Supplementary Information: Assessing the utility of residue-residue contact information in a sequence and structure rich era

Hetunandan Kamisetty, Sergey Ovchinnikov, David Baker

Learning

The basis of methods that predict contact information from Multiple Sequence Alignments is that proteins with statistically significant sequence similarity have similar structure and that this common structural constraint affects their compositional variability. Thus, a statistical model of the constraints on the compositional variability might recover structural constraints (contacts). Historically, these methods were based on Mutual information (MI) and its variations and tended to suffer from the problem of conflating direct and indirect statistical coupling: if positions (A,B) and (B,C) have a high MI; so do (A,C). Thus, even if only spatially proximal positions co-evolve in composition, non-adjacent positions can display a high MI due to transitivity.

More recent methods use a global statistical model as described in Eq. 2. The use of a global statistical model allows disentangling direct interactions versus indirect transitive effects. Here, we briefly review alternate approaches to learning this model.

Maximum Likelihood Estimation

A common approach to learning statistical models from data is the maximum likelihood procedure [1] which for the case of the model in Eq. 2 having parameters Θ =($\mathbf{v}$,$\mathbf{w}$) is

$$\hat{\mathbf{v}}, \hat{\mathbf{w}} = \arg\max_{\mathbf{v},\mathbf{w}} ll(\mathbf{v}, \mathbf{w}|D) - R(\mathbf{v}, \mathbf{w})$$

where $ll(\mathbf{v}, \mathbf{w}|D) = \sum_{n=1}^{N} \log P(x^n|\mathbf{v}, \mathbf{w})$ is the log likelihood of the model given a set of protein sequences $D = [x^1, x^2..x^N]$ and $R(\mathbf{v}, \mathbf{w})$ is an optional penalty term added to prevent over-fitting, especially in high-dimensions. Since the maximum likelihood procedure finds parameters that maximize the joint probability distribution of the observed data, it uses all the moments of the distribution to match the observed correlations. In the limit, the maximum likelihood procedure is guaranteed to converge to the true parameters (ie learning is *consistent*) if the model is identifiable. For the model described in Eq 2, the likelihood depends on the value of Z, the partition function which is, unfortunately, computationally intractable in the general case. Approximations to Z based on the Bethe Free Energy and its generalizations that use message passing have been used in discrete models of proteins [2, 3] but can be prohibitively slow when used as an inner step to learn $\mathbf{v}$ and $\mathbf{w}$ from sequence alignments [3].

Method of Moment Estimation

In cases where the likelihood is unknown, hard to compute or hard to optimize, a common alternative is the method of moments[4]. If the model has p parameters, the method of moments computes p moments

of the distribution as a function of the parameters, equates them to the corresponding observed correlations and solves the resulting equations to estimate the parameters. In the limit of infinite data, like Maximum Likelihood, an exact moment matching procedure is also consistent[5, Chapter 33].

The moments of the model in Eq. 2 are $P(X_i)$ (for each of the 21 valid choices of X_j), $P(X_i, X_j)$ etc. The method of moments approach learns parameters of the distribution such that these moments match the corresponding observables (frequencies, $M(X_i)$, and correlations, $F(X_i, X_j)$, respectively). To do this, one requires a relationship between the parameters of the model and its moments. For the distribution in Eq 2, the log partition function is its cumulant generating function[6]; its derivatives therefore provide such a link:

$$\frac{\partial}{\partial \mathbf{v}_i} \log Z = P(X_i)$$

$$\frac{\partial^2}{\partial \mathbf{v}_i \mathbf{v}_j} \log Z = P(X_i, X_j) - P(X_i)P(X_j)$$

This set of equations can be solved by then equating the moments to the corresponding observables to obtain estimates of the parameters. The advantage of such an approach is that it is a consistent learning procedure. The disadvantage is that the relation between the marginal probabilities and the parameters depends on the partition function that is the source of the computational intractability of the learning procedure. In contrast, as described in the Methods, the pseudo-likelihood [7] method employed by GREMLIN avoids this source of computational intractability by modeling the conditional distributions. Modeling conditional distributions is sufficient to recover the true parameters of the distribution. Intuitively, this is because if all the conditional distributions above are exactly modeled, the joint distribution must be exactly modeled as well. Incidentally, a similar justification proves the convergence properties of Gibbs Sampling[8].

Choice of Prior Family

Previously we suggested the use of an $l_1 + l_2$ based regularizer to learn generative models of protein sequences [9]. This choice of regularization is equivalent to using a sparsity-promoting Laplace prior on the strengths of the interactions. While that appears to learn parameters that are closer to their true strength, the task of contact prediction is simpler as we are only interested in the *relative* ranking of the interactions and not their actual values. On the set of 15 PFAM domains used in [10], we found that with APC, using a Gaussian prior on the individual values of the parameters (ie l_2 regularization) was as accurate as using a Laplace prior in predicting contacts. Additionally, using Gaussian priors results in a continuously differentiable objective function with smaller computational costs. We therefore choose this form for prior information throughout this paper. The development of alternate approaches to the APC that account for entropic effects might allow sparsity-inducing priors to improve upon extant methods in data-scarce settings.

