Preprint of:

A.J. Preto, J.G. Almeida, J. Schaarschmidt, L.C. Xue, I.S. Moreira and A.M.J.J. Bonvin. Computational tools for the structural characterization of proteins and their complexes from sequence-evolutionary data. *Encyclopedia of Analytical Chemistry* In press (2018),

Computational tools for the structural characterization of proteins and their complexes from sequence-evolutionary data

Antonio J. Preto[1,#], Jose G. Almeida[1,#], Joerg Schaarschmidt[2], Li C. Xue[2], Irina S. Moreira[1,2], Alexandre M.J.J. Bonvin[2*]

[1] CNC - Center for Neuroscience and Cell Biology; Rua Larga, FMUC, Polo I, 1ºandar, Universidade de Coimbra, 3004-517, Coimbra, Portugal.

[2] Bijvoet Center for Biomolecular Research, Faculty of Science - Chemistry, Utrecht University, Utrecht, 3584CH, the Netherlands

[#] Joint first authors

[*] Corresponding author: a.m.j.j.bonvin@uu.nl

Abstract

Structural characterization of proteins and their complexes is a fundamental part in understanding any biological phenomena. Yet, the experimental determination of the three-dimensional structure of proteins and their complexes remains a challenging undertaking. In order to complement the experimental approaches, computational methods have been developed based on a variety of algorithms and models to fill the gap between the amount of sequences and structures. In this chapter, we review the most common methodological approaches currently used in the field, highlighting *ab initio* structure prediction methods and methods for the prediction and structural modelling of protein-protein interfaces. We particularly focus on the use of evolutionary information to guide the modelling process.

Keywords

Machine-learning; Multiple-sequence alignment; Evolutionary conservation; Protein structure; Protein complexes; Structure modelling.

Introduction

Proteins are the fundamental units of cells. They cover an extraordinary functional diversity, which can be closely associated with their three-dimensional structure, encoded within the protein sequence [1]. Key aspects of proteins' function are their interactions with a large variety of other molecules such as small ligands, peptides, other proteins and nucleic acids. For example, substrate binding is an essential step for catalysis in enzymes, peptide binding is necessary to trigger signalling cascades in the majority of receptors, and protein-protein interactions are fundamental for example in protein activation and deactivation through Post Translational Modifications (PTM) such as phosphorylation, or ubiquitination [2,3]. Furthermore protein-nucleic acid interactions are paramount in maintaining the genome as well as in regulating gene expression [4]. In this Chapter, we focus on proteins and their complexes.

Interactions between proteins are determined by Protein-Protein Interfaces (PPIs) [5,6] . These are constituted of residues with specific chemical-physical properties which induce their coupling [7]. Interactions between interface residues are thus fundamental, both from structural and functional point of views, making those a prime target of study [8]. The highest level of detail that can be obtained for PPIs is the overall three-dimensional (3D) structure of the protein complex, as it reveals which residues are at the interface and what sort of interactions are occurring. Also, important residues such as Hot-Spots (HS), which are key determinant of the interaction [9,10], can be inferred from the structure of the complex [11].

Experimental structure determination is typically performed by one of the three main techniques, namely X-ray crystallography, Nuclear Magnetic Resonance (NMR) and Cryo-Electron Microscopy (Cryo-EM). So far, the majority of structures have been determined by X-ray crystallography. These are deposited in the Protein Data bank (PDB, www.rcsb.org/ or www.pdbe.org/) [12], the central repository for experimentally determined biomolecular structures. NMR has contributed about 10% of the experimental structures. Next to determining 3D structure, one main advantage of NMR is its ability to characterise the dynamics of biomolecules [13]. As such it is the method of choice to characterise intrinsically or partially disordered proteins. For large systems, however, spectral crowding becomes a limitation [14,15] even though sophisticated data analysis methods [16] have been developed to tackle this problem. Cryo-EM is characterized by imaging radiation-sensitive entities – cells, viruses and macromolecules – under cryogenic conditions using a transmission electron microscope [17]. While Cryo-EM does not require crystallisation, it used to have until recently relatively low resolution for proteins when compared to X-ray crystallography. This has been changing rapidly

in the last years thanks to breakthroughs in both software and hardware, in particular with the advent of direct electron detectors. The best resolution obtained so far for a Cryo-EM structure is 1.8 Å for glutamate dehydrogenase [18]. Cryo-EM is of particular interest when studying membrane proteins (MPs) as these are challenging to determine by X-ray crystallography. Recent Cryo-EM methods have been able to determine the structures of MPs with resolutions slightly above 3 Å. Examples are the transient receptor potential channel 1 at 3.4 Å [19] or the chloride conducting ion channel at 3.7 Å [20]. X-ray crystallography has, however, achieved resolutions as low as 1.4 Å for membrane proteins [21]. Cryo-electron tomography, a type of Cryo-EM in which the sample is rotated and tilted to provide a more comprehensive overview of larger structures, has given valuable contributions for the understanding of cellular organization and substructures. This technique allows the study of large macromolecules and cellular components under conditions much more similar to the native state when compared with previous methods [22]. The EMDataBank [23] – publicly available at http://emdatabank.org/index.html – collects the protein structures solved by Cryo-EM.

While structural elucidation of proteins is the first approach to understand how they interact with different systems, experimental determination of their structure, and especially of their complexes, can be challenging [24]. Next to the classical methods described above, high-throughput methods providing limited structural information on the complex of interest, such as mass spectrometry, have emerged [25]. Computational methods have also continuously evolved over the years, becoming a particularly advanced field. Both the variety of approaches as well as their efficiency are now reaching levels that allow to not only complement but also present a viable alternative to experimental methods [26-40]. Computational structural modelling methods rely on various approaches, such as homology modelling, Molecular Dynamics (MD) [41] , or Machine-Learning (ML) algorithms [42,43]. Integrative modelling methods that incorporate a variety of sparse data from both experimental and bioinformatics methods have also emerged in recent years [44-46]. These integrative approaches are key in characterizing large and complex molecular machineries for which no single experimental technique can provide all the necessary information. Most methods, and in particular ML ones, have strongly benefited from the explosion in genomic sequence information, which, theoretically, should provide enough information to determine several aspects of protein structure and function. This has become particularly relevant with the determination of evolutionary aspects from sequence information, which allows researchers to identify important residues by their conservation profile [47]: Residues or motives relevant for a protein's function are more likely to be conserved. This

information has been used to analyse intra- [48] and inter-protein couplings [49], predict protein structure [50] and drive the docking processes [51].

In the following sections, we: 1. Discuss the approaches used to assess evolutionary information in proteins, 2. provide a short overview of methods to predict protein 3D structure from sequence data, 3. provide a brief overview of computational methods for the characterization of soluble complexes and 4. of membrane protein-related complexes. Note that we do not claim to provide an exhaustive overview of all methods in the field but have selected some relevant ones to discuss in the context of this review

1. Approaches to assess evolutionary information of protein sequences

Structural prediction approaches benefit from both structural and sequence-related information. The latter has drawn great attention from a computational and biological perspective since it allows to determine how evolution drives protein structure and function from sequence. Two key concepts are the basis for understanding how protein sequence can determine structure and function of proteins: *Multiple Sequence Alignments* (MSA) and *Position-Specific Scoring Matrices* (PSSMs). In short, MSAs match protein sequences from an evolutionary point of view, comparing and aligning them based on their similarity [52]. PSSMs make use of MSA and provides an easy way of determining how likely an amino acid is to be represented at a given position. To do so, three different classifications for each aligned residue pair can be used – match (when the sequences share identical or a similar residues), mismatch (when the residues are different) and gap (when there is no corresponding residue). Given their central role, it is important to consider the general aspects underlying both methodologies.

MSAs started off as techniques to perform global alignments [53] by matching sequences using their full length. Problems arose since sequences might share homology only on some regions and, even if there are several highly homologous regions, these can be shuffled, distant or repeated [54,55]. To address this problem, local alignment techniques that do not require the full length of the protein or nucleic acid sequence were developed [56]. These focus on finding only the common sub-sequences across different systems. Methods capable of finding sub-sequences with common residue pairs [57] or using only “exclusive” residue pairs [58] – residue pairs which are not present in any other sub-sequence determined by the method – were developed. Even though the theory underlying local alignment makes them seem better algorithms to find common sub-sequences in different protein sequences, these methods often encounter problems in dealing with highly gapped common sub-sequences [59].

Simossis *et al*. [60] consider three essential steps when performing MSAs: i) selecting sequences (building a database of sequences to be aligned and compared), ii) selecting an adequate scoring function that allows the comparison of sequences or sub-sequences and iii) iteratively applying this scoring function to build and optimize the alignment. When comparing already known proteins, sequence selection is typically not needed, unless some sequences are detrimental for the result of the final MSA due to clear amino-acid differences. However, when using MSAs for example to calculate PSSMs, databases including thousands or millions [61] of sequences can be used to search for the right protein sequences. This search is usually done considering homology and by using methods such as the Basic Local Alignment Search Tool (BLAST) [62]. Selecting the appropriate scoring function is key in constructing an optimal MSA. These typically work column-wise (analysing each column of aligned residues at a time) and are usually the summation of all pair-wise scores. Several scoring functions are available to consequentially evaluate the MSA through iterations. From MSAs, evolution and conservation scores can be derived. They represent how evolution drove the conservation or absence of a residue at a position in the MSA. Several methods can be used and, for demonstration purposes, the Rate4Site algorithm for functional conservation calculation is briefly described in

Figure 1. Rate4Site pipeline. Starting with an MSA, Rate4Site calculates phylogenetic trees using the Neighbour-Joining (NJ) algorithm 67. This algorithm sequentially joins sequences that are closer and therefore more similar, with each other by creating a new node in each iteration. This node always connects to the tree constructed up to that point in the algorithm. Considering that a single position in the MSA has the same evolutionary rate (score), Rate4Site determines the maximum conservation rate that would explain the conditional probability of the data given that rate.

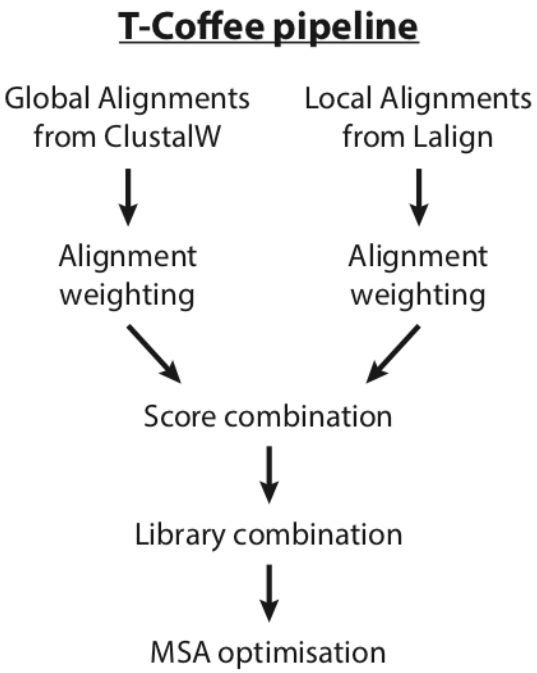

Figure to illustrate how computational calculation of residue conservation can be performed. The Rate4Site algorithm uses the maximum likelihood principle to determine the maximum rate of evolution that would explain the presence of a residue at a position of the MSA.

