Published in final edited form as:

Int J Data Min Bioinform. 2008 December 11; 2008: 703–708. doi:10.1109/ICMLA.2008.74.

Prediction of inter-residue contact clusters from hydrophobic cores

Peng Chen,

Department of Systems and Computer Science, Howard University, 2400 Sixth Street, NW
Washington, DC 20059, USA

Chunmei Liu,

Department of Systems and Computer Science, Howard University, 2400 Sixth Street, NW
Washington, DC 20059, USA

Legand Burge,

Department of Systems and Computer Science, Howard University, 2400 Sixth Street, NW
Washington, DC 20059, USA

Mohammad Mahmood,

Department of Mathematics, Howard University, 2400 Sixth Street, NW Washington, DC 20059,
USA

William Southerland, and

Department of Biochemistry, Howard University, 2400 Sixth Street, NW Washington, DC 20059,
USA

Clay Gloster

Department of Electrical and Computer Engineering, Howard University, 2400 Sixth Street, NW
Washington, DC 20059, USA

Peng Chen: pchen1978@gmail.com; Chunmei Liu: chunmei@scs.howard.edu; Legand Burge: blegand@scs.howard.edu;
Mohammad Mahmood: mmahmood@howard.edu; William Southerland: wsoutherland@howard.edu; Clay Gloster:
csgloster@howard.edu

Abstract

A contact map is a key factor representing a specific protein structure. To simplify the protein contact map prediction, we predict the inter-residue contact clusters centered at the groups of their surrounding inter-residue contacts. In this paper, we adopt a Support Vector Machine (SVM)-based approach to predict the inter-residue contact cluster centers. The input of the SVM predictor includes sequence profile, evolutionary rate and predicted secondary structure. The SVM predictor is based on hydrophobic cores that may be considered as locations of the inter-residue contact clusters. About 35% of clustering centers of inter-residue contacts can be predicted accurately.

Keywords

SVM; support vector machine; contact cluster; contact cluster centre; hydrophobic core

1 Introduction

With the fast development of genetic and biological techniques, more and more new proteins are being sequenced. Nevertheless, there always exists a colossal gap between the number of sequenced primary sequences and the number of proteins with unsolved three-dimensional (3D) structures. The slow development on exploring protein's 3D structures is due to the complicated problem itself and the lack of effective experimental techniques. Therefore, it is necessary to use computational techniques to find efficient tools for predicting the 3D structure of proteins. Among them, the prediction based on contact map is accurate.

The native contact map of a protein chain is a matrix in which element (i, j) equals to 1 if the two residues i and j are in contact in the native structure, and 0 otherwise. It is well known that a contact map of protein, which may be advanced to the protein tertiary structure prediction and protein folding, is an intermediate state from primary structure to tertiary structure of the protein (Laskowski et al., 1993; Niggemann and Steipe, 2000). Gromiha and Selvaraj (2004) reported that long-range inter-residue interactions play an important role in the folding and stability of proteins and they thus can be used to predict the 3D structure of proteins. Vendruscolo et al. (1997) argued that even a corrupted contact map can be used to reconstruct its corresponding 3D protein structure.

