Direct-coupling analysis of residue coevolution captures native contacts across many protein families

Faruck Morcos[a,1], Andrea Pagnani[b,1], Bryan Lunt[a], Arianna Bertolino[c], Debora S. Marks[d], Chris Sander[e], Riccardo Zecchina[b,f], José N. Onuchic[a,g,2], Terence Hwa[a,2], and Martin Weigt[b,h,2]

[a]Center for Theoretical Biological Physics, University of California at San Diego, La Jolla, CA 92093-0374; [b]Human Genetics Foundation, Via Nizza 52, 10126 Turin, Italy; [c]Institute for Scientific Interchange, Viale Settimio Severo 65, 10133 Turin, Italy; [d]Department of Systems Biology, Harvard Medical School, 20 Longwood Avenue, Boston, MA 02115; [e]Memorial Sloan–Kettering Cancer Center, Computational Biology Center, 1275 York Avenue, New York, NY 10065; [f]Center for Computational Studies and Dipartimento di Fisica, Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129 Turin, Italy; [g]Center for Theoretical Biological Physics, Rice University, Houston, TX 77005-1827; and [h]Laboratoire de Génomique des Microorganismes, Unité Mixte de Recherche 7238, Université Pierre et Marie Curie, 15 rue de l'École de Médecine, 75006 Paris, France

Contributed by José N. Onuchic, October 12, 2011 (sent for review July 22, 2011)

The similarity in the three-dimensional structures of homologous proteins imposes strong constraints on their sequence variability. It has long been suggested that the resulting correlations among amino acid compositions at different sequence positions can be exploited to infer spatial contacts within the tertiary protein structure. Crucial to this inference is the ability to disentangle direct and indirect correlations, as accomplished by the recently introduced direct-coupling analysis (DCA). Here we develop a computationally efficient implementation of DCA, which allows us to evaluate the accuracy of contact prediction by DCA for a large number of protein domains, based purely on sequence information. DCA is shown to yield a large number of correctly predicted contacts, recapitulating the global structure of the contact map for the majority of the protein domains examined. Furthermore, our analysis captures clear signals beyond intradomain residue contacts, arising, e.g., from alternative protein conformations, ligand-mediated residue couplings, and interdomain interactions in protein oligomers. Our findings suggest that contacts predicted by DCA can be used as a reliable guide to facilitate computational predictions of alternative protein conformations, protein complex formation, and even the de novo prediction of protein domain structures, contingent on the existence of a large number of homologous sequences which are being rapidly made available due to advances in genome sequencing.

statistical sequence analysis | residue–residue covariation | contact map prediction | maximum-entropy modeling

Correlated substitution patterns between residues of a protein family have been exploited to reveal information on the structures of proteins (1–10). However, such studies require a large number (e.g., the order of 1,000) of homologous yet variable protein sequences. In the past, most studies of this type have therefore been limited to a few exemplary proteins for which a large number of such sequences happened to be already available. However, rapid advances in genome sequencing will soon be able to generate this many sequences for the majority of common bacterial proteins (11). Sequencing a large number of simple eukaryotes such as yeast can in principle generate a similar number of common eukaryotic protein sequences. In this paper, we provide a systematic evaluation of the information contained in correlated substitution patterns for predicting residue contacts, a first step toward a purely sequence-based approach to protein structure prediction.

The basic hypothesis connecting correlated substitution patterns and residue–residue contacts is very simple: If two residues of a protein or a pair of interacting proteins form a contact, a destabilizing amino acid substitution at one position is expected to be compensated by a substitution of the other position over the evolutionary timescale, in order for the residue pair to maintain attractive interaction. To test this hypothesis, the bacterial two-component signaling (TCS) proteins (12) have been used because of the large number of TCS protein sequences, which already numbered in the thousands 5-y ago (13). Simple covariance-based analysis was first applied to characterize interactions between residues belonging to partner proteins of the TCS pathways (14, 15); it was found to partially predict correct interprotein residue contacts, but also many residue pairs which are far apart. A major shortcoming of covariance analysis is that correlations between substitution patterns of interacting residues induce secondary correlations between noninteracting residues. This problem was subsequently overcome by the direct-coupling analysis (DCA) (16, 17), which aims at disentangling direct from indirect correlations. The top 10 residue pairs identified by DCA were all shown to be true contacts between the TCS proteins, and they were used to guide the accurate prediction (3-Å rmsd) of the interacting TCS protein complex (18, 19). Furthermore, DCA was used to shed light on interaction specificity and interpathway cross-talk in bacterial signal transduction (20).

Due to rapid advances in sequencing technology, there exists by now a large number of bacterial genome projects, approximately 1,700 completed and 8,300 ongoing (11). These genome sequences can be used to compute correlated substitution patterns for a large number of common bacterial proteins and interacting protein pairs, even if they are not duplicated (i.e., present at one copy per genome on average). DCA can then be used in principle to infer the interacting residues and eventually predict tertiary and quaternary protein structures for the majority of bacterial proteins, as has been done so far for the TCS proteins. Here we address a critical question for this line of pursuit—how well does DCA identify native residue contacts in proteins other than TCS?

Previously, a message-passing algorithm was used to implement DCA (16). This approach, here referred to as mpDCA, was rather costly computationally because it is based on a slowly converging iterative scheme. This cost makes it unfeasible to apply mpDCA to large-scale analysis across many protein families. Here we will introduce mfDCA, an algorithm based on the mean-field approximation of DCA. The mfDCA is 10^3 to 10^4 times faster than mpDCA, and hence can be used to analyze many long protein sequences rapidly. By analyzing 131 large domain families for which accurate structural information is available, we show

Author contributions: F.M., A.P., J.N.O., T.H., and M.W. designed research; F.M., A.P., B.L., A.B., D.S.M., R.Z., and M.W. performed research; A.P. and M.W. contributed new reagents/analytic tools; F.M., B.L., A.B., D.S.M., C.S., J.N.O., T.H., and M.W. analyzed data; and F.M., A.P., C.S., J.N.O., T.H., and M.W. wrote the paper.

The authors declare no conflict of interest.

[1]F.M. and A.P. contributed equally to this work.

[2]To whom correspondence may be addressed. E-mail: jonuchic@ucsd.edu, hwa@ucsd.edu, or martin.weigt@upmc.fr.

See Author Summary on page 19459.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1111471108/-/DCSupplemental.

that mfDCA captures a large number of intradomain contacts across these domain families. Together, the predicted contacts are able to recapitulate the global structure of the contact map. Many cases, where mfDCA finds strong correlation between distant residue pairs, have interesting biological reasons, including interdomain contacts, alternative structures of the same domain, and common interactions of residues with a ligand. The mfDCA results are found to outperform those generated by simple covariance analysis as well as a recent approximate Bayesian analysis (10).

Results and Discussion

A Fast DCA Algorithm. In this study, we wish to characterize the correlation between the amino acid occupancy of residue positions as a predictor of spatial proximity of these residues in folded proteins. Starting with a multiple-sequence alignment (MSA) of a large number of sequences of a given protein domain, extracted using Pfam's hidden Markov models (HMMs) (21, 22), the basic quantities in this context are the frequency count $f_i(A)$ for a single MSA column i, characterizing the relative frequency of finding amino acid A in this column, and the frequency count $f_{ij}(A,B)$ for pairs of MSA columns i and j, characterizing the frequency that amino acids A and B coappear in the same protein sequence in MSA columns i and j. Alignment gaps are considered as the 21st amino acid. Mathematical definitions of these counts are provided in *Methods*.

The raw statistical correlation obtained above suffers from a sampling bias, resulting from phylogeny, multiple-strain sequencing, and a biased selection of sequenced species. The problem has been discussed extensively in the literature (10, 23–26). In this study, we implemented a simple sampling correction, by counting sequences with more than 80% identity and reweighting them in the frequency counts. All the frequency calculations and results reported below are obtained using this sampling correction; the number of nonredundant sequences is measured as the effective sequence number M_{eff} after reweighting (see *Methods*). The comparison to results without reweighting and to reweighting at 70% in *SI Appendix*, Fig. S1 shows that reweighting systematically improves the performance of DCA, but results are robust with respect to precise value of reweighting.

A simple measure of correlation between these two columns is the mutual information (MI), defined by Eq. **3** in *Methods*. As we will show, the MI turns out to be an unreliable predictor of spatial proximity. Central to our approach is the disentanglement of direct and indirect correlations, which is attempted via DCA, which takes the full set of $f_i(A)$ and $f_{ij}(A,B)$ as inputs, and infers "direct statistical couplings," which generate the empirically measured correlations. Their strength is quantified by the direct information (DI) for each pair of MSA columns; see Eq. **12** in *Methods* and ref. 16. However, the message-passing algorithm used to implement DCA in ref. 16, mpDCA, was computationally intensive, thus limiting its use in large-scale studies. Here we developed a much faster heuristic algorithm based on a mean-field approach; see *Methods*. This algorithm, termed mfDCA, is able to perform DCA for alignments of up to about 500 amino acids per row, as compared to 60–70 amino acids in the message-passing approach. For the same protein length, mfDCA is about 10^3 to 10^4 times faster, which results mainly from the fact that the costly iterative parameter learning in mpDCA can be solved analytically in a single step in mfDCA. This performance gain enabled us to systematically analyze hundreds of protein domains and examine the extent to which a high DI value is a predictor of spatial proximity in a folded protein. Many residue-position pairs, which are close neighbors along the sequence, also show high MI and/or DI. To evaluate nontrivial predictions, we therefore restricted our analysis throughout the paper to pairs, which are separated by at least five positions along the protein's backbone.

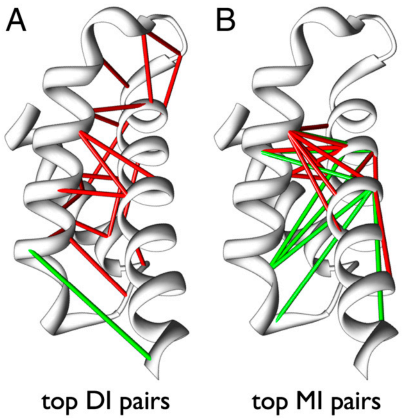

Fig. 1. Contact predictions for the family of domains homologous to Region 2 of the bacterial Sigma factor (Pfam ID PF04542) mapped to the sequence of the SigmaE factor of *E. coli* (encoded by *rpoE*) (PDB ID 1OR7). *A* shows the top 20 DI predictions, and *B* shows the top 20 MI predictions for residue–residue contacts, both with a minimum separation of five positions along the backbone. Each pair with distance <8 Å is connected by a red link, and the more distant pairs are connected by the green links.

Intradomain Contacts. We shall first illustrate the correlation between the DI values and the spatial proximity of residue pairs through a specific example, namely the domain family homologous to the DNA-recognition domain (region 2) of the bacterial Sigma-70 factor (Pfam ID PF04542). The mfDCA was used to compute the DI values using an M_{eff} of approximately 3,700 nonredundant sequences—i.e., below a threshold of 80% sequence identity. The MSA columns with the 20 largest DI and MI values are mapped to the sequence of the SigmaE factor of *Escherichia coli* (encoded by *rpoE*) whose structure has been solved to 2-Å resolution [Protein Data Bank (PDB) ID 1OR7; ref. 27]. The residue pairs with the 20 highest ranked DI values are connected by bonds of different colors in Fig. 1*A*. Those residue pairs with minimum atomic distances <8 Å are defined as "contacts" and are shown in red, the others in green.* Because only one out of the top 20 DI pairs is green, DI is seen as a good predicator of spatial contact, characterized by a true positive (TP) rate of 95% for this protein. A similar analysis using the 20 highest MI values (Fig. 1*B*) yielded 13 contacts (TP = 65%), illustrating a reduced predictive power by the simple covariance analysis. Furthermore, we see that the DI predictions are more evenly distributed over the entire domain, whereas many of the MI predictions are associated with a few residues; this difference is significant for contact map prediction and will be elaborated upon below.

In order to test the generality of the predictive power of DI ranking as contacts, we applied the above analysis to 131 predominantly bacterial domain families (with >90% of the sequences belonging to bacterial organisms). These families were selected according to the following two criteria (see *Methods* for details): (*i*) The family contains $M_{\text{eff}} > 1{,}000$ nonredundant sequences after applying sampling correction for >80% identity, in order to ensure statistical enrichment, and (*ii*) there exist at least two available high-quality X-ray crystal structures (independent PDB entries of resolution <3 Å), so that the degree of spatial proximity between each residue pair can be evaluated. The selected domain families encompassed a total of 856 different PDB structures (see *SI Appendix*, Table S1). Note that M_{eff} is found to be typically in the range of one-third to one-half of the total sequence number M (see *SI Appendix*, Fig. S2).

*The choice of the relatively large value of 8-Å minimum atom distance as a cutoff value for contacts is supported later in the discussion of Fig. 2*B*, where the distance distribution of the top DI pairings is analyzed.

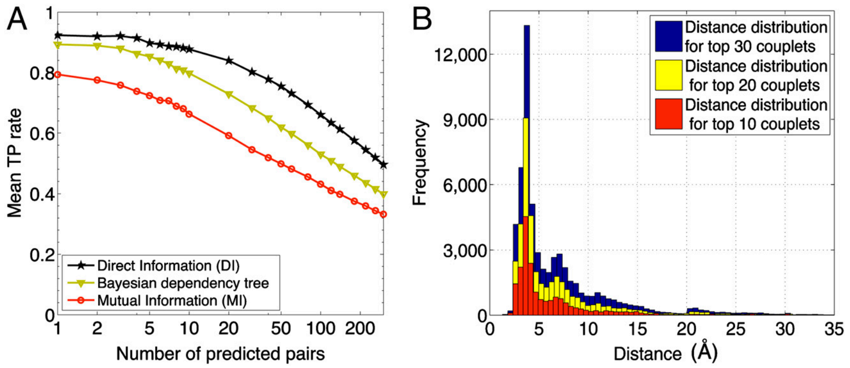

Fig. 2. (*A*) Mean TP rate for 131 domain families, as a function of the number of top-ranked contacts and histogram of the distances of all predicted structures for each of the 131 domains studied. DI results (★) clearly outperform the other two methods: MI (red ●) and an approximate Bayesian approach (yellow ▼) developed by Burger and van Nimwegen (10). Their method aims at disentangling direct and indirect correlations by averaging over tree-shaped residue–residue coupling networks, and it contains a phylogeny correction. The method can also reach length-400 multiple alignments as mfDCA does; our implementation follows closely the description in ref. 6. However, coupling trees do not allow for multiple coupling paths between two residues as DCA does, possibly accounting for its lower TP rates compared to mfDCA. (*B*) The mfDCA predictions for the top 10, 20, and 30 residue pairs show a bimodal distribution of intradomain distances with two frequency peaks around 3–5 and 7–8 Å.

We computed the DI values for each residue pair of the 131 domain families and evaluated the degree to which high-ranking DI pairs corresponded to actual contacts (minimum atomic distances <8 Å), based on the available structures for each domain. The results are shown in Fig. 2*A* (black star). The *x* axis represents the number of top-ranked DI pairs (separation >5 positions along the sequence) considered and the *y* axis is the average fraction of pairs up to this DI ranking that are true contacts. The latter was calculated using the best-predicted structure[†] (i.e., the PDB structure with the highest TP value) for each of the 131 families. Similar results were obtained when considering all the available structures; see below. In contrast, results computed using MI ranking (red circle) gave significantly reduced TP rates.[‡] Also shown in Fig. 2*A* are results generated by an approximate Bayesian approach, which has been established as the currently best-performing algorithm in identifying contacts from sequence correlation analysis (10). The Bayesian approach (yellow triangle) is seen to perform better than the simple covariance analysis (MI), but TP rates are not as high as the ones obtained by mfDCA. Analogous results for the relative performance of these methods are also observed for a collection of 25 eukaryotic proteins analyzed (see *SI Appendix*, Fig. S3), suggesting that the applicability of DCA is not restricted to bacterial proteins.

As seen in Fig. 2*A*, on average 84% of the top 20 DI pairs found by mfDCA (black star, black solid curve) are true contacts. The average TP rate is indicative of the TP of typical domain families, as the individual TPs for the 131 families examined are distributed mostly in the range of 0.7–1.0; see *SI Appendix*, Fig. S4*A* evaluated using the best-predicted structure and *SI Appendix*, Fig. S4*B* when all 856 structures are used. This figure also shows little difference in the quality of the prediction using the top 10, 20, or 30 DI pairs, and coherent results between the best-predicted and all 856 structures, despite the somewhat uneven distribution of available PDB structures over the 131 domain families. The distribution of the actual (minimum atomic) intradomain distances between residue pairs with the top 10, 20, and 30 DI ranking are shown in Fig. 2*B*, using the complete set of 856 PDB structures. The distribution exhibits a strong peak around 3–5 Å with a weaker secondary peak around 7–8 Å, for all three sets of DI rankings used. This double-peak structure is a characteristic feature of the DCA results. It is not observed in the background distribution of all residue pairs (see *SI Appendix*, Fig. S5, which has a single maximum around 20–25 Å). In Fig. 2*B*, this background is reflected by a small bump in the histograms for the top 20 and 30 DI ranking pairs. The two short-distance peaks are consistent with the biophysics of molecular contacts: The first peak presumably arises from short-ranged interactions like hydrogen bonding or pairings involved in secondary structure formation, whereas the second peak likely corresponds to long-ranged, possibly water-mediated contacts (28–30). The observation of this second, biologically reasonable peak in Fig. 2*B* also motivates the choice of 8 Å as a cutoff distance for what is considered a residue–residue contact in Figs. 1 and 2*A*.

