BIOINFORMATICS ORIGINAL PAPER Vol. 24 no. 14 2008, pages 1575–1582
doi:10.1093/bioinformatics/btn248

Structural bioinformatics

Using inferred residue contacts to distinguish between correct and incorrect protein models

Christopher S. Miller[1] and David Eisenberg[1,2,*]

[1]UCLA-DOE Institute for Genomics & Proteomics, Molecular Biology Institute and [2]Howard Hughes Medical Institute, Departments of Chemistry & Biochemistry & Biological Chemistry, Box 951570, UCLA, Los Angeles, CA 90095, USA

Received on March 15, 2008; revised on May 7, 2008; accepted on May 25, 2008

Associate Editor: Burkhard Rost

ABSTRACT

Motivation: The *de novo* prediction of 3D protein structure is enjoying a period of dramatic improvements. Often, a remaining difficulty is to select the model closest to the true structure from a group of low-energy candidates. To what extent can inter-residue contact predictions from multiple sequence alignments, information which is orthogonal to that used in most structure prediction algorithms, be used to identify those models most similar to the native protein structure?

Results: We present a Bayesian inference procedure to identify residue pairs that are spatially proximal in a protein structure. The method takes as input a multiple sequence alignment, and outputs an accurate posterior probability of proximity for each residue pair. We exploit a recent metagenomic sequencing project to create large, diverse and informative multiple sequence alignments for a test set of 1656 known protein structures. The method infers spatially proximal residue pairs in this test set with good accuracy: top-ranked predictions achieve an average accuracy of 38% (for an average 21-fold improvement over random predictions) in cross-validation tests. Notably, the accuracy of predicted 3D models generated by a range of structure prediction algorithms strongly correlates with how well the models satisfy probable residue contacts inferred via our method. This correlation allows for confident rejection of incorrect structural models.

Availability: An implementation of the method is freely available at http://www.doe-mbi.ucla.edu/services

Contact: david@mbi.ucla.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

In his 1972 Nobel prize lecture, Christian Anfinsen delighted that 'an increasing sophistication in the theoretical treatment of the energetics of polypeptide chain folding are beginning to make more realistic the idea of the a priori prediction of protein conformation' (Anfinsen, 1973). In recent years, Anfinsen's prescient observation is bearing fruit: the *de novo* prediction of the 3D structure of smaller proteins has enjoyed a series of increasingly impressive successes (Moult *et al.*, 2007; Qian *et al.*, 2007; Zhang, 2007). However, more sophisticated energy functions have not removed the computationally limiting task of exploring and evaluating the multitude of ways a linear peptide sequence can fold (Schueler-Furman *et al.*, 2005). As such, when the full range of structural space cannot be adequately sampled, or when competing energy functions result in divergent predictions of 3D structure, additional information is necessary to aid in selecting the model closest to the native structure. One source of additional information not explicitly considered by most structure prediction algorithms is prediction of pairwise residue contacts. It has been hypothesized that with relatively accurate contact predictions, one could at the very least rank predicted 3D models based on their satisfaction of predicted contacts (Grana *et al.*, 2005). In some cases, contacts predicted from sequence (Ortiz *et al.*, 1998), conservation (Schueler-Furman and Baker, 2003) or from threading to known structures (Zhang *et al.*, 2003) have successfully been incorporated directly into structure prediction.

Most pioneering efforts in contact prediction consisted in searching alignments of sequence homologs for correlated mutations between pairs of columns. The hypothesis behind using correlated mutations for contact prediction is attractive: covariant changes over evolutionary time between two residues could be due to shared structural constraints imposed by proximity in the 3D-fold of a protein. Altschuh *et al.* (1987) used the correlated mutations concept to search for identical 'patterns of change' among columns in an alignment of seven homologs of tobacco mosaic virus coat protein. Although the statistical power of patterns among just seven sequences might seem small by modern standards, the authors were able to propose that residues with similar patterns of conservation were spatially proximal. Later efforts attempted to adjust for effects of sample size or skewed phylogeny by performing many randomized trials, recalculating the test statistic of covariation and comparing the resulting distribution with the observed value (Korber *et al.*, 1993; Noivirt *et al.*, 2005; Shackelford and Karplus, 2007; Wollenberg and Atchley, 2000).

Several researchers have sought to explicitly recognize the physicochemical similarities among certain amino acids which allow for conservative substitutions. One method uses a substitution matrix to construct a matrix for all pairs of amino acids in a multiple sequence alignment column, and uses as a test statistic the correlation between such matrices from pairs of columns (Gobel *et al.*, 1994). Other approaches completely recode amino acids as vectors of physicochemical properties before looking for correlated mutations (Neher, 1994; Vicatos *et al.*, 2005).

*To whom correspondence should be addressed.

An extreme strategy is to reduce amino acids into just two classes at a time and search for correlated changes in a phylogenetic tree between these two states (Pollock *et al.*, 1999).

Ranganathan and coworkers use a thermodynamics-inspired interpretation of correlated mutations to find 'statistically coupled' residue pairs (Lockless and Ranganathan, 1999). As with most methods (Izarzugaza *et al.*, 2007), the majority of residue pairs identified are not in contact in 3D structures. However, Ranganathan and coworkers argue that these distant residue pairs are indeed energetically coupled via pathways of connectivity passing through the structure (Lockless and Ranganathan, 1999; Suel *et al.*, 2003), a hypothesis reiterated by others (Yeang and Haussler, 2007) but disputed by some (Fodor and Aldrich, 2004b).

