 2 \\ D & 2 & 2 & 1 \end{array}$$

$$M'^{1a3a}_{4,L/1} = \begin{array}{cccc} & A & C & D \\ A & 1 & 1 & 0 \\ C & 1 & 3 & 1 \\ D & 0 & 1 & 1 \end{array}$$

$$\begin{aligned} D^A_{4,L/1}(1a3a) &= \sum_{i=1}^{3}(M'^{1a3a}_{4,L/1}(i,A) - M^{1a3a}_{4,L/1}(i,A))) \\ &= (\sum_{i=1}^{3} M'^{1a3a}_{4,L/1}(i,A)) - (\sum_{i=1}^{3} M^{1a3a}_{4,L/1}(i,A))) \\ &= (1+1+0) - (3+0+2) \\ &= 2 - 5 = -3 \end{aligned}$$

We can simplify the formula so that parts are similar to the example of the true positive analysis. It is a complex way to describe that we subtract the sum of one column of false positives from the sum of one column of true positives (for amino acid A). The final score can be either negative or positive. When it is positive, it means that there are more false positive predictions than true positive predictions (for amino acid A). These amino acids are more difficult to predict. By calculating a Pearson and Spearman correlation between $D^A_{4,L/1}$ and F^A (example of 6.2.1) we can check if the amino acid prediction difficulty (of A) is related to the frequency in the golden truth.

6.3 Analysis of pairs

Another method of analysis is looking at specific amino acid pairs, instead of single amino acids in PSICOV predictions. The major difference in analysis is that we are not creating a vector for a single amino acid, but a vector for any possible pair of amino acids. We have similar expectations for results of predicted pairs as for single amino acids.

6.3.1 Analysis of predicted true positive pairs

For each target protein sequence p, sequence separation $d \in \{4, 8, 11, 23\}$, and selected number of predictions $l \in \{L, L/2, L/5, L/10\}$, construct a 20x20 matrix $M^p_{d,l}$ such that $M^p_{d,l}(i, j)$ contains the number of true positive pairs which are equal to the pair of amino acids (i, j) (or (j, i)).

$$V^{i,j}_{d,l} = (M^1_{d,l}(i, j), M^2_{d,l}(i, j), ..., M^{150}_{d,l}(i, j)) \tag{6.4}$$

Consider a vector $V^{i,j}_{d,l}$ (see 6.4), such that $V^{i,j}_{d,l}(k)$ is the number of true positive predictions of amino acid pair i, j for the k-th target protein PSICOV experiment. Analogously we create a vector $C^{i,j}_d$, where $C^{i,j}_d(k)$ is $I^k_d(i, j)$, that is the true interatomic frequency of amino acid pair (i, j) for the k-th target protein. We compute a Pearson and Spearman correlation between $V^{i,j}_{d,l}$ and $C^{i,j}_d$.

Example of true positive pair analysis

We consider the same fictional matrices as shown in the example of the true positive analysis (true positive predictions M, golden truth I). Again we have only two of the 150 targets.

$$M^{1a3a}_{4,L/1} = \begin{array}{cccc} & A & C & D \\ A & 3 & 0 & 2 \\ C & 0 & 1 & 2 \\ D & 2 & 2 & 1 \end{array}, ..., M^{5ptp}_{4,L/1} = \begin{array}{cccc} & A & C & D \\ A & 1 & 1 & 2 \\ C & 1 & 3 & 1 \\ D & 2 & 1 & 1 \end{array}$$

$$I^{1a3a}_4 = \begin{array}{cccc} & A & C & D \\ A & 3 & 2 & 2 \\ C & 2 & 4 & 2 \\ D & 2 & 2 & 3 \end{array}, ..., I^{5ptp}_4 = \begin{array}{cccc} & A & C & D \\ A & 1 & 3 & 2 \\ C & 3 & 4 & 1 \\ D & 2 & 1 & 1 \end{array}$$

This time it is just a matter of matching the right predictions with the right vector elements. It seems easier, but the amount of vectors (one for every amino acid pair) may cause a bottleneck.

$$\begin{aligned} V^{A,C}_{4,L/1} &= (M^{1a3a}_{4,L/1}(A, C), ..., M^{5ptp}_{4,L/1}(A, C)) \\ &= (0, ..., 1) \\ C^{A,C}_4 &= (I^{1a3a}_4(A, C), ..., I^{5ptp}_4(A, C)) \\ &= (2, ..., 3) \end{aligned}$$

A Pearson and Spearman correlation can be calculated between $V^{A,C}_{4,L/1}$ and $C^{A,C}_4$.

6.3.2 Analysis of predicted false positive pairs

For the analysis of predicted false positive pairs we make again a distinction between the false positive frequencies and prediction difficulties.

Amino acid pairs, false positive vs frequency analysis

We construct again a 20x20 matrix $M'^{k}_{d,l}$ (for each target protein p, sequence separation d, and selected number of predictions l), such that $M'^{k}_{d,l}(i,j)$ contains the number of false positive pairs which are equal to the pair of amino acids (i,j) (or (j,i)). Consider a vector $V'^{i,j}_{d,l}$, such that $V'^{i,j}_{d,l}(k)$ is equal to $M'^{k}_{d,l}(i,j)$, that is the occurrence of false positive predicted amino acid pair (i,j) of the k-th target protein.

We want to correlate vector $V'^{i,j}_{d,l}$ with frequency vector $F^{i,j}$, where $F^{i,j}(p)$ is the product of amino acid frequencies i and j, divided by the square of the lenght of p. Thereto, we compute a Pearson and Spearman correlation between $V'^{i,j}_{d,l}$ and $F^{i,j}$.

Example of the frequency vector for pairs

Vector $V'^{i,j}_{d,l}$ is similar to $V^{i,j}_{d,l}$ (see section 6.3.1 for the example). For $V'^{i,j}_{d,l}$ we use matrices of false positive predictions.

```
Fictional targets:
1a3a: CCDAAADDACADDCA, ...,5ptp: ADCADDDACADDACCD
```
$\texttt{F}^{\texttt{A,C}}$ = ($\texttt{F}^{\texttt{A,C}}$(1a3a), ...,$\texttt{F}^{\texttt{A,C}}$(5ptp)) = ($\frac{6\cdot 6}{15^2}$, ..., $\frac{5\cdot 4}{16^2}$)

We compute Pearson and Spearman correlation between $V'^{A,C}_{4,L/1}$ and $F^{A,C}$ ($F^{A,C}$ as illustrated above, all target sequence frequencies for amino acid pair A,C).

Amino acid pairs, prediction difficulty vs frequency analysis

$$D'^{i,j}_{d,l}(k) = (M'^{k}_{d,l}(i,j) - M^{k}_{d,l}(i,j)) \qquad (6.5)$$

Vector $D'^{i,j}_{d,l}$ (see 6.5) is similar to equation 6.3. $D'^{i,j}_{d,l}(k)$ is the relative difference of amino acid pair occurrence (i,j) in the false positive and true positive predictions of the k-th target protein. We compute a Pearson and Spearman correlation between $D'^{i,j}_{d,l}$ and frequency vector for pairs $F^{i,j}$.