<Figure 1 near here>

Scoring matrices are used to assess the global score of the MSA, which either calculate the likelihood of a residue to be present at a given position based on the specific information of the aligned sequences (most reliable option) or use pre-calculated likelihoods (scoring matrices) to assign a score to each position. Possibly, the best-known pre-calculated scoring matrices are substitution matrices based on the observed substitution frequencies in sequence alignments. For example, mutations resulting in substitution between residues with identical nature – hydrophobic-hydrophobic mutations (leucine to isoleucine, for example) – can be considered more likely to occur than those that do not keep the nature of the amino-acid at that position. One of the earliest pre-calculated scoring matrices is the *Point Accepted Mutation* (PAM) matrix [63], which generates each residue pair score considering the probability of one residue mutating to a different one considering all possible paths resulting in that specific substitution (e.g, alanine can mutate directly to arginine or it can first mutate to isoleucine and subsequently to arginine). It is based on the concept that each mutation acts as a Markov process – a process independent of any other previous processes. Several PAM matrices were created on this basis, such as PAM250, which considers 250 mutations for 100 residues in the sequence, or PAM1

which considers only 1 mutation for the same sequence length. *BLOcks of amino acid Substitution Matrix* (BLOSUM) [64] is another pre-calculated substitution matrix, used in BLAST. Instead of global alignments, local alignments are used. Several sequence databases with different homology percentages are available, generating different BLOSUMs, such as BLOSUM80, with a database composed of sequences with 80% homology, and BLOSUM52, with a database composed of sequences with 52% homology. Selecting the proper substitution matrix for a given problem is important: While PAMs consider the evolutionary process of similar proteins' aligned sequences, BLOSUMs do not consider an explicit evolutionary model and are used only in very similar regions of the alignment [65]. For example, when scoring similar sequences, BLOSUMs with higher homology or PAMs with fewer mutations per 100 mutations should be used. When scoring sequences that share little homology, BLOSUMs with little homology or PAMs with several mutations per 100 mutations should be used. To calculate likelihoods from the MSA itself and combine them with the information from the pre-calculated scoring matrices, the frequency or count of a given residue in a column can also be calculated and combined with the BLOSUM or PAM scores. The combination of both calculated and pre-calculated counts or likelihoods is referred to as pseudocounts or pseudolikelihoods, respectively. This enables the combination of context information (from the MSA itself) and previously obtained knowledge. Additionally, it prevents scores from being 0 when the MSA-derived counts and likelihoods are 0 (for example, if at a given position in an MSA no leucine residues are observed, its count and likelihood is 0, but its pseudocount and -likelihood is never 0). This is important as it would be extremely unlikely for a residue to never be represented at a given position of an MSA if all sequences were considered.

Hidden Markov Models (HMM) are increasingly popular algorithms in bioinformatics that can also be used to derive MSA profiles. They offer great advantage as theoretically, they can work with both aligned and unaligned data and provide a solid statistical basis to sequence alignment. To generate profiles, HMMs are trained with a set of sequences to determine how likely a transition from a residue to another residue or a gap – a position with no matched residue – is in an MSA, considering its current and next states. The available states are *deletion* – a position is skipped in the MSA for a single/minority of available sequences – *insert* – a position is skipped in the MSA for most available sequences – and *match* – all sequences have a residue in that position. By deriving these probabilities for each position and for each possible amino acid at that position, a HMM can be built to score a sequence and build its profile. Furthermore, if the HMM is good enough, it can also be used to actually build the alignment for new sequences, considering the transition probabilities for each state [66]. After scoring all residues, techniques

such as the sum-of-pair score [64], which sums all residue pairs scores, can be used to obtain a global MSA score that effectively describes the fitness of a MSA. Upon selecting or constructing an appropriate scoring function, the MSA algorithm will then iteratively construct and improve the MSA.

Phylogenetic trees can be used to define the order according to which the sequences are aligned by placing more homologous sequences closer to each other and sequentially aligning each sequence according to the previous [60]. This leads to the generation of a local alignment in which the sub-sequences are taken as the starting points in the global alignment. Upon construction of this preliminary MSA, other techniques can be used, most of which function iteratively. An illustration of the generation of MSAs with the *Tree-based Consistency Objective Function for alignment Evaluation* (T-Coffee) algorithm is shown in Figure 2.

<Figure 2 near here>

Once an MSA has been obtained, a convenient and comprehensive way to represent it is to create so called "profiles". Numerically, this can be done using PSSMs that construct vectors with 20 elements corresponding to each single amino acid at a specific position of the MSA. Platforms such as Consurf [50] use colour to represent the conservation of a residue at a specific position as illustrated in Figure 3. Although one can argue that there is some correlation between conservation and PSSMs, PSSMs are not the same as conservation. While a conservation score – such as the one provided by Rate4Site [67] – is a single value for each position in the provided sequence, a PSSM holds 20 different values describing all possible residues at a specific position.

<Figure 3 near here>

An important concept when dealing with the evolutionary information of protein sequences is co-evolution. From a structural point of view, it can be perceived as the coupled evolution of residue pairs in a protein sequence and it depends on homology, and therefore on sequence similarity, between the assessed protein sequences. Protein sequence similarity must be addressed carefully, since, for example, proteins that evolved in different species can converge on a similar amino acid sequence (orthologs), while other proteins can evolve from a same biological background and yet assume different functions (paralogs). These two particular cases can become an hindrance in co-evolutionary information assessment [68,69].

Some of the common sequence-based co-evolution approaches currently available are:

- **McLachlan-Based Substitution Correlation** (McBASC) uses a similarity matrix to predict protein contacts (amino acid pairs within a given a distance threshold) conservation by analysing correlated mutations [70];
- **Mutual Information** (MI) measures dependency and covariance between variables by calculating the ratio between their joint occurrence probability and their independent occurrence probability [71];
- **Chi-square** describes coupling probabilities similar to MI. However, instead of using a logarithmical ratio, it uses the mathematical square [72];
- **Pearson correlation**, considers the positions that are not gaps at the sequences and the similarity scores (originated from a PSSM matrix) for the possible positions as well as the standard deviation at those positions, in order to calculate amino acid pairs' associated co-evolutionary scores [72];
- **Joint Shannon Entropy** is similar to MI and Chi-square and is used to define conservation [72];
- **Direct-Coupling Analysis** (DCA) calculates the frequencies of residue couplings by assessing the amount of times a pair of residues is present in each alignment at specific locations and by calculating the frequency of the individual residues in each location. From this information, a covariance matrix is calculated, reporting on residue coupling conservation [73];
- **Mean-field Direct-Coupling Analysis** (mfDCA) uses an approach similar to the one used in DCA by combining it with the maximum entropy principle in order to minimize the biasing of the model [73];
- **Protein Sparse Inverse COVariance** (PSICOV) starts by building a covariance matrix in which directly coupled sites are inferred according to an MSA alignment and from which covariance scores are calculated[71,74].

2. Prediction of protein structure from sequence data

Compared to determining the 3D structure of a protein of interest, determining its sequence is rather straightforward. This is reflected by the availability of roughly 90.000.000 protein sequences in the UniProtKB database [75] compared to approximately 75.000 distinct experimental protein structures in the PDB [12]. As already stated, given the challenges and costs associated with the experimental structure determination of proteins, predicting their putative structure from sequence can be a viable alternative. In this section, we briefly introduce computational methods and strategies to predict protein structure from sequence and how

evolutionary information can contribute to this. Table 1 lists all mentioned structure prediction methods as well as a short description for each.

A. General approaches

Starting from the sequence of a protein, the first step is to determine whether a structure or partial structure for this sequence has already been experimentally determined and deposited in the PDB. If no structure for the protein of interest exists, a search against the sequences of all structures deposited in the PDB is typically performed to identify homologues. Commonly used methods to search for these are BLAST [76] to identify close homologues or HHpred [77] for more remote ones. Based on the assumption that the fold of a protein is typically conserved between proteins sharing more than 20% sequence identity, the homologous structures can be used to model the structure of the protein of interest in a process referred to as homology modelling or *Template Based Modelling* (TBM). Note that there are examples in which a single mutation is capable of completely changing the fold of a protein[78], and homology modelling will thus fail. But these are fortunately rare exceptions. A multitude of methods like MODELLER [79] or Rosetta [80] exist that use the 3D coordinates of the template structures with various strategies on how to process these and how missing segments in the alignment are handled [79]. The major challenges in this process are the identification of a suitable template and the generation of a correct alignment. This can be alleviated by running several rounds of TBM with several templates and selecting models from this pool. Other approaches include hybridization strategies where fragments from multiple templates are combined in the modelling protocol [81] or the use of restraints derived from template structures to guide de novo modelling approaches [82].

In cases where no suitable template can be identified, structure prediction methods referred to as *de novo* or *ab initio* structure prediction methods present means of generating structural models for the target protein. Yet, these are generally less reliable especially for larger proteins above 100 amino acids in length [83]. Strategies for the *de novo* structure prediction include in silico protein folding using MD approaches [84], fold recognition based on secondary structure prediction [85], and fragment assembly using a Monte Carlo search algorithm [86,87] [88]. The MD approach in particular is computationally expensive and requires simplifications and adaptations to be not only applicable to very small proteins. In contrast, the fold recognition approach is less demanding but requires a high-quality prediction of secondary structure and a resolved non-homologous structure with a similar fold. The fragment-based approaches such as implemented by Rosetta [89,90] or iTasser [91] have proven quite successful for small proteins. However, given the large conformational space that needs to be sampled in *de novo* structure prediction even for

small proteins, these methods benefit from additional information to reduce the sampling space. These can be computational predictions like protein secondary structure based on position-specific scoring (e.g. Position-Specific Iterative-blast-based secondary structure PREDiction – PSIPRED [92] – or BCL:Jufo9d [93], a tool capable of predicting both secondary and transmembrane spans, which are highly relevant for membrane protein structure prediction) or predicted surface accessibility (e.g. SOLVPRED [94], SANN [95], or ACCpro [96]). Of particular value is however the incorporation of experimental data like chemical crosslinks from mass spectrometry, NOE distances from NMR, or correlated mutations. In the following section we will focus on incorporation of predicted data from bioinformatics analysis.

B. Use of evolutionary information in structure prediction

Extracting information from evolution has also become a very popular approach to solve some structure prediction and characterization problems. An example is Consurf, which provides a way of evaluating the quality of a homology model by calculating evolutionary information and mapping this information onto the 3D structure. First Consurf performs a BLAST search to find homologous sequences to that of the input query. Then, a clustering method is used to search for redundant sequences, which are in turn excluded from the subsequent alignment step. The resulting sequences are aligned using an MSA method [50] . The alignment is used to construct a phylogenetic tree used to assess the velocity at which a residue at each specific position evolves by assigning a rank between 1 and 9. Residues with rank 5 evolve at an intermediate rate, those with rank 1 are the residues the fastest evolving and residues with rank 9 are the most conserved. The method used to attribute these scores involves a Bayesian algorithm that assesses residue evolution rates along the tree [50].

EVFold [48,97] is another platform using evolutionary information for the prediction of protein structure. It uses predicted residue-residue contacts based on co-evolution analysis as restraints to calculate 3D models. Providing a protein sequence, the server fetches sequences of similar proteins and performs an MSA on these. The MSA is performed using HMMs allowing the construction of alignments of large protein sets. A mean-field direct-coupling analysis (see section 1 above) is performed to identify evolutionary couplings, pairs of conserved residues on the protein that consistently interact in the assessed homologues. The output comprises the number of Evolutionary Couplings (ECs), the overall evolutionary strength (cumulative strength), EC strength and residue conservation. Both EVFold and Consurf rely heavily on MSA [62,98-100]. BCL::Fold, a method similar to EVFold has been developed mainly for membrane proteins. It combines evolutionary conservation with other features such as standard values for residue

volume, hydrophobicity, isoelectric points and PSSMs, among others. These are served to ML algorithms to predict the membrane protein structure [101].

