Charles University
Faculty of Science

Study programme: Bioinformatics
Branch of study: Bioinformatics

Hana Pařízková

Bioinformatic methods of detection of protein coevolution
Bioinformatické metody detekce koevoluce proteinů

Bachelor's thesis

Supervisor: doc. Ing. Bohdan Schneider, CSc.

Prague 2018

Prohlášení

Prohlašuji, že jsem závěrečnou práci zpracovala samostatně a že jsem uvedla všechny použité informační zdroje a literaturu. Tato práce ani její podstatná část nebyla předložena k získání jiného nebo stejného akademického titulu.

V Praze, 24. 4. 2018 Hana Pařízková

Acknowledgement

I thank doc. Ing. Bohdan Schneider, CSc., the supervisor of this thesis, for his professional guidance and useful remarks.

I am very grateful to all the people who introduced to me the field of bioinformatics and helped me to discover its charm and beauty.

My biggest thanks goes to my dearest partner Vašek. Not only the many inspiring discussions, but especially your everlasting support, patience and love made writing the thesis much easier.

Abstract

The term coevolution describes the situation when two or more species or biomolecules reciprocally affect each others' evolution. On the protein level, it is thought to be the main mechanism ensuring correct folding, interactions and function of a protein, and it can be observed both on the level of interacting protein families and individual amino acid residues. Coevolution studies have been proved to be a powerful tool for prediction of protein structure, function, interaction partners, etc. In this thesis, different algorithms used for detection of protein coevolution are described, as well as their applications and limitations.

Keywords: coevolution, protein family, protein structure prediction, interaction partners, correlated mutations, mirrortree, mutual information, direct coupling analysis

Abstrakt

Slovem koevoluce popisujeme stav, kdy dva či více druhů nebo biomolekul vzájemně ovlivňují svou evoluci. Na proteinové úrovni je koevoluce považována za jeden z hlavních mechanismů zajišťujících správné sbalení, interakce a funkci proteinů. Pozorována může být jak na úrovni interagujících proteinových rodin, tak na úrovni jednotlivých aminokyselinových residuí. Studium koevoluce může být užitečným nástrojem při predikci struktury proteinů, jejich funkce, interakčních partnerů, apod. V této práci jsou popsány algoritmy, které jsou používány k detekci koevoluce proteinů, stejně jako jejich možné aplikace a omezení.

Klíčová slova: koevoluce, proteinová rodina, predikce struktury proteinů, interakční partneři, korelované mutace, mirrortree, vzájemná informace, analýza přímého párování

Contents

1. Introduction

1.1 What is coevolution?

The term 'coevolution' is commonly defined as 'reciprocal evolutionary change in interacting species' [1], i.e. it describes the situation when two or more species or biomolecules reciprocally affect each others' evolution. Coevolution may be observed at the level of:

- **Species:** Coevolution is the underlying principle of the so called 'Red Queen hypothesis' [2], which describes e.g. the relations between a host and a parasite, or between symbiotic species. It is also the underlying mechanism of the phenomenon of mimicry [3, 4].
- **Populations:** The gene pool of a population may be adapted so that an average individual has as high fitness as possible. This was described e.g. by T. Dobzhansky for populations of *Drosophila pseudoobscura* [5, 6].
- **Chromosomes:** Coevolution operating on chromosomal level is not very common, because in most cases chromosomes are not inherited as a whole due to recombination. However, if recombination is prevented by massive chromosomal changes (e.g. inversions), coevolution of chromosomal types may occur as described by Dobzhansky [5, 6].
- **Biomolecules:** Physically interacting and/or functionally related biomolecules (proteins, DNA, RNAs, etc.) tend to have similar evolutionary histories [7, 8].
- **Residues:** Finally, coevolution is acting on functionally or physically interacting residues (amino acids, nucleotides) of biomolecules, typically to maintain the function or structure of the molecule [9, 10].

1.2 A brief history

The basic description of a coevolutionary process may be found already in the works of Charles Darwin: in the *Origin of Species* [11] and mainly in his work on orchids and their pollinators [12]. Here Darwin describes orchid *Angræcum sesquipedale* (now also known as Darwin's orchid; see figure 1.1(a)) from Madagascar which has an extremely long nectary (up to 30 cm) with only the lower 4 cm filled with nectar. Darwin suggests that there must exist a moth with proboscis approx. 26 cm long in Madagascar (no such moth was known at that time) and depicts the mutual dependence of the moth and the orchid on each other: *'If such great moths were to become extinct in Madagascar, assuredly the* Angræcum *would become extinct.'* In 1903 such a moth was really discovered by Walter Rotschild and Karl Jordan and named *Xanthopan morganii praedicta* [13] (see figure 1.1(b)).

In 1879 Fritz Müller was the first one to describe the already known phenomenon of mimicry by means of natural selection, and thus coevolution [3], including also quantitative statements about the benefit of involved species with respect to relative population sizes.

(a) (b)

Figure 1.1: *Angræcum sesquipedale* (a) and its pollinator *Xanthopan morganii praedicta* (b). The existence of an orchid with an extremely long nectary led Darwin to prediction of a moth with an equally long proboscis. Both pictures downloaded from Wikimedia Commons (`https://commons.wikimedia.org`).

The first description of coevolution operating also at lower than species level was made by Theodosius Dobzhansky in the middle of 20th century [5, 6]. He studied different chromosomal types of *Drosophila pseudoobscura*. In natural populations, carriers of different chromosomal types differed in their fitness, and heterozygotes showed higher fitness than homozygotes (phenomenon known as *heterosis*). But interestingly, no heterosis effect occurred when flies from two isolated populations were crossed. Dobzhansky suggests that the chromosomal types in a population are 'coadapted' so that heterozygotes have higher fitness.

The term 'coevolution' is usually attributed to Paul Ehrlich who studied relations between butterflies and their food plants and defined coevolution as 'reciprocal selective responses between ecologically closely linked organisms' [14].

1.3 Coevolution of proteins

Cellular processes are mostly sustained by protein interactions and protein-catalyzed reactions. In general, both of these processes are highly specific: proteins recognize very limited range of targets and bind them in highly regular manner, and the enzymes are specialized to perform one or only a few reactions. The specificity of the interactions is determined by structural and physicochemical properties of the protein. As a result, the sequence and structure of the proteins are under certain evolutionary constraints [15, 16]: The amino acids must pack correctly against each other to sustain such a structure, that the protein can interact with (and only with) correct partner in a correct way; the amino acids in the catalytic center must perform the correct reaction, so that the whole biochemical pathway the protein is involved in will not stop, etc. As a result, neither the amino acid residues, nor the whole proteins act in isolation, and coevolution

may be observed both on the level of individual amino acid positions and on the level of whole proteins [7, 9, 10].

However, the process of speciation, common to all protein families, causes evolutionary histories of all protein families to look somewhat similar. As a result, correlations resembling the signature of coevolution may be observed even for proteins/residues which are not coevolving at all. Thus, if we want to search for coevolving proteins/residues, we must carefully distinguish the true coevolutionary signal from the 'phylogenetic noise'.

In this review, we will describe different approaches of coevolution detection, as well as their applications and limitations. All of these topics will be discussed on the level of both protein-protein and residue-residue interactions.

1.4 Terminology

Throughout the thesis, we will use the following terminology:

- Words *residue*, *site* or *position* all refer to a position in the protein primary sequence, regardless of the amino acid type occurring there; we suppose that a given residue is homologous throughout the whole protein family.
- The *row* of a multiple sequence alignment (MSA) corresponds to the primary sequence of a single protein; the *column* corresponds to amino acids occurring at given position in individual proteins.
- The *size* of an MSA refers to number of sequences, i.e. number of rows; the *length* to the length of these sequences, i.e. number of columns.
- *Intra-protein pairs* are pairs of residues located both in one protein molecule; *inter-protein pairs* are pairs of residues coming one from one protein and second from another one.

1.5 Objectives of the thesis

The objectives of this thesis are:

- to review different approaches for detection of protein coevolution, both for protein-protein and residue-residue interactions,
- to discuss the reliability of these methods, and
- to show the possible applications, as well as limitations of coevolution prediction.

2. Residue-residue coevolution

2.1 Introduction

An amino acid residue is the smallest unit on which evolution may operate in a protein. Each site's evolution is constrained by its relations with plenty of other residues: the residue must pack correctly against them, may cooperate with them to perform the catalytic activity of the protein or be crucial in binding or recognition of other macromolecules, etc. [17]. A mutation of the residue then changes also the 'evolutionary landscape' of these related residues – mutations that would be highly unfavourable previously may become advantageous and vice versa. In other words, a substitution at one site will potentially affect the substitution rates at other sites, and the mutations at these sites will tend to cluster together [18]. This mechanism is called *compensatory mutations* and it is used for the most commonly used definition of amino acid coevolution. According to this definition, two sites are coevolving if they undergo compensatory mutations.

In general, protein sequence changes much more rapidly than its structure [19]. Compensatory mutations thus most often compensate for the changes in volume, charge or hydrophobicity of amino acids, ensuring thus the correct folding and function of the protein. Experimentally, they were observed in many studies, e.g. [20, 21, 22, 23, 24, 25]. The compensatory mutation may occur either in the same protein or in its interaction partner and complicated chains of interactions may lead to compensatory mutations occurring at spatially distant residues, not only at residues being in direct contact [23, 26, 27].

Since late 1980's several research groups have computationally studied the phenomenon of intra-protein compensatory mutations (e.g. [28, 29, 30, 31]). Technical differences in their approaches led to somewhat conflicting results regarding the mechanism, importance and intensity of this phenomenon. However, all of them concluded that residues showing similar substitution patterns are generally closer to each other in the three-dimensional structure than average.

Studies searching for inter-molecular pairs of residues showing correlated mutations are less common (e.g. [32, 33, 34, 35]). However, they also show that residues with similar substitution patterns tend to be in closer proximity than average. Nonetheless, not all protein-protein interfaces do show coevolution [35].

However, there exists also a broader definition of coevolution. According to this definition, coevolution is viewed simply as similarity of evolutionary histories [36]. Of course, compensatory mutations cause the evolutionary histories to mirror each other, and are thus one of the possible causes of the coevolution in the broader sense. Another possible cause of such similarity is a situation when structurally or functionally important regions are changing simultaneously in order to obtain new function, e.g. after gene duplication [37]. Yet another source of coevolution may be so called *heterotachy*. Heterotachy is defined as within-site variation in substitution rate over time [38]. If some structural or functional constraints are relaxed in a lineage, residues involved in these constraints may mutate more freely, and thus they cumulate mutations. This may be then observed as correlated (i.e. simultaneously happening) mutations, although the mutations are not functionally dependent on each other [36].

When searching for coevolving residues, we usually want to define coevolution in the narrower sense, as compensatory mutations carry the structural and functional meaning we are usually looking for. However, it is not easy (if not impossible) to distinguish true compensatory mutations from other events which are exhibited similarly.

In the following sections, we will give an overview of the methods used for detection of coevolving residues. They will be divided into two groups: older local, covariance-based methods and newer global approaches. The most important difference between these two groups is that while the older methods simply look at one pair of residues at a time, compute a test statistics and based on its value mark the pair as coevolving or not, the newer ones are optimizing the whole set of possible relations. As a result, the global methods are able to disentangle which of the observed correlations stem from direct physical contact of given residues and which do not. Of course, most of the modern approaches use some of the older ones at some time in the computation.

Selected implementations, available either as web servers or downloadable programs, of the below described algorithms are listed in table 2.1, and comparison of their prediction accuracy is depicted on figure 2.1 – both can be found at the end of this chapter.

2.2 Local methods

Detection of coevolving sites is naturally possible only if we have some (at least indirect) information about the evolutionary history of the protein family. This information is usually represented by an MSA, some methods also use a phylogenetic tree. The approaches for the local detection of correlated mutations further differ by the statistics used to measure non-independence (e.g. correlation coefficient, mutual information), by the way they account for biochemical properties of different amino acids (if they do), the size of detected groups and the way they assess significance of the results. In the following sections, we will briefly discuss all of the points mentioned above.

2.2.1 Test statistics

Here, we will shortly introduce some of the test statistics commonly used in the local detection of coevolution. We will focus mainly on those of them that proved to be useful and thus they are used also in the contemporary methods of coevolution detection.

In the following text, N is the number of sequences in the MSA, $p[i]$ is the amino acid located at position i in protein p, $s(A, B)$ is the similarity of amino acids A and B (given by an amino acid similarity matrix, see section 2.2.2), v_i is a vector of length N of some physicochemical property (e.g. volume, hydrophobicity; see section 2.2.2) of amino acids at position i and $P(X^i)$ is the probability of finding amino acid X in the i-th column.

Pearson correlation coefficient [30, 39] Degree of coevolution R_{ij} between residues i and j may be computed as weighted Pearson correlation coefficient of

the pairwise similarities of amino acids at positions i and j:

$$R_{ij} = \frac{1}{N^2} \sum_{p=1}^{N} \sum_{q=1}^{N} \frac{w_{pq}(s(p[i], q[i]) - \mu^i)(s(p[j], q[j]) - \mu^j)}{\sigma^i \sigma^j} \tag{2.1}$$

where w_{pq} is a measure of distance of the p-th and q-th sequence (e.g. fraction of residue mismatches over the whole alignment), μ^i is the mean value of pairwise similarities of residues at position i and σ^i is their standard deviation.

