

How Pairwise Coevolutionary Models Capture the Collective Residue Variability in Proteins?

Matteo Figliuzzi,¹ Pierre Barrat-Charlaix,¹ and Martin Weigt^{*,1}

¹Sorbonne Université, CNRS, Institut de Biologie Paris Seine, Computational and Quantitative Biology – UMR7238, 75005 Paris, France

*Corresponding author: E-mail: martin.weigt@upmc.fr.

Associate editor: Claus Wilke

Abstract

Global coevolutionary models of homologous protein families, as constructed by direct coupling analysis (DCA), have recently gained popularity in particular due to their capacity to accurately predict residue–residue contacts from sequence information alone, and thereby to facilitate tertiary and quaternary protein structure prediction. More recently, they have also been used to predict fitness effects of amino-acid substitutions in proteins, and to predict evolutionary conserved protein–protein interactions. These models are based on two currently unjustified hypotheses: 1) correlations in the amino-acid usage of different positions are resulting collectively from networks of direct couplings; and 2) pairwise couplings are sufficient to capture the amino-acid variability. Here, we propose a highly precise inference scheme based on Boltzmann-machine learning, which allows us to systematically address these hypotheses. We show how correlations are built up in a highly collective way by a large number of coupling paths, which are based on the proteins three-dimensional structure. We further find that pairwise coevolutionary models capture the collective residue variability across homologous proteins even for quantities which are not imposed by the inference procedure, like three-residue correlations, the clustered structure of protein families in sequence space or the sequence distances between homologs. These findings strongly suggest that pairwise coevolutionary models are actually sufficient to accurately capture the residue variability in homologous protein families.

Key words: coevolution, direct coupling analysis, global statistical inference, Boltzmann machine learning.

Introduction

In the course of evolution, proteins may substitute the vast majority of their amino acids without losing their three-dimensional structure and their biological functionality. Rapidly growing sequence databases provide us with ample examples of such evolutionary related, that is, homologous proteins, frequently already classified into protein families and aligned into large multiple-sequence alignments (MSA). Typical pairwise sequence identities between homologous proteins go down to 20–30%, or even below (Finn et al. 2014). Such low sequence identities are astonishing since even very few random mutations may destabilize a protein or disrupt its functionality.

Assigning a newly sequenced gene or protein to one of these families helps us to infer functional annotations. Structural homology modeling, for example, belongs to the most powerful tools for protein-structure prediction (Arnold et al. 2006; Webb and Sali 2014). However, beyond the transfer of information, the variability of sequences across homologs itself contains information about evolutionary pressures acting in them, and statistical sequence models may unveil that information (Durbin et al. 1998; de Juan et al. 2013).

A first level of information is contained in the variability of individual residues: low variability, that is, conservation, frequently identifies functionally or structurally important sites in a protein. This information is used by so-called profile

models (Durbin et al. 1998), which reproduce independently the amino-acid statistics in individual MSA columns. They belong to the most successful tools in bioinformatics; they are at the basis of most techniques for multiple-sequence alignment and homology detection, partially as profile Hidden-Markov models accounting also for amino-acid insertions and deletions (Eddy 1998).

A second level of information is contained in the covariation between pairs of residues, measurable via the correlated amino-acid usage in pairs of MSA columns (de Juan et al. 2013; Cocco et al. 2017). Covariation cannot be captured by profile models, as they treat residues independently. To overcome this limitation, global statistical models with pairwise couplings—exploiting residue conservation and covariation—have recently become popular. Methods like the direct coupling analysis (DCA) (Weigt et al. 2009; Morcos et al. 2011), PsiCov (Jones et al. 2012), or Gremlin (Balakrishnan et al. 2011) allow for the prediction of residue–residue contacts using sequence information alone, and can be used to predict three-dimensional protein structures (Marks et al. 2012; Ovchinnikov et al. 2017) and to assemble protein complexes (Schug et al. 2009; Hopf et al. 2014; Ovchinnikov et al. 2014). Currently, these methods are the central element of various of the best-performing residue-contact predictors in the CASP competition for protein structure prediction (Jones et al. 2015; Wang et al. 2017).

Despite their success in practical applications, not much is known about the reasons for this success and their intrinsic limitations. Typically, two hypotheses are made: 1) The correlated amino-acid usage in two MSA columns may result from a direct residue–residue contact in the protein structure, causing coordinated amino-acid changes to maintain the protein’s stability. It may also result indirectly via intermediate residues, making the direct use of covariation for contact prediction impractical. The success of global models is attributed to their capacity to extract direct couplings from indirect correlations. 2) Using the maximum-entropy principle, the simplest models reproducing pairwise residue covariation depend on statistical couplings between residue pairs. Whether or not this model is sufficient to capture also higher-order covariation remains currently unclear.

So far, these two points have not been investigated systematically. The reason is relatively simple: The inference of pairwise models exactly reproducing the empirical conservation and covariation statistics extracted from an MSA requires to sum over all 20^L sequences of aligned length L , an unfeasible task for sequences of typical sizes $L = 50$ – 500 . Approximation schemes like mean-field approximation (Morcos et al. 2011), Gaussian approximation (Jones et al. 2012), or pseudo-likelihood maximization (Balakrishnan et al. 2011; Ekeberg et al. 2013) have been introduced; they perform very well in contact prediction. Their approximate character prohibits, however, the analysis of higher-order correlations and collective effects, since even the pairwise statistics are not well reproduced. More precise approaches have been proposed recently (Sutto et al. 2015; Barton et al. 2016; Haldane et al. 2016), but their high computational cost has limited applications mostly to anecdotal cases so far.

Understanding these basic questions is essential for understanding the success of global coevolutionary models beyond “black box” applications, but also for recognizing their current limitations and thus potentially to open a way towards improved statistical modeling schemes. To this end, we implement a highly precise approach for parameter inference in pairwise statistical models. Applying this approach to a number of very large protein families (containing sufficient sequences to reliably measure higher-order statistical features), we demonstrate that indirectly generated pair correlations are highly collective effects of entire networks of direct couplings, which are based on the structural vicinity between residues.

However, the most interesting finding of the article is the unexpected accuracy of DCA at reproducing higher-order statistical features, which are not fitted by our approach. These nonfitted features include connected three-point correlations, the distance distributions between natural sequences and between artificial sequences sampled from the model, or the clustered organization of sequences in sequence space. Currently we do not find indications, that more involved models (e.g., including three-residue interactions) are needed to reproduce the full sequence statistics: pairwise models are not only necessary as argued above, but seem to be sufficient to describe the sequence variability between homologous proteins.

