BIOINFORMATICS ORIGINAL PAPER

Vol. 25 no. 2 2009, pages 204–210
doi:10.1093/bioinformatics/btn618

Structural bioinformatics

Identification of structurally conserved residues of proteins in absence of structural homologs using neural network ensemble

Ganesan Pugalenthi[1], Ke Tang[2], P. N. Suganthan[1,*] and Saikat Chakrabarti[3,*]

[1]School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798, [2]Department of Computer Science and Technology, University of Science and Technology of China, Hefei, Anhui, China and [3]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA

Received on August 22, 2008; revised and accepted on November 25, 2008

Advance Access publication November 27, 2008

Associate Editor: Anna Tramontano

ABSTRACT

Motivation: So far various bioinformatics and machine learning techniques applied for identification of sequence and functionally conserved residues in proteins. Although few computational methods are available for the prediction of structurally conserved residues from protein structure, almost all methods require homologous structural information and structure-based alignments, which still prove to be a bottleneck in protein structure comparison studies. In this work, we developed a neural network approach for identification of structurally important residues from a single protein structure without using homologous structural information and structural alignment.

Results: A neural network ensemble (NNE) method that utilizes negative correlation learning (NCL) approach was developed for identification of structurally conserved residues (SCRs) in proteins using features that represent amino acid conservation and composition, physico-chemical properties and structural properties. The NCL-NNE method was applied to 6042 SCRs that have been extracted from 496 protein domains. This method obtained high prediction sensitivity (92.8%) and quality (Matthew's correlation coefficient is 0.852) in identification of SCRs. Further benchmarking using 60 protein domains containing 1657 SCRs that were not part of the training and testing datasets shows that the NCL-NNE can correctly predict SCRs with ∼90% sensitivity. These results suggest the usefulness of NCL-NNE for facilitating the identification of SCRs utilizing information derived from a single protein structure. Therefore, this method could be extremely effective in large-scale benchmarking studies where reliable structural homologs and alignments are limited.

Availability: The executable for the NCL-NNE algorithm is available at http://www3.ntu.edu.sg/home/EPNSugan/index_files/SCR.htm

Contact: epnsugan@ntu.edu.sg; chakraba@ncbi.nlm.nih.gov.

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

1 INTRODUCTION

The overall fold is very important to maintain a suitable framework for protein function. It is well established that the structure of proteins is determined by their amino acid sequences (Anfinsen, 1973). Although there is an exponential increase in the available protein structures, the number of protein folds is still very limited in nature (Chothia, 1992). In other words, many evolutionary and functionally related proteins with well-diverged sequences still keep the same folding pattern. This suggests that the protein folding pattern depends not only on the whole sequence but also on some small segments of residues that are conserved during evolution in both sequence and structural aspects. These residues, which might have important implications in maintenance of protein folds, are termed as SCR.

Sequence-based motifs and conserved residues are useful in understanding the conservational variation and have been successfully linked to functionally important sites indicating higher selection pressure on them (Neuwald *et al*., 1995; Saqi and Sternberg, 1994). However, SCRs identified at 3D structure level provide more meaningful information towards understanding the structure–function relationship of proteins (Paiardini *et al*., 2005; Peters *et al*., 2006; Shapiro and Brutlag, 2004). In our earlier works (Chakrabarti and Sowdhamini, 2003; Chakrabarti *et al*., 2003, 2006; Pugalenthi *et al*., 2007), we identified structurally invariant segments at superfamily level where proteins are distantly related but retain similar fold and biological functions. These structural motifs were recognized on the basis of both sequence conservation and preservation of important structural properties, such as solvent accessibility, secondary structural content, hydrogen-bonding pattern and residue compactness. They are also found to maintain a similar spatial orientation pattern, when compared across different proteins belonging to the same family or superfamily. Therefore, these SCRs might be crucial for the formation of common structural core that provides optimal environment for the protein to perform its molecular or biological function. SCRs might also provide important clues as sequence–structural signature of multiple folding units for each protein fold, and therefore can be extremely useful in protein engineering and design experiments.

Identification of SCRs is a difficult and challenging task as it requires careful examination of 3D structural homologs and development of reliable structural alignments. Additionally, many protein structures are reported to have limited or no structural homologs. For example, 566 out of 1194 superfamilies in PASS2 database have only single structural entry (Bhaduri *et al*., 2004).

*To whom correspondence should be addressed.

© 2008 The Author(s)
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.0/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Therefore, identification of SCRs from protein structure that has limited or no homologous structural information is very important and poses a challenging task.

In this study, we propose a neural network ensemble (NNE) method that utilizes negative correlation learning (NCL) for classification and prediction of SCRs. As the availability of 3D structures and structural alignments are still limited in protein comparison studies, the NCL-NNE prediction approach provides a useful option that can successfully predict important structural residues utilizing a single protein structure.

