OPEN ACCESS Freely available online

PLoS COMPUTATIONAL BIOLOGY

Disentangling Direct from Indirect Co-Evolution of Residues in Protein Alignments

Lukas Burger, Erik van Nimwegen*

Biozentrum, University of Basel, and Swiss Institute of Bioinformatics, Basel, Switzerland

Abstract

Predicting protein structure from primary sequence is one of the ultimate challenges in computational biology. Given the large amount of available sequence data, the analysis of co-evolution, i.e., statistical dependency, between columns in multiple alignments of protein domain sequences remains one of the most promising avenues for predicting residues that are contacting in the structure. A key impediment to this approach is that strong statistical dependencies are also observed for many residue pairs that are distal in the structure. Using a comprehensive analysis of protein domains with available three-dimensional structures we show that co-evolving contacts very commonly form chains that percolate through the protein structure, inducing indirect statistical dependencies between many distal pairs of residues. We characterize the distributions of length and spatial distance traveled by these co-evolving contact chains and show that they explain a large fraction of observed statistical dependencies between structurally distal pairs. We adapt a recently developed Bayesian network model into a rigorous procedure for disentangling direct from indirect statistical dependencies, and we demonstrate that this method not only successfully accomplishes this task, but also allows contacts with weak statistical dependency to be detected. To illustrate how additional information can be incorporated into our method, we incorporate a phylogenetic correction, and we develop an informative prior that takes into account that the probability for a pair of residues to contact depends strongly on their primary-sequence distance and the amount of conservation that the corresponding columns in the multiple alignment exhibit. We show that our model including these extensions dramatically improves the accuracy of contact prediction from multiple sequence alignments.

Citation: Burger L, van Nimwegen E (2010) Disentangling Direct from Indirect Co-Evolution of Residues in Protein Alignments. PLoS Comput Biol 6(1): e1000633. doi:10.1371/journal.pcbi.1000633

Editor: Philip E. Bourne, University of California San Diego, United States of America

Received February 11, 2009; **Accepted** December 4, 2009; **Published** January 1, 2010

Funding: The work in this study was funded by the University of Basel and partially by an SNF (http://www.snf.ch) grant (number 3100A0-118318) to EvN. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

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

* E-mail: erik.vannimwegen@unibas.ch

Introduction

The identification of functionally and structurally important elements in DNA, RNA and proteins from their sequences has been a major focus of computational biology for several decades. A common approach is to create a multiple alignment of homologous sequences, which places 'equivalent' residues into the same column and as such gives a hint of the evolutionary constraints that are acting on related sequences. In particular, so-called profile hidden Markov models [1] of protein families and domains have been highly successful in identifying sequences that have similar function and fold into a common structure, making them among the most important tools in functional genomics, see e.g. [2]. These hidden Markov models typically assume that the residues occurring at a given position are probabilistically independent of the residues occurring at other positions. At the time at which these models were developed, it was entirely reasonable to ignore dependencies between residues at different positions, since the amount of available sequence data was generally insufficient to estimate joint probabilities of multiple residues. However, currently the multiple alignments of many protein families and domains include hundreds and sometimes even thousands of sequences, making it possible to systematically investigate dependencies between the residues at different positions.

As the functionality of biomolecules crucially depends on their three-dimensional structures, whose stabilities depend on interactions between residues that are near to each other in space, it is of course to be expected that significant dependencies between residues at different positions will exist. Indeed such dependencies are evident for RNA (eg [3,4]) and protein sequences [5,6]. The existence of dependencies between residues at different positions is also supported by the observation of correlated mutations in which mutations at one residue tend to be compensated by a correlated mutation in a particular other residue [5–7].

Recently there has been a significant amount of work in which multiple alignments of single protein families have been used in order to predict pairs of residues that are functionally linked or interact directly in the tertiary structure (see eg [8–14] and references therein). This work has shown that pairs of residues which show statistical dependencies are generally significantly closer in the structure than randomly chosen pairs. However, it has been repeatedly noted that there exist many highly statistically-dependent residues that are distant in space (eg [14–16]). Figure 1 illustrates these points. One of the most commonly used measures of dependency between two residues is the mutual information [4,9,14,17,18] between the distributions of amino acids occurring in the two corresponding alignment columns. We collected a comprehensive set of 2009 multiple alignments of protein domains from the Pfam database [19] for which a three dimensional

Author Summary

Whenever two residues are in close contact in the structure of a protein, their interaction will often constrain which amino acid substitutions can occur without perturbing the functionality of the protein, leading to "co-evolution" of the residues. With the large amount of data currently available, deep multiple alignments can be constructed of protein sequences that likely fold into a common structure, and several methods have been proposed for predicting contacting residues from statistical dependencies exhibited by pairs of alignment columns. Unfortunately, strong statistical dependencies are also observed between many pairs of residues that are distal in the structure. Through a comprehensive analysis of 2009 protein domains, we show that a large fraction of these distal dependencies are indirect and result from chains of contacting pairs that percolate through the protein. We present a Bayesian network model that rigorously disentangles direct from indirect dependencies and show that this greatly improves contact prediction. Additionally, we develop an informative prior that takes into account that the probability for residues to be in contact depends on their primary sequence separation, and that highly conserved residues tend to participate in a larger number of contacts. With this prior, the accuracy of the contact predictions is dramatically improved.

structure was available (see *Materials and Methods*) and calculated, for each pair (ij) of columns in each alignment, the statistical dependency using a measure, $\log(R_{ij})$, which is a finite-size corrected version of mutual information (see *Materials and Methods*). Since the distribution of $\log(R)$ values for an alignment depends strongly on the number of sequences in the alignment, their phylogenetic relationship, and the length of the alignment, $\log(R)$ values cannot be directly compared across different alignments. Therefore, we calculated the mean and variance of $\log(R)$ values for each alignment and transformed the $\log(R)$ values to Z-values (number of standard deviations from the mean). Finally, for each alignment, we divided all pairs of residues into those that are contacting in the three-dimensional structure, and those that are distant in the structure, and calculated the distribution of Z-values for these two sets of residue pairs. As in previous work (e.g. [10,20]) and as defined for CASP [21], two residues were considered in contact if their C_β distance (C_α for glycines) in the structure was smaller than 8Å. Combining the data from all alignments, the left panel of Figure 1 shows the fraction of all pairs of contacting residues (red) and distal residues (blue) larger than a given Z-value as a function of Z. The right panel shows, as a function of Z, what fraction of all residue pairs with at least this Z-value are contacting in the structure.

The left panel of Figure 1 illustrates that, indeed, a higher fraction of contacting residues shows strong statistical dependencies than distal residues. However, we also see that the difference in the Z-distribution of close and distal pairs is only moderate. Since there are generally many more distal pairs than close pairs, this implies that, even at high Z-values, the majority of residue-pairs are in fact distal in the structure (Figure 1, right panel). This result shows that simple measures of statistical dependency, such as mutual information, are poor at predicting which pairs of residues are directly contacting in the structure.

The main question is why so many structurally distal pairs show statistical dependencies in their amino-acid distributions that are stronger than those between directly contacting residues. First, whereas measures such as mutual information treat the sequences in the multiple alignments as statistically independent, in reality many of the sequences are phylogenetically closely related, which can cause 'spurious' statistical dependencies to appear between independent residue pairs which can be larger than the true statistical dependencies between contacting pairs. Several groups have investigated this confounding factor in contact prediction and several methods have been proposed for correcting these spurious phylogenetic correlations [8,9,13,14], which we will make use of below.

Although important, many strong statistical dependencies between distal residues remain even when spurious phylogenetic dependencies are corrected for (see below). Some of these distant dependencies have been suggested to be caused by homo-oligomeric interactions [14,22]. Thus, in this interpretation, some of the 'distal' pairs with strong statistical dependencies are in fact contacting in the homo-oligomer. Although it is not clear how many of the distal dependencies can be explained by this mechanism, it seems likely that only a relatively small number of residue pairs on the surface can be responsible for such homo-oligomeric interactions.

A third explanation that has been offered for the large number of distal pairs with strong statistical dependencies is that these

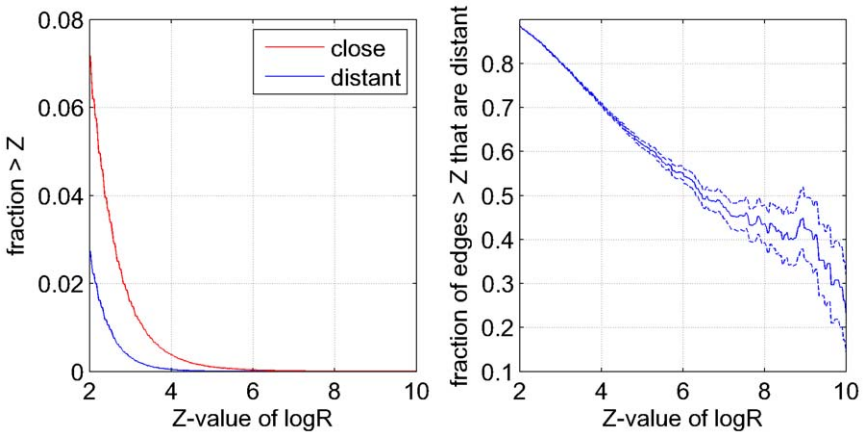

Figure 1. Statistical dependencies of structurally close and distal residue pairs. Left panel: Reverse-cumulative distribution of $\log(R)$ Z-values (horizontal axis) for structurally close (red) and distal (blue) residue pairs. Right panel: The fraction of all residue pairs that are distal in the structure as a function of their statistical dependency (Z-value).
doi:10.1371/journal.pcbi.1000633.g001

dependencies are induced by *indirect* interactions that are mediated either by intermediate molecules [15,23] or by chains of directly interacting residue pairs that run through the protein and connect distal pairs [23–25]. Indeed, for a small number of example domains, the existence of such chains of thermodynamically directly coupled residues has been demonstrated [23,24]. However, the connection between thermodynamic coupling and covariation is still under debate as there is little evidence that thermodynamic coupling of residues is limited to covarying positions [26].

In this paper, we comprehensively investigate to what extent statistical dependencies between distal pairs can be explained by indirect dependencies. The conceptual idea is illustrated in figure 2.

In this illustration, the letters reflect different residues, their distances in the figure reflect their distances in the three dimensional structure, i.e. only the pairs A–B, B–C, and D–E interact directly, and the strength of the statistical dependencies between the different pairs are represented by the thickness of the lines connecting them. Because the pairs A–B and B–C have very high statistical dependency, a strong dependency between A and C is *induced*, which is larger even than the statistical dependency of the directly interacting pair D–E. Any method that considers the statistical dependencies of each pair independently would thus erroneously assign higher confidence to the interaction of A–C than that of D–E.

It should be noted that mutual information and variants thereof have been used extensively for the inference of interacting nucleic acid pairs (see [4] for a review) in the secondary structures of RNA sequences. In these approaches too, the significance of the statistical dependency between a pair of potentially interacting positions is typically evaluated in isolation, i.e. independent of the dependencies between all other pairs. However, in contrast to protein structures, RNA secondary structures per definition consist of *disjoint pairs* of directly interacting residues, i.e. those that form Watson-Crick base pairs. Thus, for RNA secondary structures the 'percolation' of statistical dependencies to pairs that are distal in the structure cannot occur (ignoring tertiary structure).

Below we show that chains of statistically dependent contacts are very common in protein structures, explaining a significant fraction of observed dependencies between structurally distal pairs, and we characterize the distribution of lengths and distance traveled by such chains. We show that a Bayesian network model which we recently developed to predict protein-protein interactions [27] can be adapted to rigorously disentangle direct from indirect statistical dependencies between residues, and we demonstrate that such an approach much improves the prediction of pairs of residues that are in contact in the three-dimensional structure. We then investigate to what extent our Bayesian network algorithm can be further improved by incorporating a correction for the phylogenetic dependencies between sequences in the alignment [14], and by incorporating prior information regarding possible interactions. In particular we develop an informative prior that incorporates the observations that the probability for two residues to interact depends strongly on their distance in the primary sequence, and that highly conserved positions in the multiple alignment tend to interact with a higher number of other residues. We show that incorporating these additional features into our Bayesian network model dramatically improves the accuracy of the predictions.

