Prediction of Protein Long-Range Contacts Using GaMC Approach with Sequence Profile Centers

Peng Chen
Bioinformatics Research Center,
School of Computer Engineering,
Nanyang Technological University,
Singapore 639798
Email: pchen1978@gmail.com

Jinyan Li
Bioinformatics Research Center,
School of Computer Engineering,
Nanyang Technological University,
Singapore 639798
Email: jyli@ntu.edu.sg
Corresponding author

***Abstract*—In this paper, we apply an evolutionary optimization classifier, referred to as genetic algorithm-based multiple classifier (GaMC), to the long-range contacts prediction. As a result, about 44.1% contacts between long-range residues (with a sequence separation of at least 24 amino acids) are founded around the sequence profile (SP) centre when evaluating the top *L/5* (*L* is the sequence length of protein) classified contacts if the SP centers are known. Meanwhile, with the knowledge of sequence profile center and the GaMC method, about 20.42% long-range contacts are correctly predicted. Results showed that SP center may be a sound pathway to predict contact map in protein structures.**

***Availability*- http://mail.ustc.edu.cn/~bigeagle/gamc.htm**

***Keywords*-Long-range contact; sequence profile; evolutionary optimization; sequence profile centre;**

I. INTRODUCTION

During the last several decades more and more protein sequences are sequenced at an incredible high speed. At the same time, experimental determination of protein structures using x-ray crystallography and NMR techniques is a complicated and time-consuming problem and not efficient enough to allow for rapid structural determination of newly high-speed discovered sequences. However, protein structures are deemed as a key step toward understanding protein functions and taking rational molecule design. Therefore, discovering the relations between protein sequence and its corresponding three-dimensional (3D) structure using computational techniques is becoming more and more urgent.

It is well known that non-local interactions of residue pairs are crucial for proteins to attain their native state [1, 2]. Fariselli1 and Casadio reported that if residue contacts for a protein are known, the major features of its 3D protein structure could be achieved by combination of this knowledge with correctly predicted motifs of secondary structure [3]. More importantly, even corrupted map with nonphysical contacts of a protein could lead to recover its 3D structure by projecting the contact map onto its closest physically allowed structural counterpart [4]. Finally, previous results may indicate that 50% corrected contact prediction, at least for proteins with less than 150 amino acids, with 8Å distance cutoff ought to suffice that reconstruction [5].

A lot of previous works focused on residue contacts prediction using various methods, such as method with the use of evolutionary information [6], Self-Organizing Map (SOM) integrated by genetic programming (GP) [7], neural networks (NN) [8, 9], general input-output hidden Markov models (GIOHMMs) [5], support vector machine (SVM) [10, 11] and so on. Punta and Rost reported that about 30% of the predicted contacts were correct (in accuracy) with the residue separation at least six residues, where about 10% of the observed contacts are predicted (in coverage) [12]. Vullo's two-stage predictor achieved 19.8% prediction accuracy for minimum contact separation of 24 residues, when choosing the top $L/5$ contacts for evaluating prediction performance [13]. Wu and Zhang proposed a comprehensive assessment of sequence-based and template-based methods for contact map prediction and achieved accuracy around 20% for long-range contacts [14]. Currently, the most accurate contact predictor, NNcon, achieved 18% accuracy based on an evaluation of CASP8 dataset [15]. These methods solved the problem of residue contacts prediction from different angles, however, the development of computational approaches to predict inter-residues contacts is still at its embryonic stage. Therefore, fully exploring inter-residues contacts through proteins and designing novel approaches to predict residue contacts of proteins is extremely necessary.

In this paper, we propose a novel way to solve the residue contact problem with the analysis of sequence profile centers (SPC) [16]. The start point is based on the SPCs that represent average sequence profiles. One sequence profile is an encoding vector for a residue pair whose spatial distance between the members falls into one distance interval such as less than 8Å or from 8 to 10Å. Afterwards, we address the question of whether or not a multiple classifier being capable of learning the correlation between sequence information and the corresponding residue contacts of protein. To do that, we propose a genetic algorithm-based multiple classifier (GaMC), apply to calculate the distance between sequence

978-1-4244-5121-0/09/$25.00 ©2009 IEEE

profile of residue pair and each sequence profile center, and finally to make a conclusion of whether one residue pair being in contact or not. Our previous experimental results show that about 44.1% long-range contacts are around at their SPC, when selecting the top $L/5$ classified contacts and the residue pair more than 24 apart if the SPCs are known. As a result, about 20.42% long-range contacts are correctly identified using GaMC method.

