Chapter 9
Prediction of Structures and Interactions from Genome Information

Sanzo Miyazawa

Abstract Predicting three dimensional residue-residue contacts from evolutionary information in protein sequences was attempted already in the early 1990s. However, contact prediction accuracies of methods evaluated in CASP experiments before CASP11 remained quite low, typically with <20% true positives. Recently, contact prediction has been significantly improved to the level that an accurate three dimensional model of a large protein can be generated on the basis of predicted contacts. This improvement was attained by disentangling direct from indirect correlations in amino acid covariations or cosubstitutions between sites in protein evolution. Here, we review statistical methods for extracting causative correlations and various approaches to describe protein structure, complex, and flexibility based on predicted contacts.

Keywords Contact prediction · Direct coupling · Amino acid covariation · Amino acid cosubstitution · Partial correlation · Maximum entropy model · Inverse Potts model · Markov random field · Boltzmann machine · Deep neural network

9.1 Introduction

The evolutionary history of protein sequences is a valuable source of information in many fields of science not only in evolutionary biology but even to understand protein structures. Residue-residue interactions that fold a protein into a unique three-dimensional (3D) structure and make it play a specific function impose structural and functional constraints in varying degrees on each amino acid. Selective constraints on amino acids are recorded in amino acid orders in homologous protein sequences and also in the evolutionary trace of amino acid substitutions. Negative

S. Miyazawa (✉)
Gunma University, Kiryu, Japan

H. Nakamura et al. (eds.), *Integrative Structural Biology with Hybrid Methods*,
Advances in Experimental Medicine and Biology 1105,
https://doi.org/10.1007/978-981-13-2200-6_9

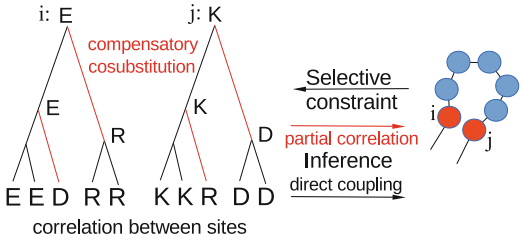

Fig. 9.1 Amino acids at sites i and j in a MSA are shown with a phylogenetic tree. Causative correlations between sites in protein evolution are extracted from the MSA or phylogenetic tree, and utilized to infer close residue pairs

effects caused by mutations at one site must be compensated by successive mutations at other sites (Yanovsky et al. 1964; Fitch and Markowitz 1970; Maisnier-Patin and Andersson 2004), causing covariations/cosubstitutions/coevolution between sites (Tufféry and Darlu 2000; Fleishman et al. 2004; Dutheil et al. 2005; Dutheil and Galtier 2007), otherwise most negative mutants will be eliminated from a gene pool and never reach fixation in population. Such structural and functional constraints arise from interactions between sites mostly in close spatial proximity. Thus, it has been suggested and also shown that the types of amino acids (Lapedes et al. 1999, 2002, 2012; Russ et al. 2005; Skerker et al. 2008; Burger and van Nimwegen 2008; Weigt et al. 2009; Halabi et al. 2009; Burger and van Nimwegen 2010; Morcos et al. 2011; Marks et al. 2011) and amino acid substitutions (Altschuh et al. 1988; Göbel et al. 1994; Shindyalov et al. 1994; Pollock and Taylor 1997; Pollock et al. 1999; Atchley et al. 2000; Fariselli et al. 2001; Fodor and Aldrich 2004; Fleishman et al. 2004; Dutheil et al. 2005; Martin et al. 2005; Fares and Travers 2006; Doron-Faigenboim and Pupko 2007; Dutheil and Galtier 2007; Dunn et al. 2008; Poon et al. 2008; Dutheil 2012; Gulyás-Kovács 2012) are correlated between sites that are close in a protein 3D structure. However, until CASP11, contact prediction accuracy remained quite low, typically with $\leq$20% true positives for top-L/5 long-range contacts in free modeling targets (Kosciolek and Jones 2016); L denotes protein length. Recently contact prediction has been significantly improved to the level that an accurate three dimensional model of a large protein ($\simeq$250 residues) can be generated on the basis of predicted contacts (Moult et al. 2016). These improvements were attained primarily by disentangling direct from indirect correlations in amino acid covariations or cosubstitutions between sites in protein evolution, and secondarily by reducing phylogenetic biases in a multiple sequence alignment (MSA) or removing them on the basis of a phylogenetic tree; see Fig. 9.1.

Here, we review statistical methods for extracting causative correlations in amino acid covariations/cosubstitutions between sites, and various approaches to describe protein structure, complex and flexibility based on predicted contacts. Mathematical formulation of each statistical method is concisely described in the unified manner in an appendix, the full version of which will be found in the article (Miyazawa 2017a) submitted to arXiv.

9.2 Statistical Methods to Extract Causative Interactions Between Sites

The primary task to develop a robust method toward contact prediction is to detect causative correlations, which reflect evolutionary constraints, in amino acid covariations between sites in a multiple sequence alignment (MSA) or in amino acid cosubstitutions between sites in branches of a phylogenetic tree; see Table 9.1. The former was called direct coupling analysis (DCA) (Morcos et al. 2011).

Table 9.1 Statistical methods for disentangling direct from indirect correlations between sites

Category	
Method name	Method/algorithm
(A) Direct coupling analysis of amino acid covariations between sites in a MSA	
Boltzmann machine	Markov chain Monte Carlo to calculate marginal probabilities and gradient descent to estimate fields and couplings
CMI (Lapedes et al. 2012)	Boltzmann machine to estimate conditional mutual information
mpDCA (Weigt et al. 2009)	Message-passing algorithm to estimate marginal probabilities and gradient descent to estimate fields and couplings
mfDCA (Morcos et al. 2011; Marks et al. 2011)	Mean field approximation to estimate the partition function
PSICOV (Jones et al. 2012)	Graphical lasso (Gaussian approximation with an exponential prior) with a shrinkage method for a covariance matrix
GaussDCA (Baldassi et al. 2014)	A multivariate Gaussian model with a normal-inverse-Wishart prior
plmDCA (Ekeberg et al. 2013, 2014)	Pseudo-likelihood maximization with Gaussian priors (ℓ_2 regularizers)
GREMLIN (Balakrishnan et al. 2011; Kamisetty et al. 2013)	Pseudo-likelihood maximization with ℓ_1 regularization terms (Balakrishnan et al. 2011) or with Gaussian priors (Kamisetty et al. 2013) which depend on site pair
ACE (Cocco and Monasson 2011, 2012; Barton et al. 2016)	Adaptive cluster expansion of cross-entropy with Gaussian priors
Persistent VI & Fadeout	Variational inference with sparsity-inducing prior, horseshoe (Ingraham and Marks 2016)
Sutto et al. (2015)	Boltzmann machine with ℓ_2 regularization terms
DI (Taylor and Sadowski 2011)	Partial correlation of normalized mutual informations between sites
(B) Partial correlation analysis of amino acid cosubstitutions between sites in a phylogenetic tree	
pcSV (Miyazawa 2013)	Partial correlation coefficients of coevolutionary substitutions between sites within branches in a phylogenetic tree

9.2.1 *Direct Coupling Analysis for Amino Acid Covariations Between Sites in a Multiple Sequence Alignment*

The direct coupling analysis is based on the maximum entropy model for the distribution of protein sequences, which satisfies the observed statistics in a MSA.

9.2.1.1 Maximum Entropy Model for the Distribution of Protein Sequences

Let us consider probability distributions $P(\sigma)$ of amino acid sequences, $\sigma \equiv (\sigma_1, \ldots, \sigma_L)^T$ with $\sigma_i \in \{\text{amino acids, deletion}\}$, single-site and two-site marginal probabilities of which are equal to a given frequency $P_i(a_k)$ of amino acid a_k at each site i and a given frequency $P_{ij}(a_k, a_l)$ of amino acid pair (a_k, a_l) for site pair (i, j), respectively.

$$P(\sigma_i = a_k) \equiv \sum_{\sigma} P(\sigma)\delta_{\sigma_i a_k} = P_i(a_k) \tag{9.1}$$

$$P(\sigma_i = a_k, \sigma_j = a_l) \equiv \sum_{\sigma} P(\sigma)\delta_{\sigma_i a_k}\delta_{\sigma_j a_l} = P_{ij}(a_k, a_l) \tag{9.2}$$

where $a_k \in \{\text{amino acids, deletion}\}$, $k = 1, \ldots, q$, $q \equiv |\{\text{amino acids, deletion}\}| = 21$, $i, j = 1, \ldots, L$, and $\delta_{\sigma_i a_k}$ is the Kronecker delta. The distribution P_{ME} with the maximum entropy is

$$P_{\mathrm{ME}}(\sigma|h, J) \tag{9.3}$$

$$\begin{aligned} &= \arg\max_{P(\sigma)}[-\sum_{\sigma} P(\sigma)\log P(\sigma) + \lambda(\sum_{\sigma} P(\sigma) - 1) \\ &\quad + \sum_i [h_i(a_k)(\sum_{\sigma} P(\sigma)\delta_{\sigma_i a_k} - P_i(a_k))] \\ &\quad + \sum_i \sum_{j>i} [J_{ij}(a_k, a_l)(\sum_{\sigma} P(\sigma)\delta_{\sigma_i a_k}\delta_{\sigma_j a_l} - P_{ij}(a_k, a_l))]] = \frac{1}{Z} e^{-H_{\mathrm{Potts}}(\sigma|h,J)} \end{aligned} \tag{9.4}$$

where λ, $h_i(a_k)$, and $J_{ij}(a_k, a_l)$ are Lagrange multipliers, and a Hamiltonian H_{Potts}, which is called that of the Potts model for $q > 2$ (or the Ising model for $q = 2$), and a partition function Z are defined as

$$-H_{\mathrm{Potts}}(\sigma|h, J) = \sum_i h_i(\sigma_i) + \sum_{i<j} J_{ij}(\sigma_i, \sigma_j), \quad Z = \sum_{\sigma} e^{-H_{\mathrm{Potts}}(\sigma|h,J)} \tag{9.5}$$

where $h_i(a_k)$ and $J_{ij}(a_k, a_l)$ are interaction potentials called fields and couplings.

Although pairwise frequencies $P_{ij}(a_k, a_l)$ reflect not only direct but indirect correlations in amino acid covariations between sites, couplings $J_{ij}(a_k, a_l)$ reflect causative correlations only. Thus, it is essential to estimate fields and couplings from marginal probabilities. This model is called the inverse Potts model.

