Assessing the utility of coevolution-based residue–residue contact predictions in a sequence- and structure-rich era

Hetunandan Kamisetty[a,b], **Sergey Ovchinnikov**[a,b,c], **and David Baker**[a,b,1]

[a]Howard Hughes Medical Institute, [b]Department of Biochemistry, and [c]Molecular and Cellular Biology Program, University of Washington, Seattle, WA 98195

Contributed by David Baker, July 29, 2013 (sent for review March 1, 2013)

Recently developed methods have shown considerable promise in predicting residue–residue contacts in protein 3D structures using evolutionary covariance information. However, these methods require large numbers of evolutionarily related sequences to robustly assess the extent of residue covariation, and the larger the protein family, the more likely that contact information is unnecessary because a reasonable model can be built based on the structure of a homolog. Here we describe a method that integrates sequence coevolution and structural context information using a pseudolikelihood approach, allowing more accurate contact predictions from fewer homologous sequences. We rigorously assess the utility of predicted contacts for protein structure prediction using large and representative sequence and structure databases from recent structure prediction experiments. We find that contact predictions are likely to be accurate when the number of aligned sequences (with sequence redundancy reduced to 90%) is greater than five times the length of the protein, and that accurate predictions are likely to be useful for structure modeling if the aligned sequences are more similar to the protein of interest than to the closest homolog of known structure. These conditions are currently met by 422 of the protein families collected in the Pfam database.

protein coevolution | maximum-entropy model | markov random field

There has been long-standing interest in the prediction of residue–residue contacts based on the covariance of residue-substitution patterns in multiple aligned sequences (1). For many years, these methods met with relatively little success, but with the increase in the number of known protein sequences and improvements in methods such approaches have recently demonstrated considerable promise. The methodological improvements distinguish between direct couplings and indirect correlations that arise from chains of the direct couplings (i.e., if A is coupled to B, and B to C, one might erroneously conclude A is coupled to C). Two recent methods, Direct Coupling Analysis (DCA) and Protein Sparse InverseCOVariance (PSICOV) (2, 3), achieve this separation by inverting a residue–residue covariance matrix.

In parallel with the growth of the sequence databases, there has been a considerable increase in the number of known structures over the past decade. Comparative modeling methods, which predict protein structure based on homologs of known structures, have become increasingly powerful, and can generate models more accurate than those produced by de novo modeling in most cases. Thus, the growth in the databases over the last decade represents something of a catch-22 for contact prediction: there are many more sequences, so such predictions can be made much more accurately, but there are few protein families with the many sequences required for accurate contact prediction whose structures cannot be modeled relatively accurately using comparative modeling methods.

In this paper we begin by examining the approximations involved in the residue–residue covariation matrix inversion used by PSICOV and DCA, and show that more accurate contact predictions can be obtained using fewer sequences by going beyond the second-order approximation to the underlying distribution implicit in both methods. We then rigorously assess the utility and limits of contact prediction for protein tertiary structure modeling by evaluating the extent to which predicted contacts can contribute to modeling in the presence of the homologous structure information likely to be available.

Results and Discussion

Previous work (2–4) has demonstrated that contacts between residues in the 3D structure of a protein can be predicted with considerable accuracy for large protein families based on the evolutionary covariance observed in multiple sequence alignments. We briefly review the basis for these approaches to motivate the method described in this paper.

Given a set of data drawn from an (unknown) multivariate probability distribution $P(X_1, \ldots, X_L)$, the marginal frequencies $M_i(X_i)$, and the pair correlations $F_{i,j}(X_i, X_j)$ can be readily computed. If the underlying distribution is Gaussian:

$$P(X_1,..,X_L = \mathbf{x}) \propto \exp\left(-(\mathbf{x}-\mu)^T \mathbf{\Omega}(\mathbf{x}-\mu)\right). \qquad [1]$$

The parameters of the distribution can be readily obtained from the frequencies and pair correlations: $\mu = [M_1, \ldots, M_L]$ and $\mathbf{\Omega} = (F - M'M)^{-1}/2$. Although the observed $F_{i,j}$ are subject to chaining effects (if A is correlated with B, and B with C, then A will appear correlated with C), the Ω_{ij} are the direct couplings between variables with chaining effects eliminated.

Although they differ significantly in derivation and their estimation procedures, both DCA and PSICOV use the recipe of approximating the direct couplings with an estimate of the inverse covariance matrix, as is appropriate for a Gaussian. Our approach, called GREMLIN, avoids this approximation and instead obtains model parameters from the conditional correlations $(F(X_1|X_2, X_3, \ldots X_L))$ using our pseudolikelihood framework (5), optimizing the learning procedure for contact prediction.

Significance

We develop an improved method for predicting residue–residue contacts in protein structures that achieves higher accuracy than previous methods by integrating structural context and sequence coevolution information. We then determine the conditions under which these predicted contacts are likely to be useful for structure modeling and identify more than 400 protein families where these conditions are currently met.

Author contributions: H.K. and D.B. designed research; H.K. and S.O. performed research; H.K. and S.O. contributed new reagents/analytic tools; H.K., S.O., and D.B. analyzed data; and H.K. and D.B. wrote the paper.