Especially in the last CASP experiment the combination of co-evolutionary coupling with ML algorithms dramatically improved the performance of residue-residue contact prediction algorithms [102]. Precisions reaching over 90% were obtained for a substantial subset of targets for predefined numbers of predicted contacts (one fifth or half the target sequence length). Such predictions are a suitable source of information to drive *de novo* structure prediction and significantly increase the quality of the resulting 3D models as for example demonstrated by RaptorX [103] and illustrated in Figure 4. The most successful *ab initio* structure prediction methods combine such contact predictions with secondary structure predictions to fold the proteins.

<Figure 4 near here>

3. Structure prediction of soluble protein-protein complexes

A large variety of computational methods were developed to study interface-related features in protein complexes and predict their 3D structure, which is a prerequisite to understand how proteins interact and what drives the formation of their complexes. Here, we highlight the use of evolutionary information and focus on methods for: a) Interface prediction, b) prediction of interface-related properties, and c) template-based modelling (TBM) of complexes. We do not review the docking field as it has been addressed in several other instances [104-108]. Table 2 summarizes all mentioned protein-protein complex structure prediction and characterization methods, as well as a short description for each.

A. Interface Prediction

An aspect of interface prediction that is widely regarded as critical for the identification of interfacial residues is evolutionary conservation – interfacial residues are typically more conserved than non-interfacial surface residues [109]. In section 2 we have already discussed evolutionary conservation and how it can be used to predict 3D structure [47] . The same concepts are also relevant for the study of interfaces.

A method which combines MSAs with interface prediction is EVComplex [49]. Using sequences from two interacting proteins, EVComplex uses EVCouplings [74] to predict intramolecular and intermolecular residue-residue contacts. In the original paper [49], those were subsequently used to predict the structure of the protein complex by distance-restrained docking using HADDOCK[127,128]. A key aspect of EVComplex is that it considers co-occurring mutations as

accurate predictors of contacts at the interface, unlike EVFold, which considers intra-monomer interactions and mutations to predict intra-monomer couplings. *Protein-Protein Interaction Prediction Platform* (PPIPP) [110] uses propensity scores based on the presence of a given residue compared to any other residue at the interface. To solve the lack of partner information, the model was trained by comparing residues in intermolecular protein-protein interfaces with intra-protein contacts. PPIPP is built on 24 *Artificial Neural Networks* (ANN) and returns the average score as final score, using PSSMs as one of its main features.

PS-HomPPI and NPS-HomPPI [111] use only sequence to predict interfacial residues based on homologous interacting proteins for which experimentally determined 3D structures exist. This method can be partner-specific (PS-HomPPI) or non-partner-specific (NPS-HomPPI), but data suggests that partner specificity increases the accuracy of the results. In fact, an important note made by Xue *et al.* in their 2015 review [112] is, that partner information is very valuable for protein interface prediction, which is often overlooked. A comparison of the results obtained through PPIPP [110] and PAIRpred [113] with the ones from PSIVER [114] (sequence-based) and SPPIDER [115] (structure-based) proved that partner information greatly improves the predictions made. Some methods use structural information from the monomer in their predictions. For example, SPPIDER [115] is a ML approach that predicts interfacial residues based on the predicted relative solvent accessibility (which uses the unbound monomer solvent accessibility and other structural features). *WHat Information Does Surface Conservation Yield?* (WHISCY) [116] uses structure to define surface residues and to smooth the prediction, and calculates conservation for all surface residues. ProMate [117] is a structure-based method, which uses several features such as secondary structure, length of non-secondary structure protein regions and pairwise amino acid residues distribution to calculate an interface propensity value for each residue. Part of the development of ProMate involved the elimination of redundant or highly correlated features, which reduces computation and search space. PAIRpred [113] is a hybrid approach, using both sequence and structure-based features: the structure-based features consist of relative *Solvent Accessible Surface Area* (SASA), residue depth, half sphere amino acid composition and a protrusion index, while the sequence-based features are based on PSSMs and predicted relative accessible surface area. All these are combined through a Support Vector Machine (SVM) to predict protein-protein interactions.

Another approach is that of meta-predictors, which combine multiple individual predictors to generate a consensus prediction [118,119]. One example is *Consensus Prediction Of interface Residues in Transient complexes* (CPORT) [119] , which combines the predictions of several different algorithms using a consensus strategy.

InterEvScore [120] is a tool used to predict protein-protein interactions with the use of multi-body interactions and co-evolutionary information. This software attributes scores based on 2 and 3-body potentials (a quantification of how a residue interacts with one or two other residues, respectively), which are essentially co-evolution weighted residue-residue contact propensities.

Evolutionary features, regardless of their popularity, have a considerable disadvantage – they are quite successful only when a high number of homologs is available [121,122]. As such, methods that do depend less on evolutionary conservation are bound to be more robust across all sorts of protein sequences. This includes the method developed by Wang *et al.* [122] for intramolecular contact prediction, which processes sequences using an ultra-deep Artificial Neural Networks (ANN) and Convolutional Neural Networks (CNN) to improve the prediction of protein contacts obtained through evolutionary conservation, and NeBcon, a meta-server which uses both co-evolution and ML methods to produce better classifications [121].

The Server for Efficient Mapping Assessment (TSEMA) is based on protein sequence, but instead of predicting which residues interact for a single protein, it uses two protein families as input. It then uses this information to predict which proteins interact between each protein family. While not providing residue information, TSEMA is capable of providing partner information, which has proven to be quite relevant as will be discussed ahead in this section.

B. Characterization of Interface-related properties

There are plenty of interface-related properties that can be used to describe the interface of a complex, such as H-bonds, salt bridges, hydrophobic interactions, SASA, number of nearby atoms, total number of interface atoms, polar and apolar area in the interface, hot spots and hot-regions, among many others. Several interesting aspects of protein interface characterization and prediction are discussed in a recent review by Gromiha and Yugandhar [123]. Servers such as *bioCOmplexes COntact Maps* (COCOMAPS) [124] allow easy calculations of various interface features such as the SASA, and its decomposition in per residue and per polar and apolar area, as well as H-bonds and number of nearby atoms, making it a relatively comprehensible and quick method for interface characterization. Intersurf [125], besides providing information on residue SASA, also outputs the numbers of surface and buried atoms. This information is reported for both monomers and complexes. All the information gathered about protein-protein interfaces can subsequently be used to predict which interface regions or residues are the most important for the binding, namely HS, as represented in Figure 6, and hot-regions. Several methods have been developed for this purpose over the years, among which SpotOn [11] is the most recent one with the highest reported sensitivity (0.95) and accuracy (0.98)

for HS prediction so far. It uses both structural and sequence features such as SASA, H-bonds and PSSMs to predict which residues should be considered HS or NS [126]. As for hot-regions – characterized as HS clusters (HS are not randomly distributed across the protein-protein interface but rather clustered [127]) – HotRegion [128] can be used, and it is based on HotPoint [129] to predict HS across all PDB entries.

<Figure 6 near here>

C. Template-based modelling of protein-protein complexes

Although many advances have been made in the field of ab-initio docking in the past decades, Template-based modelling (TBM) is still the most reliable way of modelling 3D structures of protein complexes [130]. In fact, a 2012 article concluded that complex templates are available to model approximately all complexes from structural information on proteins, using large databases on protein-protein interactions [131].

TBM is based on the assumption that homologous proteins interact in a similar way, even though this assumption is sometimes violated. TBM methods typically retrieve experimentally determined structures of protein complexes that are homologous to the query proteins, and use them as templates to infer the structure of the complex formed by the query proteins. For the sake of simplicity, we refer to the structure of the homologous complex as a template in this section even though it differs from templates used for single proteins. The development of TBM algorithms has to consider three aspects: i) How to retrieve reliable templates from the PDB, ii) how to use the template information (do we use the global structural information or only use the local interface information), and iii) how to determine which template is the best suited and which model is the best? Existing TBM methods can be largely grouped into five strategies: Superimposition, dimeric threading [132,133], direct homology modelling of the complex, e.g., with Modeller[79], Interface Structure Alignment (ISA) based methods [134,135], and interface residues or contacts restrained docking approach[136,137].

Superimposition is the simplest among the five strategies. It takes the unbound structures of the query proteins and superimposes them onto the template structures by minimizing the RMSD (Root Mean Squared Deviation). This is essentially a global structure-structure alignment. Superimposition treats the unbound query structures as rigid and hence fails to model any conformational changes, which are quite common as protein molecules are flexible and often undergo various amounts of conformational changes upon binding. It also tends to generate many steric clashes and thus requires a subsequent refinement of the models.

Dimeric threading is essentially a sequence-structure alignment based approach. It "threads" the query protein sequences unto the templates by optimizing a scoring function (which can be energy- or conservation-based or hybrid). The advantage of dimeric threading is that the conformational changes upon binding are naturally taken into account provided they are reflected in the selected template. However, the nature of sequence-structure alignment ignores the rich information of the individual query protein structures and depends on the alignment quality, which also defines the quality of the final models.

Direct homology modelling of a complex is also widely used, e.g. using MODELLER. Taking query-template alignment files and the template structures as input, MODELLER predicts structures of protein complexes by satisfaction of spatial restraints. MODELLER can also model conformational changes if such changes are present in the templates.

The interface structure alignment-based method superimposes the query protein structures onto template interfaces (as compared to the whole template structures used in superimposition), hence taking advantage of the structural and sequence conservation of the interface area.

Interface residues and contacts restrained docking was introduced in HADDOCK [138,139], a flexible macromolecular docking software which treats external information as restraints to guide the docking process. In the context of template-based modelling of complexes, HADDOCK uses contacts derived from identified homologous templates as distance restraints to guide the docking process[131]. Compared to structure superimposition approaches the resulting models are typically clash-free. This approach can also seamlessly integrate distance restraints from multiple templates: Templates can be clustered based their structural similarity, and from each cluster interfacial distance ranges can be calculated and used as input of HADDOCK. Another recent template-based approach models the quaternary protein structure of homo- and hetero-oligomers by homology, combining interface conservation, structural clustering and other interface-related features [140].

4. Structural characterization of membrane protein-protein complexes

While several of the methods described above – especially those that use information about the interface, such as interface characterization methods and information-driven docking with HADDOCK [141] – can be applied to Membrane Proteins (MPs) without the need for any major adaptations, in general, interface prediction and modelling of MPs complexes are more

challenging. We only provide a short overview here and refer to Almeida *et al.* [142] for a more detailed review. Some approaches have been developed specifically for interfacial residue prediction in MPs. ML algorithms such as SVMs have been applied to this problem [143] . They use PSSMs and evolutionary rates as calculated by *Robust EVolutionary COnservation Measur*e (REVCOM) [144], and information on membrane spanning regions as provided by the TMDET webserver [145]. Other methods using similar features but other ML algorithm types such as Random Forests (RFs) [146] have been reported [146]. TMH-Expo [147] is a method developed for intra-monomer contact prediction in multi-spanning helical MPs using Artificial Neural Networks (ANNs). Even though this method was created for intra-monomer contacts, an adaptation of this algorithm for MP interfaces could be of great use in MP complex prediction. Table 3 contains the methods mentioned in this section, as well as a short explanation on each.

