BMC Bioinformatics

BioMed Central

Predicting binding sites of hydrolase-inhibitor complexes by combining several methods

BMC Bioinformatics 2004, 5:205 doi:10.1186/1471-2105-5-205

Taner Z Sen (taner@iastate.edu)
Andrzej Kloczkowski (kloczkow@iastate.edu)
Robert L Jernigan (jernigan@iastate.edu)
Changhui Yan (chhyan@iastate.edu)
Vasant Honavar (honavar@iastate.edu)
Kai-Ming Ho (kmh@iastate.edu)
Cai-Zhuang Wang (cwang@iastate.edu)
Yungok Ihm (youngok@iastate.edu)
Haibo Cao (caohb@iastate.edu)
Xun Gu (xgu@iastate.edu)
Drena Dobbs (ddobbs@iastate.edu)

ISSN 1471-2105

Article type Methodology article

Submission date 23 Sep 2004

Acceptance date 17 Dec 2004

Publication date 17 Dec 2004

Article URL http://www.biomedcentral.com/1471-2105/5/205

Predicting binding sites of hydrolase-inhibitor complexes by combining several methods

Taner Z. Sen[*1,2], Andrzej Kloczkowski[1], Robert L. Jernigan[1,2,4], Changhui Yan[3,4], Vasant Honavar[1,3,4], Kai-Ming Ho[1,4,5], Cai-Zhuang Wang[4,5], Yungok Ihm[4,5], Haibo Cao[4,5], Xun Gu[1,4,6], Drena Dobbs[1,4,6]

1 L .H. Baker Center for Bioinformatics and Biological Statistics, Iowa State University Ames, IA 50011, USA

2 Department of Biochemistry, Biophysics, and Molecular Biology, Iowa State University, Ames, IA 50011, USA

3 Department of Computer Science, Iowa State University, Ames, IA 50011, USA

4 Bioinformatics and Computational Biology Program, Iowa State University, Ames, IA 50011, USA

5 Department of Physics and Astronomy, Iowa State University, Ames, IA 50011, USA

6 Department of Genetics, Development and Cell Biology, Iowa State University Ames, IA 50011, USA

e-mail: Taner Z. Sen, taner@iastate.edu; Andrzej Kloczkowski, kloczkow@iastate.edu; Robert L. Jernigan, jernigan@iastate.edu; Changhui Yan, chhyan@iastate.edu; Vasant Honavar, honavar@iastate.edu; Kai-Ming Ho, kmh@iastate.edu; Cai-Zhuang Wang, cwang@iastate.edu; Yungok Ihm, youngok@iastate.edu; Haibo Cao, caohb@iastate.edu; Xun Gu, xgu@iastate.edu; Drena Dobbs, ddobbs@iastate.edu

* Corresponding author

Abstract

Background: Protein-protein interactions play a critical role in protein function. Completion of many genomes is being followed rapidly by major efforts to identify interacting protein pairs experimentally in order to decipher the networks of interacting, coordinated-in-action proteins. Identification of protein-protein interaction sites and detection of specific amino acids that contribute to the specificity and the strength of protein interactions is an important problem with broad applications ranging from rational drug design to the analysis of metabolic and signal transduction networks.

Results: In order to increase the power of predictive methods for protein-protein interaction sites, we have developed a consensus methodology for combining four different methods. These approaches include: data mining using Support Vector Machines, threading through protein structures, prediction of conserved residues on the protein surface by analysis of phylogenetic trees, and the Conservatism of Conservatism method of Mirny and Shakhnovich. Results obtained on a dataset of hydrolase-inhibitor complexes demonstrate that the combination of all four methods yield improved predictions over the individual methods.

Conclusions: We developed a consensus method for predicting protein-protein interface residues by combining sequence and structure-based methods. The success of our consensus approach suggests that similar methodologies can be developed to improve prediction accuracies for other bioinformatic problems.

Background

Protein-protein interactions play a critical role in protein function. Completion of many genomes is being followed rapidly by major efforts to identify experimentally interacting protein pairs in order to decipher the networks of interacting, coordinated-in-action proteins. Identification of protein-protein interaction sites and detection of specific residues that contribute to the specificity and strength of protein interactions is an important problem[1-3] with broad applications ranging from rational drug design to the analysis of metabolic and signal transduction networks. Experimental detection of residues on protein-protein interaction surfaces can come either from determination of the structure of protein-protein complexes or from various functional assays. The ability to predict interface residues at protein binding sites using computational methods can be used to guide the design of such functional experiments and to enhance gene annotations by identifying specific protein interaction domains within genes at a finer level of detail than is currently possible.

Computational efforts to identify protein interaction surfaces[4-6] have been limited to date, and are needed because experimental determinations of protein structures and protein-protein complexes, lag behind the numbers of protein sequences. In particular, computational methods for identifying residues that participate in protein-protein interactions can be expected to assume an increasingly important role[4,5]. Based on the different characteristics of known protein-protein interaction sites[7], several methods have been proposed for predicting interface residues using a combination of sequence and structural information. These include methods based on the presence of “proline

brackets''[8], patch analysis using a 6-parameter scoring function[9,10], analysis of the hydrophobicity distribution around a target residue[7,11], multiple sequence alignments[12-14], structure-based multimeric threading[15], and analysis of amino acid characteristics of spatial neighbors to a target residue using neural networks[16,17]. Our recent work has focused on prediction of interface residues by utilizing analyses of sequence neighbors to a target residue using SVM and Bayesian classifiers[2,3].

There is an acute need for multi-faceted approaches that utilize available databases of protein sequences, structures, protein complexes, phylogenies, as well as other sources of information for the data-driven discovery of sequence and structural correlates of protein-protein interactions[4,5]. By exploiting available databases of protein complexes, the data-driven discovery of sequence and structural correlates for protein-protein interactions offers a potentially powerful approach.