Many previous works have addressed the prediction of contact map (Thomas et al., 1996; Fariselli et al., 2001; Pollastri and Baldi, 2002; MacCallum, 2004; Punta and Rost, 2005; Vicatos et al., 2005; Vullo et al., 2006; Chen et al., 2007; Cheng and Baldi, 2007; Chen et al., 2008a, 2008b). Thomas et al. (1996) proposed an approach to predict protein contacts based on mutational behaviour of pairs of amino acid residues, which is deduced from multiple sequence alignments. Fariselli et al. (2001) predicted contact maps of proteins with neural-network-based methods, using input coding of increasing complexity including evolutionary information, sequence conservation, correlated mutations and predicted secondary structures. Pollastri and Baldi (2002) proposed a recurrent neural network called Generalised Input-Output Hidden Markov Models (GIOHMMs) for the prediction of contact maps, as well as other information processing and pattern recognition tasks. MacCallum (2004) proposed a Self-Organising Map (SOM)- and Genetic Programming (GP)-based approach to predict contact map, which uses GP to select residues and residue pairs more likely to make contacts based solely on local sequence patterns extracted with the help of SOMs. PROFcon method using a neural network declared that about 30% of the predicted contacts were corrected, considering all contacts between residue pairs that are separated by at least six residues along peptide chains (Punta and Rost, 2005). Vullo's two-stage predictor obtained 19.8% accuracy for top $L/5$ predicted contacts between residues with sequence separation of 24 residues or more, where L denotes sequence length (Vullo et al., 2006). Recently, SVMcon predictor used SVMs to predict medium- and long-range contacts integrated profiles, secondary structure, relative solvent accessibility, contact potentials and other useful features (Cheng and Baldi, 2007). The prediction of contact maps has been studied extensively, but the low prediction accuracy makes it very difficult to be used for predicting 3D structures of proteins. Therefore, it is necessary to put forward some novel constructive method for predicting contact maps.

The amino acid residues in protein structures interact with each other and form clusters (Gromiha and Selvaraj, 2004). We expect that an SVM predictor based on hydrophobic cores, which are formed by hydrophobic interactions and grouped by hydrophobic residues, can solve the problem of the prediction of contacts and contact maps. Generally, hydrophobic interaction is considered to be a dominant feature to maintain protein's 3D structures. It is a common knowledge that a region of high hydrophobicity will be energetically stabilised if it is in proximity to another high hydrophobic region, rather than close to a hydrophilic (polar) region. Similar arguments can be concluded for regions of low hydrophobicity too (Gupta et al.,

2005). Furthermore, Gromiha and Selvaraj (2004) reported that hydrophobic interaction is a dominant force in protein folding and is mainly dominated by long-range interactions using a thorough statistics. Heringa and Argos (1991) used a cut-off radius of 4.5 Å between side-chain atoms to delineate amino acid clusters and showed that most of the clusters are composed of three to four residues and are localised near the protein surface. On the other hand, Zehfus (1995) reported that an average of 65% of hydrophobic residues is involved in residue clusters and each hydrophobic cluster contains at least seven residues. Selvaraj and Gromiha (2003) indicated the vital role of long-range interactions to form the hydrophobic clusters and to stabilise the proteins. Further, they reported that the long-range interactions contribute an appreciably higher percentage in the hydrophobic clusters of β -strands or turn regions near the strands, and there is no significant contribution from medium-range interactions. However, the medium-range interactions play a dominant role in the hydrophobic clusters formed by α -helices. They argued that this observation is consistent with the previous results that α -helices are influenced by medium-range interactions and β -strands are dominated by long-range interactions (Gromiha and Selvaraj, 1998; Selvaraj and Gromiha, 2003). All these results show the relationship of hydrophobic residues and the formation of inter-residue contact clusters, which are important for the folding and stability of protein structures. Thus, we observed that most pairs of amino acids in contact are located in the neighbourhood of hydrophobic cores. Therefore, clustering the natural groups of inter-residue contacts as well as studying the correspondence between the contact clusters and pairs of residues in high hydrophobic regions can improve the prediction accuracy of inter-residue contacts.

In this work, we address the problem of locating the key inter-residue contact sites by studying the correspondence between contact clusters and pairs of residues in high hydrophobic regions. At the beginning, we construct an SVM predictor, whose input vectors contain information from sequence profiles, from evolutionary rates and the predicted secondary structures. The SVM predictor is based on hydrophobic cores that may be considered as locations of inter-residue contact clusters. Therefore, about 35% of clustering centres of inter-residue contacts can be predicted accurately.

