# Chapter 2
# Protein Structures, Interactions and Function from Evolutionary Couplings

**Thomas A. Hopf and Debora S. Marks**

**Abstract** The sequences of biomolecules such as proteins and RNA genes contain information about their three-dimensional states and functions. For over 40 years biologists have used the evolutionary conservation of this information to detect homology and predict important subsets of residues. Recent work has substantially extended this view of conservation by including the detection of evolutionary couplings, interactions, between residues, resulting in a paradigm shift in our ability to compute three-dimensional structures from sequences alone. In addition to three-dimensional structure of single proteins and RNA, this statistical analysis of evolutionary constraints can identify functional residues involved in ligand binding, biomolecule-interactions, alternative ensembles of conformations, "invisible" tertiary states of disordered proteins and allows quantitative prediction of effects of mutations. In this chapter we present an overview of the statistical inference methodologies, a survey of the resulting applications and challenges facing the field.

**Keywords** Sequence coevolution · Covariation · Evolutionary couplings · 3D structure prediction · Function prediction · Protein interactions · Disorder · Conformational changes · Mutation effects · Maximum entropy model

Parts of this chapter have been adapted from (Hopf 2016).

T.A. Hopf (✉) · D.S. Marks
Department of Systems Biology, Harvard Medical School, Boston, MA, USA
e-mail: thomas_hopf@hms.harvard.edu

D.S. Marks
e-mail: debbie@hms.harvard.edu

T.A. Hopf
Department of Cell Biology, Harvard Medical School, Boston, MA, USA

T.A. Hopf
Department of Informatics, Technische Universität München, Garching, Germany



D.J. Rigden (ed.), *From Protein Structure to Function with Bioinformatics*,
DOI 10.1007/978-94-024-1069-3_2

## 2.1 Introduction

Three-dimensional structure information is missing for a large fraction of known proteins and protein interactions, as experimental structure determination remains low-throughput whilst sequence databases grow exponentially. For instance, only about 50% of Pfam families have a solved structure for any of the family members (Finn et al. 2016) while structural coverage outside of conserved domains is even lower (Perdigao et al. 2015). Similarly, 60–80% of the approx. 10,000 and 40,000 heteromultimeric interactions in *E. coli* and human, respectively, have not yet been characterized structurally (Rajagopala et al. 2014; Mosca et al. 2014). The sustained effort to discover computational methods that have the potential to bypass the need for one-by-one experimental approaches is therefore motivated by this large experimental bottleneck. Comparative modelling transfers the coordinates from a solved protein to a target with similar sequence, based on the observation that the 3D folds of proteins remain conserved even as their amino acid sequences diverge (Webb and Sali 2014) (see also Chap. 4). In cases where no sequence-similar structural template can be identified, de novo fragment assembly methods (Qian et al. 2007) or even ab initio approaches using molecular force fields (Lindorff-Larsen et al. 2011) are an alternative for small proteins (<150 residues) (see also Chap 1). The applicability of these methods is however limited by the enormous size of conformational space that has to be searched as well as the accuracy of the available empirical force fields.

A conceptually different way of approaching the protein structure prediction problem is to mine the information contained in sequences. The evolutionary constraint to maintain residue interactions required for stable and functional proteins causes the coevolution of contacting amino acids. The idea therefore seems simple—find covarying positions in aligned protein sequences to identify residue pairs that correspond to physical contacts in the 3D structure, by analogy to the successful use of this approach in determining RNA secondary structure (Gutell et al. 1992). If correct, and if sufficient, these covarying residues could be transformed into distance constraints to construct 3D models, in a similar way to distances used in NMR structure determination.

However local covariation models applied to protein sequences did not consistently detect residues close in 3D (Shindyalov et al. 1994; Neher 1994; Gobel et al. 1994) despite some successful applications that showed enrichment of interacting residues (Skerker et al. (2008), Pazos et al. (1997)) or identification of contacts across proteins using additional biological information (Skerker et al. 2008). The apparent inability of these early covariance models to systematically identify contacting residues was attributed to a number of different reasons, including a loss of signal due to phylogenetic dependencies, the limited availability of sequence data and even the idea that we should not expect that truly coevolved residues are (mostly) close (Lapedes et al. 2012, 1997). Rather surprisingly, it turned out that changing the underlying model used to compute the couplings was the key innovation needed. This is because raw covariation frequencies or mutual

information between pairs of positions are dominated by 'indirect' transitive correlations, i.e. non-causal correlations between residues positions can be induced by a chaining of causal correlations between intervening residues positions. In a heterogeneous network, such as residues in a protein, these non-causal correlations can appear stronger than causal direct correlations, a well-understood feature of the Ising model in statistical physics where true correlations produce apparent long-range correlation at a distance (Giraud et al. 1999). The solution to this is to use a class of *global probability models* known as Potts model (a maximum entropy model) in statistical physics (Giraud et al. 1999; Lapedes et al. 1997; Ben-Naim and Lapedes 1999; Lapedes et al. 2012) and Markov Random Fields (an undirected graphical model) in computer science (Koller and Friedman 2009). Using these models the dependencies of types of amino acids in pairs of positions are computed simultaneously and consistently, rather than analysing pairs of positions independently of each other.

Application of these global statistical models was the key innovation in the identification of *evolutionary couplings* between pairs of positions in multiple sequence alignments that corresponded to contacting residues (Hopf et al. 2012; Marks et al. 2011, 2012; Morcos et al. 2011; Jones et al. 2012; Balakrishnan et al. 2011; Ekeberg et al. 2013; Lapedes et al. 2012; Michel et al. 2014). A retrospective analysis showed that even sequence data from 1999 PFAM family alignments was sufficient to infer large number of accurate residue contacts with the maximum entropy model for a few protein families (Marks et al. 2011). A pioneering Bayesian approach (Burger and van Nimwegen 2010, 2008) had some success but predictions were not as accurate with respect to residue proximity (Marks et al. 2011) and the use of belief propagation for parameter inference (Weigt et al. 2009) was computationally intractable for all but the smallest proteins. Although the methods required a sufficient number of sequences that diverged under functional selection, global statistical probability approaches such as those in Tables 2.1 and 2.2 provided a chance to obtain detailed structural and functional information for unsolved proteins of biological interest that was unprecedented.

Predicted contacts derived from evolutionary couplings have allowed the de novo prediction of protein 3D structures even for large molecules beyond the scope of previous approaches (Hopf et al. 2012; Marks et al. 2011, 2012; Hopf et al. 2015b; Ovchinnikov et al. 2014, 2015; Michel et al. 2014; Kosciolek and Jones 2014; Sulkowska et al. 2012) their complexes (Ovchinnikov et al. 2014; Hopf et al. 2014), multimeric contacts (Hopf et al. 2012; dos Santos et al. 2015), alternative conformations (Hopf et al. 2012; Toth-Petroczy et al. 2016; Morcos et al. 2013), and even the ability to predict structured states of apparently-disordered proteins (Toth-Petroczy et al. 2016). Many of these reports show, at least anecdotally, that evolutionary couplings models are able to identify functionally constrained residues over and above single column conservation and, most recently, the model has been used to make quantitative prediction of mutational changes in proteins (Hopf et al. 2017; Mann et al. 2014; Figliuzzi et al. 2016). In this chapter, we briefly describe the theoretical approach that underlies the methods, survey the most impactful

**Table 2.1** Webservers for evolutionary couplings (ECs) methods

| Method name | URL | Outputs | Inference method | Generates alignment | Refs. |
|---|---|---|---|---|---|
| EVfold | evfold.org | Alignments, EC pairs, 3D structures, functional residues | PLM | Yes | Marks et al. (2011), Toth-Petroczy et al. (2016) |
| EVcomplex | evcomplex.org | Protein complex alignments, Complex EC pairs | PLM | Yes | Hopf et al. (2014) |
| GREMLIN | gremlin.bakerlab.org | Alignments, EC pairs (incl. complexes), precomputed ECs and 3D structures | PLM | Yes | Kamisetty et al. (2013), Ovchinnikov et al. (2014) |
| DCA | dca.rice.edu | EC pairs | Mean-field | No | Morcos et al. (2011) |
| MetaPSICOV | bioinf.cs.ucl.ac.uk/MetaPSICOV | EC pairs | Sparse inv. cov., PLM, machine learning | Yes | Jones et al. (2012), (2015) |
| PconsC | c2.pcons.net | EC pairs | Sparse inv. cov., PLM, machine learning | Yes | Michel et al. (2014) |

