PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Synthetic protein alignments by CCMgen quantify noise in residue-residue contact prediction

Susann Vorberg, Stefan Seemayer¤, Johannes Söding*

Quantitative and Computational Biology Group, Max-Planck Institute for Biophysical Chemistry, Göttingen, Germany

¤ Current address: BASF SE, Ludwigshafen – Chemical Park, Germany
* soeding@mpibpc.mpg.de

OPEN ACCESS

Citation: Vorberg S, Seemayer S, Söding J (2018) Synthetic protein alignments by CCMgen quantify noise in residue-residue contact prediction. PLoS Comput Biol 14(11): e1006526. https://doi.org/10.1371/journal.pcbi.1006526

Editor: Björn Wallner, Linköping University, SWEDEN

Received: June 27, 2018

Accepted: September 24, 2018

Published: November 5, 2018

Data Availability Statement: CCMpredPy and CCMgen are available under GNU AGPL3 at https://github.com/soedinglab/ccmgen. Code reproducing the analysis results can be found at https://github.com/soedinglab/ccmgen-scripts.

Funding: SS is supported by the Deutsche Forschungsgemeinschaft (grant GRK 1721). SV is supported by a fellowship of the Deutsche Forschungsgemeinschaft through the Graduate School of Quantitative Biosciences Munich (QBM), http://qbm.genzentrum.lmu.de/. The funders had no role in study design, data collection and

Abstract

Compensatory mutations between protein residues in physical contact can manifest themselves as statistical couplings between the corresponding columns in a multiple sequence alignment (MSA) of the protein family. Conversely, large coupling coefficients predict residue contacts. Methods for de-novo protein structure prediction based on this approach are becoming increasingly reliable. Their main limitation is the strong systematic and statistical noise in the estimation of coupling coefficients, which has so far limited their application to very large protein families. While most research has focused on improving predictions by adding external information, little progress has been made to improve the statistical procedure at the core, because our lack of understanding of the sources of noise poses a major obstacle. First, we show theoretically that the expectation value of the coupling score assuming no coupling is proportional to the product of the square roots of the column entropies, and we propose a simple entropy bias correction (EntC) that subtracts out this expectation value. Second, we show that the average product correction (APC) includes the correction of the entropy bias, partly explaining its success. Third, we have developed CCMgen, the first method for simulating protein evolution and generating realistic synthetic MSAs with pairwise statistical residue couplings. Fourth, to learn exact statistical models that reliably reproduce observed alignment statistics, we developed CCMpredPy, an implementation of the persistent contrastive divergence (PCD) method for exact inference. Fifth, we demonstrate how CCMgen and CCMpredPy can facilitate the development of contact prediction methods by analysing the systematic noise contributions from phylogeny and entropy. Using the entropy bias correction, we can disentangle both sources of noise and find that entropy contributes roughly twice as much noise as phylogeny.

Author summary

Knowledge about the three-dimensional structure of proteins is key to understanding their function and role in biological processes and diseases. The experimental structure determination techniques, such as X-ray crystallography or electron cryo-microscopy, are

analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

labour intensive, time-consuming and expensive. Therefore, complementary computational methods to predict a protein's structure have become indispensable. Over the last years, immense progress has been made in predicting protein structures from their amino acid sequence by utilizing highly accurate predictions of spatial contacts between amino acid residues as constraints in folding simulations. However, contact prediction methods require large numbers of homologous protein sequences in order to discriminate between signal and noise. A major obstacle preventing progress on the statistical methodology is our limited understanding of the different components of noise that are known to affect the predictions. We provide two tools, CCMpredPy and CCMgen, that can be used to learn highly accurate statistical models for contact prediction and to simulate protein evolution according to the statistical constraints between positions of residues as specified by these models, respectively. We showcase their usefulness by quantifying the relative contribution of noise arising from entropy and phylogeny on the predicted contacts, which will facilitate the improvement of the statistical methodology.

Introduction

In the course of evolution, proteins are under selective pressure to maintain their function and correspondingly their structure. A possible mechanism to maintain structural integrity is the compensation of deleterious mutations between residue pairs in physical contact, known as compensatory mutations: Upon the mutation of one residue the contacting residue has an increased probability to mutate into a residue that will locally restabilize the protein structure, for instance by regaining a lost interaction between them. In multiple sequence alignments (MSAs) of related proteins, this effect leads to correlations between columns of residues in contact among most protein family members [1–4]. Many of these correlations are indirect, though, and arise through transitive chains of contacting residue pairs [5–8].

By applying statistical techniques that can distinguish mere correlation from direct statistical coupling of residue positions [5, 7, 9], many false positive predictions could be eliminated. The adoption of this class of statistical models, known as Markov random fields (MRFs), or Potts models in statistical physics, led to a breakthrough in de-novo (template-free) protein structure prediction: The predicted contacts proved sufficiently accurate to be used as spatial restraints to reliably predict protein 3D structures purely from sequence information [10–20].

The requirement for large MSAs for sufficiently precise predictions has limited the applicability of contact-assisted de-novo protein structure prediction, all the more because large protein families are more likely to contain at least one member whose structure has been solved and which can be used as a template for homology modelling. Therefore, most research has focused on making contact prediction reliable enough for medium-sized protein families [20–25].

The background noise effects arising in residue-residue contact prediction have been postulated to arise from three sources [5, 26–32]: random sampling noise due to the limited number of sequences, phylogenetic noise due to the evolutionary relatedness of sequences in the MSA, and entropic noise or rather bias, which biases high-entropy columns towards higher scores. Unfortunately, the relative contribution and properties of the three different sources of noise are difficult to study in real alignments, mainly because the true values of coupling parameters are not known. In addition, the stochastic noise, entropy-dependent noise and phylogenetic noise cannot be modified independently (for example by subsampling), as these noise sources

are indirect, complex consequences of learning on only a limited number of sequences that are statistically dependent on each other according to their phylogenetic relationship.

Many correction schemes for removing noise from the matrix of predicted contact scores have been examined [27, 29, 30, 33–36], and the *average product correction* (*APC*) [26] came out as a clear winner and is used in almost all recent studies. However, it is widely acknowledged in the field that our limited understanding of what noise effects APC is correcting and why it is so effectively correcting them is severely impeding progress in developing better statistical methods to predict contacting residue pairs. We repeatedly made the experience that a promising extension to the standard MRF model that considerably improved the contact prediction performance *before* applying the APC was doomed to failure because it inexplicably yielded worse results than the baseline method *after* applying APC.

Based on theoretical considerations (Material and methods), we propose a simple entropy bias correction (EntC) that is computed solely from per-column entropies of the input MSA and corrects for entropy-dependent bias without affecting noise from phylogenetic effects. We find that the EntC eliminates nearly as much noise as the APC. The observation that both corrections can be expressed as a product of two factors depending only on each column separately explains partly the success of APC and suggests that it mainly corrects for entropy noise. Whereas the APC is applied as a post-correction to the matrix of predicted contact scores, the EntC can be applied directly on the statistical couplings of the MRF model, prior to computing a contact score and other post-processing treatments.

To systematically study the sources of noise limiting the accuracy of contact predictions from MSAs and to facilitate progress in the development of better contact prediction methods, we have developed CCMgen, a method for generating realistic synthetic protein sequence alignments whose residues obey the selection pressures described by a MRF with pairwise statistical couplings between residue positions.

For that purpose, CCMgen requires an exact statistical model that will reliably reproduce the empirical alignment statistics, such as single-site, pairwise or even higher-order amino acid frequencies, of the input MSA that was used to learn the MRF model in the first place. A typical strategy to obtain estimates of the MRF model parameters would involve maximizing the logarithm of the likelihood function over all sequences in the MSA. However, the normalization factor in the likelihood function requires to sum 20^L terms, where L is the protein length, and methods to optimize the full likelihood are very slow for realistic proteins [5, 37–41]. The most popular approximation is to maximize the pseudo-likelihood instead of the likelihood, as it can be shown that it converges to the same solution for large numbers of samples and it is fast to compute [42–44]. Even though pseudo-likelihood maximization gives results of the same quality of predicted residue-residue contacts as those using the full likelihood optimization, several studies unveiled that the pseudo-likelihood model is inaccurate and not able to accurately reproduce the empirical alignment statistics [37, 45].

We provide an implementation of an alternative precise inference technique, persistent contrastive divergence (PCD) [46] with our tool CCMpredPy. Compared to pseudo-likelihood maximization, PCD achieves identical precision for contact prediction while the inferred MRF model reproduces empirical marginals much more precisely. The increased quality of the models comes at the expense of longer run times, which are however still practical for even large proteins and alignments using a single desktop computer. High quality MRF models learned with PCD might prove beneficial beyond the purpose of contact prediction when problems require exact model statistics, e.g. when studying mutational effects or designing new protein features using the model energies.

Finally, we employ CCMgen in combination with MRF models that have been learned with the PCD algorithm and our entropy bias correction to quantify the relative effect sizes of

phylogenetic and entropic bias on the precision of contact prediction. We find that the contribution of entropy noise in contact prediction is on average twice as big as that of phylogenetic noise.

Results

Persistent contrastive divergence allows accurate inference of MRFs

An exactly inferred MRF will reliably reproduce the empirical single-site and pairwise amino acid frequencies, $f_i(a)$ and $f_{ij}(a, b)$ for all positions i, j in the MSA and all amino acids $a, b \in \{1, \ldots, 20\}$ [7, 47]. Several studies demonstrated that pseudo-likelihood maximization, while being the method of choice for contact prediction, yields models that cannot accurately reproduce the empirical alignment statistics [37, 39, 45].

We developed a method that uses an inference technique called persistent contrastive divergence(PCD) [46] to learn MRF models that accurately reproduce the empirical alignment statistics. As in the study by Figliuzzi *et al.* [37], we computed for all Pfam MSAs in the PSICOV dataset the single-column and paired-column amino acid frequencies as well as covariances, $\mathrm{cov}(\delta_{a,x_i}, \delta_{b,x_j}) = f_{ij}(a, b) - f_i(a)f_j(b)$, where $\delta_{a,x}$ is the Kronecker symbol. We compared these statistics with those from sequences obtained by Markov chain Monte Carlo (MCMC) sampling from MRFs that were trained on the Pfam MSAs using either pseudo-likelihood maximization or PCD.

We find that the empirical single-site amino acid frequencies are well reproduced by both models. But whereas the empirical pairwise amino acid frequencies and covariances correlate strongly with the corresponding statistics computed from the PCD samples, the correlation is much weaker for samples obtained from pseudo-likelihood MRF models (Fig 1A and 1B and S1 Fig).

Furthermore, as in Figliuzzi *et al.* [37], we investigated how well the generated MCMC samples reproduce the alignment substructure of the original Pfam alignments with respect to the organisation of subfamilies in sequence space. We projected the protein sequences of the MCMC samples onto the first two principle components obtained from a principal component analysis (PCA) of the original Pfam MSA (for details see S2 Text). Again, we find that the alignment substructure described by the grouping of sequences that can be observed in the two-dimensional PCA space, is reproduced more reliably by MCMC samples generated from PCD models than from pseudo-likelihood models (S2b and S2d Fig).