2 Methods

2.1 Materials and data sets

We obtained 776 protein chains using PDB-REPRDB (Noguchi and Akiyama, 2003) from PDB database (Berman et al., 2000). We selected those chains from different proteins solved by X-ray crystallography with a resolution of ≤ 2.0 Å and R -factor $\leq 19\%$. The sequence identity between two selected chains is less than 25%. As a result, 286 proteins were retained after removing 66 protein chains that do not have ConSurf-HSSP (Glaser et al., 2005) files out of 352 proteins with only one peptide chain.

2.2 Cross-validation

To validate our method, we chose a two-fold cross-validation test to conduct the related experiments. We divided the data set into two disjoint subsets for training and testing with each subset of approximately the same number of protein chains. The training and testing of each SVM were conducted twice by switching the training set and the test set. The outputs from SVM were used to analyse the performance of our method.

2.3 Encoding scheme for SVM

To encode a residue of interest, we use sequence profile obtained from HSSP database (Dodge and Schneider, 1998), where each residue is represented by 20 elements whose values are evaluated from multiple sequence alignment and their potential structural homologues. We then add evolutionary rate (1 element), which takes into account the phylogenetic relationships

between the homologues and the stochastic nature of the evolutionary process so that the conservation level for each residue can be inferred with Maximum Likelihood (ML) criterion. In practice, the phylogenetic relationships can be obtained by querying ConSurf-HSSP database (Glaser et al., 2005). Moreover, three elements (2 for helix/strand and 1 for other) for the predicted secondary structure were extracted from the PHD predictor (Rost and Sander, 1993), which assigns a type of secondary structure to the encoded residue. Finally, we calculate a new hydrophobic value in terms of AAIndex1 database (Kawashima et al., 1999) with two steps. First, we apply the Principal Component Analysis (PCA) technique (Jolliffe, 2002) on one constructed matrix, which is the selected set of all the hydrophobic properties of amino acids. Then, we consider the eigenvector as the hydrophobic profile in terms of the maximal eigenvalue. Thus, an element for new hydrophobic value is attached in the encoded input vector. Totally, there are 25 elements representing each residue.

To predict the contact map, a pair of residues i and j is used to represent whether they are in inter-residue contact or not. Two sliding windows centred at the residue pair are applied to encode the input data of SVM predictor. As a result, the training vector of the SVM predictor contains $25 \times 9 \times 2 = 450$ elements totally with the sliding window size of 9.

2.4 Normalisation scheme

It is necessary for us to normalise the input data of SVMs to equalise its range. A general method used in the paper is derived from Karplus and Schulz (1985). The normalised data y' can be obtained with the following simple equation:

$$y' = \frac{y - \mu}{\sigma} \quad (1)$$

where μ and σ denote the mean and the standard deviation of the original data y , respectively.

2.5 Contact map

In general, a contact map of a polypeptide chain of length N is represented by an $N \times N$ matrix S , which is defined in terms of distances between pairs of residues and a given distance cut-off d (usually taken as 8 Å between their C-alpha atoms) (Gromiha and Selvaraj, 2004):

$$S_{ij} = \begin{cases} 1 & \text{if } d(i, j) < d, |i - j| \geq 6 \\ 0 & \text{Otherwise} \end{cases}, \quad (2)$$

where $d(i, j)$ denotes the distance between residues i and j . In this work, a pair of residues is in inter-residue contact if the distance cut-off is less than 8 Å and the two residues of the pair are separated by at least six residues in sequence.

2.6 Evaluation measures for performance of predictors

Generally speaking, prediction accuracy, the ratio of the number of correctly predicted related clusters to the total number of predicted related clusters in experiment, is the best index for evaluating the performance of our predictors. However, only 20.4% of the data are related clusters, which lead to a rather unbalanced distribution of positive (related clusters) and negative (non-related clusters) samples. Using such data as training input would result in an SVM classifier classifying all related clusters as non-related clusters in a protein's contact map. To obtain a balanced training set, we used the related clusters and an equal number of randomly sampled non-related clusters.

To assess our method objectively, two indices, i.e., specificity and sensitivity (Baldi et al., 2000; Yan and Dobbs, 2003; Wang et al., 2006), are introduced in this paper.