**Table 2.2** Standalone evolutionary couplings inference software

| Method name | Inference algorithm | URL | Special features | Restrictions | Ref. |
|---|---|---|---|---|---|
| plmc | PLM | github.com/debbiemarkslab/plmc | Arbitrary sequences (incl. RNA), probabilistic treatment of gaps | – | (Weinreb et al. 2016; Toth-Petroczy et al. 2016) |
| CCMpred | PLM | github.com/soedinglab/CCMpred | Can be used on GPUs | – | (Seemayer et al. 2014) |
| plmDCA | PLM | plmdca.csc.kth.se | – | Matlab required | (Ekeberg et al. 2013) |
| GREMLIN | PLM | gremlin.bakerlab.org | – | Matlab required | (Balakrishnan et al. 2011; Kamisetty et al. 2013; Ovchinnikov et al. 2014) |
| DCA | Mean-field | dca.rice.edu | – | Matlab required | (Morcos et al. 2011) |
| PSICOV | Sparse inverse covariance | bioinfadmin.cs.ucl.ac.uk/downloads/PSICOV | – | – | (Jones et al. 2012) |
| FreeContact | Mean-field, sparse inverse covariance | rostlab.org/owiki/index.php/FreeContact | Implementation of both DCA and PSICOV algorithms | – | (Kajan et al. 2014) |
| MetaPSICOV | Sparse inverse covariance, PLM, machine learning | bioinfadmin.cs.ucl.ac.uk/downloads/MetaPSICOV | Meta-predictor | – | (Jones et al. 2012; 2015) |
| PconsC | Sparse inverse covariance, PLM, machine learning | c2.pcons.net | Meta-predictor | – | (Michel et al. 2014) |

applications and finally suggest challenges for the future, some of which might be solved by the time you read this!

## 2.2 Evolutionary Couplings from Sequence Alignments

The basis of coevolution-based structure and function prediction methods is the quantification of evolutionary couplings between all amino acid types in all pairs of sites derived from a multiple sequence alignment of the protein family (Fig. 2.1). These evolutionary couplings open up a wide variety of applications (Fig. 2.2).

### *2.2.1 The Global Model*

To avoid indirect correlations of residues pairs (as described above), global methods infer a probabilistic description of the sequence alignment that explains the observed correlations using underlying causative couplings between positions. These couplings are inferred by maximising the likelihood of observing the sequences in the alignment under the maximum entropy/Markov random field probability model.

Pairwise couplings are computed between amino acids to limit the number of model parameters to $O(N^2)$, but models of higher order (e.g. triples) are in principle possible given large enough protein families.

Under the pairwise graphical model the probability of any amino acid sequence $\sigma = (\sigma_1, \ldots, \sigma_n)$ of length $N$ is defined as

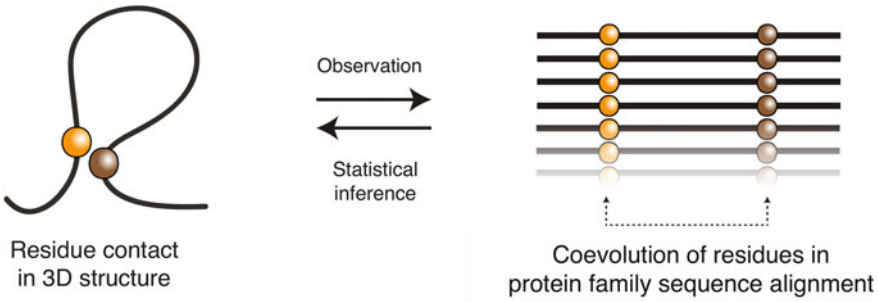


**Fig. 2.1** Residue interactions leave a coevolutionary record in protein sequences. The evolutionary constraint to maintain residue interactions, e.g. required for stable protein structures or complex formation with other molecules, creates a record of amino acid covariation in protein family sequence alignments. Mining this sequence record for residue pairs with strong evolutionary couplings using global statistical models opens a window to protein structure and function prediction (adapted from Hopf 2016, 2015b)

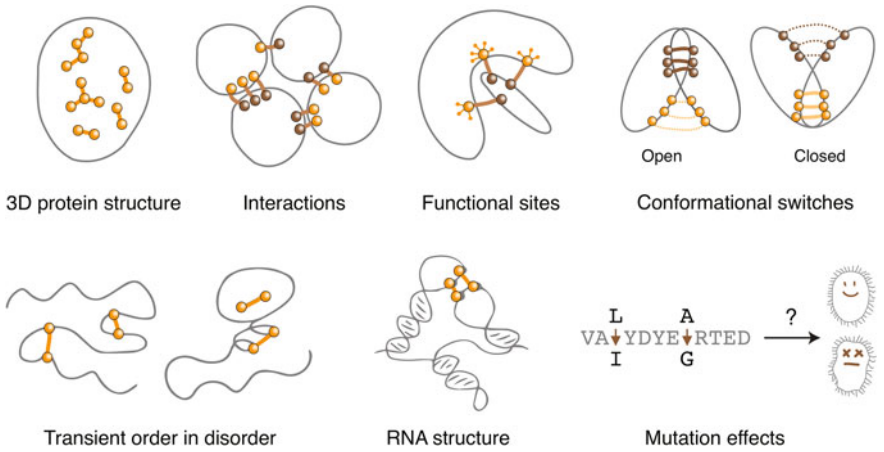


**Fig. 2.2** Applications of evolutionary couplings to predict protein structure and function. Evolutionary couplings allow to predict diverse aspects of protein structure and function that are defined by evolutionarily constrained interactions between residues, including the structures of monomers and complexes and changes in conformation. The approach can also be readily applied to other types of biomolecules, such as RNA, and used to quantify the phenotypic consequences of mutations with explicit modeling of epistatic interactions to the rest of the sequence (adapted from Marks et al. 2012)

$$P(\sigma) = \frac{1}{Z}\exp\left(\sum_{i=1}^{N}\mathbf{h}_i(\sigma_i) + \sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\mathbf{J}_{ij}(\sigma_i,\sigma_j)\right)$$

The model has two types of parameters that describe the constraint on acceptable amino acid configurations $\sigma_i$ and $\sigma_j$ at sites $i$ and $j$: bias terms $h_i$ (single-site conservation) and pair couplings $J_{ij}$ (co-conservation between pairs of sites i, j). Each variable $\sigma_i$ can assume one of the 20 amino acids as a value (most existing approaches treat gaps in the alignment as an additional 21st character, unless modelled as missing data). The *partition function* $Z$ is defined as

$$Z = \sum_{\sigma}\exp\left(\sum_{i=1}^{N}\mathbf{h}_i(\sigma_i) + \sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\mathbf{J}_{ij}(\sigma_i,\sigma_j)\right)$$

It sums over all possible $21^N$ sequences $\sigma = (\sigma_1, \ldots, \sigma_N)$ of length $N$ and ensures that $P(\sigma)$ is a valid probability distribution. Due to the exponential number of summations, calculating $Z$ is intractable for our application domain and we use a method that approximates Z using a factorization (see below).

To identify evolutionary constraints from an alignment, the inverse problem of inferring the model parameters from sequences has to be solved. Once the parameters are inferred, the pair couplings $J_{ij}$ can be used to quantify the strength of evolutionary coupling between pairs of sites $i$ and $j$.