Example of amino acid prediction difficulty vs frequency analysis for pairs

We consider the same fictional matrices as in the example for the non pair predictions difficulty vs frequency analysis. We have a matrix of false positive predictions ($M'^{1a3a}_{4,L/1}$) and one of true positive predictions ($M^{1a3a}_{4,L/1}$).

$$M^{1a3a}_{4,L/1} = \begin{array}{cccc} & A & C & D \\ A & 3 & 0 & 2 \\ C & 0 & 1 & 2 \\ D & 2 & 2 & 1 \end{array}$$

$$M'^{1a3a}_{4,L/1} = \begin{array}{cccc} & A & C & D \\ A & 1 & 1 & 0 \\ C & 1 & 3 & 1 \\ D & 0 & 1 & 1 \end{array}$$

For one element ($k = 1a3a$) of vector $D'^{i,j}_{d,l}$ we can fill in:

$$\begin{aligned} D'^{A,C}_{4,L/1}(1a3a) &= (M'^{1a3a}_{4,L/1}(A,C) - M^{1a3a}_{4,L/1}(A,C)) \\ &= 1 - 0 = 1 \end{aligned}$$

If we have a complete vector (150 elements or targets) for $D^{A,C}_{4,L/1}$, we can correlate this vector with the frequency vector for pairs $F^{A,C}$.

Chapter 7

Using the results of the analysis to adjust the output of PSICOV

After analyzing the true and false positive PSICOV predictions, we want to use our findings for improving the predictions. We will briefly discuss the results of the analysis of PSICOV predictions. For the complete set of results, I refer to the supplementary data of this thesis.[1]

7.1 Predictions versus frequency correlation results

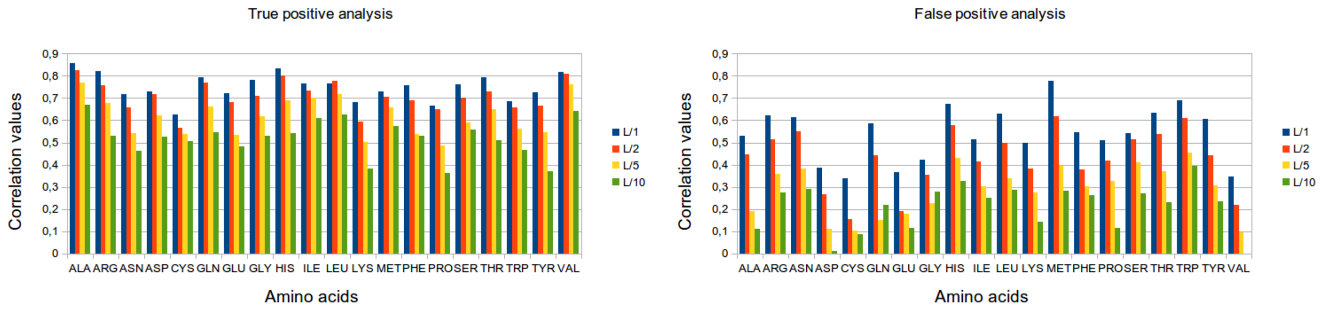


Figure 7.1: The (Pearson) correlation values out the prediction analysis. True positives versus golden truth and false positives versus frequency, as described in sections 6.1 and 6.2.1 (by sequence separation $D > 4$).

[1] https://www.dropbox.com/s/isv4wrzyyzwsw0w/analysis%26adjustments.zip

In general we observe only positive correlations between the predicted pairs of amino acids and the frequency in the golden truth or target sequence. We expected such positive correlations, but we were hoping for the presence of few outliers. A negative correlation of predicted pairs versus golden truth or target sequence frequency would have indicated amino acids that are difficult to predict. Figure 7.1 shows these correlation values. These results are somewhat expected, and do not provide relevant knowledge that could be used to improve PSICOV predictions.

7.2 Prediction difficulty versus frequency correlation results

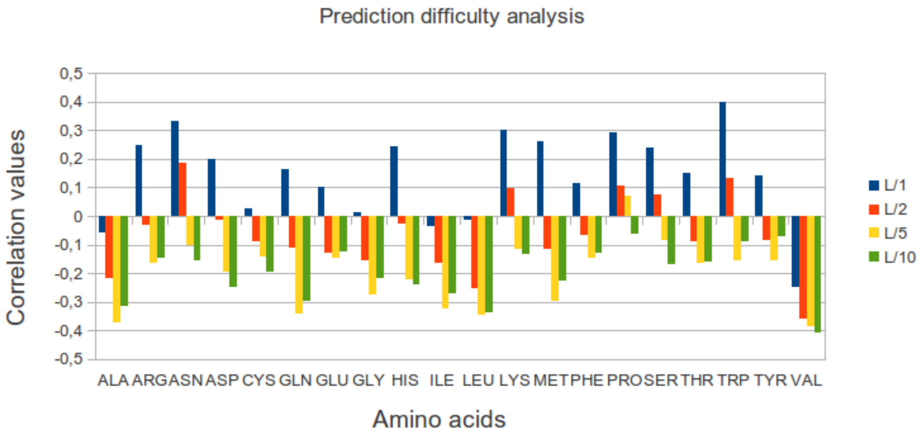


Figure 7.2: The (Pearson) correlation values between the prediction difficulty of amino acids and their target sequence frequency (by sequence separation $D > 4$).

Figure 7.2 shows how amino acid prediction difficulty correlates with the amino acid frequency in target sequences. We observe a higher number of negative correlations for larger values of l in the top-L/l lists. Results indicate that our conjecture of a negative correlation between prediction difficulty and amino acid frequencies is substantiated by the considered experiments. Interestingly, amino acid prediction difficulty versus frequency correlations strongly depend on the amino acid composition. Therefore it seems that we could use this information to adjust the output of PISCOV.

7.2.1 A closer look at prediction difficulty

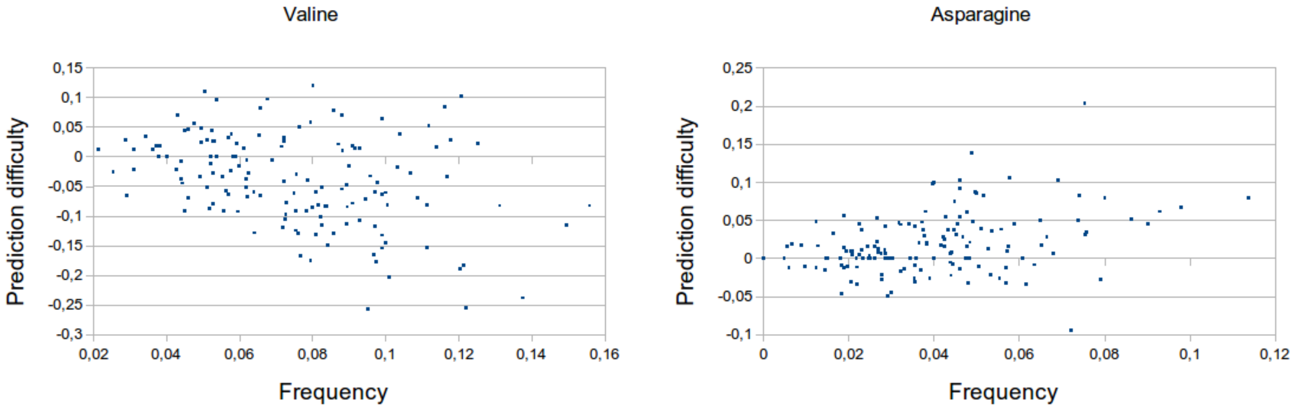


Figure 7.3: For amino acid valine and asparagine, top-L predictions and sequence separation $D > 4$, prediction difficulty values ((#false positives-#true positives)/top-L) are plotted versus frequencies are plotted for all 150 target sequences.