Equation 2.1 may be modified e.g. by using physicochemical properties of amino acids instead of their pairwise similarities [39], by omitting the weighting by w_{pq} [40, 41], or by replacing the similarity values with their ordinal ranking number [42].

OMES [24, 43] Observed Minus Expected Squared (OMES) statistics mimics the χ^2-test, well known from statistics.

The score R_{ij} for columns i and j is given by

$$R_{ij} = \sum_{X,Y} \frac{(N_{obs} - N_{exp})^2}{N} \tag{2.2}$$

where X and Y go over all distinct amino acids occurring in columns i and j, respectively, N_{obs} is the number of times each distinct pair was observed (X in column i and Y in column j) and N_{exp} is the number of times we would expect residues X and Y to co-occur in columns i and j, respectively, given their single occurrences in columns i and j. The value of N_{exp} may be computed as follows:

$$N_{exp} = \frac{N_{X^i} N_{Y^j}}{N} \tag{2.3}$$

where N_{X^i} is the number of times amino acid X occurs in column i and N_{Y^j} is the number of times amino acid Y occurs in column j.

Statistical coupling analysis [44, 45, 46] Statistical coupling analysis (SCA) is an example of a method based on a perturbation of the MSA. For each column i of the MSA, we construct a subalignment as follows: Let X be the most prevalent amino acid in column i, then the subalignment is composed from only those sequences in which there is X on i-th position. In other words, we fix i-th amino acid to X and then we examine how the composition of other positions has changed.

In the original publications [44, 45, 46], correlated mutation score ΔG_{ij} was expressed as an 'energetic term' mimicking the equation for Gibbs free energy. The idea was that it should represent the degree of thermodynamic coupling between residues i and j [44]. However, no such meaning was confirmed in latter studies [41, 47], and thus the equation was simplified to [47]:

$$\Delta G_{ij} = \sqrt{\sum_{X} (\ln P(X^i|\delta^j) - \ln P(X^i))^2} \tag{2.4}$$

where X goes over all amino acids, $P(X^i|\delta^j)$ is the probability of finding amino acid X in column i in the subalignment pertubated according to column j, and the probability $P(X^i)$ is estimated by the corresponding frequency.

The idea of MSA perturbation was used also in the algorithm of Dekker *et al.* [48]. They introduce 'explicit likelihood of subset covariation', a statistics that, briefly speaking, measures what part of subalignments of a given size would have the same composition in individual columns as the observed subalignment.

Mutual information [49, 50, 51] Mutual information (MI) measures how much information one random variable provides about another one [52]. Mutual information between residues i and j is computed as follows:

$$MI_{ij} = \sum_{X,Y} P(X^i, Y^j) \log \frac{P(X^i, Y^j)}{P(X^i)P(Y^j)} \tag{2.5}$$

where X and Y go over all amino acids occurring in columns i and j, respectively, and the probabilities are estimated by the corresponding frequencies.

The simple MI was shown to perform relatively poorly in comparison with the other covariance-based methods [33, 41, 53] (see also figure 2.1(a)). However, modifications accounting for the noise generated by the shared evolutionary history of all sequences, small number of observations and data redundancy were suggested [54, 55], and these improved its overall performance. Nowadays, MI with the corrections is probably the most often used method for detection of correlated residues in the modern algorithms (see section 2.3 for details).

Mapping substitutions on phylogenetic tree [31, 36, 37] The main pitfall of the previously described covariance-based methods is their inability to distinguish the functional signal (resulting from true coevolution) from the signal generated only because of the shared history of all sites (see also figure 5.1). Detection accuracy may be increased by mapping the substitution events on the phylogenetic tree, which allows us to compare only those changes that happened at the same evolutionary interval [37].

First of such methods was proposed in 1994 by Shindyalov *et al.* [31]. Let us have an MSA of length L and a corresponding phylogenetic tree with K branches and $(K-1)$ vertices. Leaves of the tree correspond to the sequences from the MSA, inner vertices to the ancestral sequences (predicted when computing the tree). Define matrix M of dimensions $L \times K$ as follows:

$$M_{ik} = \begin{cases} 1 & \text{if the amino acids at position } i \text{ differ at the ends of } k\text{-th branch,} \\ 0 & \text{otherwise} \end{cases}$$

(i.e. multiple or back mutations are not taken into account).

Substitutions at positions i and j are told to be correlated if they occur on the same branch k of the phylogenetic tree, i.e. if $M_{ik}M_{jk} = 1$. The total number of correlated mutations for pair i and j is then simply the sum over all branches, and using probability theory, we can estimate the probability of observing this number of correlated mutations by chance.

This approach was then transformed several times. Instead of assigning 1/0 value to each branch of the tree, we can assign physicochemical distance of amino acids observed at the ends of the branch [36, 37, 56], or we can estimate the true number of substitutions (including back and multiple mutations) [36, 56]. As the

ancestral sequences in the inner nodes of the tree are only estimated, reliability of the prediction may be increased by averaging the number of substitutions over all possible pairs of ancestral states [36, 56]. The similarity of substitution histories of residues i and j may be computed as correlation coefficient [36, 37, 56].

Other approaches for mapping substitutions on the phylogenetic tree include Bayesian mutational mapping [57] or simulation of the evolution by Markov chain process followed by maximum likelihood [58].

Other methods The previous list of algorithms is by no means complete. Other statistics suggested for coevolution detection include e.g. so called quartets [59], clustering based on pairwise similarities [60], prediction using neural networks [61], multiple interdependency [62], and many others. However, as these are not used any more in practice, they will not be discussed here in detail.

2.2.2 Properties of amino acids

The effects of a substitution depend greatly on the nature of the original and the substituted amino acid. While some substitutions do not change the stability or functionality of a protein at all, others may be fatal [20]. Thus, physicochemical or evolutionary properties of amino acids are often taken into account when searching for coevolving positions.

Physicochemical properties used include side chain volume, charge, hydrophobicity or Grantham formula combining atomic composition, polarity and volume [63]. Of course, several properties may be combined to form a vector [64]. For the comparison of two amino acids, similarity matrices may be used, e.g. by McLachlan [65], Dayhoff [66], Miyata [67] or Taylor and Jones [68]. As expected, using different properties or similarity matrices leads to different results obtained (analyzed e.g. in [37] or [60]).

When physicochemical properties are used, one may be interested not only in the magnitude of the change, but also in its direction, and search for true compensatory substitutions, i.e. such substitutions that the total value of the followed property (e.g. volume) remains conserved. The amount of conservation of given property between residues i and j may be expressed as *compensation index* C_{ij} [36]:

$$C_{ij} = 1 - \frac{\|\tilde{v_i} + \tilde{v_j}\|}{\|\tilde{v_i}\| + \|\tilde{v_j}\|} \tag{2.6}$$

where $\tilde{v_i}$ is the vector of signed changes of the property at site i and $\|\tilde{v_i}\|$ is the L_2-norm of vector $\tilde{v_i}$. When the substitutions at positions i and j tend to compensate themselves (i.e. $\|\tilde{v_i} + \tilde{v_j}\|$ is close to $\vec{0}$), the compensation index is close to 1, when the changes tend to be in the same direction, the compensation index is close to 0.

Although incorporating the physicochemical properties of amino acids was shown to improve the prediction accuracy in some studies [37], there are also some arguments against. First, different similarity matrices often give very different scores to a given pair of amino acids (the correlation coefficient for McLachlan and Miyata matrices is only 0.32 [33]). Second, it is known that in some protein families, substitution even to a very similar amino acid may lead to drastic effects

[69]. Also, not all correlated mutations may be explained in terms of physico-chemical properties – it was reported that only half of all correlated pairs in SH3 domains is compensating for volume or charge [24].

2.2.3 Detecting groups of coevolving residues

All of the previously described methodologies search only for pairs of correlated residues. However, bigger groups of coevolving residues may also provide valuable information, although it is not always related to direct contacts. Such groups are often connected with the functionality or specificity of the protein – they can form e.g. the clusters that make up ligand binding site [70, 71, 72] or the chains of residues responsible for allosteric changes [73]. Thus, some researchers tried to develop methods able to detect groups of coevolving residues instead of just pairs. Extending the definition of the test statistics from a pair to bigger groups is usually possible. However, exhaustive testing of coevolution on all groups of arbitrary size is extremely computationally demanding due to the high number of possible combinations. Thus, more efficient methods have to be employed, such as principal component analysis [37] or clustering techniques [36].

2.2.4 Assessment of significance

After computing the test statistics, statistical significance of the obtained values should be determined.

The oldest studies [30, 32] simply sorted the pairs according to the test statistics and declared the top C pairs to be coevolving, where C is a function of length L of the protein (e.g. $C = L/2$). This, unfortunately, has no statistical support.

Probability theory may be used to assess the significance of the results (e.g. in [24, 31, 41, 43, 60, 62]), especially if we know the theoretical distribution of the test statistics under the null (independence) hypothesis. If this distribution is not known, is not precise or if we want to account for the specific set of proteins we are working with, simulation of independent data by bootstrap is common (e.g. in [24, 36, 37, 50, 51, 56, 58]).

2.3 Global methods

Nowadays, the most common application of coevolution detection is in protein structure prediction, i.e. we want to find residues being in direct contact – this problem is sometimes called direct coupling analysis (DCA). However, the old local methods often fail in this task as reported by many studies (e.g. [41, 64, 74, 75, 76]). Two reasons identified in 1999 by Lapedes *et al.* [77] are:

- Bias is introduced into the calculation of the test statistics due to the fact that biological sequences are generally related by a phylogenetic tree.
- Linked chains of correlations (i.e. residue i is correlated with residue j and j is correlated with k, thus i shows some correlation with k) lead to pairs of sites showing correlation although they are not spatially proximate [44, 46, 74]. (Note that this is not in contradiction with the definition of coevolution: Such residues definitely **are** affecting each other and **are** thus

coevolving. If we are searching for coevolving residues in general, we do not mind reporting such a pair.)

The first point mentioned above may be overcome by incorporating the evolutionary information into the computation (see section 2.2.1, or e.g. [54, 62]). The second problem, i.e. inferring interactions from observations of instances, has already been studied in statistical physics, machine learning (so called model learning) and statistics [78]. Thus, it comes as no surprise that knowledge and algorithms from these fields were used also to solve DCA.

2.3.1 Maximum entropy approach

Maximum entropy approach to coevolution detection was introduced, although purely theoretically, already in 1999 by [77], and further improved in many studies (e.g. [76, 78, 79, 80]). We assume that pairs being in direct contact are a subset of the pairs identified by a covariance-based method as described in section 2.2 (mutual information with corrections by [54] is used most often). The task is to determine which of the identified pairs are truly in contact and which are not.

This can be done by finding a model given by parameters e_{ij}, determining the amount of direct coupling between residues i and j, and h_i, determining the composition bias (i.e. preference for some amino acids) at position i. Of course we want the model to describe well the observed data, i.e. the probability of observing single amino acids and amino acid pairs given by the model should match the observed frequencies (we could require the same for triplets, quartets, etc., but it would be computationally untractable). As a second condition, we want the model to be as simple as possible, which is equivalent to the set of the parameters having maximum possible entropy (i.e. maximally even distribution). This is called maximum entropy model.

Finding such a model is computationally extremely difficult, however, several very diverse approximate approaches, inspired by similar problems e.g. in statistical physics or information theory, were suggested. The first one, mpDCA [79], was, though approximate, still highly computationally intensive, and is thus not used any more. However, the following approaches – mfDCA [76, 81], plmDCA [78] or GREMLIN [75, 80] – are, thanks to their accuracy and speed, nowadays widely used.

2.3.2 PSICOV

PSICOV (Protein Sparse Inverse COVariance) algorithm [82] is based on inversion of the covariance matrix.

The covariance cov_{ij}^{AB} between amino acids A on position i and B on position j can be estimated as:

$$cov_{ij}^{AB} = f(A^i, B^j) - f(A^i)f(B^j) \tag{2.7}$$

where $f(A^i)$ is the frequency of amino acid A on the i-th position, and similarly, $f(A^i, B^j)$ is the frequency of the pair of amino acids A and B occurring in columns i and j, respectively. These values are stored in one covariance matrix indexed by the pair consisting of the position and an amino acid.

As discussed earlier, covariance (or correlation) is not a good measure of direct coupling between two variables. This can be overcome by computing an inverse covariance matrix (so called *precision* or *concentration matrix*) Θ where, instead of covariances, *partial correlations* between the two variables are given. Partial correlation of variables X and Y is the correlation between X and Y conditioned on (i.e. with controlling effect of) all other variables [83, 84]. In other words, it gives us a measure of direct coupling between variables X and Y. Thus, the off-diagonal elements of Θ which are significantly greater than 0 are identifying pairs of residues which are likely to be in spatial proximity.