Let TP be the number of correctly predicted related clusters and FP be the number of predicted related clusters that are in fact non-related clusters. In addition, let TN be the number of true negatives (non-related clusters) and FN the number of false negatives. Then, the evaluation measures can be computed as follows:

$$\begin{aligned}\text{Sensitivity} &= TP / (TP + FN) \\ \text{Specificity} &= TP / (TP + FP) \\ \text{Accuracy} &= (TP + TN) / (TP + FN + TN + FP) \\ CC &= (TP \times TN - FP \times FN) / \sqrt{(TP / FN)(TP + FP)(TN + FP)(TN + FN)}.\end{aligned}\quad (3)$$

The Correlation Coefficient (CC) is a measure of how well the predicted cluster labels correlate with the actual cluster labels.

3 Results

3.1 Contact cluster

It is well known that two residues in contact always gather together on one contact map and these contact groups are called contact clusters. These contact clusters correspond to the structural elements of the protein: α -helices, β -strand pairings and tertiary interactions of α -helices or β -sheets (Weikl, 2005). The contact clusters of a protein can be divided into local and non-local clusters. Local clusters contain at least one local contact (i, j) with small contact order $CO = |i - j| < 6$, whereas non-local clusters do not contain any such local contacts. In this work, we mainly investigate the correspondence between non-local contact clusters and pairs of residues in high hydrophobic regions. As discussed in Dosztányi et al. (1997), Gromiha and Selvaraj (2004) and Weikl (2005), two contacts (i, j) and (k, l) are put into the same cluster if they are close together on the contact map, according to the distance criterion $|i - k| + |j - l| \leq 4$, where i, j, k and l denote four amino acid residues. In this work, a cluster has to contain at least five contacts. And, we discard the peripheral contacts (i, j) , which have a minimum distance $|i - k| + |j - l| = 4$ from the other contacts in the clusters. For the contact cluster set $\aleph$ for a given protein chain, its element $\aleph_i$ corresponds to a residue pair (i, j) , whose members might be fallen into high hydrophobic cores H_n . Therefore, we have a one-to-multi mapping $\aleph \rightarrow H \cup \bar{H}$, where $\bar{H}$ denotes the subset of elements of $\aleph$ that have no mapping to the set H . Figure 1 illustrates the contact map and contact clusters for the PDB entry of 1md6 (Berman et al., 2000). The protein 1md6 has one peptide chain with 154 residues. It is observed that there exists a certain correspondence between the contact cluster set and the set of pairs of residues in high hydrophobic cores. For instance, the cluster No. 16 in Figure 1 is in correspondence with the pair of high hydrophobic cores (41, 57). Here, we use “related inter-residue contact cluster” (simply as ‘related cluster’) to represent a correspondence between a pair of residues in high hydrophobic regions and a contact cluster, and “non-related inter-residue contact cluster” (or “non-related cluster”) to represent no such correspondence otherwise.

3.2 SVM predictor

To construct the SVM predictor, we calculate a new hydrophobic scale extracted from AAIndex1 database, which contains 516 physical and chemical properties of amino acids (Kawashima et al., 1999). As a result, 25 hits are obtained by querying ‘hydrophobicity’ of interest in the AAIndex1 data set, which are shown in Table 1. To extract important features from the 25 hits (or properties in AAIndex1), we applied PCA (Jolliffe, 2002) on the selected

properties. Generally, the PCA technique can reduce the dimensionality of a given set of data, and produce a new set of principal components. The first principal component accounts for the maximum variation of the original data, and the second one accounts for the next highest variation and so on. Here, the first principal component accounts for 91.4% variation of the 25 properties and the representation of this component space is considered to be a new hydrophathy scale. The new scale is shown in Table 2 and is then used to the training vector of our SVM predictor.