9.2.1.2 Log-Likelihood and Log-Posterior-Probability

Log-posterior-probability and log-likelihood for the Potts model are

$$\log P_{\text{post}}(h, J|\{\sigma\}) \propto \ell_{\text{Potts}}(\{P_i\}, \{P_{ij}\}|h, J) + \log P_0(h, J) \tag{9.6}$$

$$\ell_{\text{Potts}}(\{P_i\}, \{P_{ij}\}|h, J) = B \sum_{\sigma} P_{\text{obs}}(\sigma) \log P_{\text{ME}}(\sigma|h, J) \tag{9.7}$$

where $P_{\text{obs}}(\equiv \sum_{\tau=1}^{B} \delta_{\sigma\sigma^\tau}/B)$ is the observed distribution of σ specified with $\{P_i(a_k)\}$ and $\{P_{ij}(a_k, a_l)\}$, and B is the number of instances; sequences σ^τ are assumed here to be independently and identically distributed samples in sequence space. $P_0(h, J)$ is a prior probability of (h, J).

Let us define cross entropy (Cocco and Monasson 2012) as the negative log-posterior-probability per instance.

$$\begin{aligned} S_0(h, J|\{P_i\}, \{P_{ij}\}) &\propto -(\log P_{\text{post}}(h, J|\{\sigma\}))/B \\ &\equiv S_{\text{Potts}}(h, J|\{P_i\}, \{P_{ij}\}) + R(h, J) \end{aligned} \tag{9.8}$$

where the cross entropy S_{Potts}, which is the negative log-likelihood per instance for the Potts model, and the negative log-prior per instance R are defined as follows.

$$S_{\text{Potts}}(h, J|\{P_i\}, \{P_{ij}\}) \equiv -\ell_{\text{Potts}}(\{P_i\}, \{P_{ij}\}|h, J)/B \tag{9.9}$$

$$= \log Z(h, J) - \sum_i \sum_k h_i(a_k) P_i(a_k) - \sum_i \sum_k \sum_{j>i} \sum_l J_{ij}(a_k, a_l) P_{ij}(a_k, a_l) \tag{9.10}$$

$$R(h, J) \equiv -\log(P_0(h, J))/B \tag{9.11}$$

The maximum likelihood estimates of h and J, which minimize the cross entropy with $R = 0$, satisfy the following equations.

$$\frac{\partial \log Z(h, J)}{\partial h_i(a_k)} = P_i(a_k), \quad \frac{\partial \log Z(h, J)}{\partial J_{ij}(a_k, a_l)} = P_{ij}(a_k, a_l) \tag{9.12}$$

It is, however, hardly tractable to computationally evaluate the partition function $Z(h, J)$ for any reasonable system size as a function of h and J. Thus, approximate maximization of the log-likelihood or minimization of the cross entropy is needed to estimate h and J.

The minimum of the cross entropy with $R = 0$ for the Potts model is just the Legendre transform of $\log Z(h, J)$ from (h, J) to $(\{P_i\}, \{P_{ij}\})$, (Eq. 9.10), and is equal to the entropy of the Potts model satisfying Eqs. 9.1 and 9.2;

$$S_{\text{Potts}}(\{P_i\}, \{P_{ij}\}) \equiv \min_{h,J} S_{\text{Potts}}(h, J|\{P_i\}, \{P_{ij}\}) = \sum_{\sigma} -P(\sigma) \log P(\sigma) \quad (9.13)$$

The cross entropy $S_{\text{Potts}}(h, J|\{P_i\}, \{P_{ij}\})$ in Eq. 9.10 is invariant under a certain transformation of fields and couplings, $J_{ij}(a_k, a_l) \rightarrow J_{ij}(a_k, a_l) - J^1_{ij}(a_k) - J^1_{ji}(a_l) + J^0_{ij}, h_i(a_k) \rightarrow h_i(a_k) - h^0_i + \sum_{j \neq i} J^1_{ij}(a_k)$ for any $J^1_{ij}(a_k)$, J^0_{ij} and h^0_i. This gauge-invariance reduces the number of independent variables in the Potts model to $(q-1)L$ fields and $(q-1)L \times (q-1)L$ couplings.

A prior $P_0(h, J)$ yields regularization terms for h and J (Cocco and Monasson 2012). If a Gaussian distribution is employed for the prior, then it will yield ℓ_2 norm regularization terms. ℓ_1 norm regularization corresponds to the case of exponential priors. Given marginal probabilities, the estimates of fields and couplings are those minimizing the cross entropy.

$$(h, J) = \arg\min_{(h,J)} S_0(h, J|\{P_i\}, \{P_{ij}\}), \;\; S_0(\{P_i\}, \{P_{ij}\}) \equiv \min_{(h,J)} S_0(h, J|\{P_i\}, \{P_{ij}\}) \quad (9.14)$$

Since $S_0(\{P_i\}, \{P_{ij}\})$ is the Legendre transform of $(\log Z(h, j) + R(h, J))$ from (h, J) to $(\{P_i\}, \{P_{ij}\})$, these optimum h and J can also be calculated from

$$h_i(a_k) = -\frac{\partial S_0(\{P_i\}, \{P_{ij}\})}{\partial P_i(a_k)}, \quad J_{ij}(a_k, a_l) = -\frac{\partial S_0(\{P_i\}, \{P_{ij}\})}{\partial P_{ij}(a_k, a_l)} \quad (9.15)$$

In most methods for contact prediction, residue pairs are predicted as contacts in the decreasing order of score $(\mathcal{S}_{ij})$ calculated from fields $\{J_{ij}(a_k, a_l)|1 \leq k, l < q\}$; see Eq. 9.47.

9.2.1.3 Inverse Potts Model

The problem of inferring interactions from observations of instances has been studied as inverse statistical mechanics, particularly inverse Potts model for Eq. 9.4, in the filed of statistical physics, as a Markov random field, Markov network or undirected graphical model in the domain of physics, statistics and information science, and as Boltzmann machine in the field of machine learning.

The maximum-entropy approach to the prediction of residue-residue contacts toward protein structure prediction from residue covariation pattems was first described in 2002 by Lapedes and collaborators (Giraud et al. 1999; Lapedes et al. 1999, 2002, 2012). They estimated conditional mutual information (CMI), which was employed as a score for residue-residue contacts, for each site pair by Boltzmann leaning with Monte Carlo importance sampling to calculate equilibrium

averages and gradient descent to minimize the cross entropy and successfully predicted contacts for 11 small proteins.

Calculating marginal probabilities for given fields and couplings by Monte Carlo simulations in Boltzmann machine is very computationally intensive. To reduce a computational load, the message passing algorithm, which is exact for a tree topology of couplings but approximate for the present model, is employed instead in mpDCA (Weigt et al. 2009). Because even the message passing algorithm is too slow to be applied to a large-scale analysis across many protein families, the mean field approximation is employed in mfDCA (Morcos et al. 2011; Marks et al. 2011); $J^{MF} = -C^{-1}$, where $C_{ij}(a_k, a_l) \equiv P_{ij}(a_k, a_l) - P_i(a_k)P_i(a_l)$. In the mean field approximation, a bottleneck in computation is the calculation of the inverse of a covariance matrix C that is a $(q-1)L \times (q-1)L$ matrix. In the mean field approximation, a prior distribution in Eq. 9.11 is ignored and pseudocount is employed instead of regularization terms to make the covariance matrix invertible.

The Gaussian approximation (a continuous multivariate Gaussian model) for the probability distribution of sequences is employed together with an exponential prior (an ℓ_1 regularization term) in PSICOV (Jones et al. 2012), and with a normal-inverse-Wishart (NIW) prior, which is a conjugate distribution of the multivariate Gaussian, in GaussDCA (Baldassi et al. 2014). The use of NIW prior has a merit that fields and couplings can be analytically formulated; see Eqs. 9.30 and 9.31.

All methods based on the Gaussian approximation employ the analytical formula for couplings, $J \simeq -C^{-1} = -\Theta$, which are essentially as same as the mean field approximation with a difference that the covariance matrix (C) or precision matrix (Θ) is differently estimated based on the various priors. The mean field and Gaussian approximations may be appropriate to systems of dense and weak couplings but questionable for sparse and strong couplings that is the characteristic of residue-residue contact networks. Although the mean field and Gaussian approximations successfully predict residue-residue contacts in proteins, it has been shown (Barton et al. 2016; Cocco et al. 2017) that they do not give the accurate estimates of fields and couplings in proteins.

A pseudo-likelihood with Gaussian priors (ℓ_2 regularization terms) is maximized to estimate fields and couplings in plmDCA (Ekeberg et al. 2013, 2014) for the Potts model with sparse interactions as well as reducing computational time; see Eq. 9.38 for the symmetric plmDCA and Eq. 9.41 for the asymmetric plmDCA. The asymmetric plmDCA method (Ekeberg et al. 2014) requires less computational time and fits particularly with parallel computing.

GREMLIN (Kamisetty et al. 2013) employs together with pseudo-likelihood Gaussian priors that depend on site pair, although its earlier version (Balakrishnan et al. 2011) employed ℓ_1 regularizers, which may be more appropriate to systems of sparse couplings. The ℓ_1 regularizers appear to learn parameters that are closer to their true strength, but the ℓ_2 regularizers appear to be as good as the ℓ_1 regularizers for the task of contact prediction that requires the relative ranking of the interactions and not their actual values (Kamisetty et al. 2013).

One of approaches to surpass the pseudo-likelihood approximation for systems of sparse couplings may be the adaptive cluster expansion (ACE) of cross

entropy (Cocco and Monasson 2011, 2012; Barton et al. 2016), in which cross entropy is approximately minimized by taking account of only site clusters the incremental entropy (cluster entropy) of which by adding one more site is significant. In this method (Barton et al. 2016), a Boltzmann machine is employed to refine fields and couplings and also to calculate model correlations such as single-site and pairwise amino acid frequencies under given fields and couplings. The results of the Boltzmann machine for both biological and artificial models showed that ACE outperforms plmDCA in recovering single-site marginals (amino acid frequencies at each site) and the distribution of the total dimensionless energies ($H_{\mathrm{Potts}}(\sigma)$) (Barton et al. 2016); those models were a lattice protein, trypsin inhibitor, HIV p7 nucleocapsid protein, multi-electrode recording of cortical neurons, and Potts models on Eridös-Rényi random graphs. More importantly ACE could accurately recover the true fields h and couplings J corresponding to Potts states with $P_i(a_k) \geq 0.05$ for Potts models ($L = 50$) on Eridös-Rényi random graphs (Barton et al. 2016). On the other hand, plmDCA gave accurate estimates of couplings at weak regularization for well sampled single-site probabilities, but less accurate fields. Also, plmDCA yielded less well inferred fields and couplings for single-site and two-site probabilities not well sampled, indicating that not well populated states should be merged. As a result, the distribution of the total energies (Barton et al. 2016) and the distribution of mutations with respect to the consensus sequence were not well reproduced (Cocco et al. 2017). Similarly, the mean field approximation could not reproduce two-site marginals and even single-site marginals (Cocco et al. 2017) and the Gaussian approximation could not well reproduce the distribution of mutations with respect to the consensus sequence (Barton et al. 2016).