To understand how many sequences are actually needed for mfDCA, we randomly generated subalignments for two protein families; see *SI Appendix*, Fig. S6. For at least these two families, an effective number of M_{eff} of approximately 250 is already sufficient to reach TP rates close to one for the top predicted residue pairs, and the predictive power increases monotonously when more sequences are available. These numbers are consistent with but slightly larger than the sequence requirements reported in ref. 31 for the statistical-coupling analysis originally proposed in ref. 5.

Long-Distance High-DI Residue Pairs. The results from the previous section illustrate the ability of mfDCA to identify intradomain contacts with high sensitivity. However, a small fraction of pairs showed high DI values (in the top 20–30 ranking) but were located far away according to the available crystal structure. Here we investigate various biological reasons for the appearance of such long-distance direct correlations.

Interdomain Residue Contacts. Given the biological role of some interdomain contacts (32), we studied if the appearance of long-distance high-DI pairs may be due to interactions between proteins which form oligomeric complexes, as described previously for the dimeric response regulators of the bacterial two-component signaling system (16). To further investigate this possibility, we examined members of the 131 proteins which formed homodimers or higher-order oligomers according to the corresponding X-ray crystal structures.

A first example is the ATPase domain of the family of the nitrogen regulatory protein C (NtrC)-like sigma54-dependent transcriptional activators (Pfam PF00158). Upon activation, different subunits of this domain are known to pack in the front-

[†] The best-predicted structures were used due to the variance in the quality of PDB structures. Also, for the number of cases where substantially different structures of the same protein exist in the PDB, the existence of a single structure containing the predicted contacts substantiates them as contacts of a native conformation of that protein.

[‡] Both DI and MI benefited modestly from sampling correction; see *SI Appendix*, Fig. S1 for a comparison of the performance of these methods with/without sampling correction.

to-back orientation to form a heptameric ring, wrapping DNA around the complex (33). We compared the DCA results to the structure of NtrC1 of *Aquifex aeolicus* (PDB ID 1NY6; ref. 33). Among the top 20 DI pairs, 17 were intradomain contacts. The three remaining pairs were long-distance (>10 Å) within the domain. Strikingly, all three were within 5 Å when paired with the closest position in an adjacent subunit of the heptamer complex; see Fig. 3. These pairs appear to have coevolved to maintain the proper formation of the heptamer complex. A second example of high-DI interdomain contact is shown in *SI Appendix*, Fig. S7 for the multidrug resistance protein MexA of *Pseudomonas aeruginosa*, where nine subunits oligomerize to form a funnel-like structure across the periplasmic space for antibiotic efflux (PDB ID 1VF7; ref. 34).

We further tested the occurrence of interdomain contacts at a global level. Out of the 131 studied domain families, 21 families feature X-ray crystal structures involving oligomers with predicted interdomain contacts (see *SI Appendix*, Table S3). Among the top 20 DI pairs that are not intradomain contacts, about half of them turned out to be interdomain contacts as shown in Fig. 3*D*.

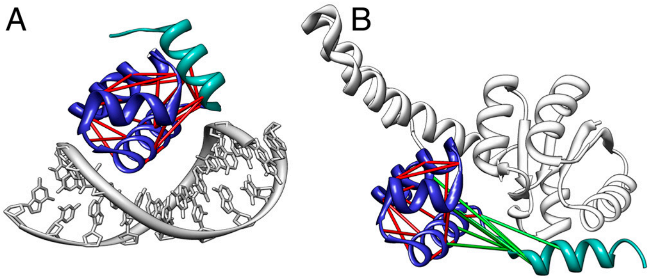

Fig. 4. The figures show the top 20 contacts predicted by DI for the family of response-regulator DNA-binding domain (GerE, PF00196) (containing both the dark- and light-blue colored regions). In *A*, the contacts are mapped to the DNA-binding domain of *E. coli* NarL, bound to the DNA target (PDB ID 1JE8). The TP rate for the top 20 DI pairs is 100%, and they are all shown as red links. In *B*, the contacts are mapped to the full-length response-regulator DosR of *M. tuberculosis* (PDB ID 3C3W), with the (unphosphorylated) response-regulator domain shown in gray. The top 20 DI pairings is only 65% in this case (13 red and 7 green links). The difference in prediction quality for the two structures can be traced back to a major reorientation of the C-terminal helix of the GerE domain (light blue) in *B*.

Alternative Domain Conformations. Another cause of long-distance high-DI pairs is the occurrence of alternative conformations for domains within the same family. As an illustration, we examine the domain family GerE (Pfam PF00196), whose members include the DNA-binding domains of many response regulators in two-component signaling systems.

Using the DNA-bound DNA-binding domain of the nitrate/nitrite response regulator NarL of *E. coli* (PDB ID 1JE8; ref. 35) as a structural template, we found that all of the top 20 DI pairs are true contacts (red bonds in Fig. 4*A*). However, when mapping the same DI pairs to the structure of the full-length transcriptional regulatory protein DosR of *Mycobacterium tuberculosis* (PDB ID 3C3W; ref. 36), seven pairs are found at distances >8 Å (green bonds in Fig. 4*B*, with the response-regulator domain shown in gray). Comparison of Fig. 4 *A* and *B* clearly shows that all of the green bonds involve pairing with the C-terminal helix (shown in light blue), which is significantly displaced in the full-length structure, presumably due to interaction with the (unphosphorylated) regulatory domain. As proposed by Wisedchaisri et al. (36), a likely scenario is that the DNA-binding domain of DosR is broken up by the interdomain interaction in the absence of phosphorylation, whereas phosphorylation of DosR restores its DNA-binding domain into the active form represented by the DNA-bound NarL structure.

It is difficult to estimate the extent to which alternative conformations may be responsible for the observed long-distance high-DI contacts, for less characterized domains for which alternative conformations may not be known. However, the example shown in Fig. 4 may motivate future studies to use these long-

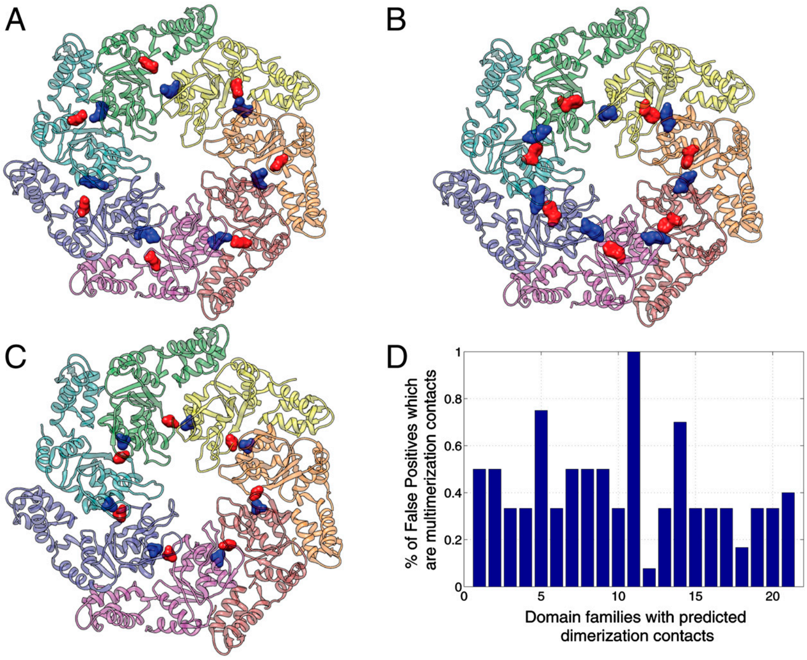
A
B
C
D
% of False Positives which are multimerization contacts
1
0.8
0.6
0.4
0.2
0
0
5
10
15
20
Domain families with predicted dimerization contacts

Fig. 3. The only three long-distance high-DI predictions found out of the top 20 DI pairs in the Sigma54 interaction domain of protein NtrC1 of *A. aeolicus* (PDB ID 1NY6) out of the top 20 predicted couplets are multimerization contacts. Structures showing each of these three interdomain contacts which are separated by less than 5 Å in a ring-like heptamer formed by Sigma54 interaction domains. (*A*) Residue pair GLU(174)-ARG(253), (*B*) residue pair PHE(226)-TYR (261), and (*C*) residue pair ALA(197)-ALA(249). (*D*) Oligomerization contacts are found in 21 structures of the 131 families studied (see *SI Appendix*, Table S3). These contacts represent a significant percentage of long-distance high-DI contacts observed in our predictions.

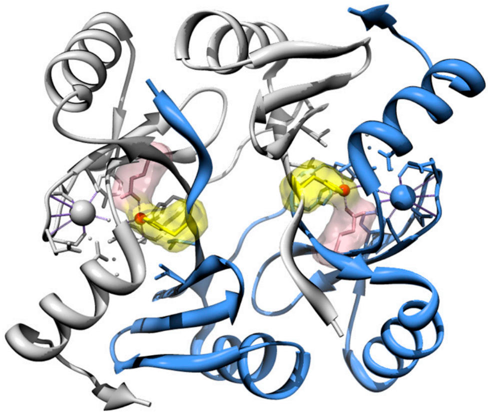

Fig. 5. The metalloenzyme domain (PF00903) of protein FosA (PDB ID 1NKI) is an example of a case where long-distance high-DI pairs are in fact residue pairs coordinating a ligand. The high-DI pair involving the residues Glu110 (pink) and His7 (yellow) coordinate a metal ion Mn(II) (red) in its dimer configuration. K^+ ions are shown as larger spheres (gray and blue), each coordinated by a monomer of the corresponding color.

distance high-DI contacts to explore possible alternative conformations.

Ligand-Mediated Interactions. Another special case of interdomain residue interactions and another cause of long-distance high-DI pairing is shown in Fig. 5. Here, mfDCA found the metalloenzyme domain family (PF00903) to have a high-DI intradomain residue pair which is separated by more than 14 Å when mapped to the glutathione transferase FosA of *P. aeruginosa* (PDB ID 1NKI; ref. 37). FosA is a metalloglutathione transferase which confers resistance to fosfomycin by catalyzing the addition of glutathione to fosfomycin. It is a homodimeric enzyme whose activity is dependent on Mn(II) and K^+, and the Mn(II) center has been proposed as part of the catalytic mechanism (37). We observed that the two residues belonging to the different subunits of the high-DI pair, Glu110 (pink) and His7 (yellow), are in direct contact (3 Å residue pair and 1.5 Å residue-ligand separation) with the Mn(II) ion (red) in the dimer configuration (Fig. 5). Thus, the "direct interaction" between these residues found by mfDCA is presumably mediated through their common interaction with a third agent, the metal ion in this case. There may well be other cases with interactions mediated by binding to other metabolites, RNA, DNA, or proteins not captured in the available crystal structures.

Contact Map Reconstruction. So far, we have focused on the top 20 DI pairs, which are largely intra- or interdomain contacts. However, one of the most striking features of the DI result in Fig. 2*A* is how gradually the average TP rate declines with increasing DI ranking. It is therefore possible to turn the question around: How many residue pairs are predicted, when we require a given minimum TP rate? For instance, one can go up to a DI ranking of 70 before the average TP rate declines to 70%, meaning that, if one were to predict contacts using the top 70 DI pairs, one would have obtained approximately 50 true contacts on average. This feature may be exploited for sequence-based structure prediction and deserves further analysis.

To become more quantitative, we define the number of acceptable pairs NAP_x as the (largest) number of DI-ranked pairs where the specified TP rate (x%) is reached for a given protein. NAP_x can be viewed as an index that characterizes the number of contact predictions at a certain acceptable quality level (given by x). We computed this index for every domain in all 856 structures in our database, for TP levels of 0.9, 0.8, and 0.7. The results are shown as cumulative distributions in Fig. 6. A casual inspection of these distributions shows that there are many structures with high NAP. Suppose the acceptable TP level is 0.7. The median of NAP_{70} is 52, meaning that, in half of the structures examined, the number of high-ranking, predictive DI pairs is at least 52. Furthermore, 70% of the structures have $\text{NAP}_{70} > 30$ and 34% of the structures have $\text{NAP}_{70} > 100$. A normalized version of Fig. 6 with respect to the length of the domain L is shown in *SI Appendix*, Fig. S8. In one extreme case involving the family of bacterial tripartite tricarboxylate receptors (PF03401), NAP_{70} was 600—i.e., 70% of the top 600 DI pairs correspond to true contacts when mapped to the best-predicted structure (PDB ID 2QPQ; ref. 38); see *SI Appendix*, Fig. S9*A*. This domain has a length of $L = 274$ and has approximately 2,300 contacts. In another example, the extracellular solute-binding family (PF00496) mapped to the structure of the periplasmic oligopeptide-binding protein OppA of *Salmonella typhimurium* (PDB ID 1JET; ref. 39) has a NAP_{70} of 497 (*SI Appendix*, Fig. S9*B*, $L = 372$, and approximately 2,530 contacts).

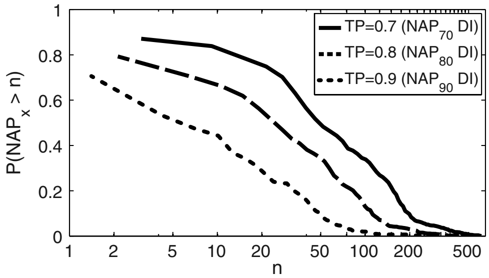

Fig. 6. Cumulative distribution of the number of acceptable pairs (NAP_x) for a given TP rate x. The curves show the probability of NAP_x to be larger than a given number n for contacts at given TP rates of 0.9, 0.8, and 0.7. The curves are computed for all 856 PDB structures in the dataset. We observe that the probability of $\text{NAP}_{70} > 30$ is 70% and $\text{NAP}_{70} > 100$ is 34%, which implies that a substantial number of protein domains can have accurate predictions that go beyond the top 30 DI pairings. We also identify some exceptional cases with $\text{NAP}_{70} > 600$.

We also computed the NAP_{70} distribution using MI; see *SI Appendix*, Fig. S10. The difference between DI and MI, about 10–20% in TP rate according to Fig. 2*A*, is seen much more significantly when displayed according to the NAP index, with the median NAP_{70} being 5 for MI and 52 for DI, which shows that DCA generates many more high-valued contact pair predictions. We also compared the performance of DCA with the approximate Bayesian method (red dashed curve in *SI Appendix*, Fig. S10), which gives a median NAP_{70} of 25 that is halfway between that of MI and mfDCA.

The large number of contacts correctly predicted by DCA prompted us to explore the extent to which DCA may be used to predict the contact maps of protein domains. For a domain with L amino acids, we calculated the inferred maps by sorting residue pairs according to their DI, and keeping the $2L$ highest-ranking pairs with minimum separation of five positions along the sequence. For the contact map prediction, we included further those pairings which have equal or larger DI than the ones mentioned above, but with shorter separation along the sequence because they may be informative about secondary structures. Fig. 7 shows two examples of such contact map predictions, for the prokaryotic promoter recognition domain of SigmaE already shown in Fig. 1 (PDB ID 1OR7, $L = 71$) and for the eukaryotic H-Ras protein (PDB ID 5P21; ref. 40, $L = 160$). The figure shows the native contact maps, together with the predictions by MI (Fig. 7, *Left*) and DI (Fig. 7, *Center*). Correctly predicted native contacts (i.e., the TPs) are indicated in red. The unpredicted native contacts taken from the X-ray crystal structures are shown in gray, and the incorrect predictions are shown in green. It is evident that, for both proteins, DI works substan-

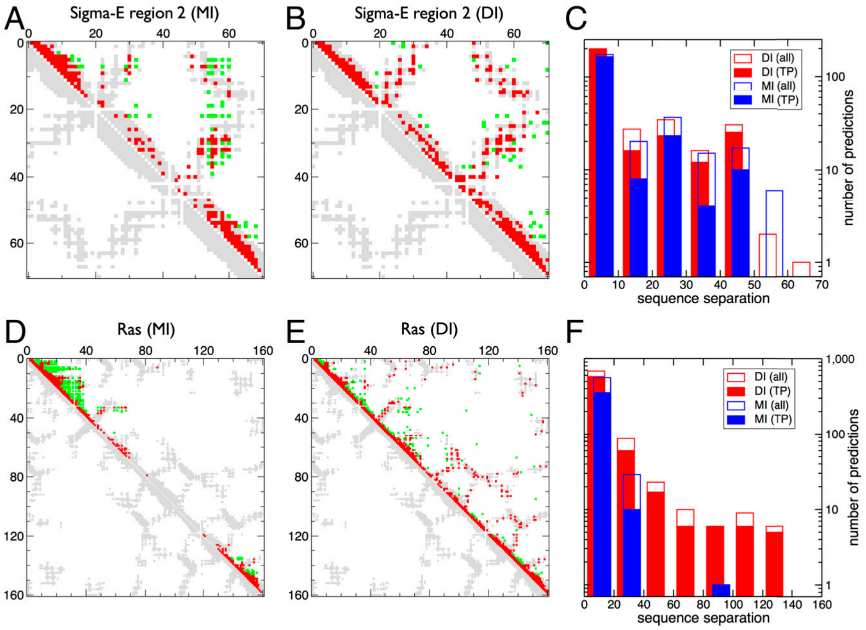

Fig. 7. Two examples of contact map predictions using MI (*A* and *D*) and mfDCA (*B* and *E*). Gray symbols represent the native map with a cutoff of 8 Å, colored symbols the computational contact predictions using MI or DI ranking (red squares for TP and green squares for spatially distant pairs). The number of pairs is determined such that there are 2*L* pairs with minimum separation five along the sequence, where *L* is the domain length. The right-most panels (*C* and *F*) bin the predictions of MI (blue) and mfDCA (red) according to their separation along the protein sequence. The overall bars count all predictions, the shaded part the TPs. Note in particular that mfDCA leads to a higher number of more accurate predictions for large separations. (*A*–*C*) The promoter recognition helix domain of the SigmaE factor (PDB ID 1OR7). (*D*–*F*) The eukaryotic signaling protein Ras (PDB ID 1P21). For better comparability of native vs. predicted contacts, the predictions are displayed only above the diagonal.

tially better than MI, both in terms of the TP rate and the representation of the native contact map. To become more quantitative, we have binned the predicted pairs according to their separation along the primary amino acid sequence (Fig. 7, *Right*). We observe that DI captures in particular a higher number and more accurately those contacts between residues, which are very distant along the sequence. Also, the DI predictions are more evenly distributed, whereas MI predictions tend to cluster together.