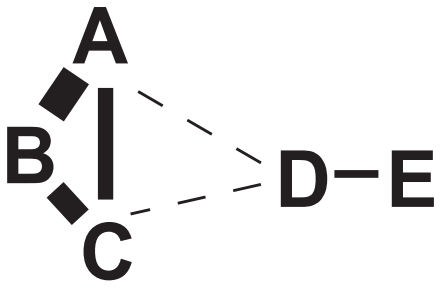

Figure 2. Statistical dependencies between pairs of residues reflect both direct and indirect interactions. The 5 letters (A through E) represent 5 residues and their distances in the figure reflect their distances in the three-dimensional structure. We assume that the pairs A–B, B–C, and D–E are in contact and interact directly. The thickness of the edges between pairs of nodes reflect the statistical dependencies between the corresponding columns in the multiple alignment.
doi:10.1371/journal.pcbi.1000633.g002

Results

Distant co-evolving pairs can frequently be explained by chains of co-evolving contacts

As mentioned above, it has been suggested that statistical dependencies between structurally distant residue pairs can be explained by chains of contacts that are all statistically dependent. However, the existence of such 'co-evolving chains' of contacts has only been demonstrated for a small number of examples [23,24]. To examine comprehensively and systematically to what extent statistical dependencies between structurally distal residues can be explained by co-evolving chains of contacts we extracted, for each multiple alignment, all pairs of residues that showed high statistical dependency ($Z_{ij} > 4$). We then divided these 'co-evolving pairs' into co-evolving contacts and co-evolving distal pairs. As illustrated in Figure 3, we then determined for each distal pair whether there exists a chain of contacts that each show stronger co-evolution than the distal pair, i.e. $Z > Z_{ij}$ for all contacts in the chain.

However, since our Z-values are in all likelihood only a very noisy measure of the true co-evolution of pairs, we expect that frequently one or more of the contacts in the chain may have a lower Z-value, even if their true co-evolution is higher than the co-evolution of pair (ij). We therefore also consider chains where some contacts (kl) have $Z_{kl} < Z_{ij}$ and define the total score $T(C)$ of a chain C as the sum of the difference in Z-value for all edges that have lower Z-value than the distal pair (ij), i.e

$$T(C) = \sum_{(kl)\in C} (Z_{ij} - Z_{kl})\Theta(Z_{ij} - Z_{kl}), \qquad (1)$$

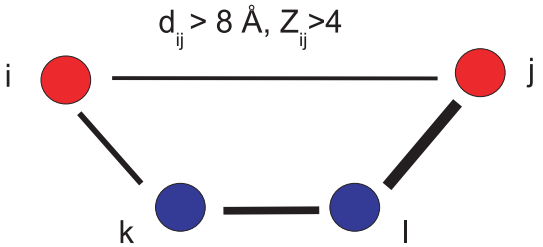

Figure 3. Illustration of a chain that explains the dependency between two distant residues *i* and *j*. The distance between the nodes illustrates the spatial separation and the thickness of the edges represents the strength of the dependence. Nodes i and j can be connected indirectly via a chain of contacts ($d < 8$Å) through nodes k and l (in blue) whose edges all have higher dependency (i.e. $Z_{ik} > Z_{ij}$, $Z_{kl} > Z_{ij}$ and $Z_{lj} > Z_{ij}$).
doi:10.1371/journal.pcbi.1000633.g003

where $\Theta(x)$ is the Heaviside-function which is one when $x \geq 0$ and zero otherwise. For each distal co-evolving pair, we determined the chain of contacts C that has minimal total score $T(C)$. Since pairs that are very distal per definition require longer chains, and since $T(C)$ generally grows with the length of the chain, we define the final score S of the best path for a given pair as the average score per contact, i.e. $S = T/n$, where n is the number of contacts in the best path.

The left panel of Figure 4 shows the cumulative distribution of the scores S of the best chains (blue curve). We see that for 6.5% of the distal co-evolving pairs, there exists a chain with score $S = 0$, i.e. where all contacts in the chain have $Z > Z_{ij}$. The median score of the best contact path is a little larger than $S = 1$, and the 25th and 75 percentiles occur at S-values of about 0.5 and 2 respectively. Note that, as all distal co-evolving pairs have $Z_{ij} > 4$, even at a score of $S = 2$ the contacts in the path have $Z > 2$ on average, meaning that they are still among the most significantly co-evolving pairs.

To assess the significance of the cumulative distribution S we performed a randomization test by randomly permuting the Z-values of all contacts of each domain 100 times and determining the S scores of the best paths that are obtained with these permuted Z-values. The red curve in the left panel of Figure 4 shows the cumulative distribution of S-scores obtained in this randomized set and it is immediately clear that the S-scores are much higher for the randomized set. The right panel of Figure 4 shows, as a value of S, the ratio between the fraction of distal pairs that can be explained by a chain with score less than S for the real and the randomized data. Especially at low values of S the ratios are enormous. For example, at $S = 0.5$ the ratio is about 100, meaning that whereas about 25% of the distal pairs can be explained by chains in the real data, in the randomized data virtually no distal pairs can be explained, i.e. only 0.25%. But strong enrichment persists until much higher values of S. For example, at $S = 1.5$ about two-thirds of distal pairs can be connected by a chain, whereas the percentage is less than 8% for the randomized data.

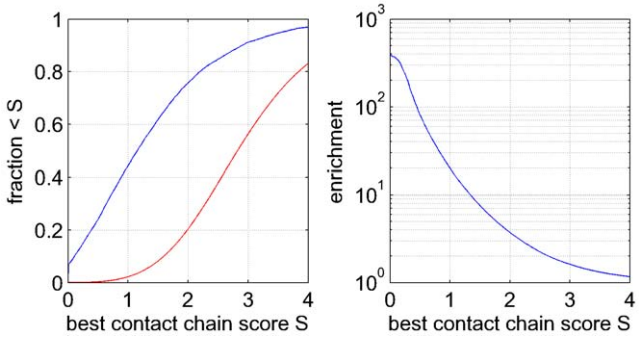

Figure 4. Most distal co-evolving pairs can be explained by chains of co-evolving contacts. Left panel: Cumulative distributions for the number of distal pairs (ij) ($d_{ij} > 8$Å) that co-evolve ($Z_{ij} > 4$) that can be explained by chains of co-evolving contacts as a function of the score S of the best chain (see text). The blue line shows the distribution for the true data and the red curve for the randomized data. Right panel: Ratio (fold-enrichment) of the fraction of distal co-evolving pairs that can be explained by chains versus the fraction that can be explained by chains from the randomized data. The vertical axis is shown on a logarithmic scale.
doi:10.1371/journal.pcbi.1000633.g004

Statistics of co-evolving contact chains

Our results show that, across essentially all protein domains for which multiple alignments and structures are available, chains of co-evolving contacts are common and explain a large fraction of statistical dependencies observed between structurally distal pairs. To gain insights in the nature of these co-evolving contact chains in protein structures, we selected all distal pairs that are explained by contact chains with scores $S < 1.5$ and obtained statistics on the number of steps and the spatial distance covered by these chains (Figure 5).

We see that the distance distribution of 'explainable' distal co-evolving pairs is roughly exponential with a length scale of about 8 Å. Since 'distal pairs' are by definition at least 8Å apart, this means that the typical length scale covered by co-evolving contact chains is about 16Å. The right panel of Figure 5 shows the mean number of steps in the shortest co-evolving contact chain as a function of the structural distance of the co-evolving distal pair. With increasing spatial separation, the number of edges in the chain steadily increases from on average 2 steps at a separation of 8Å to 15 steps at 50Å. Interestingly, the increase in the average number of steps as a function of distance is almost perfectly linear and corresponds to 3.25 ± 0.05Å per step. We thus see that 'typical' co-evolving contact paths contain about $16/3.25 \approx 5$ steps, demonstrating that statistical dependencies typically percolate along paths with multiple steps. We also note that some chains are very long, consisting of up to 20 steps, connecting residues that are as far as 60Å apart in the structure.

Bayesian network model

The insight that many of the statistical dependencies between structurally distal pairs result from chains of co-evolving contacts has important consequences for contact prediction methods. That is, any method that aims to predict contacting residues from statistical dependencies should clearly take into account indirect dependencies that are induced by such chains.

In [27] we developed a general Bayesian network model for calculating the probability of a multiple alignment of protein sequences taking into account dependencies between amino acids at all possible pairs of positions. We refer the reader to [27] for a comprehensive explanation of the method. Briefly, our model assumes that the sequences in a multiple alignment $\boldsymbol{D}$ (the data) are drawn from an (unknown) underlying joint probability distribution $P(x_1, x_2, \ldots, x_l)$ with l the width of the alignment and x_i the amino acid at position i. Profile hidden Markov models typically assume that the amino acids at different positions are independent so that one can write $P(x_1, x_2, \ldots, x_l) = \Pi_{i=1}^{l} P_i(x_i)$, with $P_i(x)$ the probability distribution of amino acids at position i. Note that, since there are 20 amino acids (disregarding gaps), such models will have $19 \times l$ parameters in total. Our model of $P(x_1, \ldots, x_l)$ allows general dependencies, such that the probability for an amino acid at position i depends on the amino acids at other positions. Note that, if the residue at i is dependent on a residue at one single other position j, there are already $20 * 19 = 380$ parameters in the distribution $P(x_i|x_j)$, and that models with dependencies on two other positions, i.e. $P(x_i|x_j, x_k)$, would have 7600 parameters for each residue. Given the current amount of sequence data, it is certainly reasonable to consider models with single dependencies, but there is hardly ever enough data to meaningfully estimate 7600 parameters per position. Our model therefore only considers pairwise conditional dependencies of the form $P(x_i|x_j)$.

Any model that considers only pairwise conditional dependencies factorizes the joint probability $P(x_1, \ldots, x_l)$ as a product $P(x_1, \ldots, x_l) = \Pi_{i=1}^{l} P(x_i|x_{\pi(i)})$, where $\pi(i)$ is the single other

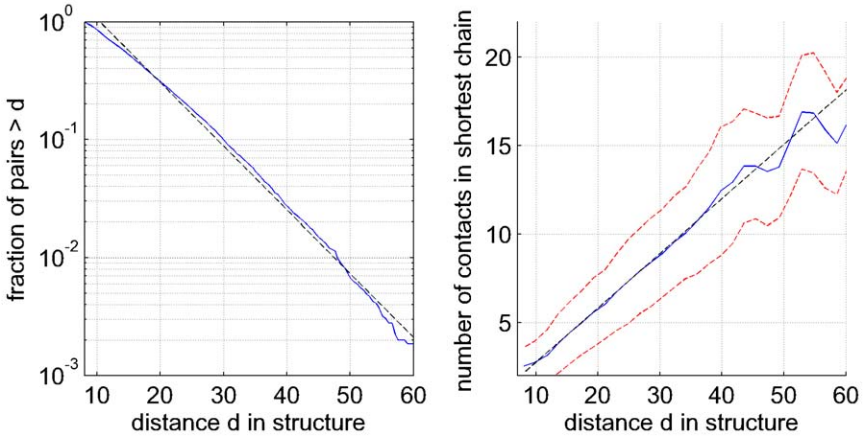

Figure 5. Statistics of co-evolving contact chains. Left panel: Reverse-cumulative distribution of the spatial distances between co-evolving pairs that can be explained by chains of co-evolving contacts of score $S<1.5$. The vertical axis is shown on a logarithmic scale. The dotted line shows a fit to an exponential distribution $P(d>x)\propto e^{-x/8}$. Right panel: Number of steps in the shortest co-evolving contact chain as a function of the spatial distance of the co-evolving pair. The blue line shows the mean distance and the red dotted lines show mean plus and minus one standard deviation. The black dotted line shows a linear fit, the fitted slope of which corresponds to an increase in distance by 3.25 ± 0.05Å per additional contact in the chain.
doi:10.1371/journal.pcbi.1000633.g005

position which the residue at position i depends on (note that independence, i.e. $P(x_i|x_{\pi(i)})=P(x_i)$ is contained in this general model). Our Bayesian network model is the most general model of this form. In particular, we do not attempt to estimate the conditional probabilities $P(x_i|x_j)$ but rather treat these conditional probabilities as nuisance parameters that we integrate out in calculating the likelihood of the alignment. In addition, and importantly, we do not consider only a single 'best' way of choosing which other position $\pi(i)$ each position i depends on, but rather we *sum* over all ways in which the dependencies can be chosen. Note that if we consider each column of the alignment as a node in a graph and connect each node i to the node it depends on, $\pi(i)$, then any consistent set of dependencies π, i.e. any set of dependencies π that does not introduce cycles in the graph, corresponds to a *spanning tree* of this graph. Thus, the sum over all consistent ways in which we can assign dependencies is in fact the sum over the set of all possible spanning trees of our graph. As explained in [27] and the *Materials and Methods* section, all integrals over the unknown conditional probabilities $P(x_i|x_j)$ can be performed analytically and, importantly, the sum over all spanning trees can be calculated as a matrix determinant using a generalization of Kirchhoff's theorem [28]. It is thus feasible to do inference with this general Bayesian network for a large number of multiple alignments, including alignments that are hundreds of columns wide.