However, the less reproducibility of couplings does not necessarily indicate the less predictability of residue-residue contacts, probably because in contact prediction the relative ranking of scores (Eq. 9.47) based on couplings is more important than their actual values. ACE with the optimum regularization strength with respect to the reproducibility of fields and couplings showed less accurate contact prediction than plmDCA and mfDCA. For ACE to show comparable performance of contact prediction with plmDCA, regularization strength had to be increased from $\gamma = 2/B = 10^{-3}$ to $\gamma = 1$ for Trypsin inhibitor, making couplings strongly damped and then the generative properties of inferred models lost (Barton et al. 2016) (Table 9.2).

9.2.2 Partial Correlation of Amino Acid Cosubstitutions Between Sites at Each Branch of a Phylogenetic Tree

In the DCA analyses on residue covariations between sites in a multiple sequence alignment (MSA), phylogenetic biases, which are sequence biases due to phylogenetic relations between species, in the MSA must be removed as well as indirect

Table 9.2 Free softwares/servers for the direct coupling analysis

Name	Methods	URL
EVcouplings (Marks et al. 2011)	mfDCA	http://evfold.org
EVcouplings, plmc (Toth-Petroczy et al. 2016; Weinreb et al. 2016)	mf/plmDCA	https://github.com/debbiemarkslab
DCA (Morcos et al. 2011; Marks et al. 2011)	mfDCA	http://dca.rice.edu/portal/dca/home
GaussDCA (Baldassi et al. 2014)	GaussDCA	http://areeweb.polito.it/ricerca/cmp/code
FreeContact (Kaján et al. 2014)	mfDCA, PSICONV	http://rostlab.org/owiki/index.php/FreeContact
plmDCA (Ekeberg et al. 2013, 2014)	plmDCA	http://plmdca.csc.kth.se/ https://github.com/pagnani/plmDCA
CCMpred (Seemayer et al. 2014)	plmDCA	Performance-optimized software https://github.com/soedinglab/ccmpred
GREMLIN (Balakrishnan et al. 2011; Kamisetty et al. 2013)	GREMLIN	http://gremlin.bakerlab.org/
ACE (Cocco and Monasson 2011, 2012; Barton et al. 2016)	ACE	https://github.com/johnbarton/ACE
Persistent-vi (Ingraham and Marks 2016)	Persistent VI	https://github.com/debbiemarkslab

correlations between sites, but instead are reduced by taking weighted averages over homologous sequences in the calculation of single and pairwise frequencies of amino acids.

Needless to say, it is supposed that observed pattems of covariation were caused by molecular coevolution between sites. Whatever caused covariations found in the MSA, it has been confirmed that they can be utilized to predict residue pairs in close proximity in a three dimensional structure. Talavera et al. (2015) claimed, however, that covarying substitutions were mostly found on different branches of the phylogenetic tree, indicating that they might or might not be attributable to coevolution.

In order to remove phylogenetic biases and also to respond to such a claim above, it is meaningful to study covarying substitutions between sites in a phylogenetic tree-dependent manner. Such an alternative approach was taken to infer coevolving site pairs from direct correlations between sites in concurrent and compensatory substitutions within the same branches of a phylogenetic tree (Miyazawa 2013). In this method, substitution probability and mean changes of physico-chemical properties of side chain accompanied by amino acid substitutions at each site in each branch of the tree are estimated with the likelihood of each substitution to detect concurrent and compensatory substitutions. Then, partial correlation coefficients of the vectors of their characteristic changes accompanied by substitutions, substitution probability and mean changes of physico-chemical properties, along branches between sites are calculated to extract direct correlations in coevolutionary

substitutions and employed as a score for residue-residue contact. The accuracy of contact prediction by this method was comparable with that by mfDCA (Miyazawa 2013). This method, however, has a drawback to be computationally intensive, because an optimum phylogenetic tree must be estimated.

9.3 Machine Learning Methods to Augment the Contact Prediction Accuracy Based on Amino Acid Coevolution

All the DCA methods such as mfDCA, plmDCA, GREMLIN, and PSICOV predict significantly nonoverlapping sets of contacts (Jones et al. 2015; Kosciolek and Jones 2016; Wuyun et al. 2016). Then, increasing prediction accuracy by combining their predictions together with other sequence/structure information have been attempted (Skwark et al. 2013, 2014, 2016; Kosciolek and Jones 2014, 2016; Jones et al. 2015; Wang et al. 2017; Shendure and Ji 2017); see Table 9.3.

PconsC (Skwark et al. 2013) combines the predictions of PSICOV and plmDCA into a machine learning method, random forests, and employs alignments with HHblits (Remmert et al. 2012) and jackHMMer (Johnson et al. 2010) at four different e-value cut-offs. Five-layer neural network is employed instead of random forests in PconsC2 (Skwark et al. 2014), and plmDCA and GaussDCA are employed in PconsC3 (Skwark et al. 2016). A receptive field consisting of 11×11 predicted contacts around each residue pair is taken into account in each layer except the first one.

Table 9.3 Machine learning methods that combine predicted direct couplings with other sequence/structure information

Name	Basic method	Post-processing
PconsC3 (Skwark et al. 2016)	plmDCA, GaussDCA	5 layer DNN; http://c3.pcons.net. PconsC (Skwark et al. 2013), PconsC2 (Skwark et al. 2014)
MetaPSICOV (Kosciolek and Jones 2014, 2016; Jones et al. 2015)	PSICOV, mfDCA, GREMLIN/CCMpred	A two stage neural network predictor; CONSIP2 pipeline http://bioinf.cs.ucl.ac.uk/MetaPSICOV
RaptorX (Wang et al. 2017)	CCMpred	Ultra-deep learning model consisting of 1- and 2-dimensional convolutional residual neural networks http://raptorx.uchicago.edu/ContactMap/
iFold (CASP12 2017)		Deep neural network (DNN)
EPSILON-CP	PSICOV, GREMLIN, mfDCA, CCMpred, GaussDCA	4 hidden layer neural network with 400-200-200-50 neurons (Shendure and Ji 2017)

MetaPSICOV (Jones et al. 2015; Kosciolek and Jones 2016) combines the predictions of PSICOV, mfDCA, and CCMpred/GREMLIN into the first stage of a two-stage neural network predictor together with a well-established "classic" machine learning contact predictor, which utilizes many features such as amino acid profiles, predicted secondary structure and solvent accessibility along with sequence separation predicted, as an additional source of information for a little depth of MSAs. The second stage analyses the output of the first stage to eliminate outliers and to fill in the gaps in the contact map. On a set of 40 target domains with a median family size of around 40 effective sequences in CASPII, CONSIP2 server achieved an average top-$L/5$ long-range contact precision of 27% (Kosciolek and Jones 2016).

Wang et al. (2017) have also shown that a ultra-deep neural network (RaptorX) can significantly improve contact prediction based on amino acid coevolution. They have modeled short-range and long-range correlations in sequential and structural features with respect to complex sequence-structure relationships in proteins by one-dimensional and two-dimensional deep neural networks (DNN), respectively. Both the DNNs are convolutional residual neural networks. The 1D DNN performs convolutional transformations, with respect to residue position, of sequential features such as position-dependent scoring matrix, predicted 3-state secondary structure and 3-state solvent accessibility. The 2D DNN does 2D convolutional transformations of pairwise features such as coevolutional information calculated by CCMpred, mutual information, pairwise contact potentials as well as the output of the 1D DNN converted by a similar operation to outer product. Residual neural networks are employed because they can pass both linear and nonlinear informations from initial input to final output, making their training relatively easy.

9.4 Performance of Contact Prediction

New statistical methods based on the direct coupling analysis are confirmed in various benchmarking studies (Moult et al. 2016; CASP12 2017; Kamisetty et al. 2013; Wuyun et al. 2016) to show remarkable accuracy of contact prediction, although deep, stable alignments are required. They can more accurately detect a higher number of contacts between residues, which are very distant along sequence (Morcos et al. 2011). The top-scoring residue couplings are not only sufficiently accurate but also well-distributed to define the 3D protein fold with remarkable accuracy (Marks et al. 2011); this observation was quantified by computing, from sequence alone, all-atom 3D structures of 15 test proteins from different fold classes, ranging in size from 50 to 260 residues, including a G-protein coupled receptor. The contact prediction performs relatively better on β proteins than on α proteins (Miyazawa 2013). These initial findings on a limited number of proteins were confirmed as a general trend in a large-scale comparative assessment of contact prediction methods (Wuyun et al. 2016; Adhikari et al. 2016).

In CASP12, RaptorX performed the best in terms of F1 score for top $L/2$ long- and medium-range contacts of 38 free-modeling (FM) targets; the total F1 score of RaptorX was better by about 7.6% and 10.0% than the second and third best servers, iFold_1 and the revised MetaPSICOV, respectively (Wang et al. 2017; CASP12 2017). Tested on 105 CASP11 targets, 76 past CAMEO hard targets, and 398 membrane proteins, the average top $L(L/10)$ long-range prediction accuracies of RaptorX are 0.47(0.77) in comparison with 0.30(0.59) for MetaPSICOV and 0.21(0.47) for CCMpred (Wang et al. 2017; CASP12 2017).

9.4.1 MSA Dependence of Contact Prediction Accuracy

In the direct-coupling-based methods, the accuracy of predicted contacts depends on the depth (Miyazawa 2013; Kamisetty et al. 2013; Wuyun et al. 2016) and quality of multiple sequence alignment (MSA) for a target. $5 \times L$ (protein length) aligned sequences may be desirable for accurate contact predictions (Kamisetty et al. 2013), although attempts to improve prediction methods for fewer aligned sequences have been made (Skwark et al. 2013, 2014, 2016; Wang et al. 2017). PconsC3 can be used for families with as little as 100 effective sequence members (Skwark et al. 2016). Also, RaptorX (Wang et al. 2017) attained top- $L/2$-accuracy >0.3 for long-rang contacts even by using MSAs with 20 effective sequence members.

Deepest MSAs including a target sequence were built with various values of E-value cutoff (Skwark et al. 2013) and coverage parameters (Jones et al. 2015; Kosciolek and Jones 2016) in sequence search and alignment programs based on the hidden Markov models such as HHblits and jackHMMer. Although prediction performance tends to increase in general as alignment depth is deeper (Miyazawa 2013), it was reported (Kosciolek and Jones 2016) that in the case of transmembrane domains, building too deep alignments could result in unrelated sequences or drifted domains being included. To increase alignment quality, E-value and coverage parameters may be carefully tuned for each alignment (Kosciolek and Jones 2016). In the case of alignments that might contain regions of partial matches, a too stringent sequence coverage requirement could result in missing related sequences. On the other hand, a too permissive sequence coverage requirement could pick up unrelated sequences, permitting many partial matches. A trade-off is required between the effective number of sequences and sequence coverage, and an appropriate E-value must be chosen not to much decrease both alignment depth and sequence coverage (Hopf et al. 2012).