To clearly illustrate the SVM predictor, we make a comparison between the number of contact clusters and the number of related clusters of inter-residue contacts for our protein data set. If a pair of residues in high hydrophobic regions is in correspondence with one clustering class of inter-residue contacts, the target value (for training input vector) of SVM predictor is set to 1; otherwise, the target is 0. As a result of the thorough statistical analysis of the correlation between pairs of residues in high hydrophobic regions and the clustering classes of inter-residues contacts over the data set, about 94.4% clustering classes are in accordance with the pairs of high hydrophobic regions. That is to say, about 94.4% contact clusters are mapping to *H* and the rest are mapping to *H*. The correspondence can be seen in Figure 2.

From Figure 2, it can be seen that the longer the sequence of a protein, the lesser the number of related clusters involved in high hydrophobic regions in correspondence with the number of natural contact clusters. Particularly, almost all proteins with more than 400 residues show non-correlation. In this case, the tendency may affect the accuracy of our contact prediction.

3.3 Performance of SVM

To predict the inter-residue contact centres, we propose a novel inter-residue contact prediction approach in which its input information is based on hydrophobic cores of proteins. In this approach, each training vector of SVM predictor contains two sliding windows of neighbouring residues in sequence. Their central residues *i* and *j* correlate with the correspondence between the residue pair and one high hydrophobic core, and the corresponding target value 1 or 0 denotes the residue pair (*i, j*) corresponding with a related or a non-related cluster.

After running SVM training process, the trained SVM predictor is used to test protein chains to locate their inter-residue contact centres. The performance of SVM is shown in Figure 3.

It is evident that on average for the 286 proteins in our data set, the accuracy can be up to 63.4%, the average specificity is 35% owing to the unbalanced training data set, the average sensitivity is 82.7%, and the average CC is 17%. It can be seen that the CCs for most protein chains are greater than zero except for only six protein chains. That is to say, our method works perfectly for almost all proteins. Since the number of negative training data is much more than positive samples, such high sensitivity denotes that most related clusters can be predicted.

3.4 Discussion

This paper addressed the problem of predicting inter-residue contacts. To simplify this complicated problem, we proposed a new method to solve the problem of inter-residue contacts instead. It can reduce the computational complexity dramatically. In particular, this approach also provides useful information in the contact map prediction and furthers the prediction of 3D structure of protein. In particular, this approach is based on the fact that native contacts are grouped into contact clusters in protein's contact map and at the same time pairs of residues in high hydrophobic regions may cover almost all these clusters. We designed our predictor in such a way that its input information is based on the pairs of residues in high hydrophobic regions.

Previous approaches always directly predicted the inter-residue contacts, whereas this paper provides a method that predicts the inter-residue contact cluster centres in contact maps of proteins with a higher accuracy than previous methods. Dosztányi et al. (1997) provided a method that achieved an accuracy of 65% using only sequence information and an accuracy of 68% using evolutionary information extracted from multiple sequence alignment. It should be noted that this approach is to identify residues in stabilisation centres along each protein peptide chain. However, our work is to predict residue–residue pairs in contact clusters. Since there are no previous similar approaches to predicting the centres, we make comparisons between our research and other inter-residue contacts approaches indirectly. For instance, for PROFcon method, about 30% of the predicted contacts are correct when considering the top $L/2$ predicted contacts between residues with sequence separation of six residues or more, where L denotes the number of residues in the protein chain (Punta and Rost, 2005). In our case, the number of observed inter-residue contact cluster centres is about L and 35% of the predicted contact clusters are correct ('specificity' in our paper). We cannot directly compare the performance of our method with that of others, but the main advantage of our approach is its simplicity and its higher accuracy.

Furthermore, as we consider contact cluster centres of a protein chain from a protein oligomer, the performance decreased slightly owing to the interactions between chains of a protein oligomer. However, even a chain with several domains may reduce the prediction performance.

To clearly illustrate the result of our approach, as shown in Figure 4, we chose the protein with PDB number 1md6 to compare the predicted and the natural inter-residue contact cluster centres. The protein 1md6 belongs to mainly beta class in CATH database (Pearl and Todd, 2005) because 13 out of the 17 secondary structures are formed into three beta sheets.