2 METHODS

2.1 The dataset

The dataset used for training and testing our algorithm was obtained from MegaMotifBase database (Pugalenthi *et al*., 2008a), which contains protein structural motifs for structurally aligned protein domains related at the superfamily level. These structural motifs were identified by screening the superfamily alignment (structural alignment) positions for conservation of important structural properties, such as solvent accessibility, secondary structural content, hydrogen-bonding pattern and residue compactness. In addition to the structural motif definition for the superfamily, this database also provides structural motif information for each individual structure by consulting the structural alignments. Thus, SCRs for individual domain can be extracted from the structural motif definitions provided by the MegaMotifBase database (Pugalenthi *et al*., 2008a).

In this study, we used 191 superfamilies for classification. Out of 191, 131 superfamilies belonging to 14 All-α, 25 All-β, 47 α/β, 37 α + β, 4 small domains, 3 multidomain protein and 1 membrane/cell surface protein classes were selected for training and testing. From 131 superfamilies, 496 domains were chosen for training and testing. From the remaining 60 superfamilies that do not overlap with training and testing datasets, 60 protein domains were used for benchmarking study. Each protein domain sequence in our dataset has <40% sequence identity to any other sequences in the training, testing and benchmarking datasets (Supplementary Material 1 and 2).

We used 6042 SCRs (positive dataset) and 105 204 non-SCRs (negative dataset) that were obtained from the selected 496 domains for training and testing. To avoid imbalance between positive and negative (residues) datasets, we randomly selected 3021 SCRs from 6042 positive samples and 3021 non-SCRs from the negative samples for training. In the same way, the test data were constructed from the remaining 3021 positive samples and 3021 residues randomly chosen from the remaining negative samples (Supplementary Material 3 and 4). In addition, 1657 SCRs were obtained from 60 protein domains for benchmarking.

2.2 Feature set

Each residue in the SCR dataset is represented by 212 features that include sequence and structural information extracted from the homologous alignment and 3D structure (Supplementary Material 5). Homologous sequences for each protein domain were obtained using five rounds of PSI-BLAST (Altschul *et al*., 1997) against NCBI non-redundant protein database, with an *E*-value cutoff of 0.001. CLUSTALW (Thompson *et al*., 1994) was used to align the homologous sequences. Sequences having <80% of the length of the query structure were removed from the alignment. Secondary structural information was assigned for all sequence homologs in the alignment using PSIPRED (McGuffin *et al*., 2000). The details of the features used in this study are briefly mentioned below.

Conservation score: sequence conservation score for each alignment position was evaluated consulting a standard 20×20 substitution matrix (Johnson and Overington, 1993).

Amino acid type and functional groups: we categorized 20 amino acids into 10 functional groups based on the presence of side chain chemical group, such as phenyl (F/W/Y), carboxyl (D/E), imidazole (H), primary amine (K), guanidino (R), thiol (C), sulfur (M), amido (Q/N), hydroxyl (S/T) and non-polar (A/G/I/L/V/P). The compositional diversities of each SCR were evaluated by calculating the frequency of 20 amino acids and 10 functional groups within the SCR alignment positions.

Structural features: structural features, such as solvent accessibility, secondary structures, hydrogen bonds and residue compactness were computed from the individual protein structure using the JOY package (Mizuguchi *et al*., 1998).

Physico-chemical properties: matrices containing quantitative values for amino acids' physico-chemical properties scaled between 0 and 1 were obtained from the UMBC AAIndex database (Kawashima *et al*., 1999). The selected physico-chemical properties include molecular weight, hydrophobicity, hydrophilicity, hydration potential, refractivity, average accessible surface area, free energy transfer, flexibility, residue volume, mutability, melting point, optical activity, side chain volume, polarity and isoelectric points.

Sequence and structural features from spatial neighbors: spatially neighboring residues were shown to have positive influence in identification of critical sites in proteins (Pugalenthi *et al*., 2008b). The residues whose C^{β} atoms were found within 5 Å distance from the C^{β} of a SCR were considered as spatial neighbors of the SCR. In case of glycine, a virtual C^{β} atom was considered. Content of amino acid type and functional groups, structural features and physico-chemical property values were computed from all spatial neighbors for each SCR.

2.3 Classification protocol

The classification model presented in this article was built through three steps. First, all the input features of training data were normalized to be between -1 and 1 by a linear function. Then, a NNE was trained by a method called NCL (Liu and Yao, 1997, 1999a, b; Yao *et al*., 2001). Finally, 'feature selection' was conducted to investigate whether we can still achieve good prediction performance with only a subset of features.

NCL approach is widely used for training the NN ensembles. Below, we briefly describe the basic ideas and steps of the NCL and readers are requested to refer the original publications (Liu and Yao, 1997, 1999a, b; Yao *et al*., 2001) for full details.

Suppose that we have a training set of size N, denoted by

$$D = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)\}$$

where $\mathbf{x} \in \mathbf{R}^d$ is the d-dimensional training samples and y is the corresponding class labels. NCL is designed to train a NN ensemble of the form:

$$F(n) = \frac{1}{M} \sum_{i=1}^{M} F_i(n) \tag{1}$$

where M is the number of the individual NNs in the ensemble. $F_i(n)$ is the output of the i-th NN on the n-th training sample and $F(n)$ is the output of the ensemble on the n-th training sample.