9.5 Contact-Guided de novo Protein Structure Prediction

It is a primary obstacle to de novo structure prediction that current methods and computers cannot make it feasible to adequately sample the vast conformational

space a protein might take in the precess of folding into the native structure (Kim et al. 2009). Thus, it is critical whether residue-residue proximities inferred with direct coupling analysis can provide sufficient information to reduce a huge search space for a protein fold, without any known 3D structural information of the protein.

Algorithms are needed to fold proteins into native folds based on contact information; see Table 9.4. Distance geometry generation (Havel et al. 1983; Braun and Go 1985) of 3D structures, which may be followed by energy minimization and molecular dynamics, will be just the primary one. In EVfold (Marks et al. 2011), contacts inferred by direct coupling analysis and predicted secondary structure information are translated into a set of distance constraints for the use of a distance geometry algorithm in the Crystallography and NMR System (CNS) (Brünger 2007). It was confirmed that the evolutionary inferred contacts can sufficiently reduce a search space in the structure predictions of 15 test proteins from different fold classes (Marks et al. 2011), and of 11 unknown and 23 known transmembrane protein structures (Hopf et al. 2012). Because distance constraints from predicted contacts may be partial in a protein sequence, they should be embedded into ab initio structure prediction methods.

Table 9.4 Contact-guided de novo protein structure prediction methods and servers

Name	Contact prediction	
EVfold (Marks et al. 2011, 2012)/EVfold_membrane (Hopf et al. 2012)	mfDCA/plmDCA	Using distance geometry algorithm (Havel et al. 1983) and simulated annealing of CNS (Brünger 2007); http://evfold.org/
DCA-fold (Sufkowska et al. 2012)	mfDCA	Simulated annealing using a coarse-grained molecular dynamics for a C_α model
FRAGFOLD/FILM3	MetaPSICOV	Combining fragment-based folding algorithm (Jones et al. 2005) with PSICOV (Kosciolek and Jones 2014) and with MetaPSICOV (Jones et al. 2015). FILM3 (Nugent and Jones 2012) is employed instead of FRAGFOLD (Jones 2001) for transmembrane proteins.
CONFOLD (Adhikari et al. 2015)	EVFOLD/FRAGFOLD (PSIPRED for 2nd structures)	Two-stage contact-guided de novo protein folding, using distance geometry simulated annealing protocol in a revised CNS v1.3. http://protein.rnet.missouri.edu/confold/
Rosetta (Kim et al. 2004; Ovchinnikov et al. 2016)	GREMLIN	Fragment assembly

Sulkowska et al. also showed that a simple hybrid method, called DCA-fold, integrating mfDCA-predicted contacts with an accurate knowledge of secondary structure is sufficient to fold proteins in the range of 1–3 Å resolution (Sufkowska et al. 2012). In this study, simulated annealing using a coarse-grained molecular dynamics model was employed for a C_α chain model, in which C_αs interact with each other with a contact potential approximated by a Gaussian function and a torsional potential depending on C_α dihedral angles at each position.

Adhikari et al. (2015) studied a way to effectively encode secondary structure information into distance and dihedral angle constrains that complement long-range contact constraints, and revised the CNS v1.3 to effectively use secondary structure constraints together with predicted long-range constraints; CONFOLD (Adhikari et al. 2015) consists of two stages. In the first stage secondary structure information is converted into distance, dihedral angle, and hydrogen bond constraints, and then best models are selected by executing the distance geometry simulated annealing. In the second stage self-conflicting contacts in the best structure predicted in the first stage are removed, constrains based on the secondary structures are refined, and again the distance geometry simulated annealing is executed.

Baker group (Ovchinnikov et al. 2016) embedded contact constraints predicted by GREMLIN (Kamisetty et al. 2013) as sigmoidal constraints to overcome noise in the Rosetta (Kim et al. 2004) conformational sampling and refinement. They found that model accuracy will be generally improved, if more than 3 L (protein length) sequences are available, and that large topologically complex proteins can be modeled with close to atomic-level accuracy without knowledge of homologous structures, if there are enough homologous sequences available.

On the other hand, a fragment-based folding algorithm FRAGFOLD was combined with PSICOV (Kosciolek and Jones 2014) and with MetaPSICOV (Jones et al. 2015; Kosciolek and Jones 2016); In this approach, predicted contacts are converted into additional energy terms for FRAGFOLD in addition to the pairwise potentials of mean force and solvation (Jones et al. 2015; Kosciolek and Jones 2016). FILM3 (Nugent and Jones 2012), with constraints based on predicted contacts and ones approximating Z-coordinate values within the lipid membrane, is employed instead of FRAGFOLD for transmembrane proteins.

RaptorX (Wang et al. 2017) employed the CNS suite (Brünger 2007) to generate 3D models from predicted contacts and secondary structure converted to distance, angle and h-bond restraints, and could yield TMscore >0.6 for 203 of 579 test proteins, while using MetaPSICOV and CCMpred could do so for 79 and 62, respectively.

9.5.1 How Many Predicted Contacts Should Be Used to Build 3D Models?

The number of feasible contacts surrounding a residue in a protein is about 6.3 (Miyazawa and Jernigan 1996), which corresponds to the maximum number of contacts per a protein, $6.3L/2$, where L denotes protein length. However, more than 50% of known 3D structures in the PDB have less than $2L$ contacts, and in the test on 15 proteins in EVfold benchmark set, less than $1.6L$ predicted contacts yielded best results (Adhikari et al. 2015). In the original EVfold, the optimal number of evolutionary constraints was in the order of $0.5L$ to $0.7L$ (Hopf et al. 2012). Because prediction accuracy tends to decrease as the rank of contact score increases, and different proteins need different numbers of predicted contacts to be folded well, protein folds were generated with a wide range of the number of predicted contacts, and then best folds were selected; from 30 to L in EVfold (Hopf et al. 2012), and from $0.4L$ to $2.2L$ in CONFOLD (Adhikari et al. 2015). In RaptorX, the top $2L$ predicted contacts irrespective of site separation were converted to distance restraints (Wang et al. 2017). On the other hand, Jones group reported (Kosciolek and Jones 2014) that artificially truncating the list of predicted contacts was likely to remove useful information to fold a protein with FRAGFOLD and PSICOV, in which the weight of a given predicted contact is determined by its positive predictive value.

9.6 Evolutionary Direct Couplings Between Residues Not Contacting in a Protein 3D Structure

Needless to say, evolutionary constraints do not only originate in intra-molecular contacts but also result from inter-molecular contacts/interactions. Even in the case of intra-molecular contacts, if there are structural variations including ones due to conformational changes in a protein family, evolutionary constraints will reflect the alternative conformations (Morcos et al. 2011; Hopf et al. 2012; Anishchenko et al. 2013). Also, intra-molecular residue couplings may contain useful information of ligand-mediated residue couplings (Morcos et al. 2011; Ovchinnikov et al. 2016). On the other hand, inter-molecular contacts may allow us to predict protein complexes, and are useful to build protein-protein interaction networks at a residue level.

9.6.1 Structural Variation Including Conformational Changes

MSA contains information on all members of the protein family, and direct couplings between residues estimated from the MSA reflect the structures of all

members. It was shown (Anishchenko et al. 2013) that 74% of top $L/2$ direct couplings residue pairs that are more than 5 Å apart in the target structures of 3883 proteins are less than 5 Å apart in at least one homolog structure.

Conformational change is an interesting case of structural variation. Many proteins adopt different conformations as part of their functions (Tokuriki and Tawfik 2009), indicating that protein flexibility is as important as structure on biological function. Protein flexibility around the energy minimum can be studied by sampling around the native structure in normal mode/principal component analysis, coarse-grained elastic network model, and short-timescale MD simulations. However, distant conformers that require large conformational transitions are difficult to predict. If conformational changes are essential on protein functions, evolutionary constraints will reflect the multiple conformations. Toth-Petroczy et al. (2016) showed that coevolutionary information may reveal alternative structural states of disorderd regions.

Morcos et al. (2011) found that some of top predicted contacts in the response-regulator DNA-binding domain family (GerE, PF00196) conflict with the structure (PDB ID 3C3W) of the full-length response-regulator DosR of M. tuberculosis, but are compatible with the structure (PDB ID 1JE8) of DNA-binding domain of *E. coli* NarL.

Sutto et al. (2015) combined coevolutionary data and molecular dynamics simulations to study protein conformational heterogeneity; the Boltzmann-learning algorithm with ℓ_2 regularization terms was employed to extract direct couplings between sites in homologous protein sequences, and a set of conformations consistent with the observed residue couplings were generated by exhaustive sampling simulations based on a coarse-grained protein model. Although the most representative structure was consistent with the experimental fold, the various regions of the sequence showed different stability, indicating conformational changes (Sutto et al. 2015).

Sfriso et al. (2016) made an automated pipeline based on discrete molecular dynamics guided by predicted contacts for the systematic identification of functional conformations in proteins, and identified alternative conformers in 70 of 92 proteins in a validation set of proteins in PDB; various conformational transitions are relevant to those conformers, such as open-closed, rotation, rotation-closed, concerted, and miscellanea of complex motions.

9.6.2 Homo-Oligomer Contacts

Intra-molecular contacts that conflict with the native fold may indicate homo-oligomer contacts (Anishchenko et al. 2013). Such a case was confirmed for homo-oligomer contacts in the ATPase domain of nitrogen regulatory protein C-like sigma-54 dependent transcriptional activators (Morcos et al. 2011) and between transmembrane helices (Hopf et al. 2012). It was pointed out (Hopf et al. 2012) that

the identification of evolutionary couplings due to homo-oligomerization is not only meaningful in itself but also useful because their removal improves the accuracy of the structure prediction for the monomer.

9.6.3 Residue Couplings Mediated by Binding to a Third Agent

Direct couplings between residues found by the DCA analysis can be mediated (Morcos et al. 2011) by their interactions with a third agent, i.e., ligands, substrates, RNA, DNA, and other metabolites. This indicates that binding sites with such a agent may be found as residue sites directly coupled but not in contact.

If interactions with a third agent requires too specific residue type at a certain site, then the residue type will be well conserved at the binding sites. This often occurs, and has been utilized to identify binding sites. However, the interactions for binding are less specific but certainly restricted, direct couplings between residues around the binding sites may occurs.

Hopf et al. (2012) devised a total evolutionary coupling score, which is defined as EC values summed over all high-ranking pairs involving a given residue and normalized by their average over all high-ranking pairs, and showed that residues with high total coupling scores line substrate-binding sites and affect signaling or transport in transmembrane proteins, Adrb2 and Opsd.

9.7 Heterogeneous Protein-Protein Contacts

An application of the direct coupling analysis to predict the structures of protein complexes is straightforward. In place of a MSA of a single protein family, a single MSA that is built by concatenating the multiple MSAs of multiple protein families every species can be employed to extract direct couplings between sites of different proteins by removing indirect intra- and inter-protein couplings (Pazos et al. 1997; Skerker et al. 2008; Weigt et al. 2009; Hopf et al. 2012).