The authors declare no conflict of interest.

Freely available online through the PNAS open access option.

Data deposition: GREMLIN is available at http://gremlin.bakerlab.org.

[1]To whom correspondence should be addressed. E-mail: dabaker@u.washington.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1314045110/-/DCSupplemental.

A recently published method, PseudoLikelihood Maximization Direct Coupling Analysis (PLMDCA) (6), also uses a pseudolikelihood approach. Here, we go beyond PLMDCA improving the robustness of predictions with fewer sequences by incorporating prior information on pairs likely to be in contact.

Accuracy of Contact Prediction. It was recently shown that pseudolikelihood-based contact prediction is more accurate than DCA on a set of Pfam domains with deep alignments (6). Here, we first carry out a more comprehensive comparison of contact prediction methods (GREMLIN, PSICOV, DCA, MIc, and PLMDCA) on a larger set of families with varying alignment depths. Second, we explore the utility of alternate sources of prior information in improving the accuracy of pseudolikelihood-based contact prediction.

We constructed a dataset of 329 protein targets selected from the Continuous, Automated Model Evaluation (CAMEO) server (7) (Dataset S1, details in *Materials*). These targets represent a snapshot of recently deposited protein structures into the Protein Data Bank and encompass a wide range of protein sizes, folds, and evolutionary histories. For each protein in the list, we constructed alignments of evolutionarily related proteins and used each method to predict contacts. We computed the accuracy of contact predictions for each method by computing the fraction of predicted contacts that had a $C\beta$–$C\beta$ ($C\alpha$ in the case of Glycine) distance less than 8 Å. To discount the effects of local secondary structure on accuracy, we restricted ourselves to positions at least 12 residues apart in the target sequence (results when restricted to positions at least 24 residues apart are similar; *SI Appendix*, Fig. S1). The methods assign a score to each contact prediction, and the predictions were ranked for each protein target based on these values.

Over the CAMEO protein test set, GREMLIN's pseudolikelihood method is more accurate than DCA, PSICOV, and MIc when no prior information was used (Fig. 1 *A*–*C*). A simple prior based on sequence separation and predicted secondary structure improves GREMLIN's accuracy further (Fig. 1*D*). For the top ranked $L/2$ predictions (where L is the length of the target protein), the accuracy of GREMLIN with this simple prior is higher than that of the next most accurate method for a majority of targets (Fig. 1*E*). PSICOV accuracy is higher on average than DCA [with average product correction (APC); ref. 8] for the first $L/2$ predictions but falls below DCA at larger numbers of predictions. PLMDCA was much slower than other methods compared here; we therefore compared it to GREMLIN on smaller datasets in Fig. 2 and *SI Appendix*, Fig. S7.

Difference in Accuracy as a Function of the Number of Sequences. For targets that had at least 20 L sequences (using the number of nonredundant sequences at 90% sequence identity as a measure of alignment depth), we generated subalignments and recomputed predictions for each of these. Fig. 1 *F* and *G* show the average accuracy of the methods across these targets varying the depth of the alignment: GREMLIN's pseudolikelihood-based approach is more accurate than other methods across alignment depths, even when no prior information is used. Using prior information further improves the accuracy of predictions, especially when there are few sequences in the alignment.

The Effects of Prior Information. We further investigated the utility of predicted secondary structure-based prior (using PSIPRED; ref. 9) as well as a more powerful prior using SVMCON—a contact prediction method that uses profile–profile similarity and secondary structure (10). To reduce the possible overlap with the protein structures used in the training of PSIPRED and SVMCON, we evaluated their contribution on a set of hard targets that are unlikely to have homologs with known structure (Dataset S2). When no prior information was used, GREMLIN and PLMDCA had similar accuracy (*SI Appendix*, Fig. S7) although GREMLIN was 5–20× faster. Using prior information based on secondary structure and sequence separation improved accuracy of predictions (Fig. 2*A*). Using a prior based on SVMCON improved results further with the resulting method outperforming SVMCON on most targets (Fig. 2 *C* and *D*).