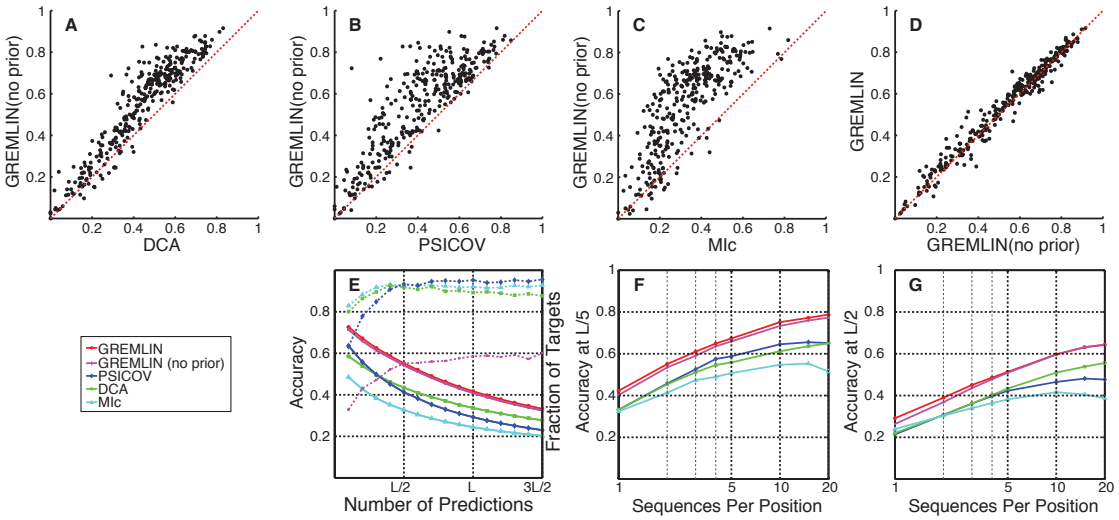


Figure S1

Accuracy of contact prediction when restricted to positions at least 24 residues apart. (A)-(D): 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. We varied the number of sequences in the input alignment for a subset of 75 targets with deep alignments. Average accuracy across this set as a function of the depth of the alignment (in log-scale) over this subset, at $L/5$ (F) and $L/2$ predictions (G)

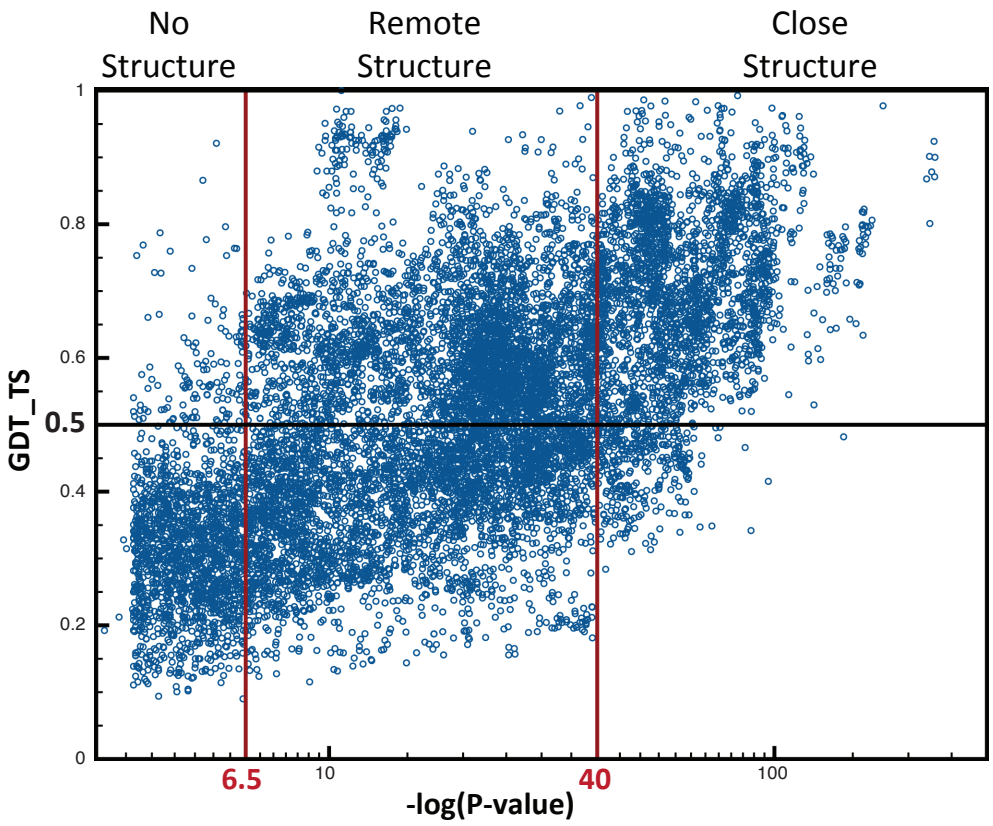


Figure S2

Significance of hit found by HHsearch vs GDT-TS of hit to native crystal structure across targets with crystal structures of resolution < 2.1Å in the CAMEO set. We classified PFAM families into three classes based on the cutoffs shown in red.