It has been argued that for the purpose of predicting residue contacts an approximate model such as those obtained by maximizing the pseudo-likelihood for a MRF is sufficiently accurate to infer the correct topology of the interaction network of residues [45]. Fig 1C shows the mean precision of the predicted contacts from a PCD model and a pseudo-likelihood model versus the number of predictions per columns in the MSA. The precision for one MSA is the fraction of correctly predicted contacting pairs of positions (i, j) out of all predicted pairs. The correctly predicted pairs (i, j) are those for which the $C_\beta - C_\beta$ distance in the reference protein structure of the Pfam MSA is below 8Å. Residue pairs that are separated by less than six positions along the protein sequence are not considered for the evaluation as they typically correspond to contacts within secondary structure elements and reflect local geometrical constraints. Indeed, predicted contacts from a PCD model achieve equal precision as predictions from a pseudo-likelihood model. S3 Fig shows further analysis, comparing the APC-corrected contact scores from pseudo-likelihood and PCD models.

However, more complex problems such as prediction of mutational effects or generating realistic samples of sequences, require exact model statistics. Several methods have been developed that exactly infer MRF models, such as bmDCA and ACE [5, 37–41], but they are

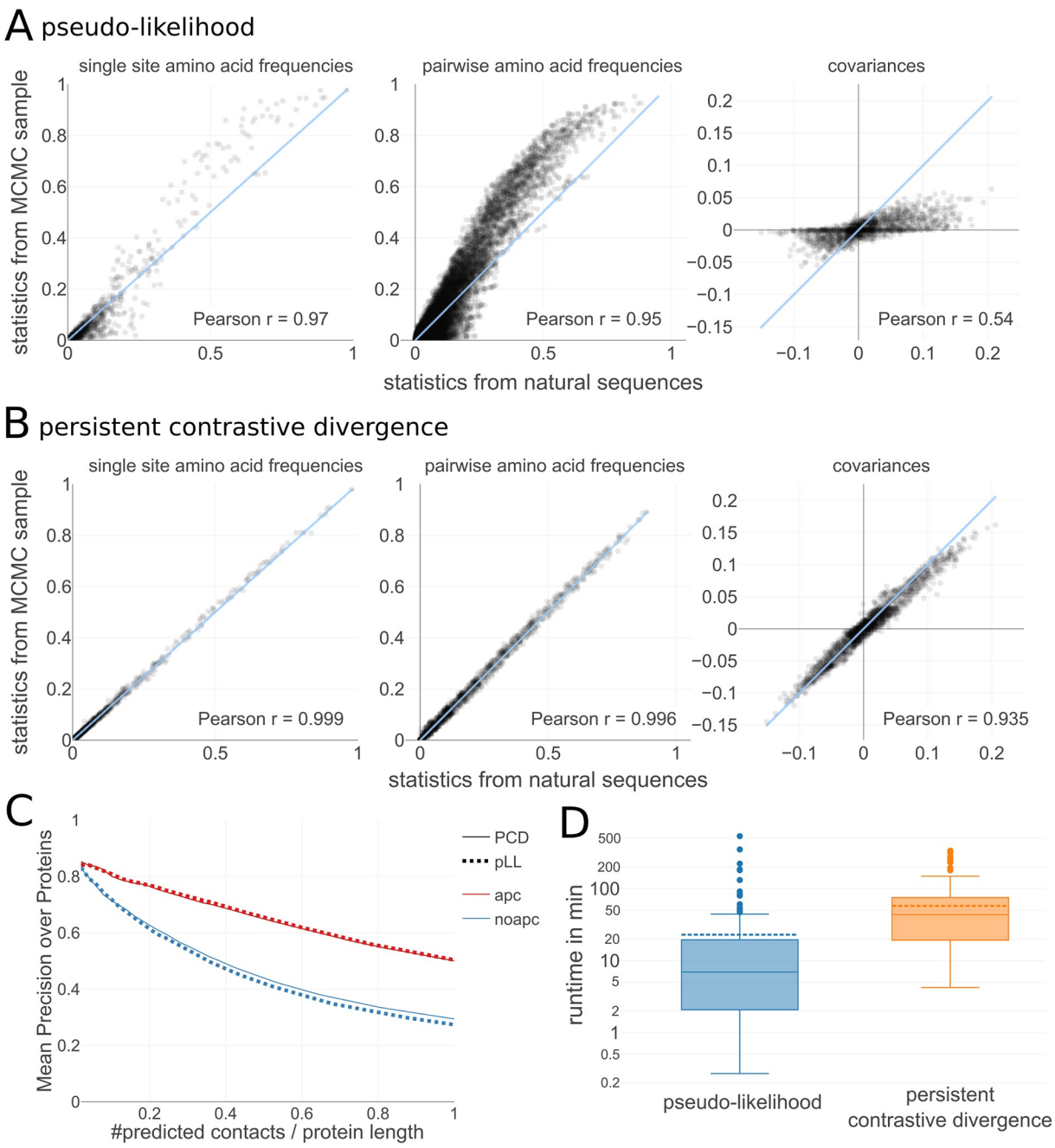

Fig 1. Persistent contrastive divergence permits the inference of high-quality models. A and **B** Comparing single amino acid frequencies (left), pairwise amino acid frequencies (center) and covariances (right) computed from natural sequences (Pfam alignment) and from Markov chain Monte Carlo (MCMC) samples generated from Markov random field (MRF) models trained with either pseudo-likelihood maximization (**A**) or persistent contrastive divergence (PCD) (**B**) for protein 1bkrA in the PSICOV dataset. **C** Mean precision of contacts, predicted as (APC corrected) L_2 norm of pair couplings of a MRF trained with either pseudo-likelihood maximization or PCD. **D** Distribution of run times in minutes when learning MRF models with CCMpredPy (run on 4 cores). The median runtime in minutes with pseudo-likelihood is 7 minutes and with PCD is 43.5 minutes. Dashed line in boxplots represents the mean, solid line represents the median of the distribution. **C** and **D** are computed over the 150 proteins in the PSICOV dataset.

https://doi.org/10.1371/journal.pcbi.1006526.g001

computationally intensive which renders them impractical for real proteins. In comparison, our PCD-based CCMpredPy method is only about a magnitude slower than pseudo-likelihood maximization (Fig 1D).

Correcting for entropy bias removes a major source of noise

A major obstacle for improving the statistical methods for residue-residue contact prediction is our lack of understanding of the sources of noise. The background noise effects have been postulated to arise from at least three sources, whose size and properties are difficult to quantify: phylogenetic, entropic and sampling noise.

Phylogenetic noise originates from the violation of the assumption of independence of sequences in the MSA [48]. This assumption has been made by all methods that have been employed for contact prediction so far. To understand the origin of phylogenetic noise, consider the example in Fig 2. The MSA is composed of two subtrees whose last common ancestor sequences, DSMF and ETMF, had a mutation at the second and first position respectively. All descendants of the first ancestral sequence whose first two residues have not mutated in the meantime will have a DS at first and second position, while all descendants of the other ancestral sequence whose first two residues have not mutated yet will have a ET at those positions. Therefore, pairs DS and ET are more likely than would be expected from the frequencies of D and E in the first column and of S and T in the second column. The first and second position will therefore appear to be statistically coupled even though they are not.

Entropic bias describes the tight correlation of the expectation value of the contact score c_{ij} between columns i and j of a MSA under the assumption of no coupling between both columns with the product of the square roots of column entropies $s_i = -\sum_{a=1}^{20} f_{ia} \log f_{ia}$:

$$\mathrm{E}[c_{ij}] \underset{\sim}{\propto} s_i^{\frac{1}{2}} s_j^{\frac{1}{2}} . \tag{1}$$

Put simply, higher column entropies lead to higher expected contact scores c_{ij} even if no coupling exists. To understand the origin of this bias, we need a bit of notation. From the MSA we compute coefficients $w_{ij}(a, b)$ that quantify the statistical coupling between residue $a \in \{1, \ldots, 20\}$ occurring in column i and residue $b \in \{1, \ldots, 20\}$ in column j of the same sequence (Materials and methods). A coefficient $w_{ij}(a, b) = 0.1$ signifies that residue a in column i and residue b in column j in the same sequence is exp(0.1) times more likely to occur than what would be expected if the amino acids in both columns were independent of each other.

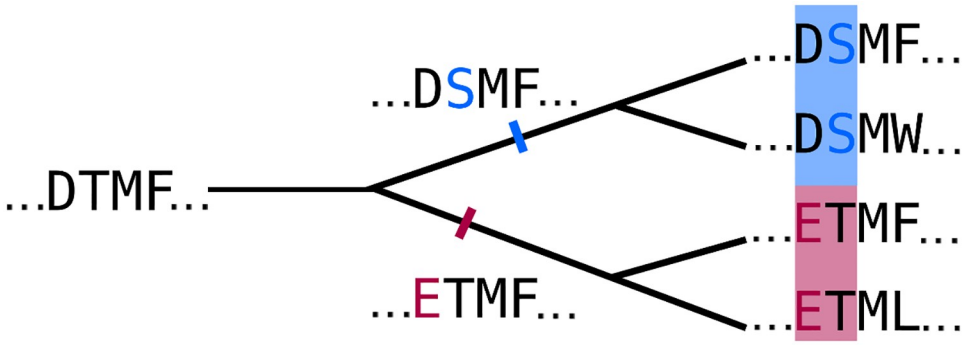

Fig 2. The phylogenetic dependence of closely related sequences can produce covariation signals. Here, two independent mutation events (highlighted in red and blue) in two branches of the tree result in a covariation signal for the two affected positions.

https://doi.org/10.1371/journal.pcbi.1006526.g002

To predict contacts, we estimate the coupling coefficients $w_{ij}(a, b)$, for example by maximizing the pseudo-likelihood, and obtain estimates $\tilde{w}_{ij}(a, b)$, from which we can calculate a score to predict contacts. The commonly used contact score between columns i and j of a MSA is simply the norm of the 400-dimensional vector $\tilde{\mathbf{w}}_{ij}$,

$$c_{ij} := \|\tilde{\mathbf{w}}_{ij}\| = \left(\sum_{a,b=1}^{20} \tilde{w}_{ij}(a, b)^2 \right)^{1/2}. \quad (2)$$

It sums up the squared coupling coefficients over all possibly coupled amino acid pairs.

Let us assume that a MSA has no statistically coupled residue pairs, meaning that the true coupling coefficients are all zero. But the estimation of the coefficients results in errors, which contribute a systematic bias, as we will now see. The regularization of the MRF will ensure that the coupling coefficients $w_{ij}(a, b)$ for those amino acid pairs (a, b) without counts will be zero and will therefore not contribute to the overall contact score c_{ij} for this residue pair. For those pairs (a, b) with one or more counts, the $w_{ij}(a, b)$ will be distributed around zero but will rarely be exactly zero, just as $f_{ij}(a, b)$ is rarely exactly equal to $f_i(a) \times f_j(b)$. So each amino acid pair (a, b) that occurs at least in one sequence will make a contribution $\mathrm{E}[\tilde{w}_{ij}(a, b)^2]$ to the sum in Eq 2. These contributions to c_{ij} stemming from noisy estimates $w_{ij}(a, b)$ create a bias that will increase with the number of pairs (a, b) of bins over which the N counts are distributed. Columns with high entropy tend to disperse the counts of amino acid pairs over more bins (a, b) than columns with low entropy. It is shown in Materials and Methods that the expectation value of this bias on c_{ij} can be approximated by a term proportional to product of the square roots of the entropies of the two columns.

