PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Optimal data collection for correlated mutation analysis

Haim Ashkenazy,[1,2] Ron Unger,[2] and Yossef Kliger[1]*

[1] Compugen LTD, Tel Aviv 69512, Israel

[2] The Mina and Everard Goodman Faculty of Life Sciences, Bar Ilan University, Ramat-Gan 52900, Israel

ABSTRACT

The main objective of correlated mutation analysis (CMA) is to predict intra-protein residue–residue interactions from sequence alone. Despite considerable progress in algorithms and computer capabilities, the performance of CMA methods remains quite low. Here we examine whether, and to what extent, the quality of CMA methods depends on the sequences that are included in the multiple sequence alignment (MSA). The results revealed a strong correlation between the number of homologs in an MSA and CMA prediction strength. Furthermore, many of the current methods include only orthologs in the MSA, we found that it is beneficial to include both orthologs and paralogs in the MSA. Remarkably, even remote homologs contribute to the improved accuracy. Based on our findings we put forward an automated data collection procedure, with a minimal coverage of 50% between the query protein and its orthologs and paralogs. This procedure improves accuracy even in the absence of manual curation. In this era of massive sequencing and exploding sequence data, our results suggest that correlated mutation-based methods have not reached their inherent performance limitations and that the role of CMA in structural biology is far from being fulfilled.

Proteins 2009; 74:545–555.

Key words: *ab-initio* structure prediction; correlated mutations; protein structure prediction; residue covariation; contact prediction.

INTRODUCTION

Correlated mutation analysis (CMA) has been suggested as a tool for predicting physiologically relevant residue–residue interactions in proteins.[1] This analysis has become one of the most studied methods for sequence-based structure prediction. The main goal of correlated mutations analysis is to identify pairs of residues that coevolve more often than expected from a pair of noninteracting residues in the protein. Coevolving residues may have either functional or structural dependencies, which provide selective pressure in favor of a mutation in one residue to compensate for a mutation in the other residue. Indeed, the statistical analysis of Lahn and coworkers[2] revealed that interacting residues tend to coevolve.

Over nearly two decades, various methods were suggested for statistically characterizing pairs of residue that have undergone correlated mutations.[1,3–7] Some of the methods calculate the odds for correlated mutations based on the probabilities for these mutations as defined by substitution matrices.[1,3,7] Others give each amino acid replacement an equal importance.[6] The low accuracy of these basic methods led to some suggested improvements to achieve higher accuracy.[8–10] A major challenge in correlated mutations analysis is dealing the phylogenetic history that introduces a significant amount of false-positive predictions and several methods were suggested in an attempt to reduce the phylogenetic noise.[10–15] Despite the considerable increase in the amount of relevant data from genome sequencing projects, *ab initio* contact prediction still seems unreachable.

The main method of comparing different methods for sequence-based contact prediction is based on protein contact-maps, two-dimensional matrices in which each residue–residue interaction is marked and which can be viewed as a low-resolution reflection of the 3D structure of the protein. Comparisons between several methods on different data sets[3,9,16] revealed that the method developed by Valencia and coworkers[1,17] more than a decade ago remains one of the most accurate methods for contact map prediction. This method is based on defining an exchange (similarities) matrix for each position in the multiple sequence alignment (MSA) and calculating the Pearson correlation coefficient between the exchange matrices at any two positions. The exchange matrix dimensions are equal to the depth of the MSA and each matrix element contains the similarity value of the residues observed in their related sequences, according to the McLachlan matrix.[18] The correlation coefficient

Additional Supporting Information may be found in the online version of this article.

*Correspondence to: Yossef Kliger, PhD, Compugen Ltd. 72 Pinchas Rosen, Tel Aviv 69512, Israel.
E-mail: kliger@compugen.co.il or yossef.kliger@gmail.com

Received 17 December 2007; Revised 3 May 2008; Accepted 2 June 2008
Published online 24 July 2008 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.22168

between two positions is a measure for correlated mutation of these positions: the higher the correlation coefficient is, the higher the probability that the given pair has coevolved.

All correlated mutation prediction methods are based on homologous sequences aligned relative to each other. Correlated mutation information is then extracted from such MSA. Correlation between the performances of CMA for contact map prediction and the number of sequences comprised in the MSA (MSA depth) has been shown years ago[19]; however, the massive growth of the sequences public databases justifies a new systematic exploring of this important aspect. In addition, simulated analysis revealed that even uncorrelated sites exhibit background correlation, hence making the signal to noise ratio a major obstacle for the prediction of intramolecular contact-maps, particularly in alignments with less than 125 sequences.[6] In current databases only a relatively small number of proteins have more than 125 orthologous proteins. Thus, we explored the option of increasing the MSA depth by considering also sequences with varied level of relatedness (i.e., remote vs. close homologs, and orthologs vs. paralogs).

Different proteins differ in the amount of information present in their MSA. The amount of information relates to the number of homologous proteins considered and their level of relatedness. This raises the question of how to build an optimal MSA. To deal with this question, many studies in the filed use PSI-BLAST[20] to collect the homologous sequences and differ from each other in their filtration methods.[5,13,21,22] Other studies are using preprepared MSAs downloaded mainly from HSSP[7,23] or Pfam[3,9,16,24]. These data preparation methods are not perfect and therefore many researchers in the field perform manual curation of MSAs[4,13,14] to improve performance. However, manual curation of MSAs impedes high throughput analysis, which is critical for robust performance analysis. Here, we explored whether there is a preferable automated data preparation procedure for constructing the MSAs (without any manual curation). The constructed MSA will be used as input for correlated mutations-based predictions.

RESULTS AND DISCUSSION

Correlated mutation-based prediction methods are based on MSAs, in which the sequence of the protein of interest (reference sequence) and its homologs are aligned together. Here, we tested the relationship between the number of sequences comprising an MSA and the performance of the correlated mutations analysis for contact map prediction. Toward this goal, the algorithm of Valencia and coworkers[17] was chosen as a benchmark method and evaluated for all proteins in a culled PDB[25] set (682 proteins).

It is common practice to focus on small portion of the residue–residue interactions sharing high reliability and evaluate the performances of CMA prediction methods based on these interactions.[26,27] The main motivation is that these few highly reliable predicted contacts can be used as constraints for protein *ab-initio* structure prediction[28] as well as to filter potential models generated by, for example threading methods.[29,30] However, this approach for performance evaluation does not reflect the degree of coverage provided by the prediction method. We therefore evaluate the accuracy of the prediction methods in the current study using two approaches: (i) based on the precision of the top $L/10$ or the top $L/5$ predictions (where L is the protein length); and (ii) based on all predicted interactions. In the latter method, evaluation is based on both the precision—the fraction of correct predictions among all predictions—and the recall—the fraction of interactions detected among all interactions. In addition, it is common to assume that coevolvement is partially because of relative proximity between residues and not necessarily due to direct physical contact. Therefore, we evaluate the performances also by using continuous measure of proximity between the predicted residues as was suggested by Pazos *et al.*[23] (Xd measure) and also used by others.[26,27,31] Using this method, we have evaluated the top $L/2$ predictions.