Posterior probability of a pairwise interaction

In our model the joint probability of a multiple alignment is given as the sum over all possible spanning trees of node-dependencies, where each spanning tree is weighted according to the product of statistical dependencies across all edges in the tree (see *Materials and Methods*). Here the statistical dependence between any pair of positions (ij) is given by the ratio $R_{ij}=P(D_{ij})/[P(D_i)P(D_j)]$ of the joint probability of the alignment columns $P(D_{ij})$ and the product $P(D_i)P(D_j)$ of their marginal probabilities. Since the number of edges in any spanning tree is limited, there is a natural 'competition' in this model between the edges to be included in the spanning tree. Therefore, spanning trees with the highest statistical weight will only use edges whose statistical dependence can *not* be explained by chains of other edges with higher dependency, and edges between pairs with indirect statistical dependency will thus only appear in spanning trees with relatively low statistical weight. The posterior probability $P((ij)|D)$, given the data D, for a pair (ij) to interact directly can thus very naturally be quantified within our model by calculating the sum of the statistical weights of all spanning trees in which the edge between the pair (ij) exists. The calculation of this posterior is illustrated in Figure 6.

Note that in this calculation $P((ij)|D)$ depends on the statistical dependencies between all pairs of positions and that all possible spanning trees are included in the calculation. Roughly speaking, a high posterior $P((ij)|D)$ indicates that the edge (i,j) is included in most spanning trees that have high probability. In this way indirect dependencies are accounted for in a rigorous way, derived from first principles, and without any free parameters.

Posterior probabilities significantly improve contact predictions

To compare the performance of the traditional mutual information-based measurement with the predictions of our model, we calculated mutual information I_{ij}, our analogous measure $\log(R_{ij})$, as well as the posterior probabilities $P((ij)|D)$ for each pair of positions (ij) for each domain in our set of 2009 Pfam alignments with available three dimensional structure.

Different domains have widely varying widths and also widely varying numbers of sequences in the alignments. With regard to the former, it is well-known that the number of pairs that are in contact in three-dimensional protein structures increases with the length of the protein sequence. To compare prediction accuracies for proteins with different lengths, the consensus, also used by the CASP assessors [21], has been to compare the number of predictions per residue. However, although there is a large variation across domains, we find that the number of contacts scales slightly super-linearly, with an exponent of roughly 1.1 for all pairs of residues, and up to 1.6 if we consider only pairs of residues that are distal in the primary sequence (see Figure S1). That is, the number of contacts per residue grows with the length of the domain, making it problematic to use predictions-per-residue as a common reference for domains of different length. We therefore decided to compare prediction accuracies as a function of the number of predictions

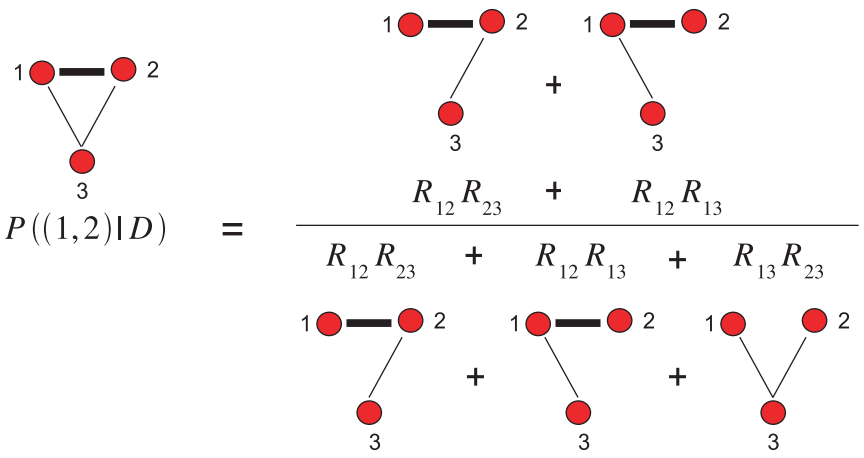

Figure 6. Illustration of the calculation of the posterior probability. For the sake of simplicity, we here show an example for an alignment with only 3 columns. The posterior probability for edge (1,2) is the statistical weight of all spanning trees that contain this edge relative to the weight of all possible spanning trees.
doi:10.1371/journal.pcbi.1000633.g006

relative to the total number of contacts in the protein. In particular, we compare predictions for different proteins at the same *sensitivity*, i.e. the fraction of all true contacts that are predicted.

As mentioned previously, $\log(R)$ values typically increase with the number of sequences in the alignment and also depend on the phylogenetic distances of the sequences present in the alignment, such that $\log(R)$ values cannot be directly compared across different domains. Therefore, for each domain we produced three lists of predicted edges, one sorted by mutual information, one by $\log(R)$, and one by posterior probability $P((ij)|D)$. For different fractions x, we selected the top edges from each list such that the fraction of all true edges among the predictions (sensitivity) equals x, separately for each domain. For each value of x and all three measures, we then calculated the average positive predictive value, i.e. the fraction of all predicted edges that are truly in contact in the three-dimensional structure of the domain, by averaging over all domains. These results are shown in the left panel of Figure 7.

Not surprisingly, residues that are close in the primary sequence are much more likely to contact each other in the structure than distant pairs, see [20] and figure 11 below. In particular, residues that are neighbors in the primary sequence are (by the definition used) *always* contacts and residues at distance 2 are contacting almost 90% of the time, whereas contacts between residues more distal in the primary sequence are relatively rare. Therefore, if one considers all contacts, the accuracy of the predictions is dominated by the large number of contacts between residues at primary sequence distances 1 and 2, which almost always exist, and are therefore not informative regarding protein structure. Therefore, the middle panel of Figure 7 shows the results when considering only pairs that are at least 3 residues apart in primary sequence. In addition, following the practice established in the contact prediction literature, we also show results when considering only pairs at least 12 residues apart in primary sequence (Figure 7, right panel) and at least 24 residues apart (Figure S2).

As expected, the accuracy of predictions for mutual information and $\log(R)$ are very similar and demonstrate that these two measures can be considered equivalent in this context (we will only refer to $\log(R)$ from hereon). Most importantly, Figure 7 shows that the predictions based on posterior probabilities (red curves) outperform the other methods by a large margin, i.e. with an almost 50% larger PPV at some sensitivities. This confirms that rigorous treatment of indirect dependencies strongly improves contact predictions. It should be noted, however, that at cut-offs where the positive predictive value is reasonably high, sensitivities are only on the order of one percent. It is thus clear that at high PPV, our method in its current form can only predict a minor fraction of all true interacting pairs, which is in accordance with results from previous studies [10,14].

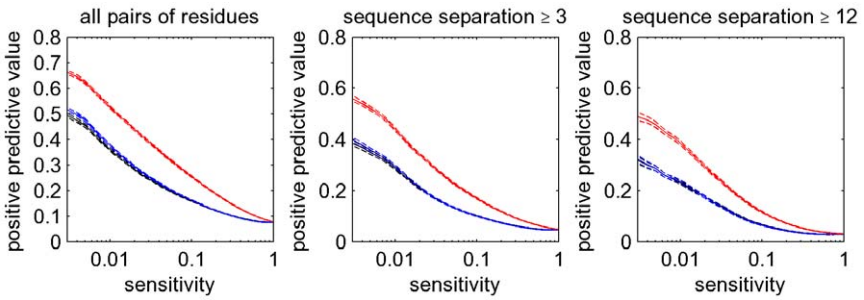

Figure 7. Accuracy of contact predictions for all 2009 alignments. Shown are the performances of mutual information (black), $\log(R)$ (blue), and the posterior probabilities (red). The vertical axis shows mean positive predictive value (PPV, solid line) plus and minus one standard error (dashed lines) as a function of sensitivity (horizontal axis, shown on a logarithmic scale). The left panel shows predictions for all residue pairs, the middle using only predictions for residues separated by at least 3 positions in the primary sequence, and the right panel for pairs separated by at least 12 positions.
doi:10.1371/journal.pcbi.1000633.g007

For completeness, we also considered the accuracy of prediction that would be obtained if, instead of summing over all possible spanning trees, we determine the maximum-likelihood tree and use only the links in this tree in our predictions, i.e. as done in [15]. As shown in Figure S3, although this leads to an improvement over using $\log(R)$, the accuracy of the posterior probability measure by far outperforms the predictions based on the maximum-likelihood tree. This nicely demonstrates the value of summing over all possible spanning trees which is employed in the calculation of the posterior for a given edge.

The posterior removes indirect dependencies and predicts contacts with weaker statistical dependency

To demonstrate that our model successfully prevents the prediction of interactions between pairs with indirect dependency, we collected all distal pairs that showed significant statistical dependence ($Z>4$) and ordered them by the score of the best co-evolving contact chain that can explain their statistical dependency, i.e. as shown in Figure 4. Figure 8 shows the reverse-cumulative distributions of the posteriors that these distal pairs obtain in our model for different cut-offs on the best path score S, as well as the distribution of posteriors of all contacting pairs with $Z>4$.

First of all, we see that co-evolving contacts have dramatically higher posteriors than distal pairs in general, which confirms the improved accuracy of contact predictions that our method accomplishes. Moreover, we see that distal pairs that can be explained with the most strongly co-evolving contact chains, i.e. with the lowest scores S, obtain the lowest posterior probabilities. For example, less than 10% of the distal pairs with a chain at score $S=0$ have a posterior larger than 0.2 and virtually no pair has a posterior as large as 0.5. As the score S of the best chains increases, so generally do the posteriors. This confirms that the posterior as calculated by our model correctly captures the extent to which a statistical dependency is direct.

Instead of selecting all distal co-evolving pairs with contact chains below some score S, we also selected all co-evolving pairs with S scores larger than various cut-offs and determined the distributions of their posteriors. These distributions are shown in Figure S4 and illustrate that distal co-evolving pairs with sufficiently large score S obtain posteriors comparable with those of co-evolving contacts. This suggests that the particular subset of distal co-evolving pairs that cannot be explained by any chain of contacts are likely true interacting residues, which may for example form contacts in the interaction surface of oligomers of the domain.

To further demonstrate that our Bayesian network model correctly distinguishes direct from indirect interactions, we also investigated the extent to which the posterior identifies structurally close pairs independent of the direct statistical dependency of the pair. We divided all pairs into bins according to their $\log(R)$ Z-value and calculated, for each bin, the distribution of structural distances of all pairs, and for the subset of pairs that have posterior probability larger than 0.2. Figure 9 shows, as a function of the Z-value of the pairs, the median, 25th, and 75th percentiles of the structural distance distributions of all pairs (blue) and those with posterior larger than 0.2 (red).

At large Z-values the red and blue curves are essentially identical. In this regime, we are only looking at the most strongly dependent residues in each alignment and any spanning tree of high likelihood must contain edges between these pairs of residues, i.e. almost all of these edges have high posterior probabilities. However, already at Z-values as high as 8, the median distance of all pairs starts to increase rapidly, from roughly 8Å to more than 20Å at Z-value 0. This illustrates again that even at very high values of $\log(R)$ a substantial fraction of pairs are distal in the structure. In contrast, the subset of residues with high posterior probability remains close over the whole range of Z-values, down to Z-values of almost 0. In fact, strikingly, there is very little change in the distribution of structural distances for Z-values from 0 to 8. This is very significant because it demonstrates that, independent of the amount of direct statistical dependency between a pair of positions, a high posterior is indicative of close structural distance. Moreover, it demonstrates that our Bayesian network model can detect truly interacting pairs of residues even if they show only a small amount of statistical dependency.