Recent years have seen machine learning approaches successfully applied to contact prediction, often resulting in more accurate predictions (Cheng and Baldi, 2007; Fariselli *et al.*, 2001; Hamilton *et al.*, 2004; Pollastri and Baldi, 2002; Punta and Rost, 2005; Shackelford and Karplus, 2007; Vullo *et al.*, 2006). This trend is partly in response to the recognition that other properties of a target sequence and its multiple sequence alignment contain information about residue contacts. For example, empirically derived amino acid propensity matrices have successfully been used on their own to predict residue contacts with accuracy comparable to some correlated mutations methods (Singer *et al.*, 2002), even though hydropathy alone may account for a large portion of the information used (Cline *et al.*, 2002). Simply examining the conservation of two positions is surprisingly effective at contact prediction (Fodor and Aldrich, 2004a). The best performing machine learning contact prediction methods use multiple features to decide if two residues are in contact (Izarzugaza *et al.*, 2007).

The aim of this study is to examine to what extent inferred residue contacts can aid in the evaluation of predicted 3D structural models (Fig. 1). To overcome bias from small sample sizes, we use data from a recent metagenomic sequencing project (Yooseph *et al.*, 2007) to greatly expand our database of sequence homologs. To integrate multiple lines of evidence for each potential residue pair in contact, we use a Bayesian inference procedure cross-validated on a large test set of known structures. The evidence input into the Bayesian inference procedure comes from multiple sequence alignments, and includes a novel correlated mutations method. We show that combining multiple lines of evidence with the Bayesian inference procedure boosts accuracy of contact prediction, and that the resulting predicted contacts contain information about the quality of 3D structural models.

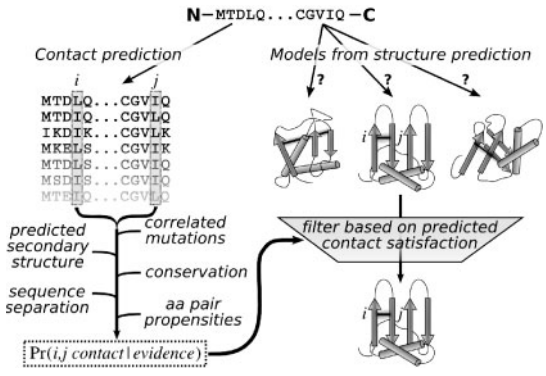

Fig. 1. Overview of strategy used to identify correct structure prediction models. Homologs for a sequence of interest are identified by searching an expanded sequence database that includes metagenomics data. Evidence gathered from the resulting multiple sequence alignments is combined via a Bayesian inference procedure to produce posterior probabilities of contact for all residue pairs (left). Models generated from structure prediction algorithms are filtered to select models that satisfy the most probable predicted residue contacts (right).

2 METHODS

2.1 Test set of known 3D structures

To evaluate the performance of contact prediction methods, we created a test set of 1656 representative known 3D protein structures. Details of test set construction are provided in Methods S1.1 in Supplementary Material.

2.1.1 Construction of multiple sequence alignments For each protein in the test set, we gathered homologous sequences from the non-redundant nr database provided by the National Center for Biotechnology Information, as well as from a database of non-redundant proteins identified in the metagenomic global ocean sampling (GOS) expedition dataset (Yooseph *et al.*, 2007). For each structure, homologs were filtered to remove sequences with >80% identity using CD-HIT (Li and Godzik, 2006). For those structures with at least 100 remaining homologs, alignments were constructed with either Kalign (Lassmann and Sonnhammer, 2005) or MUSCLE (Edgar, 2004). For comparison, the same procedure was applied to the same test set proteins without the inclusion of the metagenomics data.

2.1.2 Evaluation of contact predictions We define two residues in contact if their Cβ atoms are $\leq$8 Å apart (Cα for Glycine). We choose this definition only because it is used by much of the contact prediction literature, and we wish to facilitate comparison with other experiments. Two measures are used to evaluate predicted contacts. Accuracy is the number of residue pairs in contact divided by the total number of predictions considered. Fold improvement over random is defined as the accuracy divided by the expected accuracy if residue pairs are picked at random in the test structure of interest. For all evaluations, residue pairs were separated by at least six residues in primary sequence. For convenience, coverage (the fraction of true contacts predicted) and evaluations done with minimum residue separation of 12 residues (where random accuracies are lesser) are shown in Supplementary Table 1, though those results do not change the conclusions presented here.

2.2 Evidence used in Bayesian inference of contacts

Several lines of evidence are used to infer residue pairs in contact (Fig. 1, left). The measurement of individual lines of evidence is described here. The method used for integrating these lines of evidence is described in Section 2.3.

2.2.1 Evidence based on correlated mutations Two methods were used to detect correlated mutations in multiple sequence alignments. Both methods require special consideration for gaps in the input alignment. We chose not to make predictions on residue pairs in which one or more columns contained more than 30% gaps in the alignment. In calculating residue frequencies involving the remaining columns, if either character of a residue pair was a gap in the alignment, both characters were ignored.

The first correlated mutations algorithm is an in-house implementation of the method of Martin *et al.* (2005), which normalizes the mutual information (MI) between two columns in an alignment by their joint entropy. This normalization addresses the observation that methods such as MI are not independent of individual column conservation (Fodor and Aldrich, 2004a). We did not require the minimum entropy threshold imposed by Martin *et al.*

We call our second method, used to detect correlated mutations, the MI vector similarity method. It is ultimately also based on MI between two

columns in a multiple sequence alignment. The method rewards residue pairs which have similar patterns of covariation with all other residues in the protein. For this method, we first recode the amino acids into seven broader groups empirically derived from 'trusted' alignments (Wrabl and Grishin, 2005). For each column i in a multiple sequence alignment, we use the symbol defined by the unordered triplet set of amino acid groups at position $i-1$, i and $i+1$ to compute the MI with an unordered triplet set of amino acid groups centered at all other positions. The final score between columns i and j is the Euclidian distance between the vector of MI between the triplet at i to all positions and the vector of MI between the triplet at j to all positions. We ignore those triplet symbols having any gap when we compute frequencies necessary for MI.