***Parameter inference***. All the widely used current methods use an approximation to maximum likelihood estimation, which finds the set of parameters that maximizes the probability of observing the data. For the pairwise probability model defined above and a sequence alignment $\Sigma$ with sequences $\sigma$, the likelihood function $\mathcal{L}(\mathbf{h}, \mathbf{J})$ of the model parameters $h$ and $J$ is given by

$$\begin{aligned}\mathcal{L}(\mathbf{h}, \mathbf{J}) =& P(\Sigma|\mathbf{h}, \mathbf{J}) = \prod_{\sigma \in \Sigma} P(\sigma|\mathbf{h}, \mathbf{J}) \\ =& \prod_{\sigma \in \Sigma} \frac{1}{Z(\mathbf{h}, \mathbf{J})} \exp\left(\sum_{i=1}^{N} \mathbf{h}_i(\sigma_i) + \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \mathbf{J}_{ij}(\sigma_i, \sigma_j)\right)\end{aligned}$$

However, since straightforward calculation of the likelihood function is prohibited by the intractability of *Z(h,J)*, several approaches have been taken to approximate parameter inference. These include gradient ascent with Monte Carlo sampling (Lapedes et al. 2012), message passing (Weigt et al. 2009) and mean-field (Marks et al. 2011; Morcos et al. 2011; Michel et al. 2014; Jones et al. 2012; Stein et al. 2015), but most current applications use pseudo-likelihood approximations to the full likelihood (Besag 1975; Balakrishnan et al. 2011; Ekeberg et al. 2013; Kamisetty et al. 2013; Michel et al. 2014; Hopf et al. 2015a, b; 2014; Toth-Petroczy et al. 2016; Weinreb et al. 2016; Ovchinnikov et al. 2014, 2015).

When adopting the pseudo-likelihood maximization (PLM) approach, the full likelihood for each sequence $\sigma = (\sigma_1, \ldots, \sigma_n)$ is approximated by a product of conditional likelihoods for each site $i$, i.e.

$$P(\sigma_1, \ldots, \sigma_N|\mathbf{h}, \mathbf{J}) \approx \prod_{i=1}^{N} P(\sigma_i|\sigma \backslash \sigma_i, \mathbf{h}, \mathbf{J})$$

The conditioning of the probability to observe a selected amino acid $\sigma_i$ in site $i$ on the rest of the sequence $(\sigma \setminus \sigma_i)$ leads to the cancellation of the global partition function *Z(h,J)*. Instead, the pseudo-likelihood normalizes locally over all possible 21 amino acid configurations at each site $i$. This factorization of the full likelihood function reduces the computational complexity of the parameter inference from $O(21^N)$ to $O(|\Sigma|N^2)$. The set of parameters minimizing the pseudo-likelihood is identified using standard iterative optimization algorithms.

***Regularization***. In addition, all published methods use some form of regularization to avoid overfitting to the data, as there are orders of magnitude more parameters in the model than there are effectively-independent samples (Number of parameters = $N$ ($N$-1)/2 (q-1)$^2$ + $N$($q$-1) for protein length $N$ and $q$ = 21 amino acid states). For example, the model has approximately $2*10^6$ parameters for a protein of length $N$ = 100 whereas most protein families only contain $10^2$ to $10^5$ effective (i.e. redundancy-reduced) sequences. This gap increases quadratically as the protein length $N$ increases. The EVcouplings method and others (Kamisetty et al. 2013) typically employ parameter type-specific $l_2$-regularization (equivalent to a Gaussian prior) while the mean-field methods uses pseudocounts (Marks et al. 2011; Morcos

et al. 2011) and sparse inverse covariance method uses $l_1$ (Jones et al. 2012). Finally, since the phylogenetic relationships between sequences mean that they are not independent and identically distributed, most methods for computing evolutionary couplings methods address the issue by sequence reweighting schemes (Weigt et al. 2009; Marks et al. 2011; Morcos et al. 2011; Ekeberg et al. 2013) and we expect this approach to be improved in the future to account more quantitatively for phylogenetic tree structure.

***Positional constraints from evolutionary couplings***

After inference, the coupling parameter matrices $J_{ij}$ contain the family-specific constraints on all 20 × 20 amino acid pair configurations $\sigma_i$ and $\sigma_j$ for each possible combination of positions $i$ and $j$. The last remaining step in the calculation of positional constraints from the evolutionary couplings between pairs of sites is to summarize the $20^2$ numbers in each $J_{ij}$ matrix into a single number that quantifies the total coupling for pair *(i, j)*. The preferred method for this summary statistic is the Frobenius norm.

Of each coupling matrix $J_{ij}$ (after first centring the means of rows and columns around zero, $J'_{ij}$)

$$\mathbf{J}'_{ij}(k,l) = \mathbf{J}_{ij}(k,l) - \mathbf{J}_{ij}(\cdot,l) - \mathbf{J}_{ij}(k,\cdot) + \mathbf{J}_{ij}(\cdot,\cdot)$$

where · means average across these entries,

$$FN(i,j) = \left\|\mathbf{J}_{ij}\right\|_2 = \sqrt{\sum_k \sum_l \mathbf{J}'_{ij}(k,l)^2}$$

which sums across all $21^2$ amino acid combinations *k, l*.

Since the $J_{ij}$ parameters summarized in the *FN* matrix are confounded by factors such as finite sampling and phylogenetic relationships between samples, the empirically derived *average product correction* (APC) is applied to the *FN* matrix to remove background coupling that arises due to noise (Dunn et al. 2008; Jones et al. 2012; Ekeberg et al. 2013). The correction assumes that, on average, each site should only have couplings to a limited subset of all sites. For each site pair *(i, j)*, the APC therefore approximates the noise (background coupling of both sites) with the product of the row and column averages of the FN score matrix (•) and subtracts these from the raw pair scores *FN(i, j)*:

$$EC(i,j) = FN(i,j) - \frac{FN(i,\cdot)FN(\cdot,j)}{FN(\cdot,\cdot)}$$

The final result after applying the correction is the symmetric *N* x *N* evolutionary coupling score matrix (*N* = length of protein). Each entry *EC(i, j)* estimates the strength of evolutionary coupling between a pair of sites *(i, j)*: larger positive values indicate strong evolutionary co-constraints; values around zero indicate that the model could not detect any coupling. The most significant evolutionary couplings can then be selected based on the shape of the score distribution by estimating the

degree to which each pair score is an outlier (Hopf et al. 2014; Ovchinnikov et al. 2014; Toth-Petroczy et al. 2016).

## 2.3 Three-Dimensional Protein Structures from Evolutionary Couplings

Starting from evolutionary couplings inferred from sequence alignments of protein families, one could then test if the couplings provide sufficient information to predict the 3D structure of proteins (Fig. 2.3a). The first publication on proteins folded with evolutionary couplings was using the EVfold method in 2011, and included a diverse set of proteins from 15 families (Marks et al. 2011). The resulting computed 3D structures were typically within 3–5 Å Cα-RMSD from the known experimental structures of these proteins. To our knowledge, this was the first time longer proteins, including some with more than 200 residues, had been folded without comparative modelling, fragments or known long-range contacts to anywhere near this degree of accuracy. Initially, the approach for computing couplings from the sequence alignment was based on a mean field approximation to find the parameters of the maximum entropy model, which was later updated to the more accurate PLM method described above. 3D structures were generated from evolutionary couplings using standard NMR distance geometry and simulated annealing software that use only little compute time, as the number of generated candidate models was only approx. 200–400 per protein. Simple geometric rules were then used to rank the prediction candidates and choose the most favoured models.

Many other groups have since used this or similar approaches to predict accurate long-range contacts from sequences, benchmarking against known contacts in observed 3D structures; such accurate predictions are typically available for thousands of families (Hopf et al. 2012; Michel et al. 2014; Kosciolek and Jones 2014; Ovchinnikov et al. 2015; Toth-Petroczy et al. 2016). The available methods choose different ways of thresholding the number of predicted couplings they display in contact maps and number of couplings used for structure prediction, but overall their strategies and outputs are very similar. Many of the observed differences are just as likely due to different input alignments as they are to do with the algorithms for inferring the couplings. The webservers of EVfold, PSICOV and GREMLIN provide downloads of coupling files that can be used to define restraints for the folding software of your choice. Of the available methods, to date only EVfold will fold on demand for particular sequences of interest, though other methods offer precomputed structures for a limited set of protein families (see Table 2.1 for an overview of available webservers and Table 2.2 for standalone evolutionary couplings software).