Results and Discussion

Here we are using a dataset of 7 hydrolase complexes from the PDB, together with their sequence homologs. The application of our consensus method to other types of complexes, *e.g*. antibody-antigen complexes is currently under study and will be published later. It should be noted, however, that prediction of binding sites for other types of protein complexes, especially those involved in cell signaling, is likely to be more difficult than for the hydrolase-inhibitor complexes.

Figure 1 shows an example of the consensus method prediction mapped on the structure of proteinase B from *S. griseus* in a complex with turkey ovomucoid inhibitor (PDB 3sgb[18]). The inhibitor (3sgb_I) is shown at the top in wire frame and the proteinase B chain (3sgb_E), is shown at bottom. Actual interface residues in the proteinase B chain, i.e., amino acids that form the binding site between proteinase B and the inhibitor, were extracted from the PDB structure (see Materials and Methods). Predicted interface and non-interface residues, identified by the consensus method, are shown as color coded atoms as follows: Red spheres = true positives (TP), actual interface residues that are predicted as such; Gray strands = true negatives (TN), non-interface residues that are predicted as such; Yellow spheres = false negatives (FN), interface residues that are misclassified as non-interface residues; Blue spheres = false positives (FP), non-interface residues that are misclassified as interface residues. Note that the binding site in proteinase B is strongly indicated, with 14 out of 15 interface residues correctly classified, along with 2 false positives.

The primary amino acid sequence for proteinase B chain and the interface residue prediction results for the four individual methods and the consensus method are shown in Figure 2. Actual interface residues are identified highlighted in red. The five lines below the amino acid sequence show the locations of interface residues predicted by the different methods (described in detail below): P = Phylogeny; C = Conservatism of Conservatism (CoC); S = Data mining by SVM; T = Threading; E = Consensus. Similar Figures for each protein studied in this work are provided in Supplementary Materials [see Additional files 1, 2, 3, 4, 5, and 6].

The prediction results for all methods are shown in Table 1 and Table 2. Table 1 shows a complete summary of the classification performance on the proteinase B chain for all 5 methods including the overall Sensitivity (Sen) and Specificity (Spec); Sensitivity (Sen+) and Specificity (Spec+) for interface residues (the "positive" class); and Correlation Coefficient (see Materials and Methods for definitions of these performance parameters). Table 2 shows the overall average performance results for all seven protein complexes studied in this work. Two kinds of averages are considered: the numerical average over each of 7 proteins in the dataset, i.e., the average on a "per protein" basis ($<\ldots>_p$); and the average over the total number of residues, i.e., the average on a "per residue" basis ($<\ldots>_r$).

Sequence and structure conservation

Amino acid sequences are conserved for many different reasons related to the structure and function of proteins: for stability[19,20], enzyme active sites, subunit interfaces,

facilitation of an essential motion (hinges), and binding sites. Developing methods to identify the reason for conservation of individual highly conserved residues is a difficult problem. This is one of the reasons that a combination of approaches may be more likely to permit identification of residues that participate in protein-protein interactions. Even identifying the conserved residues themselves is not completely straightforward, and as will be seen, different approaches will indicate the same residue being conserved to different extents. In this study, we take advantage of this by using several methods to identify sequence and structure conservation. Here we use two principal methods for this purpose, one based on phylogeny to identify sequence conservation and one based on Conservatism of Conservatism[21] to identify structure conservation. These two methods often identify different residues as being conserved.

Phylogeny

To identify protein residues that are conserved – perhaps due to their functional role in forming specific protein-protein interactions - we use ClustalX[22] multiple sequence alignments of protein sequences to generate phylogenetic trees (see Materials and Methods). Conserved residues are defined as those that are identical at a given position in more than 85% alignments, i.e., only 15% substitutions or gaps were allowed. This 85% cutoff value is found to give optimal results (data not shown). Because phylogenetic trees of closely related sequences result in many residues that satisfy this condition (due to the high conservation of sequences, apparently important for protein folding, located in the protein core) we filter the results to focus on surface residues by removing conserved

residues residing inside the protein core, i.e., having low solvent accessibility (see Materials and Methods).

As shown in Figure 2, the phylogenetic method does not classify any of the amino acids in proteinase B chain (3sgb_E) as interface residues, i.e., TP = 0 and FP = 0. Thus, for the phylogenetic method prediction, the correlation coefficient (CC), which can range from –1 to +1, converges to zero, whereas overall specificity converges to 0.905. The latter misleading statistic is due to the large number of negative examples (non-interface residues), which are correctly classified. In cases such as this (with unbalanced numbers of positive and negative examples), *sensitivity+* and *specificity+* measures are especially useful because they more clearly reflect the ability of a method to detect "positive" interface residues. (See the Methods section for definition and further discussion of performance measures). Note that even though Figure 2 shows that the phylogenetic method does not identify any interface residues in this particular example, the results summarized in Table 1 for all seven proteins demonstrate that the ability of the phylogenetic method to correctly predict non-interface residues (reflected in the high overall sensitivity and specificity values), and in combination with other methods, to lead to significantly improved predictions.

Conservatism of Conservatism

To detect structurally conserved residues that are possible binding sites we have used the Conservatism of Conservatism method (CoC) developed by Mirny and Shakhnovich.[21] We use structural alignments generated by FSSP (fold classification based on structure-

structure alignment of proteins) developed by Holm and Sander[23]) to identify protein families with folds similar to that of the each of the 7 proteins. For each family, HSSP[24] (homology-derived secondary structure of proteins) alignments are used to calculate the sequence entropy at each position of the alignment. The HSSP profile is based on the multiple alignment of a sequence and its potential structural homologues[25]. The structural alignment generated by FSSP is used to calculate the value of CoC (see Materials and Methods). Each residue in the protein chain was ranked according to its CoC value at a given position in the sequence. The top 75% of total residues ranked according to their CoC values are defined as conserved. We filter the results of the CoC ranking by removing all structurally conserved residues located inside the protein core by only choosing the residues that have a relative accessibility of at least 25 as calculated by DSSP[26] (dictionary of protein secondary structure). Interface residues in proteinase B predicted by this method are indicated by a "C" in Figure 2. The overall performance of the CoC method is summarized in the second row of Tables 1 and 2. Although the correlation coefficient of the COC method is in the same range of those obtained by phylogeny and support vector machines, 0.37, the sensitivity+ value, 0.71, is surpassed only by the consensus value. Therefore, a larger fraction of interface residues is predicted by CoC than the other three methods. However, the CoC method alone is not sufficient to successfully predict binding sites, and combining this method with other prediction techniques in the consensus method gives improved results (Tables 1 and 2).