2.2.2 Other forms of evidence Other lines of evidence collected for each residue pair include conservation of the pair, predicted secondary structure of the pair, residue separation in primary sequence and amino acid composition. Details are provided in Methods S1.2 in Supplementary Material.

2.3 Contact prediction with Bayesian inference

At the heart of the contact prediction is a simple Bayesian inference procedure used to integrate the various lines of evidence collected about two residues in a multiple sequence alignment. This inference attempts to compute the probability that two residues are close, given the evidence ev collected about them:

$$\Pr(close|ev) = \frac{\Pr(ev|close)\Pr(close)}{\Pr(ev|close)\Pr(close) + \Pr(ev|far)\Pr(far)} \quad (1)$$

where $\Pr(close)+\Pr(far)=1$. To compute the posterior probability that two residues are close, we need an estimate of the likelihood function $\Pr(ev|close)$ (as well as $\Pr(ev|far)$) and an estimate of the prior probability $\Pr(close)$.

2.3.1 Cross-fold validation Training the Bayesian inference procedure requires estimating the prior probability that two residues are close as well as estimating the likelihood functions for close and far residue pairs. The prior probability estimate (see Section 2.3.2) is not heavily dependent on minor variations in the composition of the training/test set, and thus we use the entire test set to estimate $\Pr(close)$. To train the likelihood function $\Pr(ev|close)$ (and $\Pr(ev|far)$), however, we used full cross-validation on our test set. For each structure in the test set, we first remove all homologous sequences in the rest of the test set by filtering out all other structures whose sequences have a BLAST e-value ≤ 0.1. We then estimate the likelihood function for each test case using only the remaining, non-homologous structures. Usually, this removed only a small fraction of structures from the test set.

2.3.2 The prior probability To establish the prior probability of two residues being close in the 3D structure, we ask how many residues are in the protein chain for which contacts are being predicted. We expect two residues picked at random from shorter chains to have a higher chance of being close than two residues picked from a longer chain. To model this explicitly, we used all observed pairs with residue separation ≥ 6 from the structures in the test set to fit two parameters in the following power function:

$$\Pr(close) = \Pr(close|L) = aL^b \quad (2)$$

where L is the chain length of the protein and least squares optimization resulted in the parameters $a=1.31$ and $b=-0.77$ (Supplementary Fig. 1). All residue pairs in a protein have the same prior probability of contact.

2.3.3 The likelihood function We used full N-fold cross-validation with our test set (see Section 2.3.1) to estimate a likelihood function for each test case. For each structure and corresponding alignment not related in sequence to the structure being tested, we collect evidence for all possible residue pairs. Because the continuous evidence was not readily modeled by standard distributions, we discretize some lines of evidence into bins (see Supplementary Methods S1.3). Thus for each residue pair in the training set, we note first whether the residue pair is close or far, and then add an observed count in the appropriate bin describing the pair for each line of evidence. Ideally, the likelihood would be a full 6D joint distribution. However, the number of bins for which we must estimate probabilities explodes quickly in six dimensions when compared to the number of observations (residue pairs) in the training set. Thus we choose to treat some lines of evidence as independent, modeling the joint distribution as three independent distributions:

$$\begin{aligned}\Pr(ev|close) = &\Pr(cm1, cm2, cons|close) \\ &*\Pr(ss, aaPair|close) \\ &*\Pr(rs|close)\end{aligned} \quad (3)$$

where $cm1$ and $cm2$ are the percentile ranks of the two correlated mutations methods for a given structure, $cons$ is the conservation in the alignment of the pair, ss is the secondary structure of the pair, $aaPair$ are the residue identities of the pair and rs is the residue separation in primary sequence. The estimated probabilities are simple maximum likelihood estimators: the observed counts for each set of evidence are divided by the total counts across all close pairs. An analogous procedure is used to estimate $\Pr(ev|far)$.

2.4 Satisfaction of predicted contacts by predicted 3D structural models

We chose to evaluate contact satisfaction by models (3D predictions) submitted to the most recent CASP experiment (Moult *et al.*, 2007). For each CASP target, we searched the filtered nr and GOS databases for homologs with BLAST as in Section 2.1 (E-value $\leq$ 1 e–10). We filtered out homologs with $\geq$80% identity with CD-HIT and built multiple sequence alignments with Kalign as in Section 2.2. This procedure resulted in 32 CASP targets with alignments of at least 100 sequences, to which we applied our Bayesian inference procedure to infer residue contacts. We built a likelihood function for each of the 32 targets by first removing all homologs from the training set as in Section 2.3.3. To evaluate the similarity of models with the true 3D structures, we downloaded GDT_TS scores (Zemla, 2003) from the CASP website (http://predictioncenter.org/casp7).

To evaluate predicted structural models for their satisfaction of predicted residue contacts, we devised a simple scoring scheme that rewards 3D models that place highly probable contacts close together. As elsewhere, we define a contact threshold of $\leq$8 Å, and compute the contact satisfaction score as

$$satisfaction = \frac{1}{2L}\left(\sum_{i=1}^{2L}\Pr(contact_i)(8 - distance_i)\right) \quad (4)$$

where $\Pr(contact_i)$is the posterior probability of predicted contact i, and $distance_i$ of contact i is in Angstroms in the predicted 3D model. The top $2L$ predicted contacts are used in the summation, where L is the chain length of the protein. This scheme, while simple, rewards models which have multiple predicted contacts close together in a way weighted both by the posterior probability of the contact and by the actual distance in the 3D model. All models with missing residues in more than three predicted contacts were removed from the evaluation, as were models without Cβ atoms. When normalized, the lowest scoring model for each target was set to a normalized score of 0 and the highest scoring model was set to 1.

3 RESULTS

3.1 Prediction of residue contacts

3.1.1 Contact prediction based on correlated mutations Residue interactions in a folded protein are not exclusively binary: any given residue may interact in complex ways with multiple other residues.