If we take a closer look at the analysis on prediction difficulty, we can make interesting observations. Prediction difficulty of an amino acid versus the frequency of the same amino acid in its target yield positive or negative correlations for different amino acids (figure 7.3). This type of variety can be used to distinguish amino acids that are easier or harder to predict. By means of adjustments to the output of PSICOV, we could 'help' amino acids that are harder to predict by PSICOV.

7.3 Methods to adjust PSICOV predictions

We try to apply our results to adjust the predictions of PSICOV using the following simple linear transformation. For each pair (i, j) the adjusted output $APSICOV(i, j)$ is

$$APSICOV(i, j) = PSICOV(i, j) - \gamma T(i, j) \tag{7.1}$$

The term $T(i, j)$ quantify how easy pairs containing one of the amino acids at position i and j are to be predicted by PSICOV, and the γ determines the impact of term $T(i, j)$.

We consider four definitions of T, yielding four variants of APSICOV.

1. Average Frequency. $-T(i, j) =$ the average frequency of the amino acids at positions i, j.
2. Average Correlation. $-T(i, j) =$ the average (Pearson) correlation of true positives versus golden truth of amino acids occurring at positions i, j. See section 6.1 for the explanation about retrieving the correlation values.
3. Average Prediction Difficulty. $T(i, j) =$ the average prediction difficulty (Pearson correlation of prediction difficulty versus frequency) of the amino acids at positions i, j. See section 6.2.2 for the explanation about retrieving the correlations values.
4. Pairing Preference. $-T(i, j) = R(i, j)$, where $R(i, j)$ is the pairing preference of the amino acids at positions i, j, as described in [8].

The pairing preference values are fitted into positive values by a logistic function for a suitable scale.

7.3.1 Adjustments in practice: calculation of APSICOV

Recalculated PSICOV predictions

The analysis on PSICOV predictions is based on the likelihood score of predicted contact pairs, which has a range of approximately -3 till $+19$. The proportions of correlation coefficients and the likelihood are not the same and can not be compared. We recalculated the predictions using the parameter $-p$ for an estimated positive predictive value by fitting a logistic function to the observed distribution of scores [13]. These results are scores with values between 0 and 1. Little has changed in the top-L/l lists, but we were not able to recalculate four targets (*1d4o,1kqr,1mug* and *1xkr*) due to lack of sequences in their MSA and/or lack of sequence diversity.

Training and testing the model

For a proper implementation of the adjustments in practice, we use a non-overlapping training and testing set. Here we use a leave-one-out cross validation (LOOCV) procedure. Specifically, all but one target protein are used to compute the values of T. The remaining target protein is used for prediction using APSICOV.

The following example based on prediction difficulty of amino acids, illustrate the LOOCV procedure. Consider the following vectors. For every target p and amino acid aa we can fill a vector X with frequencies F, and vector Y using prediction difficulties D. When creating a vector for target px, we use all available information except frequencies and prediction difficulties found in px.

$$X_{aa}^{px} = [F_{aa}^{p1}, ...F_{aa}^{p146}] - F_{aa}^{px}$$

$$Y_{aa}^{px} = [D_{aa}^{p1}, ...D_{aa}^{p146}] - D_{aa}^{px}$$

The following step is calculation of correlation coefficients $C_{aa}^{px}(X_{aa}^{px}, Y_{aa}^{px})$ for all combinations of targets and amino acids. Using this information we want to improve the predictions.

$$AP_{i,j}^{px} = P_{i,j}^{px} - \gamma \frac{C_{aa(i)}^{px} + C_{aa(j)}^{px}}{2} \tag{7.2}$$

As you can see in equation 7.2, we use the average of correlation values of the involved amino acids. $aa(i)$ and $aa(j)$ represents the amino acids type aa at position i or j.

Chapter 8

Results of APSICOV

In this chapter will be shown the results of the proposed methods to adjust PSICOV's predictions.

Quality assessment: mean precision

When comparing different methods of contact pair prediction we use mean precision as a measure for quality:

$$\frac{1}{146}\sum_{p=1}^{146}\frac{Tp_{L/l}^{p}}{Tp_{L/l}^{p}+Fp_{L/l}^{p}}=\frac{1}{146}\sum_{p=1}^{146}\frac{Tp_{L/l}^{p}}{L/l} \tag{8.1}$$

For every target p we divide the number of true positives by the sum of all true and false positives (equation 8.1). This is equal to the deviation of true positives by the top-L/l, which includes all predictions. By subtracting the mean precision of PSICOV to the mean precision of APSICOV (for a given T, and γ), the difference in performance called *gain(mean precision)* quantifies the gain achieved by APSICOV.

8.1 Adjustments by method

8.1.1 Average frequency

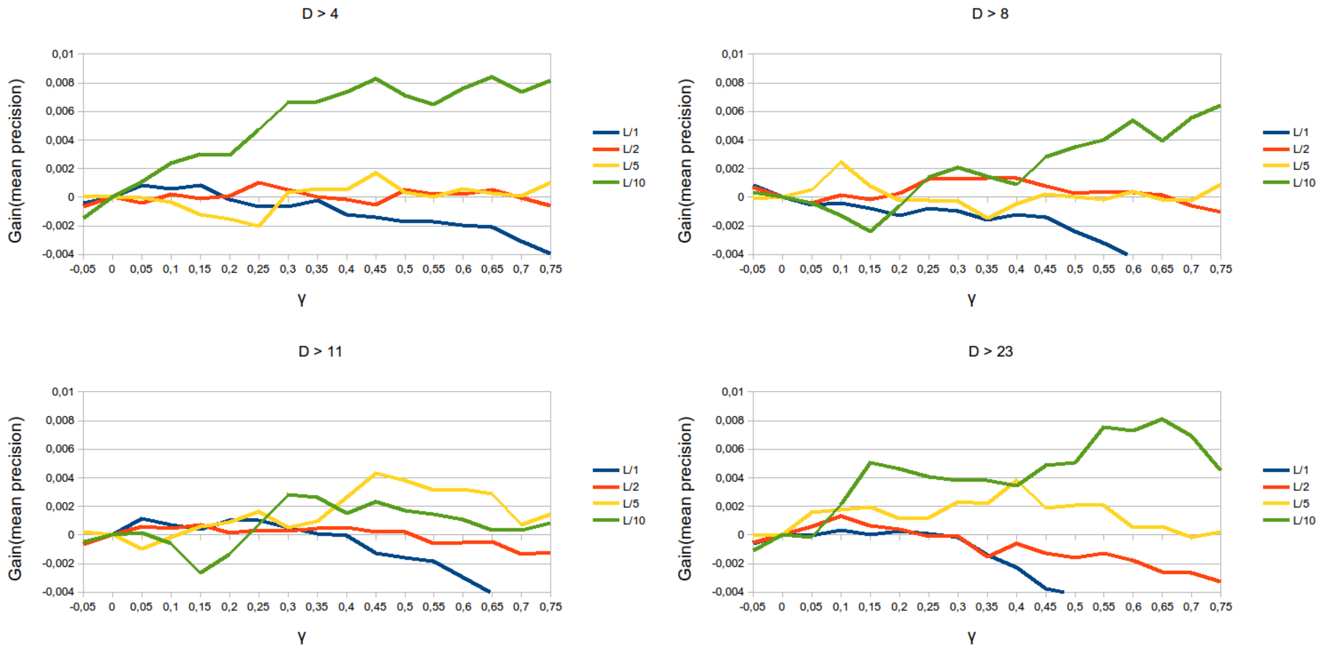


Figure 8.1: Gain of mean precision on PSICOV predictions using average frequencies of amino acids for different values of γ, by sequence separation D.