MSA depth correlates with prediction accuracy

Generally speaking, one should expect a positive correlation between the MSA depth and the predictor accuracy. The MSA depth is limited by the common practice to use MSAs comprising only of orthologous sequences.[12,13,15] Here we construct such a list by first generating the list of homologous proteins based on a PSI-Blast[20] run of the protein of interest against the NCBI Genbank database.[32] We then use the simplest, yet far from perfect, approach for screening only orthologs, namely, by keeping only the first hit of each organism.[13] Finally, only orthologous sequences having an alignment length that covers at least 90% of the reference sequence length (0.9-orthologs) are maintained. In an attempt to avoid covariation due to a common phylogenetic origin of closely related sequences we have filtered out sequences having more than 90% identity with each other (i.e., there are no two sequences that share more than 90% pair-wise identity in the MSA) using NRDB90.[33] For this analysis, we used only proteins having more than 256 orthologous sequences (26 proteins from the Culled-PDB data set). MSAs with varying depths were built for each of these proteins by randomly selecting a fixed number of orthologous sequences out of the sequences pool. In this way, the algorithm was tested on MSAs ranging in depth from 8 to 256. This process was repeated 10 times for each of the 26 MSAs. The accura-

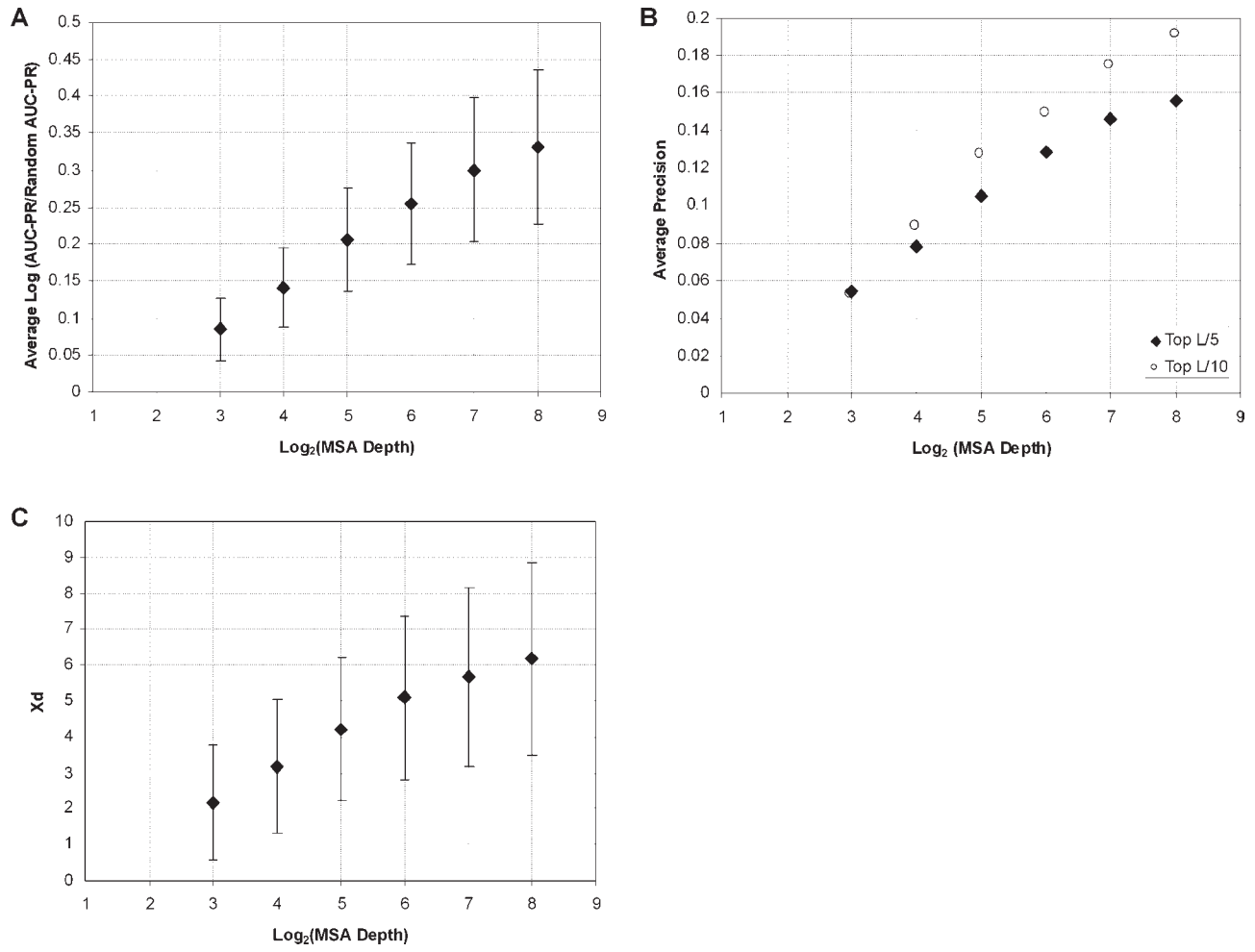

Figure 1

MSA depth correlates with predictor accuracy. Average performance calculated for 26 proteins, 10 randomizations each, based on their 0.9-orthologous sequences (using valencia and coworkers' algorithm), as evaluated using three different evaluation methods. **A**: Log of the ratio between the Area Under the Curve (AUC) of Precision-Recall graph (AUC-PR) for a CMA predictor, and the corresponding AUC-PR of a random predictor. **B**: Precision of top predictions (Top $L/5$ and $L/10$ where L = protein length). The standard deviations of the average precision is of the same order of magnitude as the standard deviation shown in (A) and (C), and are not displayed in the Figure to enhance clarity. **C**: Xd score when considering top $L/2$ predictions.

cies were evaluated by calculating the area under the Precision-Recall Curve (AUC-PR) which reflects the performances of the algorithm based on all predicted and existing interactions. The curve for a random predictor in Precision-Recall space is a horizontal line with a precision equal to the number of interactions observed in the 3D structure of the proteins divided by the number of all possible residue pairs. To eliminate the accuracy differences that are due to different protein lengths, the ratio between the AUC-PR of the Valencia and coworkers' algorithm and a random predictor was calculated. In addition, we evaluated the accuracy by calculating the precision achieved for the top $L/5$ or $L/10$ of the predicted interactions and by using continuous measure of proximity between the predictions and real contacts for top $L/2$ predictions (Xd measure) as commonly used at CASP evaluations.[26,27] The averaged performances achieved for these proteins using the Valencia and coworkers' algorithm and based on MSAs with varying depths are shown in Figure 1.