Spatially proximal residues may be dependent on this network of constraints in similar ways. This relationship can be described for each residue by a vector of correlated mutation scores with all other residues. We designed a MI vector similarity method to detect similarities between the correlated mutation vectors of two residues (Section 2.2.1).

We used this MI vector similarity method to predict residue contacts in a large test set of known structures, and found that it performs favorably when compared to other methods we tried. Consistent with previous studies (Grana *et al*., 2005), we chose to evaluate the accuracy and fold improvement over random (IOR) predictions of the top $2L$ and $L/10$ predictions for each test structure of length L (Fig. 2 and Supplementary Table 1). For the top $L/10$ predictions, we find the method of (Martin *et al*., 2005), which normalizes MI by the joint entropy between two columns in a sequence alignment, performs somewhat better than both the MI vector similarity method and a method which examines pairwise column conservation (Fodor and Aldrich, 2004a). However, this advantage disappears when examining $2L$ predictions, and all three methods identify correct contacts with comparable levels of accuracy.

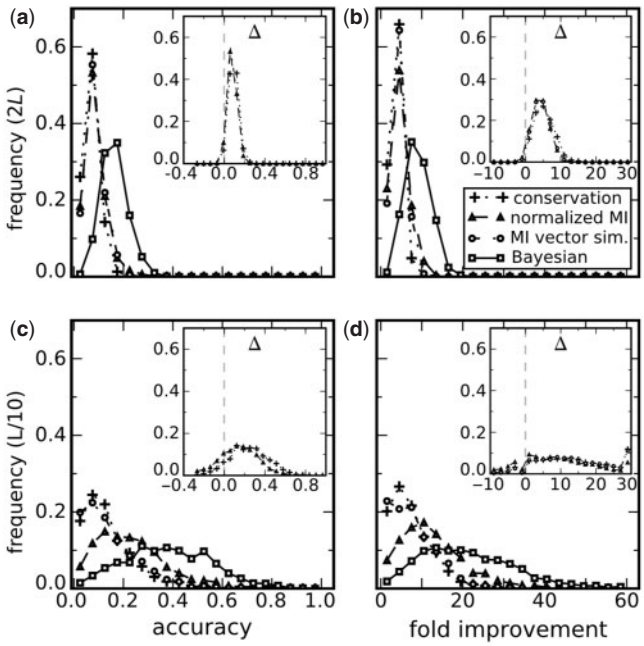

Fig. 2. Performance of various contact prediction methods. Accuracy (**a**, **c**) and fold-IOR (**b**, **d**) were computed for each prediction method for each known protein structure in the test set. The resulting distributions calculated from the entire test set are shown. The number of predicted contacts evaluated is proportional to the length L of each protein, with either $2L$ (a, b) or $L/10$ (c, d) predicted contacts examined. In all cases, methods based on individual forms of evidence perform poorer than the Bayesian inference method (open squares), which integrates multiple forms of evidence. Insets show distributions of the difference in prediction accuracy (a, c) and fold IOR (b, d), computed by comparing the performance of the Bayesian method with methods relying on a single form of evidence for individual proteins in the test set. For individual test cases, the Bayesian inference method usually performs better than any of the three methods based on only one form of evidence.

3.1.2 Improved contact prediction based on Bayesian inference It is of importance that, although they perform similarly on average, each of the three methods in Section 3.1.1 predicts largely distinct residue contacts (Supplementary Table 2). For any given structure in the test set, on average <1% of the top $2L$ contacts predicted by each method are common to all three methods. This led us to ask whether a method that combined multiple lines of evidence could more accurately predict residue contacts.

We chose to integrate multiple lines of evidence collected about each potentially contacting residue pair with a Bayesian inference procedure (see Section 2.3). The end result of this procedure is a posterior probability of contact for each residue pair. When compared to the correlated mutations-based methods for contact prediction or conservation, the Bayesian inference procedure is more accurate and has a higher fold IOR (Fig. 2 and Supplementary Table 1). Examination of the likelihood function provides a detailed description of the interplay between individual lines of evidence, and shows how these lines of evidence complement each other to produce more accurate contact predictions (Supplementary Fig. 2).

The improved performance of the Bayesian inference method is due to a systematic improvement across almost all individual proteins in the test set. For each test case, we computed the improvement (or decline) in accuracy and fold IOR observed when using the Bayesian inference procedure versus any of the other methods. The distribution of these differences shows that, for individual test cases, the Bayesian inference procedure almost always performs better than any of the other methods for predicting contacts (Fig. 2, insets). For example, while making $2L$ predictions, the Bayesian inference procedure makes more accurate predictions on 98% of test cases when compared to the next best method (Fig. 2a inset, density to the right of the dashed line).

3.1.3 Alignments augmented by metagenomics data produce slightly more accurate predictions Other investigators have reported an improved ability to predict residue contacts with increased alignment size (e.g. Martin *et al*., 2005; Shackelford and Karplus, 2007; Tillier and Lui, 2003). Recent studies have exploited metagenomic shotgun sequencing to dramatically expand the size of protein sequence space (Tringe and Rubin, 2005). We found that incorporating homologs from a non-redundant database of ~3.2 million predicted protein sequences from the GOS expedition (Yooseph *et al*., 2007), substantially increased the size of our multiple sequence alignments. The median percent GOS sequences of a test set alignment was 41%. An increase in alignment size correlates with an increase in fold-IOR of contact predictions made via Bayesian inference (Spearman correlation $r_s = 0.38$; Supplementary Fig. 3).

When alignments were built without the added GOS sequences, the loss in prediction performance was small but statistically significant (Supplementary Table 3). For the two MI-based methods, the added benefit of the GOS sequences was highly statistically significant ($P \leq 1e{-}34$ for 2L predictions, one-sided paired t-test). The Bayesian method does not benefit as clearly from the addition of these sequences (Supplementary Table 3). Interestingly, smaller alignments benefit the most significantly from additional GOS sequences, while already large alignments tend not to show performance gains (Supplementary Tables 4–6). We speculate that in these cases, additional sequences may make a correct alignment more difficult.