II. Methods

A. Datasets and cross-validation

We obtained the protein chain set from PDB-REPRDB [17], which selects protein chains from PDB based on PDB Rel. 2007_11_14, and updated on 15 April 2009. We selected proteins chains that are resolved by X-ray crystallography with resolution $\leq$2.0Å, and R-factor $\leq$19%. The sequence identity between each two chains is less than 25%. As a result, we achieved 193 proteins with single chain that have Consurf-Hssp files [18].

To validate our approach, a two-fold cross-validation strategy was employed to conduct the related subsequent experiments. In this case, predictor was trained on one subset and tested on another one and vice versa.

B. Feature spaces

We firstly encoded input vectors for each pair of residues, i and j, then respectively stretch the two residues from N- to C-termini. Meanwhile, two corresponding sliding windows with an odd size of window length, referred to as *win*, are used to encoding input vectors. They are respectively centered at residue i and j, where *win* is set to 9 in our work. Due to the improvement of contacts prediction by the application of segment connecting the residues of i and j [11, 12, 19, 20], we took a third central window with five consecutive residues centered at the residue site *int ((i+j)/2)*.

To begin with, we used the property of residue sequence profile (SP) obtained from HSSP database [21], where each residue was represented by 20 elements whose values were evaluated from multiple sequence alignment and their potential structural homologs. As discussed above, the three windows contain *(9+5+9)=23* residues, each of which corresponds to a sequence profile vector with 20 elements. Totally, the input vector for one residue pair contains *20×23=460* elements, that is, one input vector includes 460 features or variables.

C. Definition of multi-class contacts

Usually, contact map of a polypeptide chain with sequence length N is represented by an $N\times N$ matrix, ***CM***. It is defined in terms of spatial distances between C-alpha atoms of residues and a predefined cutoff distance d. Usually, d is set as 8Å. So contact map for two-class contact can be defined as:

$$CM_{ij} = \begin{cases} 1 & if \quad d(i,j) \leq d \\ 0 & Otherwise, \end{cases} , |i-j| \geq 24 \qquad (1)$$

where $d(i,j)$ denotes the distance between residues i and j.

In this paper, we took another contact expression, called as multi-class contact, as follows:

$$CM_{i,j} = \begin{cases} 0 & if \quad d(i,j) \leq d \\ 1 & d < d(i,j) \leq d_1 \\ 2 & d_1 < d(i,j) \leq d_2 \\ \vdots & \vdots \\ m & d_{m-1} < d(i,j) \leq d_m \\ \vdots & \vdots \\ \mathrm{n} & d_{n-1} < d(i,j) \end{cases} , |i-j| \geq 24 \qquad (2)$$

In this novel expression two residues, separated more than 24 residues in sequence and therefore named as long-range residue pair, are in long-range contact if their spatial distance is less than or equal to 8Å. The residue pair is assigned as contact distance-class 0. To continuously assign contact distance-class *1, 2, …, or n* to other residue pairs, similar representations can be done. For a protein chain, for instance, there are M long-range residue pairs whose spatial distances are more than 8Å. By ranking the M residue pairs in order of spatial distance, n distance intervals are achieved whose numbers are roughly the same, about *int(M/n)*. For example, residue pairs in the *m-th* interval are to belong to contact distance-class m. In this case, their distances are more than d_{m-1} and less than or equal to d_m. Therefore, distance between one pair of residues belonging to contact distance-class m is farther than that between pair of residues belonging to contact distance-class l, if $m > l$ and *m, l* $\leq n$. Finally, multi-class contact map can be constructed and then be applied to a multiple classifier system which performed better than a single classifier system.

D. Description of sequence profile centre

Provided that a sequence profile S be the encoding vector for representing one residue pair. So, as discussed above, one sequence profile vector includes 460 elements. Another definition is sequence profile center (SPC), which simply is an average calculation of all the sequence profile vectors belonging to one contact distance-class for one protein chain. The definition of sequence profile center C_i in one protein chain j for contact distance-class i is given as follows:

$$C_i^j = \frac{1}{m_i} \sum_{l=1}^{m_i} S_i(l) \qquad (3)$$

where $S_i(l)$ denotes the $l-th$ sequence profile whose corresponding residue pair is to belong to contact distance-class i; m_i is the number of residue pairs belonging to

contact distance-class i; and $i \leq n$, where n is the number of the contact distance-classes.

Then, we can calculate the distance between one SP and each SPC. Generally, the label i of SPC C_i is assigned to a SP if the SPC is the nearest than other SPCs to the SP. Some other representations for profile center or centroid can be found in literature [22].