To assess the utility of evolutionary couplings for structure prediction, it is important to distinguish between predicting residue contacts and folding the protein. It is possible to have quite accurate residue contact predictions when

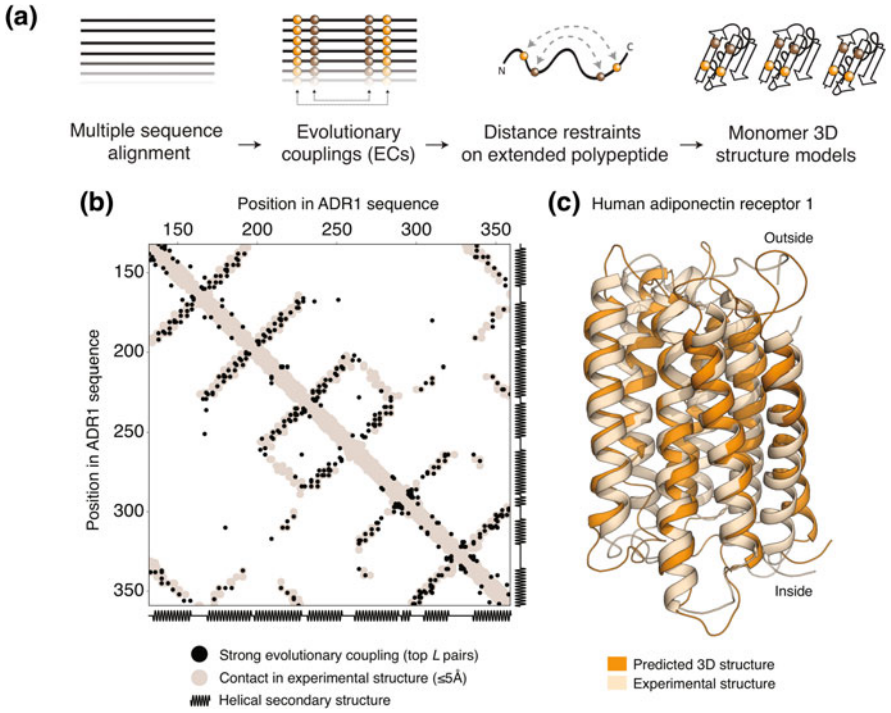


**Fig. 2.3** Protein 3D structure predicted from evolutionary sequences. **a** The 3D structure of a protein can be predicted from a multiple sequence alignment of the protein family by calculating evolutionary couplings between pairs of sites using a global probability model of the sequences. Assuming that residue pairs with strong couplings are close in 3D, the structure can then be computed by restraining the distances of these pairs in an extended polypeptide (adapted from Hopf 2016, 2015b; Marks et al. 2012) **b** Evolutionary couplings (*black dots*) for the human adiponectin receptor 1 (ADR1) largely correspond to residue contacts in the experimental 3D structure (*light brown dots*, precisions of 0.49 (5Å distance cutoff) and 0.77 (8Å cutoff), PDB 3wxv). **c** Models generated by EC-based 3D structure prediction (*dark orange* cartoon, best model) show good agreement with the experimental structure of ADR1 (*pale orange* cartoon, 2.4 Å Cα-RMSD over 192 residues, PDB 3wxv)

comparing evolutionary couplings to experimental structures, and still one may not be able to successfully fold the protein. For instance, predicted contacts may be clustered in one area of the protein, or only local along the chain and therefore missing key long-range contacts that define the overall topology of the molecule, such as contacts connecting the N- and C-termini. Only folding is therefore a definitive test if the computed evolutionary couplings contain sufficient information about the 3D structure of the protein.

While evolutionary couplings give valuable information about the 3D conformation of proteins, they also provide information over and above structure, such as functional residues that are particularly enriched for couplings with other residues (Fig. 2.2). Examples for strong coupling in functional sites include the active site of

trypsin, or the ligand binding pocket of the GPCR rhodopsin, where Lys-296 binds the retinal cofactor and has several strong couplings to other residues (Marks et al. 2011; Hopf et al. 2012). While it may be possible to identify some of these residues by single-site conservation alone, others may appear less conserved, and couplings offer the advantage of identifying the relevant interaction partners.

### 2.3.1 Transmembrane Proteins

Transmembrane proteins are of special biological interest as they mediate information transfer and molecule exchange across the cellular membranes in all forms of life, but are especially challenging to investigate experimentally when compared to globular proteins (see also Chap. 5). Given the resulting lack of experimental structures for the majority of membrane proteins, the most natural leverage of the evolutionary couplings approach was to predict their 3D structures, especially for large multipass proteins of high biomedical interest.

The first work to do so predicted evolutionary couplings and 3D structure for over 40 large membrane proteins, 25 of which were from families that had members with known structures and 18 of which were de novo predictions for families without any structure (Hopf et al. 2012). The blindly predicted structures on the test set of 25 proteins could be compared to known 3D coordinates and resulted in 3–6Å Cα-RMSD over at least 80% of the membrane domain. In similar work, the prediction of a test set of 28 proteins resulted in TM scores of at least 0.5 for most proteins (Jones et al. 2012). More recently, we updated various components of the EVfold prediction pipeline, including sequence alignment generation and inference of evolutionary couplings using PLM. Together with the increased number of sequences since the original publication in 2012, this leads to significant increases in prediction accuracy compared to the original method (average TM score increase of 0.08 on set of 25 proteins, highest TM score 0.82). We expect prediction accuracy to continue improving in the future as more sequences become available and better methods for folding are implemented.

For several examples from our set of de novo predictions, experimental structures have been published since. In general, our predictions show reasonable agreement with the experiment and have identified the correct overall 3D topology (TM score $\geq 0.5$) (Hopf 2016). Amongst these examples, the experimental structures confirmed that we correctly predicted the structural similarity of the unsolved complex 1 subunit 1 (MT-ND1) to the other subunits of the complex despite no detectable homology on the sequence level (Baradaran et al. 2013). We also correctly predicted the fold of the human adiponectin receptor 1 (Fig. 2.3) (TM score 0.69 from model in 2012, TM score = 0.79 in 2016), and successfully identified the cluster of activate site residues on the cytoplasmic side of the membrane (Tanabe et al. 2015). Both cases highlight the predictive power of evolutionary couplings to study the structure and function of proteins with limited experimental data.

### 2.3.2 Protein Interactions and Complexes

The coevolution of interacting residues is not only necessary to maintain the 3D structures of individual proteins, but also to maintain protein interactions and complexes. Based on this premise, others and we developed a general method for computing evolutionary couplings between proteins. The largest scale results identified interacting residues for over 50 protein interactions and the resulting 3D structure for a subset (Hopf et al. 2014; Ovchinnikov et al. 2014) (Fig. 2.4a, Tables 2.1 and 2.2) and many others have now computed a more limited number of interactions that often concentrate on disentangling paralog pairs of histidine kinase and response regulators (Cheng et al. 2016; Boyd et al. 2016; Feinauer et al. 2016; Bitbol et al. 2016; Gueudre et al. 2016).

For those methods with general applicability, the approaches are very similar. First, one must pair the sequences of putatively interacting proteins within each species to create a concatenated sequence alignment of the complex. Second, one computes both the couplings within (intra-protein evolutionary couplings) and between (inter-protein evolutionary couplings) the subunits simultaneously. This way, both the individual proteins can be predicted as well as the complex, using the inter-protein couplings as restraints in a docking protocol. Both EVcomplex and GREMLIN compute the couplings using pseudo-likelihood maximization, and differ only in their alignments, ranking and docking protocols.

However, the scope of both methods is currently limited by the generation of correctly paired sequence alignments that have sufficient sequence diversity. Correctly pairing the sequences when there are paralogs in a species depends on being able to identify the correct interacting proteins. Both EVcomplex and GREMLIN use the observation that interacting proteins are often encoded on the same operon. We have estimated that this excludes 80% of interacting proteins from EC-based prediction, even in *E. coli*. More recent approaches are being developed that aim to solve this issue, but their general applicability outside of a couple of systems still has to be demonstrated.

A second more pernicious assumption of this approach is that that the interactions, as well as the proteins themselves are conserved across evolution. While this may be a reasonable assumption for the components of ATP synthase, how conserved interactions are may be unknown for a large number of protein pairs. We expect to see significant algorithmic developments in this area so that the models can be used to ask the question rather than assume the answer.