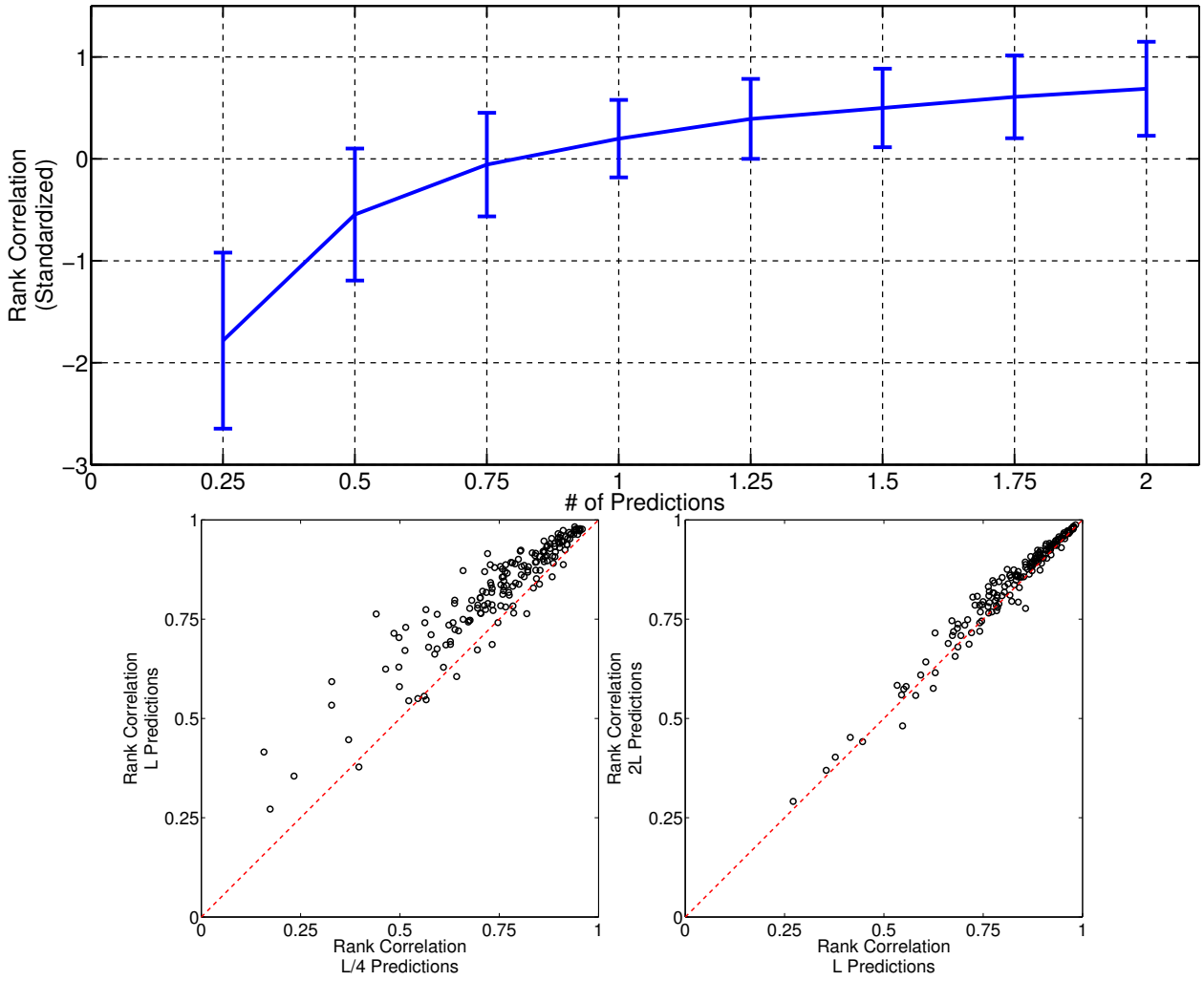