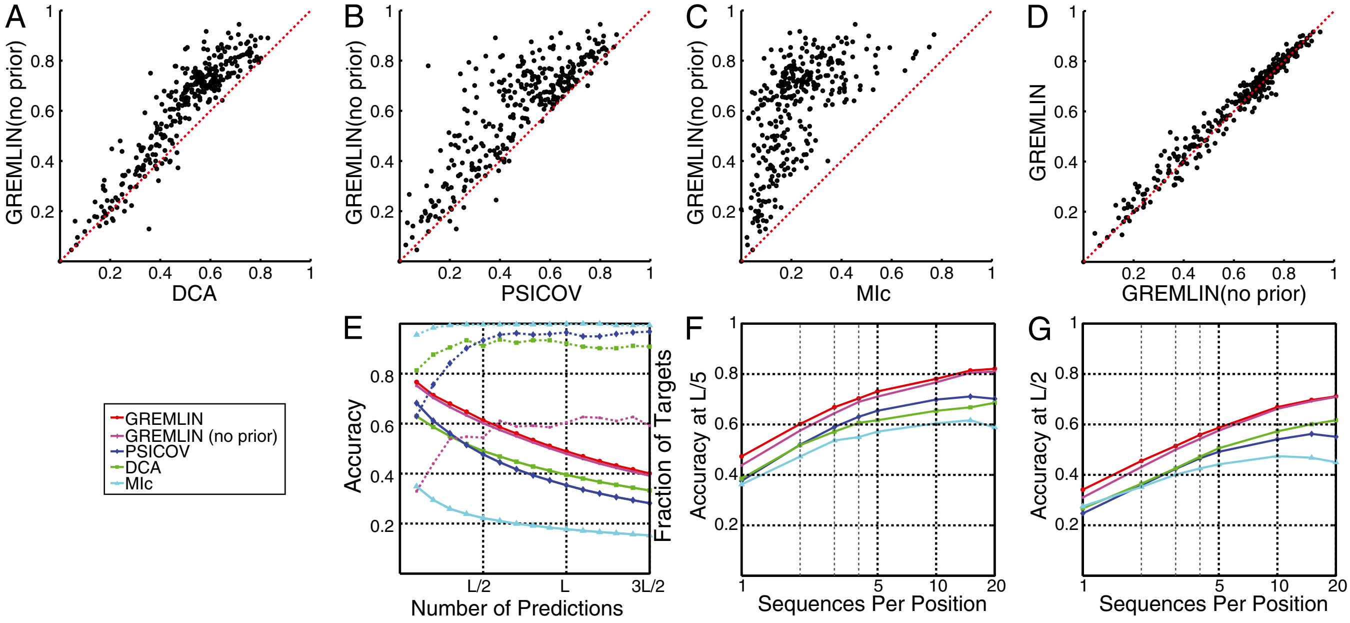

Fig. 1. Accuracy of contact prediction. Comparison of GREMLIN with DCA (*A*), PSICOV (*B*), MIc (*C*), and GREMLIN when no prior information is used (*D*). Each point corresponds to a protein, the axes indicate the accuracy of the top ranked $L/2$ $C\beta$–$C\beta$ contacts predicted by the indicated methods. (*E*) (solid lines) Average accuracy for varying numbers of predictions; (broken lines) fraction of targets where GREMLIN was more accurate than the indicated method. Dependence of accuracy of the top $L/5$ (*F*) and $L/2$ (*G*) predictions on the alignment depth for a subset of 75 targets with deep alignments.

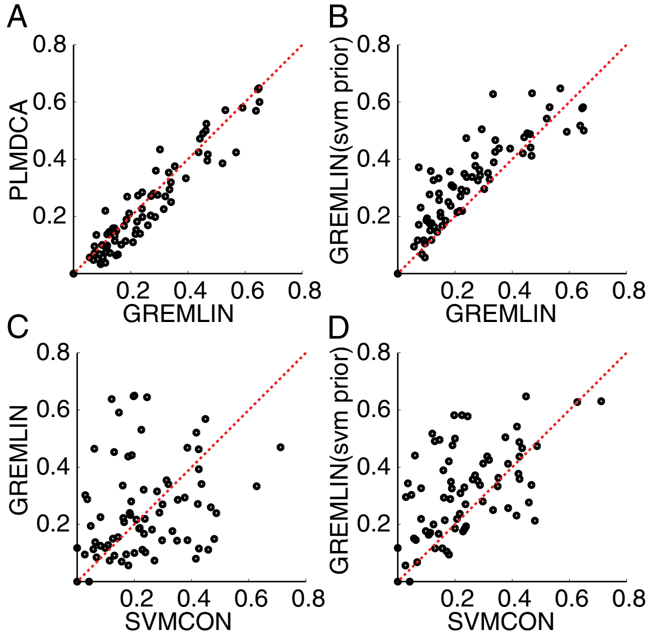

Fig. 2. Improved contact prediction by integration coevolution and predicted structure-feature information. Accuracy of the top *L*/2 predictions between positions at least 12 residues apart, with and without priors on a dataset of 73 proteins that do not have homologs of known structure (Dataset S2; results between positions at least 24 residues apart are included in the *SI Appendix,* Fig. S7 *A–D*). (*A*) Using secondary structure and sequence separation priors, GREMLIN achieves higher accuracy than PLMDCA; (*C*) SVMCON and GREMLIN predictive accuracy are not highly correlated. (*B* and *D*) Integrating a Support Vector Machine (SVM) based prior into GREMLIN improves upon both methods alone.

Integrating the information from profile similarity-based methods into sequence coevolution-based methods using GREMLIN thus performs better than either individual method.

Utility and Limits of Sequence Covariance-Based Contact Prediction. Recent studies (4, 11, 12) have shown that for targets with deep alignments, the predicted contacts are sufficiently accurate to predict the 3D structures of proteins. However, because of the steady increase in the structures deposited in the Protein Data Bank (PDB; ref. 13), any given target protein of interest is also likely to have a related protein with known structure. How useful are covariance-based contact predictions for structure modeling given the homologous structure information likely to be available?