However, the empirical covariance matrices for protein sequences are very sparse (i.e. there are a lot of 0's) and thus singular, and precise inverse matrix does not exist. Thus, approximate inverse matrix must be estimated.

The final score $\mathcal{S}_{ij}$ for residues i and j giving the 'amount of direct contact' between i and j can be then computed simply as

$$\mathcal{S}_{ij} = \sum_{A,B} |\Theta_{ij}^{AB}| \tag{2.8}$$

where A and B run over all amino acids and Θ_{ij}^{AB} is the partial correlation of amino acid A at position i with amino acid B at position j.

2.3.3 Bayesian network approach

Bayesian network approach [74] is based on the idea of chains of contacts as described in the introduction to section 2.3. In [74] it was shown that these chains of contacts are responsible for a large part of not directly interacting, but correlated residues. Thus, directly interacting residues may be identified by excluding those correlated pairs which can be explained by chains of other correlations.

The implementation is based on the previous article by the same authors [85]. A *dependency tree* is determining the interacting residues: edge going from vertex i to vertex j means that residue j depends on residue i. For simplicity, each residue (except the root of the tree) depends on exactly one other residue. Using Bayesian statistics, statistical weight of each dependency tree is determined. Trees with higher statistical weight should be composed mainly from those edges whose dependency cannot be explained by chains of other edges. The posterior probability of residues i and j interacting directly can then be quantified by calculating the sum of the statistical weights of all the dependency trees in which the edge (i, j) appears.

A big advantage of this approach is that, in contrast to the maximum entropy approach, we do not use any free parameters, which results in much shorter computation time when compared both with DCA algorithms [74] and PSICOV [82]. Also, incorporating prior information about known contacts is easy. However, it was reported that accuracy of both the methods using maximum entropy and PSICOV is higher [76, 81, 82], and the Bayesian network approach is nowadays scarcely used.

2.4 Fusion methods

To further increase the accuracy of residue-residue contact prediction, several methods combining one or more coevolution algorithms with physicochemical information, sequence conservation, secondary structure prediction, molecular modelling, machine learning and deep learning approaches have been proposed [86, 87, 88, 89, 90, 91, 92, 93]. These so called *fusion methods* were reported to be more accurate than coevolution methods described in the previous sections, with metaPSICOV [89] being the most reliable [94].

2.5 Specifics of detecting inter-protein coevolving pairs

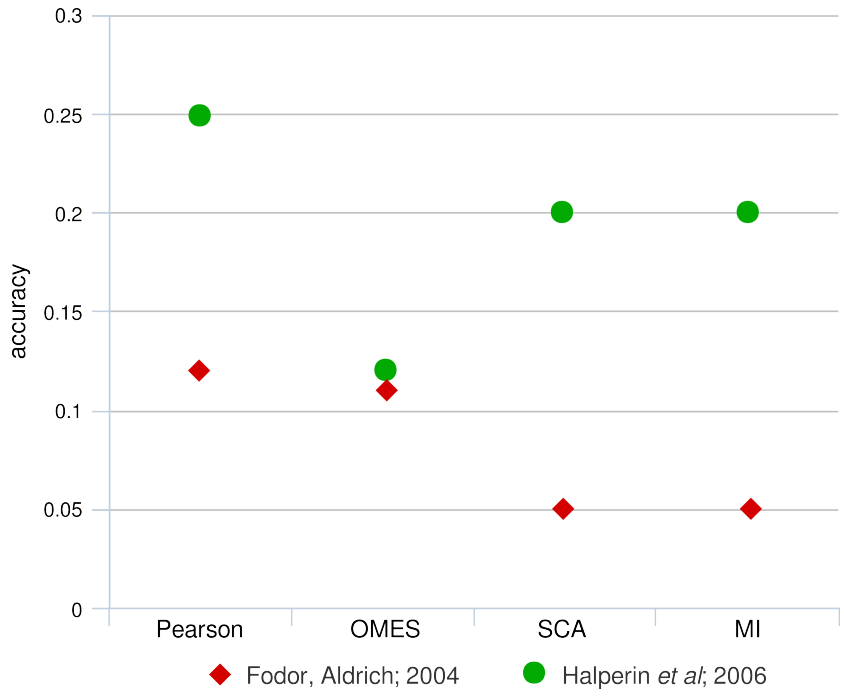
Most of the methods described above concentrate on detecting intra-molecular correlated pairs. However, in some applications (e.g. detecting docking interface between a receptor and its ligand) we need to know inter-protein coevolving residues. Unfortunately, detecting inter-protein coevolution has some specifics.

There are scarcely any methods designed specifically for detecting inter-protein coevolving pairs; an example may be the approach described by Thattai *et al.* [34]. Usually, inter-protein coevolution is detected by concatenating the sequences of the two proteins and using some of the methods described above. The simple covariance-based methods' success in identifying inter-protein coevolving pairs varied [32, 33, 43, 95], however, the newer, direct contact searching methods were shown to be in general successful in this task [35, 81, 96, 97]. Still, there are two issues which have to be overcome:

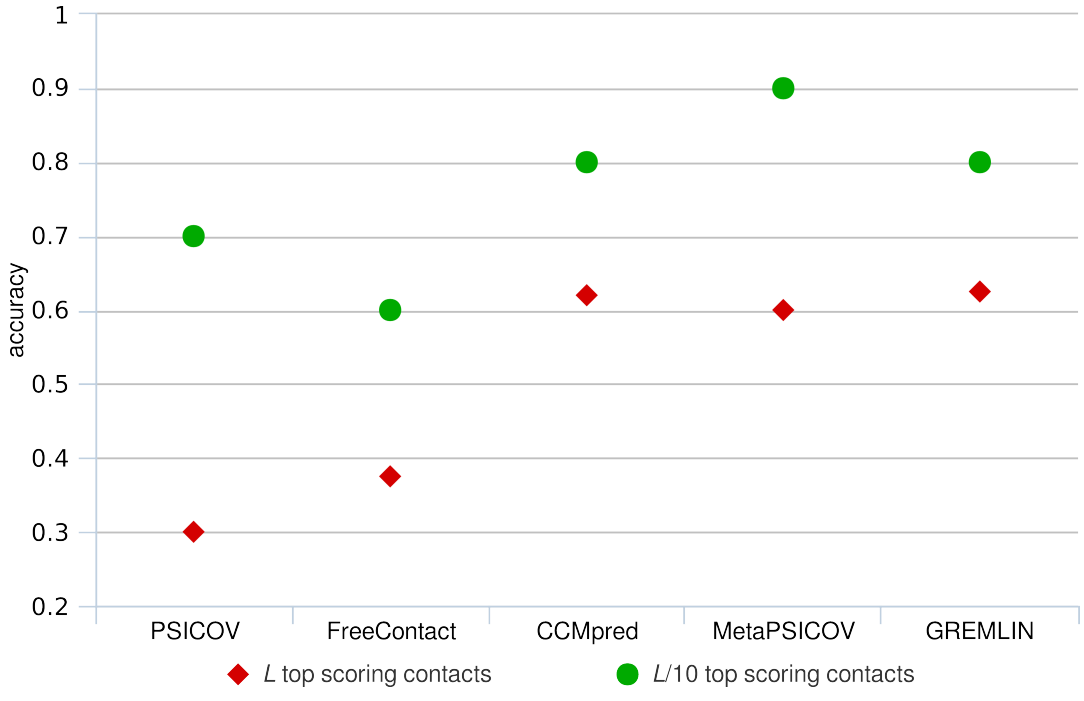
1. **MSA quality**: The concatenated sequence is longer than a sequence of a single protein. To avoid false positives, this must be compensated by larger number of homologous sequences in the data set [35].
2. **Forming orthologous pairs**: If there is only one homolog of both studied genes in each organism, forming interacting pairs properly is easy. However, if more homologs are present, there is a need to infer which two of these homologs are really interacting *in vivo*. The most stringent approach is to exclude all such organisms from the analysis, as we cannot be sure which of the sequences form the actual interacting pairs [98]. This approach, however, may greatly limit the number of analyzed sequences, and thus also reduce the power of the test. In procaryotes, the genome structure may be exploited, as the functionally linked genes are often located in the same operon [35]. Another possibility is to detect interacting pairs by protein-protein coevolution algorithms which are described in the next chapter.

Name	Algorithm	Reference	Website
Residue-residue level			
CCMpred	GREMLIN, plmDCA	2.3.1, [99]	https://github.com/soedinglab/ccmpred
CoMap	mapping substitutions on phyl. tree	2.2.1, [36, 56]	jydu.github.io/comap/
DCA	mfDCA	2.3.1, [81]	dca.rice.edu/portal/dca/
EVfold	mfDCA, plmDCA	2.3.1, [76]	http://evfold.org/evfold-web/evfold.do
FreeContact	mfDCA, PSICOV	2.3.1, 2.3.2 [100]	https://rostlab.org/owiki/index.php/FreeContact
i-COMS	mfDCA, MI, plmDCA, PSICOV	2.2.1, 2.3.1, 2.3.2, [101]	i-coms.leloir.org.ar/
MetaPSICOV	fusion	2.4, [89]	bioinf.cs.ucl.ac.uk/MetaPSICOV/
MISTIC2	MI	2.2.1, [102]	mistic2.leloir.org.ar
PconsC3	plmDCA, PSICOV	2.3.1, 2.3.2, [86, 103]	pconsc3.bioinfo.se/
plmDCA	plmDCA	2.3.1, [78]	plmdca.csc.kth.se/
PSICOV	PSICOV	2.3.2, [82]	bioinfadmin.cs.ucl.ac.uk/downloads/PSICOV/
RaptorX	fusion	2.4, [104]	raptorx.uchicago.edu/
Protein-protein level			
MirrorTree	mirrortree	3.2.2, [105]	csbg.cnb.csic.es/mtserver/
MMM	identifying interacting pairs	3.3, [106]	wwwlabs.uhnresearch.ca/tillier/MMMWEBvII/MMMWEBvII.php
PPIDFT	biochemical distances by Fourier transform	3.2.4, [107]	https://github.com/cyinbox/PPI

Table 2.1: A selection of web servers and downloadable programs for coevolution analysis. All links checked on 24 April 2018.



(a)



(b)

Figure 2.1: Comparison of the accuracy (true positives rate) of different algorithms for coevolution prediction.
(a) Local methods (see section 2.2), data from two different studies: Fodor and Aldrich [41] (analysis of 224 Pfam families) and Halperin *et al.* [33] (analysis of 15 yeast fusion protein families), results for 75 top scoring positions given in both cases.
(b) Global and fusion methods (see sections 2.3 and 2.4), data from a study by de Oliveira *et al.* [94] (analysis of 3458 protein families), average results for top L and top $L/10$ scoring positions, L is the length of the protein.

3. Protein-protein coevolution

3.1 Introduction

It has been shown that phylogenetic trees of interacting or functionally related proteins tend to have similar topologies [7, 108], as well as that interacting proteins exhibit similar evolutionary rates [109]. Both these facts are a result of protein-protein coevolution, as described in sections 1.3 and 2.1. Interestingly, similarity of evolutionary rates was observed also for functionally related, though not directly interacting proteins [8]. As a result, protein-protein coevolution may be seen as an indicator of protein-protein interaction, or at least functional dependency.

There are two main classes of protein-protein coevolution algorithms. One of them tries to find out if two protein families as a whole are or are not coevolving (these algorithms will be discussed in section 3.2). The second one attempts to disentangle which proteins from the first family coevolve with which proteins from the second family (section 3.3).

All of the methods described below may be used also to detect domain-domain coevolution. Selected implementations of the algorithms are listed in table 2.1.

3.2 Coevolution of protein families

3.2.1 Phylogenetic profiling

A simple, yet often powerful method detecting protein-protein coevolution is the so-called phylogenetic profiling (see figure 3.1(a)). A phylogenetic profile is the pattern of presence/absence of given protein in a set of genomes [110, 111]. Two proteins may be coevolving if they have similar phylogenetic profiles. The underlying assumption is that if the two proteins need each other to perform a given function, they necessarily must be present in the same organisms – the situation when a pair of genes is lost or gained together independently several times may be seen as an extreme consequence of coevolution [112, 113].

Originally, phylogenetic profiles were just binary vectors, as described in the previous paragraph [110, 111]. The 0/1 information was later successfully replaced e.g. by Protein BLAST E-values [114, 115, 116] or number of paralogs appearing in the genome [117], the latter being of special use for complex eucaryotic gene families. Similarity of phylogenetic profiles may be evaluated using e.g. number of mismatches [111], co-ocurrence probability [118], Pearson correlation coefficient [119], Euclidean distance [117] or mutual information [114, 120].

Despite its name, classical phylogenetic profiles do not make any use of the phylogenetic information, and thus cannot distinguish correlations stemming from the common ancestry from those that indicate several independent gains/losses of the gene. By including evolutionary information, as done e.g. in [121, 122, 123], the accuracy was significantly increased.

However, although useful in some cases, phylogenetic profiles have many weaknesses. First of all, the method is applicable only to completely sequenced genomes/proteomes (otherwise, we cannot be sure of the absence of the protein).

