Predicting contact map using Radial Basis Function Neural Network with Conformational Energy Function

Peng Chen

Intelligent Computing Lab,
Hefei Institute of Intelligent Machines,
Chinese Academy of Sciences,
Hefei, Anhui, 230031, China

Department of Automation,
University of Science and Technology of China,
Hefei, Anhui, 230026, China
E-mail: bigeagle@ mail.ustc.edu.cn

De-Shuang Huang*, Xing-Ming Zhao and Xueling Li

Intelligent Computing Lab,
Hefei Institute of Intelligent Machines,
Chinese Academy of Sciences,
Hefei, Anhui, 230031, China
E-mail: dshuang@iim.ac.cn
*Corresponding author

Abstract: Contact map, which is important to understand and reconstruct protein's three-dimensional (3D) structure, may be helpful to solve the protein's 3D structure. This paper presents a novel approach to predict the contact map using Radial Basis Function Neural Network (RBFNN) optimised by Conformational Energy Function (CEF) based on chemico-physical knowledge of amino acids. Finally, the results are trimmed by Short-Range Contact Function (SRCF). Consequently, it can be found that our proposed method is better than the existing methods such as PROFcon and the PE-based method. Particularly, this method can accurately predict 35% of contacts at a distance cutoff of 8 Å.

Keywords: contact map; Radial Basis Function Neural Network; RBFNN; Conformational Energy Function; CEF; Principal Component Analysis; PCA; Short-Range Contact Function; SRCF; bioinformatics.

Reference to this paper should be made as follows: Chen, P., Huang, D-S., Zhao, X-M. and Li, X. (2008) 'Predicting contact map using Radial Basis Function Neural Network with Conformational Energy Function', *Int. J. Bioinformatics Research and Applications*, Vol. 4, No. 2, pp.123–136.

Biographical notes: P. Chen received his MSc Degree of Control Theory and Control Engineering at KunMing University of Science and Technology (KMUST), KunMing, China. He is currently studying for his PhD in Automation Department from the University of Science and Technology of

China (USTC). His current research interests include intelligent computing theory, bioinformatics, and software.

D.S. Huang is a Professor at the Hefei Institute of Intelligent Machines, Chinese Academy of Sciences, China. He received his PhD in Electronic Engineering in 1993 from Xidian University, Xian, China. His current research is to focus on pattern recognition, biological and artificial neural networks, evolutionary computation and bioinformatics.

X.M. Zhao received his PhD in Automation Department from the University of Science and Technology of China (USTC). His current research interests include neural networks and bioinformatics.

Xueling Li is currently an Assistant Professor at Hefei Institute of Intelligent Machines, Chinese Academy of Sciences. She received her MSc in Zoology from East China Normal University, Shanghai, China and received her PhD in Microelectronics and Solid State Electronics from Shanghai Institute of Microsystem and Information Technology, Chinese Academy of Sciences, Shanghai, China. Her current research interests include bioinformatics and computational biology.

1 Introduction

Although bioinformatics has been achieving tremendous development since the end of last century, and more and more primary sequences of proteins and peptides were being sequenced, at the same time the number of solved three-dimensional (3D) structures of proteins is much smaller than the number of sequenced proteins. Because of the absence of efficient measure technologies and the resulting slow development on exploring protein's spatial structures, it really needs to find another efficient route to simulate and predict the 3D structures of proteins by computer. Among them, the prediction based on contact map is a relatively simple way to study the protein's 3D structure.

It is well known that contact map is an intermediate state from primary structure to tertiary structure, which may be advanced to the protein's tertiary structure and protein folding prediction. Usually, the non-local interactions (the contact separation is generally greater than six residues), especially the long-range interactions (the contact separation between a pair of residues is generally more than 24 residues), are crucial for proteins to attain their native state (Niggemann and Steipe, 2000). Specifically, identifying pairs of non-sequential amino acid residues that interact in 3D space can provide a set of topological constraints that can be utilised in protein folding recognition.

Previous work focused on contact map prediction using various methods, such as methods with the use of Neural Networks (NN) (Fariselli et al., 2001), General Input-Output Hidden Markov Models (GIOHMMs) (Pollastri and Baldi, 2002), hybrid method of RBFNN with Genetic Algorithm (GA) (Zhang and Huang, 2004), Self-Organizing Map (SOM) integrated by Genetic Programming (GP) (MacCallum, 2004), predictor with evolutionary information (Vicatos et al., 2005), and so on. Unfortunately, no satisfactory level in prediction accuracy of contact maps has been achieved over different proteins. PROFcon method declared that about 30% of the predicted contacts were correct (accuracy) at a prediction threshold at which about

10% of the observed contacts are predicted, considering all contacts between residue pairs that are separated by at least six residues (Punta and Rost, 2005). Although obtaining higher prediction accuracy, the PROFcon method achieves much lower coverage or sensitivity (only about 10%). Vullo's two-stage predictor obtained 19.8% accuracy for minimum contact separation of 23, when choosing the top length/five predicted contacts (Vullo et al., 2006). Despite achieving such tremendous success, the prediction on contact map can not be satisfied to reconstruct the protein's 3D structure. Therefore, all these methods and their derived tools can not perfectly resolve the contact problem. As a result, it is necessary for us to develop novel ideas and approaches to predict the contact map.