Data mining for binding residues

We have generated a support vector machine (SVM) classifier to determine whether or not a surface residue is located in the interaction site using information about the sequence neighbors of a target residue. An 11-residue window consisting of the residue and its 10 sequence neighbors (5 on each side) is chosen empirically. Each amino acid in the 11 residue window is represented using 20 values obtained from the HSSP profile of the sequence. Each target residue is therefore associated with a 220 (11×20) element vector. The SVM learning algorithm is given a set of labeled examples of the form (X, Y) where X is the 220 element vector representing a target residue and Y is its corresponding class label, either interface or non-interface residue. The SVM algorithm generates a classifier which takes as input a 220 element vector that encodes a target residue to be classified and outputs a class label. Our previous study[2] reported results for classifiers constructed using a combined set of 115 proteins belonging to six different categories of complexes: antibody-antigen, protease-inhibitor, enzyme complexes, large protease complexes, G-proteins, cell cycle signaling proteins, signal transduction, and miscellaneous. In another study[3]**,** we trained separate classifiers for each major category of complexes (e.g., protease-inhibitor complexes). In the case of protease-inhibitor complexes, leave-one-out experiments were performed on a set of 19 proteins. In each experiment, an SVM classifier was trained using a set of surface residues, labeled as interface or non-interface, from 18 of the 19 proteins. The resulting classifier was used to classify the surface residues of the remaining target protein into interface residue and non-interface residue categories. The interface residues obtained for 3sgb_E are reproduced in Figure 2 and marked by "S". The performance of the SVM classifier for

the current test set of complexes is summarized in Tables 1 and 2. The results show that SVM yields relatively high sensitivity+ (0.51) and specificity+ (0.41).

Threading of sequences through structures of interface surfaces

Structural threading was performed for the set of 7 protein complexes using a recently developed threading algorithm[27], which was first used in the CASP5[28] competition. For each complex structure, we first extract the interfacial region, essentially as described earlier. Residue-residue contacts in the interfacial region are described with contact matrices. The total energy in this threading method is the sum of all pair-wise contact energies for the conformation. Detailed residue–level contact potentials were obtained from the Li, Tang and Wingreen[29] parameterization of the Miyazawa and Jernigan[30] matrix. We represent a protein sequence vector **s** by the hydrophobicity values of its amino acids h_i obtained in this factorization and protein structure by the contact matrix $\mathbf{\Gamma}$. The problem of finding the best alignment of a query sequence **s** with a structure having contact matrix $\mathbf{\Gamma}$ is to find the transformation from **s** to **s'** that optimizes the energy function. The optimum **s'** is the dominant eigenvector $\mathbf{v_0}$ of the contact matrix $\mathbf{\Gamma}$. There is a strong correlation between a protein sequence and the dominant eigenvector of its native structure's contact matrix. Here the transformation we seek is obtained by maximizing the correlation between **s'** and $\mathbf{v_0}$. This is an alignment problem, and a dynamic programming method from sequence alignment has been adapted to solve this problem[27].

For each sequence, threading is performed against structures in our template database and alignment results used only when the score exceeds a length-dependent threshold. From the alignments, residues involved in contacts at the interface are identified using a scale based on the number of times a particular residue is indicated and the strength of the threading score. The predicted binding sites for 3sgb_E by the threading method are marked in Figure 1 by "T" and the prediction results are summarized in Tables 1 and 2. The threading-based approach is somewhat more successful than other methods based on its sensitivity+, selectivity+, and correlation coefficient values**,** but still not as good as the performance obtained by combining it with methods in the consensus approach.

Consensus method for predicting protein binding sites

Based on the results from the predictions with the four independent methods, we have developed a simple consensus method to obtain a better prediction. In the consensus method results presented here, an amino acid is considered to be an interface residue if any of the following conditions are met:

i) at least three independent methods classify it as an "interface residue"

ii) any two methods (except the Phylogeny-Threading pair) predict it

For this set of proteins, the parameters for combining results in the consensus method have been empirically determined without a systematic comparison of the strengths and weaknesses of each method. We employ this simple approach because it provides demonstrable improvement in prediction performance over the individual methods. The consensus interface residue predictions are indicated by an "E" in Figure 1, and performance results are summarized in the last rows of Tables 1 and 2. The consensus

method generally results in an enhanced correlation coefficient and sensitivity+, demonstrating the superior performance of the consensus method for identifying interface residues in this protein set. Predictions for each protein, provided in Supplementary Materials [see Additional files 1, 2, 3, 4, 5, and 6], illustrate that the improvements can be even more pronounced when the individual predictions of all four methods are relatively weak. This suggests that combining diverse prediction methods may be an excellent approach for the prediction of the binding sites in protein complexes.

Conclusions

Each of the four prediction methods presented in this paper sheds a different light on the conservation and prediction of protein interaction sites, but none of the methods taken separately is as powerful as the combination of all four methods. The simple consensus approach presented here could perhaps be improved by generating an ensemble predictor with more detailed probabilities. Our current work is directed at this approach. It is clear that the present subject is an active field of research[31-38].