Results

Direct Coupling Analysis—Methodology and Approximate Solutions

The aim of global coevolutionary sequence models as constructed by DCA is to provide a protein family-specific probability distribution

$$P(\underline{A}) \propto \exp \left(\sum_{j>i} J_{ij}(A_i, A_j) + \sum_{i=1}^N h_i(A_i) \right) \quad (1)$$

for all full-length amino-acid sequences $\underline{A} = (A_1, \dots, A_L)$. To model sequence variability in an MSA, couplings $J_{ij}(A, B)$ and biases (fields) $h_i(A)$ have to be fitted such that model $P(\underline{A})$ reproduces the empirically observed frequencies $f_i(A)$ of occurrence of amino acid A in the i th MSA column, and cooccurrence $f_{ij}(A, B)$ of amino acids A and B in positions i and j of the same sequence. In other words, the DCA model has to satisfy

$$P_i(A) = f_i(A) \quad \text{and} \quad P_{ij}(A, B) = f_{ij}(A, B) \quad (2)$$

for all columns i, j and all amino acids A, B , with P_i and P_{ij} being marginal distributions of model $P(\underline{A})$, cf. [supplementary methods](#) and [section 1, Supplementary Material](#) online.

Equation (2) has two important consequences. First, a precisely inferred DCA model reproduces also pairwise connected correlations (or covariances) $c_{ij}(A, B) = f_{ij}(A, B) - f_i(A)f_j(B)$ found in the MSA. This is a crucial difference with profile models, which show vanishing connected correlations by construction. Second, the inference of $P(\underline{A})$ via [equation \(2\)](#) does *not* use all the information contained in the MSA, but only the pairwise statistics. For this reason, model $P(\underline{A})$ has *a priori* no reason to reproduce any higher-order statistics contained in the alignment. In particular, even though a model of the form of [equation \(1\)](#) will contain higher-order correlations, such as three-residue correlations, these may differ significantly from those found in the original MSA.

To infer DCA parameters, we need to estimate marginal probabilities for single positions and position pairs from model $P(\underline{A})$. Exact calculations of these marginals require to perform exponential sums over q^L terms, with L being the sequence length, and $q = 21$ enumerating amino acids and the alignment gap. These sums are infeasible even for short protein sequences, and have been replaced by approximate expressions, for example, via mean-field (Morcos et al. 2011), Gaussian (Jones et al. 2012), or pseudo-likelihood approximations (Balakrishnan et al. 2011; Ekeberg et al. 2013). These approximations are sufficiently accurate for residue-contact prediction, which is topological in nature: only the existence of a strong direct statistical coupling has to be detected, not necessarily its precise numerical value. As a consequence, these methods do not reproduce the empirical frequencies and thus do not satisfy [equation \(2\)](#), cf. [figure 1](#) for pseudo-likelihood maximization (plmDCA [Ekeberg et al. 2013]). More precise methods based on an adaptive cluster expansion (Barton et al. 2016) or Boltzmann machine learning using Markov-chain Monte Carlo sampling (Ackley et al. 1985) for estimating marginal distributions have

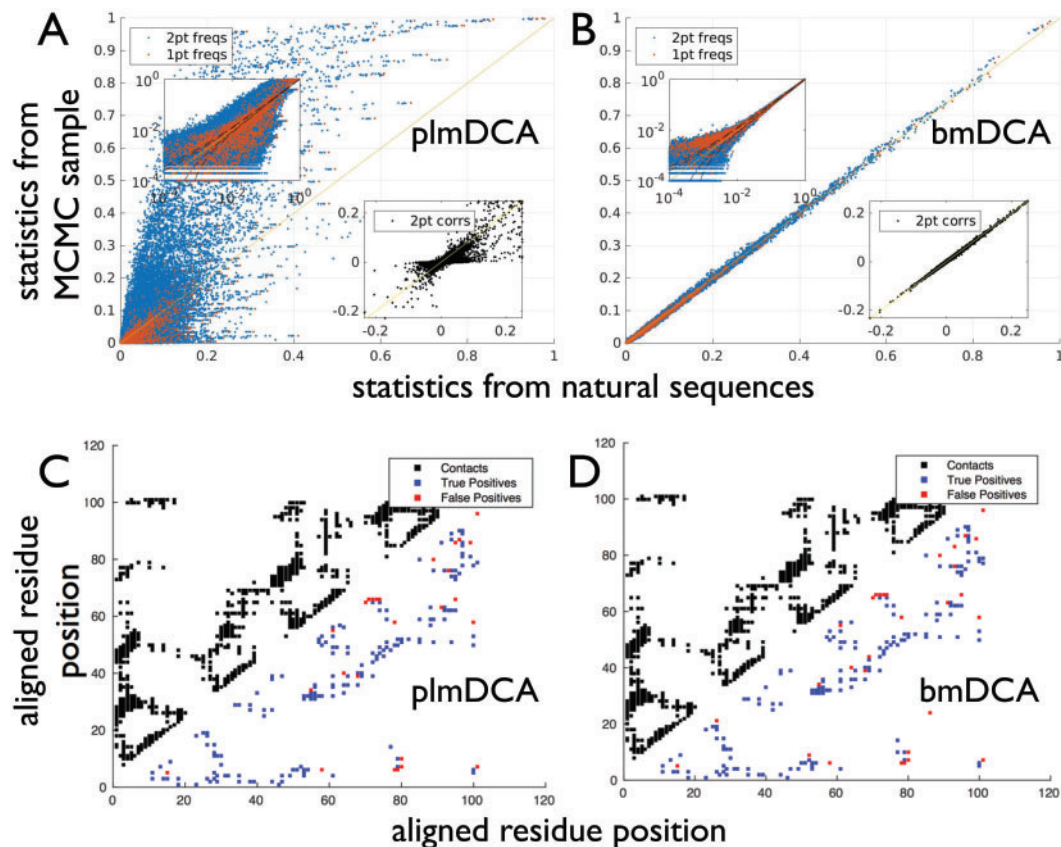


Fig. 1. Fitting accuracy and contact prediction for DCA models inferred using pseudo-likelihood maximization (plmDCA) and Boltzmann-machine learning (bmDCA): While the Potts model inferred by plmDCA (A) fails to reproduce the one- and two-residue frequencies (main panel) and the connected two-point correlations (lower left insert) in the PF00072 protein family, the model inferred using our bmDCA algorithm (B) is very accurate. Slight deviations visible for very small frequencies in log-scale (upper right insert) are results of the ℓ_2 -regularization penalizing strongly negative couplings. Despite these differences, the contact predictions (C for plmDCA and D for bmDCA) relying on the strongest $2L = 224$ DCA couplings (with $|i - j| > 4$) are close to identical: native contacts (all-atom distance below 8 Å) are shown above the diagonal, predicted contacts (below the diagonal). Very similar results are observed across all studied protein families, cf. [supplementary sections 5.1 and 5.5, Supplementary Material online](#) (color online).