Discussion.

We have shown the ability of DCA to identify with high-accuracy residue pairs in domain families that might have coevolved together and hence are representative of physical proximity in the three-dimensional fold of the domain. We have done an extensive evaluation of these capabilities for a large number of families and individual PDB structures. We found that DCA is not only able to identify intradomain contacts but also interdomain residue pairs that are part of oligomerization interfaces. Although we focused on bacterial proteins, this methodology can be applied to the ever-increasing number of eukaryotic sequences. Our initial results suggest that mfDCA performance is conserved for nonbacterial proteins. One potential application is the identification of interaction interfaces for homodimers that could ultimately help in complex structure prediction, e.g., the cases in Fig. 3 and *SI Appendix*, Fig. S7. Our results might open unexplored avenues of research for which full contact maps could be estimated and used as input data for de novo protein structure identification, which is particularly interesting in the case of interdomain contacts in multidomain proteins. Ultimately, this methodology can be utilized with pairs of proteins rather than single proteins to identify potential protein–protein interactions. An example of this approach was introduced in ref. 16, however, the current mathematical formulation of the method as well as its computational implementation allows an analysis to a much larger scale.

Despite the accuracy of the extracted signal, mfDCA cannot be expected to extract all biological information contained in the pair correlations. This idea can be illustrated by comparing the mfDCA results to those of statistical-coupling analysis (SCA), developed by Lockless and Ranganathan (5) and used to identify "coevolving protein sectors" (41). We have applied mfDCA to the data of ref. 41 for the Trypsin protein family (Serine protease), where SCA identified three sectors related to different functionalities of the protein, which cover almost 30% of all residues. The mfDCA leads to an 83.3% TP rate for the top 30 contact predictions (PDB ID 3TGI; ref. 42)—i.e., to a performance which is comparable to the other protein families analyzed here. Out of the resulting 25 true contact pairs, only eight are found within the identified sectors. Among them, three are disulfide bonds (C42 : C58, C136 : C201, C191 : C220) and another two are inside a catalytic triad crucial for the catalytic activity of the protein family (H57 : S195, D102 : S195). The other 17 true contacts predicted by mfDCA are distributed over the protein fold, without obvious relation to the sectors (see *SI Appendix*, Table S4). The difference in prediction can be traced back to differences in the algorithmic approaches: SCA uses clustering to identify larger groups of coevolving sites (sectors), whereas DCA uses maximum-entropy modeling to extract pairs of directly coupled residues. Thus, the two algorithms extract different and, in both cases, biologically important information. It remains a future challenge to develop techniques unifying SCA and DCA, and to extract even more coevolutionary information from multiple-sequence alignments.

Methods

Data Extraction. Sequence datasets were extracted primarily from Pfam families with more than 1,000 nonredundant sequences. We decided to focus on families that are predominantly bacterial (i.e., more than 90% of the family sequences belong to bacterial organisms). Another requirement in this dataset is that such families must have at least two known X-ray crystal structures with a resolution of 3 Å or better. The PDB (43) was accessed to obtain crystal structures of proteins. An additional criterion to improve statistical significance when picking sequences that belong to a particular Pfam (22) family, was to use a stricter E-value threshold than the standard used by the software package HMMER (21) to classify domain membership. An in-house mapping application was developed to map domain family alignments and predicted couplets to specific residues in PDB structures. Some of the data extraction tools used in this study are described in more detail in ref. 17. A total of 131 families were selected that complied with all these criteria. A list of these Pfam families and the 856 PDB structures analyzed can be accessed in the *SI Appendix*, Tables S1 and S2).

For each family, the protein sequences are collected in one MSA denoted by $\{(A_1^a,\ldots,A_L^a)|\ a = 1,\ldots,M\}$, where L denotes the number of MSA columns (i.e., the length of the protein domains). Alignments are local alignments to the Pfam HMM; because of the large number of proteins in each MSA, we refrained from refinements using global alignment techniques.

Sequence Statistics and Reweighting. As already mentioned in *Results and Discussion*, the main inputs of DCA are reweighted frequency counts for single MSA columns and column pairs:

$$f_i(A) = \frac{1}{M_{\text{eff}}+\lambda}\left(\frac{\lambda}{q} + \sum_{a=1}^{M}\frac{1}{m^a}\delta_{A,A_i^a}\right)$$

$$f_{ij}(A,B) = \frac{1}{M_{\text{eff}}+\lambda}\left(\frac{\lambda}{q^2} + \sum_{a=1}^{M}\frac{1}{m^a}\delta_{A,A_i^a}\delta_{B,A_j^a}\right). \qquad [1]$$

In this equation, $\delta_{A,B}$ denotes the Kronecker symbol, which equals one if $A = B$, and zero otherwise. Furthermore, we have defined $q = 21$ for the number of different amino acids (also counting the gap), and a pseudocount λ (44), whose value will be discussed below. The weighting of the influence of a single sequence by the factor $1/m^a$ aims at correcting for the sampling bias. It is determined by the number

$$m^a = |\{b \in \{1,\ldots,M\}|\text{seqid}(A^a,A^b) > 80\%\}| \qquad [2]$$

of sequences $A^b = (A_1^b,\ldots,A_L^b)$, $b \in \{1,\ldots,M\}$, which have more than 80% sequence identity (seqid) with $A^a = (A_1^a,\ldots,A_L^a)$, where a itself is counted. The same reweighting, but with a 100% sequence-identity threshold, would remove multiple counts of repeated sequences. Reweighting systematically improves the results (see *SI Appendix*, Fig. S1), with only a weak dependence on the precise threshold value (in the range of 70–90%) and the specific protein family. Last, we introduced the effective sequence number $M_{\text{eff}} = \sum_{a=1}^{M} 1/m^a$ as the sum over all sequence weights. These counts allow for calculating the mutual information,

$$\text{MI}_{ij} = \sum_{A,B} f_{ij}(A,B)\ \ln\frac{f_{ij}(A,B)}{f_i(A)f_j(B)}, \qquad [3]$$

which equals zero if and only if i and j are uncorrelated, and is positive otherwise.

Maximum-Entropy Modeling. To disentangle direct and indirect couplings, we aim at inferring a statistical model $P(A_1,\ldots,A_L)$ for entire protein sequences $(A_1,\ldots,A_L)$. To achieve coherence with data, we require this model to generate the empirical frequency counts as marginals,

$$\forall\ i,A_i\colon \sum_{\{A_k|k\neq i\}} P(A_1,\ldots,A_L) \equiv f_i(A_i)$$

$$\forall\ i,j,A_i,A_j\colon \sum_{\{A_k|k\neq i,j\}} P(A_1,\ldots,A_L) \equiv f_{ij}(A_i,A_j). \qquad [4]$$

Besides this constraint, we aim at the most general, least-constrained model $P(A_1,\ldots,A_L)$. This model can be achieved by applying the maximum-entropy principle (45, 46), and it leads to an explicit mathematical form of $P(A_1,\ldots,A_L)$ as a Boltzmann distribution with pairwise couplings $e_{ij}(A,B)$ and local biases (fields) $h_i(A)$:

$$P(A_1,\ldots,A_L) = \frac{1}{Z}\ \exp\left\{\sum_{i<j} e_{ij}(A_i,A_j) + \sum_i h_i(A_i)\right\}. \qquad [5]$$

The model parameters have to be fitted such that [**4**] is satisfied. In this fitting procedure, one has to consider that Eq. **5** contains more free parameters than there are independent conditions in [**4**], which allows one to change couplings and fields together without changing the sum in the exponent. Therefore, multiple but equivalent solutions for the fitting are possible. To remove this freedom, we consider all couplings and fields measured relative to the last amino acid $A = q$, and set

$$\forall\ i,j,A\colon\ e_{ij}(A,\ q) = e_{ij}(q,A) = 0, \quad h_i(q) = 0. \qquad [6]$$

Details on the maximum-entropy approach are given in the *SI Appendix*.

Small-Coupling Expansion. Eq. **5** contains the normalization factor Z, in statistical physics also called the partition function, which is defined as

$$Z = \sum_{A_1,\ldots,A_L} \exp\left\{\sum_{i<j} e_{ij}(A_i,A_j) + \sum_i h_i(A_i)\right\} \qquad [7]$$

and includes a sum of q^L terms. Its direct calculation is infeasible for any realistic protein length and approximations have to be used. In a prior paper (16), several of us introduced a message-passing approach, which allows the treatment of about 70 MSA columns simultaneously in about 2-d running time on a standard desktop computer (larger MSAs need preprocessing to decrease the number of columns before running message passing). Here we introduce a much more efficient scheme, which for $L = 70$ is about 3–4 orders of magnitude faster, and which allows one to directly analyze alignments with $L \le 1{,}000$ ($L \le 500$ on a standard computer because of limited working memory). The total algorithmic complexity is $O(q^3N^3)$. The major speedup compared to the iterative message-passing solver results from the fact that parameter inference can be done in a single computational step in the new algorithm.

The approach is based on a small-coupling expansion (47, 48), which is explained in detail in the *SI Appendix*: The exponential of $\Sigma_{i<j}e_{ij}(A_i,A_j)$ in Eq. **7** is expanded into a Taylor series. Keeping only the linear order of this expansion, we obtain the well-known mean-field equations

$$\frac{f_i(A)}{f_i(q)} = \exp\left\{h_i(A) + \sum_A\sum_{j\neq i} e_{ij}(A,B)f_j(B)\right\}, \qquad [8]$$

containing the single-column counts, as well as a simple relation between the coupling $e_{ij}(A,B)$ and the pair counts $f_{ij}(A,B)$ for all $i,j = 1,\ldots,L$ and $A,B = 1,\ldots,q-1$

$$e_{ij}(A,B) = -(C^{-1})_{ij}(A,B) \qquad [9]$$

where

$$C_{ij}(A,B) = f_{ij}(A,B) - f_i(A)f_j(B). \qquad [10]$$

Eqs. **6** and **9** completely determine the couplings in terms of the data. Note that the connected-correlation matrix C defined in Eq. **10** is a $(q-1)L \times (q-1)L$ matrix; the pairs (i,A) and (j,B) have to be understood as joint single indices in the inversion in Eq. **9**.

In general, when constructed without pseudocounts ($\lambda = 0$), this matrix is not invertible, and formally Eq. **9** leads to infinite couplings. Even introducing site-specific reduced amino acid alphabets (only those actually observed in the corresponding MSA column) is found to be not sufficient for invertibility. The matrix can, however, be regularized by setting $\lambda > 0$. For small λ, elements diverging in the $\lambda \to 0$ limit dominate the DI calculation discussed in the next paragraph. To avoid such spurious high DI values, we have to go to relatively large pseudocounts; $\lambda = M_{\text{eff}}$ is found to be a reasonable value throughout families and is used exclusively in this paper. *SI Appendix*, Fig. S11 shows a sensitivity analysis for different values of the pseudocount for two domain families. The mean TP rates are computed for pseudocount values $\lambda = w \cdot M_{\text{eff}}$, with the weights w ranging from 0.11 to 9. The optimum value of λ is found for $1 \le w \le 1.5$. Therefore, we used $\lambda = M_{\text{eff}}$ throughout this study.

Because of the long run time of the message-passing approach (mpDCA), we could not compare its performance for all proteins studied in this paper. *SI Appendix*, Fig. S12 contains two examples: Trypsin (PF00089) and Trypsin inhibitor (PF00014). In both cases, mfDCA outperforms mpDCA. Furthermore, it is straightforward to include into DCA also the next order of the small-coupling expansion beyond the mean-field approximation (which corresponds to the so-called Thouless, Anderson, and Palmer (TAP) equations in spin-glass physics; ref. 49). We do not find any systematic improvement of the resulting algorithm, called tapDCA, when compared to mfDCA; see *SI Appendix*, Fig. S12.

Direct Information. After having estimated the direct coupling $e_{ij}(A,B)$ through Eq. **8**, we need a strategy for ranking the $L(L-1)$ possible interactions according to their direct-coupling strength. Following the idea that MI is a good measure for correlations, in ref. 16 we introduced a quantity called direct information. It can be understood as the amount of MI between columns i and j, which results from direct coupling alone.

To this end, we introduce for each column pair (i,j) an isolated two-site model

$$P_{ij}^{(\mathrm{dir})}(A,B) = \frac{1}{Z_{ij}} \exp\{e_{ij}(A,B) + \tilde{h}_i(A) + \tilde{h}_j(B)\}, \qquad [11]$$

where the couplings $e_{ij}(A,B)$ are computed using Eq. **8**, and the auxiliary fields $\tilde{h}$ are given implicitly by compatibility with the empirical single-residue counts:

$$f_i(A) = \sum_B P_{ij}^{(\mathrm{dir})}(A,B), \qquad f_j(B) = \sum_A P_{ij}^{(\mathrm{dir})}(A,B). \qquad [12]$$

As before, in order to reduce the number of free parameters to the number of independent constraints, these fields are required to fulfill $\tilde{h}_i(q) = \tilde{h}_j(q) = 0$. Note that the auxiliary fields have to be determined for each pair (i,j) independently to fit Eq. **12**. Finally, we define the DI as the MI of model

$$\mathrm{DI}_{ij} = \sum_{AB} P_{ij}^{(\mathrm{dir})}(A,B)\ \ln \frac{P_{ij}^{(\mathrm{dir})}(A,B)}{f_i(A)\, f_j(B)}. \qquad [13]$$

Algorithmic Implementation. The algorithmic implementation of the mean-field approximation is sketched in the following steps:

1. Estimate the frequency counts $f_i(A)$ and $f_{ij}(A,B)$ from the MSA, using the pseudocount $\lambda = M_{\mathrm{eff}}$ in Eqs. **1** and **2**.
2. Determine the empirical estimate of the connected-correlation matrix Eq. **10**.
3. Determine the couplings $e_{ij}(A,B)$ according to the second of Eq. **9**.
4. For each column pair $i < j$, estimate the direct information DI_{ij} by solving Eqs. **11** and **12** for $P_{ij}^{(\mathrm{dir})}(A,B)$, and plug the result into Eq. **13**.

An implementation of the code in Matlab is available upon request.

Note Added in Proof. Our direct-coupling analysis was recently used to infer all-atom protein 3D structures, indicating that the high quality of contact prediction reported here is capable of translating to good quality predicted 3D folds (50).

ACKNOWLEDGMENTS. We thank Hendrik Szurmant, Joanna Sulkowska, and Lucy Colwell for useful discussions during the course of this work. This work was supported by a European Commission Grant 267915 (to R.Z.), the Center for Theoretical Biological Physics sponsored by the National Science Foundation (NSF Grant PHY-0822283), and by NSF Grant MCB-1051438 (to J.N.O.).