The factorization of the EntC into two factors depending only on each column separately explains partly the success of APC and suggests that it mainly corrects for entropy noise (Materials and methods).

Sampling noise on the estimated coupling coefficients would remain, even if we correct for entropic bias and phylogenetic effects, because with a finite sample of sequences we cannot estimate fractions arbitrarily accurately. For example even if the sequences could be assumed to be independent of each other, the probability of an amino acid pair (a, b) that has been observed $n \ll N$ times out of N is only estimated to a relative accuracy of approximately $\sigma/\mu = \sqrt{n(1 - n/N)}/n \approx 1/\sqrt{n}$, according to the standard deviation of the binomial distribution. More precisely, whereas the entropy bias describes the systematic offset of the contact score c_{ij} stemming from the non-zero expectation values $\mathrm{E}[\tilde{w}_{ij}(a, b)^2]$, the sampling noise originates from the *variance* of the coefficients, $\mathrm{var}[\tilde{w}_{ij}(a, b)^2]$, which is due to the finite number of measurements (sequences) N taken.

The APC and the EntC in action. Fig 3A shows the contact scores c_{ij} (see Eq 2) in grey scale computed from a MRF that has been trained from a typical example MSA. The striping patterns in horizontal and vertical directions reflect strong systematic row- and column-dependent score biases. Some positions seem to obtain generally higher scores than others. Without correction, ranking by these scores would severely overpredict contacts between positions with positive score bias and underpredict contacts between positions with negative bias.

Applying the APC eliminates the systematic effects leading to the striping patterns (Fig 3B). It thereby greatly improves the performance of all contact prediction scores for local pairwise measures such as mutual information as well as for global statistical coupling methods such as the MRF-based contact score referred to here [8, 26, 49–51].

To disentangle the entropic bias from the phylogenetic noise, we first propose an entropy-dependent correction, EntC, of the contact scores c_{ij} that depends solely on the per-column

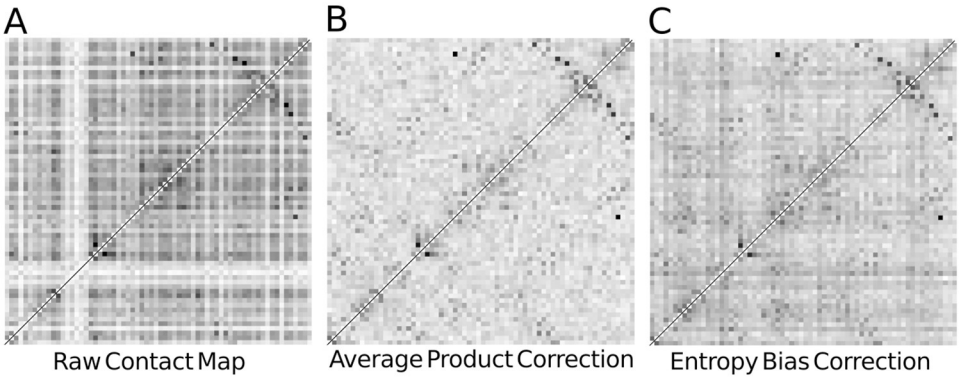

Fig 3. Entropy correction eliminates major source of noise. Raw and corrected contact score matrices for protein 1c9oA. The gray scale indicates the contact scores for each residue pair (i, j) for raw uncorrected scores computed from a Markov random field (MRF) model trained with persistent contrastive divergence (PCD) (**A**, Eq 2), average product corrected scores (**B**, Eq 17) and entropy corrected scores (**C**, Eq 18). The striping pattern in **A** arises from systematic score biases, which originate mainly from entropy.

https://doi.org/10.1371/journal.pcbi.1006526.g003

entropies s_i of the MSA from which the MRF was trained,

$$c_{ij}^{\mathrm{EntC}} = c_{ij} - \alpha s_i^{\frac{1}{2}} s_j^{\frac{1}{2}} \tag{3}$$

with an analytically determined constant α that depends on all column entropies $s_1, \ldots, s_L$. For a motivation of this score, see Materials and methods. Fig 3C shows how the entropy correction removes almost as much of the striping effects as APC.

Fig 4A reveals how strongly these two corrections correlate (see S4 Fig for the whole data set). Moreover, the correlation is high (Pearson correlation $\rho = 0.972$) also between the column-specific factors in the APC and EntC, that is, $c_{i\cdot}/\sqrt{c_{\cdot\cdot}}$ and $\sqrt{\alpha s_i}$ (compare Eqs 17 and 18). To analyze whether the correlation of APC and EntC is influenced by the phylogenetic dependence between the sequences, we generated synthetic MSAs with CCMgen using a binary and a star tree topology and trained an MRF on these two alignments (for details see next section). The correlation between the column-column APC and EntC correction terms remains as high

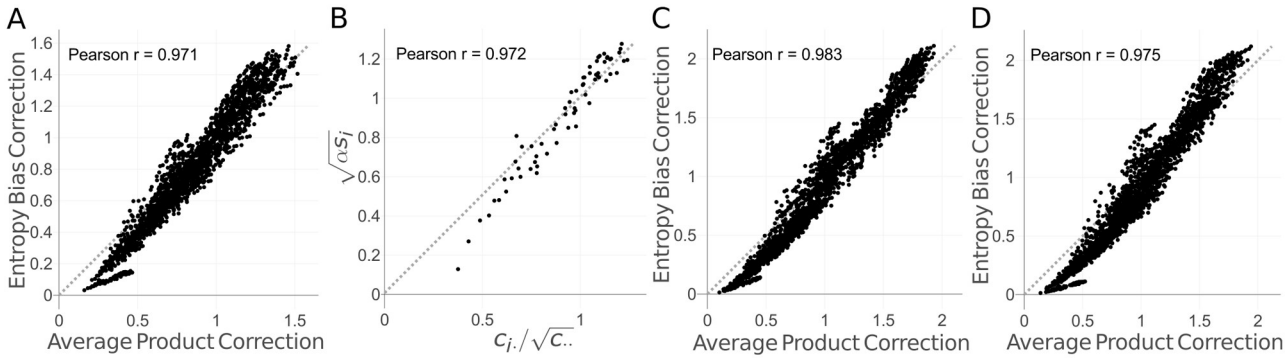

Fig 4. The average product correction (APC) and entropy bias correction (EntC) term correlate strongly. A For protein 1c9oA in the PSICOV data set the correction defined by the APC correlates well with the correction determined by our entropy bias correction strategy when learning a Markov random field (MRF) model with persistent contrastive divergence (PCD). **B** Also the factors appearing in the APC and EntC corrections, $c_{i\cdot}/\sqrt{c_{\cdot\cdot}}$ and $\sqrt{\alpha s_i}$., correlate well. **C,D** The Pearson correlation is similarly large when the MRF model is learnt from synthetic alignments generated with CCMgen using binary tree topologies (**C**) or using star tree topologies (**D**), both in the example of protein 1c9oA shown here and also across all 150 proteins in the Pfam dataset (S4 Fig).

https://doi.org/10.1371/journal.pcbi.1006526.g004

as for the real Pfam MSA (Pearson correlations 0.983 (binary tree) and 0.975 (star tree) versus 0.971 (Pfam MSA)) (see Fig 4C and 4D). These results suggest that the APC predominantly corrects out entropy bias [26] rather than phylogenetic bias. Our theoretical analysis in Materials and Methods further supports this (see also the Discussion).

However, the relative contributions of entropic and phylogenetic noise limiting the precision of contact prediction are yet unclear. In the following we will use our tool CCMgen to distinguish between both sources of noise.

Quantifying noise effects with CCMgen reveals entropy as dominating source of noise

Our workflow to analyse the relative contributions of noise sources is described in Fig 5. First, we estimate the parameters of a second order MRF model with PCD using CCMpredPy for each of the 150 Pfam MSAs in the PSICOV data set. To obtain models with few but precise constraints, we set coupling parameters to zero for non-contacting residue pairs (C_β distance >12Å) during parameter learning.

In a second step, we use CCMgen with the learned model parameters to generate realistic synthetic MSAs of interdependent sequences with pairwise statistical couplings between some positions as they are observed in MSAs between residues in physical contact. CCMgen provides full control over the generation of the synthetic MSAs by allowing us to specify the evolutionary times and phylogeny along which the sequences are sampled. We sample two sets of synthetic MSAs: one set with a star tree topology and the other with a binary tree topology (Fig 6). Given sufficient evolutionary time, the phylogenetic dependencies between sequences drawn according to the star tree topology should be negligible, whereas sequences drawn along the binary tree are expected to contain stronger interdependencies.

Because the accuracy of predictions strongly depends on alignment depth and diversity [49, 52], we ensured that the synthetic alignments contain the same number of sequences and have similar diversities as the original Pfam alignments (for details see Material and methods). These provisions justify a direct comparison of the results for sampling sequences along the star and binary topologies.

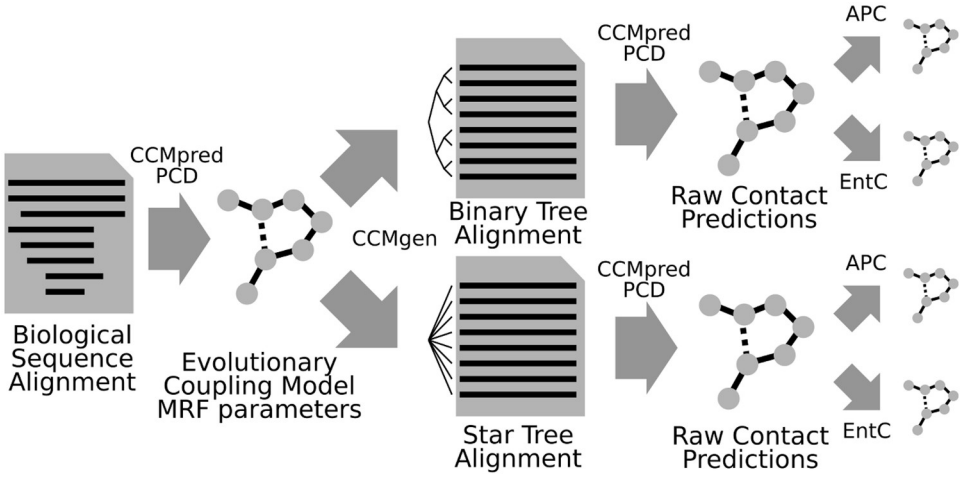

Fig 5. Workflow for quantifying noise effects. Artificial multiple sequence alignments (MSAs) are generated with CCMgen using a binary tree phylogeny for sequences with strong interdependencies and using a star tree phylogeny for nearly independently sampled sequences. Then, contacts are predicted and post-corrected using the average product correction (APC) and entropy bias correction (EntC)for both sets of alignments.

https://doi.org/10.1371/journal.pcbi.1006526.g005

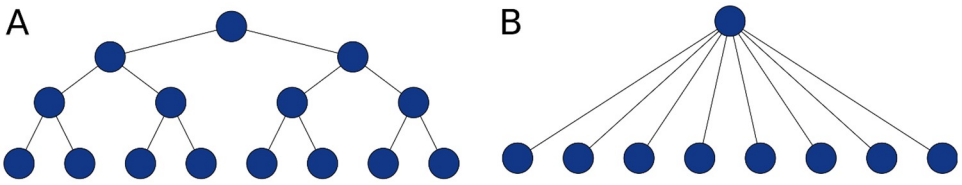

Fig 6. Idealized phylogenetic tree topologies available with CCMgen. CCMgen can generate multiple sequence alignments (MSAs) based on a Markov random field (MRF) model and a phylogenetic tree supplied either as Newick file or as one of the two shown, idealized topologies: **A** binary tree and **B** star-shaped tree.

https://doi.org/10.1371/journal.pcbi.1006526.g006

Third, we run CCMpredPy on each of the synthetic MSAs and predict residue-residue contacts by ranking the pairs according to the descending raw contact scores (Eq 2), or by the APC-corrected contact scores (Eq 17) or by entropy corrected scores (Eq 18). Since we know the ground truth of which pairs are coupled from the MRF model used for generating the synthetic MSAs, we can use these alignments to investigate and quantify the effect of phylogenetic noise on the precision of residue-residue contact prediction.