been proposed recently (Sutto et al. 2015; Haldane et al. 2016). While decreasing deviations from equation (2) substantially (i.e., fitting quality), they are typically much more computationally expensive and not suitable for large-scale application to hundreds or thousands of protein families.

Accurate Fitting Is Needed to Reproduce the Empirical Residue Covariation in Homologous Protein Families

Since the aim of the current article is to unveil the way DCA disentangles direct couplings and indirect correlations, and to investigate if it captures higher-order statistical observables estimated from the MSA, we have implemented an efficient version of Boltzmann machine (BM) learning described in “Materials and Methods” and, in full detail, in [supplementary section 2, Supplementary Material online](#). In short, BM learning estimates the pairwise marginal distributions of $P(\mathbf{A})$ by Monte-Carlo sampling, and iteratively updates model parameters until equation (2) is satisfied. In contrast to approximations such as applied in plmDCA, the inference of parameters using BM learning can be made arbitrarily accurate, provided that Monte-Carlo samples are large enough and sufficient iterations are performed. In analogy to earlier notation, we will use bmDCA for the

resulting implementation of DCA. As is shown in [figure 1](#) and in [table 1](#), bmDCA reaches very accurate fitting, approaching the statistical uncertainties related to the finite sample size (i.e., the sequence number in each MSA), even for the very large protein families studied here. Obviously bmDCA has a higher computational cost than plmDCA: While plmDCA achieves inference typically in few minutes, bmDCA needs few hours to several days for one family, in dependence of the sequence length and the required fitting accuracy.

Interestingly, the increased fitting accuracy does not improve the contact prediction beyond the one of plmDCA, the currently best unsupervised DCA contact predictor, cf. [figure 1C and D](#). Couplings $J_{ij}(A, B)$ are highly correlated between PLM and BM (Pearson correlations of 90–98% across all studied protein families), in particular large couplings are robust and lead to very similar contact predictions. However, the model statistics depends collectively on all $\mathcal{O}(q^2L^2)$ parameters and can thus differ substantially even for small differences in the individual parameters. This sensitivity (so-called criticality) has also been observed in other models inferred from large-scale biological data, cf. (Mora and Bialek 2011).

Table 1. Results for the Ten Selected Protein Families.

Protein Family				Fitting Quality		Contact Prediction		Three-Point Correlations		Collectivity of Correlations		
Pfam	L	M	PDB	PLM	BM	PLM	BM	PLM	BM	corr(DI, MI)	corr(L2I, MI)	ν
PF00004	132	39277	4D81	0.630	0.954	0.672	0.672	0.333	0.980	0.33	0.42	1.2
PF00005	137	68891	1L7V	0.546	0.948	0.599	0.586	0.718	0.978	0.51	0.65	1.4
PF00041	85	42721	3UP1	0.897	0.973	0.715	0.671	0.893	0.991	0.61	0.77	1.7
PF00072	112	73063	3ILH	0.670	0.978	0.836	0.842	0.803	0.988	0.52	0.69	1.4
PF00076	69	51964	2CQD	0.868	0.977	0.877	0.833	0.963	0.993	0.53	0.72	1.5
PF00096	23	38996	2LVH	0.954	0.987	0.657	0.711	ND	ND	0.95	0.99	2.3
PF00153	97	54582	2LCK	0.800	0.967	0.601	0.563	0.517	0.986	0.45	0.57	1.1
PF01535	31	60101	4G23	0.902	0.994	0.630	0.739	0.120	0.996	0.70	0.91	1.5
PF02518	111	80714	3G7E	0.624	0.970	0.423	0.396	−0.228	0.986	0.47	0.60	1.6
PF07679	90	36141	1FHG	0.823	0.955	0.826	0.826	0.797	0.993	0.48	0.58	1.8

NOTE.—The first four columns give the ID of the selected protein families together with the sequence length L , alignment depth M and a representative protein structure. The fitting quality measures the Pearson correlation between connected two-point correlations in the natural data, and in a sample drawn from the Potts models inferred by plmDCA and bmDCA (better quality emphasized in boldface). The contact prediction gives the fraction of true positives (all-atom distance $< 8 \text{ \AA}$) within the first 2L predictions. Columns 9 and 10 provide the Pearson correlation between connected three-point correlations observed in natural and in sampled sequences (due to the dominance of insignificantly small terms, only those with $c_{ijk}^{MSA}(A, B, C) > 0.01$ are considered). PF00096, with only 23 aligned positions is the shortest considered protein family, has no significant three-point correlations, neither in the data nor in the Potts model. The last three columns quantify the collective nature of correlations: the Pearson correlation of direct information/mutual information as compared to the length-two information/mutual information, and the exponent of the approximate power-law decay of the strongest paths (in terms of their path information) with their ranking.

Indirect Correlations Result Collectively from Networks of Direct Couplings

bmDCA provides a highly accurate approach to describe the sequence variability of homologous proteins via a pairwise coevolutionary model. This implementation allows us to ask fundamental questions about how DCA works, its capacities and its possible limitations, without being biased by the specificities of approximate DCA implementations.

The success of global models as inferred by DCA is typically attributed to the idea that they disentangle statistical correlations, which are empirically observed in an MSA and measured via the mutual information (MI), into a network of direct couplings between residues. The strongest direct couplings are biologically interpretable as residue–residue contacts in the three-dimensional protein structure. However, this idea, even if stated in many papers on the subject, has never been examined in detail, and important questions remain unanswered: can indirect effects be explained by a few strong coupling chains, or are they distributed over networks of numerous small couplings? Are these networks structurally interpretable, that is, in relation to a proteins contact map?