1. Altschuh D, Lesk A, Bloomer A, Klug A (1987) Correlation of co-ordinated amino acid substitutions with function in viruses related to tobacco mosaic virus. *J Mol Biol* 193:693–707.
2. Göbel U, Sander C, Schneider R, Valencia A (1994) Correlated mutations and residue contacts in proteins. *Proteins Struct Funct Genet* 18:309–317.
3. Neher E (1994) How frequent are correlated changes in families of protein sequences? *Proc Natl Acad Sci USA* 91:98–102.
4. Shindyalov IN, Kolchanov NA, Sander C (1994) Can three-dimensional contacts in protein structures be predicted by analysis of correlated mutations? *Protein Eng* 7:349–358.
5. Lockless SW, Ranganathan R (1999) Evolutionarily conserved pathways of energetic connectivity in protein families. *Science* 286:295–299.
6. Atchley WR, Wollenberg KR, Fitch WM, Terhalle W, Dress AW (2000) Correlations among amino acid sites in bHLH protein domains: An information theoretic analysis. *Mol Biol Evol* 17:164–178.
7. Fodor AA, Aldrich RW (2004) Influence of conservation on calculations of amino acid covariance in multiple sequence alignments. *Proteins Struct Func Bioinf* 56:211–221.
8. Liu Z, Chen J, Thirumalai D (2009) On the accuracy of inferring energetic coupling between distant sites in protein families from evolutionary imprints: Illustrations using lattice model. *Proteins Struct Func Bioinf* 77:823–831.
9. Lashuel HA, Pappu R (2009) Amyloids go genomic: Insights regarding the sequence determinants of prion formation from genome-wide studies. *Chembiochem* 10:1951–1954.
10. Burger L, van Nimwegen E (2010) Disentangling direct from indirect co-evolution of residues in protein alignments. *PLoS Comput Biol* 6:e1000633.
11. Liolios K, et al. (2009) The Genomes On Line Database (GOLD) in 2009: Status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Res* 38:D346–D354.
12. Hoch JA (2000) Two-component and phosphorelay signal-transduction. *Curr Opin Microbiol* 3:165–170.
13. Ulrich LE, Zhulin IB (2009) The MiST2 database: A comprehensive genomics resource on microbial signal transduction. *Nucleic Acids Res* 38:D401–D407.
14. White RA, Szurmant H, Hoch JA, Hwa T (2007) Features of protein-protein interactions in two-component signaling deduced from genomic libraries. *Methods Enzymol* 422:75–101.
15. Skerker JM, et al. (2008) Rewiring the specificity of two-component signal transduction systems. *Cell* 133:1043–1054.
16. Weigt M, White RA, Szurmant H, Hoch JA, Hwa T (2009) Identification of direct residue contacts in protein–protein interaction by message passing. *Proc Natl Acad Sci USA* 106:67–72.
17. Lunt B, et al. (2010) Inference of direct residue contacts in two-component signaling. *Methods Enzymol* 471:17–41.
18. Schug A, Weigt M, Onuchic JN, Hwa T, Szurmant H (2009) High-resolution protein complexes from integrating genomic information with molecular simulation. *Proc Natl Acad Sci USA* 106:22124–22129.
19. Schug A, Weigt M, Hoch J, Onuchic J (2010) Computational modeling of phosphotransfer complexes in two-component signaling. *Methods Enzymol* 471:43–58.
20. Procaccini A, Lunt B, Szurmant H, Hwa T, Weigt M (2011) Dissecting the specificity of protein-protein interaction in bacterial two-component signaling: Orphans and crosstalks. *PLoS One* 6:e19729.
21. Eddy SR (1998) Profile hidden Markov models. *Bioinformatics* 14:755–763.
22. Finn RD, et al. (2010) The Pfam protein families database. *Nucleic Acids Res* 38: D211–D222.
23. Wollenberg KR, Atchley WR (2000) Separation of phylogenetic and functional associations in biological sequences by using the parametric bootstrap. *Proc Natl Acad Sci USA* 97:3288–3291.
24. Tillier ERM, Lui TWH (2003) Using multiple interdependency to separate functional from phylogenetic correlations in protein alignments. *Bioinformatics* 19:750–755.
25. Gouveia-Oliveira R, Pedersen AG (2007) Finding coevolving amino acid residues using row and column weighting of mutual information and multi-dimensional amino acid representation. *Algorithms Mol Biol* 2:12–24.
26. Dunn SD, Wahl LM, Gloor GB (2007) Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *Bioinformatics* 24:333–340.
27. Campbell E, et al. (2003) Crystal structure of *Escherichia coli* sigmaE with the cytoplasmic domain of its anti-sigma RseA. *Mol Cell* 11:1067–1078.
28. Tanaka S, Scheraga HA (1976) Medium- and long-range interaction parameters between amino acids for predicting three-dimensional structures of proteins. *Macromolecules* 9:945–950.
29. Go N, Taketomi H (1978) Respective roles of short- and long-range interactions in protein folding. *Proc Natl Acad Sci USA* 75:559–563.
30. Miyazawa S, Jernigan RL (2003) Long- and short-range interactions in native protein structures are consistent/minimally frustrated in sequence space. *Proteins Struct Funct Genet* 50:35–43.
31. Dima RI, Thirumalai D (2006) Determination of network of residues that regulate allostery in protein families using sequence analysis. *Protein Sci* 15:258–268.
32. Myers RS, Amaro RE, Luthey-Schulten ZA, Davisson VJ (2005) Reaction coupling through interdomain contacts in imidazole glycerol phosphate synthase. *Biochemistry* 44:11974–11985.
33. Lee S-Y (2003) Regulation of the transcriptional activator NtrC1: Structural studies of the regulatory and AAA+ ATPase domains. *Genes Dev* 17:2552–2563.
34. Akama H, et al. (2004) Crystal structure of the membrane fusion protein, MexA, of the multidrug transporter in *Pseudomonas aeruginosa. J Biol Chem* 279:25939–25942.
35. Maris AE, et al. (2002) Dimerization allows DNA target site recognition by the NarL response regulator. *Nat Struct Biol* 9:771–778.
36. Wisedchaisri G, Wu M, Sherman DR, Hol WGJ (2008) Crystal structures of the response regulator DosR from *Mycobacterium tuberculosis* suggest a helix rearrangement mechanism for phosphorylation activation. *J Mol Biol* 378:227–242.
37. Rigsby RE, Rife CL, Fillgrove KL, Newcomer ME, Armstrong RN (2004) Phosphonoformate: A minimal transition state analogue inhibitor of the fosfomycin resistance protein, FosA. *Biochemistry* 43:13666–13673.
38. Herrou J, et al. (2007) Structure-based mechanism of ligand binding for periplasmic solute-binding protein of the Bug family. *J Mol Biol* 373:954–964.
39. Tame JRH, Sleigh SH, Wilkinson AJ, Ladbury JE (1996) the role of water in sequence independent ligand binding by an oligopeptide transporter protein. *Nat Struct Biol* 3:998–1001.

40. Pai EF, et al. (1990) Refined crystal structure of the triphosphate conformation of H-ras p21 at 1.35 A resolution: Implications for the mechanism of GTP hydrolysis. *EMBO J* 9:2351–2359.

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

42. Pasternak A, Ringe D, Hedstrom L (1999) Comparison of anionic and cationic trypsinogens: The anionic activation domain is more flexible in solution and differs in its mode of BPTI binding in the crystal structure. *Protein Sci* 8:253–258.

43. Berman HM, et al. (2000) The Protein Data Bank. *Nucleic Acids Res* 28:235–242.

44. Durbin R, Eddy S, Krogh A, Mitchison G (1998) *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids* (Cambridge Univ Press, New York), pp 319–321.

45. Jaynes ET (1957) Information theory and statistical mechanics. *Phys Rev* 106:620–630.

46. Jaynes ET (1957) Information theory and statistical mechanics. II. *Phys Rev* 108:171–190.

47. Plefka T (1982) Convergence condition of the TAP equation for the infinite-ranged Ising spin glass model. *J Phys A Math Gen* 15:1971–1978.

48. Georges A, Yedidia J (1991) How to expand around mean-field theory using high-temperature expansions. *J Phys A Math Gen* 24:2173–2192.

49. Thouless DJ, Anderson PW, Palmer RG (1977) Solution of "Solvable model of a spin glass". *Philos Mag* 35:593–601.

50. Marks DS, et al. 3D protein structure predicted from sequence., arXiv:1110.5091v2 [q-bio.BM].

Supplementary text: Direct-coupling analysis of residue co-evolution captures native contacts across many protein families

F. Morcos, A. Pagnani, B. Lunt, A. Bertolino, D. Marks, C. Sander, R. Zecchina, J.N. Onuchic, T. Hwa, and M. Weigt

I. INPUT DATA

Data are given as a multiple sequence alignment (MSA), i.e. a rectangular array with entries coming from a 21-letter alphabet (20 amino acids, 1 gap):

$$\mathbf{A} = (A_i^a)\,, \quad i = 1, ..., L, \quad a = 1, ..., M \tag{1}$$

with L being the number of residues in each MSA row (the protein length), and M the number of MSA rows (the number of proteins). For simplicity of notation we assume that the $q = 21$ amino acids are translated into consecutive numbers 1,...,q.

II. SEQUENCE STATISTICS

The aim of the analysis is to detect statistical coupling between the amino-acid occupancies of any two columns of the MSA $\mathbf{A}$. For doing so, we first introduce single site and pair frequency counts,

$$f_i(A) = \frac{1}{M}\sum_{a=1}^{M} \delta_{A,A_i^a}; \quad f_{ij}(A,B) = \frac{1}{M}\sum_{a=1}^{M} \delta_{A,A_i^a}\delta_{B,A_j^a}, \tag{2}$$

with $1 \leq i,j \leq L$, $1 \leq A,B \leq q$, and δ denoting the Kronecker symbol, which equals one if the two indices coincide, and zero else. The first count determines the fraction of proteins which show amino acid A in column i (residue position), the second one the fraction of MSA rows where amino acids A and B co-appear in positions i and j.

A. Reweighted frequency counts

These simple frequency counts represent faithfully the statistical properties of the MSA if and only if rows are drawn independently from the same distribution. Biological sequence data show a strong sampling bias due phylogenetic relations between species, due to the sequencing of different strains of the same species, and due to a bias in the selection of species which are currently sequenced. As a simple correction, we use a reweighting scheme, which we have introduced in [1, 2].

First, we define a similarity threshold $0 < x < 1$: Two sequences of identity (number of positions with coinciding amino acids) larger than xL are considered to carry almost the same information, smaller sequence identities are considered to carry substantially independent information. In practical tests we have found that values of x around 0.7-0.9 lead to very similar results, we use $x = 0.8$.

Second, for each sequence $A^a = (A_1^a, ..., A_L^a)$ we determine the number of similar sequences $A^b = (A_1^b, ..., A_L^b)$ via

$$m^a = \left|\left\{ b \mid 1 \leq b \leq M,\ \mathrm{seqid}(A^a, A^b) \geq xL \right\}\right| \ . \tag{3}$$

Note that this count is always at least one, since sequence A^a is counted itself in m^a. For each sequence, we use the weight $1/m^a$ in the frequency counts, i.e., sequences without similar sequences take weight one, and sequences featuring similar sequences are down-weighted. We redefine the frequency counts as

$$\begin{aligned} f_i(A) &= \frac{1}{\lambda + M_{eff}}\left(\frac{\lambda}{q} + \sum_{a=1}^{M}\frac{1}{m^a}\delta_{A,A_i^a}\right) \qquad (4)\\ f_{ij}(A,B) &= \frac{1}{\lambda + M_{eff}}\left(\frac{\lambda}{q^2} + \sum_{a=1}^{M}\frac{1}{m^a}\delta_{A,A_i^a}\delta_{B,A_j^a}\right) \ . \end{aligned}$$

This equation also contains a pseudo-count λ, which is a standard tool in estimating probabilities from counts in biological sequence analysis [3]. It serves to regularize parameters in the case of insufficient data availability, and has an interpretation in terms of Bayesian inference. The total weight of all sequences, $M_{eff} = \sum_{a=1}^{M} 1/m^a$, can be understood as the effective number of independent sequences.

Note that using $x = 1$ would reweight each sequence by the number of times it appears in the MSA, removing thus simple repeats. Lower values for x aim at giving a smaller weight to regions which are more densely sampled, and a higher weight to regions which are less densely sampled.

B. Mutual information as a correlation measure

If two MSA columns i and j were statistically independent, the joint distribution $f_{ij}(A,B)$ would factorize into $f_i(A) \times f_j(B)$, any deviation from this factorization signals correlations between the columns. Such correlation can be quantified by the mutual information

$$MI_{ij} = \sum_{A,B} f_{ij}(A,B) \ln \frac{f_{ij}(A,B)}{f_i(A)f_j(B)} \ . \tag{5}$$

It equals zero if and only if $f_{ij}(A,B)$ factorizes into the single marginals, and it is positive whenever $f_{ij}(A,B)$ does not factorize.

III. MAXIMUM-ENTROPY MODELING

As discussed in the main text, inter-column correlation may be caused by direct statistical coupling, but also by indirect correlation effects via intermediate MSA columns. As shown in [1], such direct and indirect effects may be disentangled: The idea is to infer a global statistical model $P(A_1, ..., A_L)$ for entire amino-acid sequences of the protein domain under study. This model has to be coherent to the empirical data, i.e. to generate the empirical single- and two-site frequency counts:

$$\begin{aligned} P_i(A_i) &= \sum_{\{A_k|k\neq i\}} P(A_1,...,A_L) = f_i(A_i) \\ P_{ij}(A_i,A_j) &= \sum_{\{A_k|k\neq i,j\}} P(A_1,...,A_L) = f_{ij}(A_i,A_j) \ . \end{aligned} \qquad (6)$$

Beyond these constraints, we aim at the most general, i.e. least constrained model $P(A_1, ..., A_L)$. It can be determined using the distribution maximizing the entropy

$$S = -\sum_{\{A_i|i=1,...,L\}} P(A_1,...,A_L) \ln P(A_1,...,A_L) \qquad (7)$$

while satisfying the constraints in Eqs. (6). The solution to this optimization problem is standard [4]: after introducing constraints via Lagrange multipliers, we find the analytical form of the distribution:

$$P(A_1,...,A_L) = \frac{1}{Z} \exp\left\{ \sum_{i<j} e_{ij}(A_i,A_j) + \sum_i h_i(A_i) \right\} \ . \qquad (8)$$

The Lagrange multipliers $h_i(A)$ and $e_{ij}(A, B)$ have a simple interpretation in terms of local amino-acid biases (local fields in statistical-physics language) and statistical residue couplings (coupling strength in statistical-physics language). Their numerical values have to be tuned such that the constraints given by Eqs. (6) are respected. The normalization constant

$$Z = \sum_{\{A_i|i=1,...,L\}} \exp\left\{ \sum_{i<j} e_{ij}(A_i,A_j) + \sum_i h_i(A_i) \right\} \qquad (9)$$

is called *partition function* in statistical physics. For later convenience, we also introduce the *Hamiltonian*

$$\mathcal{H} = -\sum_{1\leq i<j\leq L} e_{ij}(A_i,A_j) - \sum_{i=1}^{L} h_i(A_i) \ , \qquad (10)$$

such that our probabilistic model reads $P(A_1, ..., A_L) = \exp\{-\mathcal{H}\}/Z$.

The major problem in this context is the determination of the marginal distributions $P_i(A)$ and $P_{ij}(A, B)$ from $P(A_1, ..., A_L)$. Doing this exactly by tracing over all other variables A_i as written in Eqs. (6) would require an exponential time, which grows like q^L with the length of the aligned proteins. Different strategies have already been suggested for tackling this problem (most of them for the restricted Ising model having $q = 2$): In [1] we used a message-passing algorithm originally proposed in [5], [6] uses improved Monte Carlo sampling, [7–9] suggest perturbative expansion schemes, whereas [10] uses pseudo-likelihoods decoupling inference for different sites. For an overview over the relative performance of these algorithms on artificial data see [11].

It is important to note that the partition function itself contains all necessary information on the marginals, in particular we have

$$\begin{aligned} \frac{\partial \ln Z}{\partial h_i(A)} &= -P_i(A) \\ \frac{\partial^2 \ln Z}{\partial h_i(A)\, \partial h_j(B)} &= -P_{ij}(A,B) + P_i(A)\, P_j(B) \ . \end{aligned} \qquad (11)$$

For later convenience we introduce the connected correlations

$$C_{ij}(A,B) = P_{ij}(A,B) - P_i(A)\, P_j(B) \ , \qquad (12)$$

where indices i, j run from $1, ..., L$, whereas A, B from $1, .., q - 1$. The significance of excluding $A, B = q$ will become clear below. Note that we will consider $C_{ij}(A, B)$ as a $L(q-1) \times L(q-1)$-dimensional matrix, i.e. each pair (i, A) is interpreted as a parametrization of a single, joint index.

A. The number of independent parameters

The statistical model in Eq. (8) has $\binom{N}{2}q^2 + Nq$ parameters, but not all of them are independent. In fact, the consistency conditions in Eqs. (6) are also not independent, since the single-site marginals are implied by the two-site marginals, and all distributions are normalized. Careful inspections unveils $\binom{N}{2}(q-1)^2 + N(q-1)$ independent consistency conditions. We may therefore fix a part of the parameters in Eq. (8). Without loss of generality, we set

$$e_{ij}(A,q) = e_{ij}(q,A) = h_i(q) = 0 \qquad (13)$$

for all $i, j = 1, .., L$ and $A = 1, ...q$. Intuitively, this corresponds to a situation where all couplings and biases are measured with respect to the state q. The number of remaining parameters matches now the number of constraints, and the solution of the maximum-entropy model is unique.

B. Small-coupling expansion

The algorithmic approach is based on a systematic small-coupling expansion, i.e., on a Taylor expansion around zero coupling. This expansion was introduced in [12] by Plefka for disordered Ising models (Ising spin-glasses, corresponding to binary variables with $q = 2$). A more elegant derivation was proposed Georges and Yedidia [13], we generalize their approach to the case of Potts models with $q > 2$.