To address this question, we characterized the difference in the fit to predicted contacts of comparative models built from templates and the corresponding native structure for a large set of proteins with recently solved structures compiled from the CAMEO and Critical Assessment of Protein Structure Prediction (CASP10) experiments. If the comparative models fit the contact information as well as the native structure, this added information is likely to be not useful for model improvement. Conversely, if the contact information fits the native structure better, it should be useful for improving comparative models. For each target in the CAMEO dataset, we queried the PDB for homologs using HHsearch, and constructed homology models for the aligned residues from the hits that covered at least 75% of the protein. Because these models are generated from proteins, they are protein-like in their secondary structure composition and can be thought of as a sample of the conformational landscape of protein-like structures around the query protein. We determined the fit to predicted contacts using GREMLIN scores (*Methods*). The GREMLIN score (or any other measure of structure quality) is only useful for improving starting models if the native structure has a better score than these models. To assess this, we computed the difference between the scores of models and those of the corresponding native structure (GREMLINΔ) for each of the 329 proteins in the dataset. When GREMLINΔ > 0 for a model, GREMLIN scores correctly discriminate between native structure and model.

We examined the ability of GREMLINΔ to (*i*) rank alternate models and (*ii*) discriminate between models and the native structure. The distributions of GREMLINΔ for each target in the CAMEO protein dataset fall into three categories of interest based on these criteria. Category I (40% of targets, e.g., Fig. 3*A*, *Left*): GREMLINΔ does not properly discriminate between alternative models, nor between models and the native structure of the target protein—and hence the contact information is not useful for structure prediction. Category II (29% of cases; Fig. 3*A*, *Middle*): GREMLINΔ properly discriminates among models, but not between the best models and the native structure—and hence could be useful for model ranking, but not for increasing the accuracy of the best models. Category III (10% of cases; Fig. 3*A*, *Right*): GREMLINΔ properly discriminates among models and between the best models and the native structure—and hence should be useful for improving comparative model accuracy. A positive control using *L* perfect contacts (as ranked by GREMLIN) resulted in 97% of targets having correct ranking and discrimination (*SI Appendix*, Fig. S9) indicating that perfect contact information is adequate for this task for nearly all targets.

On a subset of 68 targets that had at least 50 models and 20 *L* sequences, increasing the alignment depth improved model ranking, as measured by multiple metrics (*SI Appendix*, Fig. S5), approaching its highest value for most targets with as few as 5 *L* sequences. This suggests that accurate global ranking is possible for the majority of targets with at least 5 *L* sequences. However, for most targets, HHsearch was already able to identify a homolog with similar accuracy without using this information (*SI Appendix*, Fig. S4); a ranking method that uses HHsearch and GREMLIN scores might improve upon both. This suggests that although infrequent, category III might represent the case of most utility for contact-based structure prediction.

Can category III be distinguished from the more frequently occurring categories I and II without knowing the native structure? We hypothesized that there might be a relationship with the closeness of the sequences in the input alignment to the native sequence compared with the templates—if the input sequences are closer to the query sequence, they could be more likely to provide information specific to its native structure. For each protein in the CAMEO dataset, we determined the difference between the HHscore of the alignment with itself and to the closest template(*HH*Δ). *HH*Δ is zero when the query and template alignments are identical; it increases as the difference in the alignments increases reaching 1 when there is no homolog with known structure.

For targets with high-resolution crystal structures, there is a clear relationship between the extent the contacts are discriminative for the native structure (GREMLINΔ), and the closeness of the input sequences to the native sequence (*HH*Δ). When *HH*Δ is small (Fig. 3*B*, blue bars), GREMLIN discrimination is rarely better than random, but when *HH*Δ is large (Fig. 3*B*, red bars), the GREMLIN score of the native structure is significantly better than that of models even in cases where the templates are likely from the same fold. *HH*Δ can be computed for a query protein of unknown structure, which makes it a useful indicator of the utility of covariance information: in cases where the best model has a high *HH*Δ, covariance information is likely to distinguish the native structure from the model. However, this happens relatively infrequently (4 of 329 CAMEO targets). A similar analysis on targets from the CASP10 experiment (Dataset S3) identified 6 of 67 targets with > *L* sequences that had *HH*Δ ≥ 0.5 for the top-ranked model.