NCL employs the standard back-propagation algorithm to train the individual NNs in parallel. The key to the success of NCL is the use of the error function. NCL uses the sum of the mean squared error (MSE) and a penalty term as the error function during the learning process. When the n-th training sample is presented, the i-th NN is trained to minimize the error function:

$$E_i(n) = \frac{1}{2}(F_i(n) - y_n)^2 + \lambda p_i(n) \tag{2}$$

where λ is a positive parameter controlling the tradeoff between the MSE (accuracy) and y is the class label and the penalty term (diversity) can be calculated by:

$$p_i(n) = (F_i(n) - F(n)) \sum_{j \neq i} (F_j(n) - F(n)) \tag{3}$$

It can be seen from Equation (3) that the penalty term explicitly encourages the i-th NN to be negatively correlated with the remaining NNs in the ensemble. By this means, diversity among the individual NNs is achieved. It can also be seen that with $\lambda = 0$ we would have an ensemble exactly equivalent to training a set of NNs independently of one another. When λ is increased, more and more emphasis would be placed on seeking the negative correlation.

The NN ensemble used in this work has five NNs. Each individual NN is a feed forward network with one hidden layer. The number of hidden neurons is set to five for all individual NNs. When employing NCL to train the NN ensemble λ is set to 1 and the number of learning epochs is set to 100.

2.4 Feature selection

Since the number of features in this study is high, we conducted feature selection to decrease the size of the features by omitting the non-effective features. We designed a wrapper approach to conduct feature selection for our dataset. A feature selection method typically consists of two main components: a selection criterion and a search scheme. The selection criterion measures the usefulness of any feature subset, and feature selection seeks the feature subset that optimizes the selection criterion. The search scheme determines how to search for the optimal feature subset among all possible combinations of features. In this work, the Matthew's Correlation Coefficient (MCC) defined in Equation (7) was used as the selection criterion, and we adopted a sequential backward elimination (or recursive feature elimination) (Webb, 2002) search scheme in this study. The feature selection procedure is described briefly in the following.

We trained the NN ensemble using the whole feature set (i.e. the original training dataset). After that, the trained NN ensemble was preserved and the MCC was calculated. Then, starting from the whole feature set, features were iteratively pruned. For each individual NN, we removed the input neurons (and the weights associated to them) that correspond to the omitted features, while keeping all the other structure of the NN unchanged. The output of the NN ensemble was obtained using Equation (1) and the MCC was calculated based on it. At each iteration, the feature whose omission led to the largest MCC was pruned. The feature selection procedure terminates when a predefined number of features have been pruned.

2.5 Performance measures

Four different parameters have been used to measure the performance of the prediction method. These four parameters can be derived from the four scalar values: TP (true positives: number of correctly classified SCR), TN (true negatives: number of correctly classified non-SCR), FP (false positives: number of non-SCR incorrectly classified as SCR) and FN (false negatives: number of SCR incorrectly classified as non-SCR). Using the following formulas, we calculated sensitivity, specificity, positive prediction value (PPV) and MCC.

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP+FN}} \tag{4}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN+FP}} \tag{5}$$

$$\text{PPV} = \frac{\text{TP}}{\text{TP+FP}} \tag{6}$$

$$\text{MCC} = \frac{\text{TP}*\text{TN} - \text{FP}*\text{FN}}{\sqrt{(\text{TP+FN})(\text{TP+FP})(\text{TN+FP})(\text{TN+FN})}} \tag{7}$$

3 RESULTS

3.1 Distribution of SCRs

We collected 6042 SCRs from 131 protein superfamilies enlisted in MegaMotifBase database (Pugalenthi *et al.*, 2008a). Generally SCRs maintain a basal level of sequence conservation ($\geq$30% sequence identity), but there could be more conserved residues in proteins than the SCRs. As shown in Table 1, there are totally 54 958 residues from 496 domains that fall within the residue conservation range of 30–100%. Out of 54 958 residues, only 6042 are SCRs. Therefore, only 11% of sequentially conserved residues account for the SCRs. This observation suggests that although the SCRs are conserved in the sequence, it is difficult to specifically identify the SCRs just by looking at the residue conservation score.

Table 1. Residue conservation between SCR and non-SCR residues

Residue conservation (%)	No. of residues	No. of SCR
30–40	31648	2518
41–50	15848	1976
51–60	5226	1122
61–70	1620	254
71–80	276	88
81–90	238	58
91–100	102	26
Total	54958	6042

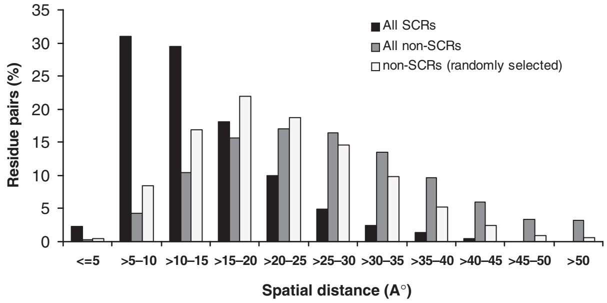

Fig. 1. Distribution of spatial distances between pairs of SCRs. Spatial distance between two SCRs was calculated utilizing the C^{β}–C^{β} atom coordinates supplied in the individual PDB (Berman *et al.*, 2000) file.