The first adjustment method results in very small gains, that is, the difference of the mean precision of APSICOV with the mean precision of PSICOV across various values of γ is often small. For L/10 we observe a higher improvement than for other top-L lists, but even in that case the gain of mean precision is rather small (< 0.01). This shows that APSICOV based on average frequency adjustment is not effective.

8.1.2 Average correlation (of TP versus golden truth)

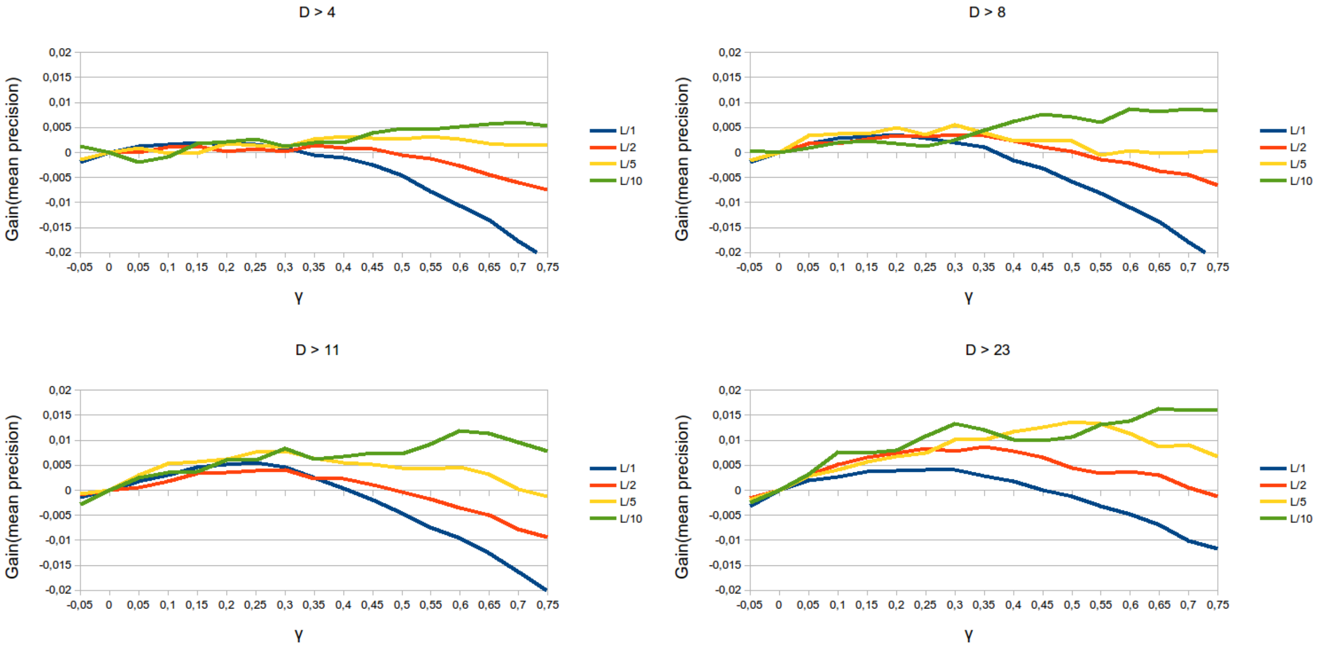


Figure 8.2: Gain of mean precision on PSICOV predictions using average correlation value of amino acids (true positives predictions versus frequency in the golden truth) for different values of γ, by sequence separation D.

If we look at the adjustments based on average correlation values, we see great differences in performance if we change the sequence separation threshold D. Especially for sequence separation $D > 23$ we see that this method leads to better mean precision values, but the gain is rather small.

8.1.3 Average prediction difficulty

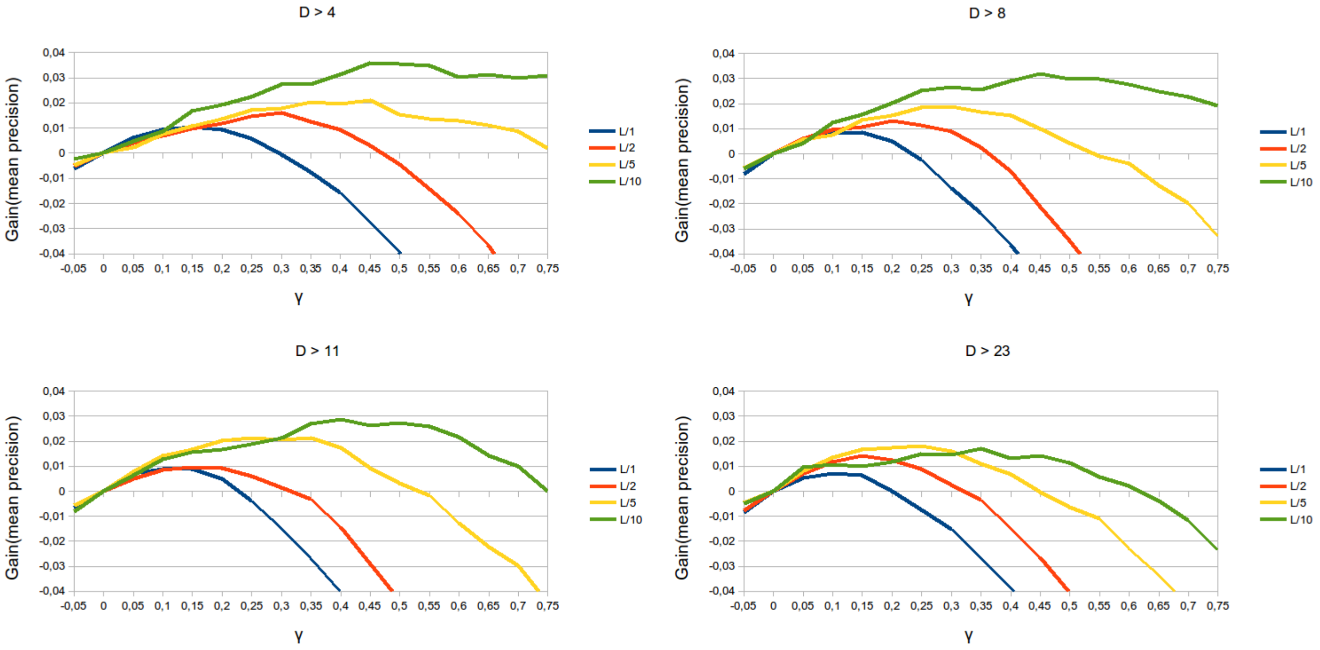


Figure 8.3: Gain of mean precision on PSICOV predictions using prediction difficulty of amino acids for different values of γ, by sequence separation D.

Prediction difficulty of amino acids clearly offers opportunities in improvement of PSICOV predictions. A stable growth in gain of mean precision shows the best impact of weighing the PSICOV predictions by prediction difficulty of amino acids.