For testing our method, the SP centers for test protein chain, due to unknown 3D structure, need to be extracted from the training protein chains. All test chain use the same SP centers. So, the definition of SP center C_i of contact distance-class i for test chains is given as follows:

$$C_i = \frac{1}{m} \sum_{j=1}^{m} C_i^j \tag{4}$$

where m denotes the number of training protein chains.

E. GaMC predictor architecture

In this paper, we analyze the long-range contacts using our proposed GaMC predictor, and its original input vectors are transformed in such a way that the classification rate is significantly enhanced while retaining the efficiency and simplicity of the original vectors. For particular problem in the prediction of the long-range contacts, the normalized SP vectors are regarded as input data into the GaMC predictor. Then we consider one sample vector as a variable set, and proceed to search for an optimal transformation for these variables based on genetic optimization. After obtaining the optimal transformation, multiple sub-classifiers based on distance dissimilarity is used to classify test samples. Finally, this method derives a modified multiple classifier system by fusing the outputs of a number of independent classifiers.

1) Chromosome encoding: Genetic algorithm [23] is an adaptive heuristic search algorithm, which has been commonly applied in optimization problems for searching optimal solutions within a solution population. The technique behaves in an analogous manner to Darwinian evolution by maintaining a population of solutions based on a fitness function, and strives to obtain the individuals with the maximum or minimum fitness value within the population. A string represents each candidate in the population, which is associated with a fitness value that reflects its capability to survive into the next generation during the evolution process.

To study the long-range contacts, we let V be a feature space set $V = (v_1, v_2, \ldots, v_m)$, where v_i is a feature variable and m is the dimension of feature vectors. Each residue pair within a protein is represented as a feature vector of V. We want to train a GA-based classifier that can correctly classify the feature vectors into K classes $C_1, C_2, \ldots, C_K$. In this work, we focus on the problem of five contact distance-classes; one class for long-range residue pairs in contacts and the other four classes for long-range residues pairs not in contacts. Our goal is to search for an optimal feature selector T that maximizes the classification rate based on the corresponding selected features. To obtain the optimal feature selector T, GAs are applied to search through the space of feature transformers with a fitness function. To do that, firstly, a vector v_i of the feature space V is represented as a chromosome string S_i. A chromosome is composed of three kinds of *transformers* represented by characters a, b, and c, and the size of a chromosome is the same as a feature vector.

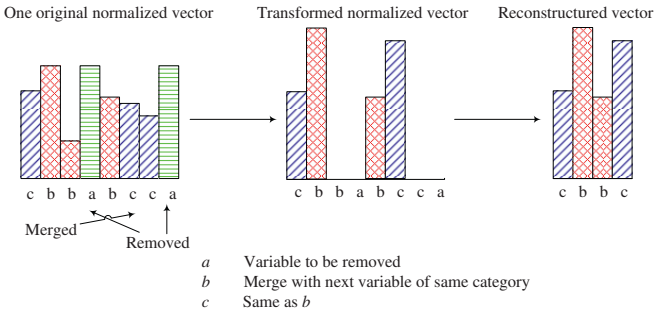


Figure 1. Transformation for Input Vector. One histogram bin in each original normalized vector denotes one feature or variable and the height corresponds to the magnitude of the feature. The transformed vector should be also normalized but not shown here for the clear comparison between the original vector and the resulted vector.

The schema for chromosome encoding is as follows: (1) Character a in a chromosome indicates that the value in the corresponding position in all feature vectors in V will be removed; (2) Two consecutive b's or c's indicate that the values in the corresponding positions will merged together. For instance, for a feature vector of 8 dimensions, its corresponding chromosome is a string of 8 characters from the ternary alphabet *(a, b, c)*. For instance, Figure 1 illustrates the selection process for a feature vector (x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8). In this case, the corresponding chromosome is '*cbbabcca*'. After being applied the transformers, the elements of the sample feature vector that remain or are merged are concatenated and normalized to form a new vector of four dimensions. The new normalized vector will be used for long-range contacts classification.

2) Definition of fitness function: For each transformation T^m associated with the string S_m, we can construct the transformed input function $f_i(v_t^m|T^m)$ for input training vector x_i. For distance-class C_k, we can define the following centroid function based on T^m as:

$$f^k(v_t^m) = \frac{1}{|C_k|} \sum_{x_i \in C_k} f_i(v_t^m|T^m) \tag{5}$$

where $|C_k|$ is the cardinality of distance-class C_k

Given these centroid functions, a new distance-class structure $C_l (l = 1, 2, ..., K)$can be imposed on the input vector x as follows:

$$x \in C_l \quad if \quad d(f, \overline{f^k}) \leq d(f, \overline{f^l}), l = 1, 2, ..., K \qquad (6)$$

where f is the function of the unknown model x and $d(\cdot\cdot)$ is a measure of dissimilarity between two functions.