In this work, firstly, several principal components are extracted from known protein structures based on a special input-encoding scheme, and an initial matrix of contact map is constructed by the Principal Component Analysis (PCA) algorithm. Secondly, the initial contact matrix integrated with other information about residue pairs is applied to the RBFNN predictor. In particular, this method uses the Short-Range Contact Function (SRCF) combined with chemico-physical characteristics of amino acids to trim the output of RBFNN predictor. The SRCF is a 20-by-20 statistical matrix, where each element describes the occurrence frequency of each pair of residue types being in short-, medium-, or long-range contact through a large protein set.

2 Methods

2.1 Training and test sequences

We obtained our protein chain set from PDB-REPRDB (Noguchi and Akiyama, 2003), which is a database of protein chains from PDB based on PDB Rel. 2005,05,29, and updated on 10 June 2005. We selected those chains from different proteins that are resolved by X-ray crystallography with resolution ≤2.0 Å, and R-factor ≤19%. The sequence identity between two selected chains is less than 25%. As a result, we selected 480 protein chains that have ConSurf-HSSP files (Glaser et al., 2005). To validate our proposed approach, we split the selected protein chain set into two subsets; one set of 287 protein chains is used as the training set and another set containing 193 protein chains as the test set. The experiments run twice and the averaged outputs of the classifier can be obtained.

2.2 The definition of contact map

In this paper a widely used definition on contact is given, where it was assumed that the coordinates of alpha carbon atom are chosen to represent the spatial position of each residue in one protein. A contact matrix A for a protein with *num* residues is a $num \times num$ binary matrix whose element $A(i, j) = 1$ if the residues i and j are in contact, and $A(i, j) = 0$ otherwise. Then we shall have the following equation:

$$\begin{cases} A(i,j) = 1, & \text{if } d(i,j) < d \\ A(i,j) = 0, & \text{if } d(i,j) \geq d \end{cases} \tag{1}$$

where $d(i, j)$ denotes the distance between the ith and jth residues, d denotes a threshold value (always d = 4–8 Å) and d = 8 Å in our work.

2.3 Principal Component Analysis (PCA)

It has been shown that a matrix B, which is expanded by the parameters including several larger eigenvalues and their corresponding eigenvectors of a contact matrix A, can approximate A perfectly. The PCA algorithm can be used for reducing the dimensions of contact matrix and extracting useful information from known proteins (Huang, 1996a, 1996b; Krzanowski, 1988).

2.4 Radial Basis Function Neural Network (RBFNN)

In general, multilayer perceptrons exhibit improved generalisation properties, especially for regions that are not represented sufficiently in the training set. In contrast, multiplayer perceptrons learn slower than their RBF counterparts. Because of simple topological structure and rapid training procedure, the RBFNNs provide a powerful technique for generating multivariate nonlinear mapping, and they work best when many training vectors are available (Huang, 1996a, 1999a, 1999b, 1999c; Theodoridis and Koutroumbas, 2003; Wang and Zhu, 2000). Currently, the RBFNNs have been successfully adopted to solve these multivariate models of nonlinear classification problem, such as the prediction in protein's inter-residue contacts map.

Generally, an RBFNN consists of three layers:

- Input layer, where all the input vectors of the training and test samples are input
- Hidden layer, where at each node the input feature vector is put through a Radial Basis Function (RBF) that is centred on a corresponding exemplar vector
- Output layer, where all the inputs from the hidden layer are combined to indicate the class of input sample.

The weights between the hidden and the output layer are adjusted during the training process for the purpose of minimising the cost or goal function. The whole architecture of the RBFNN is therefore fixed by determining the number of the hidden layer neurons and the weights between the hidden and output layers. For an input vector, $X = [x_1, x_2, \ldots, x_n]$, $T \in R_n$, and with K hidden layer neurons, the activation function,

$$f(x) = \exp(-(x - c_i)^T (x - c_i)) / 2\sigma^2, \tag{2}$$

is described by a centre, $C_i \in R_n$, and a width parameter, σ_i, where $i = 1, 2, \ldots, K$. The $exp(*)$ denotes the Gaussian radial basis function. Now, the output of an output neuron j is given by

$$g_j(X) = \sum_{i=1}^{k} w_{i,j} f_i(X) + b_j, \tag{3}$$

where $w_{ij} \in R$ is the weight between the hidden neuron i and the output neuron j ($j = 1, \ldots, J$), b_j is a possible bias to help to approximate the target value.

If the centres are not preselected, they have to be estimated during the training phase along with the weights W and the variances σ^2 (if the latter are also considered unknown).