8.1.4 Pairing preference

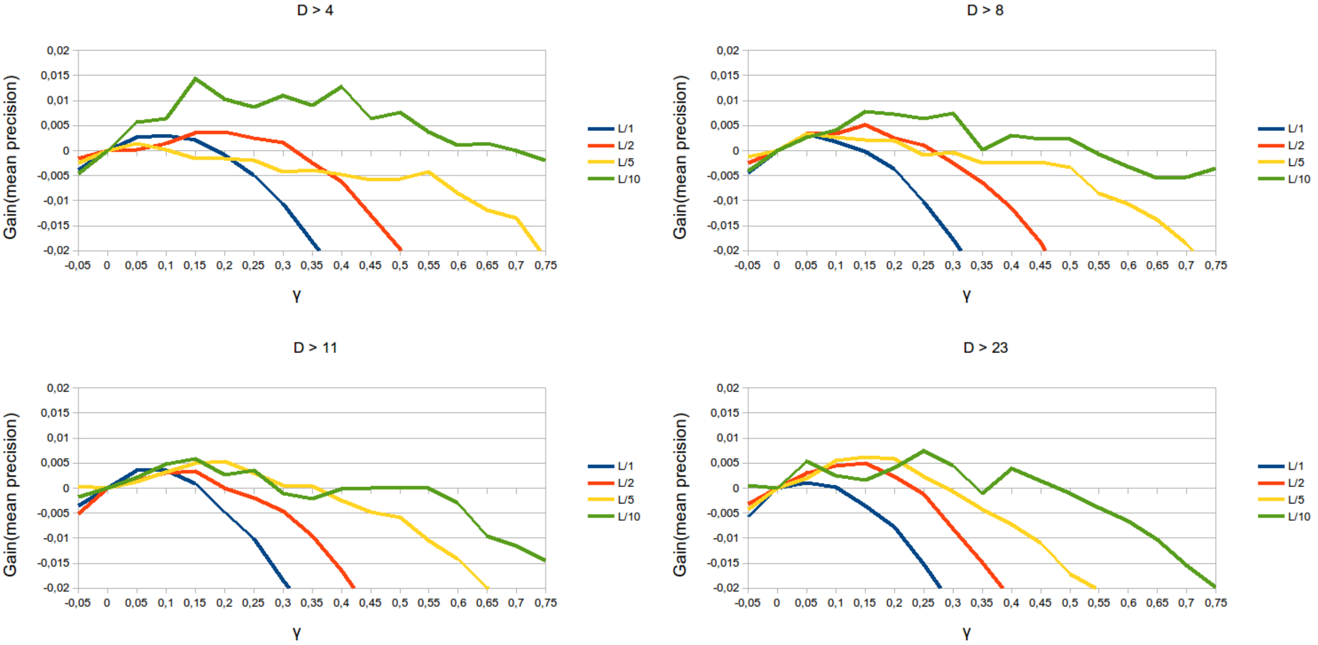


Figure 8.4: Gain of mean precision on PSICOV predictions using pairing preference of amino acids for different values of γ, by sequence separation D.

We found that adjusting PSICOV by pairing preference has some minimal improvements on the mean precision values. The gain of mean precision is maybe lower than expected, but the results do show that characteristics such as pairing preference can be helpful in predicting contact pairs.

8.2 A summary of all results

Table 8.1 shows an overview of the gain in average precision after corrections on PSICOV predictions. The value of γ is chosen by the best gain for sequence separation of 4 and top-L predictions.

	[i - j] > 4				[i - j] > 8			
	L	L/2	L/5	L/10	L	L/2	L/5	L/10
PSICOV	0.4606	0.6045	0.7406	0.7916	0.4280	0.5783	0.7279	0.7844
Avg Freq ($\gamma = 0.05$)	0.4614	0.6041	0.7406	0.7927	0.4274	0.5779	0.7285	0.7840
Avg Corr ($\gamma = 0.15$)	0.4626	0.6057	0.7405	0.7933	0.4312	0.5809	0.7317	0.7866
Pred Diff ($\gamma = 0.15$)	0.4709	0.6142	0.7513	0.8083	0.4364	0.5888	0.7415	0.8000
Pair Pref ($\gamma = 0.10$)	0.4636	0.6060	0.7409	0.7981	0.4298	0.5816	0.7307	0.7885

	[i - j] > 11				[i - j] > 23			
	L	L/2	L/5	L/10	L	L/2	L/5	L/10
PSICOV	0.4053	0.5571	0.7098	0.7777	0.3343	0.4748	0.6503	0.7367
Avg Freq ($\gamma = 0.05$)	0.4064	0.5577	0.7088	0.7779	0.3342	0.4754	0.6519	0.7366
Avg Corr ($\gamma = 0.15$)	0.4099	0.5604	0.7155	0.7815	0.3381	0.4814	0.6560	0.7441
Pred Diff ($\gamma = 0.15$)	0.4142	0.5668	0.7265	0.7934	0.3408	0.4890	0.6671	0.7467
Pair Pref ($\gamma = 0.10$)	0.4088	0.5602	0.7130	0.7825	0.3345	0.4793	0.6559	0.7391

Table 8.1: Mean precision scores compared. For PSICOV and the improved methods the mean precision values for the top-L/l contacts divided by sequence separation ranges where Cβ-Cβ distance < 8Å.

Looking at all the mean precision values, we observe that the correction based on prediction difficulty scores the best. Corrections by other adjustment methods improves the PSICOV predictions barely.

8.3 Wilcoxon signed-rank test

To test whether or not the improvements are significant, we use a Wilcoxon signed-rank test. We use a Python library to implement the statistical tests. As stated in the Scipy documentation: *"The Wilcoxon signed-rank test tests the null hypothesis that two related paired samples come from the same distribution and in particular, whether the distribution of the differences x - y is symmetric about zero."*[1]

Consider the following vectors:

$$I_{d,l} = [MpI_{d,l}^{p1}, ..., MpI_{d,l}^{p146}]$$

$$P_{d,l} = [MpP_{d,l}^{p1}, ..., MpP_{d,l}^{p146}]$$

Vector $I_{d,l}$ consists of adjusted precision values for each target with top-L/l predictions l and sequence separation d. Vector $P_{d,l}$ consists of original precision values of PSICOV predictions, using the same parameters. We calculate the Wilcoxon signed-rank test $W_{d,l}(I_{d,l}, P_{d,l})$ between these vectors.