First we introduce the perturbed Hamiltonian

$$\mathcal{H}(\alpha) = -\alpha \sum_{1\leq i<j\leq L} e_{ij}(A_i,A_j) - \sum_{i=1}^{L} h_i(A_i) \ , \qquad (14)$$

depending on the additional parameter α. This parameter allows to interpolate between independent variables for $\alpha = 0$, and the original model for $\alpha = 1$. Furthermore we introduce the so-called *Gibbs potential*

$$-\mathcal{G}(\alpha) = \ln\left[\sum_{\{A_i|i=1,\dots,L\}} e^{-\mathcal{H}(\alpha)}\right] - \sum_{i=1}^{L}\sum_{B=1}^{q-1} h_i(B)P_i(B) \tag{15}$$

as the Legendre transform of the *free energy* $\mathcal{F} = -\ln Z$. Whereas the free energy depends canonically on the couplings and the fields, the Gibbs potential depends on the couplings and the marginal single-site distributions $P_i(A)$, i.e.

$$\mathcal{G}(\alpha) = \mathcal{G}\left(\{\alpha e_{ij}(A,B)\}_{1\le i<j\le L}^{A,B=1,..,q-1}, \{P_i(A)\}_{i=1,\dots,L}^{A=1,\dots,q-1}\right) . \tag{16}$$

This choice is particularly practical for the following derivation, since it guarantees the first of Eqs. (6) to be valid at any α. Note that the Potts variables in this expression run only up to $q-1$. Due to the gauge of the couplings and the normalization of the marginals, values for $A, B = q$ are not independent variables.

The fields can be found via the standard expression for Legendre transforms, cf. Eq. (11),

$$h_i(A) = \frac{\partial \mathcal{G}(\alpha)}{\partial P_i(A)} \; , \tag{17}$$

and

$$\left(C^{-1}\right)_{ij}(A,B) = \frac{\partial h_i(A)}{\partial P_j(B)} = \frac{\partial^2 \mathcal{G}(\alpha)}{\partial P_i(A)\,\partial P_j(B)} \; . \tag{18}$$

It is worth pointing out that the previous relations hold at any value of α and are a consequence of the functional form of the Legendre transform defined in Eq. (15). We remind that the matrix C was defined in Eq. (12) to have dimension $L(q-1)$, i.e. Potts-state indices are constrained to values up to $q-1$. This restriction makes C an invertible matrix (at least for non-zero pseudo-count λ), removing trivial linear dependencies resulting from the normalization of P_{ij}. Using this last equation, we can calculate the two-point marginal distributions P_{ij} directly from the Gibbs potential by means of two partial derivations and one matrix inversion.

Our aim is to expand this Gibbs potential up to first order in α around the independent-site case $\alpha = 0$,

$$\mathcal{G}(\alpha) = \mathcal{G}(0) + \left.\frac{d\mathcal{G}(\alpha)}{d\alpha}\right|_{\alpha=0} \alpha + \mathcal{O}(\alpha^2) \; . \tag{19}$$

In the following subsections, we calculate the still unknown terms on the right-hand side of this equations, i.e. the Gibbs potential and its first derivative in $\alpha = 0$.

C. Independent-site approximation

To start with, let us consider the Gibbs potential in $\alpha = 0$. In this case, the Gibbs potential equals the negative entropy of an ensemble of L uncoupled Potts spins $A_1, ..., A_L$ of given marginals $P_i(A_i)$. This claim results from basic statistical mechanics: The free energy equals the average energy (average Hamiltonian) minus the entropy. For $\alpha = 0$, the Legendre transform removes the complete average energy.

However, the entropy of uncoupled spins of given distribution is known to be

$$\begin{aligned}\mathcal{G}(0) &= \sum_{i=1}^{L}\sum_{A=1}^{q} P_i(A)\ln P_i(A) \\ &= \sum_{i=1}^{L}\sum_{A=1}^{q-1} P_i(A)\ln P_i(A) \\ &\quad + \sum_{i=1}^{L}\left[1-\sum_{A=1}^{q-1}P_i(A)\right]\ln\left[1-\sum_{A=1}^{q-1}P_i(A)\right] \; ;\end{aligned} \tag{20}$$

the last line eliminates terms in $P_i(q)$ and reduces the expression to the independent variables.

D. Mean-field approximation

To get the first order in Eq. (19), we have to determine $d\mathcal{G}(\alpha)/d\alpha$ in $\alpha = 0$. Recalling the definition of the Gibbs potential in Eq. (15), we write

$$\begin{aligned}\frac{d\mathcal{G}(\alpha)}{d\alpha} &= -\frac{d}{d\alpha}\ln Z(\alpha) - \sum_{i=1}^{L}\sum_{A=1}^{q-1}\frac{dh_i(A)}{d\alpha}P_i(A) \\ &= -\sum_{\{A_i\}}\left[\sum_{i<j} e_{ij}(A_i,A_j) + \sum_i \frac{dh_i(A)}{d\alpha}\right]\frac{e^{-\mathcal{H}(\alpha)}}{Z(\alpha)} \\ &\quad -\sum_{i=1}^{L}\sum_{A=1}^{q-1}\frac{dh_i(A)}{d\alpha}P_i(A) \\ &= -\left\langle \sum_{i<j} e_{ij}(A_i,A_j)\right\rangle_{\alpha} \; .\end{aligned} \tag{21}$$

The first derivative of the Gibbs potential with respect to α equals thus the average of the coupling term in the Hamiltonian. At $\alpha = 0$, this average can be done easily, since the joint distribution of all variables becomes factorized over the single sites,

$$\left.\frac{d\mathcal{G}(\alpha)}{d\alpha}\right|_{\alpha=0} = -\sum_{i<j}\sum_{A,B} e_{ij}(A,B)\,P_i(A)\,P_j(B) \; . \tag{22}$$

Plugging this and Eq. (20) into Eq. (19), we find the first-order approximation of the Gibbs potential. First and second partial derivatives with respect to the marginal distributions $P_i(A)$ provide self-consistent equations for the local fields,

$$\frac{P_i(A)}{P_i(q)} = \exp\left\{h_i(A) + \sum_{\{j|j\neq i\}}\sum_{B=1}^{q-1} e_{ij}(A,B)\,P_j(B)\right\} \tag{23}$$

and the inverse of the connected correlation matrix,

$$\left(C^{-1}\right)_{ij}(A,B)\Big|_{\alpha=0} = \begin{cases} -e_{ij}(A,B) & \text{for } i \neq j \\ \frac{\delta_{A,B}}{P_i(A)} + \frac{1}{P_i(q)} & \text{for } i = j \end{cases} . \quad (24)$$

This last equation allows for solving the original inference problem in mean-field approximation in a single step, without resorting to iterative schemes like gradient decent. Since we want to fit one- and two-site marginal of $P(A_1, ..., A_L)$ to the empirical values $f_i(A)$ and $f_{ij}(A,B)$ derived from the original protein MSA, we just need to determine the empirical connected correlation matrix

$$C_{ij}^{(emp)}(A,B) = f_{ij}(A,B) - f_i(A)\, f_j(B) \quad (25)$$

and invert this matrix to get the couplings e_{ij}. Even if matrix inversion is of complexity $\mathcal{O}(L^3)$ and thus of the same complexity as susceptibility propagation, the mean-field approximation is found to be $10^3 - 10^4$ times faster. This results from the simple fact that $> 10^3$ iteration are needed in susceptibility propagation to reach sufficient precision in fitting the empirical data by the maximum-entropy model.

IV. DIRECT INFORMATION AS A DIRECT-COUPLING MEASURE

Given the estimate of the pair couplings $e_{ij}(A,B)$ we would like to rank residue pairs according to their interaction strength. To do so, we need a meaningful mapping from the $(q-1) \times (q-1)$-dimensional coupling matrices to a single scalar parameter. A way to do this which is independent of the selected gauge, was already proposed in [1]. The quantity introduced there was called *direct information* (DI) and measures the mutual information due to the direct coupling. To do so, we isolate a pair i, j of positions and introduce a two-site model

$$P_{ij}^{(dir)}(A,B) = \frac{1}{Z_{ij}} \exp\left\{ e_{ij}(A,B) + \tilde{h}_i(A) + \tilde{h}_j(B) \right\} \quad (26)$$

with the coupling being the one inferred before. The new fields $\tilde{h}_{i/j}$ are determined by imposing the empirical single-site frequency counts as marginal distributions,

$$\begin{aligned} f_i(A) &= \sum_{B=1}^{q} P_{ij}^{(dir)}(A,B) \\ f_j(B) &= \sum_{A=1}^{q} P_{ij}^{(dir)}(A,B) \ , \end{aligned} \quad (27)$$

and Z_{ij} follows by normalization. The direct information is the mutual information associated to $P_{ij}^{(dir)}$:

$$DI_{ij} = \sum_{A,B=1}^{q} P_{ij}^{(dir)}(A,B) \ln \frac{P_{ij}^{(dir)}(A,B)}{f_i(A)\, f_j(B)} \ . \quad (28)$$

In this expression, any indirect effect is obviously removed, only the strength of the direct coupling $e_{ij}(A,B)$ is measured.

[1] M. Weigt, R.A. White, H. Szurmant, J.A. Hoch, and T. Hwa. Identification of direct residue contacts in protein-protein interaction by message passing. *Proceedings of the National Academy of Sciences*, 106(1):67–72, 2009.

[2] A. Procaccini, B. Lunt, H. Szurmant, T. Hwa, and M. Weigt. Dissecting the specificity of protein-protein interaction in bacterial two-component signaling: Orphans and crosstalks. *PLoS ONE*, 6(5):e19729, 05 2011.

[3] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison. *Biological sequence analysis: Probabilistic models of proteins and nucleic acids.* Cambridge Univ Pr, 1998.

[4] D.J.C. MacKay. *Information theory, inference, and learning algorithms.* Cambridge Univ Pr, 2003.

[5] M. Mézard and T. Mora. Constraint satisfaction problems and neural networks: A statistical physics perspective. *Journal of Physiology-Paris*, 103(1-2):107 – 113, 2009. Neuromathematics of Vision.

[6] E. Schneidman, M.J. Berry, R. Segev, and W. Bialek. Weak pairwise correlations imply strongly correlated network states in a neural population. *Nature*, 440(7087):1007–1012, 2006.

[7] S. Cocco, S. Leibler, and R. Monasson. Neuronal couplings between retinal ganglion cells inferred by efficient inverse statistical physics methods. *Proceedings of the National Academy of Sciences*, 106(33):14058–14062, 2009.

[8] V. Sessak and R. Monasson. Small-correlation expansions for the inverse Ising problem. *Journal of Physics A: Mathematical and Theoretical*, 42:055001, 2009.

[9] H. Kappen and F.B. Rodriguez. Efficient learning in boltzmann machines using linear response theory. *Neural Computation*, 10:1137, 1998.

[10] P. Ravikumar, M.J. Wainwright, and J.D. Lafferty. High-dimensional ising model selection using l1-regularized logistic regression. *Annals of Statistics*, 38:1287, 2010.

[11] Y. Roudi, J.A. Hertz, and E. Aurell. Statistical physics of pairwise probability models. *Front. Comput. Neurosci.*, 3:22, 2009.

[12] T. Plefka. Convergence condition of the tap equation for the infinite-ranged ising spin glass model. *Journal of Physics A: Mathematical and General*, 15(6):1971, 1982.

[13] A. Georges and J.S. Yedidia. How to expand around mean-field theory using high-temperature expansions. *Journal of Physics A: Mathematical and General*, 24(9):2173, 1991.

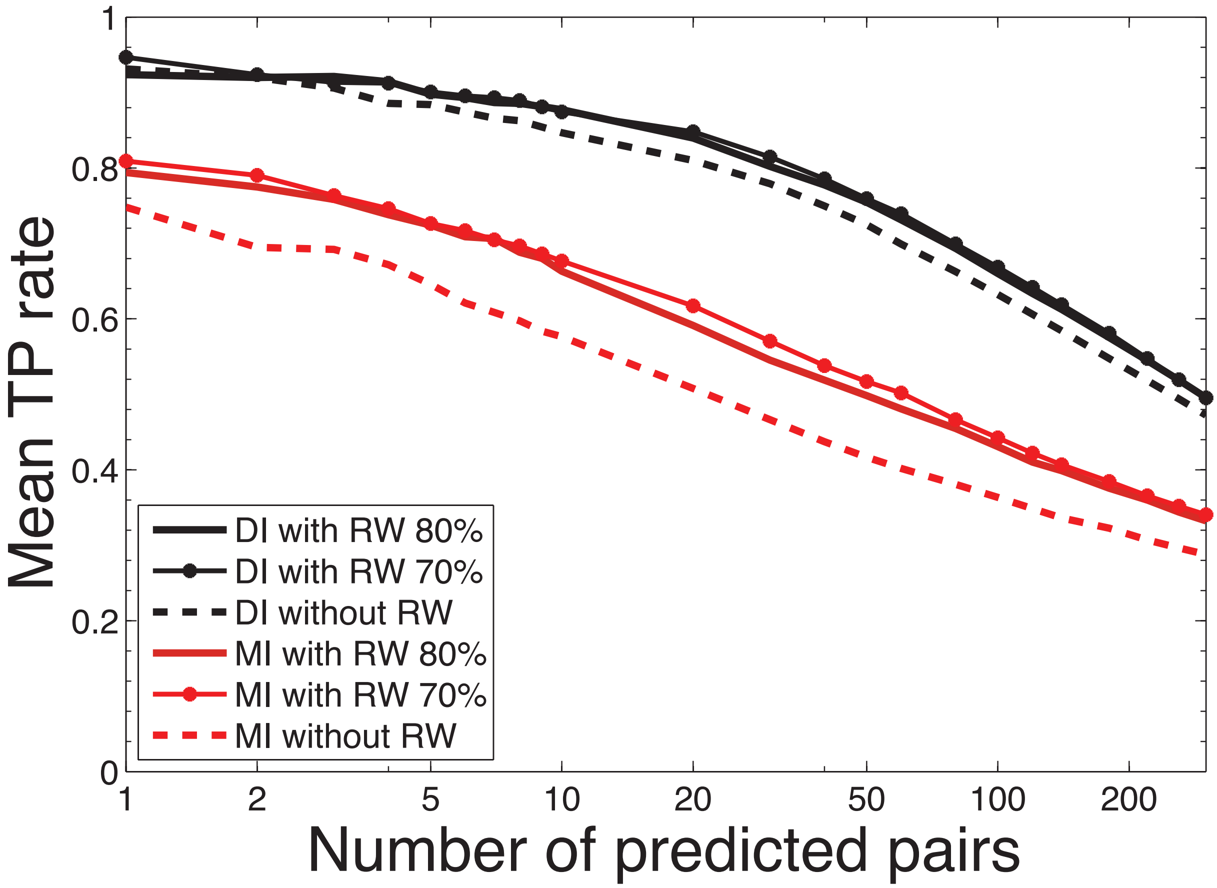

Figure S1. Mean prediction performance for 131 domain families with respect to the top number of ranked contacts. The effect of sampling correction by re-weighting (RW), i.e. clustering redundant sequences for > 80% identity is beneficial for both MI and DI methods. Results with sampling correction (solid lines) are always better than their counterparts without re-weighting (dashed lines). Using a different threshold e.g, from 80% to 70% does not have a significant influence on the mean TP performance.

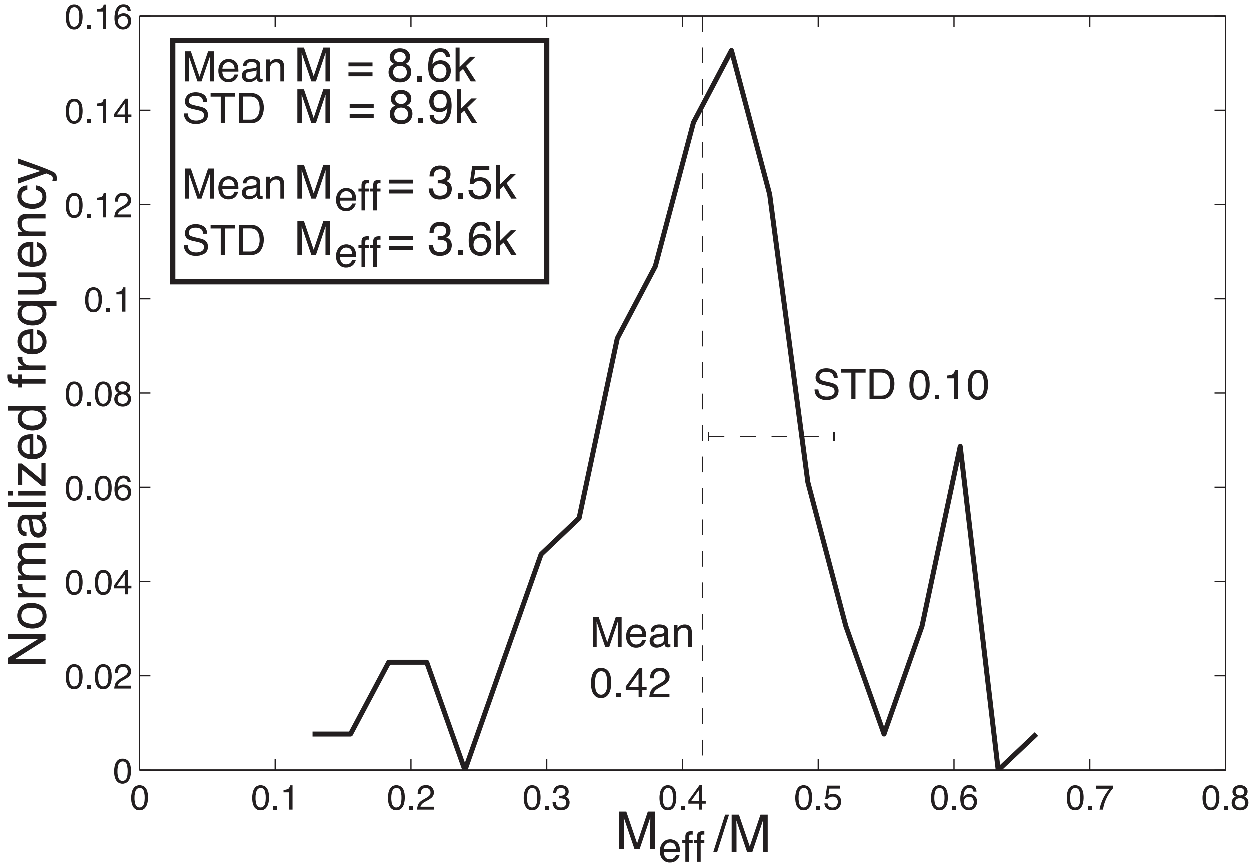

Figure S2. Distribution of the ratio Meff /M for the dataset of 131 domain families used in this study. MSA for all these families have a mean value of 8,600 sequences with a mean of 3,600 effective sequences.