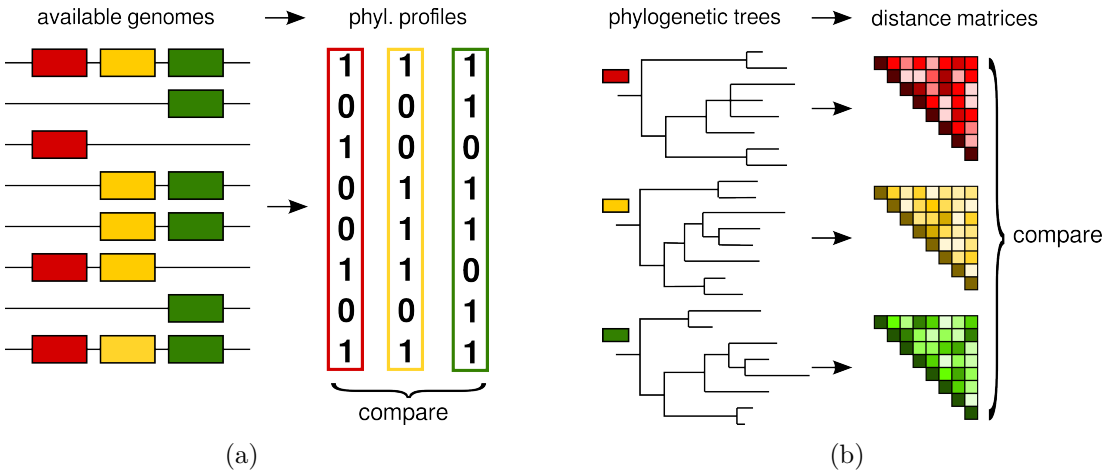


Figure 3.1: Schematic representation of two computational methods for detection of protein-protein coevolution. (a) Phylogenetic profiling: Absence/presence of given gene in a set of genomes is encoded as 0/1 vector, and these vectors are then compared. (b) Mirrortree: For each protein family, distance matrix is computed and compared with distance matrices of other proteins.

Next, it cannot be used with ubiquitous proteins present in almost all organisms, as well as with proteins specific for a given genome.

3.2.2 Tree similarity

As interacting proteins tend to have similar topologies of phylogenetic trees [7, 108], we can detect protein-protein coevolution through the similarity of their phylogenetic trees (see figure 3.1(b)). This approach is usually called *mirrortree*, after one of the first of such methodologies [124].

The first approaches quantified the similarity of two phylogenetic trees as Pearson correlation coefficient between distance matrices of the two protein families [124, 125] – in other words, the test statistics is identical to equation 2.1 with the only difference being that now we are iterating over all pairs of *sequences* instead of residues and we omit the weighting factor w_{pq}. The distance matrices may be constructed either from the pairwise sequence similarities (in this case, there is no need for phylogenetic tree construction) [124, 125], or by conversion from the phylogenetic tree [126]. The pair of protein families is considered to be coevolving if the value of the correlation coefficient is higher than a given threshold (value 0.8 was suggested in [124]).

However, phylogenetic trees of all proteins are likely to look somewhat similar as they basically respect the 'tree of life'. Thus, the original mirrortree approach has high false positives rate, and more complex approaches with corrections for the shared evolution are needed for more reliable prediction. Roughly speaking, we want to subtract the 'phylogenetic distance' of the species from the observed distances of their proteins, and thus exclude the phylogenetic relationships from the analysis. The phylogenetic distances may be obtained either from a canonical tree of life (constructed e.g. from the sequences of 16S rRNA) [126, 127], or inferred from the actual data by averaging them or by identifying the main tendencies by principal component analysis [127, 128]. Such modified algorithms

were reported to produce much less false positives than the original mirrortree method, on the other hand, they also have much lower sensitivity [127, 128].

Another source of noise in the computation is the fact that a given protein often interacts with many others. As a result, the coevolution signal within its tree is composed of the influences of all the interactions. This problem may be solved, similarly as when deciphering direct physical contacts in residue-residue coevolution (see section 2.3), by looking at the whole network of protein-protein pairs, thereby taking into account each protein's coevolutionary context. The method, called ContextMirror [129], first computes the tree similarities for all pairs of proteins and then the specificity of the coevolution between two proteins is evaluated by calculating their partial correlation given all of the other proteins. ContextMirror method was shown to be highly accurate and able to predict interactions with a degree of accuracy and coverage comparable with that of high-throughput experimental techniques [129].

3.2.3 Enrichment of correlated mutations

If we perform the coevolution analysis for individual domains, we can see stronger signal for physically interacting domains than for the non-interacting ones. The same holds if we restrict the analysis to residues belonging to protein interfaces [32, 35, 130]. In other words, protein-protein coevolution is a local phenomenon that can be circumscribed to certain residues, and we can predict protein-protein interactions by searching for inter-protein pairs of correlated residues (reviewed in chapter 2, especially in section 2.5).

If the simpler local techniques (see section 2.2) with high false positives rate are used, one has to compare the distribution of correlated intra- and inter-protein pairs, and find pairs of proteins showing relative abundance of inter-protein correlated pairs. This was done e.g. in so called i2h approach [42]. If more specific global techniques (see section 2.3) are used, two proteins may be considered interacting with quite high probability if they show at least one significantly coevolving inter-protein pair [97].

3.2.4 MSA-free approaches

All of the above described methods need an MSA prior to the analysis and rely heavily on its quality. To overcome this, several MSA-free approaches were suggested.

In the method of Yin and Yau [107], the protein sequences are represented numerically by biochemical properties of individual amino acids, and the distance matrices of the two protein families are computed using discrete Fourier transform. As described previously, distance matrices of coevolving proteins should be strongly correlated.

Yet another approach detects coevolution through the expression levels: It was reported that expression levels of the genes encoding interacting proteins are strongly correlated [131, 132], and that misexpression of protein complex subunits has more severe consequences than misexpression of non-interacting proteins [133]. Thus, pairs of coevolving proteins can be detected also through correlated expression levels [134, 135]. Gene expression level may be estimated (besides

experimentally) e.g. through codon bias [134]. Although correlated expression techniques are not very accurate, they can be used e.g. for the verification of protein interactions [135].

3.3 Coevolution of individual proteins

If we are working with a protein family with a lot of paralogs with different binding specificities, we often want to know the corresponding pairs, i.e. which paralogs within one family interact with those in the other. Several computational approaches to solve this problem have been suggested.

One possibility is to use distance matrices, similarly as in the mirrortree approach (section 3.2.2). The basic assumption is that the correct 'mapping' (set of links) will yield the highest correlation between the two trees. Given two protein families, their distance matrices are aligned to each other such that the root-mean-square difference between corresponding elements is minimal. Interactions are then predicted between the proteins corresponding to aligned columns of the two matrices [136, 137]. However, for distance matrix of size $N \times N$, the exhaustive exploration of all possible mappings would need $N!$ calculations, which is unfeasible for large families. Thus, computationally effective implementations of the above mentioned principle were provided e.g. by [106, 138, 139].

Interacting and non-interacting pairs of proteins can also be distinguished using residue-residue coevolution (see section 3.2.3) [140, 141].

Another, completely different approach using Bayesian statistics was suggested by Burger and van Nimwegen [85]. This approach, similarly as the approaches described in section 3.2.3, supposes that protein-protein coevolution is exhibited through residue-residue coevolution. Using the model described in section 2.3.3 and Markov Chain Monte Carlo simulation, it samples the posterior distribution of $P(a|D)$, where a is the assignment of candidate interacting pairs and D the observed data. Those pairs appearing in assignments with the highest value of $P(a|D)$ are then considered to be interacting.

3.4 Fusion methods

Similarly as for residue-residue coevolution, also protein-protein coevolutionary information has been combined with several other methods of prediction of protein-protein interactions (such as expression level, domains fusion or protein co-localization) in order to obtain more reliable results [142, 143].

4. Applications of coevolution methods

4.1 Residue-residue level

Structure prediction The main motivation for introducing methods detecting protein coevolution at the residue level was the desire to predict protein 3D conformation from its sequence [30, 31]. It was reported that only one contact in twelve allows accurate topology modelling [144], and thus, if we were able to predict residue-residue contacts precisely, we would greatly constrain the conformational space. However, although the residues showing correlated mutational behaviour were shown to be generally closer to each other in the three-dimensional structure than average, the specificity of the oldest methods was too low to be able to predict structure from the scratch [28, 29, 30, 31].

Soon, methods combining coevolutionary information with other methods of protein structure prediction started to emerge [145]. These approaches were able to predict quite accurate structures of small globular proteins [40, 146, 147, 148, 149], however, they still failed on β-strands rich proteins and proteins longer than 110 amino acids [150].

Introduction of the modern approaches into coevolution detection led to dramatic improvement. In CASP12 benchmark (The Critical Assessment of protein Structure Prediction, [151]), significant improvement since CASP11 was observed [152], mainly because most of the participating methods included also coevolution data into the computation. RaptorX [104], a deep learning method predicting contacts by integrating evolutionary coupling and sequence conservation information through an ultra-deep neural network, was ranked the best in the benchmark.

Nowadays, even fully automated pipelines for *ab initio* protein structure prediction based on evolutionary information exist – these include e.g. PconsFold [153] or EVfold [76].

The predicted structures offer plenty of tempting applications: *In silico*, it is possible to model alternative conformations and temporary functional states [43, 96, 154, 155, 156, 157, 158], as well as to predict the potential of forming an ordered structure for apparently disordered proteins [159], both of which is not achievable using classical experimental methods like X-ray crystallography. The approximate computed structures may be used to assist in crystallographic protein structure determination, e.g. to help to solve the phasing problem [160], as well as to validate the experimentally obtained structure [161]. Coevolutionary information has assisted also in identifying domain boundaries [162, 163] and assembling of monomers into homomultimers [164, 165, 166].

All of the above mentioned applications are of special interest for proteins whose 3D structure is challenging to determine by experimental methods, e.g. membrane proteins (reviewed in [167]).

Protein docking The inter-protein predicted contacts may be used for docking prediction [32, 35, 79, 97, 168]. Similarly to the *de novo* structure prediction discussed above, techniques using coevolution proved to be powerful in this task:

In CAPRI (Critical Assessment of PRediction of Interactions, [169]) rounds 28-35, a method integrating coevolutionary inferred links with other approaches was the best one [170], and protein docking approach termed MAGMA (Molecular dynamics And Genomics for Macromolecular Assembly) was shown to predict structures of protein complexes with crystal resolution accuracy, provided that the structures of individual proteins were available [171].

Functionally important residues Another group of applications is derived from the fact that coevolution techniques are able to identify functionally important residues, e.g. residues of the catalytic site, those responsible for protein specificity or allosteric changes [37, 172, 173, 174, 175, 176]. This information may then be used for guided mutagenesis in order to e.g. change enzyme specificity [25, 95], improve the thermostability of the protein [177] or even create synthetic proteins with a given specificity [178, 179]. Based on the covariance between residues, it is even possible to predict effects of mutations [160].

Other applications Coevolutionary information has been used also in structure alignment [180], assessment of protein model quality [181, 182] or to predict the order in which macromolecular complexes assemble [183]. It was suggested that it could be used also to develop 'coevolution-aware' aligners [33] or, as coevolving sites will tend to support the same tree topology, to improve the phylogenetic reconstructions [184].

4.2 Protein-protein level

Protein-protein interactions The major application of protein-protein coevolutionary information is the prediction of protein-protein interactions. Despite being cheaper and faster than the experimental techniques, the computational methods for the prediction of protein interactions have been shown to have similar levels of accuracy [129, 185]. They can be used to predict interactions *de novo* [186], to validate the results of high-throughput experimental techniques [135, 187] or to guide experiments by restricting the number of pairs to be tested experimentally. Computational prediction of protein-protein interactions may be of special use in discovery of non-canonical interactions and crossreactivity [186, 188].

Predicted protein-protein interactions may be further used to decipher the structure of multiple-subunits complexes by predicting which subunits interact with each other [97, 189]. Domain-domain interactions are in turn useful for molecular docking [190].

Functional annotation From the coevolutionary relationships a protein is involved in, we can predict its function [126]. Coevolutionary information was thus used in genome functional annotation [115, 191, 192, 193] or to refine the prediction of function by dividing members of a pathway into subclusters [113], as well as to identify new members of a pathway [194], to discover novel pathways [114, 195] or functionally equivalent, but not homologous proteins [196]. Coevolution has been also used to discover some unexpected functional connections,

e.g. between proteins involved in redox homeostasis and circadian rhythms [197], or to provide statistical support to hypotheses of evolution, such as coevolution between male and female fertilization proteins [198, 199], predicted by the theory of sexual selection.

Other applications Besides the above mentioned, protein-protein coevolution has been also used to predict sub-cellular locations of proteins [200] and could be helpful in identification of horizontal gene transfer and other alternative phylogenetic events [126].

5. Limitations

Although the coevolution prediction methods may provide valuable biological insights under certain circumstances (see chapter 4), there are still many drawbacks that limit their performance.