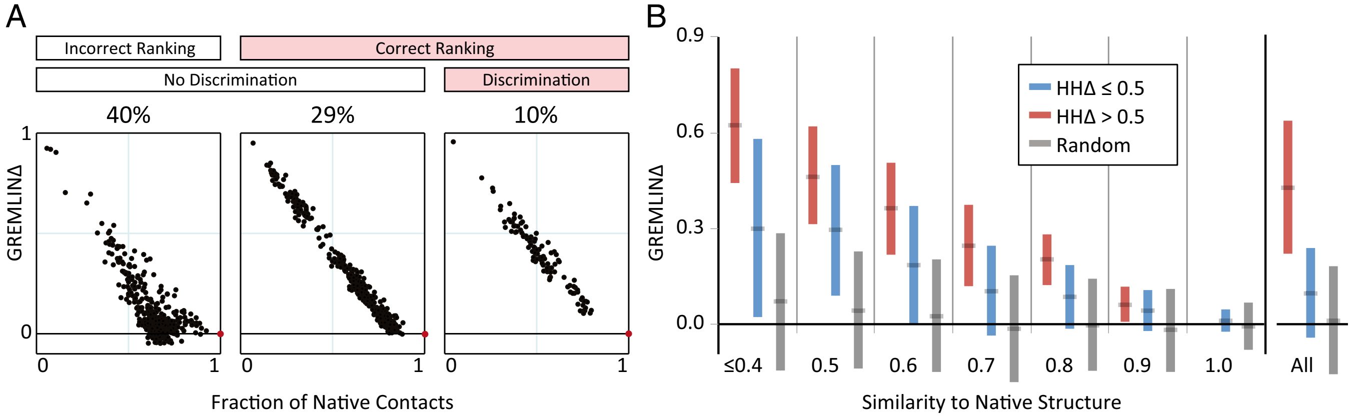

Fig. 3. Utility of contact prediction for structure modeling. (*A*) Ranking of alternate models by GREMLINΔ. Three scenarios are illustrated; each represents a distinct protein target, black dots indicate alternate models, red dots indicate native structures. (*Left*) GREMLINΔ is not useful in selecting the closest model and does not correctly discriminate between native (target pdb:4hwnA) and homology models; (*Middle*) GREMLINΔ ranks homology models correctly (top five models within 0.05 of best five on average; R^2 between GREMLIN score and fraction of native contacts > 0.8) but adds no additional information (target pdb:4fn4D); (*Right*) GREMLINΔ discriminates between best model and native structure (target pdb:4hxtA). In an additional 6% of the targets, GREMLINΔ correctly discriminated the native from the homology models but there were not enough models to reliably establish accuracy of ranking. (*B*)*HH*Δ predicts GREMLINΔ: GREMLINΔ versus structural similarity of homolog to native structure computed by TM-align (14) (for homologs of all targets with high-resolution crystal structures < 2.1 Å). When $HH\Delta \le 0.5$ (blue bars), GREMLINΔ is rarely better than random (green bars, constructed by pooling 100 permutations of predicted scores for each target). When $HH\Delta > 0.5$ (red bars), GREMLINΔ is significantly positive and contact scores successfully discriminate between native and homology model even when the homolog is likely to be from the same fold (similarity $\in [0.5, 0.8)$). Error bars show mean and SD of distributions in all cases.

Frequency of Utility. Our analysis suggests that covariance-based contact predictions are likely to rank homolog templates accurately if the alignments contain at least 5 L nonredundant sequences; if the alignment samples sequences closer to the query than the homolog, these predicted contacts should be useful for building improved structure models. We determined how frequently this scenario occurs by studying the large set of protein families in the Pfam database.

We classified families with at least 50 residues into three groups (*SI Appendix*, Fig. S2): insignificant homology to protein of known structure, remote homology to known structure, and close homology to known structure, based on HHsearches against the PDB (Fig. 4, *Left, Center,* and *Right*). We subdivided these groups based on the number of sequences in the family; the number with more than 5 L sequences, for which contact predictions are likely to be accurate, is indicated by the upper green bars. This subset of families was further subdivided based on $HH\Delta$ (Fig. 4, *Lower*); as above, an $HH\Delta > 0.5$ (Fig. 4, lower dark green bars) indicates that an alignment constructed from the sequences in the family is significantly more similar to itself than to an alignment constructed from the closest homolog of known structure in the PDB, in which case predicted contacts are likely to be useful for structure prediction. The cases of interest, where predicted contacts are likely to be accurate and useful, are indicated in red text in the lower panel. Overall, the number of families for which predicted contacts are likely to be useful is small but not insignificant (422/12,452) and corresponds to roughly 14% of the families with adequate number of sequences [Dataset S4; this is likely a conservative estimate—the improvement in accuracy with the SVM prior could allow improved structure prediction for families with as few as 2.5 L sequences (*SI Appendix*, Fig. S8)]. Predicted contacts are likely most useful for membrane proteins (12, 15), protein assemblies (16), and other systems where high-resolution structural information is currently sparse.

Conclusion

We have shown that integrating coevolution and predicted structure feature-based information using a pseudolikelihood approach improves accuracy of residue–residue contact prediction. Analysis of the accuracy and utility of predicted contact information suggests that a large fraction (24%) of the families in the Pfam database currently have enough sequences for contact prediction to be accurate. This fraction is likely to increase as high-throughput sequencing continues to rapidly expand the sequence databases and methods for contact prediction continue to improve. However, predicted contacts are likely to be currently useful for structure prediction only for a relatively small fraction of these proteins (14%); how this fraction evolves in the future depends on the relative rates of increase of the sequence and structure databases.