Conclusions

As illustrated in this review, the prediction of the structure of proteins and their complexes and the characterisation of their interfaces has greatly benefited from the development of algorithms and computational methods that can harvest information from sequence data. Sequence information alone has generated an incredible amount of mathematical- and informatics-based models and algorithms, capable of producing relevant information on sequence conservation. MSAs, PSSMs and (co-)evolutionary scores, in particular, have been key concepts in deciphering the information contained in sequence and adapting it to more complex bioinformatics tasks and to structure prediction of proteins and their complexes.

Funding

Li Xue is supported by Veni grant 722.014.005 from the NWO (Netherlands Organisation for Scientific Research). Irina S. Moreira acknowledges support by the Fundação para a Ciência e a Tecnologia (FCT) Investigator programme - IF/00578/2014 (co-financed by European Social Fund and Programa Operacional Potencial Humano), and a Marie Skłodowska-Curie Individual Fellowship MSCA-IF-2015 [MEMBRANEPROT 659826]. This work was also financed by the European Regional Development Fund (ERDF), through the Centro 2020 Regional Operational Programme under project CENTRO-01-0145-FEDER-000008: BrainHealth 2020, and through the COMPETE 2020 - Operational Programme for Competitiveness and Internationalisation and Portuguese national funds via FCT, under project POCI-01-0145-FEDER-007440.

References

1 Nelson, D. L. & Cox, M. M. *Lehninger Principles of Biochemistry*. (W. H. Freeman, 2012).

2 Ito, K. Impact of post-translational modifications of proteins on the inflammatory process. *Biochemical Society Transactions* **35**, 281-283, doi:10.1042/bst0350281 (2007).

3 Lamond, A. I. Molecular biology of the cell, 4th edition. *Nature* **417**, 383-383, doi:Doi 10.1038/417383a (2002).

4 Markiv, A., Rambaruth, N. D. & Dwek, M. V. Beyond the genome and proteome: targeting protein modifications in cancer. *Current opinion in pharmacology* **12**, 408-413, doi:10.1016/j.coph.2012.04.003 (2012).

5 Jones, S. & Thornton, J. M. Principles of protein-protein interactions. *Proc Natl Acad Sci U S A* **93**, 13-20 (1996).

6 Jones, S. & Thornton, J. M. Analysis of protein-protein interaction sites using surface patches. *J Mol Biol* **272**, 121-132, doi:10.1006/jmbi.1997.1234 (1997).

7 Yan, C., Wu, F., Jernigan, R. L., Dobbs, D. & Honavar, V. Characterization of Protein–Protein Interfaces. *The protein journal* **27**, 59-70, doi:10.1007/s10930-007-9108-x (2008).

8 Bendell, C. J. *et al.* Transient protein-protein interface prediction: datasets, features, algorithms, and the RAD-T predictor. *BMC Bioinformatics* **15**, 82-82, doi:10.1186/1471-2105-15-82 (2014).

9 Moreira, I. S., Ramos, R. M., Martins, J. M., Fernandes, P. A. & Ramos, M. J. Are hot-spots occluded from water? *Journal of biomolecular structure & dynamics* **32**, 186-197, doi:10.1080/07391102.2012.758598 (2014).

10 Moreira, I. S., Fernandes, P. A. & Ramos, M. J. Hot spots--a review of the protein-protein interface determinant amino-acid residues. *Proteins* **68**, 803-812, doi:10.1002/prot.21396 (2007).

11 Moreira, I. S. *et al.* SpotOn: a web server for protein-protein binding hot-spots. *Sci. Rep. (accepted)* (2017).

12 Berman, H. M. *et al.* The Protein Data Bank. *Nucleic acids research* **28**, doi:10.1093/nar/28.1.235 (2000).

13 Barrett, P. J. *et al.* The Quiet Renaissance of Protein NMR. *Biochemistry* **52**, 1303-1320, doi:10.1021/bi4000436 (2013).

14 Wider, G. & Wuthrich, K. NMR spectroscopy of large molecules and multimolecular assemblies in solution. *Curr Opin Struct Biol* **9**, 594-601 (1999).

15 Foster, M. P., McElroy, C. A. & Amero, C. D. Solution NMR of large molecules and assemblies. *Biochemistry* **46**, 331-340, doi:10.1021/bi0621314 (2007).

16 Frueh, D. P. Practical aspects of NMR signal assignment in larger and challenging proteins. *Progress in nuclear magnetic resonance spectroscopy* **78**, 47-75, doi:10.1016/j.pnmrs.2013.12.001 (2014).

17 Milne, J. L. *et al.* Cryo-electron microscopy--a primer for the non-microscopist. *The FEBS journal* **280**, 28-45, doi:10.1111/febs.12078 (2013).

18 Merk, A. *et al.* Breaking Cryo-EM Resolution Barriers to Facilitate Drug Discovery. *Cell* **165**, 1698-1707, doi:10.1016/j.cell.2016.05.040 (2016).

19 Liao, M., Cao, E., Julius, D. & Cheng, Y. Structure of the TRPV1 ion channel determined by electron cryo-microscopy. *Nature* **504**, 107-112, doi:10.1038/nature12822 (2013).

20 Park, E., Campbell, E. B. & MacKinnon, R. Structure of a CLC chloride ion channel by cryo-electron microscopy. *Nature* **541**, 500-505, doi:10.1038/nature20812 (2017).

21 Khademi, S. *et al.* Mechanism of ammonia transport by Amt/MEP/Rh: structure of AmtB at 1.35 A. *Science* **305**, 1587-1594, doi:10.1126/science.1101952 (2004).

22 Lučić, V., Rigort, A. & Baumeister, W. Cryo-electron tomography: The challenge of doing structural biology in situ. *The Journal of Cell Biology* **202**, 407-419, doi:10.1083/jcb.201304193 (2013).
23 Lawson, C. L. *et al.* EMDataBank unified data resource for 3DEM. *Nucleic Acids Res* **44**, D396-403, doi:10.1093/nar/gkv1126 (2016).
24 Almeida, J. G., Preto, A. J., Koukos, P. I., Bonvin, A. & Moreira, I. S. MEMBRANE PROTEINS STRUCTURES: A review on computational modeling tools. *Biochim Biophys Acta*, doi:10.1016/j.bbamem.2017.07.008 (2017).
25 Williamson, M. P. & Sutcliffe, M. J. Protein-protein interactions. *Biochemical Society transactions* **38**, 875-878, doi:10.1042/BST0380875 (2010).
26 Whisstock, J. C. & Lesk, A. M. Prediction of protein function from protein sequence and structure. *Q Rev Biophys* **36**, 307-340 (2003).
27 Petersen, T. N. *et al.* Prediction of protein secondary structure at 80% accuracy. *Proteins* **41**, 17-20 (2000).
28 Rost, B. & Sander, C. Prediction of protein secondary structure at better than 70% accuracy. *J Mol Biol* **232**, 584-599, doi:10.1006/jmbi.1993.1413 (1993).
29 Kretsinger, R. H., Ison, R. E. & Hovmoller, S. Prediction of protein structure. *Methods Enzymol* **383**, 1-27, doi:10.1016/S0076-6879(04)83001-5 (2004).
30 Shortle, D. Prediction of protein structure. *Current biology : CB* **10**, R49-51 (2000).
31 Argos, P. & Rao, J. K. Prediction of protein structure. *Methods Enzymol* **130**, 185-207 (1986).
32 Edwards, Y. J. & Cottage, A. Prediction of protein structure and function by using bioinformatics. *Methods Mol Biol* **175**, 341-375, doi:10.1385/1-59259-235-X:341 (2001).
33 Nanni, L., Brahnam, S. & Lumini, A. Prediction of protein structure classes by incorporating different protein descriptors into general Chou's pseudo amino acid composition. *Journal of theoretical biology* **360**, 109-116, doi:10.1016/j.jtbi.2014.07.003 (2014).
34 Hartlmuller, C., Gobl, C. & Madl, T. Prediction of Protein Structure Using Surface Accessibility Data. *Angewandte Chemie*, doi:10.1002/anie.201604788 (2016).
35 Al-Lazikani, B., Hill, E. E. & Morea, V. Protein structure prediction. *Methods Mol Biol* **453**, 33-85, doi:10.1007/978-1-60327-429-6_2 (2008).
36 Westhead, D. R. & Thornton, J. M. Protein structure prediction. *Current opinion in biotechnology* **9**, 383-389 (1998).
37 Benner, S. A., Geroff, D. L. & Rozzell, J. D. Protein structure prediction. *Science* **274**, 1448b-1449b, doi:10.1126/science.274.5292.1448b (1996).
38 Barton, G. J. & Russell, R. B. Protein structure prediction. *Nature* **361**, 505-506, doi:10.1038/361505b0 (1993).
39 Robson, B. & Garnier, J. Protein structure prediction. *Nature* **361**, 506, doi:10.1038/361506a0 (1993).
40 Garnier, J. Protein structure prediction. *Biochimie* **72**, 513-524 (1990).
41 Hui, W. Q., Cheng, Q., Liu, T. Y. & Ouyang, Q. Homology modeling, docking, and molecular dynamics simulation of the receptor GALR2 and its interactions with galanin and a positive allosteric modulator. *Journal of molecular modeling* **22**, 90, doi:10.1007/s00894-016-2944-x (2016).
42 Nugent, T. De novo membrane protein structure prediction. *Methods Mol Biol* **1215**, 331-350, doi:10.1007/978-1-4939-1465-4_15 (2015).
43 Nugent, T. & Jones, D. T. Accurate de novo structure prediction of large transmembrane protein domains using fragment-assembly and correlated mutation analysis. *Proc Natl Acad Sci U S A* **109**, E1540-1547, doi:10.1073/pnas.1120036109 (2012).
44 Joseph, A. P., Polles, G., Alber, F. & Topf, M. Integrative modelling of cellular assemblies. *Curr Opin Struct Biol* **46**, 102-109, doi:10.1016/j.sbi.2017.07.001 (2017).