As most of the nowadays approaches (both on the residue-residue and protein-protein level) rely solely on the MSA, it comes as no surprise that their performance is crucially dependent on its quality [80, 152]. The MSA should contain as many sequences as possible and we should strive to include sequences from diverse organisms, so that the MSA carries enough statistical significance [167]. The MSA should also respect the true course of evolution; using a phylogenetic tree to guide the MSA can both improve its quality and the confidence in the results [46, 166].

The number of sequences needed for reliable residue-residue coevolution prediction was estimated to $5L$, where L is the lenght of the alignment [80]. According to this measure, approximately 25 % of the proteins families on Pfam database [201] would have a sufficient number of sequences for reliable coevolution prediction [80]. However, we can assume that this number will grow quickly thanks to the fast progress in sequencing technologies. Oliveira *et al.* [94] estimated the number of sequences needed for true positives rate greater than 50 % to be ca 1500 for PSICOV [82], ca 1100 for FreeContact [100] and EV-fold [76] and ca 400 for CCMpred [99], metaPSICOV [89] and GREMLIN [80]. To conclude, only large and well sequenced protein families may become a subject of a coevolution analysis.

Besides alignment size, the results of a coevolution analysis greatly depend also on the choice of organisms. This was reported especially for phylogenetic profiles [116, 202, 203] and mirrortree [204] methods. To overcome these problems, methods automatically looking for species were the coevolutionary signal is particularly strong were developed [205, 206].

As discussed in sections 1.3 and 2.1, correlations resembling coevolutionary signal may stem also from other processes: the common evolutionary history of the sequences, heterotachy, etc. When searching for intra-molecular pairs of coevolving residues, it is very difficult to distinguish correlations caused by intra-molecular contact and those caused by monomer-monomer interactions [80, 160]. Generally speaking, coevolution predictions are bad for all-α and membrane proteins, probably mainly due to the low number of available sequences [94].

Last but not least, questions about the correctness of the current hypothesis of coevolution were raised recently. Talavera *et al.* [207] showed that covariation approaches such as MI are unable to differentiate between very different evolutionary scenarios, one including correlated mutations and one not (see figure 5.1), and claim that the signal detected by these methods arise mainly from small number of independent changes at otherwise highly conserved sites (the tendency of covariance-based methods to give high scores to conserved sites was observed also previously e.g. in [56, 208]). This, however, explains why the covariation methods are quite successful in identifying contacting residues, as highly conserved sites tend to be clustered in the protein core. Talavera *et al.* also show by computer simulation that the primary effect of a coevolutionary pressure is

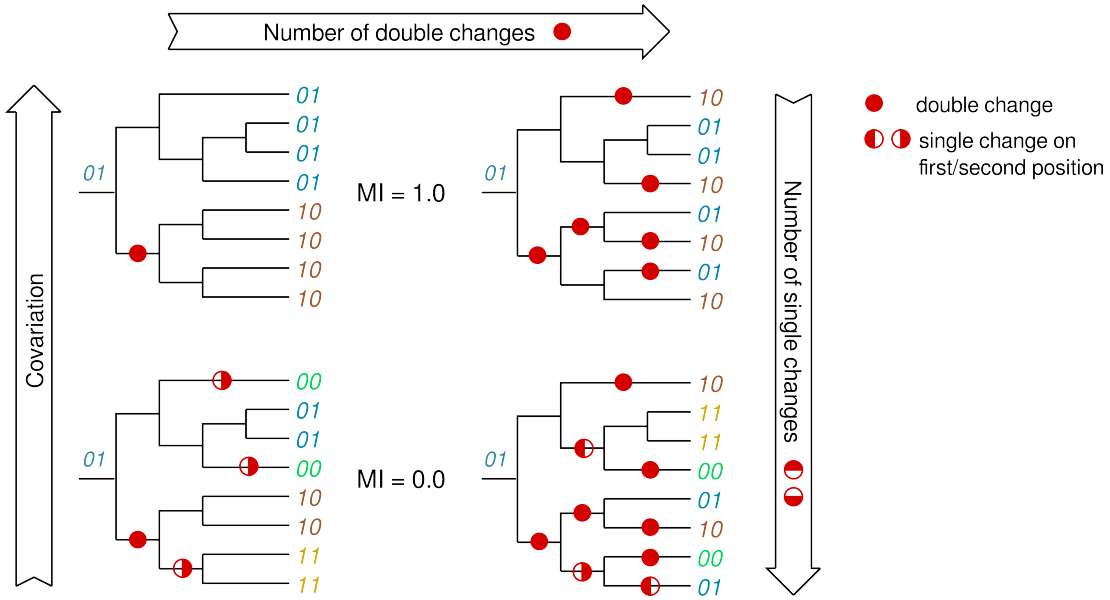


Figure 5.1: Different evolutionary scenarios for a hypothetical molecule with two binary sites. Each site can be in state 0 or 1, observed states in the leaves of the tree shown at the ends of branches. Before the first split, the molecule is in state 01. Mutual information of the two sites is 1.0 for the top two trees (if we know the state of the first site, we know the state of the second site with 100 % accuracy) and 0.0 for the bottom two trees (the two sites are independent). Number of double changes (full circles) grows from left to right, number of single changes (half circles) grows from top to the bottom, covariation grows from bottom to the top. Reproduced after [207].

reduction in substitution rates and formulate the following 'coevolution paradox': The strength of coevolution required to cause coordinated changes means that the evolutionary rate is so low that such changes almost cannot be observed.

Similar observation was made for protein-protein coevolution by Hakes *et al.* [209]. They claim that compensatory mutations are very unlikely to be responsible for the correlated evolution of proteins and that this correlation is caused mainly by the fact that interacting proteins are under similar evolutionary constraints.

Both of the above mentioned papers raise the question whether the hypothesis underlying the current coevolution methods is justifiable. The coevolution approaches surely are successful in many applications, however, the reason why they are successful may be totally different from what we thought.

6. Conclusion

Coevolution, coordinated evolution of two species or biomolecules, is one of the key concepts of the evolutionary theory. In this thesis, we summarized different computational approaches to coevolution detection on protein level. In chapter 2 we described the methods used for the detection of residue-residue coevolution and in chapter 3 the methods used for the detection of protein-protein coevolution. In chapter 4 we mentioned some of the possible applications of these methods. Finally, in chapter 5 we identified some of their weaknesses.

We have seen many diverse algorithms detecting coevolution on both residue-residue and protein-protein level. This shows not only the long-lasting interest of the scientific community in this problem, but also its difficulty. Comparison of the reliability of different approaches and their suitability to different datasets is crucial. The accuracy and limitations of residue-residue coevolution methods have been studied quite extensively. However, we were not able to find a single study comparing the performance of the protein-protein approaches. Similarly, while there are plenty of publicly available programs for residue-residue coevolution, this is not true for the protein-protein level.

Coevolution has been shown to be a powerful tool in solving many biological questions. Nonetheless, since the field is relatively new and dynamic, there still remain many problems waiting to be overcome. The following problems are in our opinion the most important ones:

1. The current programs require MSAs with large number of sequences, which limits the coevolution analysis only to large and well-sequenced protein families. Programs able to reliably detect coevolutionary relationships in smaller datasets would be extremely useful.

2. Most of the nowadays used algorithms do not use the phylogenetic tree of the protein(s) under study at all. This makes it difficult to distinguish the true coevolution from other processes. Development of algorithms employing the evolutionary information could lead to substantial improvement in the field.

3. We believe that it would be useful to take the process of coevolution into account also when solving other problems, such as inferring phylogeny. The present-day evolutionary models consider all the sites to evolve independently, but this does not have to be true.

4. As questions about the true nature of the coevolutionary process were raised recently, the process should be studied carefully to confirm or refute the contemporary hypothesis.

The author has used several of the described methods in a study of fish interferon gamma and its receptors. The paper is now under review.

List of Abbreviations

DCA	direct coupling analysis
MI	mutual information
MSA	multiple sequence alignment
OMES	observed minus expected squared
SCA	statistical coupling analysis

Bibliography

[1] J.N. Thompson. *The Coevolutionary Process*. University of Chicago Press, 2009.

[2] L. Van Valen. A new evolutionary law. *Evolutionary theory*, 1:1–30, 1973.

[3] F. Müller. *Ituna* and *Thyridia*: a remarkable case of mimicry in butterflies. *Proceedings of the Entomological Society of London*, pages 20–29, 1879. Translated by R. Meldola.

[4] J.M. Smith and D. Harper. *Animal signals*, pages 86–87. Oxford series in ecology and evolution. Oxford University Press, 2003.

[5] T. Dobzhansky. Observations and experiments on natural selection in *Drosophila*. *Hereditas*, 35(S1):210–224, 1949.

[6] T. Dobzhansky. Genetics of natural populations. XIX. Origin of heterosis through natural selection in populations of *Drosophila pseudoobscura*. *Genetics*, 35(3):288–302, 1950.

[7] K.J. Fryxell. The coevolution of gene family trees. *Trends in Genetics*, 12(9):364–369, 1996.

[8] Y. Chen and N.V. Dokholyan. The coordinated evolution of yeast proteins is constrained by functional modularity. *Trends in Genetics*, 22(8):416–419, 2006.

[9] W.R. Moyle, R.K. Campbell, R.V. Myers, M.P. Bernard, Y. Han, and X. Wang. Co-evolution of ligand-receptor pairs. *Nature*, 368(6468):251–255, 1994.

[10] S. Atwell, M. Ultsch, A.M. De Vos, and J.A. Wells. Structural plasticity in a remodeled protein-protein interface. *Science*, 278(5340):1125–1128, 1997.

[11] C. Darwin and G. Beer. *On the origin of species*. Revised edition. Oxford University Press, 2008.

[12] C. Darwin. *On the various contrivances by which British and foreign orchids are fertilised by insects: and on the good effects of intercrossing*. John Murray, 1862. Available online at `http://darwin-online.org.uk/content/frameset?pageseq=1&itemID=F800&viewtype=text`. Accessed: 2018-04-24.

[13] L.W.R. Rothschild, K. Jordan, and Zoological Museum in Tring (England). *A revision of the lepidopterous family Sphingidae*, volume Novitates zoologicae. Vol. IX. Supplement, v. 2. Hazell, Watson & Viney, Ld., London, 1903. Available online at `https://www.biodiversitylibrary.org/bibliography/5651`. Accessed: 2018-04-24.

[14] P.R. Ehrlich and P.H. Raven. Butterflies and plants: A study in coevolution. *Evolution*, 18(4):586–608, 1964.

[15] R.E. Dickerson. The structure of cytochrome c and the rates of molecular evolution. *Journal of Molecular Evolution*, 1(1):26–45, 1971.

[16] E. Zuckerkandl. Evolutionary processes and evolutionary noise at the molecular level. *Journal of Molecular Evolution*, 7(4):269–311, 1976.

[17] M. Socolich, S.W. Lockless, W.P. Russ, H. Lee, K.H. Gardner, and R. Ranganathan. Evolutionary information for specifying a protein fold. *Nature*, 437:512–518, 2005.

[18] W.M. Fitch and E. Markowitz. An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. *Biochemical Genetics*, 4(5):579–593, 1970.

[19] C. Chothia and A.M. Lesk. The relation between the divergence of sequence and structure in proteins. *The EMBO Journal*, 5(4):823–826, 1986.

[20] C. Yanofsky, V. Horn, and D. Thorpe. Protein structure relationships revealed by mutational analysis. *Science*, 146:1593–1594, 1964.

[21] K. Oosawa and M. Simon. Analysis of mutations in the transmembrane region of the aspartate chemoreceptor in *Escherichia coli*. *Proceedings of the National Academy of Sciences*, 83(18):6930–6934, 1986.

[22] T. Vernet, D.C. Tessier, H.E. Khouri, and D. Altschuh. Correlation of co-ordinated amino acid changes at the two-domain interface of cysteine proteases with protein stability. *Journal of Molecular Biology*, 224(2):501–509, 1992.

[23] L.M. Gregoret and R.T. Sauer. Additivity of mutant effects assessed by binomial mutagenesis. *Proceedings of the National Academy of Sciences*, 90(9):4246–4250, 1993.

[24] S.M. Larson, A.A. Di Nardo, and A.R. Davidson. Analysis of covariation in an SH3 domain sequence alignment: applications in tertiary contact prediction and the design of compensating hydrophobic core substitutions. *Journal of Molecular Biology*, 303(3):433–446, 2000.

[25] T. Kortemme, L.A. Joachimiak, A.N. Bullock, A.D. Schuler, B.L. Stoddard, and D. Baker. Computational redesign of protein-protein interaction specificity. *Nature Structural & Molecular Biology*, 11:371–379, 2004.

[26] A.M. Lesk and C. Chothia. How different amino acid sequences determine similar protein structures: The structure and evolutionary dynamics of the globins. *Journal of Molecular Biology*, 136(3):225–270, 1980.

[27] A.A. Pakula and R.T. Sauer. Amino acid substitutions that increase the thermal stability of the λ Cro protein. *Proteins: Structure, Function, and Bioinformatics*, 5(3):202–210, 1989.