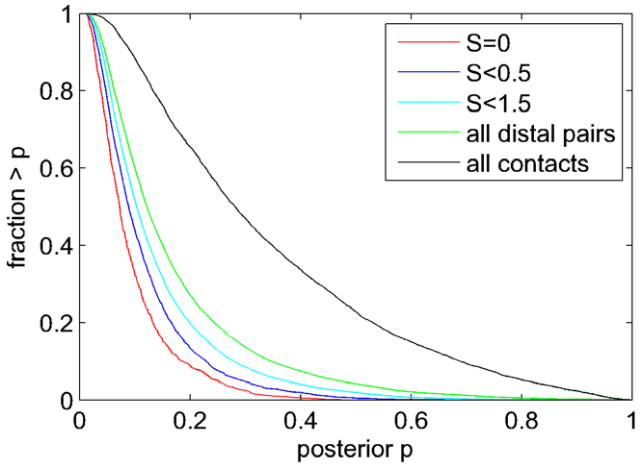

Figure 8. Posteriors reflect the extent to which co-evolving pairs can be explained by contact chains. Shown are the reverse cumulative distributions of the posteriors of distal co-evolving pairs ($Z>4$) that can be explained by contact chains of scores $S=0$ (red), $S<0.5$ (dark blue), $S<1.5$ (light blue), and for all distal co-evolving pairs (green). For comparison the reverse cumulative distribution of posteriors for co-evolving contacts ($Z>4$) is also shown (black).
doi:10.1371/journal.pcbi.1000633.g008

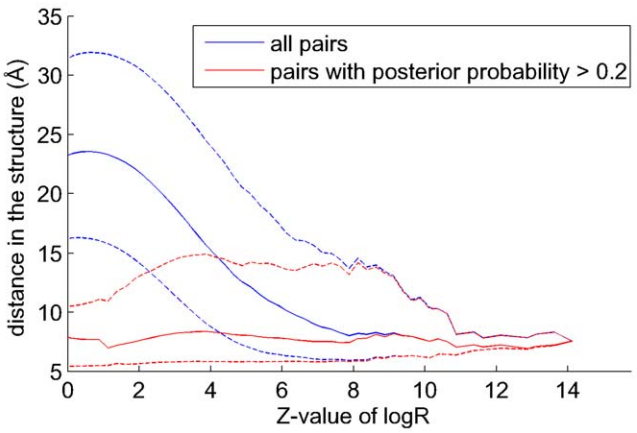

Figure 9. The posterior predicts structurally close pairs independent of their direct statistical dependence. The structural distance distribution (vertical axis) is shown for all pairs (blue) and for pairs with posterior probability larger than 0.2 (red) as a function of the Z-value of the $\log(R)$ statistic (horizontal axis). The solid lines show the medians of the distributions and the dashed lines the 25th and 75th percentiles.
doi:10.1371/journal.pcbi.1000633.g009

The Bayesian network model with phylogenetic correction significantly outperforms existing methods

One of the key problems in contact prediction is the large number of distal pairs with high statistical dependency. In the foregoing sections we have shown that many of these distal co-evolving pairs are indirect, induced by chains of dependencies between contacting residues, and we have shown that our Bayesian network model can rigorously disentangle direct from indirect dependencies, thereby greatly improving contact predictions. In the remaining sections we develop a number of extensions of our basic method to further improve the predictions.

As mentioned in the introduction, the phylogenetic relationships of the underlying sequences is a major confounding factor when determining the statistical dependency between several residues (nicely explained in eg [9,13]) and it is a difficult task to 'subtract' from the apparent statistical dependency between two residues the part that is purely due to phylogeny. The best way to address this difficulty would of course be to construct a phylogenetic tree of all sequences in the multiple alignment and to explicitly model the evolution of the sequences along the tree, using an evolutionary model that takes dependencies between positions into account. Unfortunately, it appears that such a rigorous approach is computationally intractable for several reasons. First, one would either have to accurately reconstruct the phylogenetic tree, which is very challenging for large sets of sequences, or sum over all possible trees, which is computationally infeasible. The second issue is the evolutionary model. In our Bayesian network model, the conditional probabilities $P(x_i|x_j)$ are different at every pair (ij), introducing 380 parameters per pair, which are integrated over. However, for the evolutionary case analytic integration is no longer possible, which makes such models intractable. Indeed, models that treat dependencies between residues in an explicit phylogenetic setting [12,15] consider much simpler evolutionary models in which only correlations in the overall *rates* of mutations at different positions are considered and not the specific identities of the mutations.

As an alternative to explicit phylogenetic methods, recently a number of simple *ad hoc* phylogenetic corrections have been proposed, which do not involve a reconstruction of the phylogenetic tree, which can be efficiently calculated, and which clearly improve contact predictions [13,14]. One of these corrections, the so-called *average-product correction* APC has been shown to provide the most accurate contact predictions [14]. It is based on the idea that the statistical dependency between every pair of columns is the sum of a true statistical dependency and a background dependency due to the phylogenetic relationships. In the APC it is assumed that the background dependency is a product of independent factors associated with the two positions. Since a given position will interact with only a small fraction of other positions, the background dependencies can be estimated by calculating, for each column, its average statistical dependence with all other columns. The background dependence for each pair is then subtracted to obtain a corrected statistical dependency. As described in *Materials and Methods*, we adapted the APC to our Bayesian model, essentially replacing $\log(R)$ with a corrected version $\log(R^c)$ that subtracts out the background dependency. These $\log(R^c)$ values can then be used, analogously to $\log(R)$ values, to determine corrected posterior probabilities (see *Materials and Methods*).

In figure 10, we show the accuracy of our predictions using the corrected posterior probabilities (in blue) and compare it with predictions based on mutual information using the average-product correction APC (in black). The latter has been recently shown to outperform other existing methods [14]. The red curves show the performance of the method without the phylogenetic correction, i.e. as was shown in Figure 7. It is clear that the predictions based on posterior probability combined with the phylogenetic correction significantly outperform the current best methods. For example, considering pairs at primary sequence separation at least 3, the sensitivities at PPV of 0.5 are 0.5% for the uncorrected posterior, about 1% for the APC, and about 2% for the corrected posterior. The clear improvement in prediction accuracy is also evident for pairs with primary sequence separation of at least 24 amino acids (Figure S5).

Although Figure 10 combines results of the predictions on protein domains of differing sizes, the fact that the true interactions are a much smaller fraction of all possible interactions for long sequences makes the prediction task significantly harder for long sequences, see e.g. [29]. In Figures S6, S7, S8, and S9, we show the performance of the various methods separately for short, medium length, and long sequences. We find that, independent of the length of the sequences, our method clearly outperforms current methods.

Co-evolution of residue pairs is independent of primary sequence separation

In protein structure prediction, where prediction of contacts at large sequence separations is particularly important [21], it is well-known that contact prediction accuracy generally decreases with increasing sequence separation ([20,21], also seen in figure 10).

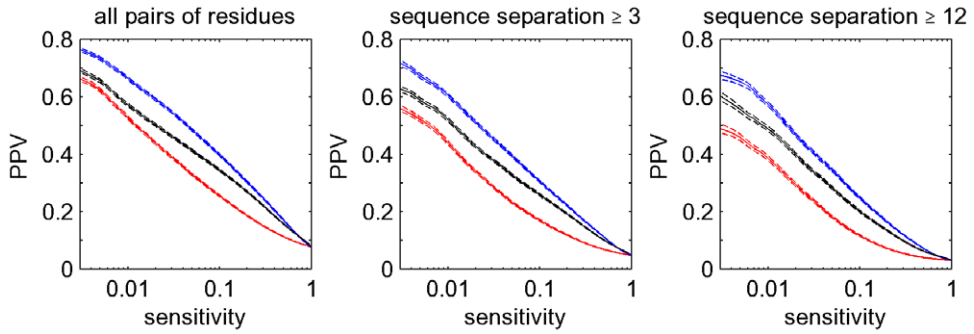

Figure 10. Improved accuracy of contact predictions when a phylogenetic correction is included. In blue, we show the performance of the phylogenetically-corrected posterior probabilities, in black the performance of the predictions based on the average-product corrected (APC) mutual information [14], and in red the performance of the posterior probability without phylogenetic correction. Curves were calculated as in figure 7.
doi:10.1371/journal.pcbi.1000633.g010

This is a direct consequence of the fact that the fraction of contacts decreases rapidly as a function of sequence separation (roughly as $1/d$, where d is the primary sequence separation, see the left panel in figure 11), which makes the prediction problem much more difficult for contacts at large primary sequence separations. Vice versa, because contacts at large primary distances are rare, they are most informative for protein structure prediction [21].

The left panel of Figure 11 shows that there are several regimes in the distribution of contact-density at different primary sequence distances. First, residues at distance 1 and 2 are almost always contacts and thus contain very little information about protein structure. In contrast, at distances 3 and 4 the fraction of contacts has already dropped to roughly 50%, i.e. about 1 bit of information per contact, and the fraction then drops quickly, reaching about 5% at primary sequence separation 10. For distances between 10 and 30 the fraction stays roughly constant at 5% and for even larger distances it drops approximately as $1/d$.

Clearly, the information contained in Figure 11 regarding protein structures can be used to improve contact prediction, i.e. by assigning prior probabilities to different contacts based on their distance in primary sequence. However, before pursuing this we ask to what extent contacts at different primary sequence distances show statistical evidence of co-evolution. The almost ubiquitous contacts at primary sequence distances 1 and 2 are probably mainly the result of geometrical constraints, the contacts at intermediate distances are likely often part of the same secondary structure, and the very distal contacts might correspond to contacts between different secondary structure elements. Given the different nature of these contacts at different primary sequence separations, one might expect very different distributions of statistical dependencies, and this would clearly affect contact prediction.

To investigate this, we determined the distribution of the Z-values of corrected $\log(R^c)$ for all *contacts* at each primary sequence separation d (Figure 11, right panel). Interestingly, the distribution of statistical dependencies is almost *constant* across the entire range of primary sequence distances. The only significant deviation is a slight peak at sequence separation 4, corresponding to residues on the same side of alpha helices ([30] and data not shown), which apparently have slightly increased statistical dependency compared to other contacts. However, far more important for the purpose of predicting protein structure is that, with regard to the statistical dependency between alignment columns, all contacts appear to be essentially equal, so that the evidence of statistical dependency between residues can be treated completely independently of the prior information regarding which contacts are more or less likely to exist based on general structural considerations. From a biological and evolutionary perspective this result shows that, interestingly, different 'types' of contacts apparently lead to similar evolutionary constraints.

Influence of entropy on contact prediction

An important, but poorly understood issue in covariation-based contact prediction is the influence of conservation on prediction accuracy. The 'conservation' shown by a position in a multiple alignment can be most generally quantified by the entropy of the amino acid distribution in the column. It is well known that this column entropy can vary immensely along protein sequences, most probably due to functional and structural constraints. One would intuitively expect that a position that is contacting many other residues would generally have to satisfy more constraints and would thus be expected to show relatively low entropy.

To investigate this, we calculated, for each position in each domain, the column entropy and the number of contacts of the corresponding residue. As shown in the left panel of Figure 12 there is indeed a clear negative correlation between the column entropy and the number of contacts. For very low entropies, i.e. less than 1, the average number of contacts is constant and approximately 10.5. As the entropy increases from 1 to about 2.75 (which is close to the entropy of a uniform distribution of amino acids) the average number of contacts drops to almost 6. That is, very low entropy columns have on average almost twice as many contacts as high entropy columns. Since the number of residues in a sphere of 8Å around the C_β atom of an amino acid (which is exactly our definition of a contact) is commonly used as a measure for how strongly a residue is buried in the core of the protein (e.g. [31]), the left panel of Figure 12 reiterates the well-known dependence between surface accessibility and conservation [32].