	[i - j] > 4				[i - j] > 8			
	L	L/2	L/5	L/10	L	L/2	L/5	L/10
PSICOV vs. Avg Freq ($\gamma = 0.05$)	2.70^{-2}	**4.74^{-1}**	**3.97^{-1}**	**5.75^{-1}**	**1.22^{-1}**	**5.73^{-1}**	**6.12^{-1}**	**5.75^{-1}**
PSICOV vs. Avg Corr ($\gamma = 0.15$)	2.17^{-2}	**2.42^{-1}**	**6.87^{-1}**	**4.47^{-1}**	3.86^{-4}	4.06^{-3}	2.61^{-2}	**1.89^{-1}**
PSICOV vs. Pred Diff ($\gamma = 0.15$)	2.48^{-9}	3.31^{-7}	4.69^{-4}	3.10^{-4}	3.36^{-7}	2.75^{-7}	1.34^{-5}	1.26^{-4}
PSICOV vs. Pair Pref ($\gamma = 0.10$)	2.61^{-3}	**4.45^{-1}**	**9.93^{-1}**	2.89^{-2}	**7.46^{-2}**	9.33^{-3}	**1.56^{-1}**	**5.42^{-2}**

	[i - j] > 11				[i - j] > 23			
	L	L/2	L/5	L/10	L	L/2	L/5	L/10
PSICOV vs. Avg Freq ($\gamma = 0.05$)	5.42^{-3}	**4.71^{-1}**	**2.60^{-1}**	**8.93^{-1}**	**4.91^{-1}**	**4.40^{-1}**	**5.93^{-2}**	**8.93^{-1}**
PSICOV vs. Avg Corr ($\gamma = 0.15$)	5.24^{-6}	4.26^{-3}	6.06^{-3}	4.29^{-2}	2.29^{-4}	3.18^{-5}	7.88^{-4}	3.33^{-2}
PSICOV vs. Pred Diff ($\gamma = 0.15$)	2.62^{-7}	2.35^{-6}	5.16^{-7}	1.08^{-4}	5.13^{-6}	5.68^{-7}	4.29^{-7}	1.67^{-2}
PSICOV vs. Pair Pref ($\gamma = 0.10$)	2.21^{-4}	2.29^{-2}	**1.02^{-1}**	4.37^{-2}	**2.91^{-1}**	1.06^{-2}	1.68^{-2}	**5.66^{-1}**

Table 8.2: The Wilcoxon signed-rank test applied to PSICOV and variants of APSICOV. For all top-L/l predictions and sequence separations we tested the difference in the precisions of the adjusted method versus those of the original PSICOV over the considered targets. P-values greater than 0.05 are in bold.

We see that the results after adjusting PSICOV using prediction difficulty are significant, with most of the p-values smaller than 0.05. For the other adjustment methods the improvement is not overall significant.

[1] http://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.wilcoxon.html

8.4 Prediction difficulty: best practice

In the above comparison we chose a fixed value of γ for each adjustment method by hand. However, we could use the training data, that was employed to compute the value of the adjustment, also for estimating a best value of γ for each target protein, l and D. In this section I will show the problem retrieving the best results using top-L and Top-$L/10$ lists by a sequence separation of $D > 4$.

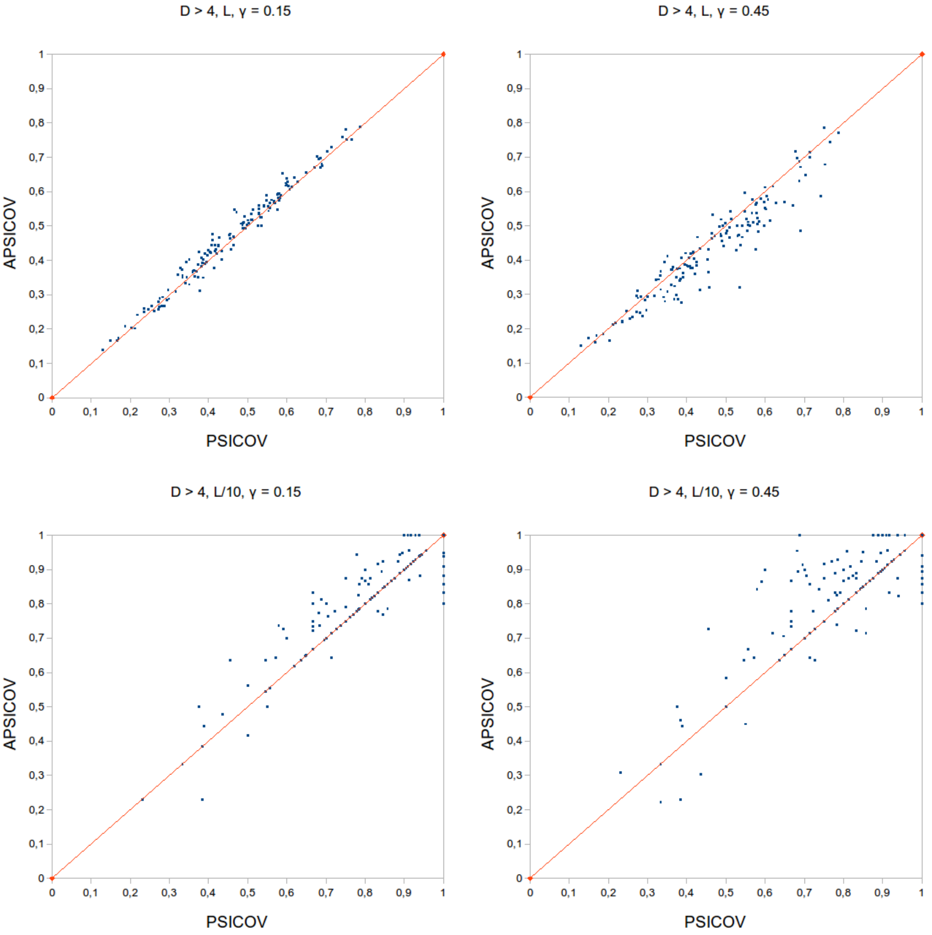


Figure 8.5: Mean precision values, of all targets individually, of the adjusted (APSICOV, using prediction difficulty) predictions versus the original PSICOV predictions.

In figure 8.5 we observe the variety in mean precision values if we change the influence by different values for γ. To ensure that adjusted PSICOV predictions provide the maximum precision for any set of parameters, we must implement a smarter algorithm.

8.4.1 Proposed γ^* algorithm

Algorithm 2 Compute maximum precision using γ^*

```
Input: all targets and chosen adjustment method,
    by $l \in \{L, L/2, L/5, L/10\}$ and $d \in \{4, 8, 11, 23\}$
Output: maximum precision for any set of parameters
for target $t, l, d$ do
    compute training model $T_{t,l,d}$, without using target $t$;
    for $\gamma \in \{0.05, 0.1, ..., 1\}$ do
        compute precision value $P^{\gamma}_{t,l,d}$, using $\gamma T_{t,l,d}$;
    end for
    choose $\gamma^*$ with the maximum precision $P^{\gamma}_{t,l,d}$;
    use $\gamma^*$ and $T_{t,l,d}$ to adjust PSICOV predictions on $t, l, d$;
end for
```

Algorithm 2 estimates a best value of γ for every target protein and and any set of parameters using the training data. For this reason it is a computationally intensive and time consuming task. Given any set of parameters, the γ^* algorithm calculates the greatest average APSICOV precision on the training set.

8.4.2 Results using γ^* algorithm

	[i - j] > 4				[i - j] > 8			
	L	L/2	L/5	L/10	L	L/2	L/5	L/10
PSICOV	0.4606	0.6045	0.7406	0.7916	0.4280	0.5783	0.7279	0.7844
Avg Freq (γ^*)	0.4595	0.6050	0.7424	0,7933	0.4274	0.5768	0.7304	0,7908
Avg Corr (γ^*)	0.4626	0.6029	0.7393	0.7936	0.4314	0.5799	0.7334	0.7839
Pred Diff (γ^*)	0.4709	0.6207	0.7616	0.8216	0.4364	0.5913	0.7460	0.8164
Pair Pref (γ^*)	0.4636	0.6075	0.7420	0.8059	0.4313	0.5835	0.7313	0.7831

	[i - j] > 11				[i - j] > 23			
	L	L/2	L/5	L/10	L	L/2	L/5	L/10
PSICOV	0.4053	0.5571	0.7098	0.7777	0.3343	0.4748	0.6503	0.7367
Avg Freq (γ^*)	0.4048	0.5565	0.7141	0.7783	0.3332	0.4762	0.6541	0.7437
Avg Corr (γ^*)	0.4109	0.5610	0.7158	0.7896	0.3385	0.4836	0.6633	0.7494
Pred Diff (γ^*)	0.4124	0.5654	0.7222	0.8064	0.3416	0.4890	0.6684	0.7536
Pair Pref (γ^*)	0.4057	0.5606	0.7151	0.7836	0.3354	0.4798	0.6509	0.7441

Table 8.3: Mean precision scores using γ^*. For PSICOV and the improved methods the mean precision values for the top-L/l contacts divided by sequence separation ranges where Cβ-Cβ distance < 8Å.