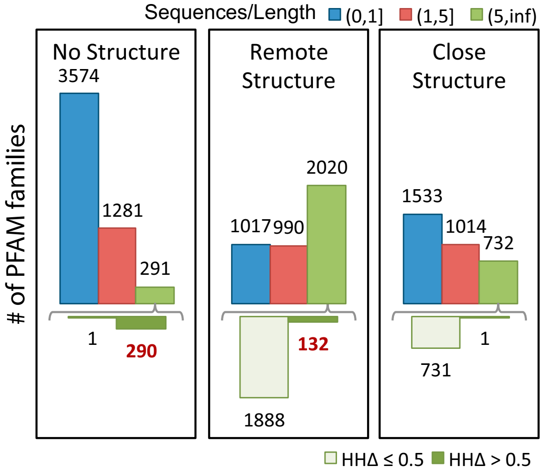

Fig. 4. Frequency of utility of contact prediction. The protein families in the Pfam database were divided into three groups based on the HHsearch *P* value of the closest protein of known structure (*Left*, HHsearch *P* value > $10^{-6.5}$; *Middle*, HHsearch *P* value between 10^{-40} and $10^{-6.5}$; *Right*, HHsearch *P* value > 10^{-40}). Within each group, the number of families with sequences/length less than 1, between 1 and 5, and greater than 5 are shown in blue, red, and green, respectively (Upper bars). For families with > 5 sequences per position (Upper green bars), distribution of $HH\Delta$ to the closest protein of known structure is shown in the lower panel. In cases where the difference in profiles is large ($HH\Delta > 0.5$: right bar in each group, *Lower*), these predictions are likely to improve on comparative models.

Methods

Overview of Computational Method. GREMLIN uses a global statistical model with the following functional form:

$$P(\mathbf{X}=\mathbf{x})=\frac{1}{Z}\exp\left(\sum_{i=1}^{L}\left[\mathbf{v}_i(x_i)+\sum_{j>i}^{L}\mathbf{w}_{i,j}(x_i,x_j)\right]\right). \qquad [2]$$

When learned from a multiple sequence alignment of related proteins, as first described in ref. 17, the random variables X_i represent the amino acid composition at position i, $\mathbf{v}_i$ is a set of parameters of the distribution that encodes the individual propensity for each amino acid at position i of the protein, and $\mathbf{w}_{i,j}$ is the set of parameters modeling the statistical coupling in amino acid propensities between positions i, j of the protein. Z, the partition function, is a global normalizer to ensure the probabilities sum to 1. This is a maximum-entropy model and is also referred to as a Markov Random Field (5).

Given a set of aligned protein sequences, one approach to learning **v**, **w** is by matching the moments of the distribution–$P(X_i)$,$P(X_i,X_j)-P(X_i)P(X_j)$ etc.—to the corresponding empirical correlations $M(X_i)$, $F-M'M$, etc., observed in the alignment. Solving these equations exactly results in a consistent learning procedure—in the limit of infinite data, the learnt parameters tend to the true parameters. For the model in Eq. **2**, however, determining the exact solution is computationally intractable. DCA uses a Taylor series expansion of Eq. **2** (*SI Appendix, Learning*) and a mean-field approximation on the truncated Taylor series. The parameters under this approximation are then solved by inverting a covariance matrix (2). PSICOV uses a different sequence of approximations by first constructing a distribution with binary indicator variables and then using the approach of ref. 18 to relate the properties of this distribution to those of a related Gaussian distribution. Although PSICOV does not explicitly match moments, the resulting learning procedure also estimates Ω; DCA uses a large pseudocount to ensure that the inverted matrix is not singular (thus being similar to l_2 regularization) and PSICOV directly estimates the inverse with a mixture of l_1 and l_2 regularization (3).

The GREMLIN Learning Algorithm. GREMLIN uses a learning procedure based on optimizing the pseudolikelihood (5) of **v**, **w**, which, in log space is expressed as the sum of conditional distributions as follows:

$$pll(\mathbf{v},\mathbf{w}|D)=\sum_{n=1}^{N}\sum_{i=1}^{L}\log P(x_i^n|x_{-i}^n,\mathbf{v},\mathbf{w}).$$

Each conditional distribution models the probability of the observed amino acid at position i in the n^{th} sequence of the alignment, x_i^n, in the context of the amino acids at all other positions in that sequence, x_{-i}^n and depends on the parameters **v**,**w** as:

$$P(x_i^n|x_{-i}^n,\mathbf{v},\mathbf{w})=\frac{1}{Z_i}\exp\left(\mathbf{v}_i(x_i^n)+\sum_{j=1,j\neq i}^{L}\mathbf{w}_{i,j}\left(x_i^n,x_j^n\right)\right).$$

The pseudolikelihood models the conditional distributions of the original joint distribution instead of the joint distribution itself. The global partition function cancels out in the conditional distribution leaving only a per-position local partition function Z_i that is trivial to compute making the pseudolikelihood tractable; it is also concave in **v**, **w** making it easy to maximize. Estimating the parameters by maximizing the pseudolikelihood is a consistent procedure (19).