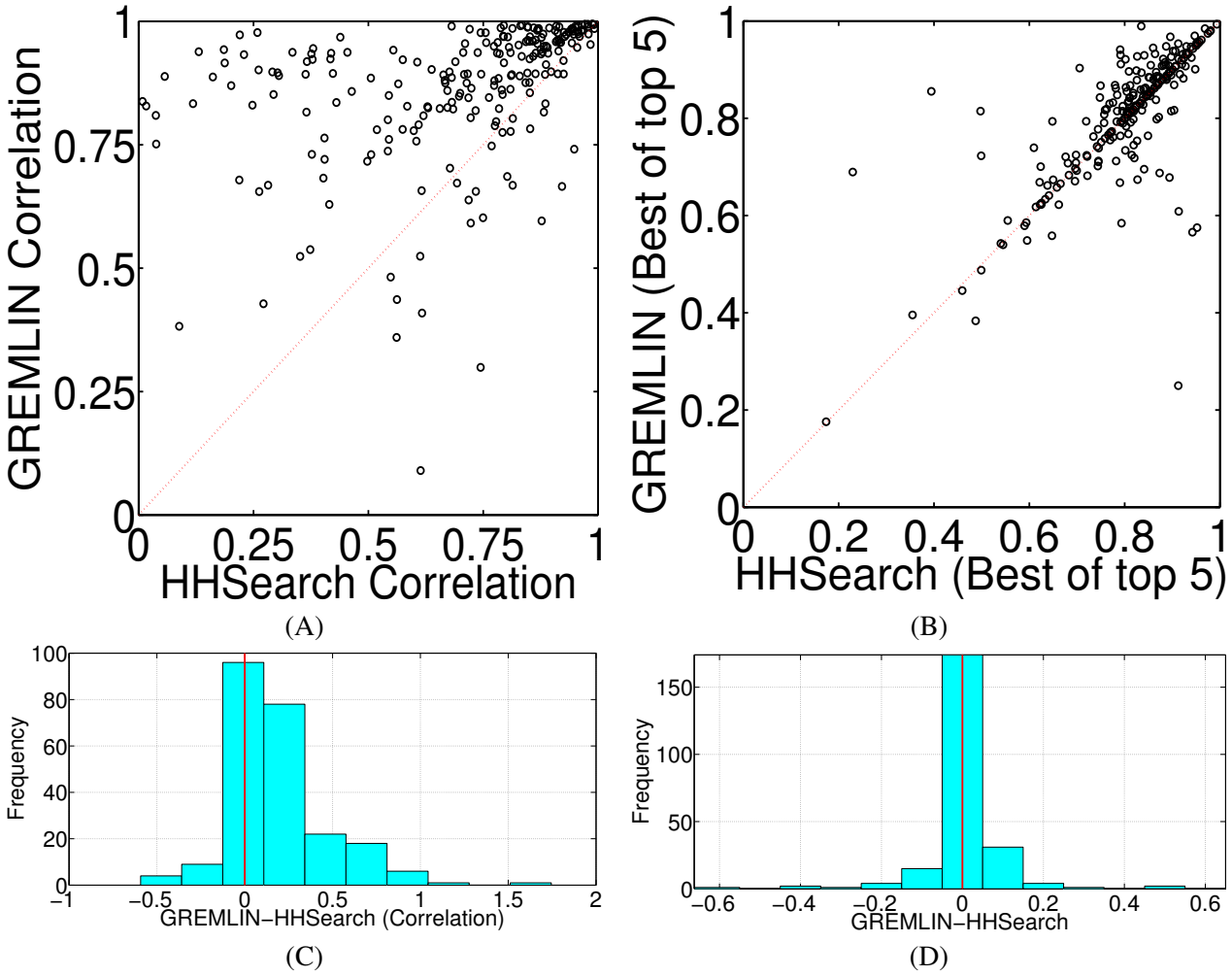