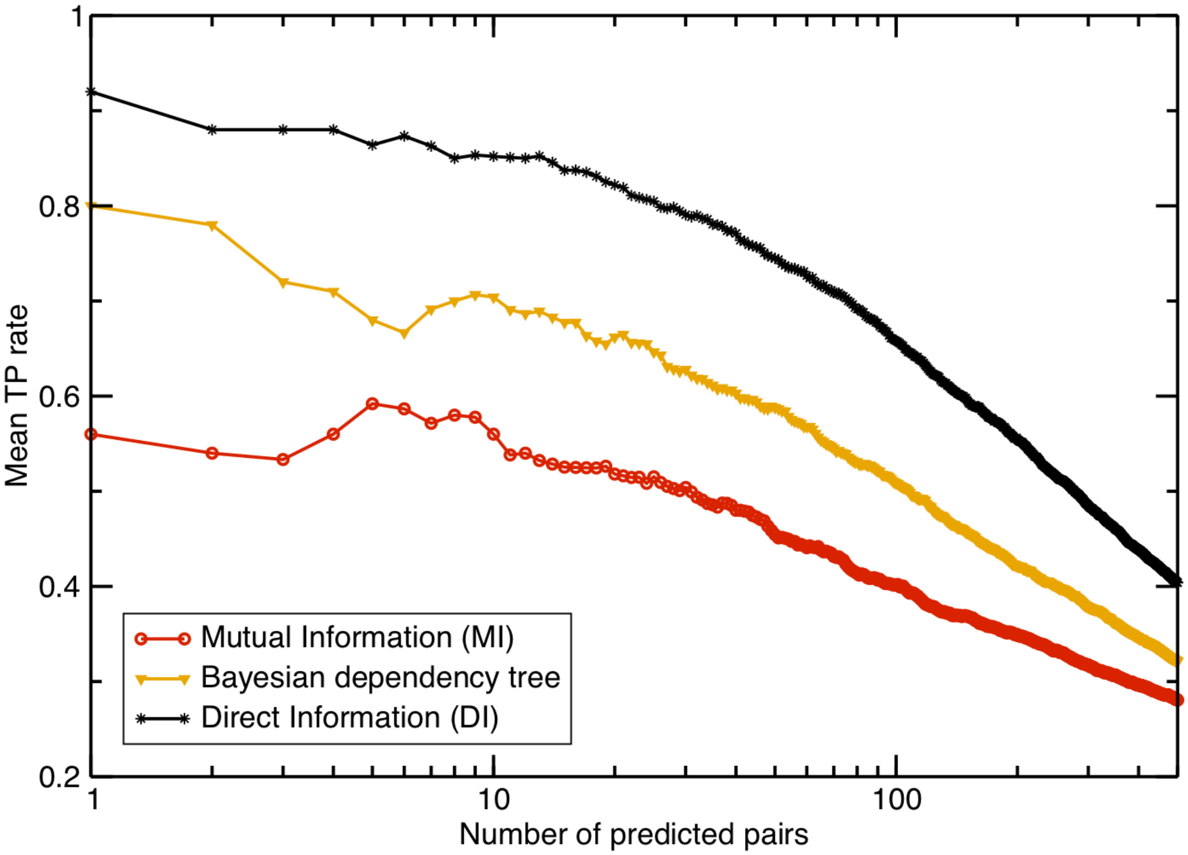

Figure S3. Mean prediction performance for 25 eukaryotic domain families with more than 2000 sequences. The figure shows equivalent results as the ones obtained for bacterial sequences (Fig. 2A and Fig. S5). This suggests that the applicability of DI-based predictions to eukayotic is plausible.

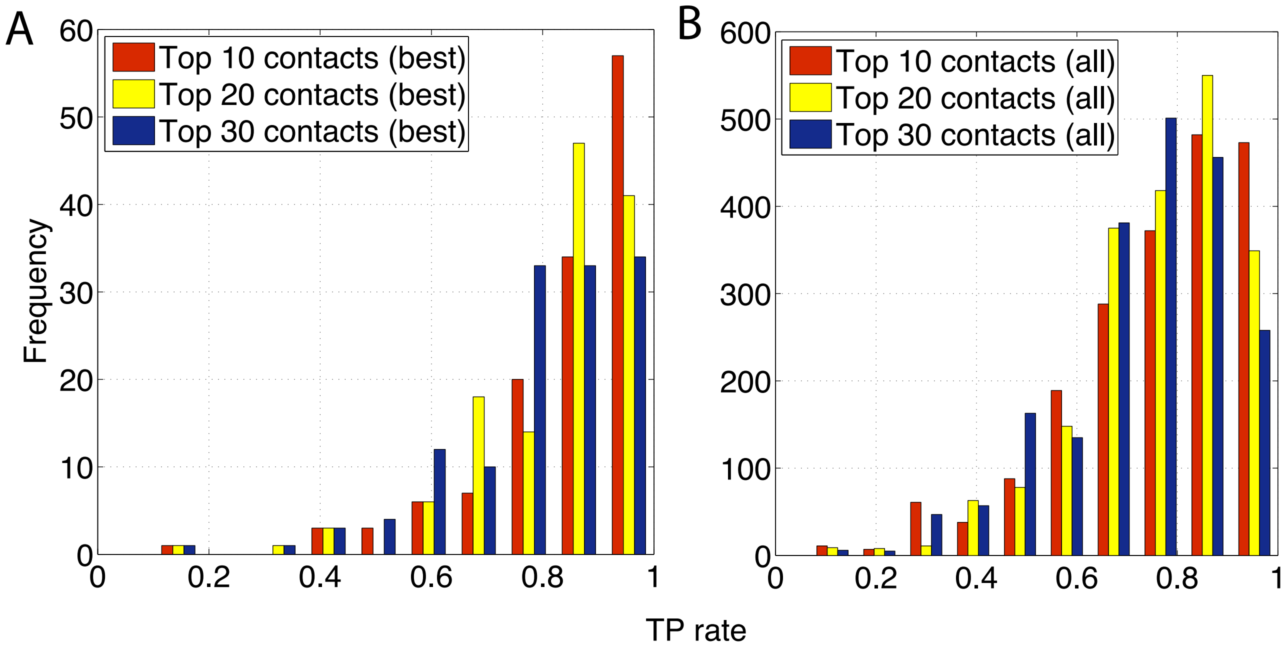

Figure S4. A) Distribution of TP rates for the 131 domains studied and computed with the best predicted structures per domain using mfDCA with sampling correction. Results are shown for the top 10,20 and 30 predicted pairs. B) Distribution of TP rates for the 131 domains studied and all PDB structures using mfDCA and sampling correction. Top 10,20 and 30 pairs seem to have a peak of the TP rate distribution around 0.8-0.9.

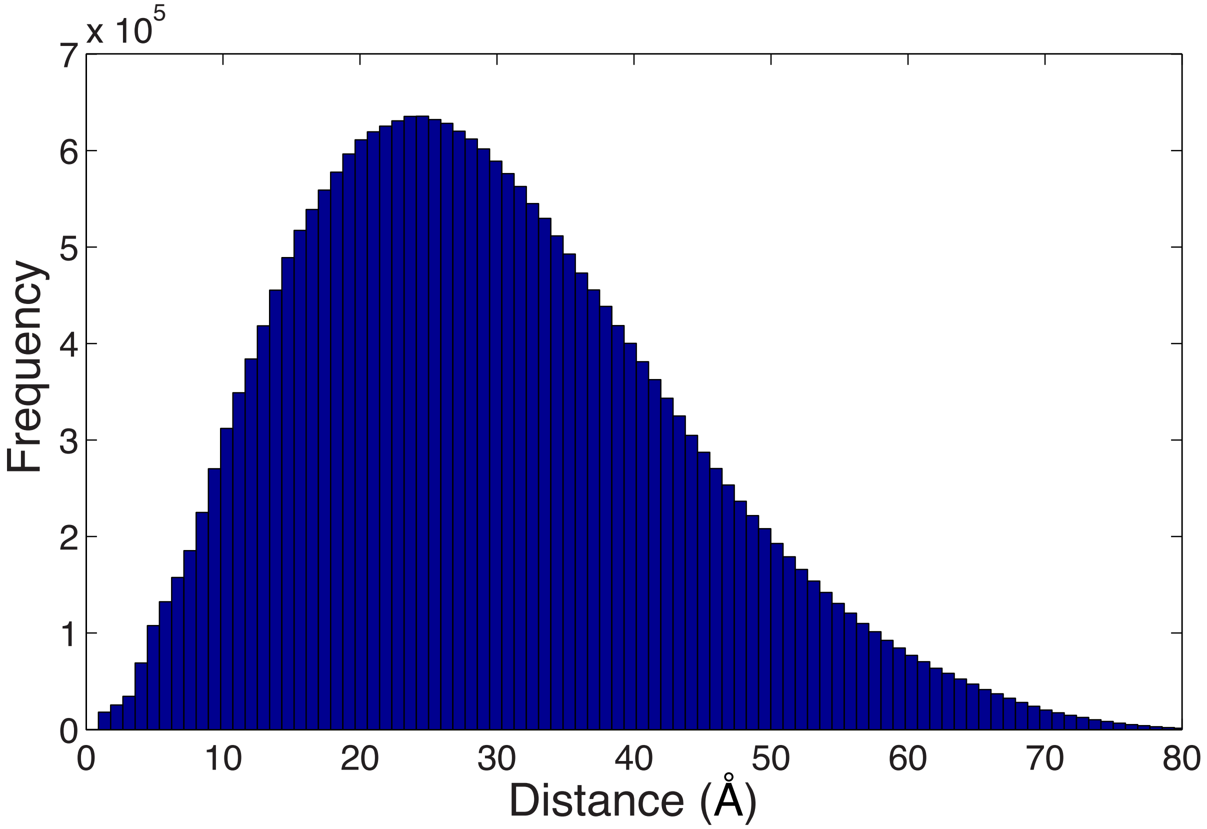

Figure S5. Histogram of all background pairwise atomic distances for 10 random PDB structures in our dataset. The peak of the distribution around 25 Å explains a small bump observed in Figure 2B near the same distance (20-25 Å) in the distribution.

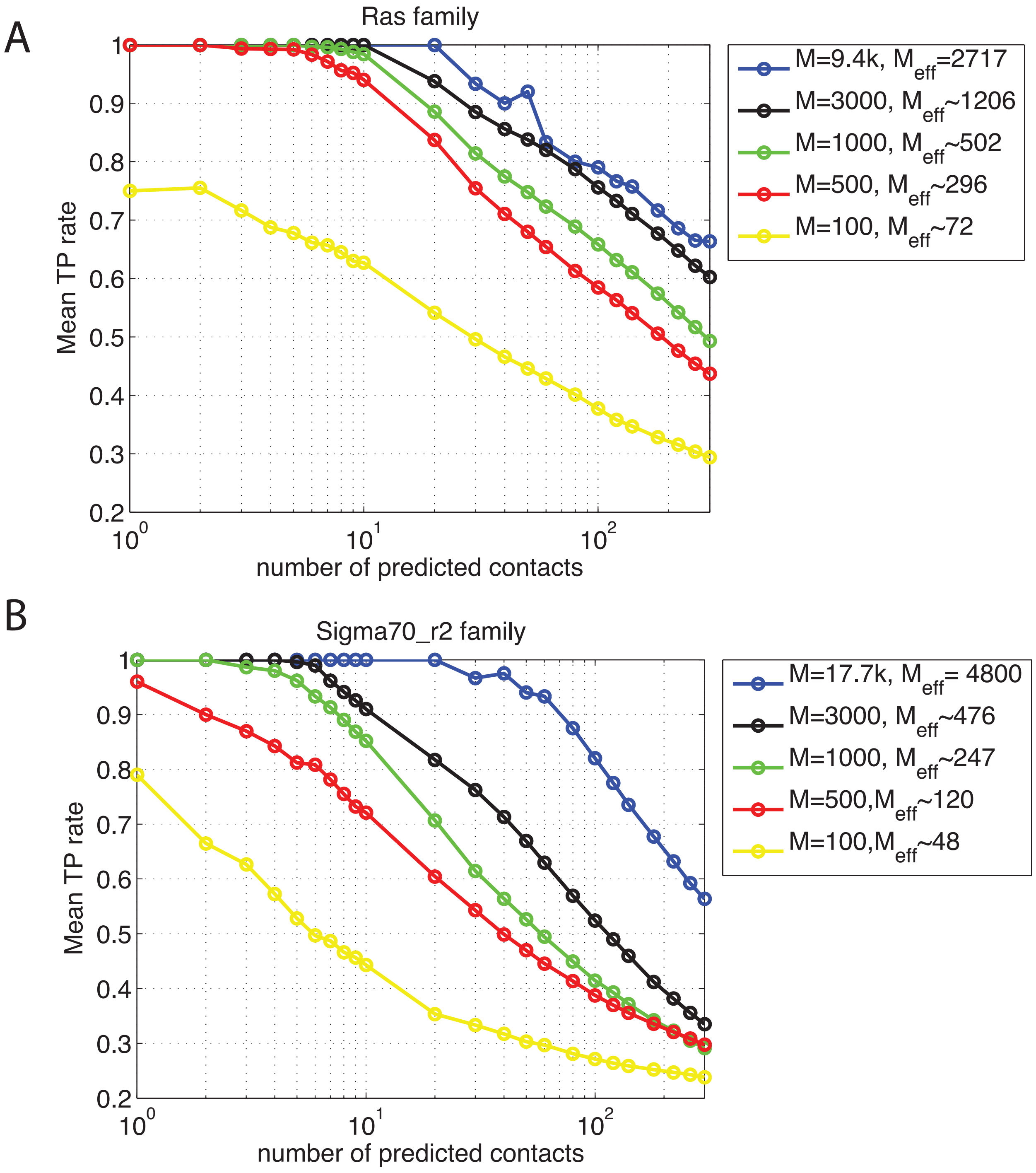

Figure S6. Sensitivity analysis of the performance of mfDCA for random sub-alignments of different lengths. Results are shown for two domain families: (A) the Ras domain family (PF00071) and (B) the DNA-recognition domain (Region 2) of the bacterial Sigma-70 factor (Pfam ID PF04542) were selected to assess prediction performance for sequence alignments of size M=100, 500, 1000 and 3000, corresponding to Meff values ranging from 72 to 1206. Curves are averaged over 100 randomly generated sub-alignments fore each M. A number of Meff ~ 250 appears to be necessary to get sensitive results, while using Meff ~ 1000 reaches results similar to the ones using full alignments.