Regularization and Priors. The GREMLIN learning objective includes a regularization term of the form:

$$R(\mathbf{v},\mathbf{w})=\lambda_v||\mathbf{v}||_2^2+\sum_{i,j}\lambda_w^{i,j}||\mathbf{w}_{i,j}||_2^2,$$

where $||()||_2$ refers to the l_2 vector norm of the parameter and $\lambda_w^{i,j}$ defines a Gaussian prior probability with zero mean and variance $=1/\lambda_w^{i,j}$ on the entries in $w_{i,j}$. When all $\lambda_w^{i,j}$ have the same value as in ref. 6, the learning procedure relies completely on the data to encode any information specific to protein topologies. In data-scarce settings, the accuracy of the learning procedure can be boosted by varying $\lambda_w^{i,j}$ to favor residue pairs that are likely to be in contact according to our prior knowledge of protein topologies. We do this using a per-residue pair $\lambda_w^{i,j}$ that depends on $\pi(i,j)$, the prior probability of i, j being in contact as follows:

$$\lambda_w^{i,j}=\lambda_c\left(1-\lambda_p\log(\pi(i,j))\right).$$

This framework is flexible and allows incorporation of prior information from various sources. We experimented with two priors: a simple prior based on secondary structure and sequence separation, π_{ss} (*SI Appendix,* Fig. S6), used by default and π_{svm}, obtained by treating the probability of being in contact as estimated by SVMCON (10) as a prior. Here, (λ_c,λ_p) were set to (0.20, 0.20) for π_{ss} and to (0.20, 1) for π_{svm} based on tests with a set of targets from CASP9.

GREMLINΔ. We define the GREMLIN score of a structure as

$$Score(model)=\frac{\sum_{(i,j)\in S}||w_{ij}||_{APC}f_{a_i,a_j}(d_{ij})}{\sum_{(i,j)\in S}||w_{ij}||_{APC}},$$

where $||w_{ij}||_{APC}$ is the APC-corrected l_2 norm (8) of $\mathbf{w}_{i,j}$ as computed by GREMLIN; d_{ij} is the distance between the $C\beta$ atoms of the corresponding residues in the model with amino acids a_i, a_j; and S is the set of the top L-predicted contacts with $|i-j|\geq 6$. f_{a_i,a_j} (d_{ij}) is 1 if d_{ij} is less than a cutoff which depends on the amino-acid pairs a_i,a_j, and 0 if d_{ij} is very large (Dataset S5). Using the top L-predicted contacts ranks structures better than using fewer contacts (*SI Appendix,* Fig. S3); increasing this to a larger number does not improve results significantly. We define GREMLINΔ as

$$\text{GREMLIN}\Delta(query,hit)=\frac{Score(query)-Score(hit)}{\max(Score(query),Score(hit))}.$$

The GREMLINΔ of a set of models is the lowest GREMLINΔ in the set.

***HH*Δ.** We used the Viterbi scores $(Vit())$ as computed by HHalign (20) to determine if the alignment of sequences was closer to the query protein than a homolog.

$$HH\Delta(query,hit)=\frac{Vit(query,query)-Vit(query,hit)}{Vit(query,query)}.$$

Materials

CAMEO Targets. The CAMEO webserver of the Protein Model Portal (7) evaluates the accuracy of structure prediction methods against recently released structures from the PDB on a continuous basis. The ROSETTA structure prediction software had predicted the structure of 440 targets at the time we performed this study (first date: 27 April 2012). For targets with fewer than 700 residues, we generated alignments using HHblits, discarding those that had less than L-nonredundant sequences and targets that had very similar sequences (>90% identity) to other targets in this set, resulting in 329 targets (Dataset S1). These targets had a broad diversity of folds, sizes (mean:253, min:60, max:620) and alignment depths (mean:5516, min:100, max:30107 nonredundant sequences) reflecting the diversity of the PDB.

CASP10 Targets. We selected the 67 targets (Dataset S3) from the recently concluded CASP10 experiment with more than L nonredundant sequences.

Hard Targets. To compare the effects of prior information on accuracy, we collected an additional set of 73 targets that do not have homolog structures in the PDB (Dataset S2).

Pfam. For protein families in the Pfam database [ref. 21; version 26] with at least 50 positions, we constructed alignments using HHblits and queried the PDB for homologs using HHsearch starting with precomputed models (from the HHSearch database; ref. 22).

PISCES. We accessed the PISCES database of nonredundant protein structures (23) (accession date November 3, 2012; sequence identity: 80%, minimum resolution: 2.0 Å, maximum R_{free}: 0.25). We used statistics collected on this database to determine the secondary structure prior, π_{ss} (*SI Appendix,* Fig. S6) and the amino-acid specific distance function f_{aa_i,aa_j} (Table S5).

DCA. We used the default settings as in (2): $\mathbf{x}=0.2,\lambda=0.5$.

PSICOV. We used the suggested flags that determine ρ to achieve a target density: -p, -d 0.03.

MIc. MIc is a mutual information-based method that uses sequence profile information and sequence weighting in its computation of a score referred to as MIc. All our reported results use the default flags (24).

PLMDCA. We used the default settings of $\lambda_h = 0.01, \lambda_J = 0.01$.

SVMCON. We used SVMCON with the default settings.

Entropy Correction via APC. The average product correction (8) was suggested as a way of correcting for entropic and phylogenetic biases in the sequence alignment. Although originally used for mutual information-based scores, it has also been used with norm-based scores in PSICOV (3). We applied this correction to GREMLIN and DCA as we found that it uniformly improved their accuracy; PSICOV, PLMDCA and MIc apply it by default.