[28] D. Altschuh, A.M. Lesk, A.C. Bloomer, and A. Klug. Correlation of co-ordinated amino acid substitutions with function in viruses related to tobacco mosaic virus. *Journal of Molecular Biology*, 193(4):693–707, 1987.

[29] D. Altschuh, T. Vernet, P. Berti1, D. Moras, and K. Nagai. Coordinated amino acid changes in homologous protein families. *Protein Engineering*, 2(3):193–199, 1988.

[30] U. Göbel, C. Sander, R. Schneider, and A. Valencia. Correlated mutations and residue contacts in proteins. *Proteins: Structure, Function, and Bioinformatics*, 18(4):309–317, 1994.

[31] I.N. Shindyalov, N.A. Kolchanov, and C. Sander. Can three-dimensional contacts in protein structures be predicted by analysis of correlated mutations? *Protein Engineering*, 7(3):349–358, 1994.

[32] F. Pazos, M. Helmer-Citterich, G. Ausiello, and A. Valencia. Correlated mutations contain information about protein-protein interaction. *Journal of Molecular Biology*, 271(4):511–523, 1997.

[33] I. Halperin, H. Wolfson, and R. Nussinov. Correlated mutations: Advances and limitations. A study on fusion proteins and on the cohesin-dockerin families. *Proteins: Structure, Function, and Bioinformatics*, 63(4):832–845, 2006.

[34] M. Thattai, Y. Burak, and B.I. Shraiman. The origins of specificity in polyketide synthase protein interactions. *PLOS Computational Biology*, 3(9):1–9, 09 2007.

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

[36] J. Dutheil and N. Galtier. Detecting groups of coevolving positions in a molecule: a clustering approach. *BMC Evolutionary Biology*, 7(1):242, 2007.

[37] S.J. Fleishman, O. Yifrach, and N. Ben-Tal. An evolutionarily conserved network of amino acids mediates gating in voltage-dependent potassium channels. *Journal of Molecular Biology*, 340(2):307–318, 2004.

[38] P. Lopez, D. Casane, and H. Philippe. Heterotachy, an important process of protein evolution. *Molecular Biology and Evolution*, 19(1):1–7, 2002.

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

[40] O. Olmea, B. Rost, and A. Valencia. Effective use of sequence correlation and conservation in fold recognition. *Journal of Molecular Biology*, 293(5):1221–1239, 1999.

[41] A.A. Fodor and R.W. Aldrich. Influence of conservation on calculations of amino acid covariance in multiple sequence alignments. *Proteins: Structure, Function, and Bioinformatics*, 56(2):211–221, 2004.

[42] F. Pazos and A. Valencia. In silico two-hybrid system for the selection of physically interacting protein pairs. *Proteins: Structure, Function, and Bioinformatics*, 47(2):219–227, 2002.

[43] I. Kass and A. Horovitz. Mapping pathways of allosteric communication in GroEL by analysis of correlated mutations. *Proteins: Structure, Function, and Bioinformatics*, 48(4):611–617, 2002.

[44] S.W. Lockless and R. Ranganathan. Evolutionary conserved pathways of energetic connectivity in protein families. *Science*, 286(5438):295–299, 1999.

[45] M.E. Hatley, S.W. Lockless, S.K. Gibson, A.G. Gilman, and R. Ranganathan. Allosteric determinants in guanine nucleotide-binding proteins. *Proceedings of the National Academy of Sciences*, 100(24):14445–14450, 2003.

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

[47] A.A. Fodor and R.W. Aldrich. On evolutionary conservation of thermodynamic coupling in proteins. *Journal of Biological Chemistry*, 279(18):19046–19050, 2004.

[48] J.P. Dekker, A. Fodor, R.W. Aldrich, and G. Yellen. A perturbation-based method for calculating explicit likelihood of evolutionary co-variance in multiple sequence alignments. *Bioinformatics*, 20(10):1565–1572, 2004.

[49] B.T.M. Korber, R.M. Farber, D.H. Wolpert, and A.S. Lapedes. Covariation of mutations in the V3 loop of human immunodeficiency virus type 1 envelope protein: An information theoretic analysis. *Proceedings of the National Academy of Sciences*, 90(15):7176–7180, 1993.

[50] N.D. Clarke. Covariation of residues in the homeodomain sequence family. *Protein Science*, 4(11):2269–2278, 1995.

[51] W.R. Atchley, K.R. Wollenberg, W.M. Fitch, W. Terhalle, and A.W. Dress. Correlations among amino acid sites in bHLH protein domains: An information theoretic analysis. *Molecular Biology and Evolution*, 17(1):164–178, 2000.

[52] T.M. Cover and J.A. Thomas. *Elements of Information Theory*, pages 13–55. John Wiley & Sons, Inc., 2005.

[53] D.S. Horner, W. Pirovano, and G. Pesole. Correlated substitution analysis and the prediction of amino acid structural contacts. *Briefings in Bioinformatics*, 9(1):46–56, 2008.

[54] S. Dunn, L. Wahl, and G.B. Gloor. Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *Bioinformatics*, 24:333–340, 2008.

[55] C.M. Buslje, J. Santos, J.M. Delfino, and M. Nielsen. Correction for phylogeny, small number of observations and data redundancy improves the identification of coevolving amino acid pairs using mutual information. *Bioinformatics*, 25(9):1125–1131, 2009.

[56] J. Dutheil, T. Pupko, A. Jean-Marie, and N. Galtier. A model-based approach for detecting coevolving positions in a molecule. *Molecular Biology and Evolution*, 22(9):1919–1928, 2005.

[57] M.W. Dimmic, M.J. Hubisz, C.D. Bustamante, and R. Nielsen. Detecting coevolving amino acid sites using Bayesian mutational mapping. *Bioinformatics*, 21(suppl 1):i126–i135, 2005.

[58] D.D. Pollock, W.R. Taylor, and N.Goldman. Coevolving protein residues: maximum likelihood identification and relationship to structure. *Journal of Molecular Biology*, 287(1):187–198, 1999.

[59] B. Galitsky. Revealing the set of mutually correlated positions for the protein families of immunoglobulin fold. *In silico Biology*, 3:241–264, 2002.

[60] W.R. Taylor and K. Hatrick. Compensating changes in protein multiple sequence alignments. *Protein Engineering, Design and Selection*, 7(3):341–348, 1994.

[61] P. Fariselli and R. Casadio. A neural network based predictor of residue contacts in proteins. *Protein Engineering, Design and Selection*, 12(1):15–21, 1999.

[62] E.R.M. Tillier and T.W.H. Lui. Using multiple interdependency to separate functional from phylogenetic correlations in protein alignments. *Bioinformatics*, 19(6):750–755, 2003.

[63] R. Grantham. Amino acid difference formula to help explain protein evolution. *Science*, 185(4154):862–864, 1974.

[64] D.D. Pollock and W.R. Taylor. Effectiveness of correlation analysis in identifying protein residues undergoing correlated evolution. *Protein Engineering, Design and Selection*, 10(6):647–657, 1997.

[65] A.D. McLachlan. Tests for comparing related amino-acid sequences. Cytochrome *c* and cytochrome c_{551}. *Journal of Molecular Biology*, 61(2):409–424, 1971.

[66] M.O. Dayhoff. *Atlas of protein sequence and structure*, volume 5, pages 345–352. National Biomedical Research Foundation, 1978.

[67] T. Miyata, S. Miyazawa, and T. Yasunaga. Two types of amino acid substitutions in protein evolution. *Journal of Molecular Evolution*, 12(3):219–236, 1979.

[68] W.R. Taylor and D.T. Jones. Deriving an amino acid distance matrix. *Journal of Theoretical Biology*, 164(1):65–83, 1993.

[69] B. Galitsky, I.M. Gelfand, and A.E. Kister. Class-defining characteristics in the mouse heavy chains of variable domains. *Protein Engineering*, 12(11):919–925, 1999.

[70] G. Casari, C. Sander, and A. Valencia. A method to predict functional residues in proteins. *Nature Structural & Molecular Biology*, 2(2):171–178, 1995.

[71] O. Lichtarge, H.R. Bourne, and F.E. Cohen. An evolutionary trace method defines binding surfaces common to protein families. *Journal of Molecular Biology*, 257(2):342–358, 1996.

[72] S. Madabushi, H. Yao, M. Marsh, D.M. Kristensen, A. Philippi, M.E. Sowa, and O. Lichtarge. Structural clusters of evolutionary trace residues are statistically significant and common in proteins. *Journal of Molecular Biology*, 316(1):139–154, 2002.

[73] G.J. Rodriguez, R. Yao, O. Lichtarge, and T.G. Wensel. Evolution-guided discovery and recoding of allosteric pathway specificity determinants in psychoactive bioamine receptors. *Proceedings of the National Academy of Sciences*, 107(17):7787–7792, 2010.

[74] L. Burger and E. van Nimwegen. Disentangling direct from indirect co-evolution of residues in protein alignments. *PLOS Computational Biology*, 6(1):1–18, 2010.

[75] S. Balakrishnan, H. Kamisetty, J.G. Carbonell, S. Lee, and C.J. Langmead. Learning generative models for protein fold families. *Proteins: Structure, Function, and Bioinformatics*, 79(4):1061–1078, 2011.

[76] D.S. Marks, L.J. Colwell, R. Sheridan, T.A. Hopf, A. Pagnani, R. Zecchina, and C. Sander. Protein 3D structure computed from evolutionary sequence variation. *PLOS ONE*, 6(12):1–20, 2011.

[77] A.S. Lapedes, B. Giraud, L. Liu, and G.D. Stormo. *Correlated mutations in models of protein sequences: phylogenetic and structural effects*, volume 33 of *Lecture Notes–Monograph Series*, pages 236–256. Institute of Mathematical Statistics, Hayward, CA, 1999.

[78] M. Ekeberg, C. Lövkvist, Y. Lan, M. Weigt, and E Aurell. Improved contact prediction in proteins: Using pseudolikelihoods to infer Potts models. *Physical Review E*, 87(1):012707, 2013.

[79] M. Weigt, R.A. White, H. Szurmant, J.A. Hoch, and T. Hwa. Identification of direct residue contacts in protein–protein interaction by message passing. *Proceedings of the National Academy of Sciences*, 106(1):67–72, 2009.

[80] H. Kamisetty, S. Ovchinnikov, and D. Baker. Assessing the utility of coevolution-based residue–residue contact predictions in a sequence- and structure-rich era. *Proceedings of the National Academy of Sciences*, 110(39):15674–15679, 2013.

[81] F. Morcos, A. Pagnani, B. Lunt, A. Bertolino, D.S. Marks, C. Sander, R. Zecchina, J.N. Onuchic, T. Hwa, and M. Weigt. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proceedings of the National Academy of Sciences*, 108(49):E1293–E1301, 2011.

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

[83] Partial correlation coefficient. *Encyclopedia of Mathematics*. `http://www.encyclopediaofmath.org/index.php?title=Partial_correlation_coefficient&oldid=14288`. Accessed: 2018-04-24.

[84] P. Bühlmann and S.A. van de Geer. *Statistics for high-dimensional data: Methods, theory and applications*, pages 435–436. Springer series in statistics. Springer, 2011.

[85] L. Burger and E. van Nimwegen. Accurate prediction of protein–protein interactions from sequence alignments using a Bayesian method. *Molecular Systems Biology*, 4(1):165, 2008.

[86] M.J. Skwark, A. Abdel-Rehim, and A. Elofsson. PconsC: combination of direct information methods and alignments improves contact prediction. *Bioinformatics*, 29(14):1815–1816, 2013.

[87] Z. Wang and J. Xu. Predicting protein contact map using evolutionary and physical constraints by integer programming. *Bioinformatics*, 29(13):i266–i273, 2013.

[88] M. Schneider and O. Brock. Combining physicochemical and evolutionary information for protein contact prediction. *PLOS ONE*, 9(10):1–15, 2014.

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

[90] J. Ma, S. Wang, Z. Wang, and J. Xu. Protein contact prediction by integrating joint evolutionary coupling analysis and supervised learning. *Bioinformatics*, 31(21):3506–3513, 2015.

[91] J. Yang, Q. Jin, B. Zhang, and H. Shen. R2C: improving ab initio residue contact map prediction using dynamic fusion strategy and Gaussian noise filter. *Bioinformatics*, 32(16):2435–2443, 2016.

[92] X. Jing, Q. Dong, and R. Lu. RRCRank: a fusion method using rank strategy for residue-residue contact prediction. *BMC Bioinformatics*, 18(1):390, 2017.

[93] S. Wang, S. Sun, Z. Li, R. Zhang, and J. Xu. Accurate de novo prediction of protein contact map by ultra-deep learning model. *PLOS Computational Biology*, 13(1):1–34, 2017.

[94] S.H.P. de Oliveira, J. Shi, and C.M. Deane. Comparing co-evolution methods and their application to template-free protein structure prediction. *Bioinformatics*, 33(3):373–381, 2017.

[95] J.M. Skerker, B.S. Perchuk, A. Siryaporn, E.A. Lubin, O. Ashenberg, M. Goulian, and M.T. Laub. Rewiring the specificity of two-component signal transduction systems. *Cell*, 133(6):1043–1054, 2008.