As we can see from the lower triangles of left side in Figure 4, the 16 out of the 17 inter-residue contact clusters are made from the 17 secondary structures and the remainder one is also near the pair of secondary structures (3, 5). It is observed based on Figure 4 that our predictor can correctly distinguish 11 clusters, which were marked as ellipses, from 17 natural contact clusters. This example seems to suggest that even though there are many redundantly or falsely predicted contact clusters and has a rather lower specificity of our predictor, our method may provide a simple and effective technique for solving the problem in the prediction of protein's structure.

4 Conclusions

This paper proposed a simple but effective approach to solve the problem of the prediction of inter-residue contacts. The main idea of our method is based on the contact clusters and the SVM. The prediction of the key inter-residue contact cluster centres is also important to study the protein structure compared to predict the contact map of a protein. Although our approach is simple and efficient, some aspects should be improved. For example, balancing the unbalanced training data may improve the performance of our approach. To decrease the number of negative samples, we should take advantage of other properties of amino acids and thus reduce the number of pairs of concerned residues corresponding to the natural inter-residue contact clusters without decreasing the coverage rate of those clusters. In conclusion, improving the performance of our approach will be our future research work.

Acknowledgments

This work was supported in part by grant 2 G12 RR003048 from the RCMI Program, Division of Research Infrastructure, National Center for Research Resources, NIH and Mordecai Wyatt Johnson grant, Howard University. This work was also supported in part by the National Science Foundation of China (No. 60803107).

References

- Baldi P, Brunak S, Chauvin Y, Andersen CAF, Nielsen H. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics* 2000;16:412–424. [PubMed: 10871264]
- Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The protein data bank. *Nucleic Acids Research* 2000;28:235–242. [PubMed: 10592235]
- Chen, P.; Wang, B.; Wong, HS.; Huang, DS. Prediction of long-range contacts from sequence profile. *International Joint Conference on Neural Networks*; Orlando, Florida, USA. 2007. p. 938–943.
- Chen P, Huang DS, Zhao XM, Li X. Predicting contact map using radial basis function neural network with conformational energy function. *International Journal of Bioinformatics Research and Applications* 2008a;4(2):123–136. [PubMed: 18490258]
- Chen P, Han K, Li X, Huang DS. Predicting key long-range interaction sites by B-factors. *Protein and Peptide Letters* 2008b;15(5):478–483. [PubMed: 18537737]
- Cheng JL, Baldi P. Improved residue contact prediction using support vector machines and a large feature set. *BMC Bioinformatics* 2007;8:113. [PubMed: 17407573]
- Dodge C, Schneider R. The HSSP database of protein structure-sequence alignments and family profiles. *Nucleic Acids Res* 1998;26:313–315. [PubMed: 9399862]
- Dosztányi Z, Fiser A, Silmon I. Stabilization centers in proteins: identification, characterization and predictions. *Journal of Molecular Biology* 1997;272(4):597–612. [PubMed: 9325115]
- Fariselli P, Olmea O, Valencia A, Casadio R. Prediction of contact maps with neural networks and correlated mutations. *Protein Engineering* 2001;14:835–843. [PubMed: 11742102]
- Glaser F, Rosenberg Y, Kessel A, Pupko T, Ben-Tal N. The ConSurf-HSSP database: the mapping of evolutionary conservation among homologs onto PDB structures. *PROTEINS: Structure, Function, and Bioinformatics* 2005;58:610–617.
- Gromiha MM, Selvaraj S. Inter-residue interactions in protein folding and stability. *Prog Biophys Mol Biol* 2004;86(2):235–277. [PubMed: 15288760]
- Gromiha MM, Selvaraj S. Protein secondary structure prediction in different structural classes. *Protein Eng* 1998;11:249–251. [PubMed: 9680186]
- Gupta N, Mangal N, Biswas S. Evolution and similarity evaluation of protein structures in contact map space. *PROTEINS: Structure, Function, and Bioinformatics* 2005;59:196–204.
- Heringa J, Argos P. Side-chain clusters in protein structures and their role in protein folding. *J Mol Biol* 1991;220:151–171. [PubMed: 2067014]
- Jolliffe, IT. *Principal Component Analysis*. Springer; Berlin, Germany: 2002.
- Karplus PA, Schulz GE. Prediction of chain flexibility in proteins. *Naturwissenschaften* 1985;72:212–213.
- Kawashima S, Ogata H, Kanehisa M. AAindex: amino acid index data-base. *Nucleic Acids Res* 1999;27:368–369. [PubMed: 9847231]
- Laskowski A, Macarthur MW, Moss DS, Thornton JM. PROCHECK: a program to check the stereochemical quality of protein structure. *J Appl Cryst* 1993;26:283–291.
- MacCallum RM. Striped sheets and protein contact prediction. *Bioinformatics* 2004;20(Suppl 1):224–231.
- Niggemann M, Steipe B. Exploring local and non-local interactions for protein stability by structural motif engineering. *J Mol Biol* 2000;296:181–195. [PubMed: 10656826]
- Noguchi T, Akiyama Y. PDB-REPRDB: a database of representative protein chains from the protein data bank (PDB) in 2003. *Nucleic Acids Res* 2003;31:492–493. [PubMed: 12520060]
- Pearl F, Todd A. The CATH domain structure database and related resources Gene3D and DHS provide comprehensive domain family information for genome analysis. *Nucleic Acids Res* 2005;33 (Database Issue):247–251.
- Pollastri G, Baldi P. Prediction of contact maps by GIOHMMs and recurrent neural networks using lateral propagation from all four cardinal corners. *Bioinformatics* 2002;18(Suppl 1):S62–S70. [PubMed: 12169532]
- Punta M, Rost B. PROFcon: novel prediction of long-range contacts. *BIOINFORMATICS* 2005;21(13):2960–2968. [PubMed: 15890748]