Correlations Are Mediated Collectively by Distributed Networks of Coevolutionary Couplings

To answer the first question, we need to quantify the correlation induced by a coupling chain of arbitrary length, connecting any two residues. To this aim, we take inspiration from the concept of direct information (DI) introduced in (Weigt et al. 2009). DI is a proxy of the strength of the direct interaction J_{ij} between two residue positions i and j ; it measures the correlation that i and j would have if they were only connected by J_{ij} , cf. figure 2A. To measure the indirect correlation between i and j induced via a chain of intermediate residues, we introduce the concept of path information (PI), as illustrated again in figure 2A and defined in “Materials and Methods”. Now, for each protein family, we extracted the 100 most correlated residue pairs (highest MI). Using a

modification of Dijkstra’s shortest-path algorithm (Dijkstra 1959)—which becomes approximate due to the nonadditivity of PI but delivers highly reliable results as shown in supplementary section 3, Supplementary Material online—we extracted for each residues pairs the 15 strongest coupling paths (highest PI) connecting the two residues.

In figure 2B, we show that the decrease of the average strength of the k th strongest path is compatible with a slowly decreasing power law, $\langle PI(k) \rangle \propto k^{-\nu}$, with exponents ν between 1.1 and 2.3. While this fit is only approximate, as visible by the strong deviations for the strongest path at $k = 1$, its slow decay clearly shows that the correlation between two residues typically is not mediated by one or few coupling chains. On the contrary, indirect effects emerge collectively, in the sense that a large number of partially overlapping coupling chains have to be taken into account, each one contributing only a small fraction to the total correlation. It is important to note that the strongest path (rank $k = 1$) is on average much stronger than the others and clearly does not fall onto a power law. For the overwhelming majority of the pairs, this strongest path is the direct one containing only one coupling. Its contribution to the total correlation is, on average, about 12.5% of the total MI. This average is dominated by the shortest protein families, PF00096 and PF01535, who are expected to show less collectivity due to their small number L of aligned residues.

On the Structural Basis of Coevolutionary Coupling Networks

As a consequence of the last section, we need to consider the collective effect of multiple paths rather than trying to biologically interpret individual paths beyond the direct one. While this is technically very hard in general, the collective effect of all paths of length two is efficiently computable, cf. “Materials and Methods.” The corresponding correlation measure, named here length-two information (L2I) and illustrated in figure 2A, adds the $L - 2$ possible indirect paths of

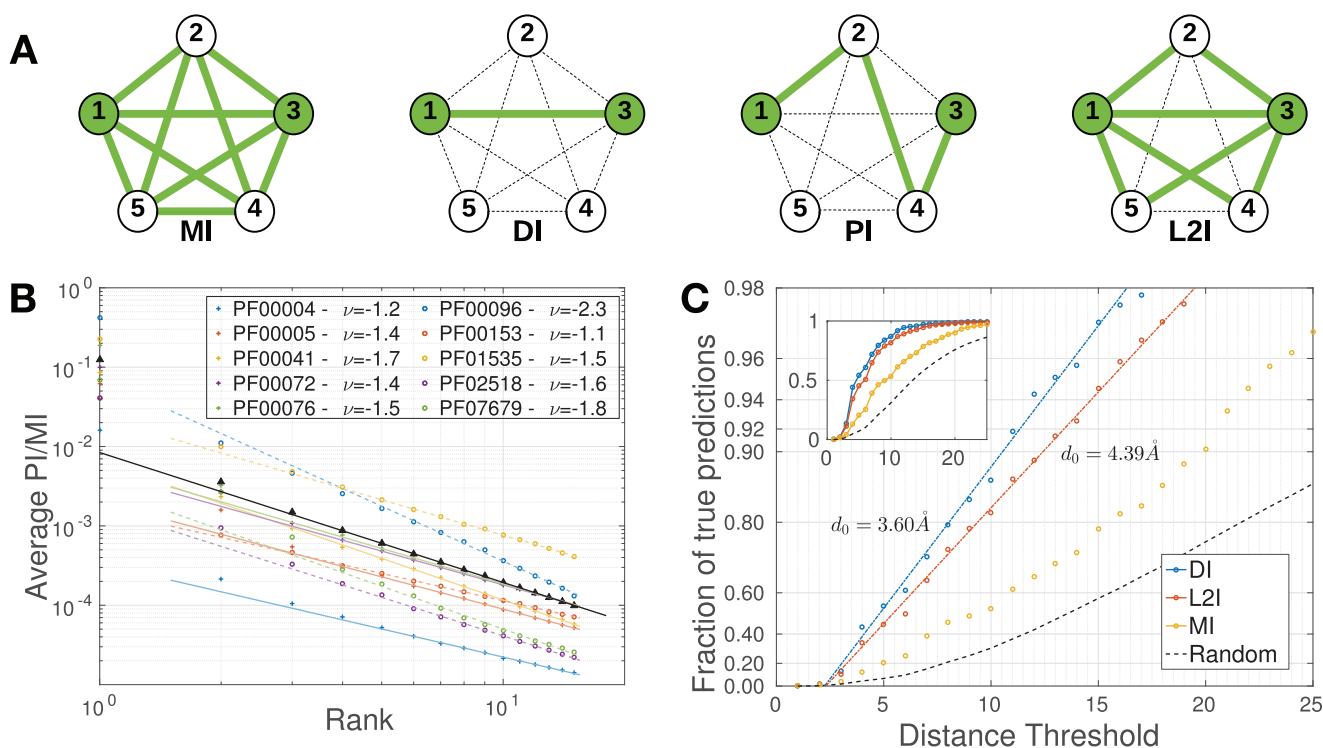


Fig. 2. Collective nature of the correlation between two residue positions: (A) Illustration of the correlation measures used in this study. While the mutual information MI depends collectively on the entire network of coevolutionary couplings, the direct information DI is obtained by taking into account only the single direct coupling between the sites of interest (e.g., 1 and 3 in the figure). All other couplings are formally set to zero. The path information PI is the direct generalization of DI to the correlation mediated by a single path (e.g., [1, 2, 4, 3] in the figure). The length-two information L2I measures the collective effect of the direct coupling and all length-two paths (e.g., [1, k , 3] with $k = 2, 4, 5$). (B) A log-log plot of the average ratio of path information to mutual information (triangular symbols and black fat line: average over all families) as a function of the corresponding path, showing a very slow (approximately power-law) decay. This illustrates the fact that indirect correlations do not depend on a single (or very few) coupling chains, but are distributed over coupling networks. (C) For the 25 highest ranking residue pairs according to DI, L2I, and MI, the fraction of pairs of distance below d , as a function of d . The scale on the y-axis is logarithmic, and chosen in a way that functions of the form $1 - e^{-d/d_0}$ will appear as straight lines, the insert shows a standard linear scale. For DI and PI, these curves show a clear exponential convergence to 1, with characteristic distance scales of 3.6 resp. 4.4 Å. MI does not show any exponential behavior, and thus no characteristic distance scale (color online).