Let L be the number of input-desired output training pairs, ($x(j)$, $y(j)$, $j = 1, \ldots, L$). We select an appropriate cost function of the output error as:

$$J = \sum_{j=1}^{L} \phi(e(j)) \tag{4}$$

where $\phi(*)$ is a differentiable error function (e.g., the square of its argument) as:

$$e(j) = y(j) - g(x(j)). \tag{5}$$

For an input-output training pair (x_l, y_l), each actual output from the jth node, g_j^l, is forced to match them by adjusting the weights w_j, the centres c_j, and the variances σ_j^2. So estimating the three parameters becomes a typical task of a nonlinear optimisation process. For example, if we adopt the gradient descent approach, the following algorithm will be resulted:

$$\begin{cases} w_i(t) = w_i(t-1) - \mu_1 \partial J / \partial w_i|_t, & i = 0, 1, \ldots, k \\ c_i(t) = c_i(t-1) - \mu_2 \partial J / \partial c_i|_t, & i = 0, 1, \ldots, k \\ \sigma_i(t) = \sigma_i(t-1) - \mu_3 \partial J / \partial \sigma_i|_t, & i = 0, 1, \ldots, k \end{cases} \tag{6}$$

where t is the current iteration step.

In this paper, the RBFNNs are used to predict the contact map, which is guaranteed to converge to a Bayesian classifier provided that it is given enough training data. The used training data are divided into two classes, i.e., the contact and non-contact class, respectively. Subsequently, the input is encoded and then the target classes can be obtained by applying the input to the RBFNN predictor.

2.5 Short-Range Contact Function (SRCF)

The SRCF based on statistical estimation can be used to estimate the interaction frequencies of all residue pairs from the observed structures in Protein Database Bank. In our work, we consider two residues to be in short-range (or medium-range or long-range) contact if their sequence separation is greater than 6 (or 12 or 24) and their spatial distance is farther than 8 Å.

Firstly, we construct a 20-by-20 matrix, A, to count the number of each residue type pairs that are in short-range contact (or medium-range or long-range contact). Secondly, another 20-by-20 matrix, B, is formed to count the number of each residue type pairs in our protein set. Finally, our adopted long-range contact function (C_i, j) can be defined as:

$$C_{i,j} = A_{i,j}/B_{i,j} \tag{7}$$

Currently, there have existed several contact functions or energies (Miyazawa and Jernigan, 1996, 1999; Gromiha and Selvaraj, 2004), but our designed statistical function is very simple and sufficiently applicable to trim the output from the RBFNN predictor.

2.6 Conformational Energy Function (CEF)

Another contact conformational energy based on statistical estimate can be added to optimise the RBFNN predictor. Our efficient conformational energy ($e_{i,j}$) can be defined as:

$$e_{i,j} = -\ln\left(\sum_m \frac{1}{N_m}\frac{N_m N_{i,j}}{N_i N_j} + \frac{h_i + h_j}{2} \times p_{i,j}\right) \tag{8}$$

where N_m is the total number of residues in protein m, $N_{i,j}$ is the number of inter-residue contacts, N_i and N_j are the numbers of residues of each type in each separate protein, h_i, h_j denote the values of hydrophobic measurements of amino acids, and the weight value, $p_{i,j}$ is changed with different proteins and is simply set to 1 in our paper.

Compared equation (8) with other definitions such as Miyazawa and Jernigan (1996, 1999), additional information about hydrophobic measurements of amino acids is used to approximately approach the natural state of macromolecules.

2.7 Encoding scheme for the RBFNN

We first encode the input vectors of the RBFNN predictor for each residue site based on a sliding window of residues centred at the current site. Each sliding window consists of a continuous set of residues with an odd size of *win*, where *win* is set to 13 in our work.

Firstly, we adopt the residue sequence profile obtained from the HSSP database (Dodge et al., 1998). Each residue is represented by 20 attributes whose values are determined from multiple sequence alignment and their potential structural homologs. Secondly, the evolutionary rate is characterised by taking into account the phylogenetic relationships between the homologs and the stochastic nature of the evolutionary process, so that the conservation level for each residue can be inferred by using the Maximum Likelihood (ML) criterion (Glaser et al., 2005). Each evolutionary rate score can then be appended to the first 20 sequence profile elements for each residue. Thirdly, each hydrophobic value can be appended to the former elements of training vector for each residue. In general, hydrophobicity plays a very important role in characterising the physico-chemical property in different areas of chemistry, medicine, and pharmacology (Kyte and Doolittle, 1982). Finally, the input vector thereby contains $22 \times 13 = 286$ elements totally.