In random predictor, the Xd measure, and the log of the ratio between the AUC-PR of a CMA method and the random predictor AUC-PR, are both expected to be around 0. As expected, panels (A) and (C) of Figure 1 reveal that in most cases the performance of the predictor based on MSAs comprising a small number of sequences is close to that of a random predictor. On the other hand, all three panels of Figure 1 reveal that increasing the MSA depth by considering more orthologous sequences resulted in significantly higher accuracies.

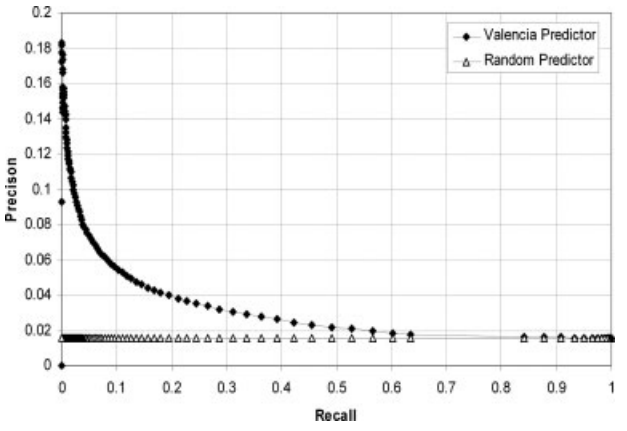

Figure 2

Performance evaluation of Valencia and coworkers' algorithm for correlated mutation based contact map prediction. Precision-Recall curve for concatenation of 97 proteins, each having MSAs comprising of more than 128 orthologous sequences (0.9-orthologs). Note that even when considering only proteins having deep MSAs the performances of correlated mutations contact map prediction is relatively low.

This observation is important since most of the proteins do not have that number of orthologous sequences in the current data base (Supplementary Fig. S1), and thus will benefit from future expansion of the databases. The fact that MSAs comprising of 256 sequences still show significant improvements as compared with MSAs comprising of 128 sequences (which are already considered to be deep) suggest that methods that are based on correlated mutations did not reach their inherent limit yet. Figure 1(B) also reveals that the accuracy for the top *L*/10 predictions is higher than accuracy for the top *L*/5 predictions. This further confirms that the correlated mutation signal truly exists.

The high standard deviation shown in Figure 1(A,C) emphasizes the large variations in accuracy achieved for different proteins as well as the differences between the information content of the MSAs. Standard deviations of the same order of magnitude were also observed in the precision of top *L*/5 and *L*/10 predictions (not shown). Thus, in addition to the dependency between MSA depth and performance, it is clear that the accuracy is highly dependent on the specific protein and the MSA.

Although it is clear from our results that the depth of the MSA correlates positively with the predictor accuracy, it is important to note that in absolute values the accuracies remain relatively low. Figure 2 shows the overall precision recall curve for all 97 proteins that have relatively deep 0.9-orthologs MSAs (of more than 128 sequences). The figure reveals a low precision of 5% can be achieved only at a recall as low as 10%. The precision of the top *L*/5 predictions reaches an average precision of 15% (with a standard deviation around 9%). The strong positive correlation between the accuracy and the MSA depth led us to further explore ways to increase the MSA depth and achieve higher performances.

MSA comprising remote orthologues

It is common to build MSA comprising sequences of similar lengths.[5,13] The reason is to enable more accurate alignment by MSA programs since MSA programs tend to open more, and perhaps unnecessary, gaps, and perform less accurately when given proteins of different lengths. However, many proteins do not have a significant number of homologs with similar length. Thus, we decided to check whether the loss of information due to the exclusion of these remote homologs is justified considering also the improvements in MSA tools in recent years.[34–37]

To test the influence of considering orthologs with varying overlap lengths, we constructed different sets of orthologous sequences that differ in the length of the alignment between the reference sequence and the PSI-Blast hit. The ratio between the alignment length and the reference sequence length can serve as a measure of the relatedness between the chosen sequence and the reference sequence. In total, three different sets were used: 0.9-ortholog (closest orthologs), 0.7-ortholog, and 0.5-ortholog (remote orthologs). These sets were filtered for highly similar sequences sharing more than 90% identity with each other. To explore the effect of considering orthologs with varying overlap length while keeping the number of sequences in the MSA fixed we choose 13 proteins that have many close orthologous sequences and many remote orthologs (orthologous with varied length). We divided the sequences of these proteins into three different unique groups according to the length of the alignment between the set sequence and reference sequence [i.e., orthologs sequences with overlap of 90% or above (0.9-orthologs set), orthologs with overlap of 70–90% (0.7 to 0.9-orthologs set), and orthologs with 50–70% overlap (0.5 to 0.7-orthologs set)]. For these proteins we built three sets of MSAs comprising of 100 sequences that were randomly chosen from the unique groups (i.e., 0.9-orthologs set, 0.7 to 0.9-orthologs set, 0.5 to 0.7-orthologs set), testing different ratios of remote orthologs: (1) 100% 0.9-orthologs, (2) 70% 0.9-orthologs + 30% 0.7 to 0.9-orthologs, and (3) 70% 0.9-orthologs + 15% 0.7 to 09-orthologs + 15% 0.5 to 0.7-orthologs. This procedure was repeated 10 times. The performances achieved based on the MSAs from these sets were calculated as specified in the previous section and are shown in Table I.

The results in Table I reveal that the performances achieved based on MSAs comprising of sequences with shorter overlap lengths are not significantly different from the performance of MSA consisting of sequences of similar lengths. These results suggest that it is beneficial

Table I
Performances Achieved Based of MSAs Comprising of Sequences with Varying Overlap Length

MSA set[a]	Average log (AUC-PR/random AUC PR)[b]	SD log (AUC-PR/random AUC PR)	Average top L/5 precision[c]	SD top L/5 precision
1. 100% 0.9-Orthologs sequences[d]	0.314	0.082	0.169	0.084
2. 70% 0.9-orthologs sequences + 30% 0.7 to 0.9-orthologs sequences[e]	0.314	0.089	0.173	0.088
3. 70% 0.9-orthologs sequences + 15% 0.7 to 0.9-orthologs sequences + 15% 0.5 to 0.7-orthologs sequences[f]	0.308	0.089	0.169	0.091

[a]MSAs of 13 proteins comprising of 100 orthologoues sequences. The sequences were randomly chosen from the orthologous pool. The process was repeated 10 times.
[b]Average log of the ratios between AUC-PR of CMA and random AUC-PR.
[c]Precision of top $L/5$ prediction (L = protein length).
[d]MSAs where all the sequences in the MSA are aligned to the reference sequence in at least 90% of their lengths.
[e]MSAs where 70% of the sequences in the MSA are aligned to the reference sequence in at least 90% of their lengths, and 30% of the sequences in the MSA are aligned to the reference sequence in at least 70%, but less than 90%, of their lengths.
[f]MSAs where 70% of the sequences in the MSA are aligned to the reference sequence in at least 90% of their lengths, 15% of the sequences in the MSA are aligned to the reference sequence in at least 70%, but less than 90%, of their lengths, and 15% of the sequences in the MSA are aligned to the reference sequence in at least 50%, but less than 70%, of their lengths.

to include orthologous sequences with shorter overlap lengths in the MSA, including sequences with alignment lengths as low as 50% of the reference sequence, since in this way the MSA depth can be increased. These findings also suggest that the performances that are being achieved today by focusing on a relatively small number of close sequences can be significantly improved by using more of the available data cloaked in remote orthologs.