Methods

Dataset of hydrolase-inhibitor complexes

The dataset of 7 hydrolase-inhibitor complexes used in this work has been derived from a larger dataset of 70 protein heterocomplexes extracted from PDB by Chakrabarti and Janin[39] and used in our previous studies[2,3]. All are proteins from hydrolase-inhibitor complexes, with six being proteinases: 1acb_E[40] (chain E of PDB structure 1acb), 1fle_E[41], 1hia_A[42], 1avw_A[43], 2sic_E[44], 3sgb_E[18]; and one being a carboxypeptidase: 4cpa[45].

Definition of surface and interface residues

Surface and interface residues for the proteins were identified based on information in the PDB coordinate files as previously described[2,3]. Briefly, solvent accessible surface areas (ASA) for each residue in the unbound protein and in the complex are calculated using DSSP[26]. A surface residue is defined as an interface residue if its calculated ASA in the complex is less than that in the monomer by at least 1 $Å^2$ [46]. In the extraction of interfacial region for threading, however, a distance-based definition of surface is used: a surface residue is defined as an interface residue if its side-chain center is within 6.5Å of the side-chain center of a residue belonging to another chain in the complex.

Based on the ASA definitions, 41% of the residues in the set of 7 proteins were surface residues, corresponding to a total of 631 surface residues. Among these surface residues, 166 were defined as interface residues and 465 as non-interface residues (i.e. surface residues that are not in the interaction sites). Thus, on average, interface residues represent 26% of surface residues, or 11% of total residues for proteins in our dataset.

Using phylogeny to identify conserved residues

Many computational tools have been developed for identifying amino acids that are important for protein function/structure, but there is no consensus regarding the best measure for evolutionary conservation[47]. Evolutionary conservation can be decomposed into three components: i) the overall selective constraints -- the number of changes observed at a site; ii) the pattern of amino acid substitutions – the number of amino acid types observed at a site; and iii) the effect of amino acid usage. We have established a reliable relationship between each measure and various aspects of structure. To explore the connection between sequence conservation and functional-structural importance, we proposed a new measure that can decompose the conservation into these three components[47]. This measure is based on phylogenetic analysis. The evolutionary rate at site k during lineage l from amino acids i to j (i,j=1,…20) can be expressed as $\lambda_{kl}(i,j) = c_k \times a_{lk} \times Q(i,j|k)$, where c_k accounts for the rate variation among sites, a_{lk} for site-specific lineage (or subtree) effect caused by functional divergence[48], and the 20×20 matrix $Q(i,j|k)$ is the (site-specific) model for amino acid substitutions. The likelihood function for a given tree can be determined according to a Markov chain model[49]. We have developed an integrated computer program (DIVERGE [50]) that can map these predicted sites onto the protein surface to examine these relationships. We use the solvent accessibility data from DSSP[26] to restrict predicted conserved residues to those located on the protein surface.

Conservatism of Conservatism

The phylogeny-based conservation of residues relies on sequence homology. It is well known, however, that many non-homologous proteins share similar folds[51]. It is therefore highly desirable to study the conservation of residues in proteins based on the structural superimposition of non-homologous proteins. In order to obtain insight into the evolutionary conservation of residues in proteins, we use the Conservatism of Conservatism method (CoC). The CoC method was developed by Mirny and Shakhnovich[21] for studying evolutionary conservation of residues in proteins with specific folds from the FSSP database[23]. With the FSSP database, Mirny and Shakhnovich performed an analysis of conserved residues in several common folds. The 20 naturally occurring amino acids were subdivided into 6 different classes, based on their physicochemical characteristics and frequencies of occurrence at different positions in multiple sequence alignments. The evolutionary conservatism within families of homologous proteins was measured through sequence entropy. Structural superimposition of different families of proteins with similar folds was used to calculate CoC for all positions of residues within a fold. Here we have applied a similar approach to identify structurally conserved residues involved in protein interactions.

For each protein, we first calculate the sequence entropy at each position within a family of related sequences from the HSSP database[25]

$$s(l) = -\sum_{i=1}^{6} p_i(l) \log p_i(l)$$

where $p_i^{(l)}$ is the frequency of the class i of residues (for each of the six classes) at position l in sequence in the multiple sequence alignment. Then we use the FSSP

database to obtain the structural alignment. The structural superimposition of different families was used to calculate the conservatism of conservatism (CoC)

$$S(l) = \frac{1}{M}\sum_{m=1}^{M} s^{m}(l)$$

where $s^{m}(l)$ is the intrafamily conservatism within the family m at position l, and M is the number of families. The CoC is the measure of the evolutionary conservation of the specific sites within the protein fold. Because the CoC method does not distinguish between residues at the protein surface evolutionarily conserved for functional reasons and residues inside the protein core that are conserved because of their importance to the folding process, we use solvent accessibility data for the unbound molecules to eliminate those conserved residues located inside the protein core.

Data mining approaches to binding site identification

Recent advances in machine learning[52] or data mining[53] offer a valuable approach to the data-driven discovery of complex relationships in computational biology[54,55]. In essence, a data mining approach uses a *representative data training set* to extract complex *a priori* unknown relationships, e.g., sequence correlates of protein-protein interactions. Examination of the resulting classifiers can help generate specific hypotheses that can be pursued using molecular and biophysical methods. For example, a classifier that is able to identify protein-protein interface residues on the basis of sequence or structural features can provide insights about sequence characteristics that contribute to important differences in function. The data mining approach for binding site identification consists of the following steps:

- Identify the surface residues in each protein.

- Label each residue in each protein as either an interface residue or a non-interface residue based on appropriate criteria for defining residues in interaction sites.
- Use a machine learning algorithm to train and evaluate a classifier to categorize a target amino acid as either an interface or a non-interface residue. Different types of information about the target residue (e.g., the identity and physicochemical properties of its sequence neighbors, whether or not the target residue is a surface residue) can be supplied as input to the classifier. A variety of machine learning algorithms[52,54] can be used for this purpose.
- Evaluate the classifier (typically using cross-validation or leave-one-out experiments) on independent test data (not used to train the classifier).
- Apply the classifier to identify putative interface residues in a protein, given its sequence (and possibly its structure), but not the sequence or structure of its interaction partner.