Further, we examined the spatial distances between all pairs of 6042 SCRs calculated from each protein domain (Fig. 1). Figure 1 provides the distance distribution of SCRs along with distances (C^{β}–C^{β} distances) calculated from all non-SCR pair as well as randomly selected non-SCR pairs (numbers equal to the SCR pairs). From this distance distribution, it can be seen that higher number SCR pair distances fall in lower distance bins (<20 Å) compared with that of non-SCRs. Therefore, it is reasonable to state that SCRs prefer a probable requirement of spatial proximity.

3.2 Prediction of SCR

We employed the NCL-NNE for classification and subsequent prediction of SCRs in proteins. The NCL-NNE was trained using the training dataset containing 3021 SCRs (positive samples) and 3021 non-SCRs (negative samples), while the performance of the classifier was tested on the testing dataset containing the remaining 3021 SCRs and 3021 randomly selected negative samples. Our NCL-NNE method achieved 92.8% sensitivities with MCC score of 0.852 in the testing data using all the 212 features that represents the compositional and conservational properties of SCRs (Table 2).

We applied a feature reduction protocol utilizing seven feature subsets to eliminate the redundant features. As seen in Table 2,

Table 2. Classification results achieved for the testing data using different feature subsets

No. of features	Sensitivity (%)	Specificity (%)	PPV (%)	MCC
5	92.59	85.95	86.83	0.787
8	90.47	91.30	91.23	0.818
10	92.19	91.63	91.68	0.836
50	91.03	93.52	93.35	0.846
100	91.72	92.73	92.66	0.845
150	92.06	92.57	92.53	0.846
200	91.00	93.20	93.05	0.842
212	92.82	92.50	92.52	0.852

Table 3. Classification results achieved for the training data using 5-fold cross-validation on different feature subsets

No. of features	Sensitivity (%)	Specificity(%)	PPV (%)	MCC
5	93.81 (1.60)	79.84 (3.37)	82.70 (2.26)	0.747 (0.014)
8	92.09 (1.49)	85.44 (2.77)	86.68 (2.12)	0.780 (0.012)
10	94.34 (1.10)	87.62 (2.24)	88.62 (1.77)	0.823 (0.012)
50	94.17 (0.95)	90.10 (0.99)	90.54 (0.79)	0.844 (0.006)
100	93.41 (1.36)	90.60 (1.23)	90.94 (0.98)	0.842 (0.008)
150	92.68 (0.51)	91.99 (0.70)	92.07 (0.62)	0.847 (0.006)
200	92.68 (0.33)	92.02 (0.41)	92.08 (0.36)	0.847 (0.004)
212	93.03 (0.10)	91.48 (0.13)	91.61 (0.12)	0.845 (0.002)

Statistical errors (standard error) associated with the average sensitivity, specificity, PPV and MCC are provided within the parenthesis.

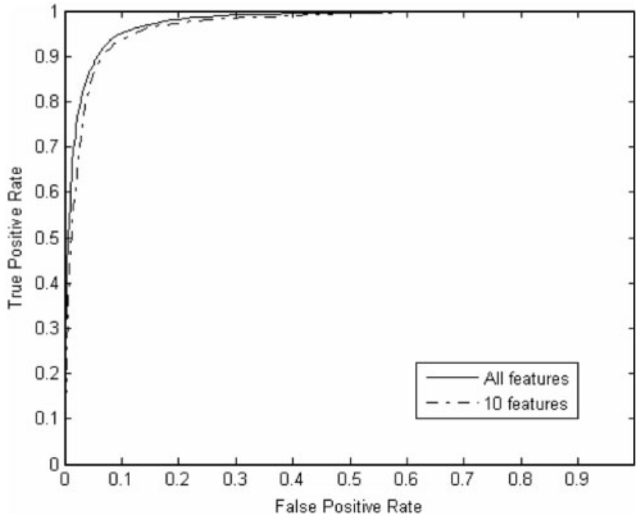

Fig. 2. ROC curves**.** ROC curves were plotted utilizing the fractions of TP and FP values derived using top 10 features and all features.

feature selection (reduction) generally does not deteriorate the classification performance much until the number of features decreases below 10. Before that, the usage of smaller number of features only leads to a very little decrease in the sensitivity and specificity rates. We also investigated the influence of the feature reduction by plotting receiver operating characteristic (ROC) curves (Fig. 2) derived from the sensitivity (TP rate) and specificity (FP rate) values for the classifiers using all the features and the 10 best performing features, respectively. Figure 2 shows that the classifiers built with the 10 features and the whole feature set perform comparably. Such observation is also supported by the similar values (0.9682 for 10 features and 0.9762 for all features) of the area under curve (AUC) obtained from the ROC curves.