45 Schneidman-Duhovny, D., Pellarin, R. & Sali, A. Uncertainty in Integrative Structural Modeling. *Current opinion in structural biology* **0**, 96-104, doi:10.1016/j.sbi.2014.08.001 (2014).
46 Rodrigues, J. P. G. L. M. a. B., A. M. J. J. . Integrative computational modeling of protein interactions. *FEBS J* **281**, doi:10.1111/febs.12771 (2014).
47 Landau, M. *et al.* ConSurf 2005: the projection of evolutionary conservation scores of residues on protein structures. *Nucleic Acids Res* **33**, W299-302, doi:10.1093/nar/gki370 (2005).
48 Braun, T., Koehler Leman, J. & Lange, O. F. Combining Evolutionary Information and an Iterative Sampling Strategy for Accurate Protein Structure Prediction. *PLoS Comput Biol* **11**, e1004661, doi:10.1371/journal.pcbi.1004661 (2015).
49 Hopf, T. A. *et al.* Sequence co-evolution gives 3D contacts and structures of protein complexes. *Elife* **3**, e03430 (2014).
50 Ashkenazy, H. *et al.* ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules. *Nucleic Acids Res* **44**, W344-350, doi:10.1093/nar/gkw408 (2016).
51 Yu, J. *et al.* InterEvDock: a docking server to predict the structure of protein-protein interactions using evolutionary information. *Nucleic Acids Res* **44**, W542-549, doi:10.1093/nar/gkw340 (2016).
52 Do, C. B. & Katoh, K. Protein multiple sequence alignment. *Methods in molecular biology (Clifton, N.J.)* **484**, 379-413, doi:10.1007/978-1-59745-398-1_25 (2008).
53 Needleman, S. B. & Wunsch, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol* **48**, 443-453 (1970).
54 Heringa, J. & Taylor, W. R. Three-dimensional domain duplication, swapping and stealing. *Curr Opin Struct Biol* **7**, 416-421 (1997).
55 Katti, M. V., Sami-Subbu, R., Ranjekar, P. K. & Gupta, V. S. Amino acid repeat patterns in protein sequences: their diversity and structural-functional implications. *Protein Sci* **9**, 1203-1209, doi:10.1110/ps.9.6.1203 (2000).
56 Smith, T. F. & Waterman, M. S. Identification of common molecular subsequences. *J Mol Biol* **147**, 195-197 (1981).
57 Waterman, M. S. & Eggert, M. A new algorithm for best subsequence alignments with application to tRNA-rRNA comparisons. *J Mol Biol* **197**, 723-728 (1987).
58 Bailey, T. L. & Elkan, C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proceedings / ... International Conference on Intelligent Systems for Molecular Biology ; ISMB. International Conference on Intelligent Systems for Molecular Biology* **2**, 28-36 (1994).
59 Notredame, C., Higgins, D. G. & Heringa, J. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol* **302**, 205-217, doi:10.1006/jmbi.2000.4042 (2000).
60 Simossis, V., Kleinjung, J. & Heringa, J. An overview of multiple sequence alignment. *Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.]* **Chapter 3**, Unit 3 7, doi:10.1002/0471250953.bi0307s03 (2003).
61 Apweiler, R. *et al.* UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res* **32**, D115-119, doi:10.1093/nar/gkh131 (2004).
62 Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J Mol Biol* **215**, 403-410, doi:10.1016/S0022-2836(05)80360-2 (1990).
63 Dayhoff, M. O., Schwartz, R. M. & Orcutt, B. C. A model of evolutionary change in proteins. *Atlas of protein sequence and structure* **5**, 345-351, doi:citeulike-article-id:4442167 (1978).
64 Henikoff, S. & Henikoff, J. G. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A* **89**, 10915-10919 (1992).

65 Mount, D. W. Comparison of the PAM and BLOSUM Amino Acid Substitution Matrices. *CSH protocols* **2008**, pdb ip59, doi:10.1101/pdb.ip59 (2008).
66 Eddy, S. R. Hidden Markov models. *Curr Opin Struct Biol* **6**, 361-365 (1996).
67 Pupko, T., Bell, R. E., Mayrose, I., Glaser, F. & Ben-Tal, N. Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. *Bioinformatics* **18 Suppl 1**, S71-77 (2002).
68 Heger, A. & Ponting, C. P. Evolutionary rate analyses of orthologs and paralogs from 12 Drosophila genomes. *Genome Research* **17**, 1837-1849, doi:10.1101/gr.6249707 (2007).
69 Vallender, E. J. Bioinformatic approaches to identifying orthologs and assessing evolutionary relationships. *Methods (San Diego, Calif.)* **49**, 50-55, doi:10.1016/j.ymeth.2009.05.010 (2009).
70 Di Lena, P., Fariselli, P., Margara, L., Vassura, M. & Casadio, R. Is there an optimal substitution matrix for contact prediction with correlated mutations? *IEEE/ACM transactions on computational biology and bioinformatics* **8**, 1017-1028, doi:10.1109/tcbb.2010.91 (2011).
71 Jones, D. T., Buchan, D. W., Cozzetto, D. & Pontil, M. PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics* **28**, 184-190, doi:10.1093/bioinformatics/btr638 (2012).
72 Baker, F. N. & Porollo, A. CoeViz: a web-based tool for coevolution analysis of protein residues. *BMC Bioinformatics* **17**, 119, doi:10.1186/s12859-016-0975-z (2016).
73 Morcos, F. *et al.* Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proceedings of the National Academy of Sciences* **108**, E1293-E1301, doi:10.1073/pnas.1111471108 (2011).
74 Marks, D. S. *et al.* Protein 3D structure computed from evolutionary sequence variation. *PLoS One* **6**, e28766, doi:10.1371/journal.pone.0028766 (2011).
75 Boutet, E. *et al.* UniProtKB/Swiss-Prot, the Manually Annotated Section of the UniProt KnowledgeBase: How to Use the Entry View. *Methods Mol Biol* **1374**, 23-54, doi:10.1007/978-1-4939-3167-5_2 (2016).
76 Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421-421, doi:10.1186/1471-2105-10-421 (2009).
77 Söding, J., Biegert, A. & Lupas, A. N. The HHpred interactive server for protein homology detection and structure prediction. *Nucleic acids research* **33**, W244-W248, doi:10.1093/nar/gki408 (2005).
78 Alexander, P. A., He, Y., Chen, Y., Orban, J. & Bryan, P. N. A minimal sequence code for switching protein structure and function. *Proc Natl Acad Sci U S A* **106**, 21149-21154, doi:10.1073/pnas.0906408106 (2009).
79 Webb, B. & Sali, A. Comparative Protein Structure Modeling Using MODELLER. *Current protocols in protein science* **86**, 2.9.1-2.9.37, doi:10.1002/cpps.20 (2016).
80 Kaufmann, K. W., Lemmon, G. H., DeLuca, S. L., Sheehan, J. H. & Meiler, J. Practically Useful: What the Rosetta Protein Modeling Suite Can Do for You. *Biochemistry* **49**, 2987-2998, doi:10.1021/bi902153g (2010).
81 Song, Y. *et al.* High-resolution comparative modeling with RosettaCM. *Structure (London, England : 1993)* **21**, 1735-1742, doi:10.1016/j.str.2013.08.005 (2013).
82 Eswar, N. *et al.* Comparative Protein Structure Modeling Using Modeller. *Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.]* **0 5**, Unit-5.6, doi:10.1002/0471250953.bi0506s15 (2006).
83 Ovchinnikov, S. *et al.* Improved de novo structure prediction in CASP11 by incorporating coevolution information into Rosetta. *Proteins* **84 Suppl 1**, 67-75, doi:10.1002/prot.24974 (2016).

84 Fukuda, I., Yonezawa, Y. & Nakamura, H. Consistent Molecular Dynamics Scheme Applying the Wolf Summation for Calculating Electrostatic Interaction of Particles. *Journal of the Physical Society of Japan* **77**, 114301, doi:10.1143/JPSJ.77.114301 (2008).
85 Russell, R. B., Copley, R. R. & Barton, G. J. Protein fold recognition by mapping predicted secondary structures. *Journal of molecular biology* **259**, 349-365, doi:10.1006/jmbi.1996.0325 (1996).
86 Rohl, C. A., Strauss, C. E., Chivian, D. & Baker, D. Modeling structurally variable regions in homologous proteins with rosetta. *Proteins* **55**, 656-677, doi:10.1002/prot.10629 (2004).
87 Xu, D. & Zhang, Y. Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. *Proteins* **80**, 1715-1735, doi:10.1002/prot.24065 (2012).
88 Karakaş, M. *et al.* BCL::Fold - De Novo Prediction of Complex and Large Protein Topologies by Assembly of Secondary Structure Elements. *PLoS ONE* **7**, e49240, doi:10.1371/journal.pone.0049240 (2012).
89 Kaufmann, K. W., Lemmon, G. H., Deluca, S. L., Sheehan, J. H. & Meiler, J. Practically useful: what the Rosetta protein modeling suite can do for you. *Biochemistry* **49**, 2987-2998, doi:10.1021/bi902153g (2010).
90 Rohl, C. A., Strauss, C. E., Misura, K. M. & Baker, D. Protein structure prediction using Rosetta. *Methods Enzymol* **383**, 66-93, doi:10.1016/S0076-6879(04)83004-0 (2004).
91 Yang, J. *et al.* The I-TASSER Suite: protein structure and function prediction. *Nature methods* **12**, 7-8, doi:10.1038/nmeth.3213 (2015).
92 McGuffin, L. J., Bryson, K. & Jones, D. T. The PSIPRED protein structure prediction server. *Bioinformatics (Oxford, England)* **16**, 404-405 (2000).
93 Leman, J. K., Mueller, R., Karakas, M., Woetzel, N. & Meiler, J. Simultaneous prediction of protein secondary structure and transmembrane spans. *Proteins* **81**, 1127-1140, doi:10.1002/prot.24258 (2013).
94 Jones, D. T., Singh, T., Kosciolek, T. & Tetchner, S. MetaPSICOV: combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins. *Bioinformatics (Oxford, England)* **31**, 999-1006, doi:10.1093/bioinformatics/btu791 (2015).
95 Joo, K., Lee, S. J. & Lee, J. Sann: solvent accessibility prediction of proteins by nearest neighbor method. *Proteins* **80**, 1791-1797, doi:10.1002/prot.24074 (2012).
96 Magnan, C. N. & Baldi, P. SSpro/ACCpro 5: almost perfect prediction of protein secondary structure and relative solvent accessibility using profiles, machine learning and structural similarity. *Bioinformatics (Oxford, England)* **30**, 2592-2597, doi:10.1093/bioinformatics/btu352 (2014).
97 Kaján, L., Hopf, T. A., Kalaš, M., Marks, D. S. & Rost, B. FreeContact: fast and free software for protein contact prediction from residue co-evolution. *BMC Bioinformatics* **15**, 85, doi:10.1186/1471-2105-15-85 (2014).
98 Higgins, D. G. & Sharp, P. M. CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. *Gene* **73**, 237-244 (1988).
99 Katoh, K., Misawa, K., Kuma, K. & Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research* **30**, 3059-3066, doi:Doi 10.1093/Nar/Gkf436 (2002).
100 Edgar, R. C. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5**, 113, doi:10.1186/1471-2105-5-113 (2004).
101 Teixeira, P. L. *et al.* Membrane protein contact and structure prediction using co-evolution in conjunction with machine learning. *PLoS ONE* **12**, e0177866, doi:10.1371/journal.pone.0177866 (2017).

102 Schaarschmidt, J. M., Bohdan; Kryshtafovych, Andriy; Bonvin, Alexandre M.J.J. Assessment of Contact Predictions in CASP12: Co-Evolution and Deep Learning Coming of Age *PROTEINS: Structure, Function, and Bioinformatics* (2017).

103 Wang, S., Sun, S., Li, Z., Zhang, R. & Xu, J. Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model. *PLoS computational biology* **13**, e1005324, doi:10.1371/journal.pcbi.1005324 (2017).

104 Moreira, I. S., Fernandes, P. A. & Ramos, M. J. Protein-Protein Docking: Dealing with the Unkown. *Journal of computational chemistry* **31**, 317-342 (2010).

105 Vakser, I. A. Protein-protein docking: from interaction to interactome. *Biophys J* **107**, 1785-1793, doi:10.1016/j.bpj.2014.08.033 (2014).

106 Oliva, R., Chermak, E. & Cavallo, L. Analysis and Ranking of Protein-Protein Docking Models Using Inter-Residue Contacts and Inter-Molecular Contact Maps. *Molecules* **20**, 12045-12060, doi:10.3390/molecules200712045 (2015).

107 Park, H., Lee, H. & Seok, C. High-resolution protein-protein docking by global optimization: recent advances and future challenges. *Curr Opin Struct Biol* **35**, 24-31, doi:10.1016/j.sbi.2015.08.001 (2015).