A critical requirement for sequences to be concatenated is, however, that respective sets of the protein sequences must have the same evolutionary history to coevolve. In other words, phylogenetic trees built from the respective sets of sequences employed for the protein families must have at least the same topology. One way to build a set of cognate pairs of protein sequences is to employ orthologous sequences for each protein family, the phylogenetic tree of which coincides with that of species. Thus, a genome-wide analysis of finding protein-protein interactions based on protein sequences is not so simple.

Weigt et al. (2009) successfully applied the direct coupling analysis to the bacterial two-component signal transduction system consisting of sensor kinase (SK) and response regulator (RR), which are believed (Skerker et al. 2008) to interact specifically with each other in most cases and often revealed by adjacency

in chromosomal location. This analysis is based on the fact that in prokaryotes cognate pairs are often encoded in the same operon. Genome-sequencing projects have revealed that most organisms contain large expansions of a relatively small number of signaling families (Skerker et al. 2008). However, it is not as simple as in prokaryotes to build a set of cognate pairs of those protein sequences in eukaryotes.

Hopf et al. (2014) developed a contact score, EVcomplex, for every inter-protein residue pair based on the overall inter-protein EC score distributions, evaluated its performance in blinded tests on 76 complexes of known 3D structure, predicted protein-protein contacts in 32 complexes of unknown structure, and then demonstrated how evolutionary direct couplings can be used to distinguish between interacting and non-interacting protein pairs in a large complex. In their analysis, protein sequence pairs that are encoded close on *E. coli* genome were employed to reduce incorrect protein pairings.

9.8 Discussion

Determination of protein structure is essential to understand protein function. However, despite significant effort to explore unknown folds in the protein structural space, protein structures determined by experiment are far less than known protein families. Only about 41–42% of the Pfam families (Finn et al. 2016) (Pfam-A release 31.0, 16712 families) include at least one member whose structure is known. The number and also the size of protein families will further grow as genome/metagenome sequencing projects proceed with next-generation sequencing technologies. Thus, accurate de novo prediction of three-dimensional structure is desirable to catch up with the high growing speed of protein families with unknown folds. Coevolutionary information can be used to predict not only proteins but also RNAs (Weinreb et al. 2016) and those complexes, together with experimental informations such as X-ray, NMR, SAS, FRET, crosslinking, Cryo-EM, and others.

Here, statistical methods for disentangling direct from indirect couplings between sites with respect to evolutionary variations/substitutions of amino acids in homologous proteins have been briefly reviewed. Dramatic improvements on contact prediction and successful 3D de novo predictions based on predicted contacts are described in details in the recent reports of CASP-11 (Moult et al. 2016) and CASP-12 meetings (CASP12 2017). Machine learning methods, particularly deep neural network (DNN) such as MetaPSICOV, iFold, and RaptorX, have shown to significantly augment contact prediction accuracy based on coevolutionary information. However, the present state-of-the-art DNN methods are, at least at the very moment, not powerful enough to extract coevolutionary information directly from homologous sequences. It was reported that without coevolutionary strength produced by CCMpred the top $L/10$ long-range prediction accuracy of RaptorX might drop by 0.15 for soluble proteins and more for membrane proteins (Wang et al. 2017), indicating that the direct coupling analysis is still essential for contact prediction.

The primary requirement for the direct coupling analysis is a high quality deep alignment. However, genome/metagenome sequencing projects provide more genetic variations from which more accurate and more comprehensive information on evolutionary constraints can be extracted. One of problems is that species being sequenced may be strongly biased to prokaryotes, making it hard to analyze eukaryotic proteins based on coevolutionary substitutions. Experiments of vitro evolution may be useful to provide sequence variations for eukaryotic proteins (Ovchinnikov et al. 2016).

For a large-scale of protein structure prediction, computationally intensive methods such as the ACE and Boltzmann machine (MCMC and mpDCA) can hardly be employed. The Gaussian approximation with a normal-inverse-Wishart prior, the Gaussian approximations with other priors (PSICOV) and mean field approximation (mfDCA) are fast enough but their performance of contact prediction tends to be compared unfavorably with the pseudo-likelihood approximation (plmDCA), indicating that they may be inappropriate for proteins with sparse couplings.

The accurate estimates of fields and couplings are very informative in evaluating the effects ($\Delta H_{\mathrm{Potts}}$) of mutations (Hopf et al. 2017), identifying protein family members and also studying folding mechanisms (Morcos et al. 2014; Jacquin et al. 2016) and protein evolution (Miyazawa 2017b). It should be also examined whether the distribution of dimensionless energies (H_{Potts}) over homologous proteins can be well reproduced. Accuracy of estimates of fields and couplings and the distribution of dimensionless energies depends on regularization parameters or the ratio of pseudocount (Barton et al. 2016; Miyazawa 2017b), and therefore they should be optimized. It was also pointed out that group L_1 regularization performs better than L_2 for the maximum pseudolikelihood method (Ingraham and Marks 2016). The ACE algorithm, which can be applied only for systems of sparse couplings, may be more favorable with respect to computational load for the estimation of fields and couplings than Boltzmann learning with Monte Carlo simulation or with message passing. However, both the methods are computationally intensive. Recently, another approach consisting of two methods named persistent-vi and Fadeout, in which the posterior probability density with horseshoe prior is approximately estimated by using variational inference and noncentered parameterization for such a sparsity-inducing prior, has shown to perform better with twofold cpu time than the maximum pseudolikelihood method with L_2 and group L_1 regularizations (Ingraham and Marks 2016).

The remarkable advances of sequencing technologies and also statistical methods are likely to bring many targets within range of the present approach in the near future, and have a potential to transform the field (Moult et al. 2016).

Appendix

An appendix described in full will be found in the article (Miyazawa 2017a) submitted to the arXiv.

Inverse Potts Model

A Gauge Employed for $h_i(a_k)$ and $J_{ij}(a_k, a_l)$

Unless specified, a following gauge is employed; we call it q-gauge, here.

$$h_i(a_q) = J_{ij}(a_k, a_q) = J_{ij}(a_q, a_l) = 0 \tag{9.16}$$

In this gauge, the amino acid a_q is the reference state for fields and couplings, and $P_i(a_q)$, $P_{ij}(a_k, a_q) = P_{ji}(a_q, a_k)$, and $P_{ij}(a_q, a_q)$ are regarded as dependent variables. Common choices for the reference state a_q are the most common (consensus) state at each site. Any gauge can be transformed to another by the following transformation.

$$J^{\mathrm{I}}_{ij}(a_k, a_l) \equiv J_{ij}(a_k, a_l) - J_{ij}(\cdot, a_l) - J_{ij}(a_k, \cdot) + J_{ij}(\cdot, \cdot) \tag{9.17}$$

$$h^{\mathrm{I}}_i(a_k) \equiv h_i(a_k) - h_i(\cdot) + \sum_{j \neq i} (J_{ij}(a_k, \cdot) - J_{ij}(\cdot, \cdot)) \tag{9.18}$$

where "$\cdot$" denotes the reference state, which may be a_q for each site (q-gauge) or the average over all states (Ising gauge).

Boltzmann Machine

Fields $h_i(a_k)$ and couplings $J_{ij}(a_k, a_l)$ are estimated by iterating the following 2-step procedures.

1. For a given set of h_i and $J_{ij}(a_k, a_l)$, marginal probabilities, $P^{\mathrm{MC}}(\sigma_i = a_k)$ and $P^{\mathrm{MC}}(\sigma_i = a_k, \sigma_i = a_l)$, are estimated by a Markov chain Monte Carlo method (the Metropolis-Hastings algorithm (Metropolis et al. 1953)) or by any other method (for example, the message passing algorithm (Weigt et al. 2009)).
2. Then, h_i and $J_{ij}(a_k, a_l)$ are updated according to the gradient of negative log-posterior-probability per instance, $\partial S_0/\partial h_i(a_k)$ or $\partial S_0/\partial J_{ij}(a_k, a_l)$, multiplied by a parameter-specific weight factor (Barton et al. 2016), $w_i(a_k)$ or $w_{ij}(a_k, a_l)$; see Eqs. 9.8 and 9.12.

$$\Delta h_i(a_k) = -(P^{\mathrm{MC}}(\sigma_i = a_k) + \frac{\partial R}{\partial h_i(a_k)} - P_i(a_k)) \cdot w_i(a_k) \tag{9.19}$$

$$\Delta J_{ij}(a_k, a_l) = -(P^{\mathrm{MC}}(\sigma_i = a_k, \sigma_i = a_l) + \frac{\partial R}{\partial J_{ij}(a_k, a_l)} - P_{ij}(a_k, a_l)) \cdot w_{ij}(a_k, a_l) \tag{9.20}$$

where weights are also updated as $w_i(a_k) \leftarrow f(w_i(a_k))$ and $w_{ij}(a_k, a_l) \leftarrow f(w_{ij}(a_k, a_l))$ according to the RPROP (Riedmiller and Braun 1993) algorithm; the function $f(w)$ is defined as

$$f(w) \equiv \begin{cases} \max(w \cdot s_-, w_{\min}) \text{ if the gradient changes its sign,} \\ \min(w \cdot s_+, w_{\max}) \text{ otherwise} \end{cases} \tag{9.21}$$

$w_{\min} = 10^{-3}$, $w_{\max} = 10$, $s_- = 0.5$, and $s_+ = 1.9 < 1/s_-$ were employed (Barton et al. 2016). After updated, $h_i(a_k)$ and $J_{ij}(a_k, a_l)$ may be modified to satisfy a given gauge.

The Boltzmann machine has a merit that model correlations are calculated.

Gaussian Approximation for $P(\sigma)$ with a Normal-Inverse-Wishart Prior

The normal-inverse-Wishart distribution (NIW) is the product of the multivariate normal distribution ($\mathcal{N}$) and the inverse-Wishart distribution ($\mathcal{W}^{-1}$), which are the conjugate priors for the mean vector and for the covariance matrix of a multivariate Gaussian distribution, respectively. The NIW is employed as a prior in GaussDCA (Baldassi et al. 2014), in which the sequence distribution $P(\sigma)$ is approximated as a Gaussian distribution. In this approximation, the q-gauge is used, and $P_i(a_q)$, $P_{ij}(a_k, a_q) = P_{ji}(a_q, a_k)$, and $P_{ij}(a_q, a_q)$ are regarded as dependent variables; see section "A Gauge Employed for $h_i(a_k)$ and $J_{ij}(a_k, a_l)$"; in GaussDCA, deletion is excluded from independent variables.