Here we have used a support vector machine (SVM) learning algorithm because SVMs are well-suited for the data-driven construction of high-dimensional patterns and are especially useful when the input is a real-valued pattern[56]. In addition, algorithms for constructing SVM classifiers effectively incorporate methods to avoid over-fitting the training data, thereby improving its generality, i.e., the performance of the resulting classifiers on test data. Support vector machine algorithms have proven effective in many applications, including text classification[57], gene expression analysis using microarray data[58], and predicting whether or not a pair of proteins is likely to interact[59].

Threading of sequences through structures of protein-protein interface surfaces

In phylogenetic and data mining approaches, the properties of the protein-protein interface are deduced by concentrating on the sequence information contained in the protein pair under investigation. However, it is well accepted that the physical origin of the specificity of protein-protein interactions comes predominantly from their structures. Thus, in any thorough investigation of protein-protein interactions, it is essential to include information from structural studies. Here we have adapted methods employed in protein structure recognition[60-63] to the problem of predicting protein-protein interface residues. In the first stage, structural models for identifying protein-protein interfaces are generated from existing protein databank (PDB) structures by extracting portions of contacts between different protein chains. We found that if we define the interaction region by the criterion that backbone C^{α} atoms on the two interacting chains are less than 15 Å apart, reasonably well connected fragments suitable for threading studies are obtained. In the second stage, after identifying a set of candidate template structures, threading is performed to examine the probability that a given model resembles the real interface. The threading algorithm is described in Cao et al.[27]. The threading alignments and scores obtained allowed us to predict which parts of each protein are in the interfacial region in the hydrolase-inhibitor complexes and to predict the most probable residue-residue contacts between the two proteins.

Ensemble predictions for combining results from multiple methods

Different approaches for identifying binding sites from amino acid sequence information yield different (sometimes contradictory, sometimes complementary) results. In such

cases, approaches for combining results from multiple predictors have a potential importance. The key idea is that results obtained by using different approaches, which we will call classifiers henceforth, may be correlated (or, more generally, statistically dependent) due to a variety of reasons including the use of a common dataset for constructing or tuning classifiers, use of intermediate variables for encoding input to the classifiers, and similarities between methods (e.g., SVM, neural networks). Regardless of the source of statistical dependency, the goal is to develop methods for weighting the output of each classifier appropriately for the purpose of producing more accurate predictions. Our method takes as input the binary (True/False) output of each classifier (e.g., SVM, CoC) and produces as output a probability that the residue under consideration is an interface residue, using the outputs produced by each of the classifiers. Algorithms for learning Bayesian (or Markov networks) can be then used to learn the network of dependences and the relevant conditional probabilities.

General evaluation measures for assessing the performance of classifiers

Let *TP* denote the number of true positives - residues predicted to be interface residues that are actually interface residues; *TN* the number of true negatives - residues predicted not to be interface residues that are in fact not interface residues; *FP* the number false positives - residues predicted to be interface residues that are not interface residues; *FN* the number of false negatives - residues predicted not to be interface residues that actually are interface residues. Let $N = TP+TN+FP+FN$. Sensitivity (recall) and Specificity (precision) are defined for the positive (+) class as well as the negative (-) class. $\text{Sensitivity}^{+} = TP/(TP+FN)$, $\text{Sensitivity}^{-} = TN/(TN+FP)$, $\text{Specificity}^{+} =$

$TP/(TP+FP)$, Specificity^{-} $=TN/(TN+FN)$. Overall sensitivity and overall specificity correspond to expected values of the corresponding measures averaged over both classes. The performance of the classifier is summarized by the correlation coefficient, which is given by

$$\frac{TP \times TN - FP \times FN}{\sqrt{(TP+FN)(TP+FP)(TN+FP)(TN+FN)}}$$

The correlation coefficient ranges from -1 to 1 and is a measure of how predictions correlate with the actual data[64]. It is important to note, that when the number of negative instances is much larger than the number of positive instances – as is the case for prediction of interface residues - the Sensitivity+ and Specificity+ measures are more appropriate for assessing prediction performance than the overall Sensitivity and Specificity measures[64]. In the extreme case when a classifier predicts every example to be negative (due to a preponderance of negative training instances) these overall performance measures would still show a high success rate despite the obvious failure of the prediction method. In such cases, the Correlation Coefficient, as well as the Sensitivity+, which is a measure of the fraction of positive instances that are correctly predicted, and Specificity+, which is a measure of the fraction of the positive predictions that are actually positive instances, may provide better performance assessment. Of course, a meaningful comparison of the performance of different classification methods depends critically on the specific application and goal.

Author's contributions

CY, VH and DD performed data mining calculations. XG performed phylogenetic calculations. KMH, CZW, YI, DD, and HC worked on threading. TZS, AK, and RLJ worked on the implementation of CoC and the development of consensus methodology. Every author contributed to the final draft of the paper.

Acknowledgments

The financial support through the NIH grant 1R21GM066387 is acknowledged by V. Honavar, D. Dobbs and R.L. Jernigan. We thank Dimitris Margaritis and other members of our research groups for helpful discussions. We also wish to thank the anonymous reviewers for valuable comments on the original version of this manuscript.

References

1. C Chothia, J Janin: **Principles of Protein-Protein Recognition.** *Nature* 1975, **256:**705-708.

2. CH Yan, V Honavar, D Dobbs: **Identification of interface residues in protease-inhibitor and antigen-antibody complexes: a support vector machine approach.** *Neural Computing & Applications* 2004, **13:**123-129.