It is well appreciated in the literature that the variation of entropy across positions has important effects on predictions based on statistical dependencies. For example, a comparative study of different prediction methods has shown that commonly used co-

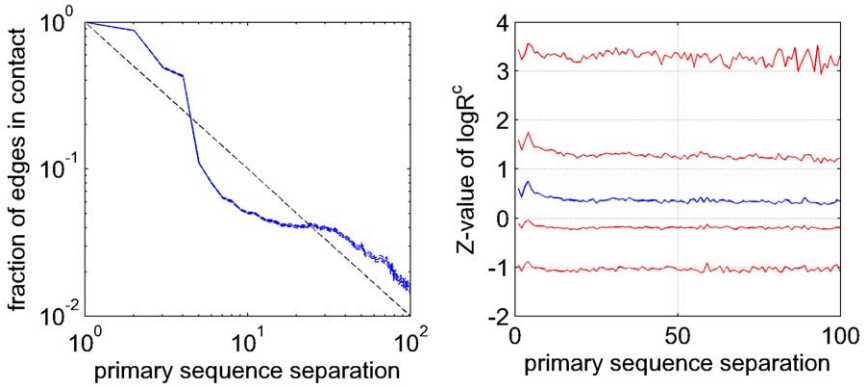

Figure 11. Occurrence of contacts and co-evolution as a function of primary sequence separation. Left panel: The fraction of residue pairs that are in contact in the structure as a function of primary sequence separation d. The solid blue line shows the mean, the dashed blue lines the mean $\pm$ one standard error. The dashed black line shows the function $1/d$. Right panel: The Z-value distribution of the $\log(R)$ statistics for all contacting pairs at different primary sequence separations. The blue line represents the median and the red lines represent the 5th, 25th, 75th and 95th percentiles, respectively. The Z-value was calculated with respect to the mean and standard deviation of the $\log(R)$ distribution of all pairs (including distal ones). In both panels only sequence separations up to 100 residues are shown as the curves become very noisy for larger sequence separations.
doi:10.1371/journal.pcbi.1000633.g011

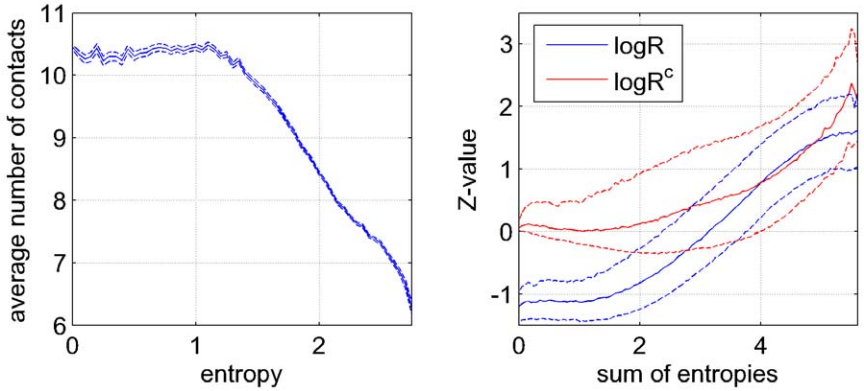

Figure 12. Contact-degree and co-evolution as a function of positional entropy. Left panel: Average number of contacts of a residue (solid line) as a function of the entropy of its alignment column. The dashed lines denote mean $\pm$ one standard error. The right panel shows the Z-value distribution of both $\log(R)$ (blue) and $\log(R^c)$ (red) for all contacting pairs versus the sum of entropies of the corresponding columns. The solid lines denote the medians and the dashed lines the 25th and 75th percentiles.
doi:10.1371/journal.pcbi.1000633.g012

variation measures differ in their sensitivity to per-site variability and generally, each method has highest accuracy within its specific preferred range of variability [10]. In analogy to our analysis of statistical dependency as a function of distance in primary sequence (Figure 11, right panel), we investigated how the statistical dependency that different contacts exhibit depends on the column entropies of the residues. As before, we transformed the $\log(R)$ values to Z-values and determined the Z-value distribution of all contacts as a function of the sum of the entropies of the corresponding columns (Figure 12, blue lines). We see that contacts indeed show a strong correlation between the sum of column entropies and statistical dependency. For low entropy columns the Z-values are mostly negative, and they become only positive at an entropy sum of about 3. It is thus clear that contact predictions that use mutual information ($\log(R)$) will preferentially predict contacts between residues of high entropy columns.

That mutual information and $\log(R)$ is low for contacts with low entropy columns is to a certain extent unavoidable. It is a basic result of information theory [17] that the mutual information between two variables cannot be larger than the minimum of the marginal entropies of the two variables. Intuitively, one could imagine a position that is so constrained by its function and its many contacts that only a single amino acid is viable at the position. Obviously, since this position shows no variation whatsoever it cannot display any signs of statistical dependency with any other column, even though it may contact many other residues. This is a basic limitation of using statistical dependency for contact prediction that cannot be avoided. However, it has been argued that modified versions of mutual information, such as the product or sum correction [14], besides correcting for the phylogenetic background signal, are also able to better identify co-evolution between less variable residues. The red lines in the right panel of Figure 12 show the mean and standard deviation of the Z-values of product-corrected statistical dependency $\log(R^c)$. We see that indeed, the correlation between the Z-values and the sum of column-entropies is significantly reduced when using $\log(R^c)$, and low entropy contacts no longer show negative Z-values on average.

Still, a clear correlation between the column-entropy sum and the statistical dependency remains even for $\log(R^c)$. On the one hand this may be the result of the inherent inability to 'detect' statistical dependency when columns are very conserved. On the other hand, it is also conceivable that those positions that have low entropy, and that form many contacts, may generally show weaker statistical dependency *per contact*. For example, it could be argued that hydrophobic residues that lie in the core of the protein and thus contact many other residues are less variable because they need to remain on the interior and therefore do not allow for changes towards non-hydrophobic residues. Such residues may not be constrained so much by their contacting residues, but rather by the necessity to stay away from the solvent-exposed protein surface, leading to relatively weak statistical dependencies with the contacting residues.

Incorporation of prior information improves prediction accuracy

So far our Bayesian method assumes that a contact between any pair of positions is a priori equally likely. However, as seen in the previous sections, the probability for a contact to occur depends strongly on the primary sequence distance between the residues and the column-entropies of the residues. We therefore developed an 'informative prior' which makes the prior probability for a contact to occur depend on both of these variables. For a given pair of positions, let d be the distance in the primary-sequence of the two positions, and let H denote the sum of the column-entropies of these positions. As described in *Materials and Methods*, we estimated the fractions $f(d,H)$ of pairs at sequence distance d and entropy-sum H that are contacts and using these fractions constructed prior probability distributions that can be easily incorporated into our method.

Figure 13 shows the results of the contact predictions performed with our Bayesian network model incorporating the informative prior and using posterior probabilities (blue lines). For comparison the results using posteriors based on $\log(R^c)$ (the blue lines in Figure 10) are shown as well (red lines). We see that, for the set of all pairs, and all pairs that are at least $d \geq 3$ apart in primary sequence, the incorporation of the prior probability dramatically improves the predictions. For example, looking at all pairs, our method can predict roughly 40% of all existing contacts at a positive predictive value of 80%. If we restrict ourselves to non-trivial contacts, i.e. those with primary-sequence distance $d \geq 3$, we find that at a positive predictive value of 50% our method reaches a sensitivity of roughly 20%. For comparison, without the prior an approximately 10 times lower sensitivity is reached at the same positive predictive value.

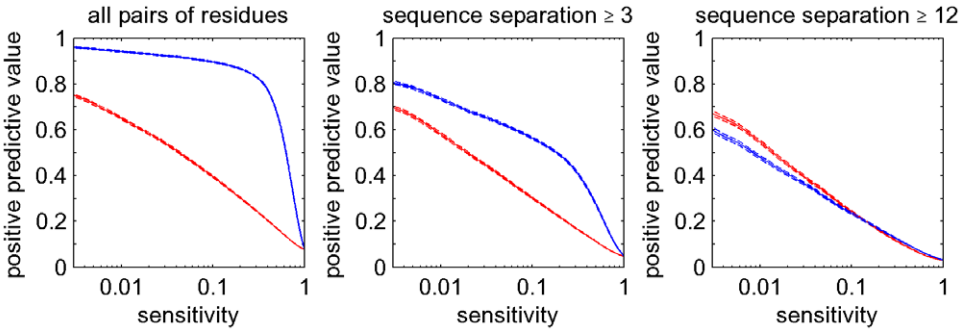

Figure 13. Improved accuracy of contact prediction when an informative prior is included. In blue, we show the performance of the posterior probabilities that take primary-sequence separation and column entropy into account. For comparison we show in red the performance of the posteriors with phylogenetic correction but uniform prior, which are the same as the blue lines in Figure 10.
doi:10.1371/journal.pcbi.1000633.g013

Somewhat surprisingly, we find that the quality of the predictions for distal pairs $d \geq 12$ is slightly reduced by the incorporation of the prior, especially at low sensitivities. We speculate that this is a result of the fact that we constructed the prior distribution assuming that $f(d,H)$ is independent of the length of the domain itself. This approximation breaks down most significantly when focusing on distal pairs because, whereas contacts at short primary distances occur in all domains, contacts at long primary distances are more common in long domains. However, it should be noted that, given that contacts at this primary-sequence distance are rare, one would most likely need to perform predictions at reasonably high sensitivity, i.e. 10% or more. In this regime, the performance with prior is comparable to or even a tiny bit better than without prior.

Discussion

One of the key problems in using co-evolution analysis to predict residue contacts is that so many structurally distal pairs show strong statistical dependencies [14–16]. A number of reasons have been proposed to explain this fact. One explanation is that sequences in multiple alignments are generally phylogenetically related and these phylogenetic relationships can induce strong apparent statistical dependencies between many pairs of columns. Although there is of yet no computationally tractable way for treating the phylogenetic dependencies in a rigorous manner, i.e. by explicitly modeling the evolution of the sequences including arbitrary dependencies, several procedures have been proposed that can correct at least for the main phylogenetic signal [8,9,14,15]. Indeed the application of such methods has been shown to very significantly improve contact predictions [9,14,15].

Still, even with the current best phylogenetic corrections, strong statistical dependencies remain evident between many structurally distal pairs. One proposed explanation that has received little attention in the contact prediction literature is that statistical dependencies between distal pairs can be induced by the percolation of statistical dependencies along chains of co-evolving contacts [23,24]. Here we have shown that such chains of co-evolving contacts are indeed pervasive across all protein domains and that they explain many if not most of the distal co-evolving pairs. Statistical analysis shows that these chains travel on average 3.25 ± 0.05Å per contact, and that the total distance covered by these chains is exponentially distributed with an average of 16Å, corresponding to a chain that consists of 5 contacts. Note that, whereas residues up to 8Å apart are generally considered contacts, our results strongly suggest that the typical distance between co-evolving contacts is only 3.25Å. Another interesting observation is that, although it is likely that contacts between residues at different distances in primary sequences are different in nature, our analysis shows that the statistical dependency shown by contacts is completely independent of their primary-sequence separation. This is an important insight because it demonstrates that co-evolutionary analysis is equally informative about close and distal contacts.

We have adapted our recently evolved Bayesian network model [27] in order to assign, to any pair of positions, a posterior probability that they interact directly. This posterior probability rigorously takes into account all possible ways in which the statistical dependence between the pair can be explained in terms of chains of other co-evolving pairs. Analysis of the predictions of this model shows that it correctly detects distal pairs that can be explained by co-evolving contact chains, and that it also allows one to detect true interacting pairs that have only weak direct statistical dependency.

Recently Halabi et al [33] have shown that, by a spectral analysis of the matrix of statistical dependencies between positions, one can identify so called 'protein sectors': sets of positions that co-evolve significantly with each other, but that are relatively independent of the positions in other sectors. Since in [33] a rather simple measure of direct statistical dependency is used, we speculate that a much more accurate identification of protein sectors could be obtained by using statistical dependencies as assessed by our posterior probabilities.

While finishing the work in this study, a paper appeared that also aims to disentangle direct from indirect interactions [22]. Like our approach, [22] models the joint probability of sequences in the multiple alignment in terms of a set of pairwise interactions. What is appealing about the approach of [22] is that it is based on the more 'physical' assumption that an interaction energy is associated with each pairwise interaction such that a total interaction energy can be calculated for each sequence, and that the probability to observe a particular sequence is given simply by the Boltzmann distribution in terms of this total energy. However, the great disadvantage of this model is that its solution requires a heuristic approximation and is computationally very expensive to calculate. For example, in [22] the authors were forced to restrict themselves to only 60 positions in the alignment, and even then the calculations for a single alignment took several days. Therefore, an application of the approach of [22] on as large a scale as in this work, with thousands of multiple alignments of up to several hundred positions, is not feasible. In addition, it is not clear how the approach of [22] could accommodate a phylogenetic correction, which would be necessary to obtain a competitive performance with this method.