length 2 (one intermediate residue) to the direct path between two residue positions. As expected, L2I captures a much higher fraction of the full mutual information than DI, cf. table 1. However, a large fraction of the mutual information is not yet covered. It is contributed by longer coupling chains: L2I depends only on $2L - 3$ out of the $L(L - 1)/2$ couplings between residue pairs. Consistent with this observation, the correlation of L2I with MI is much larger in small proteins, and decreases when going to larger proteins.

L2I allows for an interesting structural interpretation. It is well established that large DI are good predictors for native residue contacts. Is large L2I a good predictor of second neighbors in the protein structure, that is, of residue pairs which are two contacts away? To investigate this question, the blue line in figure 2C displays the fraction of true positive predictions (positive predictive value, averaged over the protein ensemble) within the highest 25 DI as a function of a distance cutoff d , which varies between 1 and 25 Å. It starts at 0 for small d , and approaches 1 exponentially with a scaling $1 - \exp(-d/d_0)$ of characteristic length $d_0 = 3.6 \text{ \AA}$. At 8 Å distance (typically used as contact definition in DCA studies),

an accurate prediction of about 85% true positives (TP) and only 15% false positives (FP) is reached. Measuring the cut-off dependent positive predictive value for the length-two information L2I, we find again an exponential behavior but with characteristic length $d_0 = 4.4 \text{ \AA}$. The fraction of TP therefore reaches 85% only between 11 and 12 Å, a distance compatible with second structural neighbors. The finding that the top DI are dominated by direct contacts, and large L2I by residue pairs which are up to second neighbors in the structure, further underlines the structural basis of coevolutionary constraints as captured by DCA. We also note that the full correlation MI—depending on coupling chains of all possible lengths—does not imply an exponential behavior in figure 2, and no characteristic length scale can be identified.

Pairwise Coevolutionary Models Accurately Reproduce the Residue Variability beyond the Fitted Two-Residue Statistics

Profile models assuming independent residues are not able to extract the full information contained in the MSA of a protein family. In particular, the inclusion of pairwise coevolutionary

couplings is required for the prediction of intra or inter-protein residue–residue contacts, which has become the most important application of coevolutionary modeling. Furthermore, studies about protein mutational effects (Levy et al. 2017) and the prediction of protein–protein interactions (Szurmant and Weigt 2018) have underlined the importance of pairwise couplings.

Is there information hidden in large MSA, which cannot be captured by pairwise models? Does one need to include higher-order couplings into the modeling? The highly accurate inference of pairwise models obtained by bmDCA, reproducing faithfully the empirical first- and second-order statistics, allows to address these questions systematically. To this aim, we use MCMC samples from the inferred models to compare statistical observables, which are not a direct consequence of the fitted covariances. These comparisons unveil the astonishing capacity of bmDCA to capture local and global statistical features, which are not explicitly fitted by the model: pairwise couplings are not only necessary for characterizing sequence variability between homologs, but they also seem to be sufficient.

First, we observe that the *three-residue statistics* is accurately reproduced by our model including only pairwise couplings: figure 3 (cf. supplementary section 5.2, Supplementary Material online, for other families) shows a density-colored scatter plot of the connected three-point correlations of the natural sequences versus the MCMC sample drawn from the model. Correlations are high across all protein families for the pairwise model, with close to perfect Pearson correlations ranging from 0.978 to 0.997, cf. table 1. Profile models, which by definition do not have any connected three-point correlation, can be seen as null model testing the strength of three-point correlations emerging due to finite sampling. As is shown in figure 3D, they are at least one order of magnitude smaller than those found empirically, underlining the significance of our findings. The only exception is family PF00096, where no significant connected three-point correlations are detectable in the MSA or in the sample. Note that we use connected correlations $c_{ijk}(A, B, C) = f_{ijk}(A, B, C) - f_{ij}(A, B)f_k(C) - f_{ik}(A, C)f_j(B) - f_{jk}(B, C)f_i(A) + 2f_i(A)f_j(B)f_k(C)$, which are intrinsically harder to reproduce than three-point frequencies $f_{ijk}(A, B, C)$. Note also that our result is far from being obvious: a Gaussian model with the same covariances would have vanishing three-point correlations, while the sequence data and the sample from our DCA model do not. Further more, it is easy to construct models with discrete variables, whose three-point correlations are not reproduced by a pairwise DCA model. This is shown in supplementary section 4, Supplementary Material online, via analytical calculations and numerical simulations.

To complement the three-point statistics, we investigated more global quantities. The first one is the clustered organization of protein families in sequence space. Figure 3A shows all sequences mapped onto their first two principal components for PF00072 (cf. Supplementary Material online for other families). We observe a clear clustering into at least three distinct subfamilies, which identify different functional subclasses of the PF00072 protein family (single domain vs.

multi-domain architectures with distinct DNA-binding domains). A sample drawn from a profile model does not reproduce this clustered structure (B), while the MCMC sample of the bmDCA model does, including the fine structure of the clusters (C). Again, this structure is not a simple consequence of the empirical covariance matrix as a sample from a Gaussian model with the same covariances would not show any clustering.

As a last measure, we compared the pairwise Hamming distances between sequences in the natural MSA and in the model-generated sequences. Again the pairwise bmDCA model is needed to reproduce the bulk of the empirical distribution of pair distances. Interestingly, a difference between the two becomes visible in the small-distance tail of the histograms in figure 3G: while natural sequences may be close to identical due to a close phylogenetic relation, small sequence distances are never observed in an equilibrium sample of the bmDCA model, that is, a part of the phylogenetic bias present in the MSA is avoided by the bmDCA model.