2.8 Evaluation measures for performance of predictors

To verify our method, two measurements have been introduced to evaluate the performance of the predictors. Here, we applied the criteria of accuracy (*Acc*) and coverage (*Cov*), which are adopted at CASP/CAFASP (Fischer et al., 2003; Punta and Rost, 2005). They are defined as follows:

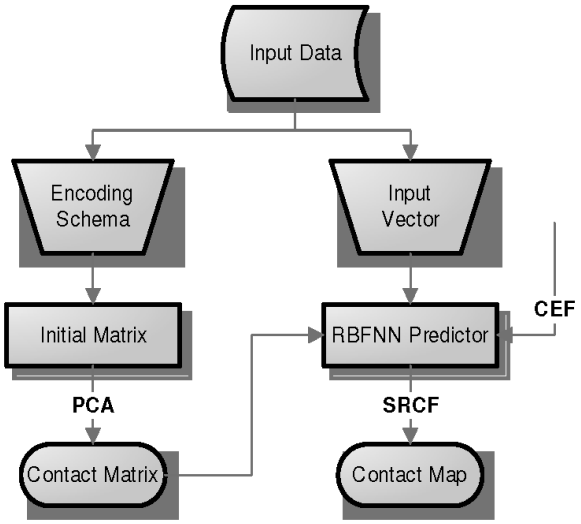
$$Acc = \frac{TP}{TP + FP}, \quad Cov = \frac{TP}{TP + FN} \tag{9}$$

where *TP* denotes the number of true positives, i.e., the two residues are assigned to long-range contact class and actually they are; *FP* denotes the number of false positives, i.e., the two residues are assigned to long-range contact class but in fact they are not in contact; and *FN* stands for the number of false negative, i.e., the two residues are assigned not to long-range contact class but in fact they are.

3 Results and discussion

To predict the contact map, we use an RBFNN predictor optimised by Conformational Energy Function (CEF) to address this problem. The flowchart of our method is described in Figure 1 based on the methods described above, the experimental results will be introduced in this section.

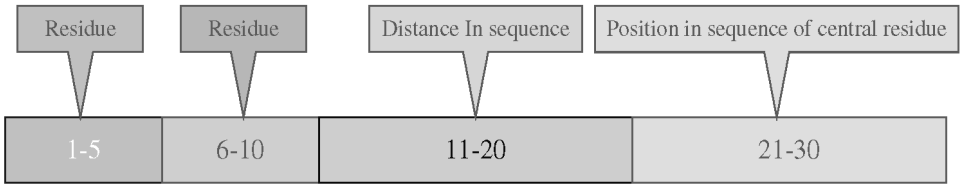
Figure 1 Flowchart of the RBFNN predictor optimised by Conformational Energy Function



3.1 The initial contact matrix and the corresponding contact map

It is well known that an input-encoding schema is thought to be a simplified, accelerated, and feasible means for achieving the initial contact matrix. To construct the initial contact map, a series of parameters are given for performing this task that includes interresidues distances in protein sequence, and the position in sequence of the central residue (the $int((i+j)/2)$ residue for residue pair i and j). Afterward, these useful parameters are to extract some information of contact map and thus can approximate the actual contact map of one protein. Finally, the initial contact matrix integrated with the RBFNN predictor can probably improve the prediction performance. The input-encoding schema for the initial contact matrix is depicted in Figure 2.

Figure 2 The encoding schema for the initial contact matrix



Our work starts from the initial matrix, and then the inclination of the contact distribution can be obtained. After that, the results are obtained by applying the helpful information to the RBFNN predictor. Firstly, the contact matrix A making use of PCA rule is defined to calculate the eigenvalue, λ_i, $i = 0, 1, \ldots, num - 1$, and its corresponding eigenvector. From these eigenvalues, we can choose four largest ones and discard the remaining ones generally which can obtain clear separation area of contact map. The selected eigenvalues are shown in Table 1.

Table 1 The selected eigenvalues

λ_0	λ_1	λ_2	λ_3
11.404	10.406	9.786	9.211

Now the contact probability, $p_{i,j}$, of residues i and j is defined as:

$$p_{i,j} = \frac{1}{N}\sum_m \sigma_m \lambda_m \times (v_i^m (v_j^m)') \times \cos\,(i-j)/(i+j)) \tag{10}$$

where λ_m is the eigenvalue, v_i^m is the corresponding eigenvector, σ_m denotes the coefficient of eigenvalue λ_m that can be changed with the length of protein sequence, where σ_m is set to 1. Particularly, we postulate $a_{i,j} = 1$, where $a_{i,j}$ is the element of contact matrix for protein 1NXB, if $p_{i,j} \geq 0.012$ by calculating equation (10), and $a_{i,j} = 0$ otherwise.

Using equation (10) with the training proteins based on the PCA rule, the initial contact map can be achieved for unknown structure protein 1NXB, where we choose 8 Å as the contact threshold of inter-residue contacts.