3. C Yan, D Dobbs, V Honavar: **A two-stage classifier for identification of protein-protein interface residues.** *Bioinformatics* 2004, **20:**i371-i378.

4. SA Teichmann, AG Murzin, C Chothia: **Determination of protein function, evolution and interactions by structural genomics.** *Curr Opin Struct Biol* 2001, **11:**354-363.

5. A Valencia, F Pazos: **Computational methods for the prediction of protein interactions.** *Curr Opin Struct Biol* 2002, **12:**368-373.

6. A Valencia, F Pazos: **Prediction of protein-protein interactions from evolutionary information.** In *Structural Bioinformatics*.Edited by Bourne PE, Weissig H. USA: John Wiley & Sons; 2003:411-426.

7. L Young, RL Jernigan, DG Covell: **A role for surface hydrophobicity in protein-protein recognition.** *Prot Sci* 1994, **3:**717-729.

8. RM Kini, HJ Evans: **Prediction of potential protein-protein interaction sites from amino acid sequence. Identification of a fibrin polymerization site.** *FEBS Lett* 1996, **385:**81-86.

9. S Jones, JM Thornton: **Prediction of protein-protein interaction sites using patch analysis.** *J Mol Biol* 1997, **272:**133-143.

10. S Jones, JM Thornton: **Analysis of protein-protein interaction sites using surface patches.** *J Mol Biol* 1997, **272:**121-132.

11. X Gallet, B Charloteaux, A Thomas, R Brasseur: **A fast method to predict protein interaction sites from sequences.** *J Mol Biol* 2000, **302:**917-926.

12. G Casari, C Sander, A Valencia: **A method to predict functional residues in proteins.** *Nat Struct Biol* 1995, **2:**171-178.

13. O Lichtarge, HR Bourne, FE Cohen: **An evolutionary trace method defines binding surfaces common to protein families.** *J Mol Biol* 1996, **257:**342-358.

14. F Pazos, M Helmer-Citterich, G Ausiello, A Valencia: **Correlated mutations contain information about protein-protein interaction.** *J Mol Biol* 1997, **271:**511-523.

15. L Lu, H Lu, J Skolnick: **MULTIPROSPECTOR: an algorithm for the prediction of protein-protein interactions by multimeric threading.** *Proteins* 2002, **49:**350-364.

16. P Fariselli, F Pazos, A Valencia, R Casadio: **Prediction of protein--protein interaction sites in heterocomplexes with neural networks.** *Eur J Biochem* 2002, **269:**1356-1361.

17. HX Zhou, Y Shan: **Prediction of protein interaction sites from sequence profile and residue neighbor list.** *Proteins* 2001, **44:**336-343.

18. RJ Read, M Fujinaga, AR Sielecki, MN James: **Structure of the complex of Streptomyces griseus protease B and the third domain of the turkey ovomucoid inhibitor at 1.8-A resolution.** *Biochemistry* 1983, **22:**4420-4433.

19. OB Ptitsyn, KL Ting: **Non-functional conserved residues in globins and their possible role as a folding nucleus.** *J Mol Biol* 1999, **291:**671-682.

20. KL Ting, RL Jernigan: **Identifying a folding nucleus for the lysozyme/alpha-lactalbumin family from sequence conservation clusters.** *J Mol Evol* 2002, **54:**425-436.

21. LA Mirny, EI Shakhnovich: **Universally conserved positions in protein folds: Reading evolutionary signals about stability, folding kinetics and function.** *J Mol Biol* 1999, **291:**177-196.

22. JD Thompson, TJ Gibson, F Plewniak, F Jeanmougin, DG Higgins: **The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools.** *Nucl Acids Res* 1997, **24:**4876-4882.

23. L Holm, C Sander: **Protein structure comparison by alignment of distance matrices.** *J Mol Biol* 1993, **233:**123-138.

24. C Sander, R Schneider: **Database of homology derived protein structures and the structural meaning of sequence alignment.** *Proteins* 1991, **9:**56-58.

25. C Dodge, R Schneider, C Sander: **The HSSP database of Protein Structure-Sequence Alignments and Family Profiles.** *Nucl Acids Res* 1998, **26:**313-315.

26. W Kabsch, C Sander: **Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features.** *Biopolymers* 1983, **22:**2577-2637.

27. H Cao, Y Ihm, CZ Wang, JR Morris, M Su, D Dobbs, KM Ho: **Three-dimensional threading approach to protein structure recognition.** *Polymer* 2004, **45:**687-697.

28. J Moult, F Fidelis, A Zemla, T Hubbard: **Critical assessment of methods of protein structure prediction (CASP)-round V.** *Proteins* 2003, **53:**334-339.

29. H Li, C Tang, NS Wingreen: **Nature of Driving Force for Protein Folding: A Result From Analyzing the Statistical Potential.** *Phys Rev Lett* 1997, **79:**765-768.

30. S Miyazawa, RL Jernigan: **Estimation of Effective Interresidue Contact Energies From Protein Crystal-Structures - Quasichemical Approximation.** *Macromolecules* 1985, **18:**534-552.

31. D Carugo, G Franzot: **Prediction of protein-protein interactions based on surface patch comparison.** *Proteomics* 2004, **4:**1727-1736.

32. H Lu, L Lu, J Skolnick: **Development of Unified Statistical Potentials Describing Protein-Protein Interactions.** *Biophys J* 2003, **84:**1895-1901.

33. L Lu, AK Arakaki, H Lu, J Skolnick: **Multimeric Threading-Based Prediction of Protein-Protein Interactions on a Genomic Scale: Application to the Saccharomyces cerevisiae Proteome.** *Genome Res* 2003, **13:**1146-1154.

34. S Martin, D Roe, JL Faulon: **Predicting protein-protein interactions using signature products.** *Bioinformatics* 2004,bth483.