Fig 7A and 7B plot the mean precision of the predicted contacts from both types of synthetic MSAs versus the number of predictions per columns in the MSA. As expected, the mean precision drops as more predictions are considered and lower ranks are included.

Both APC and EntC correction have a huge effect in reducing noise and increasing the precision of predictions. Both corrections give very similar results for MSAs generated with star topology trees, which are not expected to show phylogenetic noise, while the APC performs slightly better than the EntC on MSAs with binary tree topologies. This suggests that the APC corrects out a small part of the phylogenetic noise. This would be plausible because this noise source affects some positions more than others (Fig 2) and would thereby also cause striping, which could be corrected by APC.

We estimate the strength of the phylogenetic noise as the drop in precision between the EntC-corrected precisions on MSAs with star topology and EntC-corrected precision on the

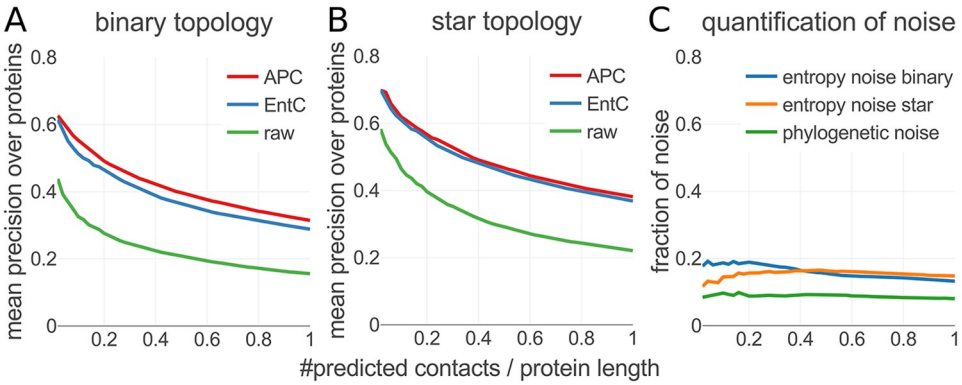

Fig 7. Effect of phylogenetic noise on contact prediction accuracy. MSAs generated with phylogenetic trees with star topologies do not contain phylogenetic noise. Therefore, the entropic bias is fully responsible for the difference in mean precision of predicted contacts between the uncorrected raw coupling scores (green) in (**B**) and the APC- or EntC-corrected scores. On MSAs generated with binary tree topologies, the precision of EntC-corrected scores drops due to phylogenetic noise (**A**). APC seems to correct out a small fraction of this noise, whereas EntC can only correct for entropy-related bias. (**C**) We can estimate the effect of phylogenetic noise as the drop in precision of the entropy-corrected scores (blue) from binary to star tree topology.

https://doi.org/10.1371/journal.pcbi.1006526.g007

MSAs with binary tree topology (Fig 7C). The strength of the entropy noise is shown in terms of the drop in precision between the EntC-corrected and uncorrected, raw contact scores, both for the star tree topology and for the binary tree topology. The contribution of entropy noise to the drop in precision is roughly two times larger than that of the phylogenetic noise.

Discussion

The average product correction explained

The success of the average product correction (APC) (Eq 17) is in part explained by three key insights: First, as we have seen, the entropy bias explains a large part of the noise in residue contact prediction. Second, as shown in Material and Methods, the EntC factorizes over columns, that is, it can be written as a product of two factors, each of which depends only on one column. Third, as we show now, the APC boils down to subtracting from the score c_{ij} an approximation to its expectation value under the null model of no couplings, if this expectation value factorizes into two terms, each of which depend only on one column,

$$\mathrm{E}[c_{ij}] \approx u_i\, u_j\,. \tag{4}$$

Taken together, these three insights explain why the APC includes the EntC, which corrects for most of the bias.

To demonstrate the third insight, we approximate

$$c_{i\bullet} = \sum_{j=1}^{L} c_{ij} \approx \sum_{j=1}^{L} E[c_{ij}] \approx \sum_{j=1}^{L} u_i u_j = u_i \langle u_\bullet \rangle \tag{5}$$

$$c_{i\bullet\bullet} = \sum_{i,j=1}^{L} c_{ij} \approx \sum_{i,j=1}^{L} E[c_{ij}] \approx \sum_{i,j=1}^{L} u_i u_j = \langle u_\bullet \rangle^2, \tag{6}$$

because the sum over L terms averages out the fluctuations around the expectation value of each term. This approximation is probably the reason why the APC works better on c_{ij} than on c_{ij}^2 because for c_{ij}^2 the values in the sum are much more dispersed and dominated by one or a few terms, which renders the above approximation much less accurate. The APC correction is then

$$\frac{c_{i\bullet} c_{\bullet j}}{c_{\bullet\bullet}} = \frac{u_i \langle u_\bullet \rangle \langle u_\bullet \rangle u_j}{\langle u_\bullet \rangle^2} = u_i u_j = \mathrm{E}\Big[c_{ij}\Big]. \tag{7}$$

Hence the APC subtracts approximately the expectation value from c_{ij} if it factorizes over columns.

EntC can overcome a major impediment to progress, the APC

As we have seen, the success of the average product correction (APC) (and other denoising techniques such as LRS [33]) depends on the specific form of the bias it can correct. The combination of pseudo-likelihood maximization, L_2 regularization, and the definition of the contact score as the norm of the coupling vector $\|\tilde{\mathbf{w}}_{ij}\|$ lead to a factorized form of the entropy, the leading cause of bias to correct for. It is plausible that changing the statistical model, its method of optimization, the regularization, or the contact score will usually result in the entropy bias to *not* factorize any more. For example, exchanging the L_2 regularization by an L_1 regularization destroys the factorization property. Therefore, even though the latter regularizer might

work better, it can still perform worse after APC because APC does not correct its entropy bias well any more. A potentially very valuable result of this work is therefore the insight into what the APC actually corrects. If we can work out the expectation value of the contact score under the L_1 regularization, for example, we could apply the appropriate entropy bias correction specifically for that model and regularization.

As another example, consider the following contact score, which uses amino acid pair-specific weights β_{ab} to upweight those pairs that are more predictive of contacts than others:

$$c'_{ij} = \sum_{a,b=1}^{20} \beta_{ab}\, w_{ij}(a,b)^2 \,. \tag{8}$$

The expectation value of this score does not factorize into separate terms for i and j any more and therefore the average product correction fails. Similarly, even neural networks would have a hard time to combine the coupling coefficients $w_{ij}(a, b)^2$ while learning to subtract the correct expectation value at the same time. This explains why it has been so difficult to improve on the popular combination of L_2 regularization and contact score $c_{ij} = \|\tilde{\mathbf{w}}_{ij}\|$ in combination with the APC.

But by subtracting the correct expectation value for each pair (a, b),

$$E[w_{ij}^2(a,b)] \approx \frac{N^2}{\lambda_w^2 (N-1)} f_{ia}(1-f_{ia}) f_{jb}(1-f_{jb}) \tag{9}$$

we should be able to overcome this roadblock. For instance we can now define a score with weights β_{ab} whose expectation value under the null model is near 0, as it should,

$$s_{ij} = \sum_{a,b=1}^{20} \beta_{ab} (w_{ij}(a,b)^2 - E[w_{ij}^2(a,b)]) \,. \tag{10}$$

This equation allows for the correction of individual couplings $w_{ij}(a, b)$. It could therefore be used to train deep neural networks directly on the EntC-corrected coupling coefficients $w_{ij}(a, b)$, combining the advantages of entropy correction with learning directly from the full set of coupling coefficients [21, 24] instead of only from their EntC-corrected norms $\|w_{ij}\|$, as given in Eq 3.

Persistent contrastive divergence facilitates inference of high quality models

Pseudo-likelihood maximization is the state-of-the-art inference technique for MRF models in contact prediction. Whereas the approximate nature of the model is sufficient for the correct ranking of residue pairs, the model is not exact in a way that it can reliably reproduce the empirical amino acid statistics of the original MSA. We implemented an alternative inference technique for MRFs, known as persistent contrastive divergence (PCD) which yields similar precision for predicted contacts but permits learning the fine statistics of the MRF model with higher precision. Even though other accurate model inference methods such as ACE [39] or bmDCA [37] can infer model parameters up to arbitrary precision, they are computationally intensive and their applicability is limited to small proteins. On the PSICOV dataset, our open source Python implementation of the PCD algorithm, CCMpredPy, was only about seven times slower than pseudo-likelihood maximization. (Its speed is proportional to the number of Markov chains and thereby depends on the required accuracy.) CCMpredPy might therefore be of use for large-scale studies that require exact models, such as investigating mutational effects or designing new protein features.

CCMgen allows the generation of realistic synthetic alignments

We developed CCMgen, the first tool for generating realistic MSAs of protein sequences for a given phylogenetic tree whose residues follow the pairwise coupling constraints from a Markov random field model. CCMgen provides full control of parameters that determine the interdependencies between sequences through the specification of the phylogenetic topology and the evolutionary rate of the sampling process. It enables to distinguish different sources of noise observed in alignments and how they affect the performance of residue-residue contact predictions. We believe CCMgen will prove to be useful for improving and validating contact prediction methods.

In this study, we demonstrated how CCMgen can be applied to analyse the noise contributions from entropy and phylogeny. Given MRF models learnt on real MSAs, we generated synthetic MSAs with statistically coupled amino acid columns from two types of phylogenetic trees, one in which the sequences are maximally independent (star topology) and one in which the statistical dependences are much stronger (binary tree). By predicting contacts from the two types of synthetic alignments and correcting the predicted contacts either with the APC or with our proposed entropy bias correction, we were able to elucidate the effect of phylogenetic and entropic noise on contact prediction accuracy.

According to the quantification of noise effects, the most important goal for residue-residue contact prediction is an accurate treatment of entropic bias, as it accounts for roughly twice the amount of correctable noise and is especially important for correctly identifying the strongest evolutionary couplings. However, phylogenetic noise has an important contribution to the predictions and only a fraction of it is probably corrected by the popular average product correction (APC). This result shows that it might be very worthwhile to develop methods for contact prediction and for learning of MRFs that can explicitly take the statistical dependencies of sequences by common descent into account.