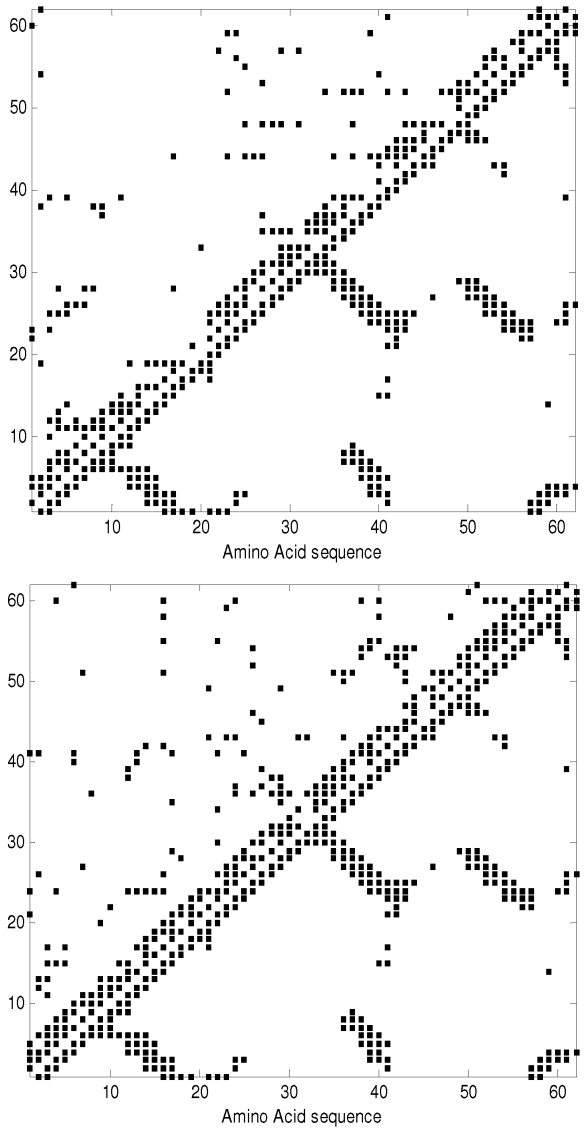
Aimed at an unknown protein 1NXB, the useful information about the length L_{test} and the amino acids in this protein sequence is extracted for comparing with other training proteins. As described above, a contact matrix A may represent the contact map for a training protein. Based on the matrix A, a coordinate mapping can link the training proteins with test protein 1NXB. For a training protein that has the sequence with the length of L_{test}, a fitting method can be applied to map the $L_{train} \times L_{train}$ data to $L_{test} \times L_{test}$ data. As a simplified denotation, v_i^m is postulated as the eigenvector according to eigenvalue, λ_i^m, in training protein A^m. As a result, the vector, v_i^m, is fitted to a vector with the length of L_{test} based on some fitting method. Moreover, the method of least squares is chosen to perform the task with fitting data between two proteins. Subsequently, a new matrix B can be constructed by vectors v_i^m and the corresponding fitting eigenvalues $\lambda_i^m, i = 1 \sim L_{test}$, to represent the protein 1NXB.

$$B = \sum_{m=1}^{N}\sum_{i=1}^{L_{training}} \eta^m \lambda_i^m v_i^m \tag{11}$$

where N is the number of training proteins, η^m is the coefficient of eigenvalue λ_i^m.

Integrating equations (10) and (11), the predicted contact map based on the PCA and the least square fitting can be achieved, and the results are shown in the left side of Figure 3 that describes the two contrasting sub-maps, the actual contact map (lower triangle) and the predicted contact map (upper triangle). In the figure, the changing tendencies of inter-residue contacts are predicted distinctly. In addition, this result also describes the contact area including main chain and non-local contact area where the inter-residue contact probability can be optimised to a maximum. So, it can be inferred that this PCA method is very useful for predicting the contact map with respect to a random predictor that produced initial contact map.

Figure 3 The initial contact map of Protein 1NXB based on PCA method (left) and the prediction of contact map based on RBFNN for Protein 1NXB (right)



3.2 The RBFNN predictor optimised by CEF

In our work, we choose the hydrophobic values using Eisenberg standard (Eisenberg et al., 1984) parameters. From equation (8), the effective conformational energy $e_{i,j}$ can be calculated and then applied to predict the contact map. By making use of the conformational energy, the optimal prediction on contact map may be achieved more quickly.

Integrating conformational energy with the entropy of the inter-residue contact probability, the RBFNN transfer function can be optimised. The likelihood goal function can be defined as follows:

$$Q(i,j) = -\sum_{n=1}^{num} p_{i,j} \log p_{i,j} + \lambda_{i,j}(e_{i,j} - s_{i,j}) \tag{12}$$

where $p_{i,j}$ is derived from equation (10), $\lambda_{i,j}$ is the parameter of Lagrange, and $s_{i,j}$ is the constraint condition from equation (8). Minimising this function with respect to $p_{i,j}$, the optimised contact probability can be achieved with running the RBFNN predictor over the all residues of training protein.

In this work, the RBFNN predictor can adopt the complex association between the sequence and the structure of protein to get the connection weights of NN. The outputs from RBFNN predictor are classified into two classes, i.e., contact or noncontact. Using a general transfer function for RBFNN predictor with the likelihood goal function equation (12), the contact map of unknown protein can be achieved conveniently. The experimental results for protein 1NXB are plotted in the right figure of Figure 3 that describes the two contrasting sub-maps, the actual contact map (lower triangle) and the predicted contact map (upper triangle). The experimental results are obtained on the condition of the threshold value of contact inter-residue, where is selected here as 8 Å.