By introducing γ^* we expect the best results for any set of parameters. In general we see higher mean precisions, but for some parameters there are some lower precisions than by using $\gamma = 0.15$. For some targets and parameters the LOOCV procedure does not find the 'best' γ.

8.4.3 Performance of the γ^* algorithm

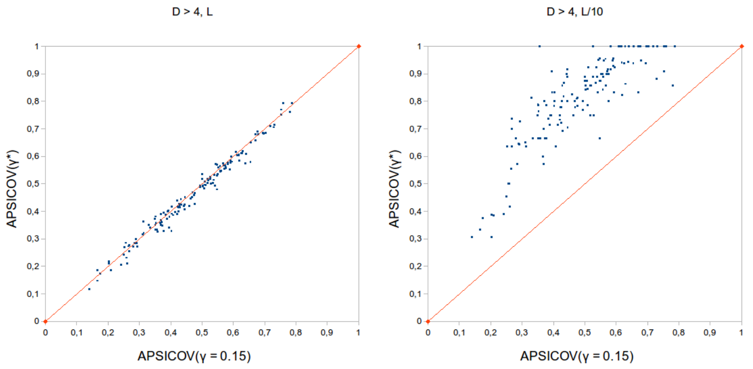


Figure 8.6: Mean precision values, of all targets individually, of APSICOV ($\gamma = 0.15$) vs APSICOV (γ^*). Adjusted by prediction difficulty.

The first results of this study were based on a self chosen value for γ (see table 8.1). Our γ^* algorithm selects a γ for every target, without prior knowledge. Figure 8.6 shows that the performance of the γ^* algorithm hardly decreases in comparison to a self chosen 'best' value for γ. At the same time, we see better results for other sets of parameters.

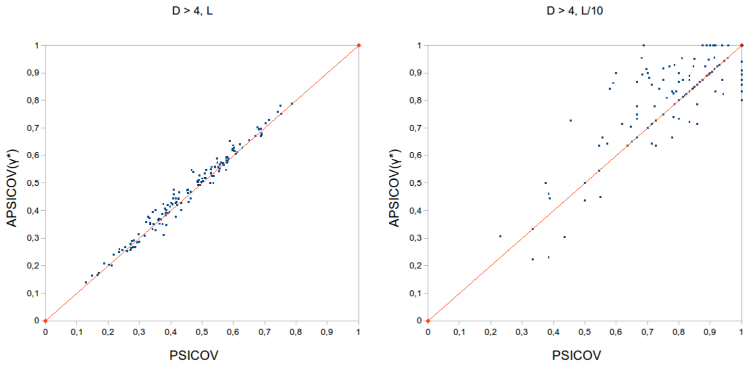


Figure 8.7: Mean precision values, of all targets individually, of PSICOV vs APSICOV (γ^*). Adjusted by prediction difficulty.

When we compare the results of the γ^* algorithm versus the original PSICOV predictions, we see improved mean precisions for almost all targets. Using γ^* we obtain results comparable with the 'best' performances as shown in figure 8.5.

Chapter 9

Discussion

In this chapter I discuss the most important observations of this study and considerations for further research.

Dataset

The quality of contact pair predictions made by PSICOV is strongly dependent on several properties of the input data. Length of a protein and the size of its family are two of them. These properties may lead to larger and more varied multiple sequence alignments, which in turn, makes it easier to find covariance. Also, the quality of MSA algorithms is a factor that affects contact pair predictions made by PSICOV.

For this study we used the same input data as provided by the PSICOV tool. The targets are carefully selected, because of the quality of their MSA's and benchmark purposes. Four out of 150 targets were not suitable for this study according the new versions of PSICOV. An interesting question is how PSICOV performs in a broader perspective. How does PSICOV perform when we use more less strictly selected targets?

Interpretation of contact pairs

As a non-biologist it is hard to keep track of the biological meaning of the data and related research questions. Contact pair prediction is a complicated working field, even within bioinformatics. One of the main issues concerns the interpretation of contact pairs. Information about contact pairs loses value if we consider the single amino acids occurring in the pair. A contact pair within a protein is a unique and specific piece of information and it is not necessarily related to other pairs. In other words, if a contact pair containing amino acids A-D is located at position (i, j), it does not necesarily mean that another pair with the same amino acids A-D but at position $(i+15, j+15)$ is also a contact pair. This fact makes it harder to implement machine learning algorithms that are only based on the amino acid composition of pairs.

Pairs of amino acids are also considered in the analysis of the PSICOV predictions. Unfortunately, the prediction difficulty analysis of pairs showed no significant correlations. The analysis on single amino acids did result in more significant correlations. Another argument for choosing single amino acids over pairs, is the available computing capacity. The amount of vectors grows by a factor 10.5 when using pairs of amino acids in the adjusted method.

Prediction difficulty as a correction model

Using prediction difficulty is an ad-hoc adjustment model for improving structural contact prediction with PSICOV, it is a straightforward method that measures the quality of all significant (top-L/l) PSICOV predictions. The output of PSICOV is a likelihood where the top-L/l list is considered as true contact pair. We separate the false positives (FP) from the true positives (TP) and define prediction difficulty by (TP-FP). You could say that the analysis of predictions has more value if we take the whole spectrum including also true negatives and false negatives into account. The problem that then arises is how you would define true and false negatives. There is no defined prediction score of bottem-L lists that determined a true negative contact pair.

Comparison of results

For improvements on PSICOV's contact pair prediction we used four different adjustment models for PSICOV. In particular, prediction difficulty is based on our own prediction analysis, and pairing preference which is based on the study of Glaser et al., showed (potential) improvements in structural contact prediction [8]. It is obvious that prediction difficulty scores better than pairing preference, which is not surprising because this model has been generated using knowledge on contact pairs and PSICOV performance on a training set. Adjustments based on amino acid frequency in target sequences and their correlations with predictions did not yield very good results. The Wilcoxon test also indicated that the improvements of this type of adjustment are not significant.

There are opportunities to improve the adjustments for both prediction difficulty and pairing preference. As discussed in the interpretation of contact pairs, we could consider a broader perspective of contact pair prediction. We can incorporate information about true and false negatives into prediction difficulty. An adjustment model based on pairing preference could also be extended. Amino acid features such as hydrophobic/hydrophilic, charged residues, and simply size are important factors which could be analyzed and used to adjust the output of PSICOV. Unfortunately, as information scientist I feel to lack fundamental knowledge in biochemistry and molecular biology to develop a thoroughly biologically grounded pairing preference adjustment model.