Materials and methods

Recap: MRFs model statistical couplings between columns in a MSA

To predict contacts between residues, a popular approach is to train a Markov random field (MRF) model describing the probability to observe a sequence $\mathbf{x} = (x_1, \ldots, x_L)$ of length L with $x_i \in \{1, \ldots, 20\}$ representing the 20 amino acids,

$$p(\mathbf{x}|\mathbf{v}, \mathbf{w}) = \frac{1}{Z} \exp\left(\sum_{i=1}^{L} v_i(x_i) + \sum_{i<j}^{L} w_{ij}(x_i, x_j) \right). \tag{11}$$

The couplings $w_{ij}(a, b)$ describe the preference to find amino acid a at position i and b at j in the same sequence in relation to the probability if these positions were independent, as parametrized by the single-column amino acid preferences $v_i(a)$. Z is the normalization constant, equal to the sum of the exp function in the numerator summed over all possible 20^L sequences.

To estimate the parameters $v_i(a)$ and $w_{ij}(a, b)$ of the MRF, the logarithm of the likelihood for all sequences in the MSA, equal to the sum over the log-likelihood of each sequence $\mathbf{x}_n$, could be maximized: $\sum_{n=1}^{N} \log\ p(\mathbf{x}_n|\mathbf{v}, \mathbf{w}) \rightarrow \max$. A regularization term that pushes all parameters towards zero needs to be added to prevent overtraining, most commonly a L_2

penalty,

$$R(\mathbf{w}) = -\frac{1}{2}\lambda \sum_{i<j}^{L} \sum_{a,b=1}^{20} w_{ij}(a,b)^2. \tag{12}$$

But the huge number 20^L of terms in Z renders an exact solution infeasible for realistic protein lengths.

A number of approximations have been developed for this general class of problems. The approach that has consistently been found to work best for residue contact prediction is the pseudo-likelihood approximation, in which we replace the likelihood with the pseudo-likelihood and maximize the regularized log pseudo-likelihood [42–44],

$$\begin{aligned} PL(\mathbf{v},\mathbf{w}) &= \prod_{n=1}^{N} \prod_{i:x_{ni}\neq 0}^{L} p(x_{ni}|\mathbf{x}_{n,\setminus i},\mathbf{v},\mathbf{w}) I(x_{ni}\neq 0) \\ PL_{\text{reg}}(\mathbf{v},\mathbf{w}) &= \prod_{n=1}^{N} \prod_{i:x_{ni}\neq 0}^{L} \frac{1}{Z_{ni}} e^{v_i(x_{ni}) + \sum_{i<j}^{L} w_{ij}(x_{ni},x_{nj})} + \exp(R(\mathbf{w})) \underset{\mathbf{v},\mathbf{w}}{\longrightarrow} \max. \end{aligned} \tag{13}$$

Here, $\mathbf{x}_{n,\setminus i}$ denotes the vector obtained from $\mathbf{x}_n$ by removing the i'th component and $Z_{ni} = \sum_{c=1}^{20} \exp\left(v_i(c) + \sum_{j:j\neq i}^{L} w_{ij}(c,x_{nj})\right)$ is a normalization constant, which can therefore be evaluated easily. The second product runs over all columns i for which x_{ni} is not a gap (represented by a 0).

Once the parameters $\mathbf{v}$, $\mathbf{w}$ are estimated from a MSA, we can predict contacts for pairs of positions i and j using their statistical couplings. The most widely used score for residue contact prediction simply takes the L_2 norm $\|\mathbf{w}_{ij}\|_2$ of the 20×20-dimensional vector $\mathbf{w}_{ij}$ with elements $w_{ij}(a, b)$ (Eq 2) [42, 43, 51, 53, 54]. In this study, we chose the regularization strength $\lambda = 0.2(L-1)$ [51].

Sequence weighting and gap treatment

Sequences in a MSA do not represent independent draws from a probabilistic model. To reduce the effects of redundant sequences, we employ a popular sequence reweighting strategy that has been found to improve contact prediction performance. Every sequence x_n of length L with $n \in \{1, \ldots, N\}$ in an alignment with N sequences has an associated weight $\omega_n = 1/m_n$, where m_n represents the number of similar sequences:

$$m_n = \sum_{m=1}^{N} I(\text{Id}(x_n, x_m) \geq 0.8)\,, \tag{14}$$

$$\text{Id}(x_n, x_m) = \frac{1}{L}\sum_{i=1}^{L} I(x_n^i = x_m^i)\,. \tag{15}$$

An identity threshold of 0.8 has been used for all analyses. Amino acid counts and frequencies are computed with respect to the sequence weights. For example,

$$f_i(a) = \sum_{n=1}^{N} \omega_n I(x_{ni} = a) \Big/ \sum_{n=1}^{N} \omega_n \tag{16}$$

is the weighted fraction of sequences that have an amino acid a in column i.

We treat gaps as missing information and not as a 21st character. An example is Eq 13, where the second product runs over all MSA columns i except those having a gap in sequence n, $x_{ni} = 0$. This gap treatment leads to very minor changes both to the results and to the equations with respect to treating gaps as 21st character, e.g. the weighted number of sequences $\sum_{n=1}^{N} \omega_n$ gets replaced by $N_i = \sum_{n=1}^{N} \omega_n I(x_{ni} \neq 0)$ (the summed weight of sequences that do not contain a gap at positions i of the MSA), or by $N_{ij} = \sum_{n=1}^{N} \omega_n I(x_{ni} \neq 0, x_{nj} \neq 0)$. See, for example, Eqs 24 and 31 (for details see subsection 3.7.2 of PhD thesis of Susann Vorberg, available from soeding@mpibpc.mpg.de).

Recap: Average product correction

The APC subtracts from each score $c_{ij} = \|\mathbf{w}_{ij}\|_2$ the product of the average score $c_{i\bullet}$ for row i times the average score $c_{j\bullet}$ for column j divided by the average score $c_{\bullet\bullet}$ over all cells [26]:

$$c_{ij}^{\mathrm{APC}} = c_{ij} - \frac{c_{i\bullet} c_{j\bullet}}{c_{\bullet\bullet}}. \tag{17}$$

The APC ensures that the average of the corrected coupling score over each column and over each row is 0. This can be verified by summing Eq 17 over all i or j. The assumption made is that, since each residue is only in contact with a small fraction of all residues, the mean coupling score over a column or row is dominated by the systematic score bias on all pairs in the column or row rather than by the coupling scores on a small fraction of contacting residues. APC can also be interpreted as an approximation to the first principal component of the raw contact matrix [33]. It therefore removes the highest variability in the raw contact matrix that is assumed to arise from background biases.

Entropy correction

We define the following entropy bias correction (EntC), which depends solely on the per-column entropies of the MSA from which the MRF was trained:

$$c_{ij}^{\mathrm{EC}} = c_{ij} - \alpha\, s_i^{\frac{1}{2}} s_j^{\frac{1}{2}} \tag{18}$$

where α is a coefficient determining the strength of the correction, and

$$s_i = -\sum_{a=1}^{20} f_i(a) \log_2 f_i(a) \tag{19}$$

is the entropy of column i.

We determine α by analytically minimizing the sum of squares of the corrected off-diagonal coupling scores,

$$\sum_{i \neq j}^{L} \left(c_{ij} - \alpha\, s_i^{\frac{1}{2}} s_j^{\frac{1}{2}} \right)^2 \rightarrow \min_{\alpha}, \tag{20}$$

By setting the derivative to zero we obtain the optimal α value,

$$\alpha = \frac{\sum_{i \neq j}^{L} c_{ij}\, s_i^{\frac{1}{2}} s_j^{\frac{1}{2}}}{\sum_{i \neq j}^{L} s_i\, s_j}. \tag{21}$$

We also investigated other correction strategies using entropy statistics computed from the input MSA, such as the joint entropy for pairs of columns or different exponents in Eq 18. The

resulting variations of the entropy correction performed comparably regarding the average correlation with APC as well as precision of contact predictions.

Quantitative motivation of the entropy correction

We are given an MSA under the model that the sequences evolved under no pair couplings, that is, $w_{ij}(a, b) = 0$ for all columns i, j and all amino acids a, b. The square of the coupling score for columns i and j is $c_{ij}^2 = \sum_{a,b=1}^{20} \tilde{w}_{ij}(a, b)^2$, where $\tilde{w}_{ij}(a, b)$ are our estimates of the coupling coefficients learnt by maximizing the regularized pseudo-likelihood $PL(\mathbf{v}, \mathbf{w})$ in Eq 13. Our task is to calculate the expectation value of the coupling scores $c_{ij} = \left(\sum_{a,b=1}^{20} \tilde{w}_{ij}(a, b)^2\right)^{1/2}$. This expectation value under the null model of no couplings will be subtracted from the score to obtain the entropy-corrected score. For simplicity, we first assume that all sequences are independent draws from an MRF (with zero pair couplings).

From Eq 13 we derive the logarithm of the regularized pseudo-likelihood,

$$\begin{aligned} PLL_{\text{reg}}(\mathbf{v}, \mathbf{w}) = & \sum_{n=1}^{N} \sum_{i:x_{ni}\neq 0}^{L} \left(v_i(x_{ni}) + \sum_{j:j\neq i}^{L} w_{ij}(x_{ni}, x_{nj}) - \log Z_{ni}(\mathbf{v}, \mathbf{w}) \right) \\ & - \frac{\lambda}{2} \sum_{i\neq j}^{L} \sum_{a,b=1}^{20} w_{ij}(a, b)^2 \,. \end{aligned} \tag{22}$$

At the local and global optimum, its partial derivatives with respect to the coupling coefficients must vanish:

$$\begin{aligned} \frac{\partial PLL_{\text{reg}}}{\partial w_{ij}(a, b)} = & \sum_{n=1}^{N} I(x_{ni} = a, x_{nj} = b) - \sum_{n:x_{ni}\neq 0}^{N} \left(\frac{1}{Z_{ni}(\mathbf{v}, \mathbf{w})} \frac{\partial Z_{ni}(\mathbf{v}, \mathbf{w})}{\partial w_{ij}(a, b)} \right) \\ & - \lambda w_{ij}(a, b) = 0 \\ \frac{\partial PLL_{\text{reg}}}{\partial w_{ij}(a, b)} = & \; n_{ijab} - \sum_{n:x_{ni}\neq 0}^{N} p(x_{ni} = a | \mathbf{x}_{n,\setminus i}, \mathbf{v}, \mathbf{w}) \, I(x_{nj} = b) - \lambda w_{ij}(a, b) = 0 \,, \end{aligned} \tag{23}$$

where $n_{ijab} := \sum_{n=1}^{N} I(x_{ni} = a, x_{nj} = b)$ counts how often a appears in column i at the same time as b in column j.