The posterior distribution for the NIW is also a NIW. Thus, the cross entropy S_0 can be represented as

$$S_0(\boldsymbol{\mu}, \Sigma | \{P_i\}, \{P_{ij}\}) = \frac{-1}{B} \log[\prod_{\tau=1}^{B} \mathcal{N}(\{\delta_{\sigma_i^\tau a_k}\} | \boldsymbol{\mu}, \Sigma) \mathcal{N}(\boldsymbol{\mu} | \boldsymbol{\mu}^0, \Sigma/\kappa) \mathcal{W}^{-1}(\Sigma | \Lambda, \nu)] \tag{9.22}$$

$$= \frac{-1}{B} \log[\mathcal{N}(\boldsymbol{\mu} | \boldsymbol{\mu}^B, \Sigma/\kappa^B) \mathcal{W}^{-1}(\Sigma | \Lambda^B, \nu^B) \tag{9.23}$$

$$(\det(2\pi\Sigma))^{-B/2} (\frac{\kappa}{\kappa^B})^{\dim \Sigma/2} \frac{(\det(\Lambda/2))^{\nu/2}}{(\det(\Lambda^B/2))^{\nu^B/2}} \frac{\Gamma_{\dim \Sigma}(\nu^B/2)}{\Gamma_{\dim \Sigma}(\nu/2)} (\det \Sigma)^{-(\nu-\nu^B)2}] \tag{9.24}$$

where $\Gamma_{\dim \Sigma}(\nu/2)$ is the multivariate Γ function, $\boldsymbol{\mu}$ is the mean vector, and $\dim \Sigma$ is the dimension of covariance matrix Σ, $\dim \Sigma = (q-1)L$ excluding deletion in GaussDCA. The normal and NIW distributions are defined as follows.

$$\mathcal{N}(\boldsymbol{\mu} | \boldsymbol{\mu}^0, \Sigma) \equiv (\det(2\pi\Sigma))^{-1/2} \exp(-\frac{(\boldsymbol{\mu} - \boldsymbol{\mu}^0)^T \Sigma^{-1} (\boldsymbol{\mu} - \boldsymbol{\mu}^0)}{2}) \tag{9.25}$$

$$\mathcal{W}^{-1}(\Sigma|\Lambda,\nu) \equiv \frac{(\det(\Lambda/2))^{\nu/2}}{\Gamma_{\dim\Sigma}(\nu/2)}(\det\Sigma)^{-(\nu+\dim\Sigma+1)/2}\exp(-\frac{1}{2}\mathrm{Tr}\Lambda\Sigma^{-1}) \quad (9.26)$$

Parameters $\boldsymbol{\mu}^B$, κ^B, ν^B, and Λ^B satisfy

$$\mu_i^B(a_k) = (\kappa\mu_i^0(a_k) + BP_i(a_k))/(\kappa+B)\ ,\ \kappa^B = \kappa + B\ ,\ \nu^B = \nu + B \quad (9.27)$$

$$\Lambda_{ij}^B(a_k,a_l) = \Lambda_{ij}(a_k,a_l) + BC_{ij}(a_k,a_l) + \frac{\kappa B}{\kappa+B}[(P_i(a_k)-\mu_i^0(a_k))(P_j(a_l)-\mu_j^0(a_l))] \quad (9.28)$$

where the Λ and ν are the scale matrix and the degree of freedom, respectively, shaping the inverse-Wishart distribution, and C is the given covariance matrix; $C_{ij}(a_k,a_l) \equiv P_{ij}(a_k,a_l) - P_i(a_k)P_i(a_l)$. The mean values of $\boldsymbol{\mu}$ and Σ under NW posterior are $\boldsymbol{\mu}^B$ and $\Lambda^B/(\nu^B - \dim\Sigma - 1)$, and their mode values are $\boldsymbol{\mu}^B$ and $\Lambda^B/(\nu^B + \dim\Sigma + 1)$, which minimize the cross entropy or maximize the posterior probability. The covariance matrix Σ can be estimated to be the exactly same value by adjusting the value of ν, whichever the mean posterior or the maximum posterior is employed for the estimation of Σ. In GaussDCA, the mean posterior estimate was employed but here the maximum posterior estimate is employed according to the present formalism.

$$(\boldsymbol{\mu},\Sigma) = \arg\min_{(\boldsymbol{\mu},\Sigma)} S_0(\boldsymbol{\mu},\Sigma|\{P_i\},\{P_{ij}\}) = (\boldsymbol{\mu}^B, \Lambda^B/(\nu^B+\dim\Sigma+1)) \quad (9.29)$$

According to GaussDCA, ν is chosen in such a way that $\sigma_{ij}(a_k,a_l)$ is nearly equal to the covariance matrix corrected by pseudocount; $\nu = \kappa + \dim\Sigma + 1$ for the mean posterior estimate in GaussDCA, but $\nu = \kappa - \dim\Sigma - 1$ for the maximum posterior estimate here.

From Eq. 9.15, the estimates of couplings and fields are calculated.

$$J_{ij}^{\mathrm{NIW}}(a_k,a_l) = -\frac{\partial S_0(\{P_i\},\{P_{ij}\})}{\partial P_{ij}(a_k,a_l)} = -\frac{(\kappa+B+1)}{\kappa+B}(\Sigma^{-1})_{ij}(a_k,a_l) \quad (9.30)$$

Because the number of instances is far greater than 1 ($B \gg 1$), these estimates of couplings are practically equal to the estimates ($J^{\mathrm{MF}} = -\Sigma^{-1}$) in the mean field approximation, which was employed in GaussDCA (Baldassi et al. 2014).

$$h_i^{\mathrm{NIW}}(a_k) = -\sum_{j\neq i}\sum_l J_{ij}^{\mathrm{NIW}}(a_k,a_l)P_j(a_l) - \frac{(\kappa+B+1)}{\kappa+B}\sum_j\sum_{l\neq q}(\Sigma^{-1})_{ij}(a_k,a_l) [\delta_{ij}\frac{\delta_{kl}-2P_i(a_l)}{2} + \frac{\kappa B}{\kappa+B}(P_j(a_l)-\mu_j^0(a_l))] \quad (9.31)$$

The $(h_i^{\mathrm{NIW}}(a_k) - h_i^{\mathrm{NIW}}(a_q))$ does not converge to $\log P_i(a_k)/P_i(a_q)$ as $J^{\mathrm{NIW}} \to 0$ but $h_i^{\mathrm{MF}}(a_k) - h_i^{\mathrm{MF}}(a_q)$ does; in other words, the mean field approximation gives a better h for the limiting case of no couplings than the present approximation. Barton et al. (2016) reported that the Gaussian approximation generally gave a better generative model than the mean field approximation.

In GaussDCA (Baldassi et al. 2014), $\boldsymbol{\mu}^0$ and Λ/κ were chosen to be as uninformative as possible, i.e., mean and covariance for a uniform distribution.

$$\mu_i^0(a_k) = 1/q, \quad \frac{\Lambda_{ij}(a_k, a_l)}{\kappa} = \frac{\delta_{ij}}{q}(\delta_{kl} - \frac{1}{q}) \tag{9.32}$$

Pseudo-likelihood Approximation

Symmetric Pseudo-likelihood Maximization

The probability of an instance $\sigma^{\boldsymbol{\tau}}$ is approximated as follows by the product of conditional probabilities of observing $\sigma_i^{\boldsymbol{\tau}}$ under the given observations $\sigma_{j\neq i}^{\boldsymbol{\tau}}$ of all other sites.

$$P(\sigma^{\boldsymbol{\tau}}) \approx \prod_i P(\sigma_i = \sigma_i^{\boldsymbol{\tau}} | \{\sigma_{j\neq i} = \sigma_j^{\boldsymbol{\tau}}\}) \tag{9.33}$$

Then, cross entropy is approximated as

$$S_0(h, J|\{P_i\}, \{P_{ij}\}) \approx S_0^{\mathrm{PLM}}(h, J|\{P_i\}, \{P_{ij}\}) \equiv \sum_i S_{0,i}(h, J|\{P_i\}, \{P_{ij}\}) \tag{9.34}$$

$$S_{0,i}(h, J|\{P_i\}, \{P_{ij}\}) \equiv \frac{-1}{B} \sum_{\boldsymbol{\tau}} \ell_i(\sigma_i = \sigma_i^{\boldsymbol{\tau}} | \{\sigma_{j\neq i} = \sigma_j^{\boldsymbol{\tau}}\}, h, J) + R_i(h, J) \tag{9.35}$$

where conditional log-likelihoods and ℓ_2 norm regularization terms employed in Ekeberg et al. (2013) are

$$\ell_i(\sigma_i = \sigma_i^{\boldsymbol{\tau}} | \{\sigma_{j\neq i} = \sigma_j^{\boldsymbol{\tau}}\}, h, J) = \log[\frac{\exp(h_i(\sigma_i^{\boldsymbol{\tau}}) + \sum_{j\neq i} J_{ij}(\sigma_i^{\boldsymbol{\tau}}, \sigma_j^{\boldsymbol{\tau}}))}{\sum_k \exp(h_i(a_k) + \sum_{j\neq i} J_{ij}(a_k, \sigma_j^{\boldsymbol{\tau}}))}] \tag{9.36}$$

$$R_i(h, J) \equiv \gamma_h \sum_k h_i(a_k)^2 + \frac{\gamma_J}{2} \sum_k \sum_{j\neq i} \sum_l J_{ij}(a_k, a_l)^2 \tag{9.37}$$

The optimum fields and couplings in this approximation are estimated by minimizing the pseudo-cross-entropy, S_0^{PLM}.

$$(h^{\mathrm{PLM}}, J^{\mathrm{PLM}}) = \arg\min_{h,J} S_0^{\mathrm{PLM}}(h, J|\{P_i\}, \{P_{ij}\}) \tag{9.38}$$

Equation 9.38 is not invariant under gauge transformation; the ℓ_2 norm regularization terms in Eq. 9.38 favors only a specific gauge that corresponds to $\gamma_J \sum_l J_{ij}(a_k, a_l) = \gamma_h h_i(a_k)$, $\gamma_J \sum_k J_{ij}(a_k, a_l) = \gamma_h h_j(a_l)$, and $\sum_k h_i(a_k) = 0$ for all i, $j(> i)$, k and l (Ekeberg et al. 2013). $\gamma_J = \gamma_h = 0.01$ that is relatively a large value independent of B was employed in Ekeberg et al. (2013). $\gamma_h = 0.01$ but $\gamma_J = q(L-1)\gamma_h$ were employed in Hopf et al. (2017), in which gapped sites in each sequence were excluded in the calculation of the Hamiltonian $H(\sigma)$, and therefore $q = 20$.

GREMLIN (Kamisetty et al. 2013) employs Gaussian prior probabilities that depend on site pairs.

$$R_i(h, J) \equiv \gamma_h \sum_k h_i(a_k)^2 + \sum_k \sum_{j \neq i} \frac{\gamma_{ij}}{2} \sum_l J_{ij}(a_k, a_l)^2 \tag{9.39}$$

$$\gamma_{ij} \equiv \gamma_c(1 - \gamma_p \log(P_{ij}^0)) \tag{9.40}$$

where P_{ij}^0 is the prior probability of site pair (i, j) being in contact.