108 Zhang, Q. *et al.* Recent Advances in Protein-Protein Docking. *Current drug targets* **17**, 1586-1594 (2016).

109 Zhang, Q. C., Petrey, D., Norel, R. & Honig, B. H. Protein interface conservation across structure space. *Proc Natl Acad Sci USA* **107** (2010).

110 Ahmad, S. & Mizuguchi, K. Partner-Aware Prediction of Interacting Residues in Protein-Protein Complexes from Sequence Data. *PLoS ONE* **6**, e29104-e29104 (2011).

111 Xue, L. C., Dobbs, D. & Honavar, V. HomPPI: a class of sequence homology based protein-protein interface prediction methods. *BMC Bioinformatics* **12**, 1-24, doi:10.1186/1471-2105-12-244 (2011).

112 Xue, L. C., Dobbs, D., Bonvin, A. M. J. J. & Honavar, V. Computational prediction of protein interfaces: A review of data driven methods. *FEBS Letters* **589**, 3516-3526, doi:10.1016/j.febslet.2015.10.003 (2015).

113 ul Amir Afsar Minhas, F., Geiss, B. J. & Ben-Hur, A. PAIRpred: Partner-specific prediction of interacting residues from sequence and structure. *Proteins* **82**, 1142-1155, doi:10.1002/prot.24479 (2014).

114 Murakami, Y. & Mizuguchi, K. Applying the Naive Bayes classifier with kernel density estimation to the prediction of protein-protein interaction sites. *Bioinformatics (Oxford, England)* **26** (2010).

115 Porollo, A. & Meller, J. Prediction-based fingerprints of protein-protein interactions. *Proteins* **66** (2007).

116 de Vries, S. J., van Dijk, A. D. & Bonvin, A. M. WHISCY: what information does surface conservation yield? Application to data-driven docking. *Proteins* **63**, 479-489, doi:10.1002/prot.20842 (2006).

117 Neuvirth, H., Raz, R. & Schreiber, G. ProMate: a structure based prediction program to identify the location of protein-protein binding sites. *J Mol Biol* **338** (2004).

118 Qin, S. & Zhou, H. X. meta-PPISP: a meta web server for protein-protein interaction site prediction. *Bioinformatics (Oxford, England)* **23** (2007).

119 de Vries, S. J. & Bonvin, A. M. J. J. CPORT: A Consensus Interface Predictor and Its Performance in Prediction-Driven Docking with HADDOCK. *PLoS ONE* **6**, e17695-e17695 (2011).

120 Andreani, J., Faure, G. & Guerois, R. InterEvScore: a novel coarse-grained interface scoring function using a multi-body statistical potential coupled to evolution. *Bioinformatics (Oxford, England)* **29**, 1742-1749, doi:10.1093/bioinformatics/btt260 (2013).

121 He, B., Mortuza, S. M., Wang, Y., Shen, H. B. & Zhang, Y. NeBcon: Protein contact map prediction using neural network training coupled with naive Bayes classifiers. *Bioinformatics (Oxford, England)*, doi:10.1093/bioinformatics/btx164 (2017).

122 Wang, S., Sun, S., Li, Z., Zhang, R. & Xu, J. Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model. *PLoS Comput Biol* **13**, e1005324, doi:10.1371/journal.pcbi.1005324 (2017).

123 Gromiha, M. M. & Yugandhar, K. Integrating computational methods and experimental data for understanding the recognition mechanism and binding affinity of protein-protein complexes. *Prog Biophys Mol Biol* **128**, 33-38, doi:10.1016/j.pbiomolbio.2017.01.001 (2017).

124 Vangone, A., Spinelli, R., Scarano, V., Cavallo, L. & Oliva, R. COCOMAPS: a web application to analyze and visualize contacts at the interface of biomolecular complexes. *Bioinformatics (Oxford, England)* **27**, 2915-2916, doi:10.1093/bioinformatics/btr484 (2011).

125 Ray, N., Cavin, X., Paul, J. C. & Maigret, B. Intersurf: dynamic interface between proteins. *J Mol Graph Model* **23**, 347-354, doi:10.1016/j.jmgm.2004.11.004 (2005).

126 Wagner, S. *et al.* Consequences of membrane protein overexpression in Escherichia coli. *Molecular & cellular proteomics : MCP* **6**, 1527-1550, doi:10.1074/mcp.M600431-MCP200 (2007).

127 Keskin, O., Ma, B. & Nussinov, R. Hot regions in protein--protein interactions: the organization and contribution of structurally conserved hot spot residues. *J Mol Biol* **345**, 1281-1294, doi:10.1016/j.jmb.2004.10.077 (2005).

128 Cukuroglu, E., Gursoy, A. & Keskin, O. HotRegion: a database of predicted hot spot clusters. *Nucleic Acids Res* **40**, D829-833, doi:10.1093/nar/gkr929 (2012).

129 Tuncbag, N., Keskin, O. & Gursoy, A. HotPoint: hot spot prediction server for protein interfaces. *Nucleic Acids Res* **38**, W402-406, doi:10.1093/nar/gkq323 (2010).

130 Szilagyi, A. & Zhang, Y. Template-based structure modeling of protein-protein interactions. *Curr Opin Struct Biol* **24**, 10-23, doi:10.1016/j.sbi.2013.11.005 (2014).

131 Kundrotas, P. J., Zhu, Z., Janin, J. & Vakser, I. A. Templates are available to model nearly all complexes of structurally characterized proteins. *Proc Natl Acad Sci U S A* **109**, 9438-9441, doi:10.1073/pnas.1200678109 (2012).

132 Lu, L., Lu, H. & Skolnick, J. MULTIPROSPECTOR: an algorithm for the prediction of protein-protein interactions by multimeric threading. *Proteins* **49**, 350-364, doi:10.1002/prot.10222 (2002).

133 Mukherjee, S. & Zhang, Y. Protein-protein complex structure predictions by multimeric threading and template recombination. *Structure (London, England : 1993)* **19**, 955-966, doi:10.1016/j.str.2011.04.006 (2011).

134 Aytuna, A. S., Gursoy, A. & Keskin, O. Prediction of protein-protein interactions by combining structure and sequence conservation in protein interfaces. *Bioinformatics (Oxford, England)* **21**, 2850-2855, doi:10.1093/bioinformatics/bti443 (2005).

135 Kuzu, G., Gursoy, A., Nussinov, R. & Keskin, O. Exploiting conformational ensembles in modeling protein-protein interactions on the proteome scale. *Journal of proteome research* **12**, 2641-2653, doi:10.1021/pr400006k (2013).

136 Rodrigues, J. P. *et al.* Defining the limits of homology modeling in information-driven protein docking. *Proteins* **81**, 2119-2128, doi:10.1002/prot.24382 (2013).

137 Xue, L. C., Rodrigues, J. P., Dobbs, D., Honavar, V. & Bonvin, A. M. J. J. Template-based protein–protein docking exploiting pairwise interfacial residue restraints. *Briefings in Bioinformatics* **18**, 458-466, doi:10.1093/bib/bbw027 (2017).

138 Dominguez, C., Boelens, R. & Bonvin, A. M. HADDOCK: a protein-protein docking approach based on biochemical or biophysical information. *Journal of the American Chemical Society* **125**, 1731-1737, doi:10.1021/ja026939x (2003).

139 de Vries, S. J., van Dijk, M. & Bonvin, A. M. The HADDOCK web server for data-driven biomolecular docking. *Nat Protoc* **5**, 883-897, doi:10.1038/nprot.2010.32 (2010).
140 Bertoni, M., Kiefer, F., Biasini, M., Bordoli, L. & Schwede, T. Modeling protein quaternary structure of homo- and hetero-oligomers beyond binary interactions by homology. *Sci Rep* **7**, 10480, doi:10.1038/s41598-017-09654-8 (2017).
141 Van Zundert, G. C. P. *et al.* The HADDOCK2.2 Web Server: User-Friendly Integrative Modeling of Biomolecular Complexes. *Journal of Molecular Biology* **428**, 720-725, doi:10.1016/j.jmb.2015.09.014 (2016).
142 Almeida, J. G., Preto, A. J., Koukos, P. I., Bonvin, A. & Moreira, I. S. Membrane proteins structures: A review on computational modeling tools. *Biochimica et biophysica acta* **1859**, 2021-2039, doi:10.1016/j.bbamem.2017.07.008 (2017).
143 Asadabadi, E. B. & Abdolmaleki, P. Predictions of Protein-Protein Interfaces within Membrane Protein Complexes. *Avicenna journal of medical biotechnology* **5**, 148-157 (2013).
144 Bordner, A. J. & Abagyan, R. REVCOM: a robust Bayesian method for evolutionary rate estimation. *Bioinformatics (Oxford, England)* **21**, 2315-2321, doi:10.1093/bioinformatics/bti347 (2005).
145 Tusnady, G. E., Dosztanyi, Z. & Simon, I. TMDET: web server for detecting transmembrane regions of proteins by using their 3D coordinates. *Bioinformatics* **21**, 1276-1277, doi:10.1093/bioinformatics/bti121 (2005).
146 Bordner, A. J. Predicting protein-protein binding sites in membrane proteins. *BMC Bioinformatics* **10**, 312, doi:10.1186/1471-2105-10-312 (2009).
147 Li, B. *et al.* Accurate Prediction of Contact Numbers for Multi-Spanning Helical Membrane Proteins. *Journal of chemical information and modeling* **56**, 423-434, doi:10.1021/acs.jcim.5b00517 (2016).
148 Larkin, M. A. *et al.* Clustal W and Clustal X version 2.0. *Bioinformatics (Oxford, England)* **23**, 2947-2948, doi:10.1093/bioinformatics/btm404 (2007).
149 Huang, X. & Miller, W. A time-efficient, linear-space local similarity algorithm. *Advances in Applied Mathematics* **12**, 337-357, doi:http://dx.doi.org/10.1016/0196-8858(91)90017-D (1991).
150 Ekiert, D. C. *et al.* Architectures of Lipid Transport Systems for the Bacterial Outer Membrane. *Cell* **169**, 273-285 e217, doi:10.1016/j.cell.2017.03.019 (2017).
151 Ko, T. P., Liao, C. C., Ku, W. Y., Chak, K. F. & Yuan, H. S. The crystal structure of the DNase domain of colicin E7 in complex with its inhibitor Im7 protein. *Structure* **7**, 91-102 (1999).
152 Moreira, I. S. *et al.* SpotOn: High Accuracy Identification of Protein-Protein Interface Hot-Spots. *Scientific Reports* **7**, 8007, doi:10.1038/s41598-017-08321-2 (2017).
153 Deisenhofer, J. Crystallographic refinement and atomic models of a human Fc fragment and its complex with fragment B of protein A from Staphylococcus aureus at 2.9- and 2.8-A resolution. *Biochemistry* **20**, 2361-2370 (1981).
154 Kallberg, M., Margaryan, G., Wang, S., Ma, J. & Xu, J. RaptorX server: a resource for template-based protein structure modeling. *Methods in molecular biology (Clifton, N.J.)* **1137**, 17-27, doi:10.1007/978-1-4939-0366-5_2 (2014).
155 Roy, A., Kucukural, A. & Zhang, Y. I-TASSER: a unified platform for automated protein structure and function prediction. *Nat Protoc* **5**, 725-738, doi:10.1038/nprot.2010.5 (2010).
156 Leman, J. K., Mueller, R., Karakas, M., Woetzel, N. & Meiler, J. Simultaneous prediction of protein secondary structure and transmembrane spans. *Proteins* **81**, 1127-1140, doi:10.1002/prot.24258 (2013).
157 Ashkenazy, H. *et al.* ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules. *Nucleic acids research* **44**, W344-350, doi:10.1093/nar/gkw408 (2016).