Under the hypothesis that none of the columns is coupled to any other and that the regularization λ is sufficiently strong, the estimated coupling coefficients $\tilde{w}_{ij}(a, b)$ will all be fairly small and scattered around zero. Therefore, the model probabilities $p(x_{ni} = a | \mathbf{x}_{n,\setminus i}, \tilde{\mathbf{v}}, \tilde{\mathbf{w}})$ can be approximated by the empirical frequency $f_{ia} := n_{ia}/N = \sum_{n=1}^{N} I(x_{ni} = a)/N$. Hence Eq 23 reduces to

$$\lambda \tilde{w}_{ij}(a, b) \approx n_{ijab} - \frac{N}{N_i} f_{ia} \sum_{n=1}^{N} I(x_{ni} \neq 0, x_{nj} = b) \,. \tag{24}$$

Because under the null model gaps at position i occur approximately independently from b at j, $\sum_{n=1}^{N} I(x_{ni} \neq 0, x_{nj} = b) \approx (1/N) \sum_{n=1}^{N} I(x_{ni} \neq 0) \times \sum_{n=1}^{N} I(x_{nj} = b) = (N_i/N) N f_{jb}$, we obtain

$$\lambda \tilde{w}_{ij}(a, b) \approx n_{ijab} - N f_{ia} f_{jb}. \tag{25}$$

We now show that the counts n_{ijab} are distributed according to a *hypergeometric distribution*,

$$p(k = n_{ijab}|f_{ia}, f_{jb}, N) = \text{Hypergeom}(k = n_{ijab}|n, K, N). \tag{26}$$

with $k = n_{ijab}|$, $n = Nf_{ia}$, and $K = Nf_{jb}$. Suppose you draw n objects (here: sequences) without replacement from a set of N objects, and K of these N objects have a certain feature (here: $x_{nj} = b$) while $N - K$ don't. Then the probability that k out of the n drawn objects have the feature is given by the hypergeometric distribution. In our case, the subset of objects = sequences that is drawn is the set of $n = Nf_{ia}$ sequences that have an a in column i. The number n_{ijab} of these sequences that also have the feature $x_{nj} = b$ is therefore distributed according to the hypergeomteric distribution.

The expectation value for a variable $k = n_{ijab}$ is $\mathrm{E}[n_{ijab}] = nK/N = N f_{ia} f_{jb}$. Therefore, the square of the coupling score c_{ij} can be expressed as

$$c_{ij}^2 = ||\mathbf{w}_{ij}||_2^2 = \sum_{a,b=1}^{20} \tilde{w}_{ij}(a,b)^2 \approx \frac{1}{\lambda^2} \sum_{a,b=1}^{20} (n_{ijab} - \mathrm{E}[n_{ijab}])^2 \,. \tag{27}$$

The expectation value of the numerator is $(n_{ijab} - \mathrm{E}[n_{ijab}])^2 = \mathrm{var}[n_{ijab}]$ which is $\frac{nK}{N}\frac{N-K}{N}\frac{N-n}{N-1}$, or, using our notation $n = Nf_{ia}$, and $K = Nf_{jb}$,

$$\begin{aligned} \mathrm{E}[c_{ij}^2] &\approx \frac{1}{\lambda^2} \sum_{a,b=1}^{20} \frac{Nf_{ia}(1-f_{ia})\, Nf_{jb}(1-f_{jb})}{N-1} \,. \\ \mathrm{E}[c_{ij}^2] &\approx \frac{N^2}{\lambda^2(N-1)} \left(\sum_{a=1}^{20} f_{ia}(1-f_{ia}) \right) \left(\sum_{b=1}^{20} f_{jb}(1-f_{jb}) \right) . \end{aligned} \tag{28}$$

Remarkably, the expectation value factorizes into a term depending only on i and one depending only on j. The factorization is in fact the reason why the APC (Eq 17) works so well, since the APC subtracts a product of two terms, $c_{i\bullet}/c_{\bullet\bullet}^{1/2} \times c_{j\bullet}/c_{\bullet\bullet}^{1/2}$, one depending only on i and the other only on j.

The factors in Eq 28 are highly correlated with the column entropies s_i (Fig 8A), so that we can write

$$\mathrm{E}[c_{ij}^2] \underset{\sim}{\propto} \frac{N^2}{\lambda^2(N-1)} s_i s_j \,. \tag{29}$$

Finally, the variance of c_{ij} is small in comparison to $\mathrm{E}[c_{ij}^2]$ because usually many approximately independent terms $(n_{ijab} - \mathrm{E}[n_{ijab}])^2$ contribute to c_{ij}^2 such that the fluctuations around the expectation value of each such term tend to average each other out. We can therefore approximate the entropic bias as

$$\mathrm{E}[c_{ij}] = (\mathrm{E}[c_{ij}^2] - \mathrm{var}[c_{ij}])^{1/2} \approx \mathrm{E}[c_{ij}^2]^{1/2} \underset{\sim}{\propto} \frac{N^{1/2}}{\lambda} s_i^{1/2} s_j^{1/2} \,. \tag{30}$$

Given that Eq 28 is a more accurate estimate for $\mathrm{E}[c_{ij}^2]$ than Eq 30 is for $\mathrm{E}[c_{ij}]$, we were expecting better results by predicting contacting residues (i, j) based on a ranking by $c_{ij}^2 - \alpha u_i u_j$ with $u_i = \sum_{a=1}^{20} f_{ia}(1 - f_{ia})$ than when we ranked by the entropy bias corrected score $c_{ij} - \alpha s_i^{1/2} s_j^{1/2}$. To our surprise, the entropy bias correction worked slightly better. Investigation of this puzzling result is left for future work.

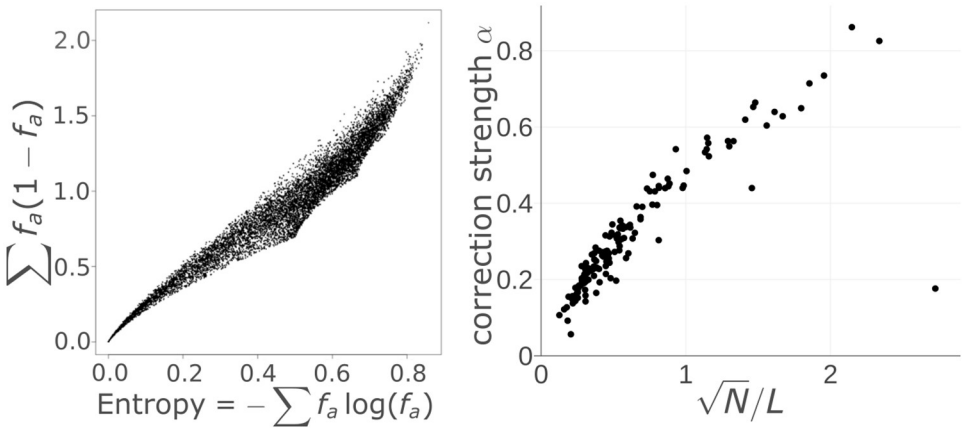

Fig 8. Entropy bias correction. (**A**) Each dot shows the entropy $-\sum_{a=1}^{20} f_a \log f_a$ of a randomly sampled probability distribution over the 20 amino acids versus $\sum_{a=1}^{20} f_a(1-f_a)$ for the same distribution. (**B**) For each of the 150 Pfam MSA we plot the strength of the entropy bias correction α from Eq 21 versus $\sqrt{N}/L$, where N is the number of sequences in the MSA and L is the number of columns. Without sequence weighting, a linear dependence would be expected from Eq 30.

https://doi.org/10.1371/journal.pcbi.1006526.g008

We used a regularization strength proportional to the number of residues L in the MSA, $\lambda = 0.2L$. Therefore, without sequence weighting, $\frac{N^{1/2}}{\lambda}$ should be proportional to the α parameter from Eq 21 that defines the optimum strength of the entropy bias. Indeed, Fig 8B shows a tight correlation of α with $\sqrt{N}/L$. We cannot expect the relationship to be strictly linear, because in our theoretical analysis we had assumed that sequences are independent and have a weight of 1, whereas in the example of Fig 8B the coupling coefficients $\tilde{w}_{ij}(a,b)$ where learned from Pfam MSAs using sequence weighting.

Learning MRFs with persistent contrastive divergence

While the log likelihood function cannot be efficiently computed because of the exponential complexity of the normalization constant Z, it is possible to approximate its gradient with an approach called contrastive divergence [55]. The gradient of the log likelihood with respect to the couplings $w_{ij}(a,b)$ can be written as

$$\frac{\partial}{\partial w_{ij}(a,b)}\left[\sum_{n=1}^{N}\left(\sum_{i=1}^{L} v_i(x_i) + \sum_{i<j}^{L} w_{ij}(x_i,x_j)\right) - \log Z\right] = N_{ij} q(x_i = a, x_j = b) - N_{ij}\, p(x_i = a, x_j = b|\mathbf{v},\mathbf{w})\,, \tag{31}$$

where $N_{ij} = \sum_{n=1}^{N} \omega_n I(x_{ni} \neq 0, x_{nj} \neq 0)$ is the summed weight of sequences that have no gap in either column i or j, $q(x_i = a, x_j = b) = \frac{1}{N_{ij}} \sum_{n=1}^{N} \omega_n I(x_{ni} = a, x_{nj} = b)$ represents the empirically observed pairwise amino acid frequencies that are normalized over $a, b \in \{1, \ldots, 20\}$, and $p(x_i = a, x_j = b|\mathbf{v},\mathbf{w})$ corresponds to the model probabilities of the MRF for observing an amino acid pair (a, b) at positions i and j. The empirical amino acid counts, given by $N_{ij}q(x_i = a, x_j = b)$, are constant and need to be computed only once from the alignment.

The marginal distributions of the MRF cannot be computed analytically as it involves the normalization constant Z. Markov chain Monte Carlo (MCMC) algorithms can be used to generate samples from probability distributions that involve the computation of complex integrals such as the normalization constant Z. Given that the Markov chains run long enough, the

equilibrium statistics of the samples will be identical to the true probability distribution statistics. Thus, an estimate of the marginal distribution of the MRF in the gradient in Eq 31, $p(x_i = a, x_j = b|\mathbf{v}, \mathbf{w})$, can be obtained by simply computing the expected amino acid counts from MCMC samples. However, MCMC methods require many sampling steps to obtain unbiased estimates from the stationary distribution which comes at high computational costs.

Hinton suggested contrastive divergence (CD) as an approximation to MCMC methods [55]. The idea is simple: instead of starting a Markov chain from a random point and running it until it has reached the stationary distribution, we run C chains in parallel, each being initialized with one of the sequences from the input MSA and we evolve them for only a small number of steps. Obviously the chains do not converge to the stationary distribution in only a few steps and the sequence samples obtained from the current configuration of the chains present biased estimates. The intuition behind CD is that even though the resulting gradient estimate from the biased samples will also be noisy and biased, it points roughly into the same direction as the true gradient of the full likelihood. Therefore the approximate CD gradient should become zero approximately where the true gradient of the likelihood becomes zero.

We apply CD and generate sequence samples to estimate the marginal probabilities by evolving Markov chains which have been initialized with randomly selected protein sequences from the original Pfam MSAs for one full step of Gibbs sampling. We set the number of Markov chains to $C = \max(500, 0.1N)$, with N being the number of sequences in the MSA, which seems to give a good trade-off between performance and runtime. Gibbs sampling requires updating at each sampling step all sequence positions x_i with $i \in \{1, \ldots, L\}$ (L = sequence length). For each position, a new amino acid a is chosen according to the conditional probability

$$p(x_i^{t+1} = a|\mathbf{x}_{-i}^t, \mathbf{v}, \mathbf{w}) \propto \exp\left(v_i(a) + \sum_{j \neq i}^{L} w_{ij}(a, x_j^t)\right). \quad (32)$$

This Gibbs sampling approach is known to generate samples $\mathbf{x}^0, \ldots, \mathbf{x}^t$ that are distributed according to the model probability in Eq 11 [56, 57]. Note that we do not update positions representing a gap and we thereby retain the gap structure of the initial sequence.