Although the trained NCL-NNE shows good performance on the testing data, it is natural to ask whether the performance of NCL-NNE depends on any specific split of training and testing data. To verify this issue, we also conducted 5-fold cross-validation procedures on the training data. For each feature subsets presented in the Table 2, the corresponding performance measures achieved in the 5-fold cross-validation procedure are also provided in Table 3. If NCL-NNE exhibits significantly different performance in the cross-validation and testing procedure, then its performance highly depends on the training data. If NCL-NNE shows similar performance in the two scenarios, we may expect the NN ensemble trained with it generalizes well to unseen data. It can be observed from Tables 2 and 3 that NCL-NNE performed more or less similar in the two scenarios. Therefore, we can conclude that NCL-NNE is not very sensitive to different training data, and thus our final NN ensemble generalizes well.

In order to check whether the high accuracy is due to the NCL-NNE classifier or the quality of the selected features, we applied a linear model on our datasets. We obtained good prediction rate using all features (sensitivity 83.45% and specificity 83.28%) and 10 features (sensitivity 90.90% and specificity 89.47%). This result shows that the quality of the best performing features selected by our ‘feature selection’ approach play an important role in successful classification. However, NCL-NNE reported higher sensitivity and specificity rates than the linear model signifying its importance for better performance.

3.3 Influence of structural features and spatial neighbors

Table 4 shows the list of 10 best performing features. Eight out of the 10 best performing features that were automatically selected by the classifier involve features that represent structural properties, such as solvent accessibility, secondary structures, hydrogen bonding and residue compactness for a given SCR and its spatial neighbors. This finding emphasizes that important structural properties retrieved from a single protein can be successfully used in machine learning classification for identification of sites that are conserved for such properties across similar protein structures. Our finding also indicates that the environment of neighboring residues of the SCRs can be an important factor towards better classification and identification of SCRs.

As shown in Figure 3a, we found that significantly higher number of SCRs prefer aliphatic, hydrophobic residues (73.06%, 76.61%, 86.58% and 80.75% of SCRs has at least one alanine, isoleucine, leucine and valine, respectively, as their structural neighbor) surrounding themselves compared with that of non-SCRs. Similarly, lower fractions of SCRs prefer charged residues (aspartic acid: 37.32%, glutamic acid: 35.50%, histidine: 24.20% and arginine: 35.37%) as their structural neighbor. Figure 3b compares the fraction of each amino acid within the spatial neighbors of SCRs and non-SCRs where fraction of each amino acid is normalized by the overall background frequency of that particular residue. Higher preference

Table 4. List of best performing features

Feature	SCR related	SCR neighbors related	Structural feature	Sequence feature
Helix content in SCR	Yes	No	Yes	No
Strand content in SCR	Yes	No	Yes	No
Coil content in the SCR	Yes	No	Yes	No
Helix content in the spatial neighbor	No	Yes	Yes	No
Solvent accessibility in SCR	Yes	No	Yes	No
Hydrogen bonding information in SCR	Yes	No	Yes	No
Residue compactness in SCR	Yes	No	Yes	No
Residue compactness in the spatial neighbor	No	Yes	Yes	No
Leucine content in spatial neighbor	No	Yes	Yes	Yes
Cysteine content in SCR	Yes	No	No	Yes

Table 5. Evaluation of performance of different feature groups

No. of features	Sensitivity using 5-fold CV (%)	Sensitivity without CV (%)	Specificity using 5-fold CV (%)	Specificity without CV (%)
Group 1	33.79 (10.03)	30.22	89.57 (2.79)	93.77
Group 2	11.35 (3.46)	7.55	97.85 (1.06)	98.83
Group 3	41.90 (9.90)	41.31	90.41 (4.79)	85.48
Group 4	97.91 (0.26)	97.38	79.11 (1.87)	79.13
Group 1 + 2	13.01 (3.66)	10.76	98.94 (0.28)	99.07
Group 1 + 3	44.15 (8.53)	32.17	93.12 (2.43)	96.12
Group 1 + 4	96.62 (0.36)	96.46	85.57 (1.14)	85.31
Group 2 + 3	26.45 (6.88)	12.45	96.79 (1.07)	99.17
Group 2 + 4	95.13 (0.61)	94.97	88.25 (1.62)	88.80
Group 3 + 4	97.85 (0.26)	97.25	79.48 (2.19)	80.47
Group 1 + 2 + 3	22.28 (4.36)	14.86	98.54 (0.51)	99.22
Group 1 + 2 + 4	92.85 (0.22)	93.11	91.89 (0.78)	92.18
Group 1 + 3 + 4	96.29 (0.63)	96.19	85.67 (1.28)	86.28
Group 2 + 3 + 4	95.10 (0.65)	94.74	88.28 (1.60)	89.73
All Groups	93.03 (0.10)	92.82	91.48 (0.13)	92.50

Statistical errors (standard error) associated with the average sensitivity are provided within the parenthesis. CV, cross-validation.