- Rost B, Sander C. Prediction of protein secondary structure at better than 70% accuracy. *J Mol Biol* 1993;232:584–599. [PubMed: 8345525]
- Selvaraj S, Gromiha MM. Role of hydrophobic clusters and long-range contact networks in the folding of $(\alpha/\beta)_8$ barrel proteins. *Biophys J* 2003;84:1919–1925. [PubMed: 12609894]
- Thomas DJ, Casari G, Sander C. The prediction of protein contacts from multiple sequence alignments. *Protein Engineering* 1996;9(11):941–948. [PubMed: 8961347]
- Vendruscolo M, Kussell E, Domany E. Recovery of protein structure from contact maps. *Fold Des* 1997;2(5):295–306. [PubMed: 9377713]
- Vicatos S, Reddy BVB, Kaznessis Y. Prediction of distant residue contacts with the use of evolutionary information. *PROTEINS: Structure, Function, and Bioinformatics* 2005;58:935–949.
- Vullo A, Walsh L, Pollastri G. A two-stage approach for improved prediction of residue contact maps. *BMC Bioinformatics* 2006;7:180. [PubMed: 16573808]
- Wang B, Chen P, Huang DS, Li J, Lok TM, Lyu MR. Predicting protein interaction sites from residue spatial sequence profile and evolution rate. *FEBS Lett* 2006;580(2):380–384. [PubMed: 16376878]
- Weikl TR. Loop-closure events during protein folding: Rationalizing the shape of phi-value distributions. *Proteins: Structure, Function, and Bioinformatics* 2005;60(4):701–711.
- Yan, C.; Dobbs, D. *Intelligent Systems Design and Applications*. Springer; Berlin, Germany: 2003. Identification of surface residues involved in protein-protein interaction – a support vector machine approach; p. 53-62.
- Zehfus MH. Automatic recognition of hydrophobic clusters and their correlation with protein folding units. *Protein Sci* 1995;4:1188–1202. [PubMed: 7549883]

Biographies

Peng Chen, PhD, is a Post-Doctoral Researcher in the Department of Systems and Computer Science at Howard University. He received his PhD from the University of Science and Technology of China in 2007. His research interests are in the areas of intelligent computing, pattern recognition and bioinformatics.