Although the disentangling of direct and indirect statistical dependencies strongly improves contact predictions, and incorporating a phylogenetic correction further improves the performance, the predictions are still far from perfect. In particular, at reasonably high positive predictive value the sensitivity amounts to less than 10% of all true contacts. Although it is clear that contact predictions based only on statistical dependencies could be further improved, for example by a more rigorous treatment of the phylogenetic dependencies, we believe that it is unlikely that such improvements would dramatically enhance the performance. First of all, simple inspection of the data shows that a large number of the pairs that are contacts in the sense that they are less than 8Å apart, really show no sign of co-evolution at all. That is, a large fraction of 'contacts' may simply not interact directly, and these obviously can never be detected using statistical dependence measurements. On the other end of the scale are residues that contact so many others that they are very strongly constrained, and show almost no variability in evolution. For such highly conserved residues it is also inherently impossible to identify their interaction partners using co-evolutionary analysis.

We thus believe that the largest further improvements to contact prediction are to be expected from incorporating information other than statistical dependency. To illustrate that additional information can be easily incorporated into our model, we developed an informative prior that takes into account that the likelihood of a contact to exist depends on the primary-sequence distance of the residues, and that highly conserved residues tend to have a higher number of contacts. The incorporation of even this simple additional information already leads to dramatic improvements in contact prediction. Clearly more powerful priors could be developed that take into account more sophisticated structural knowledge. In addition, in our current method we integrate over all possible joint probabilities for pairs of interacting residues, effectively assuming that all possible joint probability distributions are equally likely. Here too improvements could likely be made by taking into account prior knowledge on which joint probability distributions are more or less likely for interacting pairs of amino acids. Ultimately the most satisfying approach would be to combine our approach with direct structural modeling, i.e. somewhat along the lines of the approach taken in [34].

Following the plausible intuition that, the more different kinds of information are taken into account, the greater the prediction accuracy that can be obtained, several machine learning and statistical methods have been proposed that incorporate a much larger number of different features (see [20,34,35] and references therein). Besides primary sequence separation and conservation, these methods include features such as domain length, relative solvent accessibility, predicted secondary structure, the amino acid composition in short windows around the positions of interest, chemical properties of the amino acids, and contact potentials. Due to varying training and test sets and varying standards of evaluation, it is very difficult to compare the performance of our method with these approaches. However, some principal differences between these methods and ours should be noted. First, all these methods rely on training sets to fit parameters, so that additional methods are required to avoid over-fitting, whereas our method is essentially without any tunable parameters and does not require any training sets. Second, some of these methods are rather *ad hoc* 'black box' methods, e.g. neutral networks [20] or support vector machines [35], that use partially redundant sets of features, from which it is typically hard to derive mechanistic insights. In contrast, our method is derived directly from first principles. In any case, the results that we have presented show that it is crucial to take indirect dependencies into account when incorporating co-evolution information. We have provided a rigorous method for doing so and it is clear that any contact prediction method that incorporates co-evolution information would strongly benefit from using our method for disentangling direct and indirect dependencies.

Whereas we have here applied our method to predict contacting residues in a single protein, it is straight forward to use the same method for predicting contacting residues between pairs of proteins that are known to interact. That is, given two set of orthologs proteins s_1 and s_2, for which it is known that each member of set s_1 interacts with the corresponding member of set s_2, we can simply concatenate the multiple alignments of s_1 and s_2 into one longer multiple alignment, and apply our method to this longer alignment.

More generally, our method provides a computationally tractable extension of weight matrix models to take into account arbitrary pairwise dependencies, and there are a number of more general applications that we envisage pursuing in the future. First, our method can be generally used to 'score' multiple alignments in a way that includes pairwise dependencies. This could be used to discover subfamilies within large multiple alignments or to generally refine multiple alignments. Since the performance of alignment-based contact prediction methods is expected to depend strongly on the quality of the alignments, such a refinement may further improve contact prediction. Finally, another attractive application is to develop a regulatory-motif finding algorithm that takes into account arbitrary pairwise dependencies between positions.

Materials and Methods

Domain sequences and structures

Domain alignments and the mappings from domains to available structures in the PDB database were downloaded from the Pfam database [19,36]. We only used Pfam A, which is the high-quality and manually curated part of Pfam [19]. For each Pfam domain with at least one known structure, we reduced the alignment to positions corresponding to match states of the corresponding Pfam hidden Markov model with no more than 20 percent gaps. The removal of columns with many gaps is necessary as gaps can cause spurious correlations (see below) and make it difficult to compare the phylogenetic background signal between different columns. We removed from each alignment all multiple copies of identical sequences as well as sequences that had more than 50 percent gaps with respect to the match states. Additionally, alignments containing less than 100 sequences or less than 50 columns were discarded. To keep computational times limited we also removed alignments with more than 400 columns. For each Pfam alignment, all corresponding PDB files were collected according to the iPfam annotation [36] and distances between pairs of residues were determined as the distance between the C_β atoms (C_α for glycines). In the case of NMR models, the minimal distances of all models contained in the PDB entry were chosen. If a Pfam domain was present in multiple protein structures or in several chains of one protein structure, we chose the median distance over all chains and structures. For some alignments the corresponding structure did not cover all columns in the alignment and we discarded the small number of examples where the coverage was less than 50%. This resulted in 2009 domains with structurally-defined distances between residues. Finally, distance in primary sequence was defined as the distance between the match states of the alignment.

Probabilistic model

Our Bayesian network model was described in detail in [27]. Briefly, given a single column i of the alignment with observed amino acid counts n_α^i, the probability $P(D_i|w^i)$ of the column is given in terms of the (unknown) probability distribution w^i, with w_α^i the probability that letter α occurs at position i, i.e. $P(D_i|w^i)=\Pi_\alpha (w_\alpha^i)^{n_\alpha^i}$. Using a Dirichlet prior for w^i with parameter λ, we obtain the marginal probability of the column $P(D_i)$ by integrating over all possible distributions w^i. This integral can be performed analytically and the result can be expressed in terms of gamma functions:

$$P(D_i)=\frac{\Gamma(20\lambda)}{\Gamma(n+20\lambda)}\prod_\alpha \frac{\Gamma(n_\alpha^i+\lambda)}{\Gamma(\lambda)}, \tag{2}$$

where n is the number of sequences in the alignment. Similarly, the *joint* probability of the data D_{ij} in a pair of columns (ij) is given in terms of the number of times $n_{\alpha\beta}^{ij}$ that the combination of letters $(\alpha\beta)$ occurs at positions (ij), i.e.

$$P(D_{ij})=\frac{\Gamma(20^2\lambda')}{\Gamma(n+20^2\lambda')}\prod_{\alpha\beta} \frac{\Gamma(n_{\alpha\beta}^{ij}+\lambda')}{\Gamma(\lambda')}. \tag{3}$$

Here, we set the parameter λ' of the Dirichlet prior for the joint probability distribution to 0.5. As shown in [28], in the context of a dependence tree model, consistency requires that λ equals $20\lambda'$.

The statistical dependence between columns i and j is quantified by the ratio

$$R_{ij}=\frac{P(D_{ij})}{P(D_i)P(D_j)}. \tag{4}$$

The connection of $\log(R)$ to mutual information is easily established by substituting equations (2) and (3) into the logarithm of R as given by (4) and using Stirling's approximation to the logarithm of the gamma function. We then find that approximately

$$R_{ij}\propto e^{nI_{ij}} \tag{5}$$

for large n, with I_{ij} the mutual information between columns i and j. Importantly, when determining the counts $n_{\alpha\beta}^{ij}$ and n_α^i in order to determine R_{ij}, we discard all pairs of residues within a given sequence where either α or β is a gap. Treating gaps as a 21 amino acid causes strong spurious correlations between residues that are close in primary sequence since gaps usually come in blocks (data not shown).

A *dependence tree* π specifies for each position i (except for the root of the tree) a parent position $\pi(i)$ which is the residue that i depends on. To keep the notation simple, we here use the symbol π to both denote the mapping from a node to its parent node and the dependence tree itself. It can be shown [27] that, given a dependence tree, the joint probability $P(D|\pi)$ of the entire alignment can be written as

$$P(D|\pi)=\prod_i P(D_i)\prod_{j\neq r} R_{j\pi(j)}, \tag{6}$$

where the first product goes over all positions and the second over all positions except for the root r.

Finally, the probability $P(D)$ of the whole alignment is given by summing over all possible dependence trees π

$$P(D)=\prod_i P(D_i)\left(\sum_\pi P(\pi)\prod_{j\neq r} R_{j\pi(j)}\right), \tag{7}$$

where $P(\pi)$ is the prior probability of a particular spanning tree π. The last product is in fact the product of the R-values over all edges of the tree given by π and is independent of the choice of the root. If the prior probability of a spanning tree can be written as a product of probabilities $W_{j\pi(j)}$ along each edge $(j,\pi(j))$ of the tree

$$P(\pi)=\prod_{j\neq r} W_{j\pi(j)} \tag{8}$$

then equation (7) can be rewritten as

$$P(D)=\prod_i P(D_i)\left(\sum_\pi \prod_{j\neq r} M_{j\pi(j)}\right) \tag{9}$$

with $M_{j\pi(j)} \doteq R_{j\pi(j)}W_{j\pi(j)}$. Thus, the weight of each edge is simply multiplied by its prior probability. The largest term in the sum of equation (9) is the *maximum spanning tree* when a weight $\log(M_{ij})$ is assigned to each edge (ij) and this maximum spanning tree can be easily determined [37].

The sum over spanning trees in (9) can be calculated using a generalization of Kirchhoff's matrix-tree theorem [28]. For this we need to calculate the Laplacian of the matrix M_{ij}, which is defined as

$$L_{ij}=\delta_{ij}(\sum_k M_{ik})-M_{ij} \tag{10}$$

where the sum goes over all columns (or rows) of the M-matrix and δ_{ij} is the Kronecker delta function, which is one if $i=j$ and zero otherwise. We can then write the sum over all spanning trees as

$$\sum_\pi \prod_{i\neq r} M_{i\pi(i)}=\det(Q(L)) \tag{11}$$

where $Q(L)$ is the matrix L with one line and column removed (the determinant is independent of which line and column are removed). The summation over all spanning trees (there are n^{n-2} spanning trees for a full graph with n nodes) thus reduces to the calculation of a determinant, which can be done in a time proportional to n^3.

As discussed previously [27], the calculation of the determinant of the matrix M_{ij} is numerically very challenging since the entries M_{ij} vary over many orders of magnitude. In order to circumvent this problem, we rescale the entries of the matrix as suggested in [38]:

$$M_{ij}\rightarrow\beta\left(M_{ij}\right)^\alpha \tag{12}$$

with $\alpha=\dfrac{K\log(10)}{\log M_+-\log M_-}$ and $\beta=-K\log(10)\dfrac{\log M_+}{\log M_+-\log M_-}$ where $\log M_+$ ($\log M_-$) is the logarithm of the maximal (minimal) entry of the matrix M_{ij}. This function maps all M values into the interval $[10^{-K},1]$, preserves the relative ordering of entries and does not exaggerate relative differences in belief [38]. The lower bound 10^{-K} ensures that the rescaled M-matrix remains numerically non-singular. K can be set according to the numerical precision of the machine and we set $K=5$. We then use these rescaled M-values to calculate the posterior probabilities.

Calculating posteriors

Using expression (7), the posterior probability of a particular edge (kl) is given by

$$P((kl)|D) = \frac{P_{kl}(D)}{P(D)} \tag{13}$$

where

$$P_{kl}(D) = \prod_i P(D_i) \left(\sum_{\pi:(kl)\in\pi} \prod_{j\neq r} M_{j\pi(j)} \right) \tag{14}$$

which is the sum of the probabilities $P(D|\pi)P(\pi)$ for all spanning trees π that contain the edge (kl). This expression can be calculated by replacing the set of n nodes with a set of $(n-1)$ nodes, in which nodes k and l are contracted to one node, say kl, and the edge weights of this new node kl are given by $M_{kl,f} = M_{k,f} + M_{l,f}$ for all nodes $f \neq k,l$ [39]. Using this construction we can write the sum over all spanning trees containing edge (kl) as

$$P_{kl}(D) = \prod_i P(D_i) \left(M_{kl} \sum_{\pi'} \prod_{j\neq r} M_{j\pi'(j)} \right) \tag{15}$$

where the sum now goes over all spanning trees π' of the $(n-1)$ nodes. This sum over spanning trees can of course also be calculated as a determinant as described above. Roughly speaking, an edge (kl) will have high posterior if it occurs in the large majority of all spanning trees π that have high probability $P(D,\pi)$.