35. H Neuvirth, R Raz, G Schreiber: **ProMate: A Structure Based Prediction Program to Identify the Location of Protein-Protein Binding Sites*1.** *Journal of Molecular Biology* 2004, **338:**181-199.

36. JC Obenauer, MB Yaffe: **Computational prediction of protein-protein interactions.** *Methods Mol Biol* 2004, **261:**445-468.

37. Y Ofran, B Rost: **Predicted protein-protein interaction sites from local sequence information.** *FEBS Lett* 2003, **544:**236-239.

38. A Valencia, F Pazos: **Prediction of protein-protein interactions from evolutionary information.** *Methods Biochem Anal* 2003, **44:**411-426.

39. P Chakrabarti, J Janin: **Dissecting protein-protein recognition sites.** *Proteins* 2002, **47:**334-343.

40. F Frigerio, A Coda, L Pugliese, C Lionetti, E Menegatti, G Amiconi, HP Schnebli, P Ascenzi, M Bolognesi: **Crystal and molecular structure of the bovine alpha-chymotrypsin-eglin c complex at 2.0 A resolution.** *J Mol Biol* 1992, **225:**107-123.

41. M Tsunemi, Y Matsuura, S Sakakibara, Y Katsube: **Crystal structure of an elastase-specific inhibitor elafin complexed with porcine pancreatic elastase determined at 1.9 A resolution.** *Biochemistry* 1996, **35:**11570-11576.

42. PR Mittl, S Di Marco, G Fendrich, G Pohlig, J Heim, C Sommerhoff, H Fritz, JP Priestle, MG Grutter: **A new structural class of serine protease inhibitors revealed by the structure of the hirustasin-kallikrein complex.** *Structure* 1997, **5:**253-264.

43. HK Song, SW Suh: **Kunitz-type soybean trypsin inhibitor revisited: refined structure of its complex with porcine trypsin reveals an insight into the interaction between a homologous inhibitor from Erythrina caffra and tissue-type plasminogen activator1.** *J Mol Biol* 1998, **275:**347-363.

44. Y Takeuchi, Y Satow, KT Nakamura, Y Mitsui: **Refined crystal structure of the complex of subtilisin BPN' and Streptomyces subtilisin inhibitor at 1.8 A resolution.** *J Mol Biol* 1991, **221:**309-325.

45. DC Rees, WN Lipscomb: **Refined crystal structure of the potato inhibitor complex of carboxypeptidase A at 2.5 A resolution.** *J Mol Biol* 1982, **160:**475-498.

46. S Jones, JM Thornton: **Principles of protein-protein interactions.** *Proc Natl Acad Sci U S A* 1996, **93:**13-20.

47. R Durbin, S Eddy, A Krogh, G Mitchison: *Biological sequence analysis: probabilistic models of proteins and nucleic acids* Cambridge, U.K.: Cambridge University Press; 1998.

48. X Gu: **Statistical methods for testing functional divergence after gene duplication.** *Mol Biol Evol* 1999, **16:**1664-1674.

49. J Felsenstein: **Evolutionary trees from DNA sequences:a maximum likelihood approach.** *J Mol Evol* 1981, **17:**368-376.

50. X Gu, K Vander Velden: **DIVERGE: Phylogeny-based Analysis for Functional-Structural Divergence of a Protein .** *Bioinformatics* 2002, **18:**500-501.

51. DV Laurents, S Subbiah, M Levitt: **Different protein sequences can give rise to highly similar folds through different stabilizing interactions.** *Prot Sci* 1994, **3:**1938-1944.

52. T Mitchell: *Machine Learning* New York: Mc-Graw Hill; 1997.

53. IH Witten, E Frank: *Data mining: Practical machine learning tools and techniques with java implementations* San Mateo, CA: Morgan Kaufmann; 1999.

54. P Baldi, S Brunak: *Bioinformatics: The Machine Learning Approach*. 2nd edition Cambridge, MA: MIT Press; 2001.

55. NM Luscombe, D Greenbaum, M Gerstein: **What is bioinformatics? A proposed definition and overview of the field.** *Methods Inform Med* 2001, **40:**346-358.

56. V Vapnik: *Statistical learning theory* New York: Springer-Verlag; 1998.

57. MA Hearst, B Scholkopf, S Dumais, E Osuna, J Platt: **Trends and controversies - support vector machines.** *IEEE Intelligent Systems* 1998, **13:**18-28.

58. MPS Brown, WN Grundy, D Lin, N Christianini, CWS Sugnet, T Furey, M Ares Jr., D Haussler: **Knowledge based analysis of microarray gene expression data using support vector machines.** *Proc Natl Acad Sci USA* 2000, **97:**262-267.

59. JR Bock, DA Gough: **Predicting protein--protein interactions from primary structure.** *Bioinformatics* 2001, **17:**455-460.

60. A Godzik, J Skolnick: **Sequence-structure matching in globular proteins: application to supersecondary and tertiary structure determination.** *Proc Natl Acad Sci USA* 1992, **89:**12098-12102.

61. DT Jones, RT Miller, JM Thornton: **Successful protein fold recognition by optimal sequence threading validated by rigorous blind testing.** *Proteins* 1995, **23:**387-397.

62. J Meller, R Elber: **Linear programming optimization and a double statistical filter for protein threading protocols.** *Proteins* 2001, **45:**241-261.

63. S Miyazawa, RL Jernigan: **Identifying sequence-sequence pairs undetected by sequence alignments.** *Protein Eng* 2000, **13:**459-475.

64. P Baldi, S Brunak, Y Chauvin, CAF Andersen, H Nielsen: **Assessing the accuracy of prediction algorithms for classification: an overview.** *Bioinformatics* 2000, **16:**412-424.

