Protein function depends crucially on the correct folding of its three-dimensional structure. The maintenance of this structure during evolution imposes strong constraints on the variability of the amino acid sequences between homologous (i.e., closely related) proteins. For example, deleterious alterations (or mutations) in one residue may have to be compensated by mutations of the residues it interacts with, leading to correlations between the amino acid compositions at different sequence positions. In essence, a particular change at one position may correlate with certain necessary changes at other positions that enable the protein to maintain its structure and function.

Due to rapid advances in the technologies for sequencing DNA, a large number of genome projects that aim to determine the complete DNA (genome) sequence of an organism are now in existence, which is particularly the case for bacterial genomes, with approximately 1.700 completed and 8,300 ongoing (1). These genome sequences can be used to provide correlated substitution patterns for a large number of common bacterial proteins and interacting protein pairs, as well as for some abundant eukaryotic, or nonbacterial, proteins. The availability of such correlations presents opportunities to make inferences on residue contacts, that is, contacts between amino acid residues within or between proteins, and to eventually predict protein structures at various levels of organization based on sequence information alone. Crucial to this inference is the ability to disentangle direct and indirect correlations (i.e., induced through intermediate residue positions), as accomplished by the recently introduced method of direct-coupling analysis (DCA) (2), based on sophisticated tools developed in statistical physics. Here, we further develop a computationally efficient implementation of DCA, called mfDCA, an algorithm based on the mean-field approximation of DCA, which is 10^3 to 10^4 times faster than the original DCA implementation, and hence can be used to rapidly analyze large numbers of families of long protein sequences.

The performance of mfDCA is first illustrated in Fig. P1A for a specific protein structure (domain) taken from the proteome (i.e., the assembly of proteins present in an organism under specific conditions) of the bacterium Escherichia coli (3). The top 20 DCA predictions, separated by at least five positions along the sequence to exclude...
trivial predictions, are displayed as bonds of different color: The 19 residue pairs with distances < 8 Å are shown in red, and the lone pair with a larger interresidue distance is shown in green. The mfDCA thus gave a true-positive (TP) rate of 95% for the top 20 contacts predicted for this protein.

To test the general applicability of mfDCA, we analyzed 131 predominantly bacterial domain families. These families were selected (i) to be sufficiently abundant for statistical analysis (> 1,000 nonredundant sequences), and (ii) to feature at least two high-resolution X-ray crystal structures (resolution < 3 Å), so that the spatial proximity between predicted residue pairs could be evaluated. The selected domain families encompassed a total of 856 different solved structures listed in the Protein Data Bank. On average, 84% of the top 20 mfDCA predictions (a minimum separation of five along the sequence) in each of the 131 families are TP contacts, whereas simple correlation analysis using mutual information reaches only a 59% TP rate (Fig. P1B). The individual TP rates of mfDCA predictions for the 131 families are distributed mostly in the range of 70–100%. Analogous results were also observed for a collection of 25 abundant eukaryotic proteins.

The high accuracy of mfDCA predictions led us to examine whether the small fraction of “false positives” (i.e., predicted pairings located far away in the crystal structures) might also have some biological significance. We investigated various biological reasons for the appearance of such long-distance direct correlations and found the following cases.

First, direct correlation may be due to contacts between proteins in oligomeric complexes (i.e., complexes comprising different proteins). One example is the ATPase domain of the nitrogen regulatory protein C family transcriptional activators, which act to promote the activity of RNA polymerase. Upon activation, different subunits of this domain are known to form a multiunit ring, wrapping DNA around the complex (4). Among the top 20 mfDCA-predicted contacts, three pairs appeared to be long-distance (> 10 Å) within the domain; but they all turned out to be within 5 Å of the closest position in an adjacent subunit of the ring structure. More globally, out of the 131 studied domain families, 21 families feature X-ray crystal structures involving oligomers. Among the top 20 pairs that are not intradomain contacts (i.e., contacts within the same domain), about one-half turned out to be interdomain contacts (i.e., contacts between domains).

A second cause of high-scoring long-distance pairs is the occurrence of alternative domain conformations. As an illustration, we examined the domain family GerE, whose members include the DNA-binding domains of a number of proteins (response regulators) in a system known as the two-component signaling system. Using the DNA-bound domain of the nitrate/nitrite response regulator NarL of E. coli, we found that all of the top 20 DCA predictions were true contacts. However, when mapping the same pairs to the structure of the inactive, full-length response regulator of the transcriptional regulatory protein DosR of Mycobacterium tuberculosis, seven pairs were found at distances > 8 Å. A comparison between these structures shows clearly that all of the long-distance pairs involve one particular position that is significantly displaced in the full-length structure, presumably due to interaction with the (unphosphorylated) regulatory domain.

A third cause of high-scoring long-distance pairings inferred by mfDCA is observed in the metalloenzyme domain family, which binds to metal ions. It has a high-scoring residue pair, which is separated by more than 14 Å when mapped to the glutathione transferase FosA of Pseudomonas aeruginosa. We observed that the two residues belonging to the different subunits of this residue pair are, however, in direct contact with the manganese “Mn(II)” ion when the enzymes are paired, or in the dimer configuration. 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.

One of the most striking features of our results obtained using DCA is how gradually the average TP rate declines with increasing direct information ranking. This observation naturally led us to explore the information lying beyond the first 10–30 DCA predictions. For instance, how many residue contacts can be predicted if a TP rate of at least 70% is required? As can be seen in Fig. P1B, this TP rate is reached on average after the top 70 DCA predictions, where one would have obtained approximately 50 true contacts on average. In extreme cases (e.g., the family of bacterial receptors called the tripartite tricarboxylate receptors), a TP rate of > 70% was present for the first 600 DCA-predicted residue pairs, therefore including more than 400 native contacts.

The large number of contacts correctly predicted by mfDCA prompted us to explore the extent to which our method may be used to predict the contact maps of protein domains. In Fig. P1C, we show the example of the eukaryotic H-Ras protein (5). The first approximately 800 mfDCA contact predictions are displayed; they include 320 nontrivial pairs, which are separated by at least five positions along the protein backbone. The predicted contact map captures very well the native contacts. Even if not all single contacts are predicted, almost all contact clusters are identified.

In this work, we have shown the ability of mfDCA to identify, with high accuracy, residue pairs in domain families and found that mfDCA is not only able to identify intradomain contacts but also interdomain residue pairs that are part of oligomerization interfaces. One potential application is the identification of interaction interfaces for homodimers that could ultimately help in predicting complex structures. 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 used with pairs of proteins rather than single proteins to identify potential protein–protein interactions.