[96] A.E. Dago, A. Schug, A. Procaccini, J.A. Hoch, M. Weigt, and H. Szurmant. Structural basis of histidine kinase autophosphorylation deduced by integrating genomics, molecular dynamics, and mutagenesis. *Proceedings of the National Academy of Sciences*, 109(26):E1733–E1742, 2012.

[97] T.A. Hopf, C.P.O. Schärfe, J.P.G.L.M. Rodrigues, A.G. Green, O. Kohlbacher, C. Sander, A.M.J.J. Bonvin, and D.S. Marks. Sequence co-evolution gives 3D contacts and structures of protein complexes. *eLife*, 3:e03430, 2014.

[98] S.H. Tan, Z. Zhang, and S.K. Ng. ADVICE: automated detection and validation of interaction by co-evolution. *Nucleic Acids Research*, 32(Web server issue):W69–W72, 2004.

[99] S. Seemayer, M. Gruber, and J. Söding. CCMpred—fast and precise prediction of protein residue–residue contacts from correlated mutations. *Bioinformatics*, 30(21):3128–3130, 2014.

[100] L. Kaján, T.A. Hopf, M. Kalaš, D.S. Marks, and B. Rost. FreeContact: fast

and free software for protein contact prediction from residue co-evolution. *BMC Bioinformatics*, 15(1):85, 2014.

[101] J. Iserte, F.L. Simonetti, D.J. Zea, E. Teppa, and C. Marino-Buslje. I-COMS: Interprotein-COrrelated Mutations Server. *Nucleic Acids Research*, 43(W1):W320–W325, 2015.

[102] F.L. Simonetti, E. Teppa, A. Chernomoretz, M. Nielsen, and C. Marino Buslje. MISTIC: mutual information server to infer coevolution. *Nucleic Acids Research*, 41(W1):W8–W14, 2013.

[103] M.J. Skwark, M. Michel, D. Menendez Hurtado, M. Ekeberg, and A. Elofsson. Accurate contact predictions for thousands of protein families using PconsC3. *bioRxiv*, 2016.

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

[105] D. Ochoa and F. Pazos. Studying the co-evolution of protein families with the Mirrortree web server. *Bioinformatics*, 26(10):1370–1371, 2010.

[106] A. Rodionov, A. Bezginov, J. Rose, and E.R.M. Tillier. A new, fast algorithm for detecting protein coevolution using maximum compatible cliques. *Algorithms for Molecular Biology*, 6(1):17, 2011.

[107] C. Yin and S.S.T. Yau. A coevolution analysis for identifying protein-protein interactions by Fourier transform. *PLOS ONE*, 12(4):1–19, 2017.

[108] S. Pagès, A. Bélaïch, J. Bélaïch, E. Morag, R. Lamed, Y. Shoham, and E.A. Bayer. Species-specificity of the cohesin-dockerin interaction between *Clostridium thermocellum* and *Clostridium cellulolyticum*: Prediction of specificity determinants of the dockerin domain. *Proteins: Structure, Function, and Bioinformatics*, 29(4):517–527, 1997.

[109] H.B. Fraser, A.E. Hirsh, L.M. Steinmetz, C. Scharfe, and M.W. Feldman. Evolutionary rate in the protein interaction network. *Science*, 296(5568):750–752, 2002.

[110] E.M. Marcotte, M. Pellegrini, H.L. Ng, D.W. Rice, T.O. Yeates, and D. Eisenberg. Detecting protein function and protein-protein interactions from genome sequences. *Science*, 285(5428):751–753, 1999.

[111] M. Pellegrini, E.M. Marcotte, M.J. Thompson, D. Eisenberg, and T.O. Yeates. Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. *Proceedings of the National Academy of Sciences*, 96(8):4285–4288, 1999.

[112] M.A. Huynen and P. Bork. Measuring genome evolution. *Proceedings of the National Academy of Sciences*, 95(11):5849–5856, 1998.

[113] T. Ettema, J. van der Oost, and M. Huynen. Modularity in the gain and loss of genes: applications for function prediction. *Trends in Genetics*, 17(9):485–487, 2001.

[114] S.V. Date and E.M. Marcotte. Discovery of uncharacterized cellular systems by genome-wide analysis of functional linkages. *Nature Biotechnology*, 21(9):1055–1062, 2003.

[115] F. Enault, K. Suhre, C. Abergel, O. Poirot, and J.M. Claverie. Annotation

of bacterial genomes using improved phylogenomic profiles. *Bioinformatics*, 19(S1):i105–i107, 2003.

[116] E.S. Snitkin, A.M. Gustafson, J. Mellor, J. Wu, and C. DeLisi. Comparative assessment of performance and genome dependence among phylogenetic profiling methods. *BMC Bioinformatics*, 7(1):420, 2006.

[117] J.A.G. Ranea, C. Yeats, A. Grant, and C.A. Orengo. Predicting protein function with hierarchical phylogenetic profiles: The Gene3D Phylo-Tuner method applied to eukaryotic genomes. *PLOS Computational Biology*, 3(11):1–13, 2007.

[118] J. Wu, S. Kasif, and C. DeLisi. Identification of functional links between genes using phylogenetic profiles. *Bioinformatics*, 19(12):1524–1530, 2003.

[119] G.V. Glazko and A.R. Mushegian. Detection of evolutionarily stable fragments of cellular pathways by hierarchical clustering of phyletic patterns. *Genome Biology*, 5(5):R32, 2004.

[120] M. Huynen, B. Snel, W. Lathe, and P. Bork. Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. *Genome Research*, 10(8):1204–1210, 2000.

[121] D. Barker and M. Pagel. Predicting functional gene links from phylogenetic-statistical analyses of whole genomes. *PLOS Computational Biology*, 1(1):1–8, 2005.

[122] Y. Zhou, R. Wang, L. Li, X. Xia, and Z. Sun. Inferring functional linkages between proteins from evolutionary scenarios. *Journal of Molecular Biology*, 359(4):1150–1159, 2006.

[123] H.X. Ta, P. Koskinen, and L. Holm. A novel method for assigning functional linkages to proteins using enhanced phylogenetic trees. *Bioinformatics*, 27(5):700–706, 2011.

[124] F. Pazos and A. Valencia. Similarity of phylogenetic trees as indicator of protein–protein interaction. *Protein Engineering, Design and Selection*, 14(9):609–614, 2001.

[125] C. Goh, A.A. Bogan, M. Joachimiak, D. Walther, and F.E. Cohen. Co-evolution of proteins with their interaction partners. *Journal of Molecular Biology*, 299(2):283–293, 2000.

[126] F. Pazos, J.A.G. Ranea, D. Juan, and M.J.E. Sternberg. Assessing protein co-evolution in the context of the tree of life assists in the prediction of the interactome. *Journal of Molecular Biology*, 352(4):1002–1015, 2005.

[127] T. Sato, Y. Yamanishi, M. Kanehisa, and H. Toh. The inference of protein–protein interactions by co-evolutionary analysis is improved by excluding the information about the phylogenetic relationships. *Bioinformatics*, 21(17):3482–3489, 2005.

[128] T. Sato, Y. Yamanishi, K. Horimoto, M. Kanehisa, and H. Toh. Partial correlation coefficient between distance matrices as a new indicator of protein–protein interactions. *Bioinformatics*, 22(20):2488–2492, 2006.

[129] D. Juan, F. Pazos, and A. Valencia. High-confidence prediction of global interactomes based on genome-wide coevolutionary networks. *Proceedings of the National Academy of Sciences*, 105(3):934–939, 2008.

[130] M.G. Kann, B.A. Shoemaker, A.R. Panchenko, and T.M. Przytycka. Correlated evolution of interacting proteins: Looking behind the mirrortree. *Journal of Molecular Biology*, 385(1):91–98, 2009.

[131] M.B. Eisen, P.T. Spellman, P.O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences*, 95(25):14863–14868, 1998.

[132] A. Grigoriev. A relationship between gene expression and protein interactions on the proteome scale: analysis of the bacteriophage T7 and the yeast *Saccharomyces cerevisiae*. *Nucleic Acids Research*, 29(17):3513–3519, 2001.

[133] B. Papp, C. Pál, and L.D. Hurst. Dosage sensitivity and the evolution of gene families in yeast. *Nature*, 424(6945):194–197, 2003.

[134] H.B. Fraser, A.E. Hirsh, D.P. Wall, and M.B. Eisen. Coevolution of gene expression among interacting proteins. *Proceedings of the National Academy of Sciences*, 101(24):9033–9038, 2004.

[135] I. Tirosh and N. Barkai. Computational verification of protein-protein interactions by orthologous co-expression. *BMC Bioinformatics*, 6(1):40, 2005.

[136] J. Gertz, G. Elfond, A. Shustrova, M. Weisinger, M. Pellegrini, S. Cokus, and B. Rothschild. Inferring protein interactions from phylogenetic distance matrices. *Bioinformatics*, 19(16):2039–2045, 2003.

[137] A.K. Ramani and E.M. Marcotte. Exploiting the co-evolution of interacting proteins to discover interaction specificity. *Journal of Molecular Biology*, 327(1):273–284, 2003.

[138] R. Jothi, M.G. Kann, and T.M. Przytycka. Predicting protein–protein interaction by searching evolutionary tree automorphism space. *Bioinformatics*, 21(S1):i241–i250, 2005.

[139] J.M.G. Izarzugaza, D. Juan, C. Pons, F. Pazos, and A. Valencia. Enhancing the prediction of protein pairings between interacting families using orthology information. *BMC Bioinformatics*, 9(1):35, 2008.

[140] A.-F. Bitbol, R.S. Dwyer, L.J. Colwell, and N.S. Wingreen. Inferring interaction partners from protein sequences. *Proceedings of the National Academy of Sciences*, 113(43):12180–12185, 2016.

[141] T. Gueudré, C. Baldassi, M. Zamparo, M. Weigt, and A. Pagnani. Simultaneous identification of specifically interacting paralogs and interprotein contacts by direct coupling analysis. *Proceedings of the National Academy of Sciences*, 113(43):12186–12191, 2016.

[142] E.M. Marcotte, M. Pellegrini, M.J. Thompson, T.O. Yeates, and D. Eisenberg. A combined algorithm for genome-wide prediction of protein function. *Nature*, 402(6757):83–86, 1999.

[143] R. Jansen, H. Yu, D. Greenbaum, Y. Kluger, N.J. Krogan, S. Chung, A. Emili, M. Snyder, J.F. Greenblatt, and M. Gerstein. A Bayesian networks approach for predicting protein–protein interactions from genomic data. *Science*, 302(5644):449–453, 2003.

[144] D.E. Kim, F. DiMaio, R. Yu-Ruei Wang, Y. Song, and D. Baker. One contact for every twelve residues allows robust and accurate topology-level protein structure modeling. *Proteins: Structure, Function, and Bioinformatics*, 82(S2):208–218, 2014.

[145] S.A. Benner and D. Gerloff. Patterns of divergence in homologous proteins as indicators of secondary and tertiary structure: A prediction of the structure of the catalytic domain of protein kinases. *Advances in Enzyme Regulation*, 31:121–181, 1991.

[146] O. Olmea and A. Valencia. Improving contact predictions by the combination of correlated mutations and other sources of sequence information. *Folding and Design*, 2(3):S25–S32, 1997.

[147] J. Skolnick, A. Kolinski, and A.R. Ortiz. MONSSTER: a method for folding globular proteins with a small number of distance restraints. *Journal of Molecular Biology*, 265(2):217–241, 1997.

[148] A.R. Ortiz, A. Kolinski, and J. Skolnick. Fold assembly of small proteins using Monte Carlo simulations driven by restraints derived from multiple sequence alignments. *Journal of Molecular Biology*, 277:419–448, 1998.

[149] A.R. Ortiz, A. Kolinski, and J. Skolnick. Nativelike topology assembly of small proteins using predicted restraints in Monte Carlo folding simulations. *Proceedings of the National Academy of Sciences*, 95(3):1020–1025, 1998.

[150] A.R. Ortiz, A. Kolinski, P. Rotkiewicz, B. Ilkowski, and J. Skolnick. Ab initio folding of proteins using restraints derived from evolutionary information. *Proteins*, 37(S3):177–185, 1999.

[151] CASP. `http://predictioncenter.org/`. Accessed: 2018-04-24.

[152] J. Schaarschmidt, B. Monastyrskyy, A. Kryshtafovych, and A.M.J.J. Bonvin. Assessment of contact predictions in CASP12: Co-evolution and deep learning coming of age. *Proteins: Structure, Function, and Bioinformatics*, 86:51–66, 2017.

[153] M. Michel, S. Hayat, M.J. Skwark, C. Sander, D.S. Marks, and A. Elofsson. PconsFold: improved contact predictions improve protein models. *Bioinformatics*, 30(17):i482–i488, 2014.

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

[155] B. Jana, F. Morcos, and J.N. Onuchic. From structure to function: the convergence of structure based models and co-evolutionary information. *Physical Chemistry Chemical Physics*, 16:6496–6507, 2014.