Nevertheless, evolutionary couplings from sequence variation allow to predict protein interactions at residue level resolution not possible before (Fig. 2.4b), including the 3D structures of complexes that had not been solved experimentally at the time but whose subsequent characterisation confirmed the accuracy of the approach (e.g. DinJ-YafQ toxin-antitoxin interaction) (Hopf et al. 2014).

Both EVcomplex and GREMLIN also show that one can predict whether or not two proteins in a subunit interact physically, given sufficient sequence diversity and confidence in the matched alignment. In the case of the ATP synthase complex, we

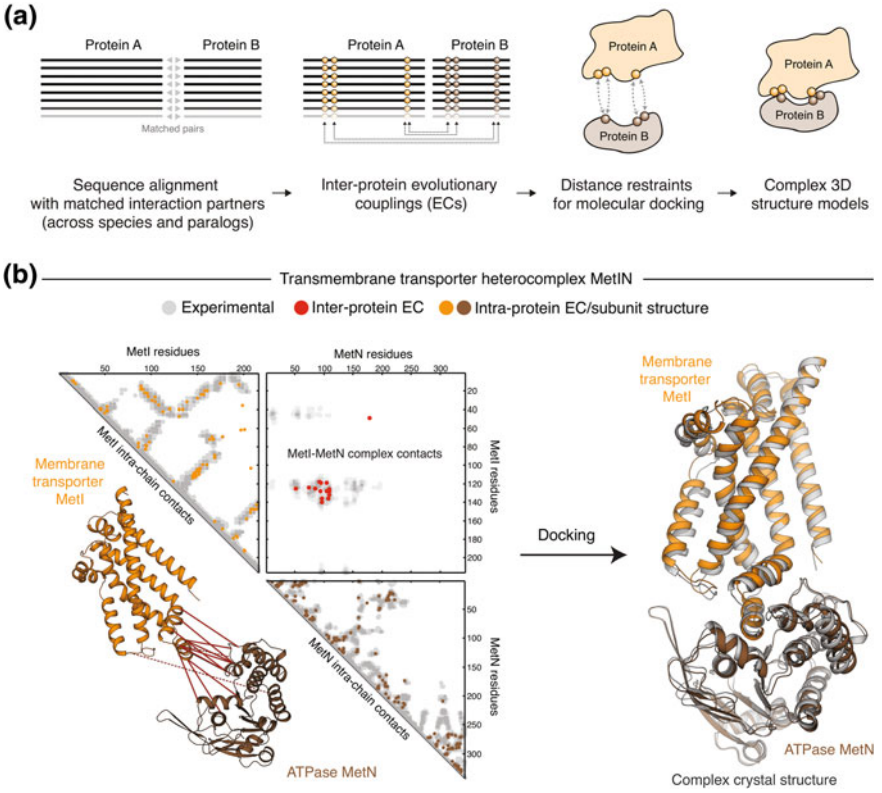


**Fig. 2.4** Protein interactions at residue level detail from evolutionary couplings**. a** Evolutionary couplings across interacting proteins can be calculated by generating a concatenated sequence alignment, where putatively interacting sequences within each species are matched with each other. Assuming coevolution due to structural proximity, the 3D structure of the complex can then be predicted from the monomer structures by docking with distance restraints on the strongly coupled pairs. **b** *Left* Evolutionary couplings (*coloured dots*) in the ABC transmembrane transporter MetIN correspond to structurally proximal residue pairs (*dark/medium/light grey dots* at 5/8/12Å distance cutoffs, PDB 3tui) both in the monomer structures (intra-protein ECs, *triangle* contact maps) as well as between the interacting subunits (inter-protein ECs, *square* contact map). The inter-protein ECs define the structural interaction between both subunits (*red lines* between *orange* and *brown* cartoons). *Right* Docking of the monomer structures (*orange/brown* cartoons) using significant inter-protein ECs leads to an accurate model of the complex (*grey cartoon*, 1.5Å interface-RMSD, PDB 3tui). (Figure adapted from Hopf 2016, 2014)

correctly identified 24 of 28 interactions with only 2 false positives and two interactions that are experimentally ambiguous. Similarly, GREMLIN correctly identified 12/23 interacting protein pairs in the ribosomal 50S subunit. The missing predictions (false negatives) may arise because the models are wrong, or, just as plausibly, the interactions could be weaker and a consequence of constraints between other subunits in the complex. Finally, recent work has also highlighted that evolutionary couplings can be applied to accurately predict the 3D structure of

RNA as well as protein-RNA interactions, in ribosomal complexes and RNaseP, from sequences alone (Weinreb et al. 2016).

### 2.3.3 Conformational Plasticity and Disordered Proteins

Many, if not most proteins may be structurally flexible, with conformational plasticity ranging from simple hinge movements or open-closed conformational switching to ordered stable structures that occur only upon binding or in the appropriate environment. Indeed, it may be the case that even protein segments that are considered highly flexible, such as histone tails, may take on a defined 3D structure in some functional states. Around half of human proteins contain substantially sized regions whose amino acid sequence is considered to indicate structural 'disorder', sometimes called 'intrinsic disorder'(van der Lee et al. 2014; Oates et al. 2013) (see also Chap. 6). These regions can range from 30 amino acid long insertions to longer regions of many hundreds of amino acids that are often present on transcription and translation factors.

Early work on evolutionary couplings showed that these methods will capture contacts from alternative 3D conformation, as demonstrated by the identification of couplings corresponding to open and closed conformations of the glycerol-3-phosphate transporter GlpT (Hopf et al. 2012) and the L-leucine binding protein (Morcos et al. 2013). More recently, this has been explored systematically with another 38 proteins known to have alternative conformations and differential contacts, demonstrating not only fold rearrangement but also, sometimes, secondary structure switching (Toth-Petroczy et al. 2016).

This recent work has extended the exploration of conformational states to proteins considered disordered. Since a small number of disordered proteins are known to become ordered in specific environments and have been captured experimentally, this gave the opportunity to investigate whether evolutionary couplings methods can detect these 3D states. After a number of methodological improvements, including iterative testing for alignment robustness, evolutionary couplings were computed to determine the potential of these proteins forming long-range contacts and secondary structure. In 40 of the 45 cases contacts were successfully predicted for known "order upon binding', including the well-known cyclin inhibitor p27 when it binds the Cyclin A-Cdk2 complex. Importantly, the method also found very little evidence of structural constraints for proteins such as the C-terminal tail of Histone H1 that had multiple lines of evidence for lack of structure (Toth-Petroczy et al. 2016). Hence, the true positive predictions for proteins with ordered conformations do not seem to be at the expense of false positives in proteins without ordered conformations.

To explore the structural potential of apparently disordered regions for which there is currently no experimental information on a proteome-wide scale, Toth-Petroczy et al. systematically surveyed all regions in the human proteome of more than 100 amino acids in length where alignments could be constructed (about

25% of all regions). This analysis resulted in predictions for ~1000 protein regions, of which 40% showed signal for some long-range structure and another 40% secondary structure. The predicted contact maps revealed that some of these disordered domains resembled zinc finger and RNA-binding domains, which could not be identified from their primary sequence (the data from this analysis is available from http://marks.hms.harvard.edu/disorder).

## 2.4 Predicting the Effect of Mutations

A major challenge in biology is being able to predict the functional effects of mutations on phenotype or fitness. New work has shown that the global statistical models of sequences can also be used to predict the effects of mutations by quantifying the change of probabilities between the mutated protein and the wild type sequence ($\Delta E = \log P(\text{mutant}) / P(\text{wild type})$) (Hopf et al. 2017; Figliuzzi et al. 2016; Mann et al. 2014). This quantity $\Delta E$, called the statistical energy difference of a mutant, is computed by summing the changes in couplings and site amino acid preferences between all pairs of positions, to give a total score that describes the effect of any single or higher-order mutation. For instance, as illustrated in the cartoon protein in Fig. 2.5a, the substitution W3L leads to a change in 4 couplings and one single site bias term. Through the evaluation of couplings to other sites, the computation explicitly models the context dependence (or epistasis) of mutations. These interactions are typically neglected by approaches using single-site conservation to quantify the effects of mutations. It is important to note that this approach uses precisely the same statistical model (e.g. PLM or DCA) as one uses to compute residue contacts from the sequence alignment, but *does not depend on computing the structure*. This allows to infer epistatic mutational landscapes for any protein with enough sequence information (Figs. 2.5b, c).