Discussion

This article unveils a number of reasons behind the success of global pairwise models in extracting information from the sequence variability of homologous protein sequences. First, we show that residue–residue correlations actually result from the collective variability of many residues, and are not the result of a few strong coupling chains. Therefore, local statistical measures taking into account only a small numbers of residues at a time (like correlation measures) are necessarily limited in their capacity to represent the data, and global modeling approaches are needed.

One of the most astonishing findings is that many features of the data, which are not explicitly fitted by a pairwise modeling, are nevertheless well reproduced by the inferred models. This includes higher-order correlations, like the connected three-point correlations considered here, and more general aspects of the distribution of amino-acid sequences like the histogram of pairwise Hamming distances between pairs of sequences or the clustered organization of the sample in sequence space. Interestingly, only the small distances between phylogenetically closely related sequences are not reproduced in a sample drawn from the inferred DCA model. This capacity to reproduce the sequence variability beyond the fitted empirical observables distinguishes the DCA model (fitting one- and two-residue frequencies) from profile models of independent residues (fitting only one-residue frequencies). While the restriction to pairwise models was initially motivated by the limited availability of sequence data—three-point correlations require to estimate frequencies for $21^3 = 9261$ combinations of amino acids or gaps—we find that even for large MSA pairwise models seem to be sufficient to capture collective effects beyond residue pairs.

Note that this argument does not rule out the existence of higher-order residue effects in the underlying evolutionary processes shaping the sequence variability in homologous protein families (cf. Merchan and Nemenman 2016; Schmidt and Hamacher 2017). However, their statistical

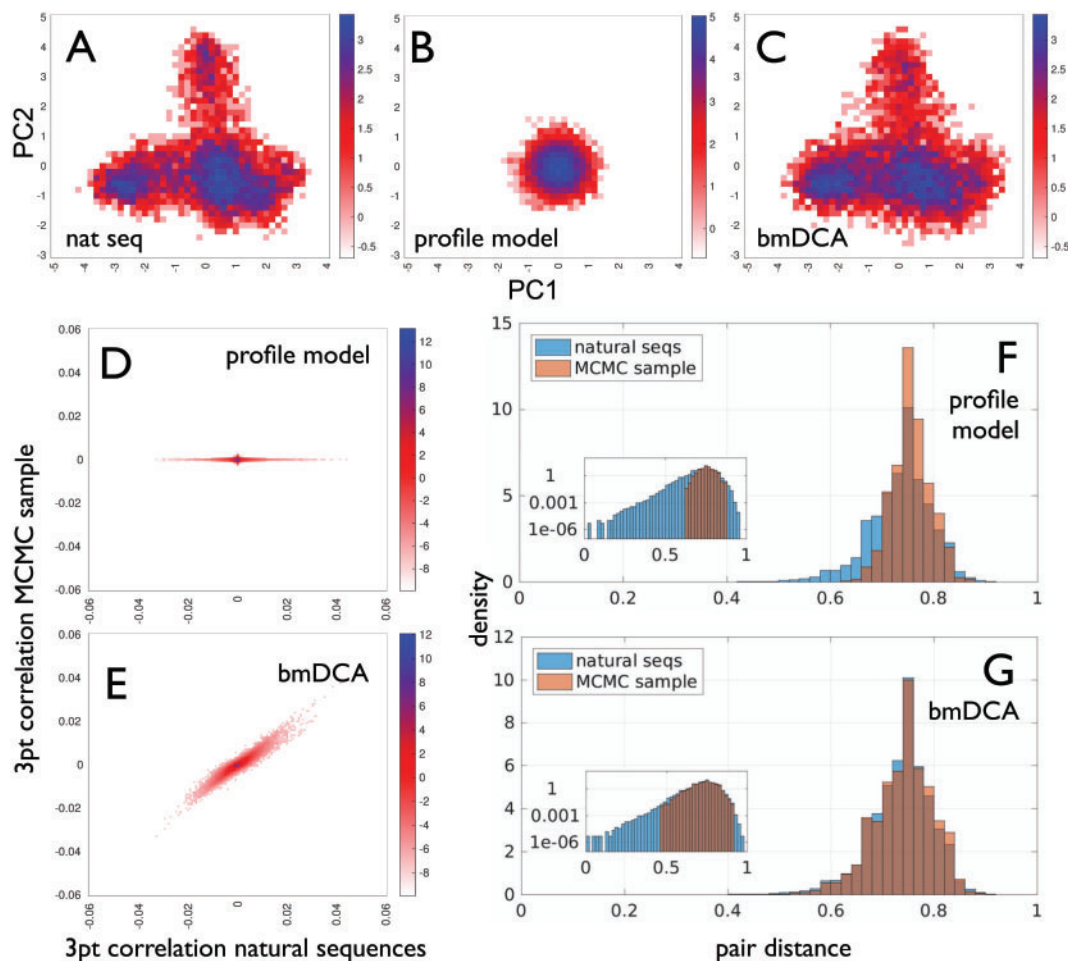


Fig. 3. Nonfitted statistical observables are captured by DCA: (A–C) Natural sequences (PF00072—A) and MCMC samples from inferred profile (B) and bmDCA (C) models are projected on the first two principal components of the natural MSA. (D and E) Three-point correlations of samples of the profile (D) and bmDCA (E) models, as compared to the three-point correlations in the natural sequences. (F and G) Histograms of all pairwise Hamming distances between natural or MCMC sampled sequences, for profile (F) and bmDCA (G) models. Surprisingly bmDCA is able to reproduce all three nonfitted statistical properties of the natural MSA, with the difference of the small distances between close homologs, while the profile model not taking into account residue–residue couplings does not. This suggests that accurately inferred pairwise models are necessary and sufficient to capture the residue variability in families of homologous proteins. Similar results are observed across all studied protein families, as is documented in [supplementary sections 5.2–5.4, Supplementary Material online](#) (color online).

signature is not strong enough to be detectable via deviations from the behavior of a pairwise model, even in the large families considered here. Random samples drawn from a DCA model based exclusively on the knowledge of the empirical one- and two-residue statistics appear to be statistically indistinguishable from natural sequences.

This finding is particularly interesting in the context of work made few years ago by the Ranganathan lab ([Russ et al. 2005; Socolich et al. 2005](#)). Using the small WW domain, they applied a number of diverse procedures to scramble MSA of natural sequences to produce artificial sequences. Scrambling MSA columns to maintain residue conservation while destroying residue correlations, lead in all tested cases to nonfolding amino-acid sequences. A procedure maintaining also pairwise correlations lead to a substantial fraction of folding and functional proteins. Later on it has been observed that the functional artificial sequences actually have the highest probabilities within pairwise coevolutionary models

([Balakrishnan et al. 2011](#)). These findings open interesting roads to evolution-guided protein design ([Reynolds et al. 2013](#)).