FIGURE CAPTIONS

Figure 1. Interface residues predictions mapped on the three dimensional structure of Proteinase B from *Streptomyces griseus* (3sgb). The target protein is shown in ribbons and atomic spheres; the inhibitor partner is shown at the top in faint wire frame. The residues are color coded as: red = true positives (TP), gray = true negatives (TN), yellow = false negatives (FN), and blue = false positives (FP). Red, yellow, and blue residues are shown in spacefill representation. Note that the actual interface residues extracted from the PDB structure include the red (TP) and yellow (FN) residues. Red and gray residues represent correct predictions of interface and non-interface residues (14 TP+ 210 TN = 224 correct predictions); yellow and blue residues represent incorrect predictions (1 FN + 2 FP= 3)

Figure 2 Comparison of individual methods for interface residue prediction with the consensus method. Results are shown for Proteinase B from *Streptomyces Griseus* (3sgb_E), the same protein shown in Figure 1. Actual interfaces are highlighted in red. Interface residues predicted by each of five different methods are indicated as follows: P = Phylogeny (none predicted for this protein), C = Conservatism of Conservatism; S = Support Vector Machine; T = Threading; and E = Consensus. Amino acid residues present in the protein sequence, but not included in the PDB structure file, are indicated by “X”s in the sequence.

Table 1 Classification results for Proteinase B from *S. griseus* (3sgb_E). TP is the number of true positive; TN is the number of true negatives; FP is the number of false positives, and FN is the number of false negatives. Overall sensitivity, overall specificity, sensitivity+, specificity+, and correlation coefficient are defined in the text.

3SGBE	TP #	TN #	FP #	FN #	Overall Sen	Overall Spe	Sen+	Spe+	CC
Phylog.	0	212	0	15	0.94	0.91*	0	-	0*
COC	15	194	18	0	0.92	0.96	1	0.45	0.64
SVM	3	205	7	12	0.92	0.90	0.20	0.30	0.20
Thread.	14	201	11	1	0.95	0.97	0.93	0.56	0.70
Cons.	14	210	2	1	0.99	0.99	0.94	0.88	0.90

Table 2. Overall Classification Performance Results Averaged over 7 Proteins.

Average results for Sensitivity+, Specificity+, overall Sensitivity, overall Specificity, and Correlation Coefficient averaged over the 7 proteins in the dataset. $< >_p$ denotes averaging over the total number of proteins, $< >_r$ denotes averaging over the total number of residues.

Method	**$<Sen+>_p$**	**$<Spe+>_p$**	**$<Spe>_p$**	**$<Spe>_r$**	**$<Sen>_p$**	**$<Sen>_r$**	**$<Cor>_p$**	**$<Cor>_r$**
Phylog.	0.39	0.71	0.90	0.89	0.91	0.89	0.43	0.37
COC	0.71	0.31	0.89	0.88	0.81	0.80	0.38	0.37
SVM	0.51	0.41	0.89	0.88	0.88	0.88	0.39	0.37
Thread.	0.59	0.57	0.91	0.89	0.92	0.91	0.53	0.48
Cons.	0.70	0.56	0.92	0.91	0.90	0.89	0.56	0.55

ADDITIONAL SUPPLEMENTARY FILES

1) 1acb_e.pdf, Adobe Portable Document Format: Comparison of individual methods for interface residue prediction for bovine α-chymotrypsin (1acbe).

2) 1avw_a.pdf, Adobe Portable Document Format: Comparison of individual methods for interface residue prediction for porcine pancreatic trypsin (1avwa).

3) 1fle_e.pdf, Adobe Portable Document Format: Comparison of individual methods for interface residue prediction for porcine pancreatic elastase (1flee).

4) 1hia_a.pdf, Adobe Portable Document Format: Comparison of individual methods for interface residue prediction for kallikrein (1hiaa).

5) 2sic_e.pdf, Adobe Portable Document Format: Comparison of individual methods for interface residue prediction for subtilisin BPN' (2sice).

6) 4cpa.pdf, Adobe Portable Document Format: Comparison of individual methods for interface residue prediction for carboxypeptidase A (4cpa).

Figure 1

```
        10        20        30        40        50        60
ISGGXXXXXXXXXDAIYSSXXXXTGRCSLGFNVTYYFLTAGHCTDXGATTWWAXXXXXXX

                                    C
                                     S
                          TTTT                T

        70        80        90       100       110       120
XXNSARTTVLGTTSGSXSFXXXXPNNDYGIVRYTNTTIPKDGTVGGQDITSAANATVGMA

           C                              C C             CC      C
         S                                                       S
                    T       T

       130       140       150       160       170       180
VTRRGSTTXXXXXXXXXXXXXXGTHSGSVTALNATVNYGGGDVVYGMIRTNXXXXXVCAGDS

                          CC C     CCCCCCC                      CC
                               S    S S       S
      T                             TTTTT                  TT T
                                    EEEEE                     E

       190       200       210       220
GGPLYSGXXXXTRAIGLTSGGSGNCSSGGTTFFQPVTEALAYGVSVY

                  CCCCCCCCCC     C          CCC
                         S       S           S
                  TTTTTTT       TT
                  EEEEEEEE       E           E
```

Figure 2

Additional files provided with this submission:

Additional file 1: 1acbe_e.pdf : 21KB
http://www.biomedcentral.com/imedia/1496560625568275/sup1.pdf

Additional file 2: 1avw_a.pdf : 19KB
http://www.biomedcentral.com/imedia/7843543675682752/sup2.pdf

Additional file 3: 1fle_e.pdf : 28KB
http://www.biomedcentral.com/imedia/2065250409568275/sup3.pdf

Additional file 4: 1hia_a.pdf : 13KB
http://www.biomedcentral.com/imedia/1532649308568275/sup4.pdf

Additional file 5: 2sic_e.pdf : 18KB
http://www.biomedcentral.com/imedia/3373843395682758/sup5.pdf

Additional file 6: 4cpa.pdf : 47KB
http://www.biomedcentral.com/imedia/7740914995682760/sup6.pdf