3.2 Using inferred residue contacts to evaluate predicted 3D structural models

Many computational methods predict 3D protein structure. We asked whether residue contacts inferred by the Bayesian inference procedure could be used to select from among many 3D structural models those most similar to the true structure. The most recent CASP experiment (Moult *et al.*, 2007) provides a rich source of computational models predicted for proteins with experimentally derived 3D structures. For each CASP target protein, we predicted residue contacts using a protocol identical to that applied to our cross-validation test set. For those 32 CASP targets with at least 100 diverse homologs from the nr and GOS databases, we applied the Bayesian inference procedure to produce posterior probabilities of contact for every residue–residue pair.

3.2.1 Predicted contact satisfaction correlates with 3D model correctness For each 3D model, we measured how well the model satisfied our top 2*L* predicted residue contacts. We devised a scoring function that rewards models that position predicted residue contacts close together (Section 2.4). If a predicted contact is present in the model, the score increases proportionally to the posterior probability of the contact. We also measured how correct each 3D model is by its GDT_TS score, which roughly measures the percentage of the model that does not deviate from the true structure (Zemla, 2003). For almost all targets, there is a striking correlation between predicted contact satisfaction and model correctness (mean Spearman rank correlation $r_s = 0.50$; Fig. 3a, Supplementary Fig. 4). If the predicted contact satisfaction scores are shuffled with respect to the GDT_TS scores for each model, the distribution of correlations centers sharply around 0, showing that the observed correlations cannot be explained simply by the presence of many similar models with similar satisfaction and GDT_TS scores (Fig. 1a, white bars). We found the correlation increased when greater numbers of predicted contacts were used (data not shown), indicating that there is information in the posterior probabilities we assign to predicted contacts. Using more predicted contacts, even when including contacts with lower posterior probabilities, appears beneficial when evaluating model correctness. This is expected with accurate posterior probabilities (Supplementary Fig. 5), where lower confidence contacts will still contribute information to the satisfaction score in a manner correctly weighted by their assigned probability.

For most models, the predicted contact satisfaction score correctly identifies the best models with high sensitivity and specificity. We defined a model as 'good' if it had a GDT_TS score ≥90% of the maximum GDT_TS score, and used receiver operating characteristic (ROC) curves to examine how well the predicted contact satisfaction score correctly classified 'good' models (Supplementary Fig. 6). Most targets showed high sensitivity with useful levels of specificity, as summarized by the distribution of the area under the curve (AUC) for all targets (Fig. 3b).

The ability of predicted contact satisfaction to correctly rank models by their closeness to the true structure is not merely due to more compact models having higher predicted contact satisfaction. To rule out this possibility, we evaluated the ability of predicted contact satisfaction to rank only the most compact models for each target, where compactness was measured by the radius of gyration (Supplementary Methods S1.1). Even when considering only the top 10% most compact models for each target, predicted contact satisfaction correlates equally well with GDT_TS (mean $r_s = 0.51$, Supplementary Fig. 7). Thus the success of our method cannot be explained only by the increased compactness of good models.

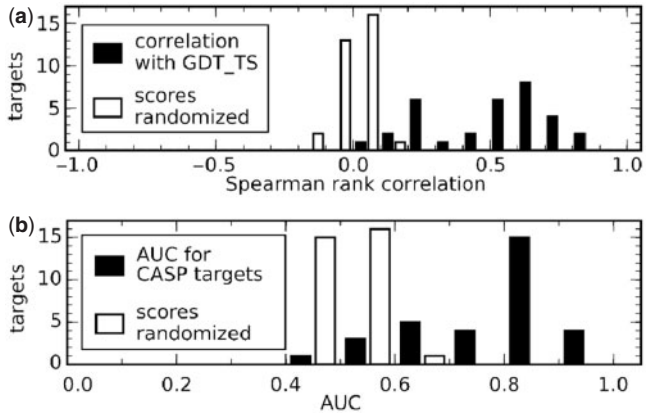

Fig. 3. Satisfaction of predicted contacts by 3D structural models correlates with the similarity of the models to the true structures. For each CASP target with contact predictions, submitted 3D models were evaluated with a score that probabilistically rewards the presence of predicted contacts in the model. (**a**) Histogram of Spearman rank correlations between the contact satisfaction score and GDT_TS, a measure of 3D model similarity to the true structure. (**b**) Histogram of AUC on ROC plots testing the ability of the predicted contact satisfaction score to identify the most correct models (those with GDT_TS ≥90% of the maximum GDT_TS). In each panel, the histogram in white has had the contact satisfaction scores shuffled with respect to the GDT_TS values. For most targets, there is strong correlation between predicted contact satisfaction and the correctness of a model as measured by GDT_TS, and a strong ability to classify good models with high sensitivity and specificity, as characterized by high AUC values.

3.2.2 Example target T0308 We examined some targets more closely to better understand why the predicted contact satisfaction score is useful in choosing more correct 3D models. One informative target was T0308 (PDB id 2H57), the crystal structure of ADP-ribosylation factor-like 6 (Wang *et al.*, 2006). The 3D structure of this 190 residue protein was generally predicted quite well, with one well populated set of models clustering between 75–80% GDT_TS and a second less populous set clustering with >85% GDT_TS (Fig. 4a). These two sets are also differentiated by their predicted contact satisfaction scores, with the more correct models generally having higher satisfaction scores (Spearman rank correlation $r_s = 0.61$). A representative model with a very high GDT_TS score is shown in Figure 4c. When the top 10 correct predicted contacts are displayed on this model, all of these contacts clearly have their Cβ atoms positioned close together. In a representative model with a lower GDT_TS score, two crucial loops and a beta strand are positioned differently, which greatly lengthens the Cβ distances between a handful of correct predicted contacts in the model (Fig. 4d), resulting in a lower satisfaction score. These two incorrectly repositioned loops help bind a guanosine triphosphate substrate analog in the 2.0 Å crystal structure (Wang *et al.*, 2006). In this case the predicted contact satisfaction scores help highlight a crucial structural difference between the two models. Though more groups submit models with incorrectly positioned loops, the predicted contact satisfaction score selects the more correct models.