Asymmetric Pseudo-likelihood Maximization

To speed up the minimization of S_0, a further approximation, in which $S_{0,i}$ is separately minimized, is employed (Ekeberg et al. 2014), and fields and couplings are estimated as follows.

$$J_{ij}^{\mathrm{PLM}}(a_k, a_l) \simeq \frac{1}{2}(J_{ij}^*(a_k, a_l) + J_{ji}^*(a_l, a_k)) \tag{9.41}$$

$$(h_i^{\mathrm{PLM}}, J_i^*) = \arg\min_{h_i, J_i} S_{0,i}(h, J|\{P_i\}, \{P_{ij}\}) \tag{9.42}$$

It is appropriate to transform h and J estimated above into a some specific gauge such as the Ising gauge.

ACE (Adaptive Cluster Expansion) of Cross-Entropy for Sparse Markov Random Field

The cross entropy $S_0(\{h_i, J_{ij}\}|\{P_i\}, \{P_{ij}\}, i, j \in \Gamma)$ of a cluster of sites Γ, which is defined as the negative log-likelihood per instance in Eq. 9.14, is approximately minimized by taking account of sets $L_k(t)$ of only significant clusters consisting of

k sites, the incremental entropy (cluster cross entropy) ΔS_Γ of which is significant ($|\Delta S_\Gamma| > t$) (Cocco and Monasson 2011, 2012; Barton et al. 2016).

$$S_0(\{P_i, P_{ij}|i, j \in \Gamma\}) \simeq \sum_{l=1}^{|\Gamma|}, \sum_{\Gamma' \in L_l(t), \Gamma' \subset \Gamma} \Delta S_0(\{P_i, P_{ij}|i, j \in \Gamma'\}) \tag{9.43}$$

$$\Delta S_0(\{P_i, P_{ij}|i, j \in \Gamma\}) \equiv S_0(\{P_i, P_{ij}|i, j \in \Gamma\}) - \sum_{\Gamma' \subset \Gamma} \Delta S_0(\{P_i, P_{ij}|i, j \in \Gamma'\}) \tag{9.44}$$

$$= \sum_{\Gamma' \subseteq \Gamma} (-1)^{|\Gamma|-|\Gamma'|} S_0(\{P_i, P_{ij}|i, j \in \Gamma'\}) \tag{9.45}$$

$L_{k+1}(t)$ is constructed from $L_k(t)$ by adding a cluster Γ consisting of $(k+1)$ sites in a lax case provided that any pair of size k clusters $\Gamma^1, \Gamma^2 \in L_k(t)$ and $\Gamma^1 \cup \Gamma^2 = \Gamma$ or in a strict case if $\Gamma' \in L_k(t)$ for $\forall \Gamma'$ such that $\Gamma' \subset \Gamma$ and $|\Gamma'| = k$. Thus, Eq. 9.43 yields sparse solutions. The cross entropies $S_0(\{P_i, P_{ij}|i, j \in \Gamma'\})$ for the small size of clusters are estimated by minimizing $S_0(\{h_i, J_{ij}\}|\{P_i, P_{ij}\}, i, j \in \Gamma')$ with respect to fields and couplings. Starting from a large value of the threshold t (typically $t = 1$), the cross-entropy $S_0(\{P_i, P_{ij}\}|i, j \in \{1, \ldots, N\})$ is calculated by gradually decreasing t until its value converges. Convergence of the algorithm may also be more difficult for alignments of long proteins or those with very strong interactions. In such cases, strong regularization may be employed.

The following regularization terms of ℓ_2 norm are employed in ACE (Barton et al. 2016), and so Eq. 9.43 is not invariant under gauge transformation.

$$-\frac{1}{B} \log P_0(h, J|i, j \in \Gamma) = \gamma_h \sum_{i \in \Gamma} \sum_k h_i(a_k)^2 + \gamma_J \sum_{i \in \Gamma} \sum_k \sum_{J > i, j \in \Gamma} \sum_l J_{ij}(a_k, a_l)^2 \tag{9.46}$$

$\gamma_h = \gamma_J \propto 1/B$ was employed (Barton et al. 2016).

The compression of the number of Potts states, $q_i \leq q$, at each site can be taken into account. All infrequently observed states or states that insignificantly contribute to site entropy can be treated as the same state, and a complete model can be recovered (Barton et al. 2016) by setting $h_i(a_k) = h_i(a_{k'}) + \log(P_i(a_k)/P_i'(a_{k'}))$, and $J_{ij}(a_k, a_l) = J'_{ij}(a_{k'}, a_{l'})$, where "$'$" denotes a corresponding aggregated state and a potential.

Starting from the output set of the fields $h_i(a_k)$ and couplings $J_{ij}(a_k, a_l)$ obtained from the cluster expansion of the cross-entropy, a Boltzmann machine is trained with $P_i(a_k)$ and $P_{ij}(a_k)$ by the RPROP algorithm (Riedmiller and Braun 1993) to refine the parameter values of h_i and $J_{ij}(a_k, a_l)$ (Barton et al. 2016); see section "Boltzmann Machine". This post-processing is also useful because model correlations are calculated.

An appropriate value of the regularization parameter for trypsin inhibitor were much larger ($\gamma = 1$) for contact prediction than those ($\gamma = 2/B = 10^{-3}$) for

recovering true fields and couplings (Barton et al. 2016), probably because the task of contact prediction requires the relative ranking of interactions rather than their actual values.

Scoring Methods for Contact Prediction

Corrected Frobenius Norm (L_{22} Matrix Norm), $\mathcal{S}^{\mathrm{CFN}}_{ij}$

For scoring, plmDCA (Ekeberg et al. 2013, 2014) employs the corrected Frobenius norm of J^{I}_{ij} transformed in the Ising gauge, in which J^{I}_{ij} does not contain anything that could have been explained by fields h_i and h_j; $J^{\mathrm{I}}_{ij}(a_k, a_l) \equiv J_{ij}(a_k, a_l) - J_{ij}(\cdot, a_l) - J_{ij}(a_k, \cdot) + J_{ij}(\cdot, \cdot)$ where $J_{ij}(\cdot, a_l) = J_{ji}(a_l, \cdot) \equiv \sum_{k=1}^{q} J_{ij}(a_k, a_l)/q$.

$$\mathcal{S}^{\mathrm{CFN}}_{ij} \equiv \mathcal{S}^{\mathrm{FN}}_{ij} - \mathcal{S}^{\mathrm{FN}}_{\cdot j}\mathcal{S}^{\mathrm{FN}}_{i\cdot}/\mathcal{S}^{\mathrm{FN}}_{\cdot\cdot}, \quad \mathcal{S}^{\mathrm{FN}}_{ij} \equiv \sqrt{\sum_{\kappa\neq\mathrm{gap}}\sum_{l\neq\mathrm{gap}} J^{\mathrm{I}}_{ij}(a_k, a_l)^2} \tag{9.47}$$

where "·" denotes average over the indicated variable. This CFN score with the gap state excluded in Eq. 9.47 performs better (Ekeberg et al. 2014; Baldassi et al. 2014) than both scores of FN and DI/EC (Weigt et al. 2009; Morcos et al. 2011; Marks et al. 2011; Hopf et al. 2012).

References

Adhikari B, Bhattacharya D, Cao R, Cheng J (2015) CONFOLD: residue-residue contact-guided ab initio protein folding. Proteins 83:1436–1449. https://doi.org/10.1002/prot.24829

Adhikari B, Nowotny J, Bhattacharya D, Hou J, Cheng J (2016) ConEVA: a toolbox for comprehensive assessment of protein contacts. BMC Bioinf 17:517. https://doi.org/10.1186/s12859-016-1404-z

Altschuh D, Vernet T, Berti P, Moras D, Nagai K (1988) Coordinated amino acid changes in homologous protein families. Protein Eng 2:193–199

Anishchenko I, Ovchinnikov S, Kamisetty H, Baker D (2013) Origins of coevolution between residues distant in protein 3D structures. Proc Natl Acad Sci USA 114:9122–9127. https://doi.org/10.1073/pnas.1702664114

Atchley WR, Wollenberg KR, Fitch WM, Terhalle W, Dress AW (2000) Correlations among amino acid sites in bHLH protein domains: an information theoretic analysis. Mol Biol Evol 17:164–178

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

Baldassi C, Zamparo M, Feinauer C, Procaccini A, Zecchina R, Weigt M, Pagnani A (2014) Fast and accurate multivariate Gaussian modeling of protein families: predicting residue contacts and protein-interaction partners. PLoS ONE 9(3):e92721. https://doi.org/10.1371/journal.pone.0092721

Barton JP, Leonardis ED, Coucke A, Cocco S (2016) ACE: adaptive cluster expansion for maximum entropy graphical model inference. Bioinformatics 32:3089–3097. https://doi.org/10.1093/bioinformatics/btw328

Braun W, Go N (1985) Calculation of protein conformations by proton-proton distance constraints: a new efficient algorithm. J Mol Biol 186:611–626. https://doi.org/10.1016/0022-2836(85)90134-2

Brünger AT (2007) Version 1.2 of the crystallography and NMR system. Nat Protoc 2:2728–2733. https://doi.org/10.1038/nprot.2007.406

Burger L, van Nimwegen E (2008) Acurate prediction of protein-protein interactions from sequence alignments using a Bayesian method. Mol Syst Biol 4:165

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

CASP12 (2017) 12th community wide experiment on the critical assessment of techniques of protein structure prediction. http://predictioncenter.org/casp12/

Cocco S, Monasson R (2011) Adaptive cluster expansion for inferring Boltzmann machines with noisy data. Phys Rev Lett 106:090601. https://doi.org/10.1103/PhysRevLett.106.090601

Cocco S, Monasson R (2012) Adaptive cluster expansion for the inverse Ising problem: convergence, algorithm and tests. J Stat Phys 147:252–314. https://doi.org/10.1007/s10955-012-0463-4

Cocco S, Feinauer C, Figliuzzi M, Monasson R, Weigt M (2017) Inverse statistical physics of protein sequences: a key issues review. arXiv:1703.01222 [q-bio.BM]

Doron-Faigenboim A, Pupko T (2007) A combined empirical and mechanistic codon model. Mol Biol Evol 24:388–397

Dunn SD, Wahl LM, Gloor GB (2008) Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. Bioinformatics 24:333–340

Dutheil J (2012) Detecting coevolving positions in a molecule: why and how to account for phylogeny. Brief Bioinf 13:228–243

Dutheil J, Galtier N (2007) Detecting groups of coevolving positions in a molecule: a clustering approach. BMC Evol Biol 7:242

Dutheil J, Pupko T, Jean-Marie A, Galtier N (2005) A model-based approach for detecting coevolving positions in a molecule. Mol Biol Evol 22:1919–1928

Ekeberg M, Lövkvist C, Lan Y, Weigt M, Aurell E (2013) Improved contact prediction in proteins: using pseudolikelihoods to infer Potts models. Phys Rev E 87:012707–1–16. https://doi.org/10.1103/PhysRevE.87.012707