A modification of CD known as persistent contrastive divergence (PCD) does not reinitialize the Markov chains at data samples every time a new gradient is computed [46]. Instead, the Markov chains are kept persistent: they are evolved between successive gradient computations. The assumption behind PCD is that the model changes only slowly between parameter updates given a sufficiently small learning rate. Consequently, the Markov chains will not be pushed too far from equilibrium after each update but rather stay close to the stationary distribution [46, 58, 59].

Tieleman and others observed that PCD performs better than CD in all practical cases tested, even though CD can be faster in the early stages of learning [46, 58, 60]. Therefore we start optimizing the full likelihood with CD and switch to PCD at later stages of learning. CCMpredPy settings for training a MRF with persistent contrastive divergence are listed in S1 Text.

Generating MCMC samples from MRFs with CCMgen

MCMC samples for the analysis in Fig 1 have been generated with CCMgen by evolving 10000 Markov chains by repeated Gibbs sampling as described in Eq 32. The Markov chains, each representing protein sequences of length L (length of protein in the PSICOV data set) have been randomly initialized with the 20 amino acids. Since the alignment substructure is

strongly impacted by the non-random distribution of gaps in the sequences (S2a Fig), in a second step the gap structure of randomly selected sequences from the original Pfam alignment is copied over (gaps represented as 21st amino acid). Thus, it is ensured that the sampling procedure reproduces the original alignment substructure as closely as possible (S2b and S2c Fig). The number of Gibbs steps before drawing samples was set to 500. Increasing the number of Gibbs steps to e.g. 1000 does not change the statistics of the MCMC samples, hence we can assume that the Markov chains have reached the equilibrium distribution. CCMgen settings for MCMC sampling are listed in S1 Text.

Sampling sequences from MRFs along phylogenies with CCMgen

Instead of evolving sequences along a linear path, the MRF model can also be used to sample protein sequences according to an arbitrary phylogenetic tree.

CCMgen can simulate the evolution of sequences along any given phylogenetic tree constrained by a MRF model, such as those calculated from CCMpred for example. The user can either supply a phylogenetic tree in Newick format that has been generated by a phylogenetic reconstruction program such as FastTree [61] on a real alignment or choose between two types of idealized trees, a binary and a star-shaped topology. For these idealized trees the user can specify the number of leaf nodes and the total depth of the tree, which is the total number of mutations per position from the sequence at the root to the leaf nodes. The root sequence can either be supplied by the user or be generated by evolving an all-alanine sequence with a number of mutations (i.e. Gibbs sampling steps according to the MRF as described in Eq 32). Sequences at subsequent child nodes are generated one by one, by duplicating the sequence at the parent node and evolving the respective child node sequences each with a number of mutations proportional to the edge length. The output of CCMgen is a MSA file with the sequences at the leaf nodes of the tree. CCMgen is released as open-source python command-line application.

Workflow for quantification of noise in contact prediction with CCMgen. We used CCMpredPy to learn MRF models for all Pfam alignments in the PSICOV data set using PCD. In order to obtain models with few but precise constraints, we set coupling parameters to zero for non-contacting residue pairs (C_β distance > 12Å) by initializing them to zero and setting the gradients to zero in the optimization. This procedure ensures that the majority of residue pairs not forming contacts in the protein structure will not be coupled in the MRF model.

We used CCMgen with the learned MRF models to generate synthetic alignments by evolving sequences along idealized star and binary tree topologies. The ancestral sequence at the root of a tree, $\mathbf{x}^t$ with $t = 0$, is obtained by evolving an all-alanine sequence for 10 steps of Gibbs sampling as described in Eq 32. The synthetic alignments have the same number of sequences as the corresponding Pfam alignments from the PSICOV set.

We also ensure that the diversity of the resulting MSAs is within 1% of the diversity of the original Pfam MSA from the PSICOV set by adjusting the depth of the trees, which is equivalent to adapting the mutation rate (S5a Fig). We measure diversity as the number of effective sequences, N_{eff}, defined as the exponential of the average column entropies, as defined in the HH-suite software package [62].

For each new CCMgen run with adapted mutation rate a new ancestral sequence is sampled by evolving an all-alanine sequence for 10 Gibbs steps. We found that the aforementioned procedure of masking non-contacting residue pairs during training of the MRF model has the advantageous effect of allowing smaller mutation rates to be used to achieve the desired diversity compared to sampling sequences with a fully parametrized model, likely due to the smaller number of constraints trapping the sampling procedure in local optima. Enforcing small

mutation rates is essential for preserving the interdependence between sequences when sampling along binary topologies, which consequently controls the amount of phylogenetic bias. The mutation rates used to obtain synthetic alignments of similar diversity as the original Pfam alignments are very similar regardless of the phylogenetic topology along which sequences were sampled (S5b Fig).

CCMpredPy and CCMgen settings for training the MRF with persistent contrastive divergence and generating the synthetic alignments along binary and star tree topologies are listed in S1 Text.

Dataset and preprocessing

We used the PSICOV data set that was published together with the PSICOV method [49] and which comprises MSAs for 150 Pfam domains with known crystal structures. For each Pfam MSA in the PSICOV set we first removed sequences with more than 75% gaps and columns with more than 50% gaps, similarly as in [50, 51, 63], to reduce the well-known impact of gaps on the analysis.

Supporting information

S1 Text. Running CCMpredPy and CCMgen. Parameter settings used with CCMpredPy and CCMgen in this study.
(PDF)

S2 Text. Analysing alignment substructure with principal component analysis (PCA).
(PDF)

S1 Fig. Comparing quality of Markov random field (MRF) models learned with pseudo-likelihood maximization and persistent contrastive divergence (PCD).
(PDF)

S2 Fig. Visualizing alignment substructure for protein 1gmxA by projecting sequences of the multiple sequence alignment (MSA) onto their first two principal components.
(PDF)

S3 Fig. Contact scores computed from Markov random field (MRF) models trained with pseudo-likelihood maximization and persistent contrastive divergence (PCD) correlate strongly.
(PDF)

S4 Fig. Distribution of Pearson correlation coefficients between average product correction (APC) and entropy bias correction (EntC) term.
(PDF)

S5 Fig. Statistics of synthetic alignments generated with CCMgen along binary and star tree topologies.
(PDF)

Acknowledgments

We thank Sergey Ovchinnikov for helpful discussion and valuable advice during the revision of the manuscript.

Author Contributions

Conceptualization: Johannes Söding.

Data curation: Susann Vorberg.

Formal analysis: Johannes Söding.

Funding acquisition: Johannes Söding.

Investigation: Susann Vorberg, Stefan Seemayer.

Methodology: Susann Vorberg, Johannes Söding.

Project administration: Johannes Söding.

Software: Susann Vorberg, Stefan Seemayer.

Supervision: Johannes Söding.

Validation: Susann Vorberg, Stefan Seemayer.

Visualization: Susann Vorberg, Stefan Seemayer.

Writing – original draft: Susann Vorberg, Stefan Seemayer.

Writing – review & editing: Susann Vorberg, Johannes Söding.

References

1. Göbel U, Sander C, Schneider R, Valencia A. Correlated mutations and residue contacts in proteins. Proteins. 1994; 18(4):309–317. https://doi.org/10.1002/prot.340180402 PMID: 8208723
2. Neher E. How frequent are correlated changes in families of protein sequences? Proc Natl Acad Sci U S A. 1994; 91(1):98–102. https://doi.org/10.1073/pnas.91.1.98 PMID: 8278414
3. Shindyalov IN, Kolchanov NA, Sander C. Can three-dimensional contacts in protein structures be predicted by analysis of correlated mutations? "Protein Eng Des Sel. 1994; 7(3):349–358. https://doi.org/10.1093/protein/7.3.349
4. Godzik A, Sander C. Conservation of residue interactions in a family of Ca-binding proteins. "Protein Eng Des Sel. 1989; 2(8):589–596. https://doi.org/10.1093/protein/2.8.589
5. Lapedes A, Giraud B, Liu L, Stormo G. Correlated mutations in models of protein sequences: phylogenetic and structural effects. Stat Mol Biol. 1999; 33:236–256.
6. Giraud B, Heumann J, Lapedes A. Superadditive correlation. Phys Rev E. 1999; 59(5):4983–4991. https://doi.org/10.1103/PhysRevE.59.4983
7. Weigt M, White RA, Szurmant H, Hoch JA, Hwa T. Identification of direct residue contacts in protein-protein interaction by message passing. Proc Natl Acad Sci U S A. 2009; 106(1):67–72. https://doi.org/10.1073/pnas.0805923106 PMID: 19116270
8. Burger L, van Nimwegen E. Disentangling direct from indirect co-evolution of residues in protein alignments. PLoS Comput Biol. 2010; 6(1):e1000633. https://doi.org/10.1371/journal.pcbi.1000633 PMID: 20052271
9. Thomas J, Ramakrishnan N, Bailey-Kellogg C. Graphical Models of Residue Coupling in Protein Families. IEEE/ACM Trans Comput Biol Bioinforma. 2008; 5(2):183–197. https://doi.org/10.1109/TCBB.2007.70225
10. Marks DS, Colwell LJ, Sheridan R, Hopf TA, Pagnani A, Zecchina R, et al. Protein 3D structure computed from evolutionary sequence variation. PLoS One. 2011; 6(12):e28766. https://doi.org/10.1371/journal.pone.0028766 PMID: 22163331
11. Marks DS, Hopf TA, Sander C. Protein structure prediction from sequence variation. Nat Biotechnol. 2012; 30(11):1072–1080. https://doi.org/10.1038/nbt.2419 PMID: 23138306
12. Hopf TA, Colwell LJ, Sheridan R, Rost B, Sander C, Marks DS. Three-dimensional structures of membrane proteins from genomic sequencing. Cell. 2012; 149(7):1607–1621. https://doi.org/10.1016/j.cell.2012.04.012 PMID: 22579045

13. Nugent T, Jones DT. Accurate de novo structure prediction of large transmembrane protein domains using fragment-assembly and correlated mutation analysis. Proc Natl Acad Sci U S A. 2012; 109(24): E1540–7. https://doi.org/10.1073/pnas.1120036109 PMID: 22645369

14. Hopf TA, Schärfe CPI, Rodrigues JPGLM, Green AG, Sander C, Bonvin AMJJ, et al. Sequence co-evolution gives 3D contacts and structures of protein complexes. Elife. 2014; 3:e03430. https://doi.org/10.7554/eLife.03430

15. Kosciolek T, Jones DT. De novo structure prediction of globular proteins aided by sequence variation-derived contacts. PLoS One. 2014; 9(3):e92197. https://doi.org/10.1371/journal.pone.0092197 PMID: 24637808

16. Ovchinnikov S, Kamisetty H, Baker D. Robust and accurate prediction of residue-residue interactions across protein interfaces using evolutionary information. Elife. 2014; 3:e02030. https://doi.org/10.7554/eLife.02030 PMID: 24842992

17. Hayat S, Sander C, Marks DS, Elofsson A. All-atom 3D structure prediction of transmembrane β-barrel proteins from sequences. Proc Natl Acad Sci U S A. 2015; 112(17):5413–5418. https://doi.org/10.1073/pnas.1419956112 PMID: 25858953