Now, we can extend the above GA method by producing a multiple classifier system and by training a specific sub-classifier for each contact distance-class of the system, due to the possible no-existence of a global optimal transformation and interval subset for all contact distance-classes. The basic idea is to develop K sub-classifiers for each individual distance-class using a GA training process, and the aim is to search for the optimal transformation and interval subset for each sub-classifier. To achieve the purpose, the normalized conditional function $f_i(v_t^{m,k}|T^{m,k})$ is used as input for the *k-th* classifier, where $T^{m,k}$ are the optimal transformation for this sub-classifier, respectively. Moreover, fusing the outputs of a number of independent classifiers can improve classification rate since the errors made by a classifier may be corrected by the others [24-26].

Recall that the *k-th* classifier is trained to identify whether or not an unknown model comes from distance-class k. The output of the *k-th* classifier is denoted as O_k and makes a final decision about whether the O_k is 1, which means that the input data comes from distance-class k, or 0, which indicates that the input data does not come from contact distance-class k. To perform the task, one fitness function for each *k-th* classifier is used to measure the discrepancies between the original distance-class structures, $C_1, C_2, \ldots, C_k$ and the imposed distance-class structure, $C_1^m, C_2^m, ..., C_K^m$, based on the string S_m. So the fitness function is defined as:

$$\aleph_k^m = \sum_{k=1}^{K} \left| C_k \bigcap C_k^m \right| + \sum_{k=1}^{K} \left| \overline{C_k} \bigcap \overline{C_k^m} \right| \qquad (7)$$

where $\overline{C_k}$ and $\overline{C_k^m}$ denote the complements of C_k and C_k^m, respectively. The first term counts the number of correct positive classifications, i.e., the number of patterns in distance-class k actually classified as belonging to the distance-class. On the other hand, the second term counts the number of correct negative classifications, i.e., the number of patterns not belonging to distance-class k that are correctly classified as not belonging to distance-class k.

Particularly, the maximal value of $\aleph_k^m$ will be obtained when the two distance-class structures exactly coincide, and its value will decrease as their discrepancy increases.

F. GA parameters

Table 1 summarizes the parameters for GA used to train our GaMC predictor.

G. Performance indexes

In fact, to evaluate the performance of our classifiers, some measurements about performance index have to be introduced. Here, we applied the criteria of accuracy (Acc) and coverage (Cov), which were adopted at CASP/CAFASP [12, 27] and defined as follows:

$$Acc = \frac{TP}{TP + FP} \quad Cov = \frac{TP}{N_{act}} \qquad (8)$$

where *TP* denotes the number of true positives, *FP* denotes the number of false positives, and N_{act} is the number of actual contacts.

Table I
THE PARAMETERS LIST USED IN GA

Parameter	Value
Population size	150
Crossover probability	0.95
Mutation probability	0.01
Crossover type	Single point
Selector type	Roulette Wheel
Scaling scheme	Linear
Termination	Best score not changed over 120 generations

III. RESULTS

A. Analyses on dataset

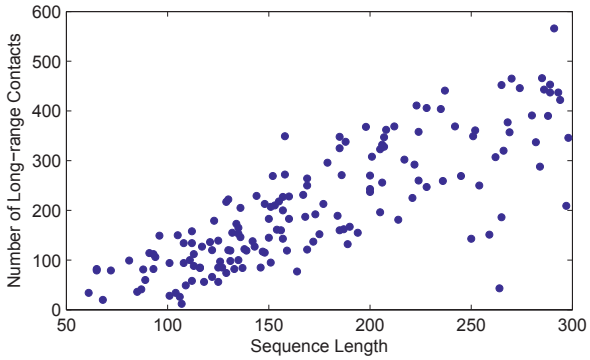


Figure 2. Relationship of the number of long-range contacts versus the corresponding protein sequence length.

For our dataset, there are 193 protein chains with 34549 residues and 42089 long-range contacts. Some proteins have more residue pairs in long-range contact and some proteins contain fewer contacts with respect to protein sequence length. But approximately protein sequence length has a linear relationship with the corresponding number of long-range contacts through a thorough statistics for the whole dataset [20]. The distribution of sequence length versus the number of long-range contacts is shown in Figure 2.