MSA comprising both orthologs and paralogs

The basis of correlated mutations analysis is the assumption that each of the proteins in the MSA has a similar structure. Thus, it is common to use MSA comprising sequences of orthologous proteins.[12,13,15] The main reason for not considering paralogous proteins is the common understanding that orthologs have similar functions (and therefore similar structures), whereas paralogous proteins may have different functions and thus a larger structural diversity. Although this is probably true, it is not clear to what extent the gain in the signal to noise ratio due to the increase in MSA depth (see Fig. 1) would be offset by the introduction of additional noise introduced by the different selection forces that act on the structures of the paralogous proteins.

Here, we tested whether the inclusion of information cloaked in paralogous sequences is beneficial for contact map prediction. Thus, we prepared three more sets: 0.9-orthologs + paralogs, 0.7-orthologs + paralogs, and 0.5-orthologs + paralogs, simply by collecting all PSI-Blast hits, rather than only the first sequence from each organism.

To explore the influence of considering paralogous sequences in the MSA on the performances, we prepared MSAs comprising of 128 sequences. Different orthologous to paralogous ratios were used. The examination was made for all proteins that had more than 128-orthologous sequences and more than 128-paralogous sequences (21 proteins from the culled PDB set). The sequences for the MSAs were randomly chosen out of the 0.9-orthologs and 0.9-paralogs sequences pool. The procedure was repeated 10 times. The performances were calculated and are shown in panels (A) and (B) of Figure 3. In addition, we have filtered from the sequences pools the sequences sharing identity of more than 90% with each other. As expected this step reduced the number of sequences for each protein. After this step, we were left with only five proteins that have more than 100 orthologs as well as more than 100 paralogs. We repeated our analysis for these proteins (while fixing the depth of each MSA to 100 sequences) and the results are shown in panels (C) and (D) of Figure 3.

All four panels of Figure 3 reveal that the performances achieved based on MSAs that includes paralogous sequences is not significantly different from the performances achieved based on orthologs only. Furthermore, even when the MSA was comprised of only paralogous sequences, the performances are not reduced dramatically. However, it is worth to mention that in three (PDB-IDs: 1gwe, 2cxn, 2c78) out of the 21 proteins (which their corresponding MSAs were not filtered to remove redundant sequences), replacing orthologous with paralogous sequences consistently reduced the performances. Although this is not surprising, because there may be cases in which the added paralogs are significantly different in their structure/function from the reference sequence. In addition, the large variations in performance achieved for different proteins, mainly because of the differences between the information contained in the different MSAs also holds here. Nevertheless, our results suggest that for most proteins it is beneficial to consider also paralogous sequences in the MSA, as including also paralogous sequences deepen the MSA.

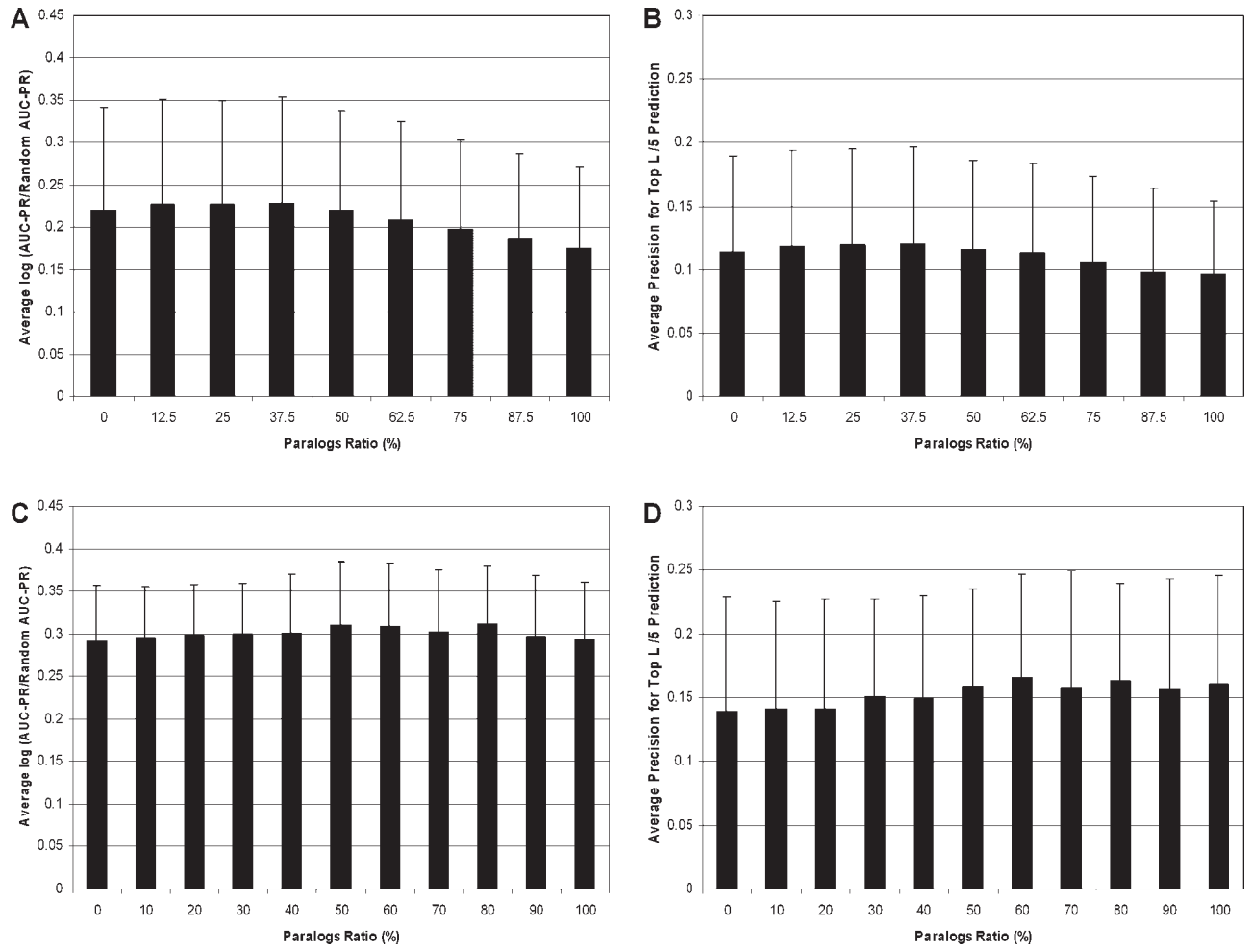

Figure 3

The influence of including paralogous sequences in MSA. Replacing orthologous with paralogous sequences does not reduce CMA performance. Proteins sequences were randomly selected for MSAs out of the 0.9-orthologs + paralogs sequence pool with 10 repeats for each protein. Sequences sharing more than 90% sequence similarity were included (**A,B**) or redundancy was removed (**C,D**). Performance was evaluated using two methods: The log of the ratio between AUC-PR for CMA and random AUC-PR (A,C) and the precision of top *L*/5 prediction (B,D), where *L* is the protein length.