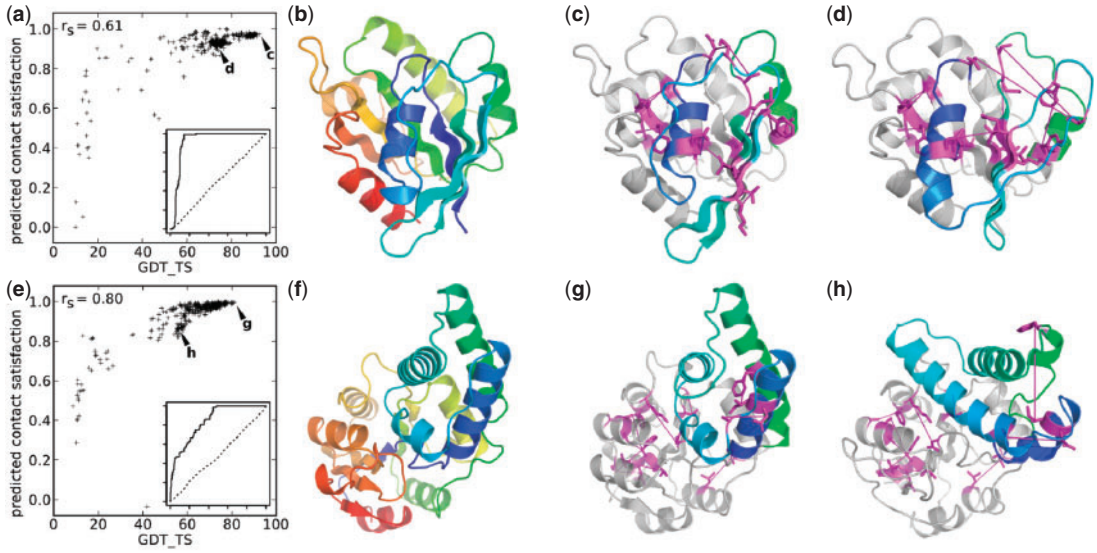

Fig. 4. Predicted contact satisfaction for 3D models of CASP targets. For each model submitted for CASP target T0308 (**a**) or T0303 (**e**), the GDT_TS score is plotted against a normalized predicted contact satisfaction score. The Spearman rank correlation (r_S) is shown. Insets are ROC curves which show that the predicted contact satisfaction score is able to discriminate good models (those with GDT_TS scores ≥90% of the maximum GDT_TS score) from the rest of the models (*y*-axis: true positive rate, *x*-axis: false positive rate, dotted line: average ROC curve for 10 sets of shuffled scores). (**b**) The crystal structure of T0308. Two submitted models are shown in (**c**) (Group 416, Model 3) and (**d**) (Group 18, Model 4), with the top 10 correct predicted contacts shown as magenta lines between Cβ atoms. The model in (c) has very good agreement with the true structure and has a high contact prediction score. The model in (d) is mostly correct, but has incorrect placement of a loop important for substrate binding (shown in color), which makes predicted contacts have long inter-molecular distances in the model and lowers the predicted contact satisfaction score. (**f**) The true crystal structure of T0303. Two submitted models are shown in (**g**) (Group 556, Model 3) and (**h**) (Group 205, Model 2), with the top 10 correct predicted contacts again shown. Both models display mostly correct packing of the distal domain in the figure, but the model in (h) packs the alpha helices incorrectly in the proximal domain (shown in color). This incorrect packing results in longer inter-atomic Cβ distances for predicted contacts and a lower contact prediction satisfaction score.

3.2.3 Example target T0303 CASP target T0303 (PDB id 2HSZ) is a predicted phosphatase with a 1.9 Å crystal structure solved by the Joint Center for Structural Genomics, (2006). Submitted 3D models for this target were moderately successful, with almost all GDT_TS values ranging from 40% to 80%. These values correlate strongly with how well the models satisfy the top 2*L* predicted contacts produced by the Bayesian inference procedure (Spearman rank correlation $r_s = 0.80$; Fig. 4e). When the top 10 correct predicted contacts are displayed on the most accurate model submitted, the dispersed contacts have short Cβ distances (Fig. 4g). Several of the models appear to have correctly positioned only slightly more than half of the protein. For example, the model shown in Figure 4h has correctly folded only the distal domain in the figure, as highlighted by the close Cβ distances of correct predicted contacts in that domain. However, the alpha helices are incorrectly packed in the proximal domain of this model, and the resulting increased length of predicted contact Cβ distances results in a lower total predicted contact satisfaction score. In this case, many models with similar incorrect packing of these helices would be filtered out by their lower predicted contact satisfaction scores (Fig. 4e).