Figure S3

Rank correlation between GREMLIN score and fraction of native contacts while varying the number of predictions used. We ranked alternate models for all CAMEO targets varying the number of predictions used in the computation of the GREMLIN scores. To account for the variability in number of sequences and models for each target, we computed standardized rank-correlations between the rank of the model using GREMLIN scores and its rank based on the distance to native. The average standardized correlation uniformly increases as the number of predictions used increases from L/4 to L (A,B). Increasing the number of predictions beyond L increases the overall standardized rank-correlations, however this increase is not significant or uniform (C).

Figure S4



Comparison of ranking accuracy between GREMLIN and HHsearch . We scored alternate models generated from templates for all CAMEO targets with more than 50 templates and compared the ability of GREMLIN scores and HHsearch scores to rank these templates correctly. We used two metrics of ranking accuracy: the correlation between the score and the fraction of native contacts – a measure of global ranking accuracy, and the accuracy of the best model (by fraction of native contacts) among the top 5 models ranked by each metric. GREMLIN was significantly more accurate globally (A,C); however the accuracy of best model in the top 5 was of comparable accuracy to those selected by HHsearch (B,D) for most targets.

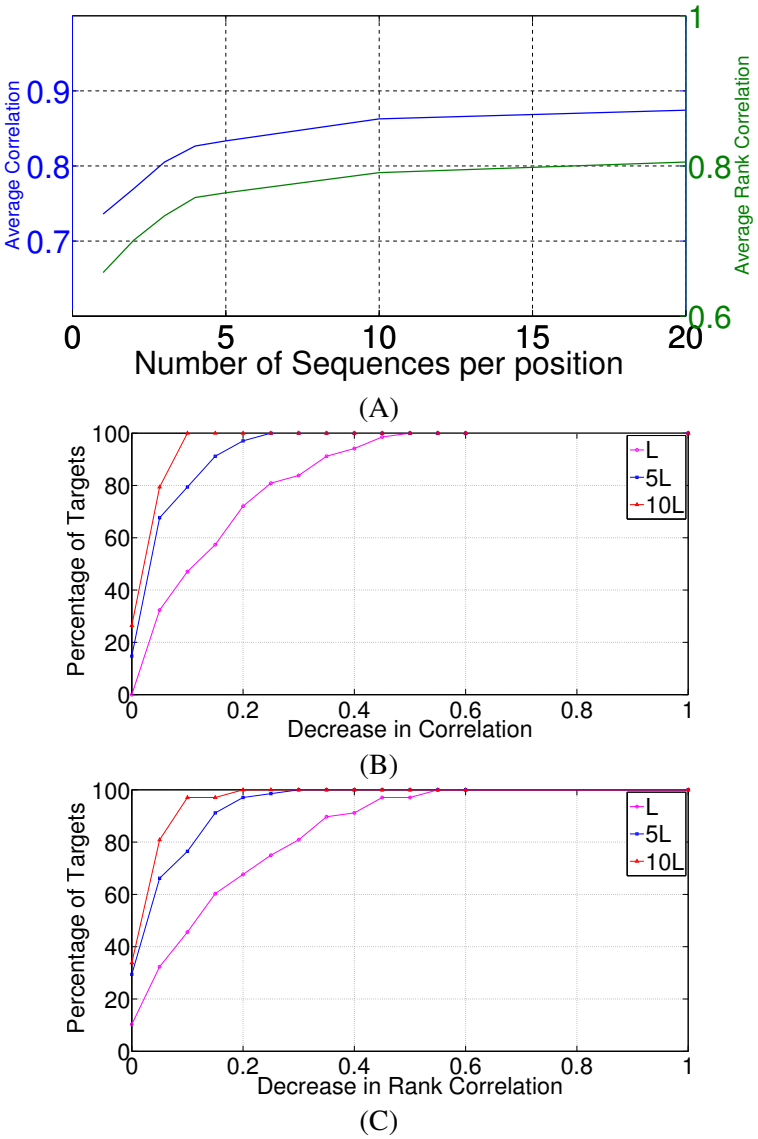


Figure S5

Ranking accuracy depends on the depth of the alignment. For 68 targets that had at least 20L sequences and more than 50 templates, we varied the depth of the alignment and ranked alternate models using GREMLIN scores in each case. The average correlation and rank-correlation between fraction of native contacts of the model and its GREMLIN score increases with increasing alignment depths (A). When there are 10L sequences, the correlation coefficient is close to the value attained with all sequences for nearly all targets (all targets within 0.1 of maximum; 80% of targets within 0.05) (B,red-line); the numbers with 5L sequences are slightly lower but comparable ($\sim$80% of targets within 0.1; $\sim$70% within 0.05) (B, blue-line). The behavior of rank-correlations is similar (C).

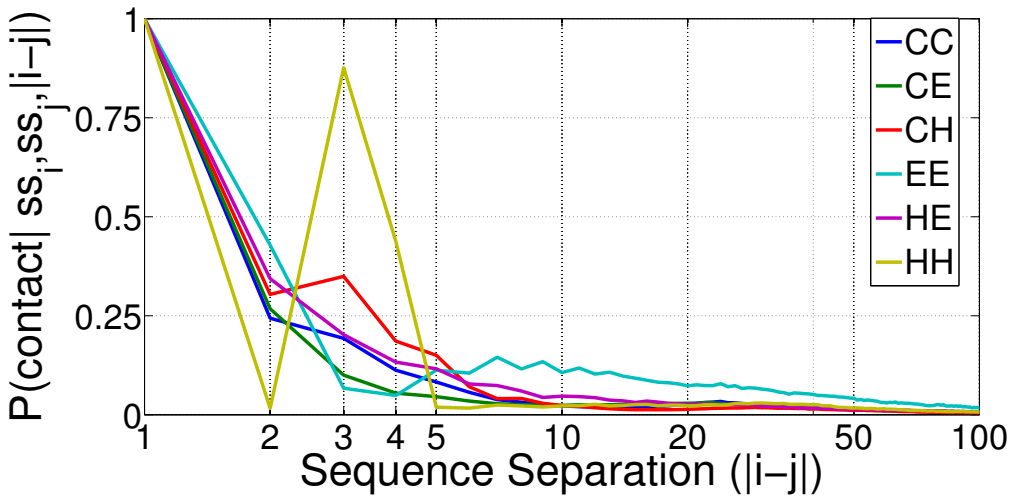


Figure S6

Prior probability based on sequence separation and predicted secondary structure: π_{ss}. We estimated the prior probability that a pair of residues are in contact (closest separation between heavy atoms less than 5Å) conditioned on their secondary structure (Helix (H), Sheet (E) and Coil (C), assigned by STRIDE[11]) and their sequence separation, from its frequency of occurence in a set of non-redundant protein structures ([12]) predicted to be monomeric by PISA[13, 14]. Most secondary-structure specific effects are limited to low sequence separation with sheet-sheet contacts being a notable exception.

For a query protein, the probability of each secondary structure element at a position $P(ss_i)$ was estimated using PSIPRED [15]; the expected probability of being in contact was then estimated by summing over the corresponding conditional probabilities as follows:

$$\pi_{ss}(i,j) = \sum_{ss_i \in \{H,E,C\}} \sum_{ss_j \in \{H,E,C\}} P(ss_i)P(ss_j)P(contact|ss_i, ss_j, |i-j|)$$

The $P(contact|ss_i, ss_j, |i-j|)$ value for sequence separation 100 was used for larger sequence separations.