Removing highly similar proteins from the MSA

Researchers in the field frequently filter out sequences sharing high similarity between each other, using different identity cutoffs.[13,21,38,39] Our objective was to explore the tradeoff between reducing noise by filtering out highly similar sequences and losing the additional information such sequences may still carry. Thus, we have evaluated the performance achieved based on MSAs comprising sequences from our largest set of sequences (i.e., 0.5-orthologs + paralogs) and the MSAs comprised of the same set where highly similar sequences have been filtered out. We have explored the effect of removing sequences sharing either more than 95% identity or more than 90% identity. The results obtained based on these nonredundant MSAs compared with those achieved based on the full MSAs are shown in Figure 4. The results reveal that indeed there is justification for removing highly similar sequences from the MSA, although the gain is relatively modest. This conclusion is further supported by the comparison between panel (A) and (C) and between panel (B) and (D) of Figure 3.

Automated procedure for selecting sequences for the analysis

It is often hard to automatically differentiate between a real ortholog and a paralog of an ortholog. Relying on the prefixes of the protein entry names (the mnemonic code on the ID line) on the Swiss-Prot knowledgebase[40]

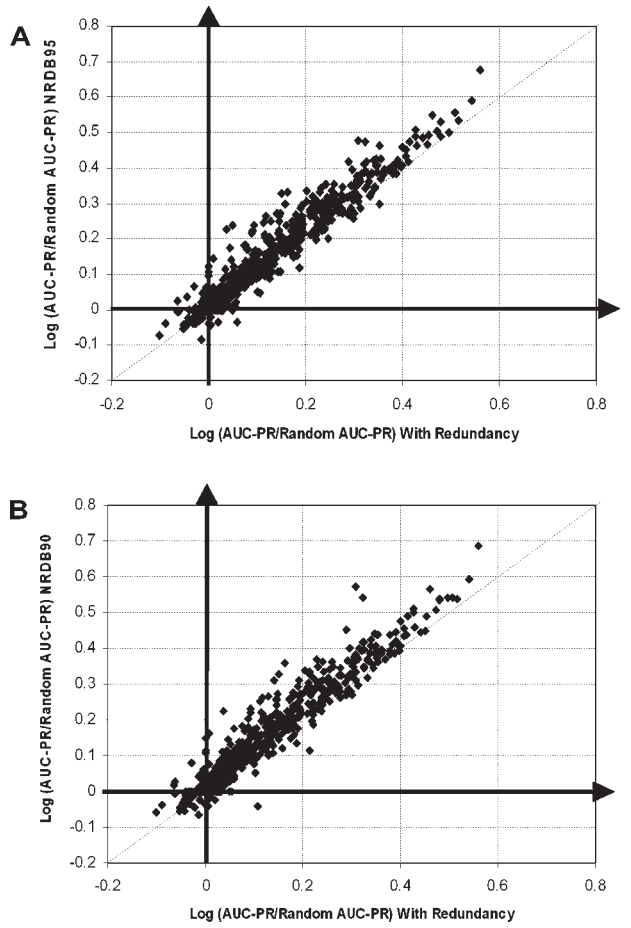

Figure 4

The effect of removing highly similar sequences from the MSA on the CMA performances. The performance achieved when using MSAs that contain highly similar sequences against the performances achieved when removing from the MSA sequences sharing (**A**) 95% (**B**) 90% sequence identity with other sequences in the MSA. For both panels most of the points are above the diagonal showing that filtering improves the accuracy of the prediction.

(SP) is another good way for making a list of orthologs.[41] Therefore, in addition to the orthologous set described in previous sections we constructed another set of orthologous sequences based on SP (SP-orthologs). It is worth noting that SP suffers from low coverage because it is manually annotated by expert curators. Thus the depths of the MSAs of most proteins (532 out of the 682) are smaller than 20 sequences.

To improve CMA performances, it is common to perform manual curation of MSAs.[4,13,14] However, this impedes high-throughput analysis, which is critical for robust performance analysis. The results obtained with the data preparation methods as were described in detail in the previous sections suggest that automated data preparation methods (without any manual curation) can be used.

Wilcoxon signed rank test for zero median perform two-sided test of the hypothesis that the difference between the matched samples of performances achieved by the different methods comes from a distribution whose median is zero. Wilcoxon signed rank tests were performed for all the results achieved based on the examined MSA sets, confirming that the differences between the performances achieved based on the different data preparation methods are statistically significant (e.g., the results achieved based on 0.9-orthologs MSAs and 0.9-orthologs + paralogs MSAs are statistically significant).

To further explore whether automated data preparation methods can be used, the performances were compared for each protein by ranking the AUC-PR achieved based on the MSAs derived from the different methods. This comparison allows us to examine the different data preparation methods for each protein from the culled PDB set. The number of times each data preparation method got the different ranks is shown in Figure 5.*

Figure 5(A) shows the number of times each data preparation method got the highest performances when considering the depth of 0.9-orthologs MSA. The leftmost bins on Figure 5(A) (depth < 16 and $16 <$ depth ≤ 128) further confirm our suggestions to increase the MSA depth by considering also sequences with varied length as well as to consider all the homologous sequences, especially when there is small number of close-orthologs. Furthermore, the bins that represent the performances that were achieved when there are relatively large amounts of orthologous sequences (orthologs-0.9 ≥ 128) reveals that the results gain with the 0.5-orthologs + paralogs are still relatively better than the results obtained with other tested MSAs. This is a remarkable finding, as the common wisdom in the field tends to assume that, paralogs and remote homologs contribute more noise than signal.[5,12,13,15]

Figure 5(B) reveals that, as expected, MSA that contains more homologous sequences (0.5-orthologs + paralogs) gains the best predictor performances. Significantly, the second best performances was gained by 0.5-orthologs MSAs, then by 0.7 orthologs + paralogs MSAs, and so forth, confirming the positive correlation between the number of proteins in an MSA and performance for most proteins. It is worth noting that the positive correlation between MSA depth and performance exists in all seven sets (Supplementary Fig. S2). The results also confirm our hypothesis that it is beneficial to increase the MSA depth by considering also sequences with varied length ($0.5 > 0.7 > 0.9$) as well as paralogs in addition to orthologous sequences. Figure 5(B) also reveals that the performance gain with SP-orthologs MSAs is the lowest for most of the proteins probably because of limited

*There are cases in which two or more data preparation methods got the same rank and therefore the sum of proteins getting a specific rank can be higher than the number of analyzed proteins.