4 DISCUSSION

The use of a Bayesian inference procedure for contact prediction has several advantages. First, multiple lines of evidence can be integrated into a single posterior probability of contact, boosting prediction accuracy (Fig. 2, Supplementary Table 1). Others have taken machine learning approaches to integrate multiple data sources for contact prediction (Cheng and Baldi, 2007; Fariselli *et al*., 2001; Hamilton *et al*., 2004; Pollastri and Baldi, 2002; Punta and Rost, 2005; Shackelford and Karplus, 2007; Vullo *et al*., 2006), and though meaningful comparisons are made difficult by varying definitions of the problem, Bayesian inference predicts contacts competitively (Supplementary Table 7). The advantage of the Bayesian inference approach presented here is the production of accurate posterior probabilities (Supplementary Fig. 5), indicating in part that the use of our prior distribution based on protein length is meaningful. Accurate posterior probabilities allow for useful incorporation of a wide range of contacts into the predicted contact satisfaction score, modulating the importance of each predicted contact correctly. For example, even though the top 2*L* predicted contacts are less accurate than the higher probability top *L*/10 predicted contacts (Fig. 2), because the posterior probabilities are accurate we still gain useful information in an appropriately weighted way by including these lower probability contacts when evaluating models. In the Bayesian inference procedure, additional lines of evidence are included simply by adding a dimension to the joint likelihood distribution during training. Understanding the dependencies between lines of evidence is a matter of reducing the joint likelihood function to a marginal distribution that considers only the evidence of interest (Supplementary Fig. 2). Once a likelihood function and prior are trained, inference of residue contacts is very computationally efficient.

Most lines of evidence used here require some parameter estimation from the multiple sequence alignments that the algorithm takes as input. For example, correlated mutations methods based on MI must estimate the probabilities of each of the 20 amino acids in each column. These estimates are limited by sample size effects in small alignments (Martin *et al.*, 2005). Augmentation of multiple sequence alignments with homologs identified in metagenomic sequencing projects helps in a small but significant way to combat errors in estimation due to small sample size, resulting in better contact predictions (Supplementary Fig. 3 and Supplementary Tables 3–6). As new sequencing technologies accelerate the pace of such metagenomics data acquisition, it will be important to include such sequences carefully when defining protein families for any purpose. We note that structures with very large alignments (>10 000 homologs) have less accurate contact predictions than expected (Supplementary Fig. 3), and already large alignments do not benefit as clearly from the addition of metagenomics sequences (Supplementary Tables 4 and 5). We suspect difficulties in correctly aligning such large families or the presence of structurally variant subfamilies may be at fault. The method presented here, like others, assumes a correct alignment. In the future, it may be beneficial to explicitly model potential alignment error in contact prediction.

We have presented a Bayesian inference procedure for residue contact prediction, and have shown its utility in selecting 3D models which are most similar to the true protein structure. While the idea of using inferred contacts to rank predicted 3D structural models has been recently discussed (Grana *et al.*, 2005), we know of few studies which explicitly attempt to evaluate the potential utility of such an approach (Eyal *et al.*, 2007; Olmea *et al.*, 1999; Schueler-Furman and Baker, 2003). Our results confirm that there is a clear benefit in using contact prediction as one component of 3D structure prediction. In the future, it will be interesting to dissect why contact prediction works so well on certain structures and is less successful on others (Fig. 2 and Supplementary Figs 4 and 6).

The real merit of using contact predictions in 3D structure prediction may not yet be revealed, though. Our analysis uses predicted contacts to post-filter 3D models. However, it has been observed that when an accurate energy function is available, the limiting step in structure prediction is exploration of conformational space (Qian *et al.*, 2007). Thus it would be desirable to use the information provided by contact prediction to guide sampling of conformational space. Structure prediction algorithms such as the highly successful ROSETTA program can exploit experimental distance restraints gathered from nuclear magnetic resonance spectroscopy (Bowers *et al.*, 2000). Extending these programs to use accurate probabilistic distance restraints such as those provided by the Bayesian inference procedure described here may prove similarly fruitful in restricting the search of conformational space (Ortiz *et al.*, 1999). Structure prediction by Zhang *et al.* (2003) benefited from threading-inferred contacts even when accuracy of the predictions was as low as 20%. Contact predictions for most structures in our test set achieve this level of accuracy (Fig. 2c), and with accurate posterior probabilities it may be possible in the future to a priori estimate the accuracy of predictions for an unknown target. As protein structure prediction continues to improve (Schueler-Furman *et al.*, 2005), especially as pertaining to longer targets, probabilistically accurate contact prediction may provide valuable information.

ACKNOWLEDGEMENTS

We thank R. Llewellyn, M. Beeby, L. Goldschmidt, and R. Riley for discussions.

Funding: C.S.M. was supported by the Ruth L. Kirschstein National Research Service Award GM07185. This work was funded in part by DOE and the Howard Hughes Medical Institute.

Conflict of Interest: none declared.

REFERENCES

Altschuh,D. *et al.* (1987) Correlation of co-ordinated amino acid substitutions with function in viruses related to tobacco mosaic virus. *J. Mol. Biol.*, **193**, 693–707.

Anfinsen,C.B. (1973) Principles that govern the folding of protein chains. *Science*, **181**, 223–230.

Bowers,P.M. *et al.* (2000) De novo protein structure determination using sparse NMR data. *J. Biomol. NMR*, **18**, 311–318.

Cheng,J. and Baldi,P. (2007) Improved residue contact prediction using support vector machines and a large feature set. *BMC Bioinformatics*, **8**, 113.

Cline,M.S. *et al.* (2002) Information-theoretic dissection of pairwise contact potentials. *Proteins*, **49**, 7–14.

Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.

Eyal,E. *et al.* (2007) A pair-to-pair amino acids substitution matrix and its applications for protein structure prediction. *Proteins*, **67**, 142–153.

Fariselli,P. *et al.* (2001) Prediction of contact maps with neural networks and correlated mutations. *Protein Eng.*, **14**, 835–843.

Fodor,A.A. and Aldrich,R.W. (2004a) Influence of conservation on calculations of amino acid covariance in multiple sequence alignments. *Proteins*, **56**, 211–221.

Fodor,A.A. and Aldrich,R.W. (2004b) On evolutionary conservation of thermodynamic coupling in proteins. *J. Biol. Chem.*, **279**, 19046–19050.

Gobel,U. *et al.* (1994) Correlated mutations and residue contacts in proteins. *Proteins*, **18**, 309–317.