158 Marks, D. S. *et al.* Protein 3D Structure Computed from Evolutionary Sequence Variation. *PLoS ONE* **6**, doi:10.1371/journal.pone.0028766 (2011).
159 Hopf, T. A. *et al.* Sequence co-evolution gives 3D contacts and structures of protein complexes. *eLife* **3**, e03430, doi:10.7554/eLife.03430 (2014).
160 Engelen, S., Trojan, L. A., Sacquin-Mora, S., Lavery, R. & Carbone, A. Joint Evolutionary Trees: A Large-Scale Method To Predict Protein Interfaces Based on Sequence Sampling. *PLoS computational biology* **5**, e1000267, doi:10.1371/journal.pcbi.1000267 (2009).
161 Ahmad, S. & Mizuguchi, K. Partner-Aware Prediction of Interacting Residues in Protein-Protein Complexes from Sequence Data. *PLoS ONE* **6**, e29104, doi:10.1371/journal.pone.0029104 (2011).
162 Xue, L. C., Dobbs, D. & Honavar, V. HomPPI: a class of sequence homology based protein-protein interface prediction methods. *BMC Bioinformatics* **12**, 244, doi:10.1186/1471-2105-12-244 (2011).
163 Minhas, F., Geiss, B. J. & Ben-Hur, A. PAIRpred: partner-specific prediction of interacting residues from sequence and structure. *Proteins* **82**, 1142-1155, doi:10.1002/prot.24479 (2014).
164 Murakami, Y. & Mizuguchi, K. Applying the Naive Bayes classifier with kernel density estimation to the prediction of protein-protein interaction sites. *Bioinformatics* **26**, 1841-1848, doi:10.1093/bioinformatics/btq302 (2010).
165 Porollo, A. & Meller, J. Prediction-based fingerprints of protein-protein interactions. *Proteins* **66**, 630-645, doi:10.1002/prot.21248 (2007).
166 Neuvirth, H., Raz, R. & Schreiber, G. ProMate: a structure based prediction program to identify the location of protein-protein binding sites. *J Mol Biol* **338**, 181-199, doi:10.1016/j.jmb.2004.02.040 (2004).
167 de Vries, S. J. & Bonvin, A. CPORT: A Consensus Interface Predictor and Its Performance in Prediction-Driven Docking with HADDOCK. *PLoS ONE* **6**, doi:10.1371/journal.pone.0017695 (2011).
168 He, B., Mortuza, S. M., Wang, Y., Shen, H. B. & Zhang, Y. NeBcon: protein contact map prediction using neural network training coupled with naive Bayes classifiers. *Bioinformatics (Oxford, England)* **33**, 2296-2306, doi:10.1093/bioinformatics/btx164 (2017).
169 Dong, G. Q., Fan, H., Schneidman-Duhovny, D., Webb, B. & Sali, A. Optimized atomic statistical potentials: assessment of protein interfaces and loops. *Bioinformatics (Oxford, England)* **29**, 3158-3166, doi:10.1093/bioinformatics/btt560 (2013).
170 Garzon, J. I. *et al.* FRODOCK: a new approach for fast rotational protein-protein docking. *Bioinformatics (Oxford, England)* **25**, 2544-2551, doi:10.1093/bioinformatics/btp447 (2009).
171 Li, B. *et al.* Accurate Prediction of Contact Numbers for Multi-Spanning Helical Membrane Proteins. *Journal of chemical information and modeling* **56**, 423-434, doi:10.1021/acs.jcim.5b00517 (2016).

Tables

Table 1. Structure prediction methods using sequence and evolutionary information described in Section 2.

METHOD	DESCRIPTION	REFERENCE
MODELLER	MODELLER is a software for homology modelling and it features methods for homolog sequence with structural information search and retrieval	[79]
RAPTORX	RAPTORX performs secondary structure prediction and tri dimensional homology modelling	[154]
I-TASSER	I-TASSER performs automated protein sequence and function predictions. Structure is generated by multiple threading alignments and function is inferred by matching the structural models with proteins of known function.	[155]
PSIPRED	PSIPRED is a method for secondary structure prediction, using PSSMs and two feed-forward neural networks.	[92]
BCL::JUFO9D	BCL:Jufo9d is a method that predicts both protein structure and transmembrane spans from sequence. To do so, it uses PSSMs as input in a feed-forward neural network which outputs a 9*9 matrix with information on two sets of three possible states – helix, strand and coil, and membrane core, interface and solution.	[156]
CONSURF	Consurf is a method used to calculate the evolutionary conservation of all residues in a protein by MSA. It uses the Rate4Site algorithm to calculate evolutionary rates from a MSA, and represents these results using alignments coloured regarding each residues conservation.	[157]
EVFOLD	EVFold is a method used to calculate residue coevolution for a protein sequence. This information is then used to predict the structure of the input protein sequence.	[158]
BCL::FOLD	BCL::Fold is a method that uses information and methods similar to EVFold, as well as idealized secondary structures to predict membrane protein structure.	[88]

Table 2. Methods for the structure prediction and characterisation of soluble protein-protein complexes from sequence information described in section 3.

METHOD	DESCRIPTION	REFERENCE
EVCOUPLINGS	EVCouplings is a method to determine protein contacts. It has been adapted in both the EVFold and EVComplex pipelines.	[159]
EVCOMPLEX	EVComplex uses EVCouplings for inter- and intramonomer contact prediction, predicting a structure for a complex formed between two proteins.	[159]
HADDOCK	Haddock performs biomolecular complex modelling and refinement.	[139]
JET	JET detects protein interfaces, residues involved in the folding process as well as residues involved in molecular recognition.	[160]
PPIPP	Employs two stage neural networks based on the interacting pairs to predict single-protein residues.	[161]
HOMPPI	HomPPI employs sequence homology-based methods to predict protein-protein interface residues. There are two variants, partner-specific (PS-HomPPI) and non partner-specific (NPS-HomPPI)	[162]
PAIRPRED	PAIRPRED predicts residue pairs within a protein complex from sequence and structural information.	[163]
PSIVER	PSIVER predicts protein-protein interaction sites through Naïve Bayes classification with kernel density estimation.	[164]
SPPIDER	SPIDDER predicts interaction sites through relative solvent accessibility and the employment of machine learning approaches.	[165]
WHISCY	WHISCY predicts protein-protein interfaces through surface conservation and structural information.	[116]
PROMATE	ProMate Identifies protein-protein binding sites from surface properties.	[166]
CPORT	CPORT combines several interface predictors into a consensus prediction.	[167]
INTEREVSCORE	InterEvScore is a scoring function that uses a coarse-grained statistical potential including two and three-body interactions, which helps report on the structural environment.	[120]
NEBCON	NeBcon predicts a contact map using neural networks, Naïve Bayes classification and co-evolutionary approaches.	[168]
COCOMAPS	COCOMAPS predicts and scores interface residues as well as protein contacts between complexes.	[124]
INTERSURF	Intersurf analyses dynamic interfaces and energy of protein complexes.	[125]
SPOTON	SpotOn identifies HS in the interface with high precision by employing machine learning methodologies.	[152]
HOTREGION	HotRegion is a database of predicted HS clusters.	[128]
HOTPOINT	HotPoint predicts HS on the interface using an empirical approach.	[129]
INTEREVDOCK	InterEvDock predicts protein-protein interactions using evolutionary information.	[51]

SOAP-PP	SOPA-PP assesses protein interfaces and loops by optimized statistical potential inclusion on a Bayesian framework.	[169]
FRODOCK	FRODOCK performs fast rotational protein-protein docking to generate predictions of protein-protein complexes.	[170]

Table 3. Methods for the structural characterization of membrane protein-protein complexes described in section 4.

METHOD	DESCRIPTION	REFERENCE
REVCOM	REVCOM estimates evolutionary conservation rates by employing Bayesian statistics.	[144]
TMDET	TMDET determines protein-membrane relative localization, identifying transmembrane regions using tri-dimensional coordinates.	[145]
TMH-EXPO	TMH-Expo maps contact numbers and identifies transmembrane helices interacting residues.	[171]

Figures

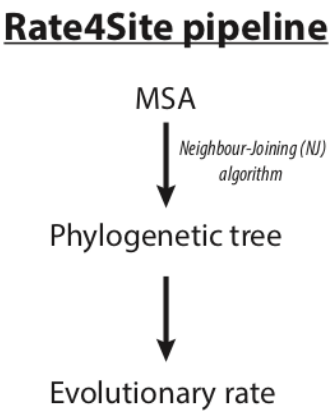

Figure 1. Rate4Site pipeline. Starting with an MSA, Rate4Site calculates phylogenetic trees using the Neighbour-Joining (NJ) algorithm [67]. This algorithm sequentially joins sequences that are closer and therefore more similar, with each other by creating a new node in each iteration. This node always connects to the tree constructed up to that point in the algorithm. Considering that a single position in the MSA has the same evolutionary rate (score), Rate4Site determines the maximum conservation rate that would explain the conditional probability of the data given that rate.

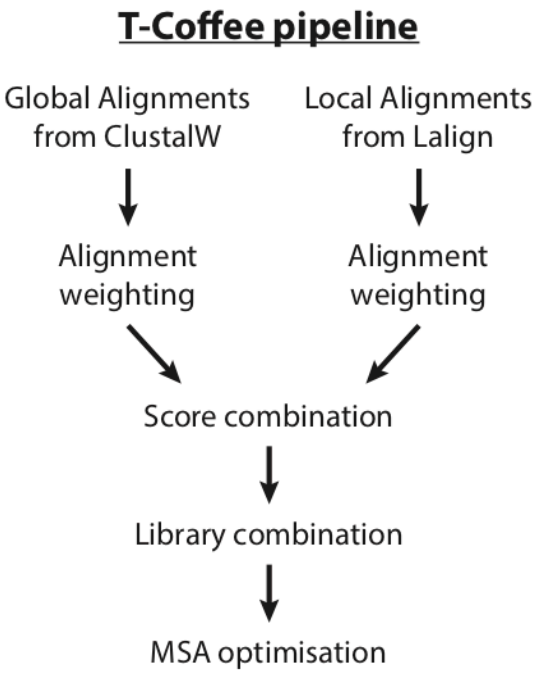

Figure 2. T-Coffee pipeline. First, T-Coffee retrieves the pairwise global and local alignments from ClustalW [148] and Lalign [149], respectively. These alignments are then weighted according to the pairwise sequence identity. To combine the two alignments, the scores for identical residue pairs from ClustalW and Lalgin at a given position are summed and considered as a single entry, while unique residue pairs at a given position are considered as separate entries. This creates a series of constraints providing a better MSA overall. Then, T-Coffee performs what the authors refer to as library extension, a heuristic process to calculate the likelihood of a pair based on triplets of matched residues – if two sequences share the same residue at a given position and, if other sequences have the same residue in that position, the weight for this residue pair will be as high as the number of triplets considering the initial residue pair. A weight of zero is attributed if a residue pair does not occur. By using a tree to calculate sequence similarity, the two most similar sequences are selected and the weights calculated during library extension are used to maximize the MSA score. Then, sequence pairs are added and residues are shifted until the final MSA is constructed. During this process, no gaps are removed after being added to the MSA.