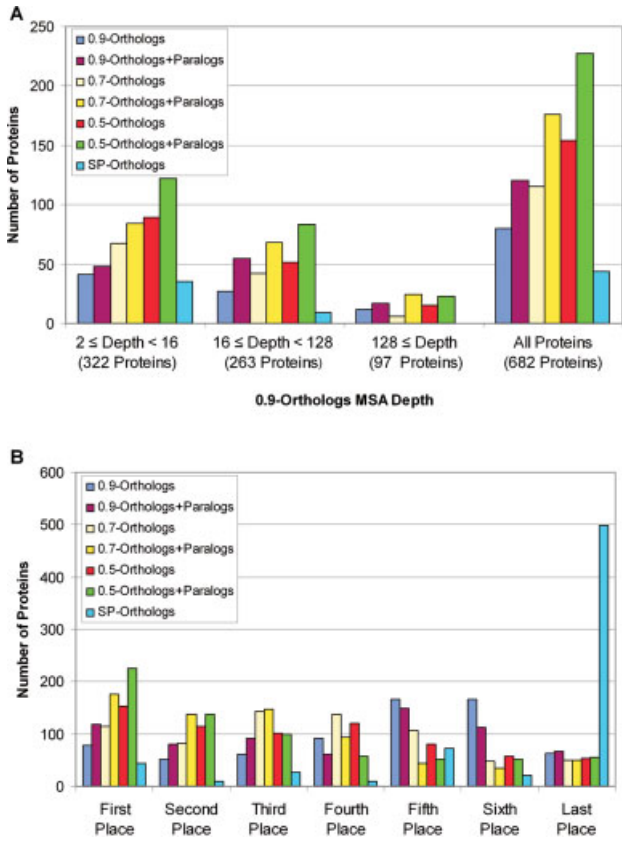

Figure 5

Comparing the performances achieved using different data preparation methods. **A**: The number of times each data preparation method was ranked first. The results are divided into three bins according to the depths of the respective 0.9-orthologs MSA. **B**: The number of times each data preparation method got the different ranks when considering all the proteins (682 proteins). The 0.5-orthologs + paralogs set was found to be superior over all other data preparation methods, but neither of the methods was optimal for all proteins.

number of annotated sequences. Unfortunately, Figure 5(B) also reveals that none of the data preparation methods is optimal for all proteins. Performances evaluation for top $L/5$ predictions (precision of $L/5$ predictions) reveals as expected, similar results, there are more cases in which the performances achieved based on 0.5-orthologs + paralogs MSAs is better than those achieved based on the other data preparation methods (data not shown).

In addition, to evaluate the results obtained by our data preparation method relative to those achieved using preprepared MSAs, we have compared the results obtained when using our optimal data preparation method (0.5-orthologs + paralogs) and Pfam[42] MSAs. The Pfam database contains information about protein domains and families and holds their MSAs. The performances that were achieved based on full Pfam MSAs compared with those achieved based on the 0.5-orthologs + paralogs MSAs are shown in Supplementary Figure S3. The results reveal that only for 19 (out of 87) proteins the performance obtained using Pfam MSAs were better than that achieved using the 0.5-orthologs + paralogs MSAs, whereas for 63 proteins the performance obtained with the 0.5-orthologs + paralogs MSAs were better. It is noteworthy that using Pfam preprepared MSAs is not always possible since Pfam does not always contain an MSA that fully cover the protein of interest, which is another advantage of our suggested data preparation method.

We believe that our data preparation method can be applied to most, if not all, CMA methods. To confirm our hypothesis we have repeated the analysis by using the same seven MSAs as described earlier (i.e., the 0.9-orthologs set, 0.9-orthologs + paralogs set, 0.7-orthologs set, 0.7-orthologs + paralogs set, 0.5-orthologs set, 0.5-orthologs + paralogs set, and SP-orthologs set) and evaluate the performance achieved by the MI method, which is based on the algorithm described by Martin *et al.*[6] To benefit from the large MSAs obtained by our data preparation method, we have slightly modified the algorithm such that it will not ignore positions that contain one gap, but rather only ignore positions containing more than 10% gaps (as suggested by the algorithm of Valencia and coworkers). The results obtained by the MI method based on the different set of MSAs are shown in Supplementary Figure S4. Importantly, the results are consistent with the conclusions made based on Valencia and coworkers' algorithm: (1) the performance obtained when using the 0.5-orthologs + paralogs set is superior over the results obtained with the other data preparation methods, and (2) none of the data preparation method is optimal for all the proteins.

Thus, based on our results we can recommend the following simple automatic procedure that does not involved manual curation to collect sequences for the CMA. The procedure outline is as follows: (1) Use PSI-Blast against the GenBank database, collecting all the sequences producing significant alignments after two iterations and share at least 35% identity with the sequence template constructed at the first iteration. (2) Collect all the sequences whose length of alignment with the reference is at least 50% of the reference sequence (i.e., both orthologous and paralogous sequences). (3) Remove redundancy of highly similar sequences sharing more than 90% identity with each other. Implementation of this procedure is available for academic use by request from the authors.

CONCLUSIONS

In this study, we have found that the inclusion of remote homologs is valuable for the analysis of correlated

mutations in many proteins. The data preparation procedure we put forward here with 50% alignment coverage of the query protein and including both orthologs and paralogs can contribute to increased accuracy of correlated mutation based analysis. It is worth noting that, due to the large diversity between proteins none of the data preparation methods outperformed all the others for all proteins tested.

We believe that our findings together with recent improvements in MSA preparation tools and CMA algorithms will contribute to significant improvement in the accuracy of contacts prediction based on CMA in the coming decade.

METHODS

Selection of proteins

The reference sequences for MSA and the PDB ID code of their corresponding crystal structures were taken according to the protein chain that appears at a nonredundant PDB set (Culled PDB[25]).

The sequences of a nonredundant set of the PDB (Culled PDB[25]) were downloaded from http://dunbrack.fccc.edu/Guoli/pisces_download.php on March 10, 2007. This set comprises of 2585 chains of X-ray-derived structures having resolution of or better than 2 Å, R-factor cutoff of 0.25, and do not share more than 20% sequence identity.

The references sequences for the 2585 chains of the culled PDB were chosen using Blast[20] against the SwissProt knowlegdebase[40] considering only proteins where the PDB sequence is at least 90% identical to a SwissProt protein sequence sharing alignment of at least 100 residues. This resulted in 1473 protein sequences.

Collecting homologs

The NCBI GenBank database[32] (comprising of 4,924,867 sequences) and the SP-Knowledgebase[40] (comprising of 247,428 sequences) were downloaded on December 24, 2006. The different sets of sequences were built by running PSI-Blast[20] against the GenBank database, collecting all the sequences producing significant alignments after two iterations and share at least 35% identity with the sequence template constructed at the first iteration. The sequences were further filtered according to the criteria specified for each set (i.e., orthologs, orthologs + paralogs) and according to the ratio between the length of the reference sequence and the length of the sequence and reference sequence alignment.