Grana,O. *et al.* (2005) CASP6 assessment of contact prediction. *Proteins*, **61** (Suppl. 7), 214–224.

Hamilton,N. *et al.* (2004) Protein contact prediction using patterns of correlation. *Proteins*, **56**, 679–684.

Izarzugaza,J.M. *et al.* (2007) Assessment of intramolecular contact predictions for CASP7. *Proteins*, **69** (Suppl. 8), 152–158.

Joint Center for Structural Genomics (2006) Crystal structure of novel predicted phosphatase from Haemophilus somnus 129PT at 1.90 A resolution (unpublished). Joint Center for Structural Genomics.

Korber,B.T. *et al.* (1993) Covariation of mutations in the V3 loop of human immunodeficiency virus type 1 envelope protein: an information theoretic analysis. *Proc. Natl Acad. Sci. USA*, **90**, 7176–7180.

Lassmann,T. and Sonnhammer,E.L. (2005) Kalign–an accurate and fast multiple sequence alignment algorithm. *BMC Bioinformatics*, **6**, 298.

Li,W. and Godzik,A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**, 1658–1659.

Lockless,S.W. and Ranganathan,R. (1999) Evolutionarily conserved pathways of energetic connectivity in protein families. *Science*, **286**, 295–299.

Martin,L.C. *et al.* (2005) Using information theory to search for co-evolving residues in proteins. *Bioinformatics*, **21**, 4116–4124.

Moult,J. *et al.* (2007) Critical assessment of methods of protein structure prediction-round VII. *Proteins*, **69** (Suppl. 8), 3–9.

Neher,E. (1994) How frequent are correlated changes in families of protein sequences? *Proc. Natl Acad. Sci. USA*, **91**, 98–102.

Noivirt,O. *et al.* (2005) Detection and reduction of evolutionary noise in correlated mutation analysis. *Protein Eng. Des. Sel.*, **18**, 247–253.

Olmea,O. *et al.* (1999) Effective use of sequence correlation and conservation in fold recognition. *J. Mol. Biol.*, **293**, 1221–1239.

Ortiz,A.R. *et al.* (1998) Nativelike topology assembly of small proteins using predicted restraints in Monte Carlo folding simulations. *Proc. Natl Acad. Sci. USA*, **95**, 1020–1025.

Ortiz,A.R. *et al.* (1999) Ab initio folding of proteins using restraints derived from evolutionary information. *Proteins*, **37**, 177–185.

Pollastri,G. and Baldi,P. (2002) Prediction of contact maps by GIOHMMs and recurrent neural networks using lateral propagation from all four cardinal corners. *Bioinformatics*, **18** (Suppl. 1), S62–S70.

Pollock,D.D. *et al*. (1999) Coevolving protein residues: maximum likelihood identification and relationship to structure. *J. Mol. Biol.*, **287**, 187–198.

Punta,M. and Rost,B. (2005) PROFcon: novel prediction of long-range contacts. *Bioinformatics*, **21**, 2960–2968.

Qian,B. *et al*. (2007) High-resolution structure prediction and the crystallographic phase problem. *Nature*, **450**, 259–264.

Schueler-Furman,O. and Baker,D. (2003) Conserved residue clustering and protein structure prediction. *Proteins*, **52**, 225–235.

Schueler-Furman,O. *et al*. (2005) Progress in modeling of protein structures and interactions. *Science*, **310**, 638–642.

Shackelford,G. and Karplus,K. (2007) Contact prediction using mutual information and neural nets. *Proteins*, **69** (Suppl. 8), 159–164.

Singer,M.S. *et al*. (2002) Prediction of protein residue contacts with a PDB-derived likelihood matrix. *Protein Eng.*, **15**, 721–725.

Suel,G.M. *et al*. (2003) Evolutionarily conserved networks of residues mediate allosteric communication in proteins. *Nat. Struct. Biol.*, **10**, 59–69.

Tillier,E.R. and Lui,T.W. (2003) Using multiple interdependency to separate functional from phylogenetic correlations in protein alignments. *Bioinformatics*, **19**, 750–755.

Tringe,S.G. and Rubin,E.M. (2005) Metagenomics: DNA sequencing of environmental samples. *Nat. Rev. Genet.*, **6**, 805–814.

Vicatos,S. *et al*. (2005) Prediction of distant residue contacts with the use of evolutionary information. *Proteins*, **58**, 935–949.

Vullo,A. *et al*. (2006) A two-stage approach for improved prediction of residue contact maps. *BMC Bioinformatics*, **7**, 180.

Wang,J. *et al*. (2006) Crystal structure of human ADP-ribosylation factor-like 6 (CASP Target) (unpublished). Structural Genomics Consortium, Toronto, Canada.

Wollenberg,K.R. and Atchley,W.R. (2000) Separation of phylogenetic and functional associations in biological sequences by using the parametric bootstrap. *Proc. Natl Acad. Sci. USA*, **97**, 3288–3291.

Wrabl,J.O. and Grishin,N.V. (2005) Grouping of amino acid types and extraction of amino acid properties from multiple sequence alignments using variance maximization. *Proteins*, **61**, 523–534.

Yeang,C.H. and Haussler,D. (2007) Detecting coevolution in and among protein domains. *PLoS Comput. Biol.*, **3**, e211.

Yooseph,S. *et al*. (2007) The Sorcerer II Global Ocean Sampling expedition: expanding the universe of protein families. *PLoS Biol.*, **5**, e16.

Zemla,A. (2003) LGA: a method for finding 3D similarities in protein structures. *Nucleic Acids Res.*, **31**, 3370–3374.

Zhang,Y. (2007) Template-based modeling and free modeling by I-TASSER in CASP7. *Proteins*, **69** (Suppl. 8), 108–117.

Zhang,Y. *et al*. (2003) TOUCHSTONE II: a new approach to ab initio protein structure prediction. *Biophys. J.*, **85**, 1145–1164.