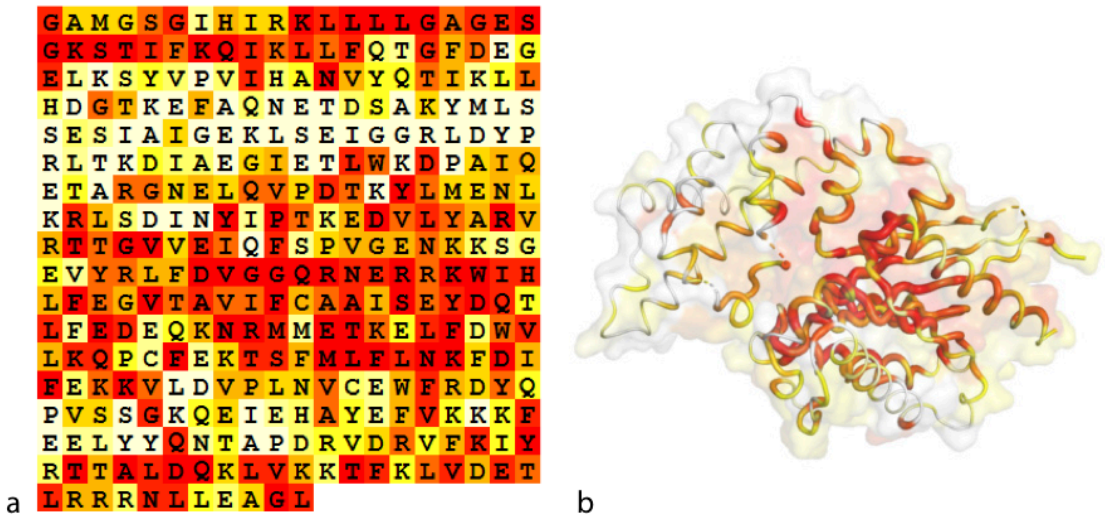

Figure 3. Visual representations of residue evolutionary conservation from Consurf for the Galpha-protein AtGPA1 from *Arabidopsis Thaliana*. (a) Sequence visualization of conservation, with each residue colored according to its conservation value. (b) Structure visualization of conservation. Both use a white (lowest)-yellow (mid)-red (highest) color scheme to represent conservation.

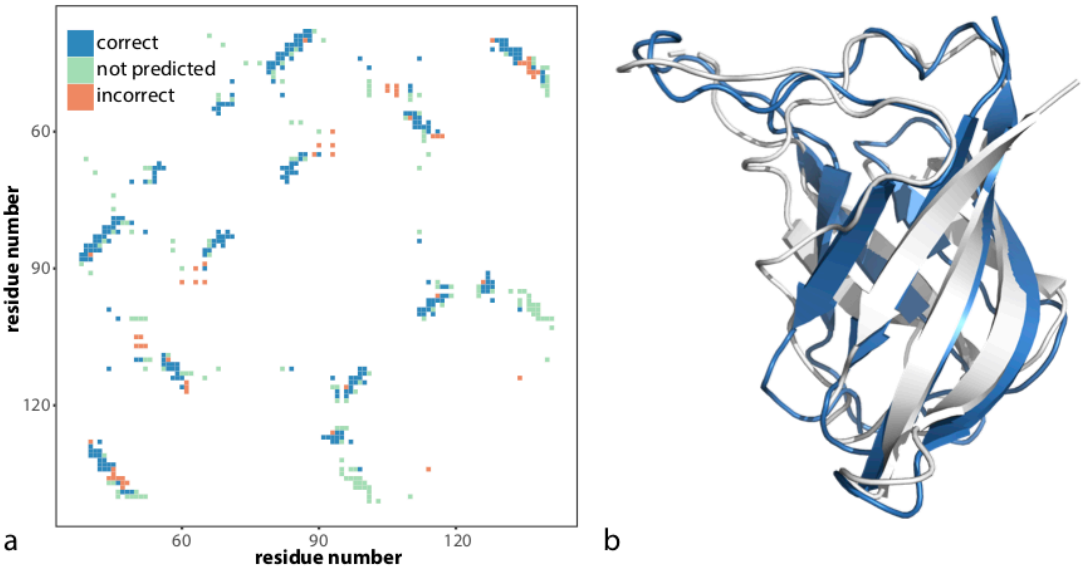

Figure 4. Structure prediction of CASP target T0866 (PDB-ID: 5UW2 [150]) assisted by predicted intramolecular contacts. (a) The contact prediction by Deepfold-Contact with the top 156 contacts (1.5* sequence length of the target) reaches a precision of 80% with the correct predictions (blue) distributed over all native contacts (blue and green), while the incorrect predictions (red) are mostly close to real contacts. (b) De novo structure prediction with Rosetta aided by this contact list yields a model (blue) closely matching the fold of the native protein (white).

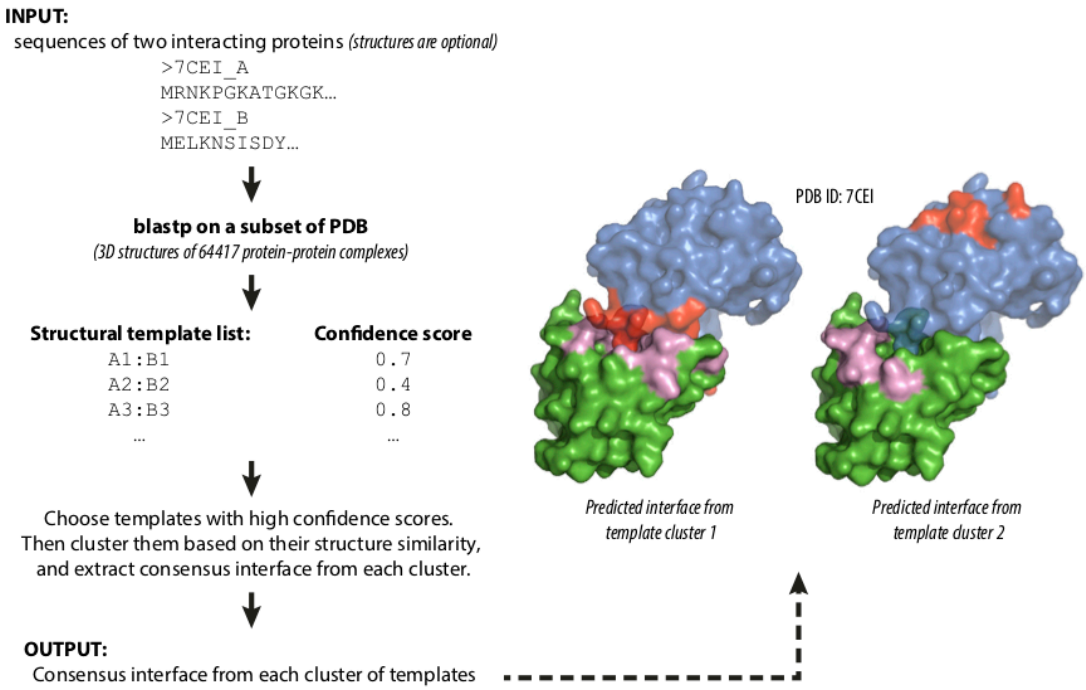

Figure 5. PS-HomPPI: partner-specific template based protein-protein interface predictor (http://ailab1.ist.psu.edu/PSHOMPPIv1.3/, [111]). PS-HomPPI v1.3 takes as input two protein sequences that are considered to interact with each other, and search the query against a protein-protein structure database (which is a subset of the PDB databank) for structural templates. For each structural template, PS-HomPPI calculates a confidence score based on the sequence similarity of the template to the query proteins. Based on the confidence score, templates are classified as Safe zone, Twilight zone and Dark zone templates. Only up to top K templates in the best available template zone are used for making interface predictions (K is 10 by default or a user-specified value). When multiple templates are available, they are clustered based on their structure similarity. PS-HomPPI calculates the consensus interface from each template cluster and return the consensus interfaces to the user. Here, the example query is 7CEI [151]. Two sets of potential interfaces are predicted by PS-HomPPI v1.3. They are mapped onto the experimentally determined structure (pink and red patches). As we can see, the first set of predicted interface aligns very well with the experimental structure, while the second set of the predicted interface is either wrong or indicates an alternative binding mode.

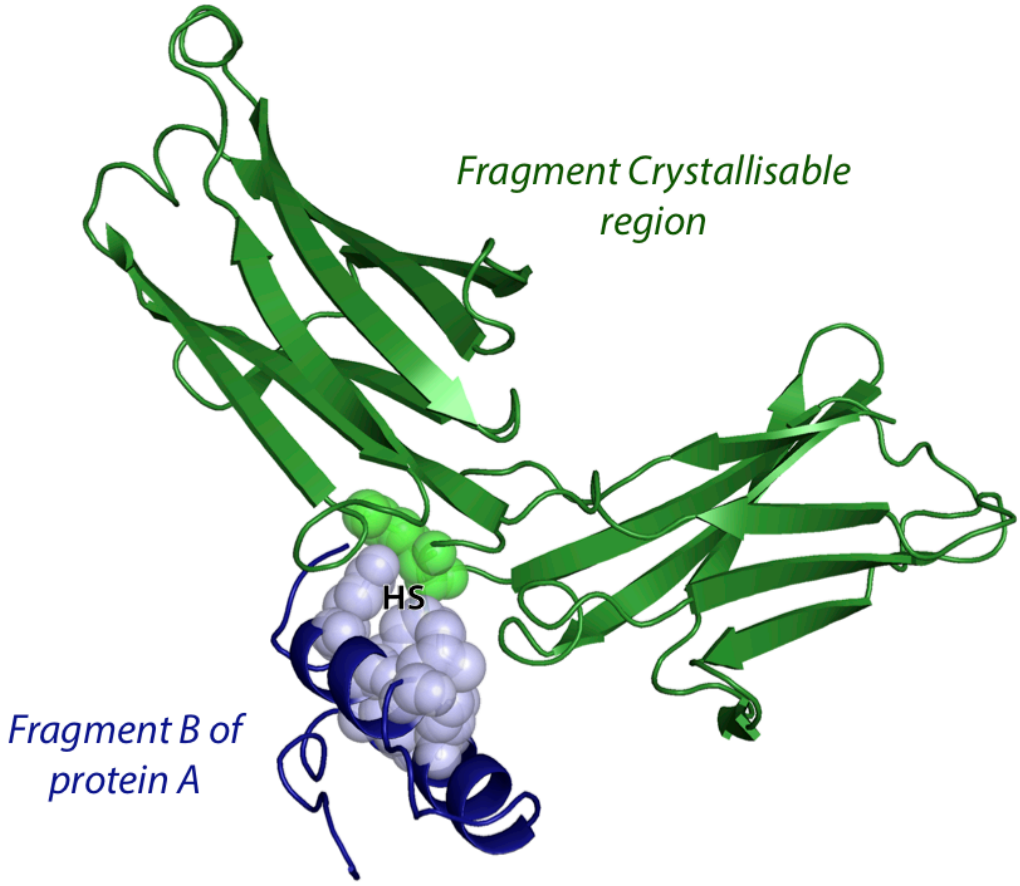

Figure 6. Structural representations of Hot-spots (HS; sphere representation colored by monomer) as predicted by SpotOn [152] for a complex (PDBID: 1FC2 [153]) formed between Fragment B of Protein A (blue) and a fragment crystallisable region of a human antibody (green). HS are key players in protein-protein interfaces as their disruption is deleterious for complex binding.