When varying the depths of the alignments we used the secondary structure predictions previously generated from the whole alignment. Results when using secondary structure predicted with sub-alignments are very similar (Fig. S10).

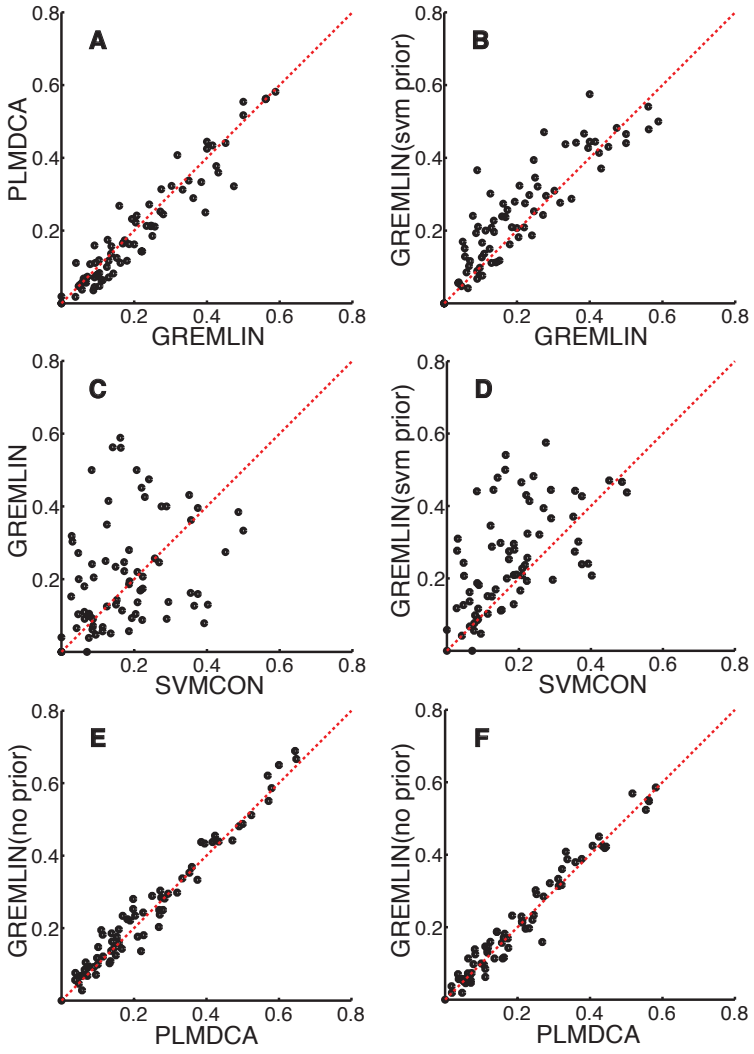


Figure S7

Effect of prior information on accuracy of contact prediction. On a set of 73 proteins that likely do not have prior homolog structures (Table IV), we compared the accuracy of the top $L/2$ predictions with and without priors. When restricted to positions at least 24 residues apart, GREMLIN achieves higher accuracy than PLMDCA by using secondary structure and sequence separation priors (A); however SVMCON has higher accuracy than GREMLIN on a large fraction of targets (C). Integrating SVM-based priors into co-evolution based GREMLIN predictions improves upon accuracy of individual methods (B,D). PLMDCA and GREMLIN (with no priors) have comparable accuracy when restricted to positions at least 12 residues apart (E) and 24 residues apart (F), although GREMLIN was 5-20x faster on this dataset.