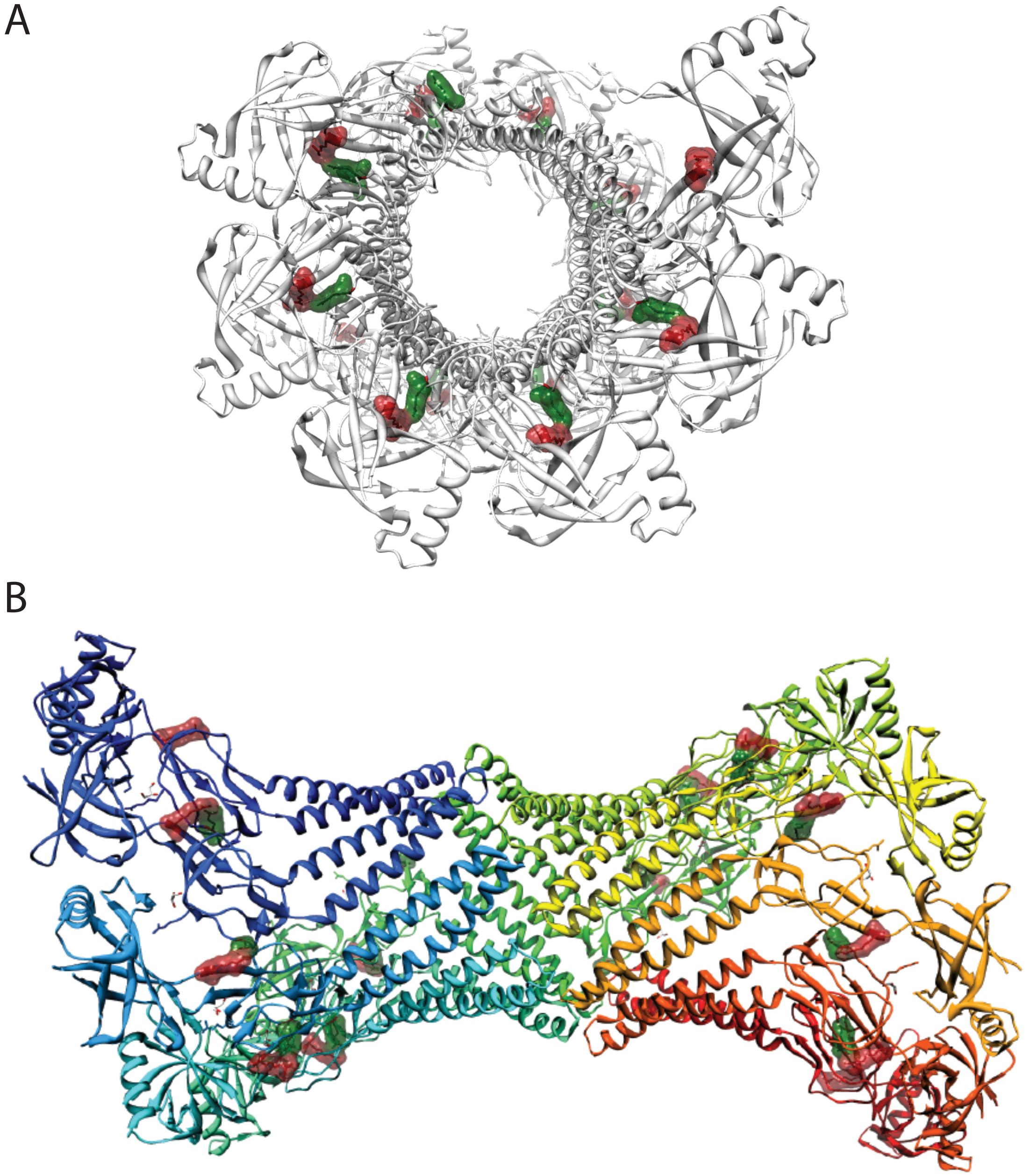

Figure S7. A) Protein MexA (PDB ID 1vf7), showing nine secretion and transporter activity domains HlyD domains (PF00529) forming a funnel like structure used as antibiotic efflux. One of two false positives in the top 20 predictions was a multimerization couplet, shown in green and red. B) Side view of the complex with domains in different colors.

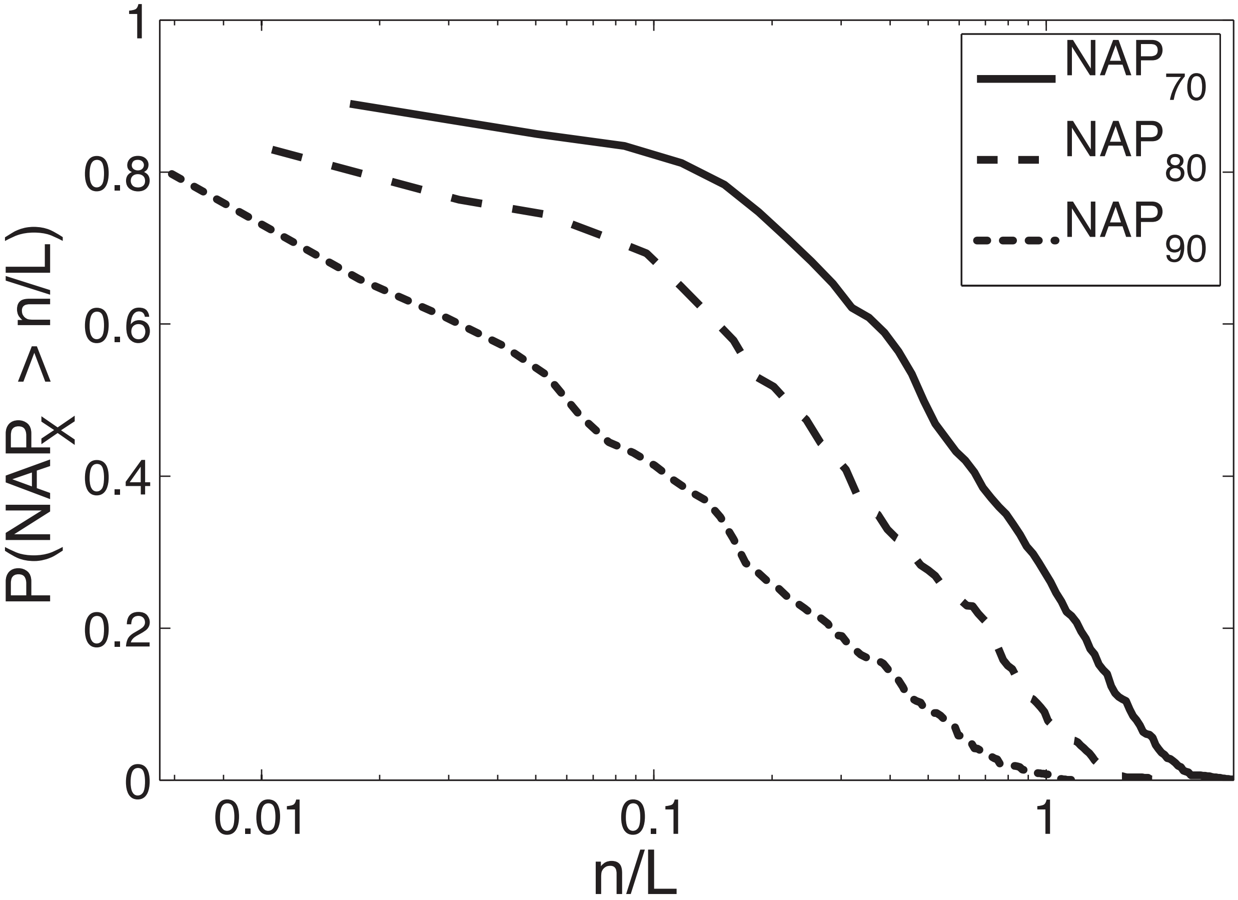

Figure S8. Cumulative distribution of the Number of Acceptable Pairs (NAPx) for a given TP rate x normalized by the length of the domain L. The curves show the probability of NAPx to be larger than a given number n for contacts at given TP rates of 0.9, 0.8 and 0.7. The curves are computed for all 856 PDB structures in the dataset.

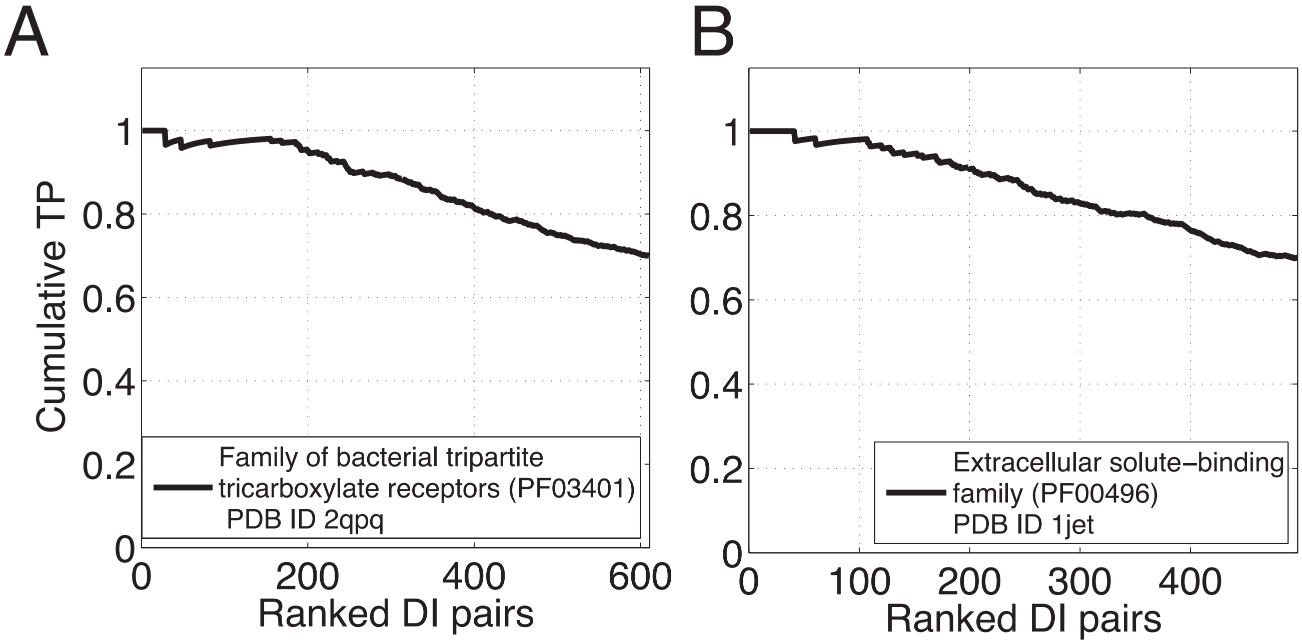

Figure S9. A) Family of bacterial tripartite tricarboxylate receptors (PF03401), NAP70 is 600, i.e.,70% of the top 600 DI pairs correspond to true contacts when mapped to structure PDB ID 2qpq. B) The extracellular solute-binding family (PF00496) mapped to the structure of the periplasmic oligopeptide-binding protein OppA of S. typhimurium (PDB ID 1jet) has a NAP70 of 497. Approximately 350 contacts are true positives.

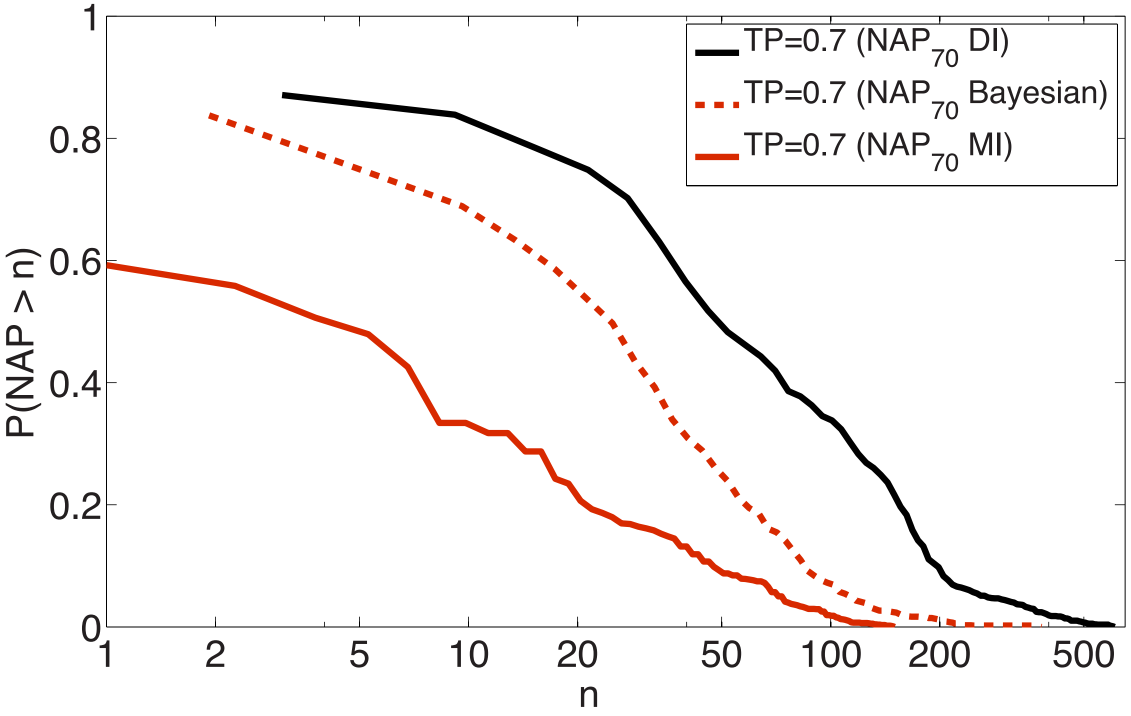

Figure S10. Comparison of the probability function of the Number of Accepted Pairs (NAP70) to be larger than a certain number of pairs for three methods: DI, Bayesian approach and MI. DI shows a clear improvement against MI (red curve) and the Bayesian approach by Burger et al. (dashed red) which becomes more evident as NAP grows larger.

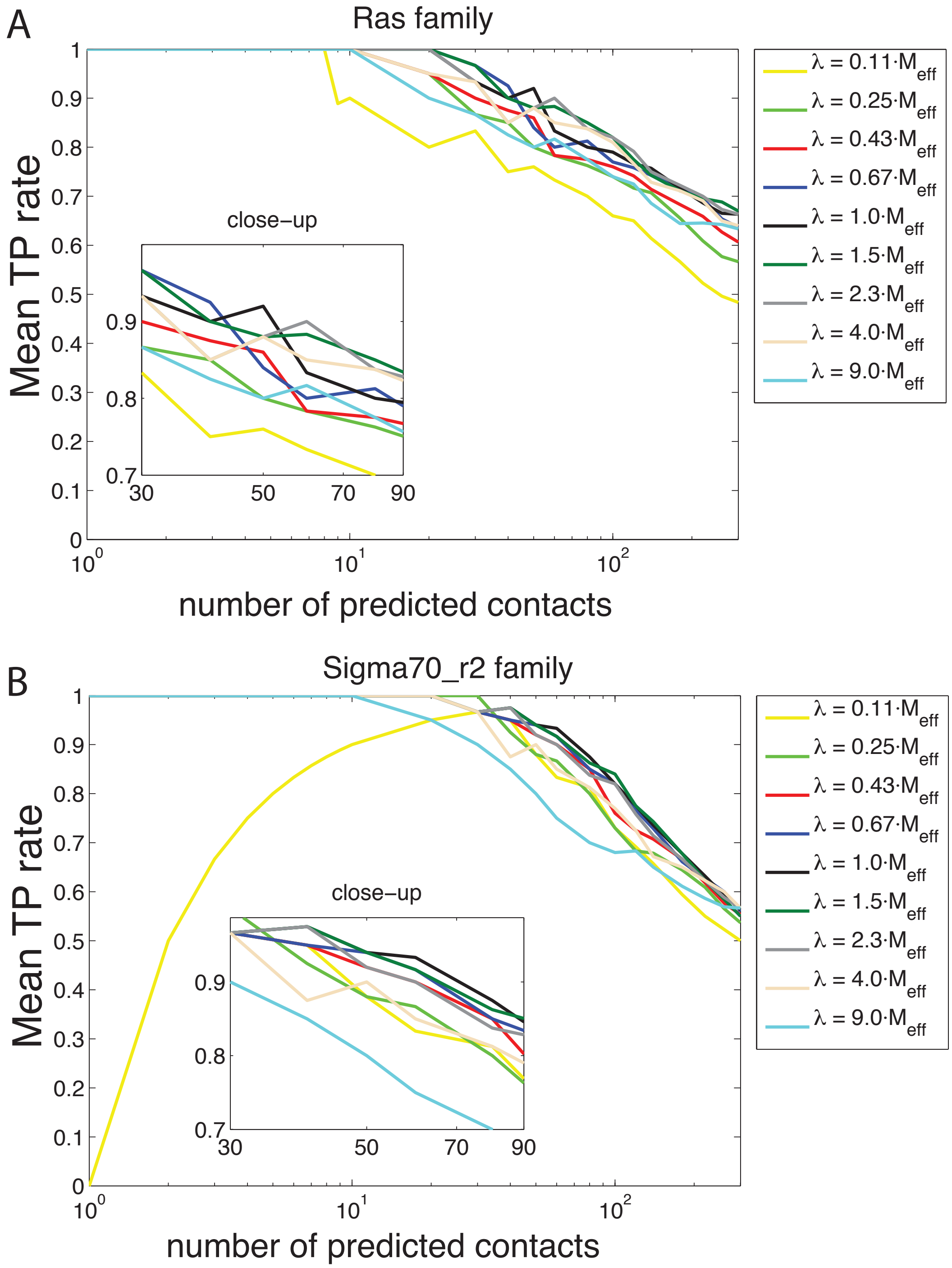

Figure S11. Performance of mfDCA for different values of the pseudocount parameter λ . Mean TP rates are shown for two domain families (A) the Ras domain family (PF00071) and (B) the DNA-recognition domain (Region 2) of the bacterial Sigma-70 factor (Pfam ID PF04542). The pseudo-count values used depend on the number of effective sequences Meff and a weighting parameter, pseudo-count weight w as λ = w Meff. Mean TP rates are computed for different w values between 0.11 and 9. A relatively small variance in performance for values of w > 0.5 is observed with the optimum between 1-1.5. λ = Meff was used as a fixed parameter in all the results shown in this study.