Problem of using different parameters

In table 8.1 the results for a fixed value of γ are shown, where this value was selected by hand by looking at the highest improvements for top-$L, D > 4$. The reason for this is that the most improvements are expected to be made on the largest set of predictions, that is, by using top-$L, D > 4$. This ensures that we do not see the best results for the other parameter settings. The problem that arises is shown in figure 8.5.

A solution is the implementation of algorithm 2 (γ^*). We are able to choose the 'best' γ for any set of parameters (the highest peak in the graphs of adjustments) using γ^*.We compute the γ with the greatest improvements for every target by any set of parameters. For some targets, the algorithm does not find the optimal value for γ, but figure 8.6 and 8.7 shows that the γ^* algorithm performs almost as good as picking a γ with prior knowledge. It even shows that it handles the variation in parameters very well.

Future work

There are still a lot of improvements to make on precise structural contact prediction. This is only the beginning of combining covariances in MSA's and residue/protein characteristics for contact pair prediction purposes. Like Jones et al. states, it is likely that PSICOV reaches a higher accuracy when further research combines structural information such as predicted secondary structures [13].

In further research we could look at specific pairs instead of single amino acids and also consider true and false negatives. As a threshold on true and false negatives, we could consider a top-L/l list where the number of predictions is equal to all contact pairs defined by the golden truth. For any pair of residues, we could fill in the spectrum of true/false positives and true/false negatives.

It would also be interesting to look at the behavior of PSICOV using new input data. Instead of using carefully selected data, we could use all proteins where its three-dimensional structure is known by experiments. How would PSICOV perform and what are the improvements after implementation of our findings?

Chapter 10

Conclusions

PSICOV is a powerful tool for precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. This study shows that there is still room for improvement in the used prediction methods. The assumption that there is a PSICOV prediction difficulty for every amino acid turns out to be a relevant piece of information. To a lesser extent, we have shown that other amino acid characteristics such as pairing preference, may also be useful and improve residue-residue contact predictions.

Jones et al. states that PSICOV could be further improved by incorporating it in a more general machine learning-based contact prediction method [13]. Information like predicted secondary structures, better alignment algorithms or sequence weighting would significantly improve the results. We have demonstrated that the assumption of Jones et al. can be confirmed. Improvements in mean precision from 0.01 to just over 0.03 are made by a correction model based on prediction difficulty. This comes down to a few (approximately from 1 up to 4) more true positive predictions for each target protein, where the sequence length of the proteins varies from 50 residues till 266 residues.

Unfortunately, our method also leads to some targets with a higher false positive rate. The discussed methods and further work must result in greater improvements in further studies. At this time, we can at least conclude that there are opportunities to be exploited to improve structural contact prediction.

A hybrid approach

Grishin noticed: "*Algorithm development studies usually reported 'post,-' rather than 'predictions,' benchmarked on proteins with know spatial structures* [10]." The biggest problem within contact pair prediction is that in all studies we use already known protein structures. There is no such thing as testing truly blind predictions. Besides Jones et al., Marks et al. suggest a combined approach of experimental and computational structural biology [17]. This study and the suggested improvements by Jones et al. and Marks et al., comes down to a hybrid algorithm using more and more available information. We must keep in mind that the main goal of contact pair prediction is the actual prediction of an unknown structure. Of course we should use the already available information, but the trained algorithms must perform just as good on 'truly' unknown protein structures.

Bibliography

[1] Banerjee, O., Ghaoui, L. E., and d'Aspremont, A. Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *Journal of Machine Learning Research 9* (2008), 485–516.

[2] Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. The protein data bank. *Nucleic Acids Research 38*, Database issue (2000), D211–D222.

[3] Branden, C., and Tooze, J. *Introduction to Protein Structure.* Garland Publishing, Inc., 1999.

[4] Campbell, N. A., and Reece, J. B. *Biology.* Pearson Education Inc., Benjamin Cummings, 2005.

[5] Ezkurdia, I., Graña, O., josé M. G. Izarzugaza, and Tress, M. L. Assessment of domain boundary predictions and the prediction of intramolecular contacts in CASP8. *Proteins 77*, Suppl. 9 (2009), 196–209.

[6] Finn, R. D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J. E., Gavin, O. L., Gunasekaran, P., Ceric, G., Forslund, K., Holm, L., Sonnhammer, E. L. L., Eddy, S. R., and Bateman, A. The pfam protein families database. *Nucleic Acids Research 38*, Database issue (2010), D211–D222.

[7] Friedman, J., Hastie, T., and Tibshirani, R. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics 9*, 3 (2008), 432–411.

[8] Glaser, F., Steinberg, D. M., Vakser, I. A., and Ben-Tal, N. Residue frequencies and pairing preferences at protein protein interfaces. *Proteins 43* (2001), 89–102.

[9] Gobel, U., Sander, C., Schneider, R., and Valencia, A. Correlated mutations and residue contacts in proteins. *Proteins 18* (1994), 309–317.

[10] Grishin, N. V. Membrane protein structure prediction for exploration. *Cell 149*, 7 (2012), 1424–1425.

[11] Gromiha, M. M., and Selvaraj, S. Inter-residue interactions in protein folding and stability. *Progress in Biophysics and Molecular Biology 86* (2004), 235–277.

[12] Hekkelman, M. L., te Beek, T. A. H., Pettifer, S. R., Thorne, D., Attwood, T. K., and Vriend, G. WIWS: a protein structure bioinformatics web service collection. *Nucleic Acids Research 38* (2010), W719–723.

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

[14] Joosten, R. P., te Beek, T. A., Krieger, E., Hekkelman, M. L., Hooft, R. W., Schneider, R., Sander, C., and Vriend, G. A series of pdb related databases for everyday needs. *Nucleic Acids Research 39*, Database issue (2011), D411–D419.

[15] Lapedes, A., Giraud, B., and Jarzynski, C. Using sequence alignments to predict protein structure and stability with high accuracy. *arXiv 29* (2012).

[16] Marks, D. S., Colwell, L. J., Sheridan, R., Hopf, T. A., Pagnani, A., Zecchina, R., and Sander, C. Protein 3d structure computed from evolutionary sequence variation. *PLoS ONE 6*, 12 (2011).

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

[18] Meinshausen, N., and Bühlmann, P. High-dimensional graphs and variable selection with lasso. *The Annals of Statistics 34*, 3 (2006), 1436–1462.

[19] Nelsen, D. L., and Cox, M. M. *Principles of Biochemistry.* W.H. Freeman and Company, 2005.

[20] Olmea, O., and Valencia, A. Improving contact predictions by the combination of correlated mutations and other sources of sequence information. *Folding and Design 2* (1997), S25–S32.

[21] PDB-101. Looking at structures: Dealing with coordinates. `rcsb.org/pdb/101/structural_view_of_biology.do`, 2013. Accessed: 2013-10-22.

[22] Pevsner, J. *Bioinformatics and functional genomics.* John Wiley & Sons, Inc., 2003.

[23] Xu, Y., Xu, D., and Liang, J. *Computational Methods for Protein Structure Prediction and Modeling: Volume 1: Basic Characterization.* Springer, 2007.

[24] Yuan, M., and Lin, Y. Model selection and estimation in the gaussian graphical model. *Biometrika 94* (2007), 19–35.

[25] Zvelebil, M., and Baum, J. O. *Understanding bioinformatics.* Garland Science, Taylor & Francis Group, 2008.