To test the applicability of our implementation of the method, EVmutation, the predicted effects of mutations have been compared against thousands of variants assayed in high-throughput multiplexed mutational scans that have emerged over the last few years, providing a large pool of ground truth for evaluation (Deng et al. 2012; Jacquier et al. 2013; Stiffler et al. 2015; Melamed et al. 2013; 2015; Rockah-Shmuel et al. 2015; Starita et al. 2015; Roscoe and Bolon 2014; Starita et al. 2013; Li et al. 2016; Melnikov et al. 2014). Whilst the exact interpretation of $\Delta E$ effects is not clear a priori, one would expect them to be related to the 'fitness' of the protein sequence. For instance, $\Delta E$ of all single mutations to a bacterial DNA methylase correlated well with an experimental scan testing their effect on bacterial fitness (Spearman's rank correlation $\rho = 0.69$) (Rockah-Shmuel et al. 2015). Similarly, EVmutation effects showed significant correlations across a wide range of 34 experimental datasets for 21 proteins and a tRNA molecule (Hopf et al. 2017). The approach generalizes to any type of biological sequence, and could also be used to predict effects for protein-RNA complexes. Using epistatic interactions with other sites particularly contributed to improved prediction accuracy in functional

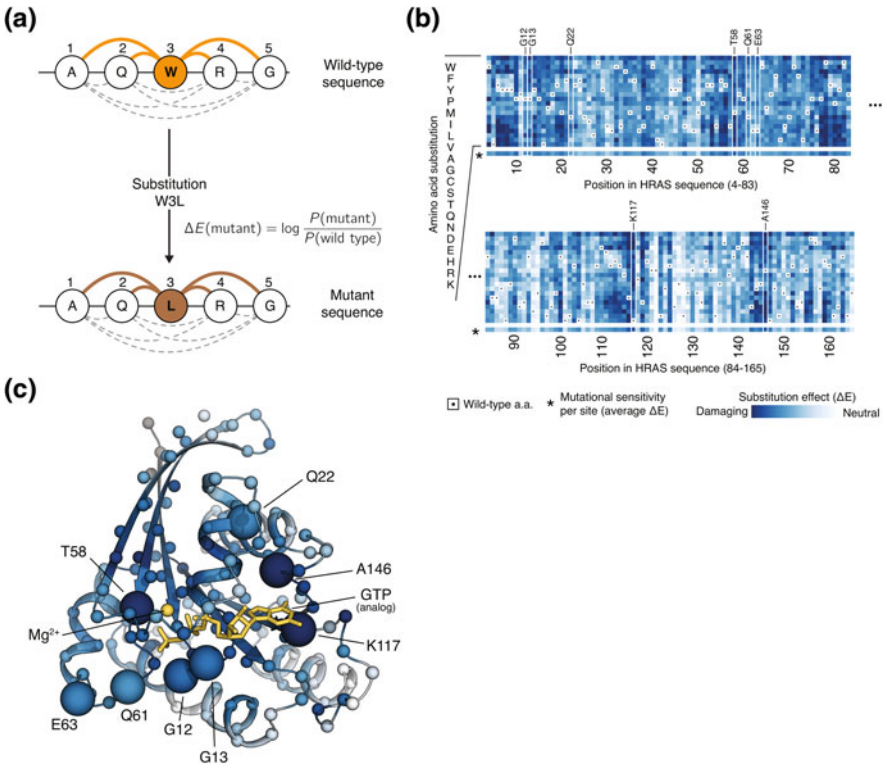


**Fig. 2.5** Prediction of mutation effects using an epistatic model of evolutionary sequences. **a** The global probability model of a protein family can be used to predict the effects of mutations by comparing the probabilities of the wild-type and mutant sequences. The calculation sums the differences in all couplings to mutated positions as well as the change in the single-site amino acid preference terms of the changed sites. Thereby, epistatic interactions with the sequence background are incorporated in the calculation (adapted from (Hopf 2016)) **b** Computed ΔE mutational landscape of the human disease gene HRAS (x-axis: position in HRAS sequence, y-axis: amino acid substitutions, *white boxes*: positions with known disease mutations). **c** Residues (small spheres) around the active site of RASH (GTP ligand analog, *yellow sticks*), including positions with known disease mutations (large spheres), are predicted as sensitive to mutation (colour scale as in (a) from *blue*/damaging to *white*/neutral)

sites, such as ligand binding and protein interaction interfaces, when compared to a model that only uses single-site conservation. When tested on human disease variants, ΔE separated them from neutral variants with similar or higher accuracy than state-of-the-art methods for variant effect classification without, however, being specifically trained on known variants for this problem (Hopf et al. 2017). This suggests that established machine learning methods could benefit from the inclusion of evolutionary statistical energies instead of positional sequence conservation.

## 2.5 Summary and Future Challenges

Over the last 5 years, approaches based on evolutionary couplings from sequence alignments have already shown their power in predicting structural constraints and 3D structures for proteins, RNA, their interactions, the potential structured states of disordered regions, as well as the effects of mutations on protein function. Readers would do well to use this chapter as a basis, but since the field will change rapidly in the next few years, they should be encouraged to search for more recent work than the snapshot presented here.

We expect to see an increase in hybrid approaches that combine evolutionary couplings with experimental methods to accelerate structure determination in such fields as cryoEM, NMR, crystallography or mass spectrometry. First promising work that demonstrates the power of this type of approach has already been published (Tang et al. 2015). Where refined 3D models are desired, there is a still a clear need for improving the structure prediction protocols, although some advances have been made here recently.

Notwithstanding the impressive impact these methods have already had, there are many challenges to be solved, not least with the probability model itself. First, an implicit assumption in the underlying model is that all sequences have been tried by evolution and the ones that we see now are the only possible functional ones, leading to many issues associated with inferring models from undersampled data. Whilst regularization during inference and heuristics for post hoc corrections address this problem somewhat, we expect advances in this area would be beneficial for more accurate models.

A second challenge for the emerging field is to develop improved criteria for assessing the quality of alignments, and the choice of alignment depth that is critically dependent on the research question being asked. If we did not know what the 3D structure of GPCRs looked like, then any family alignment however large and non-specific may be useful; on the other hand, if we want to explore the different ligand-binding pockets of the subfamilies we would need alignments that reflected that specificity. Similarly, for complexes and protein interactions the challenge is to assess the likelihood of interaction with the ambiguity that the interaction may not be conserved in all alignable sequences.

A third challenge is to blindly disambiguate evolutionary couplings that arise due to different aspects of protein function, including the blind assignment of couplings to different conformational states, or the distinction between intra- and inter-protein interactions in homomultimeric complexes in the absence of an experimental structure of the monomer.

All of these challenges are exciting questions for future research, and will help to further increase the usefulness of evolutionary couplings as a tool in exploring diverse aspects of protein structure and function.