Phylogenetic correction

Due to the phylogenetic relatedness of the sequences in the alignment, there typically will be a statistical dependence between residues even in the absence of a functional linkage of these positions. Previous work [14] showed that this dependence can be corrected for (to some extent) by assuming that, due to phylogenetic relationships, each position has a certain amount of 'background' statistical dependence with other columns. Since each position interacts only with a small fraction of all other positions this background dependence can be estimated by calculating the average mutual information of that position with all the remaining positions. In [14], two types of corrections were proposed, a multiplicative one, named APC, and a additive one, named ASC. We here briefly review the derivation of these corrections.

The idea of the ASC is that the mutual information I_{ij} between positions i and j is the sum of the true mutual information I_{ij}^{true} and background mutual informations B_i and B_j, associated with positions i and j, i.e.

$$I_{ij} = I_{ij}^{true} + B_i + B_j. \tag{16}$$

We define average mutual informations as

$$\langle I_{i\cdot} \rangle = \frac{1}{m} \sum_{j=1}^{m} I_{ij}, \tag{17}$$

with m the number of columns of the alignment. Other averages like $\langle I_{\cdot\cdot} \rangle$, $\langle B \rangle$, and so on, are defined analogously. Note that, for notational simplicity, in these averages we have adopted the convention that $I_{ii} = 0$. We can then derive the equalities

$$\langle I_{\cdot\cdot} \rangle = \langle I_{\cdot\cdot}^{true} \rangle + 2\langle B \rangle, \tag{18}$$

and

$$\langle I_{i\cdot} \rangle = \langle I_{i\cdot}^{true} \rangle + B_i + \langle B \rangle. \tag{19}$$

If one assumes that, since true interactions are relatively rare, the averages $\langle I_{\cdot\cdot}^{true} \rangle$ and $\langle I_{i\cdot}^{true} \rangle$ are much smaller than $\langle B \rangle$, we can set $\langle I_{\cdot\cdot}^{true} \rangle \approx 0$ and $\langle I_{i\cdot}^{true} \rangle \approx 0$ and have

$$\langle B \rangle = \langle I_{\cdot\cdot} \rangle / 2, \tag{20}$$

and

$$B_i = \langle I_{i\cdot} \rangle - \langle I_{\cdot\cdot} \rangle / 2. \tag{21}$$

Finally, under these assumptions the true mutual information I_{ij}^{true} is then given by

$$I_{ij}^{true} = I_{ij} - \langle I_{i\cdot} \rangle - \langle I_{j\cdot} \rangle + \langle I_{\cdot\cdot} \rangle. \tag{22}$$

Motivated by this derivation, the ASC is defined as

$$I_{ij}^{c} = I_{ij} - \langle I_{i\cdot} \rangle - \langle I_{j\cdot} \rangle + \langle I_{\cdot\cdot} \rangle. \tag{23}$$

In the product correction APC we assume that the background mutual information between i and j can be written as a *product* of contributions of the two columns, i.e.

$$I_{ij} = I_{ij}^{true} + B_i B_j. \tag{24}$$

Assuming again that the true average mutual informations are small we find

$$\langle B \rangle^2 = \langle I_{\cdot\cdot} \rangle, \tag{25}$$

and

$$B_i = \frac{\langle I_{i\cdot} \rangle}{\sqrt{\langle I_{\cdot\cdot} \rangle}}. \tag{26}$$

Using this the APC version of the mutual information is given by

$$I_{ij}^{c} = I_{ij} - \frac{\langle I_{i\cdot} \rangle \langle I_{j\cdot} \rangle}{\langle I_{\cdot\cdot} \rangle}. \tag{27}$$

Since the APC performs better than the ASC we focused on adapting the APC for our Bayesian model. As mentioned above, the logarithms of the R values are the equivalent of mutual information in our model. Therefore, naively we would simply replace I_{ij} with $\log(R_{ij})$ in equation (27) above. However, whereas the mutual information naturally has a lower bound of zero, which is reached only for independent positions, $\log(R)$ is off-set with respect to mutual information and becomes *negative* for independent positions. Note also that all posterior probabilities are invariant under a global shift of all the $\log(R)$ values by a

constant. Therefore, we substitute into equation (27) a shifted version of $\log(R)$ which is guaranteed to be non-negative. For each domain we determine the minimal value $\log(R_{min})$ and define a shifted version of $\log(R)$ as

$$S_{ij} = \log(R_{ij}) - \log(R_{min}). \tag{28}$$

Using these shifted $\log(R)$s we then define the corrected $\log(R)$ as

$$\log(R^c_{ij}) = S_{ij} - \frac{\langle S_{i\cdot}\rangle\langle S_{j\cdot}\rangle}{\langle S_{\cdot\cdot}\rangle}. \tag{29}$$

In our model with phylogenetic correction we simply replace each factor R_{ij} with R^c_{ij}.

Prior probability of spanning trees

Our Bayesian model easily allows for the incorporation of prior probabilities on each spanning tree via the edge probabilities $W_{j\pi(j)}$ in equation (9). Here, we use these edge probabilities to include the dependence on both the primary sequence separation of the positions in the pair (Figure 11), as well as the sum of the entropies of the corresponding columns (Figure 12). To estimate the fraction $f(d,H)$ of all pairs with sequence-separation d and entropy-sum H that are contacts, we separated all pairs of columns into entropy bins of width 0.2, spanning the whole range of entropies $[0,2\log(20)]$ and compared the dependence on primary sequence separation within the different bins (Figure 14, left panel).

We see that, irrespective of the column entropy sum H, the fraction $f(d,H)$ has approximately the same shape as a function of d as the overall fraction of contacts $f(d)$ which we showed in Figure 11. We find that for distances $d=4$ or less the fraction is virtually independent of entropy, i.e. $f(d,H)\approx f(d)$, while for larger distances the fractions $f(d,H)$ are roughly proportional to $f(d)$, with a proportionality constant that decreases with entropy H. That is, we assume the following general form for $f(d,H)$:

$$f(d,H) = \begin{cases} f(d) & \text{if } d <= 4 \\ f(d)g(H) & \text{if } d > 4 \end{cases} \tag{30}$$

We first estimated $f(d)$ directly from the observed fractions as shown in Figure 11 for all sequence separations up to $d=50$. As $f(d)$ is proportional to $1/d$ for sequence separations ≥ 50 and becomes very noisy for large sequence separations (data not shown), we approximate the curve as $f(d)=C/d$ for sequence separations ≥ 50 (blue line in Figure 14). The constant C is chosen so that the curve is continuous at $d=50$. We then determined the function $g(H)$ by numerically maximizing, for each fixed entropy bin H_i, the likelihood of the data, which is given by

$$P(X) = \left[\prod_{e\in E} f(d_e)X\right]\left[\prod_{e\notin E}(1-f(d_e)X)\right], \tag{31}$$

where the first product runs over all edges E with $d>4$ and $H=H_i$ that are contacts, the second product over all edges with $d>4$ and $H=H_i$ that are not contacts, and d_e stands for the primary sequence separation of edge e. The value X^* that maximizes the likelihood of the data determines the value of $g(H)$ for the bin H_i, i.e. $g(H_i)=X^*$. The resulting function $g(H)$ is shown in the right panel of figure 14. Clearly the probability of an edge decreases with the entropy-sum H, i.e. it drops by almost a factor of 5 from the lowest to the highest entropy edges.

Finally, in order to assign prior probabilities to different possible spanning trees, we assume a random graph model where each edge e occurs with a probability μ_e that is proportional to $f(d_e,H_e)$, with d_e the primary sequence separation, and H_e the entropy sum of edge e. Note that each spanning tree only contains $(l-1)$ edges for a domain of length l, and we thus have to ensure that our random graph model produces on average $(l-1)$ edges.

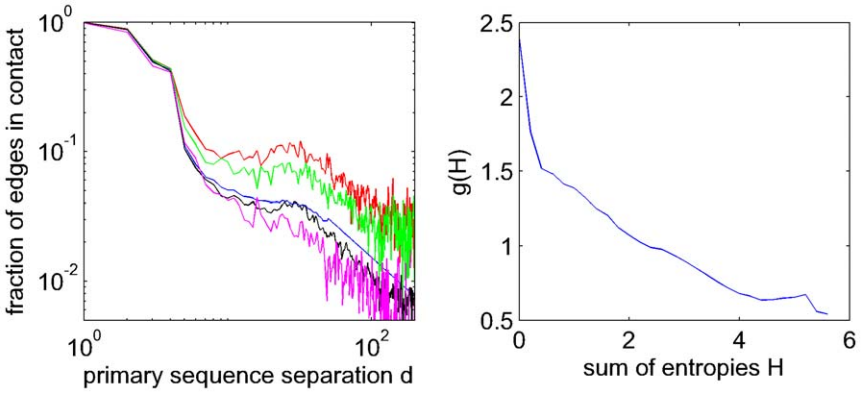

Figure 14. Estimation of prior probabilities. The left panel shows the dependence between the fraction of pairs that are in contact and primary sequence separation for all pairs (in blue) as well as for pairs whose sum of entropies lies in a given entropy bin (H∈[0,0.2) in red, H∈[0.2,0.4) in green, H∈[3.4,3.6) in black and H∈[5.4,5.6) in magenta). For the sake of clarity, only a few selected entropy bins across the entire range are shown. The right panel shows the estimated function $g(H)$, which describes how the probability of an edge to be a contact depends on the sum of entropies of the corresponding columns of the alignment (see text).
doi:10.1371/journal.pcbi.1000633.g014

As the expected number of edges in a random graph is equal to the sum over all μ_e, we set μ_e to

$$\mu_e = (l-1)\frac{f(d_e,H_e)}{\sum_e f(d_e,H_e)}. \quad (32)$$

Let G be the full graph including all $\binom{l}{2}$ edges of a particular domain and let π be one particular spanning tree π. We can now write the prior probability of the tree as

$$P(\pi) = \prod_{e\in\pi}\mu_e \prod_{e\in G\setminus\pi}(1-\mu_e) \quad (33)$$

Here, the first product runs over all edges e in the tree π and the second one over all edges in G that are not in the tree π. Since the posteriors are independent of a global rescaling of all prior probabilities $P(\pi)$, we divide $P(\pi)$ by the probability of the graph that contains no edges, to obtain

$$P(\pi) \propto \prod_{e\in\pi}\frac{\mu_e}{1-\mu_e} \quad (34)$$

which is independent of the edges that are not contained in the tree. We can thus set the edge weights $W_{j\pi(j)}$ in equation 9 to

$$W_{j\pi(j)} = \frac{\mu_{j\pi(j)}}{1-\mu_{j\pi(j)}}. \quad (35)$$

Unfortunately, we cannot directly used $W_{j\pi(j)}$ to calculate the matrix entries $M_{j\pi(j)} = R_{j\pi(j)}W_{j\pi(j)}$ in equation 9. As discussed above, the R-values relate to mutual information I through $R \propto e^{nI}$, where n is the total number of sequences in the alignment. However, even when the phylogenetic correction is employed, because the n sequences contain many phylogenetically closely-related sequences, the number of *statistically independent* sequences is generally much smaller than n. Because of this, even the corrected R-values still significantly overestimate statistical dependence. To take this into account we define the matrix entries $M_{j\pi(j)}$ as

$$M_{j\pi(j)} = \left(R_{j\pi(j)}\right)^{\alpha} W_{j\pi(j)} \quad (36)$$

where α is a free parameter, which must lie between 0 (only prior information) and 1 (original R-values). Note that, through this transformation, we are assuming that instead of n independent sequences, there are only αn effectively independent sequences. The PPV-sensitivity curves for varying values of α are shown in Figures S10, S11, and S12. For the curve in the main text, we chose $\alpha = 0.025$, so as to maximize the accuracy for pairs with $d \geq 3$ without a significant decrease in accuracy for pairs with $d \geq 12$.