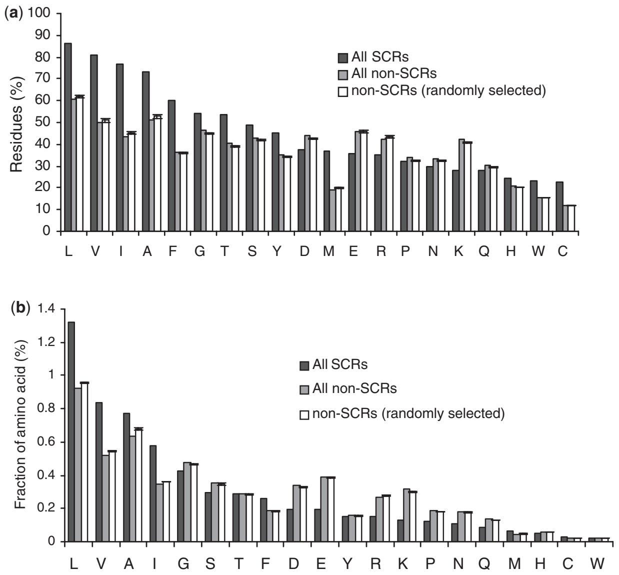

Fig. 3. Distribution of 20 amino acid type within the spatial neighbors of SCRs. (**a**) Shows the percentage of residues having at least one of the 20 amino acids within their spatial neighbor whereas (**b**) provides the fraction of each amino acid within the spatial neighbor. White bars with standard error from five trials provide data obtained from randomly selected non-SCRs.

of aliphatic, hydrophobic residues as spatial neighbors for the SCRs is also observed in Figure 3b.

We also trained our NCL-NNE on different feature subsets that were formed by grouping the qualitatively similar features together. We categorized all the 212 features into four different groups. Group 1 and group 2 contain features that represent the amino acids' composition and conservation for a given SCR and its spatial neighbors, respectively. Group 3 contains physico-chemical property features computed for a given SCR and its spatial neighbors. Similarly, group 4 represents the structural property features for a given SCR and its spatial neighbors. We test the performance of the classifier utilizing these feature groups separately as well as mixing them in all possible combinations. The performance of the NCL-NNE utilizing various combinations of feature groups is summarized in Table 5. Table 5 presents the results obtained via conducting 5-fold cross-validation on the training set and the results obtained on the testing set. As can be observed, NCL-NNE again performed similar in the two cases. Among the feature groups, group 4 alone performs quite well, in fact achieved the highest sensitivity (97.38%). This result further ascertains the importance of structural properties of the SCRs and its neighbors, and thereby supports their (structural property features) selection as best performing ones by the automated feature reduction protocol used in this study. However, the specificity of the prediction is compromised when only group 4 features were used. It can also be noticed that combination of structural property features (group 4) together with features belonging to other three groups significantly improves the specificity (92.50%), while marginally decreasing the sensitivity (92.82%) value.

3.4 Benchmarking studies

To test the capability, we applied the NCL-NNE to 60 protein domains obtained from 60 superfamilies for the prediction of SCRs. These 60 domains contain 1657 SCRs that do not overlap with the training and test datasets. The NCL-NNE prediction module correctly predicts 1497 SCRs (out of 1657 SCRs) with 90.3% sensitivity and 89.2% specificity. Further, the performance of our approach was compared with recently reported CUSP algorithm (Sandhya *et al*., 2008) that utilizes protein's structural homologs and structural alignments to distinguish structurally conserved regions. CUSP method correctly predicts 1485 SCRs with 89.6% sensitivity (Supplementary Material 6). This suggests that the result obtained

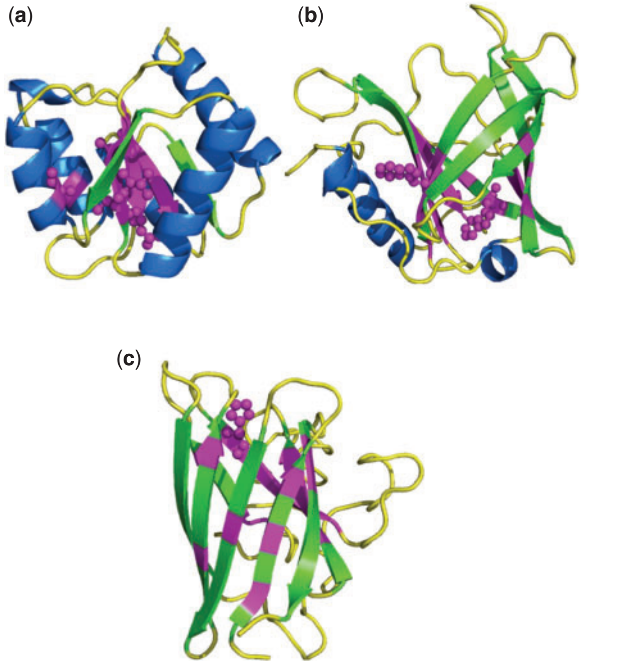

Fig. 4. Example of successful prediction of SCRs. SCRs predicted by NCL-NNE are shown in purple. Predicted SCRs that are experimentally verified are shown in ball and stick model. (**a**) Wild-type CheY from *Escherichia coli* (PDB code: 3CHY); (**b**) serum RBP (PDB code: 1JYD) and **c**) Cu–Zn superoxide dismutase (SOD) (PDB code: 2SOD). Regular secondary structures are colored in blue (helix), green (strand) and yellow (loops).

from our method is very similar to CUSP result obtained from high quality structural alignments. Importantly, our method achieves high sensitivity rate in absence of homologous structural information and structural alignments.