## References

Balakrishnan S, Kamisetty H, Carbonell JG, Lee SI, Langmead CJ (2011) Learning generative models for protein fold families. Proteins 79(4):1061–1078. doi:10.1002/prot.22934

Baradaran R, Berrisford JM, Minhas GS, Sazanov LA (2013) Crystal structure of the entire respiratory complex I. Nature 494(7438):443–448. doi:10.1038/nature11871

Ben-Naim E, Lapedes AS (1999) Genetic correlations in mutation processes. Phys Rev E Stat Phys Plasmas Fluids 59(6):7000–7007

Besag J (1975) Statistical analysis of non-lattice data. Statistician 179–195

Bitbol AF, Dwyer RS, Colwell LJ, Wingreen NS (2016) Inferring interaction partners from protein sequences. Proc Natl Acad Sci USA 113(43):12180–12185. doi:10.1073/pnas.1606762113

Boyd JS, Cheng RR, Paddock ML, Sancar C, Morcos F, Golden SS (2016) A combined computational and genetic approach uncovers network interactions of the cyanobacterial circadian clock. J Bacteriol 198(18):2439–2447. doi:10.1128/JB.00235-16

Burger L, van Nimwegen E (2008) Accurate prediction of protein-protein interactions from sequence alignments using a Bayesian method. Molecular Syst biology 4:165. doi:10.1038/msb4100203

Burger L, van Nimwegen E (2010) Disentangling direct from indirect co-evolution of residues in protein alignments. PLoS Comput Biol 6(1):e1000633. doi:10.1371/journal.pcbi.1000633

Cheng RR, Nordesjo O, Hayes RL, Levine H, Flores SC, Onuchic JN, Morcos F (2016) Connecting the sequence-space of bacterial signaling proteins to phenotypes using coevolutionary landscapes. Mol Biol Evol. doi:10.1093/molbev/msw188

Deng Z, Huang W, Bakkalbasi E, Brown NG, Adamski CJ, Rice K, Muzny D, Gibbs RA, Palzkill T (2012) Deep sequencing of systematic combinatorial libraries reveals beta-lactamase sequence constraints at high resolution. J Mol Biol 424(3–4):150–167. doi:10.1016/j.jmb.2012.09.014

dos Santos RN, Morcos F, Jana B, Andricopulo AD, Onuchic JN (2015) Dimeric interactions and complex formation using direct coevolutionary couplings. Sci Rep 5:13652. doi:10.1038/srep13652

Ekeberg M, Lovkvist C, Lan Y, Weigt M, Aurell E (2013) Improved contact prediction in proteins: using pseudolikelihoods to infer Potts models. Phys Rev E Stat Nonlin Soft Matter Phys 87(1):012707

Feinauer C, Szurmant H, Weigt M, Pagnani A (2016) Inter-protein sequence co-evolution predicts known physical interactions in Bacterial Ribosomes and the Trp Operon. PLoS ONE 11(2):e0149166. doi:10.1371/journal.pone.0149166

Figliuzzi M, Jacquier H, Schug A, Tenaillon O, Weigt M (2016) Coevolutionary landscape inference and the context-dependence of mutations in Beta-Lactamase TEM-1. Mol Biol Evol 33(1):268–280. doi:10.1093/molbev/msv211

Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, Potter SC, Punta M, Qureshi M, Sangrador-Vegas A, Salazar GA, Tate J, Bateman A (2016) The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res 44(D1):279–285. doi:10.1093/nar/gkv1344

Giraud BG, Heumann JM, Lapedes AS (1999) Superadditive correlation. Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Topics 59 (5 Pt A):4983–4991

Gobel U, Sander C, Schneider R, Valencia A (1994) Correlated mutations and residue contacts in proteins. Proteins 18(4):309–317. doi:10.1002/prot.340180402

Gueudre T, Baldassi C, Zamparo M, Weigt M, Pagnani A (2016) Simultaneous identification of specifically interacting paralogs and interprotein contacts by direct coupling analysis. Proc Natl Acad Sci USA 113(43):12186–12191. doi:10.1073/pnas.1607570113

Gutell RR, Power A, Hertz GZ, Putz EJ, Stormo GD (1992) Identifying constraints on the higher-order structure of RNA: continued development and application of comparative sequence analysis methods. Nucleic Acids Res 20(21):5785–5795

Hopf T (2016) Phenotype prediction from evolutionary sequence covariation. München, Technische Universität München, Diss 2016

Hopf TA, Colwell LJ, Sheridan R, Rost B, Sander C, Marks DS (2012) Three-dimensional structures of membrane proteins from genomic sequencing. Cell 149(7):1607–1621. doi:10.1016/j.cell.2012.04.012

Hopf TA, Ingraham JB, Poelwijk FJ, Springer M, Sander C, Marks DS (2015a) Quantification of the effect of mutations using a global probability model of natural sequence variation. arXiv preprint arXiv:151004612

Hopf TA, Ingraham JI, Poelwijk FJ, Schärfe CPI, Springer M, Sander C, Marks DS (2017) Mutational effects captured by epistatic models of evolutionary sequence variation. Nat Biotech 35:128–135. doi:10.1038/nbt.3769

Hopf TA, Morinaga S, Ihara S, Touhara K, Marks DS, Benton R (2015b) Amino acid coevolution reveals three-dimensional structure and functional domains of insect odorant receptors. Nat Commun 6:6077. doi:10.1038/ncomms7077

Hopf TA, Schärfe CP, Rodrigues JP, Green AG, Kohlbacher O, Sander C, Bonvin AM, Marks DS (2014) Sequence co-evolution gives 3D contacts and structures of protein complexes. eLife 3. doi:10.7554/eLife.03430

Jacquier H, Birgy A, Le Nagard H, Mechulam Y, Schmitt E, Glodt J, Bercot B, Petit E, Poulain J, Barnaud G, Gros PA, Tenaillon O (2013) Capturing the mutational landscape of the beta-lactamase TEM-1. Proc Natl Acad Sci USA 110(32):13067–13072. doi:10.1073/pnas.1215206110

Jones DT, Buchan DW, Cozzetto D, Pontil M (2012) PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. Bioinformatics 28(2):184–190. doi:10.1093/bioinformatics/btr638

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

Kajan L, Hopf TA, Kalas M, Marks DS, Rost B (2014) FreeContact: fast and free software for protein contact prediction from residue co-evolution. BMC Bioinformatics 15:85. doi:10.1186/1471-2105-15-85

Kamisetty H, Ovchinnikov S, Baker D (2013) Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era. Proc Natl Acad Sci USA 110(39):15674–15679. doi:10.1073/pnas.1314045110

Koller D, Friedman N (2009) Probabilistic graphical models: principles and techniques. MIT press

Kosciolek T, Jones DT (2014) De novo structure prediction of globular proteins aided by sequence variation-derived contacts. PLoS ONE 9(3):e92197. doi:10.1371/journal.pone.0092197

Lapedes A, Giraud B, Jarzynski C (2012) Using sequence alignments to predict protein structure and stability with high accuracy. arXiv preprint arXiv:12072484

Lapedes AS, Giraud BG, Liu LC, Stormo GD (1997) Correlated Mutations in Protein Sequences: Phylogenetic and Structural Effects. Santa Fe Institute

Li C, Qian W, Maclean CJ, Zhang J (2016) The fitness landscape of a tRNA gene. Science. doi:10.1126/science.aae0568

Lindorff-Larsen K, Piana S, Dror RO, Shaw DE (2011) How fast-folding proteins fold. Science 334(6055):517–520. doi:10.1126/science.1208351

Mann JK, Barton JP, Ferguson AL, Omarjee S, Walker BD, Chakraborty A, Ndung'u T (2014) The fitness landscape of HIV-1 gag: advanced modeling approaches and validation of model predictions by in vitro testing. PLoS Comput Biol 10(8):e1003776. doi:10.1371/journal.pcbi.1003776

Marks DS, Colwell LJ, Sheridan R, Hopf TA, Pagnani A, Zecchina R, Sander C (2011) Protein 3D structure computed from evolutionary sequence variation. PLoS ONE 6(12):e28766. doi:10.1371/journal.pone.0028766

Marks DS, Hopf TA, Sander C (2012) Protein structure prediction from sequence variation. Nat Biotechnol 30(11):1072–1080. doi:10.1038/nbt.2419

Melamed D, Young DL, Gamble CE, Miller CR, Fields S (2013) Deep mutational scanning of an RRM domain of the Saccharomyces cerevisiae poly(A)-binding protein. RNA 19(11):1537–1551. doi:10.1261/rna.040709.113

Melamed D, Young DL, Miller CR, Fields S (2015) Combining natural sequence variation with high throughput mutational data to reveal protein interaction sites. PLoS Genet 11(2): e1004918. doi:10.1371/journal.pgen.1004918

Melnikov A, Rogov P, Wang L, Gnirke A, Mikkelsen TS (2014) Comprehensive mutational scanning of a kinase in vivo reveals substrate-dependent fitness landscapes. Nucleic Acids Res 42(14):e112. doi:10.1093/nar/gku511