Note, however, that the finite size of the input MSA requires to use regularized inference, which penalizes large absolute parameter values. It leads to a small bias visible in [figure 1B](#): small pair frequencies are slightly but systematically overestimated by DCA. This may smoothen the inferred statistical model, cf. ([Otwinowski and Plotkin 2014](#)) for the related case of inferring epistatic fitness landscapes. As a consequence “bad” sequences may be given high probabilities in our model. Based on the findings presented in [figures 1B and 3](#), we expect these effects to be minor. When increasing the regularization strength beyond parameters used in this study, the clustered structure of sampled sequences ([fig. 3C](#)) disappears gradually. Data in large MSA allow to use small regularization, thereby simultaneously limiting overfitting of statistical noise and reducing biases in parameter inference.

This may be impossible for small MSA, so the ongoing growth of sequence databases is key for the wide applicability of global statistical sequence models.

One potentially important limitation remains: the distribution of sequences in sequence space is not only determined by functional constraints acting on amino-acid sequences, but also by phylogenetic relations between sequences. Natural sequences are, even beyond the very closely related sequences not reproduced by the DCA model, far from being an independent sample of all possible amino-acid sequences. They are correlated due to finite divergence times between homologs, and due to the human selection bias in sequenced species. Any model reproducing the full empirical statistics of the MSA describes therefore a mixture of functional and phylogenetic correlations, while an ideal model would contain the functional ones and discard the phylogenetic ones. How these can be disentangled remains an important open question.

Materials and Methods

Protein Families

We have selected ten protein families of known three-dimensional structure which belong to the largest 20 Pfam families (Finn et al. 2014), which are not repeat proteins (i.e., they are not just frequent because repeated many times on the same protein), and have an aligned sequence length below 200 amino acids (for computational reasons), cf. table 1. Sequences with more than 50 alignment gaps are removed. The resulting sequence numbers are reported in table 1. The main reason to include only large Pfam families is the possibility to accurately estimate three-point correlations. For each triplet of residue positions, there are $21^3 = 9,261$ combinations of amino acids or gaps. Nonsystematic tests in smaller protein families show that our main findings of the paper translate directly to these families.

Boltzmann Machine Learning

DCA infers a Potts model

$$P(A_1, \dots, A_L) = \frac{1}{Z} \exp \left\{ \sum_{i < j} J_{ij}(A_i, A_j) + \sum_i h_i(A_i) \right\} \quad (3)$$

reproducing the single- and two-residue frequencies found in the input MSA:

$$\begin{aligned} \sum_{\{A_k | k \neq i\}} P(A_1, \dots, A_L) &= f_i(A_i) \\ \sum_{\{A_k | k \neq i, j\}} P(A_1, \dots, A_L) &= f_{ij}(A_i, A_j) \end{aligned} \quad (4)$$

with empirical frequencies $f_i(A_i)$ and $f_{ij}(A_i, A_j)$ defined, respectively, as the fraction of sequences in the MSA having amino acid A_i (resp. A_i and A_j) in column i (resp. in columns i and j) (cf. supplementary section 1, Supplementary Material online for a precise definition of these frequency counts, including a sequence weighting to reduce phylogenetic biases). For the sake of contact prediction, this inference can be done with efficient approximation schemes like mean-field of pseudo-

likelihood maximization. The objectives of this study—to understand the collective variability of the residues—require a more precise inference based on the classical ideas of Boltzmann-machine learning (Ackley et al. 1985). It consists of an iterative procedure where

- (i) for a given set of model parameters $\{J_{ij}, h_i\}$, Markov-chain Monte Carlo (MCMC) sampling is used to estimate the one- and two-point frequencies of the model;
- (ii) parameters are adjusted when the estimated model frequencies deviate from the empirical ones.

To reduce finite-sample effects, the model parameters are subject to an ℓ_2 -regularization. The likelihood function is convex, guaranteeing convergence to a single globally optimal solution, which reproduces the empirical one- and two-point frequencies with arbitrary accuracy. The direct implementation of Boltzmann-machine learning is computationally very slow. We have therefore introduced a reparameterization of the model, which allows to replace the gradient ascent of the likelihood by a faster pseudo-Newtonian method. Technical details of the implementation are described in supplementary section 2, Supplementary Material online.

From Direct Couplings to Indirect Correlations

Quantifying the Strength of a Coupling Chain

To quantify the strength of a coupling chain, we generalize the direct information introduced in (Weigt et al. 2009). There, the direct probability

$$P_{ij}^{dir}(A_i, A_j) = \exp \{ J_{ij}(A_i, A_j) + \tilde{h}_i(A_i) + \tilde{h}_j(A_j) \} / Z_{ij} \quad (5)$$

was defined as the hypothetical distribution of two residues i and j connected only by the inferred direct coupling J_{ij} and having the empirical single-residue frequencies $f_i(A_i)$ and $f_j(A_j)$, thereby removing all indirect effects from model P . Parameters \tilde{h}_i and \tilde{h}_j are to be adjusted to ensure correct marginals. The path probability between positions i_1 and i_{L+1} through the length- L path $[i_1, i_2, \dots, i_{L+1}]$ is a direct generalization:

$$P_{[i_1 \dots i_{L+1}]}^{path}(A_{i_1}, A_{i_{L+1}}) = \sum_{\{A_{i_2} \dots A_{i_L}\}} f_{i_1}(A_{i_1}) \prod_{l=1}^L P_{i_{l+1}i_l}^{dir}(A_{i_{l+1}} | A_{i_l}), \quad (6)$$

with $P_{ij}^{dir}(A_i | A_j) = P_{ij}^{dir}(A_i, A_j) / f_j(A_j)$. Equation (6) contains the product of direct probabilities for all links in the path, in analogy to a Markov chain. The sum over all configurations taken by intermediate sites $[i_2 \dots i_L]$ is performed efficiently by dynamic programming; the definition guarantees the empirical marginals in all sites on the path.