3.3 Results from RBFNN predictor trimmed by SRCF

In the paper, we calculate the long- and SRCF in terms of equation (7). In Table 2, the upper triangle denotes the long-range contact (sequence separation greater than 24 residues) matrix and the lower triangle stands for the short-range contact matrix. Particularly, the short-range contact matrix is used in this paper to trim the RBFNN output while the long-range contact matrix will be investigated in our future work for improving the long-range contact prediction. Demonstration in Table 2 shows the results that the larger the contact value is, the more possible the residue pair is in short- or long-range contact; otherwise, the less possible. Additionally, the black underlined number in Table 2 denotes the larger contact frequency, which means that its residue pair is more possible in short- or long-range contact. On the other hand, the red number stands for smaller contact frequency, which means that less possible contact occurs to the residue pair.

Table 2 Long- (upper triangle) and short-range (lower triangle) contact matrixes for trimming the output from RBFNN predictor

	Ala	*Cys*	*Asp*	*Glu*	*Phe*	*Gly*	*His*	*Ile*	*Lys*	*Leu*	*Met*	*Asn*	*Pro*	*Gln*	*Arg*	*Ser*	*Thr*	*Val*	*Trp*	*Tyr*
Ala	1.86	*2.01*	0.95	1.01	0.83	1.29	1.82	0.85	1.94	0.92	*2.11*	0.95	0.99	1.21	0.77	1	1.33	1.41	*2.24*	0.8	1.29
Cys	0.97	*2.66*		*2.19*	0.52	**0.38**	0.67	0.88	**0.46**	0.76	**0.39**	0.93	0.52	**0.49**	0.65	**0.43**	0.56	0.75	0.69	0.9	**0.43**	0.59
Asp	0.88	0.65	0.58		0.62	**0.42**	0.61	1.28	0.74	0.83	0.84	0.82	0.5	0.81	0.78	0.51	0.95	1.03	0.95	0.91	**0.36**	0.65
Glu	0.74	0.52	**0.43**	**0.3**		**0.28**	0.62	0.85	0.53	0.74	0.69	0.75	**0.35**	0.64	0.64	**0.35**	0.69	0.7	0.72	0.88	**0.35**	0.61
Phe	1.41	0.81	0.67	0.65	1.26		1.16	1.19	0.69	1.36	0.61	1.44	0.76	0.72	0.88	0.53	0.75	0.94	0.96	1.54	0.61	0.89
Gly	1.46	0.81	1.04	0.8	1.23	1.68		*2.18*	1.03	1.41	1.02	1.59	0.92	1.37	1.37	0.97	1.14	1.53	1.5	1.67	0.79	1.23
His	0.74	0.55	0.63	**0.43**	0.61	0.71	**0.33**		0.63	0.76	**0.35**	0.8	**0.42**	0.62	0.66	0.39	0.5	0.79	0.86	0.94	**0.43**	0.71
Ile	1.8	0.84	0.75	0.88	1.46	1.21	0.68	1.84		*2.18*	0.75	1.97	0.92	0.82	0.92	0.64	0.92	1.09	1.38	*2.37*	0.76	1.3
Lys	0.9	0.58	0.75	0.74	0.72	0.9	**0.41**	0.82	**0.4**		**0.41**	0.84	**0.41**	0.62	0.59	**0.45**	**0.41**	0.75	0.73	0.98	0.33	0.71
Leu	*2.09*	0.95	0.86	0.92	1.62	1.42	0.82	1.96	0.9	*2.29*		*2.21*	0.97	0.87	1.09	0.73	1.01	1.09	1.26	*2.36*	0.75	1.26
Met	0.87	**0.45**	**0.4**	**0.37**	0.73	0.73	**0.49**	0.85	**0.48**	0.99	**0.48**		0.62	**0.46**	0.61	**0.41**	0.56	0.54	0.68	1.13	**0.4**	0.75
Asn	0.94	0.63	0.7	0.5	0.72	1.12	**0.45**	0.81	0.65	0.83	**0.46**	0.75		1.01	0.8	0.62	0.65	1	1.01	0.97	**0.48**	0.74
Pro	1.05	0.69	0.67	0.61	0.79	1.15	0.61	0.92	0.54	1.04	0.52	0.75	0.8		0.98	0.69	0.69	0.96	0.97	1.18	0.6	0.96
Gln	0.68	**0.38**	**0.48**	0.42	0.58	0.8	**0.36**	0.78	0.5	0.83	**0.43**	0.58	0.55	**0.4**		**0.39**	0.5	0.7	0.64	0.87	**0.44**	0.61
Arg	0.88	0.64	0.78	0.71	0.85	0.91	0.52	0.93	**0.49**	1.18	**0.45**	0.72	0.76	0.55	0.66		0.55	0.83	0.82	1.1	0.52	0.76
Ser	1.18	0.78	0.79	0.71	0.96	1.31	0.66	1.18	0.72	1.28	0.67	0.85	0.87	0.68	0.9	1.12		1.11	1.13	1.27	0.57	0.8
Thr	1.34	0.73	0.87	0.86	1.18	1.3	0.66	1.38	0.8	1.47	0.59	0.84	0.91	0.78	0.92	1.12	1.46		1.23	1.6	0.55	0.91
Val	*2.29*	1.18	0.92	1.02	1.72	1.48	0.8	*2.39*	1.11	*2.52*	1.02	0.93	1.18	0.95	1.19	1.3	1.7	*2.95*		*3.18*	0.86	1.43
Trp	0.91	**0.39**	**0.46**	**0.45**	0.78	0.77	0.5	0.76	**0.47**	1.05	**0.47**	**0.45**	0.7	**0.45**	0.51	0.69	0.71	1.02	0.62		**0.38**	**0.6**
Tyr	1.4	0.83	0.79	0.73	1.26	1.29	0.69	1.48	0.91	1.61	0.67	0.79	1.03	0.69	0.88	1.09	1.23	1.63	0.69	1.27		0.92
	Ala	*Cys*	*Asp*	*Glu*	*Phe*	*Gly*	*His*	*Ile*	*Lys*	*Leu*	*Met*	*Asn*	*Pro*	*Gln*	*Arg*	*Ser*	*Thr*	*Val*	*Trp*	*Tyr*		