Michel M, Hayat S, Skwark MJ, Sander C, Marks DS, Elofsson A (2014) PconsFold: improved contact predictions improve protein models. Bioinformatics 30(17):482–488. doi:10.1093/bioinformatics/btu458

Morcos F, Jana B, Hwa T, Onuchic JN (2013) Coevolutionary signals across protein lineages help capture multiple protein conformations. Proc Natl Acad Sci USA 110(51):20533–20538. doi:10.1073/pnas.1315625110

Morcos F, Pagnani A, Lunt B, Bertolino A, Marks DS, Sander C, Zecchina R, Onuchic JN, Hwa T, Weigt M (2011) Direct-coupling analysis of residue coevolution captures native contacts across many protein families. Proc Natl Acad Sci USA 108(49):1293–1301. doi:10.1073/pnas.1111471108

Mosca R, Ceol A, Stein A, Olivella R, Aloy P (2014) 3did: a catalog of domain-based interactions of known three-dimensional structure. Nucleic acids research 42 (Database issue): 374–379. doi:10.1093/nar/gkt887

Neher E (1994) How frequent are correlated changes in families of protein sequences? Proc Natl Acad Sci USA 91(1):98–102

Oates ME, Romero P, Ishida T, Ghalwash M, Mizianty MJ, Xue B, Dosztanyi Z, Uversky VN, Obradovic Z, Kurgan L, Dunker AK, Gough J (2013) D(2) P(2): database of disordered protein predictions. Nucleic acids research 41 (Database issue): 508–516. doi:10.1093/nar/gks1226

Ovchinnikov S, Kamisetty H, Baker D (2014) Robust and accurate prediction of residue-residue interactions across protein interfaces using evolutionary information. eLife 3: 02030. doi:10.7554/eLife.02030

Ovchinnikov S, Kinch L, Park H, Liao Y, Pei J, Kim DE, Kamisetty H, Grishin NV, Baker D (2015) Large-scale determination of previously unsolved protein structures using evolutionary information. eLife 4: 09248. doi:10.7554/eLife.09248

Pazos F, Helmer-Citterich M, Ausiello G, Valencia A (1997) Correlated mutations contain information about protein-protein interaction. J Mol Biol 271(4):511–523. doi:10.1006/jmbi.1997.1198

Perdigao N, Heinrich J, Stolte C, Sabir KS, Buckley MJ, Tabor B, Signal B, Gloss BS, Hammang CJ, Rost B, Schafferhans A, O'Donoghue SI (2015) Unexpected features of the dark proteome. Proc Natl Acad Sci USA 112(52):15898–15903. doi:10.1073/pnas.1508380112

Qian B, Raman S, Das R, Bradley P, McCoy AJ, Read RJ, Baker D (2007) High-resolution structure prediction and the crystallographic phase problem. Nature 450(7167):259–264. doi:10.1038/nature06249

Rajagopala SV, Sikorski P, Kumar A, Mosca R, Vlasblom J, Arnold R, Franca-Koh J, Pakala SB, Phanse S, Ceol A, Hauser R, Siszler G, Wuchty S, Emili A, Babu M, Aloy P, Pieper R, Uetz P (2014) The binary protein-protein interaction landscape of Escherichia coli. Nat Biotechnol 32 (3):285–290. doi:10.1038/nbt.2831

Rockah-Shmuel L, Toth-Petroczy A, Tawfik DS (2015) Systematic mapping of protein mutational space by prolonged drift reveals the deleterious effects of seemingly neutral mutations. PLoS Comput Biol 11(8):e1004421. doi:10.1371/journal.pcbi.1004421

Roscoe BP, Bolon DN (2014) Systematic exploration of ubiquitin sequence, E1 activation efficiency, and experimental fitness in yeast. J Mol Biol 426(15):2854–2870. doi:10.1016/j.jmb.2014.05.019

Seemayer S, Gruber M, Soding J (2014) CCMpred–fast and precise prediction of protein residue-residue contacts from correlated mutations. Bioinformatics 30(21):3128–3130. doi:10.1093/bioinformatics/btu500

Shindyalov IN, Kolchanov NA, Sander C (1994) Can three-dimensional contacts in protein structures be predicted by analysis of correlated mutations? Protein Eng 7(3):349–358

Skerker JM, Perchuk BS, Siryaporn A, Lubin EA, Ashenberg O, Goulian M, Laub MT (2008) Rewiring the specificity of two-component signal transduction systems. Cell 133(6):1043–1054. doi:10.1016/j.cell.2008.04.040

Starita LM, Pruneda JN, Lo RS, Fowler DM, Kim HJ, Hiatt JB, Shendure J, Brzovic PS, Fields S, Klevit RE (2013) Activity-enhancing mutations in an E3 ubiquitin ligase identified by high-throughput mutagenesis. Proc Natl Acad Sci USA 110(14):1263–1272. doi:10.1073/pnas.1303309110

Starita LM, Young DL, Islam M, Kitzman JO, Gullingsrud J, Hause RJ, Fowler DM, Parvin JD, Shendure J, Fields S (2015) Massively Parallel Functional Analysis of BRCA1 RING Domain Variants. Genetics. doi:10.1534/genetics.115.175802

Stein RR, Marks DS, Sander C (2015) Inferring pairwise interactions from biological data using maximum-entropy probability models. PLoS Comput Biol 11(7):e1004182. doi:10.1371/journal.pcbi.1004182

Stiffler MA, Hekstra DR, Ranganathan R (2015) Evolvability as a function of purifying selection in TEM-1 beta-Lactamase. Cell 160(5):882–892. doi:10.1016/j.cell.2015.01.035

Sulkowska JI, Morcos F, Weigt M, Hwa T, Onuchic JN (2012) Genomics-aided structure prediction. Proc Natl Acad Sci USA 109(26):10340–10345. doi:10.1073/pnas.1207864109

Tanabe H, Fujii Y, Okada-Iwabu M, Iwabu M, Nakamura Y, Hosaka T, Motoyama K, Ikeda M, Wakiyama M, Terada T, Ohsawa N, Hato M, Ogasawara S, Hino T, Murata T, Iwata S, Hirata K, Kawano Y, Yamamoto M, Kimura-Someya T, Shirouzu M, Yamauchi T, Kadowaki T, Yokoyama S (2015) Crystal structures of the human adiponectin receptors. Nature 520(7547):312–316. doi:10.1038/nature14301

Tang Y, Huang YJ, Hopf TA, Sander C, Marks DS, Montelione GT (2015) Protein structure determination by combining sparse NMR data with evolutionary couplings. Nat Methods 12(8):751–754. doi:10.1038/nmeth.3455

Toth-Petroczy A, Palmedo P, Ingraham J, Hopf TA, Berger B, Sander C, Marks DS (2016) Structured states of disordered proteins from genomic sequences. cell 167 (1):158–170 e112. doi:10.1016/j.cell.2016.09.010

van der Lee R, Buljan M, Lang B, Weatheritt RJ, Daughdrill GW, Dunker AK, Fuxreiter M, Gough J, Gsponer J, Jones DT, Kim PM, Kriwacki RW, Oldfield CJ, Pappu RV, Tompa P, Uversky VN, Wright PE, Babu MM (2014) Classification of intrinsically disordered regions and proteins. Chem Rev 114(13):6589–6631. doi:10.1021/cr400525m

Webb B, Sali A (2014) Comparative Protein Structure Modeling Using MODELLER. Curr Protoc Bioinformatics 47:5 6 1–32. doi:10.1002/0471250953.bi0506s47

Weigt M, White RA, Szurmant H, Hoch JA, Hwa T (2009) Identification of direct residue contacts in protein-protein interaction by message passing. Proc Natl Acad Sci USA 106(1):67–72. doi:10.1073/pnas.0805923106

Weinreb C, Riesselman AJ, Ingraham JB, Gross T, Sander C, Marks DS (2016) 3D RNA and Functional Interactions from Evolutionary Couplings. Cell 165(4):963–975. doi:10.1016/j.cell.2016.03.030

http://www.springer.com/978-94-024-1067-9

From Protein Structure to Function with Bioinformatics

J. Rigden, D. (Ed.)
2017, XV, 503 p. 86 illus., 75 illus. in color., Hardcover
ISBN: 978-94-024-1067-9