In order to show the structural and functional importance of SCR, we applied NCL-NNE for the prediction of SCRs from three protein structures (Fig 4). Figure 4a shows three-dimensional structure of wild-type CheY from *Escherichia coli* (PDB code: 3chy) (Lopez-Hernandez and Serrano, 1996). Our approach predicts 15 residues as SCRs (F8, L9, V10, V11, A42, F53, V54, I55, S56, D57, L68, V83, L84, M85 and V86). Out of 15 predicted SCRs, five residues V10, V11, A42, V54 and V57 (shown in purple ball and stick model in Fig. 4a) were reported in the previous studies as a part of folding nuclei, which play important role in folding of the protein (Mirny and Shakhnovich, 2001).

The structural importance of SCRs can be further explained by serum retinol binding protein (RBP), a member of the lipocalin family (PDB code: 1JYD) (Fig. 4b). Our algorithm predicts 27 SCRs (W24, A26, K29, A43, E44, F45, M53, A55, G75, H104, W105, I106, V107, T109, Y114, A115, V116, Q117, Y118, S119, C120, Y133, S134, F135, V136, F137 and S138). This structure has four conserved tryptophans (W24, W67, W91 and W105) and W24 and W105 were predicted as SCRs by NCL-NNE. Greene *et al*. (2001) conducted conservative substitutions for the four tryptophans and observed that substitutions at W67 and W91 positions do not affect the overall structural integrity. Substitution of W105, which is largely buried in the wall of the β-barrel, has minor effect on the structure. Further they reported that mutation at W24 position leads to large losses in stability and lower yields of native protein generated by *in vitro* folding.

Table 6. Execution time for NCL-NNE method

Protein PDB code	Chain identifier	Length	Execution time (in s)
1BY5	A	698	74
1EZ0	A	504	57
1CPT	–	412	43
1EZF	A	323	39
1A7T	A	227	31
1DOI	–	128	25
2HPQ	P	79	21

Though the SCRs are generally associated with structural stability, some of them might have functional role or provide optimal environment for the protein to perform its function. For example, H41 plays both catalytic and structural role in Cu–Zn superoxide dismutase (SOD) (PDB code: 2SOD; Fig. 4c). Twenty-one SCRs are predicted (A4, C6, L8, I18, V29, I33, H41, G42, F43, H44, V45, H46, D81, L82, V85, T114, M115, V116 and V117) for the Cu–Zn SOD by NCL-NNE. Previous analysis by Toyama *et al*. (2004) suggest that H41 involves in hydrogen bonding with T37 and H118 and this H41-mediated hydrogen bonds (T37-H41-H118) play crucial role in keeping the protein structure suitable for its efficient catalytic reactions.

3.5 Execution time

The execution time for our algorithm is reasonably faster. The procedure involves formulation of the features and prediction of SCRs using NCL-NNE model. In order to provide a flavor of the computation time for NCL-NNE, we randomly selected seven proteins with varying lengths and measured the user CPU time (Table 6) spent for the feature generation followed by prediction on a Pentium4 machine having 3 GHz CPU and 2 GB memory.

4 CONCLUSION

SCRs are crucial for the overall protein fold and can play important role in maintaining the suitable scaffold for the function of a protein. Identification of SCRs from single structure is a challenging task. Here, we implemented a NNE method that utilizes NCL approach for prediction of SCRs using features that represent the amino acid conservation and composition, physico-chemical properties and structural properties, such as solvent accessibility, secondary structures, hydrogen bonding and residue compactness. Validation of the NCL-NNE on the test dataset provided high sensitivity and quality of prediction (sensitivity: 92.8%, MCC: 0.852). Additional large-scale benchmarking using alignments of separate 60 protein domains shows 90.3% prediction sensitivity for the NCL-NNE. We also found that utilization of the structural features derived from the SCRs and their spatial neighbors are beneficial for successful classification and prediction. Our NCL-NNE prediction approach utilizes information derived from a single protein structure and its sequence homologs. Therefore, this method could be extremely useful for identification of SCRs in large-scale benchmarking studies where structural homologs and reliable structural alignments are still limited.

ACKNOWLEDGEMENTS

We thank Dr R. Sowdhamini and Sandhya Sankaran for providing CUSP results.