3.4 RBFNN predictor performance

Based on the above experiments, the accuracies in predicting contact map at distance cutoff from 5 Å to 12 Å are shown in Table 3. For performance at a distance cutoff of 8 Å and considering the top $L/2$ predicted contacts (where L is the sequence length of the protein), our RBFNN predictor achieves 35.3% average accuracy and 11.5% coverage. Furthermore, no SRCF to trim the output of the RBFNN predictor may worsen the prediction accuracy up to 2.2% and therefore lose 1.5% coverage performance. Moreover, integrating the CEF into the likelihood goal function of RBFNN predictor can improve the prediction accuracy up to 4.7% and thus increase the prediction coverage up to 3.3%.

Table 3 Performance vs. protein length

Length	*Distance cutoff*				*Number*
	5 Å	*8* Å	*10* Å	*12* Å	
<100	30.3%* (20.2%)#	46.5% (18.4%)	68.6% (9.1%)	83% (5.8%)	20
100–200	26.5% (7.4%)	34.2% (12.2%)	60.6% (7.5%)	84% (5.3%)	107
200–300	23.3% (9.2%)	30.5% (8.2%)	65.4% (6.6%)	81.1% (5.2%)	66
Average	25.8% (9.3%)	35.3% (11.5%)	67.1% (7.36%)	83.9% (5.32%)	
No SRCF	25% (9.1%)	33.1% (10%)	64.1% (7.29%)	81.5% (5.25%)	193
No CEF	26.2% (9.9%)	30.6% (8.21%)	60.1% (7%)	81% (5.2%))	

*The number indicates the accuracy performance for several distance cutoffs and different protein length ranges.

#The bracket number denotes the coverage for several distance cutoffs and different protein length ranges.

Compared with other methods, our predictor performs better than the PROFcon method (about 30%) for distance separation greater than six residues (Punta and Rost, 2005), and on the other hand, performance of our method is almost similar to the PE-based contact map predictor (Vullo et al., 2006), which achieved about 36.6% accuracy for distance separation greater than five residues. To compare with the PE-based predictor, we perform our RBFNN predictor on the dataset the PE-based predictor collected. Finally, our predictor achieves 37.3% accuracy for distance separation greater than five residues (Performance calculation of our RBFNN predictor in the work is based on the distance separation greater than six residues).

4 Conclusions

From the above experimental results, it can be found that the generalisation capability of the RBFNN predictor integrated with CEF was better than the previous method. Moreover, our approach was only based on the primary sequence of unknown protein.

It was shown that the construction of the initial contact matrix can be efficiently used as a part of likelihood goal function for training our RBFNN predictor to predict the contact map, which may be used to reconstruct the 3D structure of protein. All these results suggest a number of further investigations:

- RBFNN predictor may be a useful and simple approach to predict the contact map of protein. Mining the RBFNN algorithm can improve the prediction performance of contact map.
- PCA can extract some key information at helping to improve the performance of our RBFNN predictor.
- Understanding the short-range contact preference of residue pair may be in favour of trimming the output from our RBFNN predictor. Because some residue pairs are seldom in contact but some ones are often contacting in 3D space, we can remove the former residue pairs and retain the latter ones in a reasonable possibility.
- In this work, about 35% short-range contacts are correctly predicted when considering the top $L/2$ predicted contacts and the residue pair with six apart.

In summary, the better performance can be achieved apparently. Even for protein with several domains, the sequence context seems to play an important role as well. On all accounts, our approach proposed in this paper is a promising method for predicting the contact map.