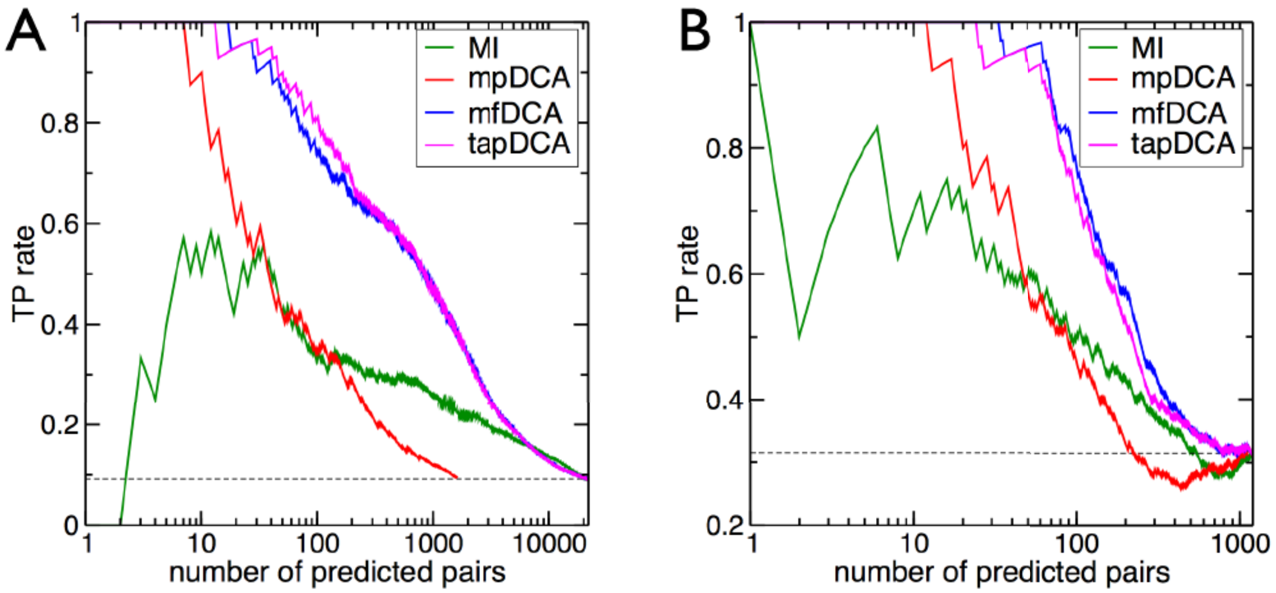

Figure S12. Comparison of different DCA approximations for (A) Trypsin (PF00089, PDB 3TGI) and (B) Trypsin inhibitor (PF00014, PDB 5PTI). Whereas all DCA algorithms outperform the contact prediction by mutual information (green line), we find the new mfDCA (blue line) to be superior to the previous mpDCA (red line). Going beyond mfDCA to the next order of the smallcoupling expansion (tapDCA, pink line), cf. Methods, does not systematically improve over mfDCA, but leads to a substantially slower algorithm. The fact that the red curve in panel A finishes at a smaller number of pairs results from the fact, that mpDCA can be run only on subalignments of up to 70 columns due to the algorithmic complexity of the approach.

Table S1. List of PDB structures analyzed in this study.

PDB IDs									
153l	1gbs	1lqp	1qgs	1vz0	2bkn	2gd9	2oqg	2z1e	3e10
154l	1gdt	1lr0	1qhg	1w55	2bko	2gj3	2oqr	2z1f	3e38
1a04	1gg4	1ls9	1qhh	1w6s	2bkp	2gjg	2oxo	2z1u	3e4r
1a0b	1gqy	1lsp	1qks	1w77	2bm4	2gkg	2oyo	2z2l	3e4v
1a0p	1gu9	1lss	1qpz	1w78	2bm5	2glk	2p19	2z2m	3e7l
1ae9	1gug	1luc	1qsa	1w8i	2bm6	2gm5	2p4g	2z4g	3e8o
1al3	1gun	1lvw	1qte	1wet	2bm7	2gms	2p5v	2z4p	3eag
1atg	1gus	1m65	1qtw	1wmi	2bnm	2gmy	2p7o	2z6r	3ec2
1b7e	1gut	1m68	1qu7	1woq	2brc	2gqp	2paq	2z8x	3ecc
1b9m	1h3l	1m6k	1qwy	1wp1	2byi	2gsk	2pbq	2z98	3ech
1b9n	1h4i	1m70	1qxx	1wpm	2c2a	2gu1	2pfx	2z9b	3ecp
1bia	1h7l	1m7j	1r1m	1wpn	2c81	2guf	2ph1	2zau	3edp
1bib	1h7q	1ma7	1r1t	1wpp	2ce0	2guh	2pjr	2zbc	3eet
1bl0	1h8z	1mb3	1r1u	1ws6	2cg4	2gup	2pkh	2zc3	3efm
1boo	1h98	1mdo	1r23	1x74	2ch7	2gxg	2pmh	2zc4	3eiw
1bsl	1h9g	1mkm	1r62	1x9h	2cvi	2gza	2pn6	2zcm	3eix
1byi	1h9j	1mkz	1r8d	1x9i	2cwq	2h1c	2pq7	2zdp	3eko
1byq	1h9k	1mm8	1r8e	1xa3	2cyy	2h98	2pt7	2zf8	3elk
1c02	1h9m	1mnz	1r9x	1xc3	2d1h	2h99	2puc	2zie	3eus
1c52	1h9s	1moq	1r9y	1xd7	2d1v	2h9b	2pud	2zif	3ex8
1c5k	1hfe	1muh	1r9z	1xi2	2d5m	2haw	2px7	2zig	3eyw
1c75	1hm9	1mur	1ra0	1xja	2d5n	2hek	2q0o	2zki	3ezu
1cb7	1hw1	1mus	1ra5	1xk6	2d5w	2heu	2q0t	2zkz	3f1c
1ccw	1hxd	1muw	1rak	1xk7	2dbb	2hkl	2q1z	2zod	3f1n
1cp2	1i0r	1mv8	1req	1xkw	2dek	2hmt	2q4f	2zov	3f1o
1crx	1i1g	1mw8	1rhc	1xkz	2df8	2hmu	2q8p	2zxj	3f1p
1crz	1i52	1mw9	1rio	1xma	2dg6	2hmv	2qb6	3b4y	3f2b
1ctj	1i58	1n2z	1rk6	1xo0	2di3	2hnh	2qb7	3b6i	3f44
1d4a	1i5n	1n9l	1rp3	1xoc	2dql	2hoe	2qb8	3b8x	3f52
1d5y	1i74	1n9n	1rrm	1xw3	2dvz	2hof	2qcz	3b9o	3f6c
1dad	1i8o	1nfp	1rtt	1y0h	2dxw	2hph	2qdf	3bcv	3f6o
1dae	1i9c	1nki	1rzu	1y1z	2dxx	2hq0	2qdl	3be6	3f6v
1dag	1icr	1nly	1rzv	1y20	2e15	2hqs	2qeu	3bem	3f8b
1dah	1id0	1nnf	1s5m	1y7m	2e1n	2hs5	2qgq	3bg2	3f8c
1dai	1id1	1nox	1s5n	1y7y	2e4n	2hsg	2qgz	3bhq	3f8f
1dak	1ihc	1nqe	1s8n	1y80	2e5f	2hsi	2qi9	3bkh	3fd3
1dd9	1ihr	1nw5	1sfx	1y82	2e7w	2hwv	2qj7	3bkv	3fgv
1dde	1ihu	1nw6	1sg0	1y9u	2e7x	2hxv	2qm1	3bm7	3fis
1di6	1ii0	1nw7	1si0	1yc9	2e7z	2i0m	2qmo	3bpk	3fms
1di7	1ii9	1nw8	1sig	1ydx	2eb7	2i5r	2qpq	3bpq	3fwy
1dlj	1ini	1nwz	1sly	1ye5	2ecu	2ia2	2qsx	3bpv	3fwz
1dts	1inj	1ny5	1sqe	1yf2	2efn	2ia4	2qwx	3bqx	3fxa
1dur	1ir6	1ny6	1sqs	1yg2	2eh3	2ibd	2qx4	3bre	3fzv
1e2x	1iuj	1o1h	1sum	1yio	2ehl	2ict	2qx6	3bs3	3g13
1e3u	1ixc	1o2d	1suu	1yiq	2ehz	2ift	2qx8	3bvp	3g5o
1e4d	1ixg	1o61	1t3t	1ylf	2ek5	2ikk	2r01	3bwg	3g7r
1e4f	1ixh	1o69	1t5b	1yoy	2esh	2ipl	2r0x	3c1q	3gdi
1e4g	1iz1	1o7l	1t72	1ysp	2esn	2ipm	2r1j	3c29	3gfa
1e8c	1j5y	1oad	1ta9	1ysq	2esr	2ipn	2r25	3c3w	3gfv
1ecl	1j6u	1oap	1td5	1yvi	2ewn	2is1	2r4t	3c48	3gfx
1efa	1jbg	1odd	1tf1	1z05	2ewv	2is2	2r6g	3c57	3gfy
1efd	1jbw	1odv	1tqg	1z19	2eyu	2is4	2r6o	3c7j	3gfz
1eg2	1je8	1oj7	1tqq	1z7u	2f00	2is6	2r6v	3c85	3gg0
1ek9	1jet	1olt	1tv8	1zat	2f2e	2is8	2ra5	3c8f	3gg1
1esz	1jeu	1opc	1tvl	1zi0	2f5x	2iu5	2rb9	3c8n	3gg2
1etk	1jev	1opx	1tzb	1zlj	2f6g	2iuy	2rc7	3c9u	3ghj

1eto
1etv
1etw
1etx
1ety
1ezw
1f07
1f1u
1f44
1f48
1f5v
1f9i
1fca
1fdn
1fep
1fia
1fip
1fp6
1fr3
1fse
1fxo
1g1l
1g1m
1g20
1g28
1g5p
1g60
1g6o
1g72
1g8k
1jft
1jh9
1jiw
1jlj
1jnu
1jpu
1jq5
1jyk
1k20
1k2v
1k38
1k4f
1k54
1k56
1kap
1kb0
1kbu
1kgs
1kmo
1kmp
1kq3
1ku3
1ku7
1kv9
1kw3
1kw6
1l3l
1lj9
1lq9
1lqk
1or7
1ot6
1ot9
1ota
1otb
1oxk
1p2f
1p31
1p3d
1p7d
1p9r
1p9w
1pb0
1pb7
1pb8
1pjr
1pnz
1po0
1pt7
1pvp
1q05
1q06
1q07
1q08
1q09
1q0a
1q35
1q7e
1qg8
1qgq
1tzc
1u07
1u2w
1u8b
1u8t
1uaa
1uc8
1uc9
1us4
1us5
1usc
1usf
1uux
1uuy
1uyl
1v4y
1v51
1v8p
1v96
1vct
1ve2
1vf7
1vgt
1vgw
1vhd
1vhv
1vim
1vj7
1vke
1vlj
1zvt
1zvu
1zzc
2a0b
2a3n
2a5h
2a5l
2a61
2aa4
2aac
2ad6
2ad7
2ad8
2aef
2aej
2afh
2am1
2anu
2ap1
2ar0
2ara
2arc
2azn
2b02
2b0p
2b13
2b3z
2b44
2bas
2bfw
2f6p
2f7a
2f7b
2f8l
2f9f
2fa1
2fa5
2fb2
2fbh
2fcj
2fdn
2fe1
2fez
2ff4
2ffu
2fhp
2fn9
2fnu
2fpo
2fsw
2fvy
2fw0
2g2c
2g6v
2g7u
2gai
2gaj
2gci
2gd0
2gd2
2iv7
2iw1
2iw4
2iwx
2jba
2jcg
2jfg
2nip
2npn
2nq2
2nq9
2nqh
2nt3
2nt4
2o08
2o0y
2o3j
2o4d
2o7i
2o7p
2o8x
2o99
2o9a
2obc
2ofy
2ogi
2ojh
2okc
2olb
2ooc
2rc8
2rca
2rde
2rii
2ril
2rsl
2uag
2v25
2v2k
2v9y
2vha
2vjq
2vk2
2vke
2vkr
2vlg
2vma
2vmb
2vpz
2vsh
2w27
2w8b
2w8i
2yve
2yx0
2yxb
2yxo
2yxz
2yye
2yz5
3can
3ccg
3cij
3cix
3ckj
3ckn
3ckv
3clo
3cnr
3cnv
3cp5
3ctp
3cuo
3cwr
3cx4
3cyi
3cyp
3cyq
3d5k
3d6z
3d7i
3dbo
3df7
3df8
3dma
3dr4
3drf
3drj
3dsg
3du1
3gp4
3gpv
3gr3
3guv
3h4o
3h5t
3h87
3hfi
3hh0
3hhh
3hl0
3hmz
3hn7
3hoi
3htv
3hvw
3pyp
3uag
4aah
4crx
4req
4uag
5req
6req
7req
8abp

Table S2. List of Pfam domain families analyzed in this study.

Pfam Domain Names
ABM
AIRS
AIRS_C
AP_endonuc_2
ATP-grasp_3
Amidohydro_3
AraC_binding
ArsA_ATPase
AsnC_trans_reg
B12-binding
BPD_transp_1
Bac_luciferase
Bug
CMD
CbiA
CheW
CoA_transf_3
Cons_hypoth95
Cytochrom_C
DHH
DHHA1
DNA_gyraseA_C
DegT_DnrJ_EryC1
EAL
FCD
FMN_red
Fe-ADH
FecCD
Fer4
Fer4_NifH
Flavin_Reduct
Flavodoxin_2
FtsA
GGDEF
GSPII_E
GSPII_F
GerE
Glycos_transf_1
Glycos_transf_2
Glyoxalase
GntR
HATPase_c
HD
HTH_1
HTH_11
HTH_3
HTH_5
HTH_8
HTH_AraC
HTH_IclR
HemolysinCabind
HisKA
HlyD
Hpt
HxlR
IclR
IspD
IstB
LacI
LysR_substrate
MCPsignal
MarR
MerR-DNA-bind
MerR
Methylase_S
MoCF_biosynth
Molybdopterin
Molydop_binding
Mur_ligase
Mur_ligase_C
Mur_ligase_M
N6_Mtase
N6_N4_Mtase
NMT1
NTP_transferase
Nitroreductase
OEP
OmpA
PAS
PASTA
PAS_3
PD40
PHP
PIN
PQQ
PadR
ParBc
Pentapeptide
Peptidase_M23
Peripla_BP_1
Peripla_BP_2
Phage_integr_N
Phage_integrase
PhoU
PilZ
Plasmid_stabil
Plug
ROK
Radical_SAM
Resolvase
Response_reg
RibD_C
RimK
Rrf2
SBP_bac_1
SBP_bac_3
SBP_bac_5
SIS
SLBB
SLT
Sigma54_activat
Sigma70_r2
Sigma70_r4
Sigma70_r4_2
Surf_Ag_VNR
TOBE
TOBE_2
TP_methylase
TetR_N
TonB
TonB_dep_Rec
Toprim
Trans_reg_C
Transpeptidase
Transposase_11
TrkA_N
TrmB
UDPG_MGDP_dh_N
UTRA
UvrD-helicase
YkuD

Table S3. Pfam domain families and their respective PDB structure with oligomerization TP contacts.

Pfam Domain	PDB structure
AsnC_trans_reg	2z4p
Bac_luciferase	3b4y
CMD	1vke
EAL	2r6o
Flavodoxin_2	1t5b
FMN_red	2a5l, 2q62
Glyoxalase	2p7o
GSPII_E	2gza
HlyD	2f1m,1t5e
Hpt	1i5n
HTH_IclR	2g7u
HxlR	2f2e
IspD	3f1c
MCPsignal	2ch7
MerR-DNA-bind	3gp4
Mur_ligase	2am1
Resolvase	2gm5
Sigma54_activat	1ny6
TOBE	1h9s
TOBE_2	2awn
TP_methylase	1vhv

Table S4. Top-30 prediction of mfDCA for the Serine protease data of (41). The first two columns specify the residue pair, the third column provides the DI value, and the last one the native distance in rat trypsin (PDB ID 3tgi). Residues belonging to the sectors defined in (41) are indicated, using the color scheme of (41).

Res. 1	Res. 2	DI	Dist/Å
136	201	0.52	2.0
32	40	0.47	2.8
191	220	0.37	2.2
189	226	0.34	3.3
57	195	0.34	2.7
42	58	0.28	2.0
44	52	0.25	4.3
30	139	0.25	2.7
72	77	0.24	3.0
72	78	0.23	8.0
59	104	0.23	3.9
51	105	0.22	3.8
190	213	0.20	3.7
34	40	0.19	3.4
116	127	0.18	23.7
26	157	0.18	4.9
45	209	0.18	3.8
117	127	0.17	23.9
46	112	0.16	4.0
71	78	0.15	8.5
71	79	0.15	6.9
117	122	0.15	13.3
161	184	0.15	3.1
138	213	0.14	4.2
116	122	0.14	13.1
53	209	0.14	3.5
189	228	0.13	3.9
100	179	0.13	2.3
102	195	0.13	6.1
27	157	0.13	3.8