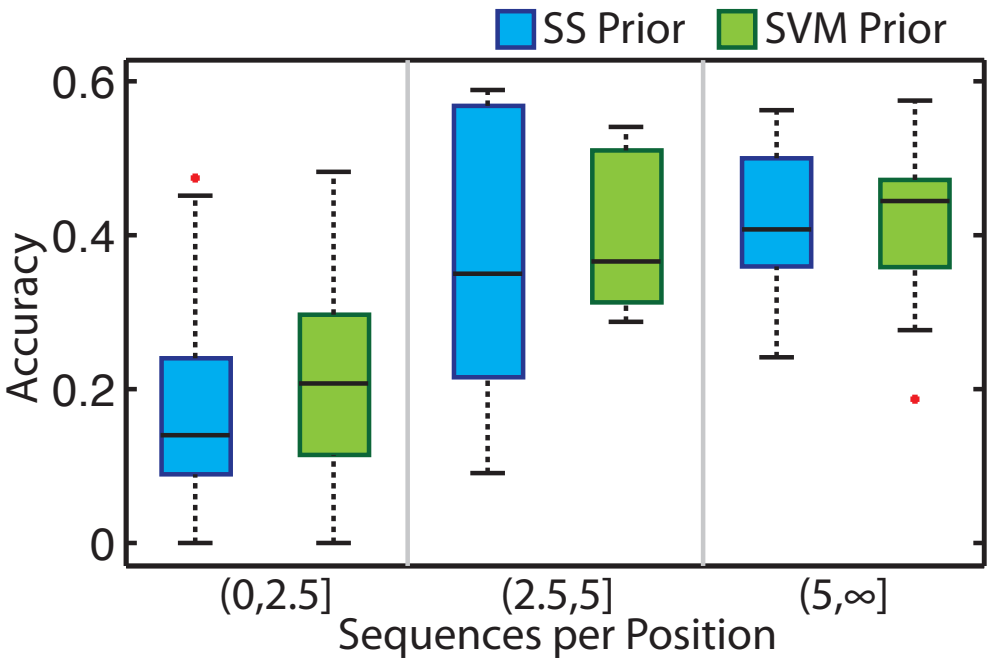
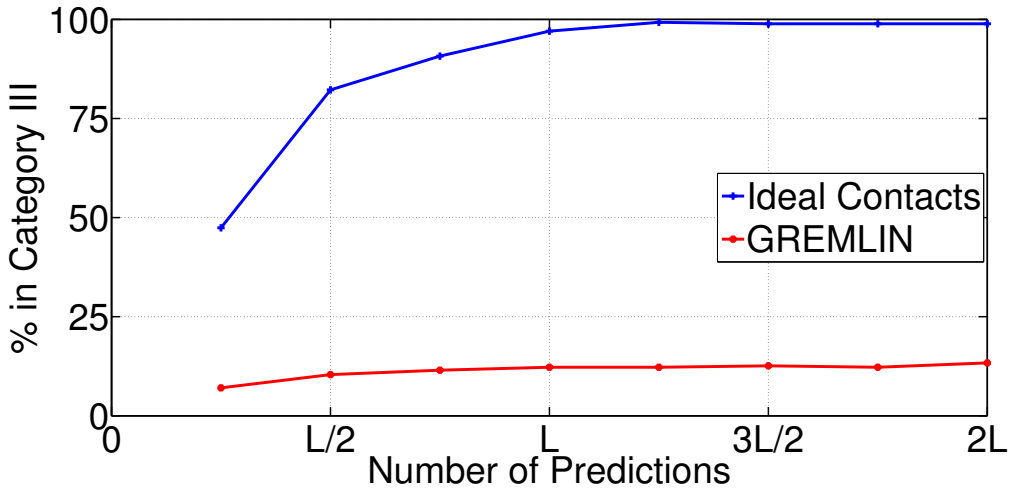
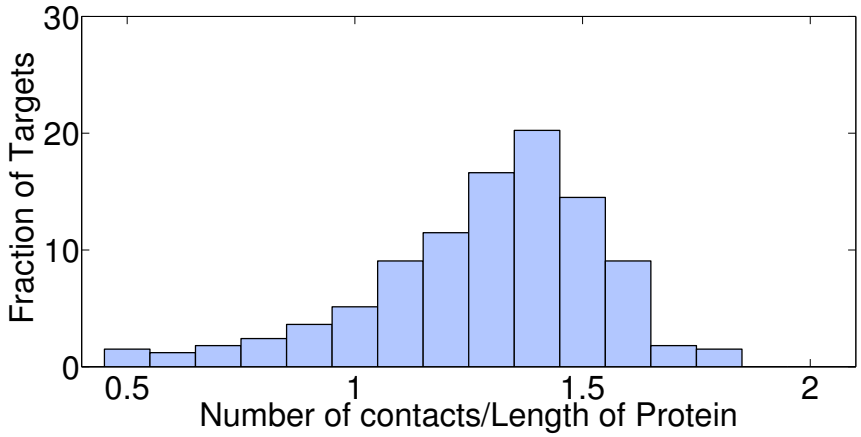


Figure S8

SVM priors improve robustness and predictions with few sequences. On the set of hard targets that likely have no prior homolog structures, using the SVM prior improves the robustness of predictions for targets with few sequences in the alignment (left and middle panels). With the SVM prior, the accuracy of predicted contacts (restricted to positions at least 24 residues apart) for targets with at least 2.5L sequences (middle, green bar) was comparable to targets with more than 5L sequences with and without priors (right). On each box, the central mark is the median, the edges of the box are the 25th and 75th percentiles, the whiskers extend to the most extreme data points with outliers plotted individually.



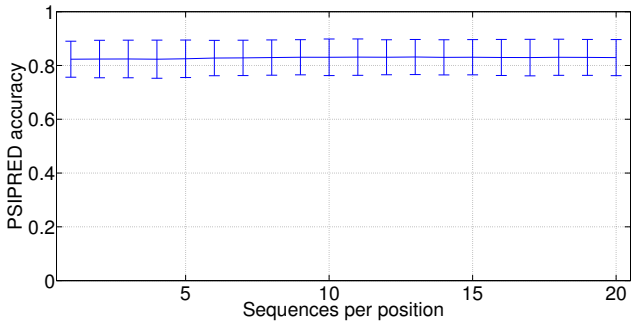
(A)