Sequence Reweighting. DCA, PSICOV and PLMDCA reweigh sequences in the input multiple sequence alignment to account for redundancy (2, 3, 6). When making predictions using GREMLIN, we reweigh sequences in a filtered alignment (filtered at 90% sequence identity) instead of using the full alignment.

HHsuite. We used HHblits to build the alignment from the clustered uniprot database (dated Mar 2012) with the following options in addition to default settings: -nodiff, -neffmax 20, -n 4, and -maxfilt 100,000. To ensure consistency of coverage across the alignment, we removed sequences that had more than 25% gaps compared with query and sites that had more than 25% gaps. In addition, for GREMLIN, we postprocessed the alignment using HHfilter to generate a nonredundant alignment at 90% sequence identity. We queried the PDB using the global alignment mode of HHsearch and the latest PDB database (date: January 1, 2012), which preceded the participation of ROSETTA in CAMEO and CASP10. We used HHalign to compute Viterbi scores. These programs are part of HHsuite (version: 2.0.15) (20).

ACKNOWLEDGMENTS. We thank Profs. Nick Grishin, David Jones, John Moult, Johannes Soeding, and Jinbo Xu for their many valuable comments, and Christof Angermuller for his assistance in using HHblits. S.O. was supported in part by PHS NRSA T32 GM07270 from National Institute of General Medical Sciences.

1. Tress ML, Valencia A (2010) Predicted residue–residue contacts can help the scoring of 3d models. *Proteins. Struct Funct Bioinf* 78(8):1980–1991.
2. Morcos F, et al. (2011) Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc Natl Acad Sci USA* 108(49):E1293–E1301.
3. Jones DT, Buchan DWA, Cozzetto D, Pontil M (2012) PSICOV: Precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics* 28(2):184–190.
4. Marks DS, et al. (2011) Protein 3D structure computed from evolutionary sequence variation. *PLoS ONE* 6(12):e28766.
5. Balakrishnan S, Kamisetty H, Carbonell JG, Lee SI, Langmead CJ (2011) Learning generative models for protein fold families. *Protiens Struct Funct Bioinf* 79(4):1061–1078.
6. Ekeberg M, Lövkvist C, Lan Y, Weigt M, Aurell E (2013) Improved contact prediction in proteins: using pseudolikelihoods to infer Potts models. *Phys Rev E Stat Nonlin Soft Matter Phys* 87(1):012707.
7. Arnold K, et al. (2009) The protein model portal. *J Struct Funct Genomics* 10(1):1–8.
8. Dunn SD, Wahl LM, Gloor GB (2008) Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *Bioinformatics* 24(3):333–340.
9. McGuffin LJ, Bryson K, Jones DT (2000) The PSIPRED protein structure prediction server. *Bioinformatics* 16(4):404–405.
10. Cheng J, Baldi P (2007) Improved residue contact prediction using support vector machines and a large feature set. *BMC Bioinf* 8(1):113.
11. Sułkowska JI, Morcos F, Weigt M, Hwa T, Onuchic JN (2012) Genomics-aided structure prediction. *Proc Natl Acad Sci USA* 109(26):10340–10345.
12. Nugent T, Jones DT (2012) Accurate de novo structure prediction of large transmembrane protein domains using fragment-assembly and correlated mutation analysis. *Proc Natl Acad Sci USA* 109(24):E1540–E1547.
13. Berman HM, et al. (2000) The protein data bank. *Nucleic Acids Res* 28(1):235–242.
14. Zhang Y, Skolnick J (2005) TM-align: A protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res* 33(7):2302–2309.
15. Hopf TA, et al. (2012) Three-dimensional structures of membrane proteins from genomic sequencing. *Cell* 149(7):1607–1621.
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(1):67–72.
17. Thomas J, Ramakrishnan N, Bailey-Kellogg C (2008) Graphical models of residue coupling in protein families. *IEEE/ACM Trans Comp Bio Bioinf* 5(2):183–197.
18. Banerjee O, El Ghaoui L, d'Aspremont A (2008) Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *J Mach Learn Res* 9:485–516.
19. Gidas B (1988) Consistency of maximum likelihood and pseudo-likelihood estimators for Gibbs distributions. *J Inst Math Its Appl* 10:129–145.
20. Söding J (2005) Protein homology detection by HMM-HMM comparison. *Bioinformatics* 21(7):951–960.
21. Bateman A, et al. (2002) The PFAM protein families database. *Nucleic Acids Res* 30(1): 276–280.
22. Biegert A, Mayer C, Remmert M, Söding J, Lupas AN (2006) The MPI Bioinformatics Toolkit for protein sequence analysis. *Nucleic Acids Res* 34(Web Server issue, Suppl 2): W335-9.
23. Wang G, Dunbrack RL, Jr. (2005) PISCES: Recent improvements to a PDB sequence culling server. *Nucleic Acids Res* 33(Web Server issue, Suppl 2):W94-8.
24. Jeong CS, Kim D (2012) Reliable and robust detection of coevolving protein residues. *Protein Engineering Design and Selection,* 25(11):705–713.