For constructing the orthologs by Swiss-Prot set we ran the PSI-Blast procedure against the SP knowledgebase considering sequences producing significant alignments after two iterations and sharing at least 35% identity with the sequence template constructed at the first iteration. Sequences with length of reference sequence-PSI-Blast hit alignment of at least 70% of the reference sequences length that have the same prefix of the Swiss-Prot entry name were selected for the SP-orthologs set.

Removing redundant sequences

A perl script that implements the NRDB algorithm[33] was used to remove sequences that share more than 90% sequence identity with each other and leaving in the MSA only one representative for such group of sequences. During this step, we made sure that the sequence of the protein of interest (reference sequence) remains in the sequences pool and was not replaced with one of its homolog.

Constructing MSA

Mafft (L-INS-i)[35] and ProbCons[34] were suggested by Nuin *et al.*[43] to be consistently the most accurate MSA programs. As our evaluation reveals (data not shown) that ProbCons v1.1 and Mafft (L-INS-i) v5.861 perform similarly, we chose to work with Mafft (L-INS-i) because it is much faster.

Pfam MSAs

The full Pfam-A[42] release 21.0 (Nov 2006) was downloaded from ftp://ftp.sanger.ac.uk/pub/databases/Pfam/database_files/old_releases/Pfam21.0/. Only MSAs that cover at least 90% of the selected PDB chain were considered. In total, we have evaluated the performances for 87 proteins. The MSAs were converted from Pfam Stockholm format to multiple fasta alignment format using AlignIO module written by Peter Schattner and which is part of the BioPerl[44] package.

Calculating correlated mutations

The correlated mutations for each protein were calculated using the program that was generously provided by Florencio Pazos[17] and results are reported as correlated coefficients between two positions. The algorithm excludes positions with more than 10% gaps or positions that are totally conserved. The exclusion of such positions impedes real comparison between performances achieved based on different sets of sequences for a given protein. Whereas when considering sequences with different level of relatedness we can expect varying number of gaps and as a results varying number of predictions and therefore bias when comparing the AUC-PR achieved based on the different sets. To overcome this obstacle we decided that the omitted pairs get correlation score of zero, based on the assumption that the coevolvement rate of these pairs is equal to the average coevolvement rate.

As suggested by Valencia and coworkers, we considered only pairs of residues with minimal separation of six res-

idues as predicting interactions with shorter sequence separation is not likely to add structural information.

We have implemented the MI method suggested by Martin *et al.*[6] such that it will not ignore positions where one of the sequences contains a gap. In our implementation, the mutual information is calculated based on all sequences that do not contain a gap in both positions. This step increased the number of predictions especially when dealing with MSAs containing remote homologs, where gaps are inevitable to construct an MSA. We have excluded positions with more than 10% gaps or positions that are totally conserved, as suggested by Valencia and coworkers.

Performance evaluation

The structure of the proteins was obtained from the PDB database.[45] The CSU program[46] was used to determine all the pairs that interact according to the protein crystal structure.

Based on the correlation coefficient values and the real interaction, a Precision-Recall curve was calculated for each protein [Recall = TP/(TP + FN) and Precision = TP/(TP + FP)]. The precision-recall curve allows us to evaluate the algorithm performances in all range of correlation scores and considering all the predicted interactions rather than focusing just on small part of the interactions as often done by those who calculates the precision for the top predictions. The precision recall curve provides an informative picture of the predictor performance even when dealing with highly skewed datasets.[47]

It is reasonable to assume that the number of interactions per residue is independent of proteins length, as we indeed found (data not shown). Hence the number of residue–residue interactions in proteins scales linearly with the protein length (N), whereas the number of all possible residue pairs is $\sim N^2$. Therefore, small proteins may seem to have better performance. To enable comparison of the performance on proteins with varied length, we have divided the area under the precision recall curve (AUC-PR),[47] by the AUC-PR calculated for a random predictor for each protein. The curve for a random predictor in Precision-Recall space is a horizontal line with a precision equal to the number of interactions observed in the proteins 3D structure divided by the number of all possible residue pairs. The AUC-PR was calculated using the AUCCalculator 0.1.[47]

In addition we have used CASP performance evaluation procedure to evaluate the performances. One of CASP performance evaluation procedures is based on calculating the precision [precision = TP/(TP + FP)] of the top $L/5$ or $L/10$ predictions, where L is the protein length.[26,27] The top predictions are the predictions that gain the higher predictor's probability estimates.

The Xd method[23] was used to also consider relative proximity between the predicted residues rather than consider only direct physical contact.

$$\mathrm{Xd} = \sum_{i=1}^{i=15} \frac{(\mathrm{Pip} - \mathrm{Pia})}{(d_i \times 15)}$$

There are 15 distance bins covering the range from 0 to 60 Å. The sum runs for all the distance bins. d_i is the upper limit representing each distance bin (normalized to 60). Pip is the percentage of predicted pairs whose distance is included in bin i. Pia is the same for all the pairs. Defined in this way, Xd > 0 indicates the positive cases where the population of predicted contacts' distances is shifted to lower distances. For this method, we considered our top $L/2$ predictions as the predicted contacts.

ACKNOWLEDGMENTS

The authors greatly thank Florencio Pazos for generously providing the program files and supplied all the needed explanations. They also thank Mark Goadrich for his fruitful insights regarding the AUC Calculator. The authors are grateful to A. Wool, E. Schreiber, A. Apatoff, A. Dan, and I. Borukhov for useful comments and helpful discussions.

REFERENCES

1. Gobel U, Sander C, Schneider R, Valencia A. Correlated mutations and residue contacts in proteins. Proteins 1994;18:309–317.
2. Choi SS, Li W, Lahn BT. Robust signals of coevolution of interacting residues in mammalian proteomes identified by phylogeny-aided structural analysis. Nat Genet 2005;37:1367–1371.
3. Eyal E, Frenkel-Morgenstern M, Sobolev V, Pietrokovski S. A pair-to-pair amino acids substitution matrix and its applications for protein structure prediction. Proteins 2007;67:142–153.
4. Fleishman SJ, Yifrach O, Ben-Tal N. An evolutionarily conserved network of amino acids mediates gating in voltage-dependent potassium channels. J Mol Biol 2004;340:307–318.
5. Kass I, Horovitz A. Mapping pathways of allosteric communication in GroEL by analysis of correlated mutations. Proteins 2002;48:611–617.
6. Martin LC, Gloor GB, Dunn SD, Wahl LM. Using information theory to search for co-evolving residues in proteins. Bioinformatics 2005;21:4116–4124.
7. Singer MS, Vriend G, Bywater RP. Prediction of protein residue contacts with a PDB-derived likelihood matrix. Protein Eng 2002; 15:721–725.
8. Fariselli P, Olmea O, Valencia A, Casadio R. Progress in predicting inter-residue contacts of proteins with neural networks and correlated mutations. Proteins 2001;Suppl 5:157–162.
9. Halperin I, Wolfson H, Nussinov R. Correlated mutations: advances and limitations. A study on fusion proteins and on the Cohesin-Dockerin families. Proteins 2006;63:832–845.
10. Noivirt O, Eisenstein M, Horovitz A. Detection and reduction of evolutionary noise in correlated mutation analysis. Protein Eng Des Sel 2005;18:247–253.
11. Dimmic MW, Hubisz MJ, Bustamante CD, Nielsen R. Detecting coevolving amino acid sites using Bayesian mutational mapping. Bioinformatics 2005;21 (Suppl 1):i126–i135.