[156] L. Sutto, S. Marsili, A. Valencia, and F.L. Gervasio. From residue coevolution to protein conformational ensembles and functional dynamics. *Proceedings of the National Academy of Sciences*, 112(44):13567–13572, 2015.

[157] P. Sfriso, M. Duran-Frigola, R. Mosca, A. Emperador, P. Aloy, and M. Orozco. Residues coevolution guides the systematic identification of alternative functional conformations in proteins. *Structure*, 24(1):116–126, 2016.

[158] B. Lakhani, K.M. Thayer, M.M. Hingorani, and D.L. Beveridge. Evolutionary covariance combined with molecular dynamics predicts a framework for allostery in the MutS DNA mismatch repair protein. *The Journal of Physical Chemistry B*, 121(9):2049–2061, 2017.

[159] A. Toth-Petroczy, P. Palmedo, J. Ingraham, T.A. Hopf, B. Berger, C. Sander, and D.S. Marks. Structured states of disordered proteins from genomic sequences. *Cell*, 167(1):158–170, 2016.

[160] T.A. Hopf, L.J. Colwell, R. Sheridan, B. Rost, C. Sander, and D.S. Marks. Three-dimensional structures of membrane proteins from genomic sequencing. *Cell*, 149(7):1607–1621, 2012.

[161] B. Zimmerman, B. Kelly, B.J. McMillan, T.C.M. Seegar, R.O. Dror, A.C. Kruse,

and S.C. Blacklow. Crystal structure of a full-length human tetraspanin reveals a cholesterol-binding pocket. *Cell*, 167(4):1041–1051, 2016.

[162] D.J. Rigden. Use of covariance analysis for the prediction of structural domain boundaries from multiple protein sequence alignments. *Protein Engineering*, 15:65–77, 2002.

[163] M.I. Sadowski. Prediction of protein domain boundaries from inverse covariances. *Proteins: Structure, Function, and Bioinformatics*, 81(2):253–260, 2013.

[164] D. Malinverni, S. Marsili, A. Barducci, and P. De Los Rios. Large-scale conformational transitions and dimerization are encoded in the amino-acid sequences of Hsp70 chaperones. *PLOS Computational Biology*, 11(6):1–15, 2015.

[165] R.N. dos Santos, F. Morcos, B. Jana, A.D. Andricopulo, and J.N. Onuchic. Dimeric interactions and complex formation using direct coevolutionary couplings. *Scientific Reports*, 5:13652, 2015.

[166] J.M. Nicoludis, B.E. Vogt, A.G. Green, C.P.I. Schärfe, D.S. Marks, and R. Gaudet. Antiparallel protocadherin homodimers use distinct affinity- and specificity-mediating regions in cadherin repeats 1-4. *eLife*, 5:e18449, 2016.

[167] J.M. Nicoludis and R. Gaudet. Applications of sequence coevolution in membrane protein biochemistry. *Biochimica et Biophysica Acta (BBA) - Biomembranes*, 1860(4):895–908, 2018.

[168] M. Tress, D. de Juan, O. Graña, M.J. Gómez, P. Gómez-Puertas, J.M. González, G. López, and A. Valencia. Scoring docking models with evolutionary information. *Proteins: Structure, Function, and Bioinformatics*, 60(2):275–280, 2005.

[169] CAPRI. `https://www.ebi.ac.uk/msd-srv/capri/`. Accessed: 2018-04-24.

[170] J. Yu, J. Andreani, F. Ochsenbein, and R. Guerois. Lessons from (co-)evolution in the docking of proteins and peptides for CAPRI Rounds 28–35. *Proteins: Structure, Function, and Bioinformatics*, 85(3):378–390, 2017.

[171] A. Schug, M. Weigt, J.A. Hoch, J.N. Onuchic, T. Hwa, and H. Szurmant. Chapter 3 - Computational modeling of phosphotransfer complexes in two-component signaling. In *Methods in Enzymology: Two-Component Signaling Systems, Part C*, volume 471 of *Methods in Enzymology*, pages 43–58. Academic Press, 2010.

[172] C. Marino Buslje, E. Teppa, T. Di Doménico, J.M. Delfino, and M. Nielsen. Networks of high mutual information define the structural proximity of catalytic sites: Implications for catalytic residue identification. *PLOS Computational Biology*, 6(11):1–8, 2010.

[173] K.A. Reynolds, R.N. McLaughlin, and R. Ranganathan. Hot spots for allosteric regulation on protein surfaces. *Cell*, 147(7):1564–1575, 2011.

[174] D. Aguilar, B. Oliva, and C. Marino Buslje. Mapping the mutual information network of enzymatic families in the protein structure to unveil functional features. *PLOS ONE*, 7(7):1–12, 2012.

[175] A. Gulyás-Kovács. Integrated analysis of residue coevolution and protein structure in ABC transporters. *PLOS ONE*, 7(5):1–19, 2012.

[176] L.J. Colwell, M.P. Brenner, and A.W. Murray. Conservation weighting functions enable covariance analyses to detect functionally important amino acids. *PLOS ONE*, 9(11):1–9, 2014.

[177] C. Wang, R. Huang, B. He, and Q. Du. Improving the thermostability of alpha-

amylase by combinatorial coevolving-site saturation mutagenesis. *BMC Bioinformatics*, 13(1):263, 2012.

[178] W.P. Russ, D.M. Lowery, P. Mishra, M.B. Yaffe, and R. Ranganathan. Natural-like function in artificial WW domains. *Nature*, 437(7058):579–583, 2005.

[179] Y. Kliger, O. Levy, A. Oren, H. Ashkenazy, Z. Tiran, A. Novik, A. Rosenberg, A. Amir, A. Wool, A. Toporik, E. Schreiber, D. Eshel, Z. Levine, Y. Cohen, C. Nold-Petry, C.A. Dinarello, and I. Borukhov. Peptides modulating conformational changes in secreted chaperones: From in silico design to preclinical proof of concept. *Proceedings of the National Academy of Sciences*, 106(33):13797–13801, 2009.

[180] S. Wang, J. Ma, J. Peng, and J. Xu. Protein structure alignment beyond spatial proximity. *Scientific Reports*, 3(1), 2013.

[181] C.S. Miller and D. Eisenberg. Using inferred residue contacts to distinguish between correct and incorrect protein models. *Bioinformatics*, 24(14):1575–1582, 2008.

[182] M.L. Tress and A. Valencia. Predicted residue–residue contacts can help the scoring of 3D models. *Proteins: Structure, Function, and Bioinformatics*, 78(8):1980–1991, 2010.

[183] S. Mallik and S. Kundu. Coevolutionary constraints in the sequence-space of macromolecular complexes reflect their self-assembly pathways. *Proteins: Structure, Function, and Bioinformatics*, 85(7):1183–1189, 2017.

[184] N. Galtier. Sampling properties of the bootstrap support in molecular phylogeny: Influence of nonindependence among sites. *Systematic Biology*, 53(1):38–46, 2004.

[185] C. von Mering, R. Krause, B. Snel, M. Cornell, S.G. Oliver, S. Fields, and P. Bork. Comparative assessment of large scale data sets of protein–protein interactions. *Nature*, 417(6887):399–403, 2002.

[186] A. Procaccini, B. Lunt, H. Szurmant, T. Hwa, and M. Weigt. Dissecting the specificity of protein-protein interaction in bacterial two-component signaling: Orphans and crosstalks. *PLOS ONE*, 6(5):1–9, 2011.

[187] L. Salwinski and D. Eisenberg. Computational methods of analysis of protein–protein interactions. *Current Opinion in Structural Biology*, 13(3):377–382, 2003.

[188] C. Goh and F.E. Cohen. Co-evolutionary analysis reveals insights into protein–protein interactions. *Journal of Molecular Biology*, 324(1):177–192, 2002.

[189] M. Gershoni, A. Fuchs, N. Shani, Y. Fridman, M. Corral-Debrinski, A. Aharoni, D. Frishman, and D. Mishmar. Coevolution predicts direct interactions between mtDNA-encoded and nDNA-encoded subunits of oxidative phosphorylation complex I. *Journal of Molecular Biology*, 404(1):158–171, 2010.

[190] R. Jothi, P.F. Cherukuri, A. Tasneem, and T.M. Przytycka. Co-evolutionary analysis of domains in interacting proteins reveals insights into domain–domain interactions mediating protein–protein interactions. *Journal of Molecular Biology*, 362(4):861–875, 2006.

[191] Y. Zheng, R.J. Roberts, and S. Kasif. Genomic functional annotation using co-evolution profiles of gene clusters. *Genome Biology*, 3(11):research0060.1–research0060.9, 2002.

[192] M. Levesque, D. Shasha, W. Kim, M.G. Surette, and P.N. Benfey. Trait-to-

Gene: A computational method for predicting the function of uncharacterized genes. *Current Biology*, 13(2):129–133, 2003.

[193] M. Strong, P. Mallick, M. Pellegrini, M.J. Thompson, and D. Eisenberg. Inference of protein function and protein linkages in *Mycobacterium tuberculosis* based on prokaryotic genome organization: a combined computational approach. *Genome Biology*, 4(9):R59, 2003.

[194] J.S. Reader, D. Metzgar, P. Schimmel, and V. de Crécy-Lagard. Identification of four genes necessary for biosynthesis of the modified nucleoside queuosine. *Journal of Biological Chemistry*, 279(8):6280–6285, 2004.

[195] A. Smit and A. Mushegian. Biosynthesis of isoprenoids via mevalonate in *Archaea*: the lost pathway. *Genome Research*, 10(10):1468–1484, 2000.

[196] E. Morett, J.O. Korbel, E. Rajan, G. Saab-Rincon, L. Olvera, M. Olvera, S. Schmidt, B. Snel, and P. Bork. Systematic discovery of analogous enzymes in thiamin biosynthesis. *Nature Biotechnology*, 21(7):790–795, 2003.

[197] R.S. Edgar, E.W. Green, Y. Zhao, G. van Ooijen, M. Olmedo, X. Qin, Y. Xu, M. Pan, U.K. Valekunja, K.A. Feeney, E.S. Maywood, M.H. Hastings, N.S. Baliga, M. Merrow, A.J. Millar, C.H. Johnson, C.P. Kyriacou, J.S. O'Neill, and A.B. Reddy. Peroxiredoxins are conserved markers of circadian rhythms. *Nature*, 485(7399):459–464, 2012.

[198] M. Watanabe, A. Ito, Y. Takada, C. Ninomiya, T. Kakizaki, Y. Takahata, K. Hatakeyama, K. Hinata, G. Suzuki, T. Takasaki, Y. Satta, H. Shiba, S. Takayama, and A. Isogai. Highly divergent sequences of the pollen self-incompatibility (S) gene in class-I S haplotypes of *brassica campestris* (syn. *rapa*) l. *FEBS Letters*, 473(2):139–144, 2000.

[199] N.L. Clark, J. Gasper, M. Sekino, S.A. Springer, C.F. Aquadro, and W.J. Swanson. Coevolution of interacting fertilization proteins. *PLOS Genetics*, 5(7):1–14, 2009.

[200] E.M. Marcotte, I. Xenarios, A.M. van der Bliek, and D. Eisenberg. Localizing proteins in the cell from their phylogenetic profiles. *Proceedings of the National Academy of Sciences*, 97(22):12115–12120, 2000.

[201] Pfam. `https://pfam.xfam.org/`. Accessed: 2018-04-24.

[202] J. Sun, J. Xu, Z. Liu, Q. Liu, A. Zhao, T. Shi, and Y. Li. Refined phylogenetic profiles method for predicting protein–protein interactions. *Bioinformatics*, 21(16):3409–3415, 2005.

[203] R. Jothi, T.M. Przytycka, and L. Aravind. Discovering functional linkages and uncharacterized cellular pathways using phylogenetic profile comparisons: a comprehensive assessment. *BMC Bioinformatics*, 8(1):173, 2007.

[204] D. Herman, D. Ochoa, D. Juan, D. Lopez, A. Valencia, and F. Pazos. Selection of organisms for the co-evolution-based study of protein interactions. *BMC Bioinformatics*, 12(1):363, 2011.

[205] K. Choi and S.M. Gomez. Comparison of phylogenetic trees through alignment of embedded evolutionary distances. *BMC Bioinformatics*, 10(1):423, 2009.

[206] E.R. Tillier and R.L. Charlebois. The human protein coevolution network. *Genome Research*, 19(10):1861–1871, 2009.

[207] D. Talavera, S.C. Lovell, and S. Whelan. Covariation is a poor measure of molecular coevolution. *Molecular Biology and Evolution*, 32(9):2456–2468, 2015.

[208] L. Oliveira, A.C.M. Paiva, and G. Vriend. Correlated mutation analyses on very large sequence families. *ChemBioChem*, 3(10):1010–1017, 2002.

[209] L. Hakes, S.C. Lovell, S.G. Oliver, and D.L. Robertson. Specificity in protein interactions and its relationship with sequence diversity and coevolution. *Proceedings of the National Academy of Sciences*, 104(19):7999–8004, 2007.