Ekeberg M, Hartonen T, Aurell E (2014) Fast pseudolikelihood maximization for direct-coupling analysis of protein structure from many homologous amino-acid sequences. J Comput Phys 276:341–356

Fares M, Travers S (2006) A novel method for detecting intramolecular coevolution. Genetics 173:9–23

Fariselli P, Olmea O, Valencia A, Casadio R (2001) Prediction of contact maps with neural networks and correlated mutations. Protein Eng 14:835–843

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. Nucl Acid Res 44:D279–D285. https://doi.org/10.1093/nar/gkv1344

Fitch WM, Markowitz E (1970) An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. Biochem Genet 4:579–593

Fleishman SJ, Yifrach O, Ben-Tal N (2004) An evolutionarily conserved network of amino acids mediates gating in voltage-dependent potassium channels. J Mol Biol 340:307–318

Fodor AA, Aldrich RW (2004) Influence of conservation on calculations of amino acid covariance in multiple sequence alignment. Proteins 56:211–221

Giraud BG, Heumann JM, Lapedes AS (1999) Superadditive correlation. Phys Rev E 59:4973–4991

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

Gulyás-Kovács A (2012) Integrated analysis of residue coevolution and protein structure in ABC transporters. PLoS ONE 7(5):e36546. https://doi.org/10.1371/journal.pone.0036546

Halabi N, Rivoire O, Leibler S, Ranganathan R (2009) Protein sectors: evolutionary units of three-dimensional structure. Cell 138:774–786

Havel TF, Kuntz ID, Crippen GM (1983) The combinatorial distance geometry method for the calculation of molecular conformation. I. A new approach to an old problem. J Theor Biol 104:359–381

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

Hopf TA, Schärfe CPI, Rodrigues JPGLM, Green AG, Kohlbacher O, Bonvin, AMJJ, Sander C, Marks DS (2014) Sequence co-evolution gives 3D contacts and structures of protein complexes. eLife 3:e03430. https://doi.org/10.7554/eLife.03430

Hopf TA, Ingraham JB, Poelwijk FJ, Schärfe CPI, Springer M, Sander C, Marks DS (2017) Mutation effects predicted from sequence co-variation. Nature Biotech 35:128–135. https://doi.org/10.1038/nbt.3769

Ingraham J, Marks D (2016) Variational inference for sparse and undirected models. arXiv:1602.03807 [stat.ML]

Jacquin H, Gilson A, Shakhnovich E, Cocco S, Monasson R (2016) Benchmarking inverse statistical approaches for protein structure and design with exactly solvable models. PLoS Comput Biol 12:e1004889. https://doi.org/10.1371/journal.pcbi.1004889

Johnson LS, Eddy SR, Portugaly E (2010) Hidden Markov model speed heuristic and iterative HMM search procedure. BMC Bioinf 11:431

Jones DT (2001) Predicting novel protein folds by using FRAGFOLD. Proteins 45(S5):127–132

Jones DT, Bryson K, Coleman A, McGuffin LJ, Sadowski MI, Sodhi JS, Ward JJ (2005) Prediction of novel and analogous folds using fragment assembly and fold recognition. Proteins 61(S7):143–151. https://doi.org/10.1002/prot.20731

Jones DT, Buchan DWA, Cozzetto D, Pontil M (2012) PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. Bioinformatics 28:184–190. https://doi.org/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 31:999–1006. https://doi.org/10.1093/bioinformatics/btu791

Kaján L, Hopf TA, Kalaš M, Marks DS, Rost B (2014) FreeContact: fast and free software for protein contact prediction from residue co-evolution. BMC Bioinf 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 110:15674–15679. https://doi.org/10.1073/pnas.1314045110

Kim DE, Chivian D, Baker D (2004) Protein structure prediction and analysis using the Rosetta server. Nucl Acid Res 32:W526–W531

Kim DE, Blum B, Bradley P, Baker D (2009) Sampling bottlenecks in *de novo* protein structure prediction. J Mol Biol 393:249–260

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

Kosciolek T, Jones DT (2016) Accurate contact predictions using covariation techniques and machine learning. Proteins 84(S1):145–151. https://doi.org/10.1002/prot.24863

Lapedes AS, Giraud BG, Liu LC, Stormo GD (1999) Correlated mutations in protein sequences: phylogenetic and structural effects. In: Seillier-Moiseiwitsch F (ed) IMS lecture notes: statistics in molecular biology and genetics: selected proceedings of the joint AMS-IMS-SIAM summer conference on statistics in molecular biology, 22–26 June 1997, pp 345–352. Institute of Mathematical Statistics

Lapedes A, Giraud B, Jarzynsk C (2002) Using sequence alignments to predict protein structure and stability with high accuracy. LANL Sciece Magagine LA-UR-02-4481

Lapedes A, Giraud B, Jarzynsk C (2012) Using sequence alignments to predict protein structure and stability with high accuracy. arXiv:1207.2484 [q-bio.QM]

Maisnier-Patin S, Andersson DI (2004) Adaptation to the deleterious effect of antimicrobial drug resistance mutations by compensatory evolution. Res Microbiol 155:360–369

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. https://doi.org/10.1371/journal.pone.0028766

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

Martin LC, Gloor GB, Dunn SD, Wahl LM (2005) Using information theory to search for co-evolving residues in proteins. Bioinformatics 21:4116–4124

Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E (1953) Equation of state calculations by fast computing machines. J Chem Phys 21:1087–1092

Miyazawa S (2013) Prediction of contact residue pairs based on co-substitution between sites in protein structures. PLoS ONE 8(1):e54252. https://doi.org/10.1371/journal.pone.0054252

Miyazawa S (2017a) Prediction of structures and interactions from genome information. arXiv:1709.08021 [q-bio.BM]

Miyazawa S (2017b) Selection originating from protein stability/foldability: relationships between protein folding free energy, sequence ensemble, and fitness. J Theor Biol 433:21–38. https://doi.org/10.1016/j.jtbi.2017.08.018

Miyazawa S, Jernigan RL (1996) Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term for simulation and threading. J Mol Biol 256:623–644. https://doi.org/10.1006/jmbi.1996.0114

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:E1293–E1301. https://doi.org/10.1073/pnas.1111471108

Morcos F, Schafer NP, Cheng RR, Onuchic JN, Wolynes PG (2014) Coevolutionary information, protein folding landscapes, and the thermodynamics of natural selection. Proc Natl Acad Sci USA 111:12408–12413. https://doi.org/10.1073/pnas.1413575111

Moult J, Fidelis K, Kryshtafovych A, Schwede T, Tramontano A (2016) Critical assessment of methods of protein structure prediction: progress and new directions in round XI. Proteins 84(S1):4–14. https://doi.org/10.1002/prot.25064

Nugent T, Jones DT (2012) Accurate *de novo* structure prediction of large transmembrane protein domains using fragmentassembly and correlated mutation analysis. Proc Natl Acad Sci USA 109:E1540–E1547. https://doi.org/10.1073/pnas.1120036109

Ovchinnikov S, Kim DE, Wang RYR, Liu Y, DiMaio F, Baker D (2016) Improved *de novo* structure prediction in CASP11 by incorporating coevolution information into Rosetta. Proteins 84(S1):67–75. https://doi.org/10.1002/prot.24974

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

Pollock DD, Taylor WR (1997) Effectiveness of correlation analysis in identifying protein residues undergoing correlated evolution. Protein Eng 10:647–657

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

Poon AFY, Lewis FI, Frost SDW, Kosakovsky Pond SL (2008) Spidermonkey: rapid detection of co-evolving sites using Bayesian graphical models. Bioinformatics 24:1949–1950

Remmert M, Biegert A, Hauser A, Söding J (2012) HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nat Methods 9:173–175

Riedmiller M, Braun H (1993) A direct adaptive method for faster backpropagation learning: the RPROP algorithm. IEEE Int Conf Neural Netw 1993:586–591

Russ WP, Lowery DM, Mishra P, Yaffe MB, Ranganathan R (2005) Natural-like function in artificial WW domains. Nature 437:579–583

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

Sfriso P, Duran-Frigola M, Mosca R, Emperador A, Aloy P, Orozco M (2016) Residues coevolution guides the systematic identification of altemative functional conformations in proteins. Structure 24:116–126. https://doi.org/10.1016/j.str.2015.10.025

Shendure J, Ji H (2017) EPSILON-CP: using deep learning to combine information from multiple sources for protein contact prediction. BMC Bioinf 18:303. https://doi.org/10.1186/s12859-017-1713-x

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

Skerker JM, Perchuk BS, Siryapom A, Lubin EA, Ashenberg O, Goulian M, Laub MT (2008) Rewiring the specificity of two-component signal transduction systems. Cell 133:1043–1054

Skwark MJ, Abdel-Rehim A, Elofsson A (2013) PconsC: combination of direct information methods and alignments improves contact prediction. Bioinformatics 29:1815–1816

Skwark MJ, Raimondi D, Michel M, Elofsson A (2014) Improved contact predictions using the recognition of protein like contact patterns. PLoS Comput Biol 10:e1003889. https://doi.org/10.1371/journal.pcbi.1003889

Skwark MJ, Michel M, Hurtado DM, Ekeberg M, Elofsson A (2016) Accurate contact predictions for thousands of protein families using PconsC3. bioRXiv. https://doi.org/10.1101/079673

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

Sutto L, Marsili S, Valencia A, Gervasio FL (2015) From residue coevolution to protein conformational ensembles and functional dynamics. Proc Natl Acad Sci USA 112:13567–13572. https://doi.org/10.1073/pnas.1508584112

Talavera D, Lovell SC, Whelan S (2015) Covariation is a poor measure of molecular coevolution. Mol Biol Evol 32:2456-2468. https://doi.org/10.1093/molbev/msv109

Taylor WR, Sadowski MI (2011) Structural constraints on the covariance matrix derived from multiple aligned protein sequences. PLoS ONE 6(12):e28265. https://doi.org/10.1371/journal.pone.0028265

Tokuriki N, Tawfik DS (2009) Protein dynamism and evolvability. Science 324:203–207

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 167:158–170. https://doi.org/10.1016/j.cell.2016.09.010

Tufféry P, Darlu P (2000) Exploring a phylogenetic approach for the detection of correlated substitutions in proteins. Mol Biol Evol 17:1753–1759

Wang S, Sun S, Li Z, Zhang R, Xu J (2017) Accurate *de novo* prediction of protein contact map by ultra-deep learning model. PLoS Comput Biol 13:e1004324. https://doi.org/10.1371/journal.pcbi.1005324

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 106:67–72. https://doi.org/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 165:1–13. https://doi.org/10.1016/j.cell.2016.03.030

Wuyun Q, Zheng W, Peng Z, Yang J (2016) A large-scale comparative assessment of methods for residue-residue contact prediction. Brief Bioinform 19:219–230. https://doi.org/10.1093/bib/bbw106

Yanovsky C, Hom V, Thorpe D Protein structure relationships revealed by mutation analysis. Science 146:1593–1594 (1964)