Supporting Information

Figure S1 Number of contacts n versus the number of residues l per protein domain for varying separations in primary sequence. The red lines are the regression lines (in log-space), corresponding to the power-laws $n = 2.43l^{1.12}$, $n = 0.16l^{1.43}$ and $n = 0.05l^{1.62}$. The dashed black line corresponds to $n = l$.
Found at: doi:10.1371/journal.pcbi.1000633.s001 (0.33 MB TIF)

Figure S2 Accuracy of contact predictions for all 2009 alignments based on mutual information (black), log*(R)* (blue), and posterior probabilities (red). For different values of sensitivity, the corresponding number of predictions for each domain and each method were selected and their positive predicted value (PPV), i.e. the fraction of correct predictions, was calculated (vertical axis). Dashed lines indicate mean PPV plus/minus one standard error. The top left panel shows predictions for all residue pairs, the top right one using only predictions for residues separated by at least 3 positions in the primary sequence, the bottom left one for pairs separated by at least 12 positions, and the bottom right panel for pairs separated by at least 24 positions.
Found at: doi:10.1371/journal.pcbi.1000633.s002 (0.32 MB TIF)

Figure S3 Comparison of prediction accuracy for log*(R)* (blue), for the log*(R)* values contained in the maximum-likelihood tree (green) and for the posterior probability (red). As the maximum-likelihood tree only predicts *l-1* edges, where *l* is the number of columns of the alignment, the different measures cannot be directly compared in terms of sensitivity (there would be finite-length effects as predictions by the maximum-likelihood tree measure cannot reach a sensitivity of *1*). Instead, we sort the predictions per domain and, for each fixed cut-off on the rank *r*, we show the average positive predictive value (solid lines) for all predictions with rank *r* or higher. The dashed lines indicate plus/minus one standard error. As the shortest domains in our dataset have length 50, all domains are included in the calculation of the green curve for ranks 1 to 49. The blue and green curves are identical for high ranks as all the highest-scoring edges are included in the maximum spanning tree. However, for decreasing ranks, the maximum-spanning tree discards edges that can be explained indirectly, which leads to an improvement in performance. Importantly, the posterior probability significantly outperforms the maximum-spanning tree predictions both for low *and* high ranks.
Found at: doi:10.1371/journal.pcbi.1000633.s003 (0.36 MB TIF)

Figure S4 Posteriors reflect the extent to which co-evolving pairs can be explained by contact chains. Shown are the reverse cumulative distributions of distal co-evolving pairs (*Z>4*) that cannot be easily explained by contact chains, i.e. where the best scoring chain has a score of *S>2* (red), *S>3* (dark blue), or *S>4* (light blue). For comparison the reverse cumulative distributions of posteriors for all co-evolving distal pairs (green) and all co-evolving contacts (black) are also shown.
Found at: doi:10.1371/journal.pcbi.1000633.s004 (0.13 MB TIF)

Figure S5 Accuracy of contact predictions for all alignments. In blue, we show the performance of the phylogenetically-corrected posterior probabilities, in black the performance of the predictions based on the average-product corrected (APC) mutual information, and in red the performance of the posterior probabilities without phylogenetic correction. Curves were calculated as described in the main text.
Found at: doi:10.1371/journal.pcbi.1000633.s005 (0.33 MB TIF)

Figure S6 Accuracy of contact predictions for alignments of length 50 to 100. In blue, we show the performance of the phylogenetically-corrected posterior probabilities, in black the performance of the predictions based on the average-product corrected (APC) mutual information, and in red the performance of the posterior probabilities without phylogenetic correction. Curves were calculated as described in the main text.
Found at: doi:10.1371/journal.pcbi.1000633.s006 (0.33 MB TIF)

Figure S7 Accuracy of contact predictions for alignments of length 101 to 200. In blue, we show the performance of the phylogenetically-corrected posterior probabilities, in black the performance of the predictions based on the average-product corrected (APC) mutual information, and in red the performance

of the posterior probabilities without phylogenetic correction. Curves were calculated as described in the main text.
Found at: doi:10.1371/journal.pcbi.1000633.s007 (0.33 MB TIF)

Figure S8 Accuracy of contact predictions for alignments of length 201 to 300. In blue, we show the performance of the phylogenetically-corrected posterior probabilities, in black the performance of the predictions based on the average-product corrected (APC) mutual information, and in red the performance of the posterior probabilities without phylogenetic correction. Curves were calculated as described in the main text.
Found at: doi:10.1371/journal.pcbi.1000633.s008 (0.33 MB TIF)

Figure S9 Accuracy of contact predictions for alignments of length 301 to 400. In blue, we show the performance of the phylogenetically-corrected posterior probabilities, in black the performance of the predictions based on the average-product corrected (APC) mutual information, and in red the performance of the posterior probabilities without phylogenetic correction. Curves were calculated as described in the main text.
Found at: doi:10.1371/journal.pcbi.1000633.s009 (0.33 MB TIF)

Figure S10 Accuracy of contact predictions including the informative prior for different values of the weighting parameter α, including the limit of using only the informative prior ($\alpha = 0$). The positive predictive value (vertical axis) is shown as a function of sensitivity (horizontal axis). Different colors correspond to different values of α (see legend) and dashed lines show mean plus and minus one standard error. For comparison, we also show the performance of the posterior when using no prior information (black). Note that the horizontal axis is shown on a logarithmic scale.
Found at: doi:10.1371/journal.pcbi.1000633.s010 (0.37 MB TIF)

Figure S11 Accuracy of contact predictions including the informative prior for different values of the weighting parameter α, including the limit of using only the informative prior ($\alpha = 0$), when considering only pairs that are at least $d = 3$ apart in primary sequence. The positive predictive value (vertical axis) is shown as a function of sensitivity (horizontal axis). Different colors correspond to different values of α (see legend) and dashed lines show mean plus and minus one standard error. For comparison, we also show the performance of the posterior when using no prior information (black). Note that the horizontal axis is shown on a logarithmic scale.
Found at: doi:10.1371/journal.pcbi.1000633.s011 (0.36 MB TIF)

Figure S12 Accuracy of contact predictions including the informative prior for different values of the weighting parameter α, including the limit of using only the informative prior ($\alpha = 0$), when considering only pairs that are at least $d = 12$ apart in primary sequence. The positive predictive value (vertical axis) is shown as a function of sensitivity (horizontal axis). Different colors correspond to different values of α (see legend) and dashed lines show mean plus and minus one standard error. For comparison, we also show the performance of the posterior when using no prior information (black). Note that the horizontal axis is shown on a logarithmic scale.
Found at: doi:10.1371/journal.pcbi.1000633.s012 (0.33 MB TIF)

Author Contributions

Conceived and designed the experiments: EvN. Performed the experiments: LB. Analyzed the data: LB. Wrote the paper: LB EvN.

References

1. Eddy S (1998) Profile hidden markov models. Bioinformatics 14: 755–763.
2. Hunter S, Apweiler R, Attwood TK, Bairoch A, Bateman A, et al. (2009) Interpro: the integrative protein signature database. Nucleic Acids Res 35: D224–228.
3. Eddy S, Durbin R (1994) RNA sequence analysis using covariance models. Nucleic Acids Research 22(11): 2079–2088.
4. Lindgreen S, Gardner P, Krogh A (2006) Measuring covariation in RNA alignments: physical realism improves information measures. Bioinformatics 22(24): 2988–2995.
5. Yanovsky C, Horn V, Thorpe D (1964) Protein structure relationships revealed by mutational analysis. Science 146: 1593–1594.
6. Fitch W, Markowitz E (1970) An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. Biochem Genet 4: 579–593.
7. Maisnier-Patin S, Andersson D (2004) Adaptation to the deleterious effect of antimicrobial drug resistance mutations by compensatory evolution. Research in Microbiology 155: 360–369.
8. Wollenberg K, Atchley W (2000) Separation of phylogenetic and functional associations in biological sequences by using the parametric bootstrap. PNAS 97: 3288–3291.
9. Tillier E, Liu T (2003) Using multiple interdependency to separate functional from phylogenetic correlations in protein alignments. Bioinformatics 19(6): 750–755.
10. Fodor A, Aldrich R (2004) Influence of conservation on calculations of amino acids covariance in multiple sequence alignments. Proteins: Structure, Function, and Bioinformatics 56: 211–221.
11. Martin L, Gloor G, Dunn S, Wahl L (2005) Using information theory to search for co-evolving residues in proteins. Bioinformatics 21(22): 4116–4124.
12. Fares M, Travers S (2006) A novel method for detecting intramolecular coevolution: Adding a further dimension to selective constraints analyses. Genetics 173: 9–23.
13. Gouveia-Oliveira R, Pedersen A (2007) Finding coevolving amino acid residues using row and column weighting of mutual information and multi-dimensional amino acid representation. Algorithms for Molecular Biology 2: 12.
14. Dunn S, Wahl L, GB G (2008) Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. Bioinformatics 24(3): 333–340.
15. Yeang CH, Haussler D (2007) Detecting coevolution in and among protein domains. PLoS Computational Biology 3: e211.
16. Pazos F, Valencia A (2008) Protein co-evolution, co-adaptation and interactions. The EMBO Journal 27: 2648–2655.
17. Cover TM, Thomas JA (1991) Elements of information theory John Wiley and Sons.
18. Chiu D, Kolodziejczak T (1991) Inferring consensus structure from nucleic acid sequences. Comput Appl Biosc 7: 347–52.
19. Bateman A, Coin L, Durbin R, Finn R, Hollich V, et al. (2004) The Pfam protein families database. Nucl Acids Res 32: D138–D141.
20. Shackelford G, Karplus K (2007) Contact prediction using mutual information and neural nets. Proteins 69(Suppl 8): 159–164.
21. Izarzugaza J, Graña O, Tress M, Valencia A, Clarke N (2007) Assessment of intramolecular contact predictions for CASP7. Proteins 69(Suppl 8): 152–158.
22. Weigt M, White R, Szurmant H, Hoch J, Hwa T (2009) Identification of direct residue contacts in protein-protein interaction by message passing. PNAS 106: 67–72.
23. Lockless S, Ranganathan R (1999) Evolutionarily conserved pathways of energetic connectivity in protein families. Science 8(286): 295–299.
24. Süel G, Lockless S, Wall M, Ranganathan R (2003) Evolutionarily conserved networks of residues mediate allosteric communication in proteins. Nature Structural Biology 10(1): 59–69.
25. Gloor G, Martin L, Wahl L, Dunn S (2005) Mutual information in protein multiple sequence alignments reveals two classes of coevolving positions. Biochemistry 44: 7156–7165.
26. Fodor A, Aldrich R (2004) On evolutionary conservation of thermodynamic coupling in proteins. Journal of Biological Chemistry 279(18): 19046–19050.
27. Burger L, van Nimwegen E (2008) Accurate prediction of protein-protein interactions from sequence alignments using a Bayesian method. Molecular Systems Biology 4: 165.
28. Meilà M, Jaakkola T (2006) Tractable Bayesian learning of tree belief networks. Statistics and Computing 16(1): 77–92.
29. Olmean O, Rost B, Valencia A (1999) Effective use of sequence correlation and conservation in fold recognition. J Mol Biol 295: 1221–1239.
30. Pollock D, Taylor W, Goldman N (1999) Coevolving protein residues: maximum likelihood identification and relationship to structure. J Mol Biol 287: 187–198.
31. Kortemme T, Baker D (2002) A simple physical model for binding energy hot spots in protein-protein complexes. PNAS 99(22): 14116–14121.
32. Rost B, Sander C (1994) Conservation and prediction of solvent accessibility in protein families. Proteins 20: 216–226.
33. Halabi N, Rivoire O, Leibler S, Ranganathan R (2009) Protein sectors: evolutionary units of three-dimensional structure. Cell 138(4): 774–786.
34. Miller C, Eisenberg D (2008) Using inferred residue contacts to distinguish between correct and incorrect protein models. Bioinformatics 24(14).

35. Cheng J, Baldi P (2007) Improved residue contact prediction using support vector machines and a large feature set. BMC Bioinformatics 8: 113.
36. Finn R, Marshall M, Bateman A (2005) iPfam: visualization of protein-protein interactions in pdb at domain and amino acid resolutions. Bioinformatics 21: 410–412.
37. Chow C, Liu C (1968) Approximating discrete probability distributions with dependence trees. IEEE Transactions on Information Theory IT-14: 462–467.
38. Cerquides J, de Màntaras RL (2003) Tractable bayesian learning of tree augmented naive bayes classifiers. Proceedings of Twentieth International conference on Machine Learning.
39. Bollobás B (1998) Modern Graph Theory. Berlin: Springer, corr. 2nd printing edition.