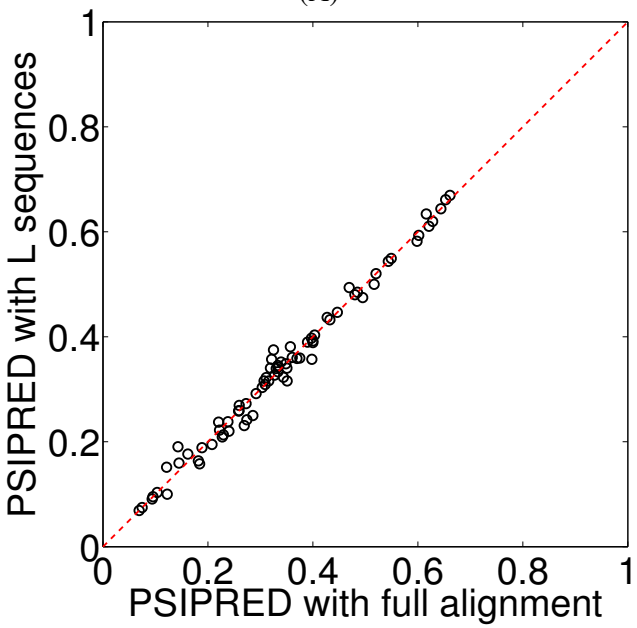
(B)

Figure S9

Contacts are adequate to rank and discriminate native structures accurately. As a positive control, for targets in the CAMEO dataset, we generated perfect predictions using only correct contacts between positions at least 6 residues apart varying the number of contacts from L/4 to 2L (in the order predicted by GREMLIN). In each case, we determined if the contacts correctly rank alternate models and discriminate between native and alternate models (Category III in Fig 3b). The fraction of targets in Category III increases monotonically as the number of contacts increases from L/4 to about 5L/4 and plateaus beyond (A, blue lines). When L contacts are used, 97% of the targets have accurate ranking and native discrimination. In contrast, due to the inaccuracies in predictions, the corresponding fraction using actual GREMLIN predictions is much lower (A, red lines). (B) suggests an explanation for this behavior: most targets in this dataset have between L and 2L contacts; thus, a large fraction of the total contacts are present in the top L correct contacts.



(A)



(B)

Figure S10

Effect of alignment depths on PSIPRED accuracy. When comparing the accuracy of GREMLIN with varying alignment depths, we used PSIPRED's secondary structure predictions with default parameters (which might use deeper alignments). To test if this affects our conclusions, we used the generated sub-alignments as input to PSIPRED and compared the accuracy of predicted secondary structure with varying alignment depths (using [16] to add pseudo-counts). The average accuracy of PSIPRED is essentially identical at these alignment depths (A): the difference between L and 20L sequences is less than 1%. Error bars show standard deviation of accuracy. The accuracy of the top L/2 GREMLIN predictions (when restricted to positions at least 12 residues apart, as previously) when L sequences are used for both PSIPRED and GREMLIN is essentially identical to the predictions when secondary structure was predicted using all sequences(A). Results at higher alignment depths are similar.

References

[1] Ronald A Fisher. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 222:309–368, 1922.

[2] H. Kamisetty, E.P. Xing, and C.J. Langmead. Free energy estimates of all-atom protein structures using generalized belief propagation. *Journal of Computational Biology*, 15(7):755–766, 2008.

[3] 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. *PNAS*, 106:67–72, Jan 2009.

[4] Karl Pearson. Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London. A*, 185:71–110, 1894.

[5] Harald Cramér. *Mathematical methods of statistics*, volume 9. Princeton university press, 1945.

[6] M.J. Wainwright and M.I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1-2):1–305, 2008.

[7] J. Besag. Efficiency of pseudolikelihood estimation for simple Gaussian fields. *Biometrika*, 64(3):616–618, 1977.

[8] G. Casella and E.I. George. Explaining the gibbs sampler. *The American Statistician*, 46(3):167–174, 1992.

[9] S. Balakrishnan, H. Kamisetty, J.G. Carbonell, S.I. Lee, and C.J. Langmead. Learning Generative Models for Protein Fold Families. *Proteins: Structure, Function, and Bioinformatics*, 79(4):1061–1078, 2011.

[10] F. Morcos, A. Pagnani, B. Lunt, A. Bertolino, D.S. Marks, C. Sander, R. Zecchina, J.N. Onuchic, T. Hwa, and M. Weigt. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *PNAS*, 108(49):E1293–E1301, 2011.

[11] Matthias Heinig and Dmitrij Frishman. Stride: a web server for secondary structure assignment from known atomic coordinates of proteins. *Nucleic acids research*, 32(suppl 2):W500–W502, 2004.

[12] G. Wang and R.L. Dunbrack. Pisces: recent improvements to a pdb sequence culling server. *Nucleic acids research*, 33(suppl 2):W94–W98, 2005.

[13] Evgeny Krissinel and Kim Henrick. Protein interfaces, surfaces and assemblies' service pisa at the european bioinformatics institute., June 2009.

[14] Evgeny Krissinel and Kim Henrick. Inference of macromolecular assemblies from crystalline state. *Journal of molecular biology*, 372(3):774–797, 2007.

[15] Liam J McGuffin, Kevin Bryson, and David T Jones. The PSIPRED protein structure prediction server. *Bioinformatics*, 16(4):404–405, 2000.

[16] Christof Angermüller, Andreas Biegert, and Johannes Söding. Discriminative modelling of context-specific amino acid substitution probabilities. *Bioinformatics*, 28(24):3240–3247, 2012.