To measure the correlation mediated by direct links or indirect paths, we use variants of the mutual information based on the direct and path probabilities. To this aim, we define the direct information (DI) as

$$DI_{ij} = \sum_{A_i, A_j=1}^q P_{ij}^{dir}(A_i, A_j) \log \frac{P_{ij}^{dir}(A_i, A_j)}{f_i(A_i)f_j(A_j)}, \quad (7)$$

and the *path information* (PI) as

$$PI_{[i...j]} = \sum_{A_i, A_j=1}^q P_{[i...j]}^{path}(A_i, A_j) \log \frac{P_{[i...j]}^{path}(A_i, A_j)}{f_i(A_i)f_j(A_j)}. \quad (8)$$

The full *mutual information* (MI) is defined by replacing P_{ij}^{dir} or $P_{[i...j]}^{path}$ by f_{ij} .

The Joint Effect of Paths of Length 2

Quantifying the strength of a group of indirect effects between two sites i and j is in general non trivial. However, it is possible if one only considers all chains of couplings that go through at most one intermediary site k . In other words, one can combine the direct path $[ij]$ and all the chains of the form $[ikj]$ ($k \neq i, j$) into a single probability distribution:

$$P_{ij}^{L2}(A_i, A_j) \propto \frac{P_{ij}^{dir}(A_i, A_j)}{z_i(A_i)z_j(A_j)} \cdot \prod_{k \neq i, j} P_{[ikj]}^{path}(A_i, A_j), \quad (9)$$

where z_i and z_j ensure P_{ij}^{L2} to have marginals f_i and f_j . The path probabilities can be simply multiplied since each intermediate residue k appears only once, and they become conditionally independent for given (A_i, A_j) . The correlation resulting from this combination of paths is the mutual information of P_{ij}^{L2} , called $L2I$.

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online. Code and raw data can be accessed via Github (<https://github.com/matteofigliuzzi/bmDCA>; last accessed January 2018).

Acknowledgments

M.W. acknowledges funding by the ANR project COEVSTAT (ANR-13-BS04-0012-01), and by the European Union's H2020 research and innovation programme MSCA-RISE-2016 under Grant Agreement No. 734439 INFERNET. This study was undertaken partially in the framework of CalSimLab, supported by the grant ANR-11-LABX-0037-01 as part of the "Investissements d'Avenir" program (ANR-11-IDEX-0004-02).

References

Ackley DH, Hinton GE, Sejnowski TJ. 1985. A learning algorithm for Boltzmann machines. *Cogn Sci*. 9(1): 147–169.

Arnold K, Bordoli L, Kopp J, Schwede T. 2006. The swiss-model workspace: a web-based environment for protein structure homology modelling. *Bioinformatics* 22(2): 195–201.

Balakrishnan S, Kamisetty H, Carbonell JG, Lee S-I, Langmead CJ. 2011. Learning generative models for protein fold families. *Proteins* 79(4): 1061–1078.

Barton JP, De Leonardi E, Coucke A, Cocco S. 2016. Ace: adaptive cluster expansion for maximum entropy graphical model inference. *Bioinformatics* 32(20): 3089–3097.

Cocco S, Feinauer C, Figliuzzi M, Monasson R, Weigt M. 2017. Inverse statistical physics of protein sequences: a key issues review. *Rep Prog Phys*. 81:3.

de Juan D, Pazos F, Valencia A. 2013. Emerging methods in protein co-evolution. *Nat Rev Genet*. 14(4): 249–261.

Dijkstra EW. 1959. A note on two problems in connexion with graphs. *Numer Math*. 1(1): 269–271.

Durbin R, Eddy SR, Krogh A, Mitchison G. 1998. Biological sequence analysis: probabilistic models of proteins and nucleic acids. Cambridge: Cambridge University Press.

Eddy SR. 1998. Profile Hidden-Markov models. *Bioinformatics* 14(9): 755–763.

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(1): 012707.

Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J, et al. 2014. Pfam: the protein families database. *Nucleic Acids Res*. 42(D1): D222–D230.

Haldane A, Flynn WF, He P, Vijayan R, Levy RM. 2016. Structural propensities of kinase family proteins from a potts model of residue co-variation. *Protein Sci*. 25(8): 1378–1384.

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:e03430.

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.

Jones DT, Singh T, Kosciolk 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.

Levy RM, Haldane A, Flynn WF. 2017. Potts Hamiltonian models of protein co-variation, free energy landscapes, and evolutionary fitness. *Curr. Opin. Struct. Biol*. 43:55–62.

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

Merchan L, Nemenman I. 2016. On the sufficiency of pairwise interactions in maximum entropy models of networks. *J Stat Phys*. 162(5): 1294–1308.

Mora T, Bialek W. 2011. Are biological systems poised at criticality?. *J Stat Phys*. 144(2): 268–302.

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): E1293–E1301.

Owiniowski J, Plotkin JB. 2014. Inferring fitness landscapes by regression produces biased estimates of epistasis. *Proc Natl Acad Sci USA*. 111(22): E2301–E2309.

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

Ovchinnikov S, Park H, Varghese N, Huang P-S, Pavlopoulos GA, Kim DE, Kamisetty H, Kyrpidis NC, Baker D. 2017. Protein structure determination using metagenome sequence data. *Science* 355(6322): 294–298.

Reynolds KA, Russ WP, Socolich M, Ranganathan R. 2013. Evolution-based design of proteins. *Methods Enzymol*. 523:213–235.

Russ WP, Lowery DM, Mishra P, Yaffe MB, Ranganathan R. 2005. Natural-like function in artificial ww domains. *Nature* 437(7058): 579–583.

Schmidt M, Hamacher K. 2017. Three-body interactions improve contact prediction within direct-coupling analysis. *Phys Rev E* 96(5): 052405.

Schug A, Weigt M, Onuchic JN, Hwa T, Szurmant H. 2009. High-resolution protein complexes from integrating genomic information with molecular simulation. *Proc Natl Acad Sci USA*. 106(52): 22124–22129.

Socolich M, Lockless SW, Russ WP, Lee H, Gardner KH, Ranganathan R. 2005. Evolutionary information for specifying a protein fold. *Nature* 437(7058): 512–518.

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(44): 13567–13572.

- Szurmant H, Weigt M. 2018. Inter-residue, inter-protein and inter-family coevolution: bridging the scales. *Curr Opin Struct Biol.* 50:26–32.
- 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(1): e1005324.
- Webb B, Sali A. 2014. Protein structure modeling with MODELLER. *Methods Mol Biol.* 1137:1–15.
- 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.