Chunmei Liu, PhD, is Assistant Professor of the Department of Systems and Computer Science at Howard University. She received her PhD in Computer Science from The University of Georgia. Her primary areas of scientific expertise include computational biology, algorithms and graph theory.

Legand Burge, PhD, is Associate Professor of the Department of Systems and Computer Science at Howard University. His research interests lie in the field of distributed computing. The primary thrust of his current research is in global resource management in large-scale distributed systems.

Mohammad Mahmood is Associate Professor of the Department of Mathematics at Howard University. His research interests include non-linear waves in optics, plasmas and fluids, water waves, etc.

William Southerland is Professor of Biochemistry at the College of Medicine of Howard University. He received his PhD from Duke University. His research is on molecular modelling, molecular dynamics and design of therapeutic agents.

Clay Gloster is Associate Professor of the Department of Electrical and Computer Engineering at Howard University. He received his PhD from North Carolina State University. His research interest includes reconfigurable and adaptive computing.

Figure 1.

Illustration of inter-residue contact clusters for chain A of protein 1md6.

The left upper side of the graph (a) shows the natural contact map at a distance cut-off of $d < 8 \text{ \AA}$; the right lower one displays the contact clusters. The graphs (b) and (c) show the hydrophobic profile of the protein chain in terms of the residues number in sequence, and the green diamonds denote the local maxima of hydrophobicity (see online version for colours)

(a)

(b)

Figure 2.

(a) The correlation between the number of contact clusters and the number of related clusters involving high hydrophobic regions; (b) The distribution of the correlation for each protein chain, where L in each subgraph denotes the sequence length of each protein. In (a), these six separations contain approximately equal number of proteins. Each bar sums up the number of contact clusters for each protein separated by sequence length. The blue bar denotes the entire clusters, and the red one denotes the number of related clusters corresponding to pairs of residues in high hydrophobic regions. In (b), circles above the black line indicate that the related

clusters are fully covered the natural contact clusters for protein chains, and circles far away from the line denotes that there are several related clusters correspond to each natural contact clusters (see online version for colours)

Figure 3.

Performances of the SVM predictor on 286 proteins. X-axis label for each subgraph denotes protein number. Moreover, the definition of Y-axis label can refer to equation (3) (see online version for colours)

Figure 4.

The comparison of the predicted inter-residue contact clusters (the upper triangle) and the natural contact clusters (the lower triangle). In the upper triangle of left side, blue diamond stands for the predicted results based on the SVM predictor. The ellipses in the upper triangle indicate the correctly predicted contact cluster. The right-side figures the secondary structures labelled from 1 to 17. Symbols like '1-16', i.e., pair of secondary structures (1, 16), illustrate which pair of secondary structures are predicted as inter-residue contact clusters (see online version for colours)

Table 1

Accession number of properties in AIndex1 by querying 'hydrophobicity' of interest

ARGP820101 [#]	CIDH920101	CIDH920102	CIDH920103	CIDH920104
CIDH920105	EISD840101	GOLD730101	JOND750101	JURD980101
MANP780101	PONP800101	PONP800102	PONP800103	PONP800104
PONP800105	PONP800106	PRAM900101	SWER830101	ZIMJ680101
PONP930101	WILM950101	WILM950102	WILM950103	WILM950104

[#]The accession number extracted from Release 8.0 of AIndex1 database possesses the property of hydrophobic scale of amino acids.

Table 2
The hydrophobic profile for the first principal component of 25 properties in AAindex1 database

AA	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
Scale	-31.95	-45.92	-7.22	-28.93	-33.56	7.11	10.05	-15.28	-12.61	15.42	8.9	4.24	22.59	20.21	20.59	18.91	17.83	14.37	12.08	3.19