Acknowledgements

This work was supported by the grants of the National Science Foundation of China, Nos. 30570368 & 30700161, the grant from the National Basic Research Program of China (973 Program), No. 2007CB311002, the grants from the National High Technology Research and Development Program of China (863 Program), Nos. 2007AA01Z167 & 2006AA02Z309, the grant of the Guide Project of Innovative Base of Chinese Academy of Sciences (CAS), No.KSCX1-YW-R-30, and the grant of Oversea Outstanding Scholars Fund of CAS, No.2005-1-18.

References

Dodge, C., Schneider, R. and Sander, C. (1998) 'The HSSP database of protein structure sequence alignments and family profiles', *Nucleic Acids Res.*, Vol. 26, pp.313–315.

Eisenberg, D., Schwarz, E., Komaromy, M. and Wall, R. (1984) 'Analysis of membrane and surface protein sequences with the hydrophobic moment plot', *J. Mol. Biol.*, Vol. 179, No. 1, pp.125–142.

Fariselli, P., Olmea, O., Valencia, A. and Casadio, R. (2001) 'Prediction of contact maps with neural networks and correlated mutations', *Protein Engineering*, Vol. 14, pp.835–843.

Fischer, D., Rychlewski, L., Dunbrack Jr., R.L., Ortiz, A.R. and Elofsson, A. (2003) 'CAFASP3: the third critical assessmentof fully automated structure prediction methods', *Proteins: Structure, Function, and Genetics*, Vol. 53, pp.503–516.

Glaser, F., Rosenberg, Y., Kessel, A.T.P. and Ben-Tal, N. (2005) 'The ConSurf-HSSP database: the mapping of evolutionary conservation among homologs onto PDB structures', *PROTEINS: Structure, Function, and Bioinformatics*, Vol. 58, pp.610–617.

Gromiha, M.M. and Selvaraj, S. (2004) 'Inter-residue interactions in protein folding and stability', *Progress in Biophysics and Molecular Biology*, Vol. 86, pp.235–277.

Huang, D.S. (1996a) *Systematic Theory of Neural Networks for Pattern Recognition*, Publishing House of Electronic Industry of China, Beijing.

Huang, D.S. (1996b) 'Generalization capabilities in feedforward neural networks for pattern recognition', *Journal of Beijing Institute of Technology*, Vol. 5, No. 2, pp.184–192.

Huang, D.S. (1999a) 'Application of generalized radial basis function networks to recognition of radar targets', *International Journal of Pattern Recognition and Artificial Intelligence*, Vol. 13, No. 6, pp.945–962.

Huang, D.S. (1999b) 'The bottleneck behaviour in linear feed-forward neural network classifiers and their breakthrough', *Journal of Computer Science and Technology*, Vol. 14, No. 1, pp.34–43.

Huang, D.S. (1999c) 'Radial basis probabilistic neural networks: model and application', *International Journal of Pattern Recognition and Artificial Intelligence*, Vol. 13, No. 7, pp.1083–1101.

Krzanowski, W.J. (1988) *Principles of Multivariate Analysis*, Oxford University Press, New York.

Kyte, J. and Doolittle, R.F. (1982) 'A simple method for displaying the hydrophobic character of a protein', *J. Mol. Biol.*, Vol. 157, pp.105–132.

MacCallum, R.M. (2004) 'Striped sheets and protein contact prediction', *Bioinformatics*, Vol. 20, Suppl. 1, pp.224–231.

Miyazawa, S. and Jernigan, R.L. (1999) 'An empirical energy potential with a reference state for protein folding and sequence recognition', *Proteins*, Vol. 36, pp.357–369.

Miyazawa, S. and Jernigan, R.L. (1996) 'Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term for simulation and threading', *J. Mol. Biol*., Vol. 256, pp.623–644.

Niggemann, M. and Steipe, B. (2000) 'Exploring local and non-local interactions for protein stability by structural motif engineering', *J. Mol. Biol.*, Vol. 296, pp.181–195.

Noguchi, T. and Akiyama, Y. (2003) 'PDB-REPRDB: a database of representative protein chains from the Protein Data Bank (PDB) in 2003', *Nucleic Acids Res.*, Vol. 31, pp.492, 493.

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

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

Theodoridis, S. and Koutroumbas, K. (2003) *Pattern Recognition*, 2nd ed., Elsevier Science Press, New York, USA.

Vicatos, S., Reddy, B.V.B. and Kaznessis, Y. (2005) 'Prediction of distant residue contacts with the use of evolutionary information', *PROTEINS: Structure, Function, and Bioinformatics*, Vol. 58, pp.935–949.

Vullo, A., Walsh, L. and Pollastri, G. (2006) 'A two-stage approach for improved prediction of residue contact maps', *BMC Bioinformatics*, Vol. 7, Article 180.

Wang, Z.O. and Zhu, T. (2000) 'An efficient learning algorithm for improving generalization performance of radial basis function neural networks', *Neural Networks*, Vol. 13, Nos. 4–5, pp.545–553.

Zhang, G.Z. and Huang, D.S. (2004) 'Prediction of inter-residue contacts map based on genetic algorithm optimized radial basis function neural network and binary input encoding scheme', *J. Computer-Aided Molecular Design*, Vol. 18, pp.791–810.