18. Hopf TA, Morinaga S, Ihara S, Touhara K, Marks DS, Benton R. Amino acid coevolution reveals three-dimensional structure and functional domains of insect odorant receptors. Nat Commun. 2015; 6:6077. https://doi.org/10.1038/ncomms7077 PMID: 25584517

19. Ovchinnikov S, Kinch L, Park H, Liao Y, Pei J, Kim DE, et al. Large scale determination of previously unsolved protein structures using evolutionary information. Elife. 2015; 4:e09248. https://doi.org/10.7554/eLife.09248 PMID: 26335199

20. Ovchinnikov S, Park H, Varghese N, Huang PS, Pavlopoulos GA, Kim DE, et al. Protein structure determination using metagenome sequence data. Science. 2017; 355(6322):294–298. https://doi.org/10.1126/science.aah4043 PMID: 28104891

21. Jones DT, Kandathil SM. High precision in protein contact prediction using fully convolutional neural networks and minimal sequence features. Bioinformatics. 2018;bty341. https://doi.org/10.1093/bioinformatics/bty341

22. He B, Mortuza SM, Wang Y, Shen HB, Zhang Y. NeBcon: Protein contact map prediction using neural network training coupled with naïve Bayes classifiers. Bioinformatics. 2017; https://doi.org/10.1093/bioinformatics/btx164

23. Michel M, Skwark MJ, Menéndez Hurtado D, Ekeberg M, Elofsson A. Predicting accurate contacts in thousands of Pfam domain families using PconsC3. Bioinformatics. 2017; 33(18):2859–2866. https://doi.org/10.1093/bioinformatics/btx332 PMID: 28535189

24. Golkov V, Skwark MJ, Golkov A, Dosovitskiy A, Brox T, Meiler J, et al. Protein contact prediction from amino acid co-evolution using convolutional networks for graph-valued images. In: Lee DD, Sugiyama M, Luxburg UV, Guyon I, Garnett R, editors. Adv. Neural Inf. Process. Syst. 29. Curran Associates, Inc.; 2016. p. 4222–4230.

25. Wang S, Sun S, Li Z, Zhang R, Xu J. Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model. PLoS Comput Biol. 2016; 13(1):e1005324. https://doi.org/10.1371/journal.pcbi.1005324

26. Dunn SD, Wahl LM, Gloor GB. Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. Bioinformatics. 2008; 24(3):333–340. https://doi.org/10.1093/bioinformatics/btm604 PMID: 18057019

27. Gouveia-Oliveira R, Pedersen AG. Finding coevolving amino acid residues using row and column weighting of mutual information and multi-dimensional amino acid representation. Algorithms Mol Biol. 2007; 2(1):1–12. https://doi.org/10.1186/1748-7188-2-12

28. Gloor GB, Martin LC, Wahl LM, Dunn SD. Mutual Information in Protein Multiple Sequence Alignments Reveals Two Classes of Coevolving Positions. Biochemistry. 2005; 44(19):7156–7165. https://doi.org/10.1021/bi050293e PMID: 15882054

29. Martin LC, Gloor GB, Dunn SD, Wahl LM. Using information theory to search for co-evolving residues in proteins. Bioinformatics. 2005; 21(22):4116–4124. https://doi.org/10.1093/bioinformatics/bti671 PMID: 16159918

30. Noivirt O, Eisenstein M, Horovitz A. Detection and reduction of evolutionary noise in correlated mutation analysis. Protein Eng Des Sel. 2005; 18(5):247–53. https://doi.org/10.1093/protein/gzi029 PMID: 15911538

31. Fodor AA, Aldrich RW. Influence of conservation on calculations of amino acid covariance in multiple sequence alignments. Proteins. 2004; 56(2):211–21. https://doi.org/10.1002/prot.20098 PMID: 15211506

32. Atchley WR, Wollenberg KR, Fitch WM, Terhalle W, Dress AW. Correlations Among Amino Acid Sites in bHLH Protein Domains: An Information Theoretic Analysis. Mol Biol Evol. 2000; 17(1):164–178. https://doi.org/10.1093/oxfordjournals.molbev.a026229 PMID: 10666716

33. Zhang H, Gao Y, Deng M, Wang C, Zhu J, Li SC, et al. Improving residue-residue contact prediction via low-rank and sparse decomposition of residue correlation matrix. Biochem Biophys Res Commun. 2016; 472(1):217–222. https://doi.org/10.1016/j.bbrc.2016.01.188 PMID: 26920058

34. Buslje CM, Santos J, Delfino JM, Nielsen M. Correction for phylogeny, small number of observations and data redundancy improves the identification of coevolving amino acid pairs using mutual information. Bioinformatics. 2009; 25(9):1125–31. https://doi.org/10.1093/bioinformatics/btp135 PMID: 19276150

35. Lee BC, Kim D. A new method for revealing correlated mutations under the structural and functional constraints in proteins. Bioinformatics. 2009; 25(19):2506–2513. https://doi.org/10.1093/bioinformatics/btp455 PMID: 19628501

36. Tillier ERM, Lui TWH. Using multiple interdependency to separate functional from phylogenetic correlations in protein alignments. Bioinformatics. 2003; 19(6):750–755. https://doi.org/10.1093/bioinformatics/btg072 PMID: 12691987

37. Figliuzzi M, Barrat-Charlaix P, Weigt M. How pairwise coevolutionary models capture the collective residue variability in proteins. Mol Biol Evol. 2018; 35(4):1018–1027. https://doi.org/10.1093/molbev/msy007 PMID: 29351669

38. Barton JP, Chakraborty AK, Cocco S, Jacquin H, Monasson R. On the entropy of protein families. Journal of Statistical Physics. 2016 Mar 1; 162(5):1267–93. https://arxiv.org/pdf/1512.08101.pdf.

39. Barton JP, De Leonardis E, Coucke A, Cocco S. ACE: adaptive cluster expansion for maximum entropy graphical model inference. Bioinformatics. 2016; 32(20):3089–3097. https://doi.org/10.1093/bioinformatics/btw328 PMID: 27329863

40. Haldane A, Flynn WF, He P, Vijayan R, Levy RM. Structural propensities of kinase family proteins from a potts model of residue co-variation. Protein Sci. 258:1378–1384.

41. Sutto L, Marsili S, Valencia A, Gervasio FL From residue coevolution to protein conformational ensembles and functional dynamics. Proc Natl Acad Sci U S A 2015, 112(44):13567–13572. https://doi.org/10.1073/pnas.1508584112 PMID: 26487681

42. Seemayer S, Gruber M, Söding J. CCMpred-fast and precise prediction of protein residue-residue contacts from correlated mutations. Bioinformatics. 2014; 30(21):3128–3130. https://doi.org/10.1093/bioinformatics/btu500 PMID: 25064567

43. Ekeberg M, Hartonen T, Aurell E. Fast pseudolikelihood maximization for direct-coupling analysis of protein structure from many homologous amino-acid sequences. J Comput Phys. 2014; 276:341–356. https://doi.org/10.1016/j.jcp.2014.07.024

44. Balakrishnan S, Kamisetty H, Carbonell JG, Lee SI, Langmead CJ. Learning generative models for protein fold families. Proteins. 2011; 79(4):1061–78. https://doi.org/10.1002/prot.22934 PMID: 21268112

45. Cocco S, Feinauer C, Figliuzzi M, Monasson R, Weigt M. Inverse statistical physics of protein sequences: a key issues review. Reports Prog Phys. 2018; 81(3):032601. https://doi.org/10.1088/1361-6633/aa9965

46. Tieleman T. Training Restricted Boltzmann Machines using Approximations to the Likelihood Gradient. Proc 25th Int Conf Mach Learn. 2008; 307:7.

47. Morcos F, Pagnani A, Lunt B, Bertolino A, Marks DS, Sander C, et al. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. Proc Natl Acad Sci U S A. 2011; 108(49):E1293–301. https://doi.org/10.1073/pnas.1111471108 PMID: 22106262

48. Qin C, Colwell LJ. Power law tails in phylogenetic systems. Proc Natl Acad Sci U S A. (in press).

49. Jones DT, Buchan DWA, Cozzetto D, Pontil M. PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. Bioinformatics. 2012; 28(2):184–90. https://doi.org/10.1093/bioinformatics/btr638 PMID: 22101153

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

51. Kamisetty H, Ovchinnikov S, Baker D. Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era. Proc Natl Acad Sci U S A. 2013; 110(39):15674–15679. https://doi.org/10.1073/pnas.1314045110 PMID: 24009338

52. Monastyrskyy B, D'Andrea D, Fidelis K, Tramontano A, Kryshtafovych A. New encouraging developments in contact prediction: Assessment of the CASP11 results. Proteins. 2015; 84(Suppl 1):131–144. https://doi.org/10.1002/prot.24943 PMID: 26474083

53. Feinauer C, Skwark MJ, Pagnani A, Aurell E. Improving contact prediction along three dimensions. PLoS Comput Biol. 2014; 10(10):e1003847. https://doi.org/10.1371/journal.pcbi.1003847 PMID: 25299132

54. Baldassi C, Zamparo M, Feinauer C, Procaccini A, Zecchina R, Weigt M, et al. Fast and accurate multivariate gaussian modeling of protein families: predicting residue contacts and protein-interaction partners. PLoS One. 2014; 9(3):e92721. https://doi.org/10.1371/journal.pone.0092721 PMID: 24663061

55. Hinton GE. Training Products of Experts by Minimizing Contrastive Divergence. Neural Comput. 2002; 14(8):1771–1800. https://doi.org/10.1162/089976602760128018 PMID: 12180402

56. Gelfand AE, Smith AFM. Sampling-Based Approaches to Calculating Marginal Densities. J Am Stat Assoc. 1990; 85(410):398–409. https://doi.org/10.1080/01621459.1990.10476213

57. Geman S, Geman D. Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. IEEE Trans Pattern Anal Mach Intell. 1984; PAMI-6(6):721–741. https://doi.org/10.1109/TPAMI.1984.4767596

58. Murphy KP. Machine Learning: A Probabilistic Perspective. MIT Press; 2012.

59. Fischer A, Igel C. An Introduction to Restricted Boltzmann Machines. Lect Notes Comput Sci Prog Pattern Recognition, Image Anal Comput Vision, Appl. 2012; 7441:14–36.

60. Swersky K, Chen B, Marlin B, de Freitas N. A tutorial on stochastic approximation algorithms for training Restricted Boltzmann Machines and Deep Belief Nets. In: 2010 Inf. Theory Appl. Work. IEEE; 2010. p. 1–10. Available from: http://ieeexplore.ieee.org/document/5454138/.

61. Price MN, Dehal PS, Arkin AP. FastTree 2—Approximately Maximum-Likelihood Trees for Large Alignments. PLoS One. 2010; 5(3):e9490. https://doi.org/10.1371/journal.pone.0009490 PMID: 20224823

62. Remmert M, Biegert A, Hauser A, Söding J. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nat Methods. 2012; 9(2):173–175. https://doi.org/10.1038/nmeth.1818

63. Skwark MJ, Raimondi D, Michel M, Elofsson A. Improved contact predictions using the recognition of protein like contact patterns. PLoS Comput Biol. 2014; 10(11):e1003889. https://doi.org/10.1371/journal.pcbi.1003889 PMID: 25375897