12. Dutheil J, Pupko T, Jean-Marie A, Galtier N. A model-based approach for detecting coevolving positions in a molecule. Mol Biol Evol 2005;22:1919–1928.
13. Gloor GB, Martin LC, Wahl LM, Dunn SD. Mutual information in protein multiple sequence alignments reveals two classes of coevolving positions. Biochemistry 2005;44:7156–7165.
14. Wollenberg KR, Atchley WR. Separation of phylogenetic and functional associations in biological sequences by using the parametric bootstrap. Proc Natl Acad Sci USA 2000;97:3288–3291.
15. Dunn SD, Wahl LM, Gloor GB. Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. Bioinformatics 2008;24:333–340.
16. Fodor AA, Aldrich RW. Influence of conservation on calculations of amino acid covariance in multiple sequence alignments. Proteins 2004;56:211–221.
17. Pazos F, Olmea O, Valencia A. A graphical interface for correlated mutations and other protein structure prediction methods. Comput Appl Biosci 1997;13:319–321.
18. McLachlan AD. Tests for comparing related amino-acid sequences. Cytochrome c and cytochrome c 551. J Mol Biol 1971;61:409–424.
19. Olmea O, Valencia A. Improving contact predictions by the combination of correlated mutations and other sources of sequence information. Fold Des 1997;2:S25–S32.
20. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389–3402.
21. Kundrotas PJ, Alexov EG. Predicting residue contacts using pragmatic correlated mutations method: reducing the false positives. BMC Bioinformatics 2006;7:503.
22. Perez-Jimenez R, Godoy-Ruiz R, Parody-Morreale A, Ibarra-Molero B, Sanchez-Ruiz JM. A simple tool to explore the distance distribution of correlated mutations in proteins. Biophys Chem 2006;119:240–246.
23. Pazos F, Helmer-Citterich M, Ausiello G, Valencia A. Correlated mutations contain information about protein-protein interaction. J Mol Biol 1997;271:511–523.
24. Vicatos S, Reddy BV, Kaznessis Y. Prediction of distant residue contacts with the use of evolutionary information. Proteins 2005;58:935–949.
25. Wang G, Dunbrack RL, Jr. PISCES: a protein sequence culling server. Bioinformatics 2003;19:1589–1591.
26. Izarzugaza JM, Grana O, Tress ML, Valencia A, Clarke ND. Assessment of intramolecular contact predictions for CASP7. Proteins 2007;69:152–158.
27. Grana O, Baker D, MacCallum RM, Meiler J, Punta M, Rost B, Tress ML, Valencia A. CASP6 assessment of contact prediction. Proteins 2005;61 (Suppl 7):214–224.
28. Ortiz AR, Kolinski A, Rotkiewicz P, Ilkowski B, Skolnick J. Ab initio folding of proteins using restraints derived from evolutionary information. Proteins 1999;37:177–185.
29. Olmea O, Rost B, Valencia A. Effective use of sequence correlation and conservation in fold recognition. J Mol Biol 1999;293:1221–1239.
30. Cheng J, Baldi P. A machine learning information retrieval approach to protein fold recognition. Bioinformatics 2006;22:1456–1463.
31. Grana O, Eyrich VA, Pazos F, Rost B, Valencia A. EVAcon: a protein contact prediction evaluation service. Nucleic Acids Res 2005;33 (Web Server issue):W347–W351.
32. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL. GenBank. Nucleic Acids Res 2006;34 (Database issue):D16–D20.
33. Holm L, Sander C. Removing near-neighbour redundancy from large protein sequence collections. Bioinformatics 1998;14:423–429.
34. Do CB, Mahabhashyam MS, Brudno M, Batzoglou S. ProbCons: probabilistic consistency-based multiple sequence alignment. Genome Res 2005;15:330–340.
35. Katoh K, Kuma K, Toh H, Miyata T. MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Res 2005;33:511–518.
36. Schwartz AS, Pachter L. Multiple alignment by sequence annealing. Bioinformatics 2007;23:e24–e29.
37. Wallace IM, O'Sullivan O, Higgins DG, Notredame C. M-Coffee: combining multiple sequence alignment methods with T-Coffee. Nucleic Acids Res 2006;34:1692–1699.
38. Dekker JP, Fodor A, Aldrich RW, Yellen G. A perturbation-based method for calculating explicit likelihood of evolutionary co-variance in multiple sequence alignments. Bioinformatics 2004;20:1565–1572.
39. Shackelford G, Karplus K. Contact prediction using mutual information and neural nets. Proteins 2007;69 (Suppl 8):159–164.
40. Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, Pilbout S, Schneider M. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res 2003;31:365–370.
41. Release 8.3 UniProtKB/Swiss-Prot Headlines: of mice and men: over 10,000 orthologous sequence pairs in UniProtKB/Swiss-Prot. 2006. http://www.uniprot.org/news/2006/07/11/release
42. Finn RD, Mistry J, Schuster-Bockler B, Griffiths-Jones S, Hollich V, Lassmann T, Moxon S, Marshall M, Khanna A, Durbin R, Eddy SR, Sonnhammer EL, Bateman A. Pfam: clans, web tools and services. Nucleic Acids Res 2006;34 (Database issue):D247–D251.
43. Nuin PAS, Wang Z, Tillier ERM. The accuracy of several multiple sequence alignment programs for proteins. BMC Bioinformatics 2006;7:471.
44. Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, Dagdigian C, Fuellen G, Gilbert JG, Korf I, Lapp H, Lehvaslaiho H, Matsalla C, Mungall CJ, Osborne BI, Pocock MR, Schattner P, Senger M, Stein LD, Stupka E, Wilkinson MD, Birney E. The Bioperl toolkit: Perl modules for the life sciences. Genome Res 2002;12:1611–1618.
45. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The protein data bank. Nucl Acids Res 2000;28:235–242.
46. Sobolev V, Sorokine A, Prilusky J, Abola EE, Edelman M. Automated analysis of interatomic contacts in proteins. Bioinformatics 1999;15:327–332.
47. Jesse D, Mark G. The relationship between Precision-Recall and ROC curves. Proceedings of the 23rd international conference on Machine learning. Pittsburgh, Pennsylvania: ACM; 2006.