Funding: A*Star (Agency for Science, Technology and Research to G.P. and P.N.S.); National Natural Science Foundation of China grant (No. 60802036 to K.T.); Intramural Research Program of the National Library of Medicine at National Institutes of Health/DHHS (to S.C.).

Conflict of Interest: none declared.

REFERENCES

Altschul,S.F. *et al*. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*., 25, 3389–3402.

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

Berman,H.M. *et al*. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.

Bhaduri,A. *et al*. (2004) PASS2: an automated database of protein alignments organised as structural superfamilies. *BMC Bioinformatics.* **5**, 35.

Chakrabarti,S. and Sowdhamini,R. (2003) Regions of minimal structural variation among members of protein domain superfamilies: application to remote homology detection and modeling using distant relationships. *FEBS Lett.*, **569**, 31–36.

Chakrabarti,S. *et al*. (2003) SMoS: a database of structural motifs of superfamily. *Protein Eng*., **16**, 791–793.

Chakrabarti,S. *et al*. (2006) SSToSS - sequence-structural templates of single-member superfamilies. *In Sillico Biol*., **6**, 0029.

Chothia,C. (1992) Proteins. One thousand families for the molecular biologist. *Nature,* **357,** 543–544.

Greene,L.H. *et al*. (2001) Role of conserved residues in structure and stability: Tryptophans of human serum retinol-binding protein, a model for the lipocalin superfamily. *Protein Sci*., **10,** 2301–2316.

Johnson,M.S. and Overington,J.P. (1993) A structural basis for sequence comparisons. An evaluation of scoring methodologies. *J. Mol. Biol.*, **233**,716–738.

Kawashima,S. *et al*. (1999) AAindex: amino acid index database. *Nucleic Acids Res*., **27**, 368–369.

Liu,Y. and Yao,X. (1997) Negatively correlated neural networks can produce best ensembles. *Aust. J. Intell. Inf. Process. Syst.*, **4**, 176–185.

Liu,Y. and Yao,X. (1999a) Ensemble learning via negative correlation. *Neural Netw.*, **12**, 1399–1404.

Liu,Y. and Yao,X. (1999b) Simultaneous training of negatively correlated neural networks in an ensemble. *IEEE Trans. Syst. Man Cybern. B Cybern.*, **29**, 716–725.

Lopez-Hernandez,E. and Serrano,L. (1996) Structure of the transition state for folding of the 129 aa protein chey resembles that of a smaller protein, ci2. *Fold. Des*., **1**, 43–55.

McGuffin,L.J. *et al*. (2000) The PSIPRED protein structure prediction server. *Bioinformatics*, **16**, 404–405.

Mirny,L. and Shakhnovich,E. (2001) Evolutionary conservation of the folding nucleus. *J. Mol. Biol*., **308**, 123–129.

Mizuguchi,K. *et al*. (1998) JOY: protein sequence-structure representation and analysis. *Bioinformatics***, 14**, 617–623.

Neuwald,A.F. *et al*. (1995) Gibbs motif sampling: detection of bacterial outer membrane protein repeats. *Protein Sci*., **4**, 1618–1632.

Paiardini,A. *et al*. (2005) CAMPO, SCR_FIND and CHC_FIND: a suite of web tools for computational structural biology. *Nucleic Acids Res*., **33**, W50–W55.

Peters,B. *et al*. (2006) Identification of similar regions of protein structures using integrated sequence and structure analysis tools. *BMC Struct. Biol*., **6**, 4.

Pugalenthi,G. *et al*. (2007) SMotif: a server for structural motifs in proteins. *Bioinformatics*. **23**, 637–638.

Pugalenthi,G. *et al*. (2008a) MegaMotifBase: a database of structural motifs in protein families and superfamilies. *Nucleic Acid Res*., **36**, D218–D221.

Pugalenthi,G. *et al*. (2008b) Identification of catalytic residues from protein structure using support vector machine with sequence and structural features. *Biochem. Biophys. Res. Commun*., **367**, 630634.

Sandhya,S. *et al*. (2008) CUSP: an algorithm to distinguish structurally conserved and unconserved regions in protein domain alignments and its application in the study of large length variations. *BMC Struct. Biol.,* **8,** 28.

Saqi,M.A. and Sternberg,M.J. (1994) Identification of sequence motifs from a set of proteins with related function. *Protein Eng*., **7**, 165–171.

Shapiro,J. and Brutlag,D. (2004) FoldMiner: structural motif discovery using an improved superposition algorithm. *Protein Sci*., **13**, 278–294.

Thompson,J.D. *et al*. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*., **22**, 4673–4680.

Toyama,A (2004) Catalytic and structural role of a metal-free histidine residue in bovine Cu-Zn Superoxide dismutase *Biochemistry*, **43**, 4670–4679.

Webb,A.R. (2002) *Statistical Pattern Recognition*. John Wiley and Sons, London.

Yao,X. *et al*. (2001) Neural network ensembles and their application to traffic flow prediction in telecommunications networks. In *Proceedings of International Joint Conference on Neural Networks.* IEEE Press, Washington DC 1, pp. 693–698.