B. Transformation of sample vectors

After GA was applied to reduce the dimensionality of input vector, the transformed input vector can be used as input vector into classifier. Meanwhile, some input variables

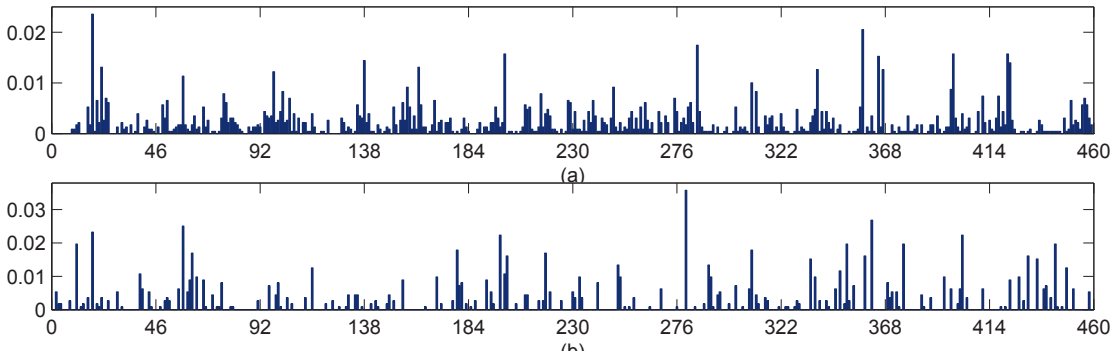


Figure 3. (a) One original input vector with 460 dimensions, (b) Transformed input vector by the transformation for contact distance-class 0 classifier. All the two vectors are equalized by normalized themselves.

were removed or merged, but it was done without decreasing the information of input. Moreover, in doing so the computational complexity was dropped dramatically. For five-distance-class problem in our experiments, five optimal non-linear transformations for each sub-classifier were achieved. Therefore, there are five-discard for representing the ratio of the number of removed to that of total variables. For instance, one discard ratio 34.348% for distance-class one classifier is illustrated in Figure 3. As a result, there are 143 original variables removed; and 65 variables merged together into 40 variables. The transformed vector can be obtained by normalizing itself after removing or merging original variables. Other transformed vectors, similar to the illustration in Figure 3, can be got for the remainder classifiers, such as for distance-classes 2, 3, 4, or 5, with respect to their corresponding transformations.

C. Analyses for proteins with respect to their numbers of long-range contacts

In our dataset, due to containing unknown or non-standard amino acid residues or chain breaks in some protein chains, and even including zero or little long-range contacts in some protein chains (seeing in Figure 2), the classification performance for these protein chains may decrease to some extent. If those protein chains are excluded, then higher classification performance may be obtained through our dataset.

D. Performance of GaMC predictor

Running the GaMC predictor can lead to predict whether or not one sample vector is belonging to long-range contact distance-class. Since the contact prediction accuracy varies significantly with individual proteins and their structure classes [7], we calculate accuracy and coverage for each test protein. For each protein chain, we select four levels of predicted contacts ranked by predicted distance between SP vector and SP center for long-range contact. The reason in doing so is that the total number of true contacts has approximately a linear relationship with the protein length [20]. And the relationship was also shown in Figure 2 for our dataset. In detail, the four levels are 'All', 'L', 'L/2', and 'L/5', respectively, where L denotes protein length. Results show that in many cases (e.g. 1hh7A, 1bxaA, 1gpr_, 1cewI, 1cznA, 1gn0A, 1igd_, 1tif_, 1s8nA) the prediction accuracies are larger than 30%.

However, the prediction accuracies for some protein chains such as 1cv8_ and 1c7kA are pretty low. Investigating through the dataset, we observe that the contact prediction accuracy is related to the prediction of SP centers, the number of long-range contacts and the quality of multiple sequence alignment as well as the proportion of beta-sheets.

Furthermore, in order to investigate the distribution of our GaMC long-range contact prediction with respect to CATH [28] domain classes, we compute the average accuracy on the five CATH structure classes (Table II). According to Table II, the contact prediction accuracy of proteins belonging to β-sheets ($\alpha-\beta$, all β) is higher than that of all α-helical proteins, which is consistent with other previous observations [7, 11]. In Table II, the average accuracy is about 20.42% when selecting the top $L/5$ classified contacts and the residue pair with 24 apart. Taking into account the inherent physical restraints of protein structures, this prediction performance may be helpful for reconstructing an *ab initio* low-resolution structure since previous experiments show that only $L/5$ or even less residues contacts are required to reconstruct a low resolution structure for a small protein [29-33]. However, the hard challenge is how to reconstruct a protein structure from even a corrupted predicted contact map [4], where contact restraints are much less reliable than the experimental contacts determined by NMR techniques.

E. Performance Comparison based on CASP7 evaluation

The CASP7 evaluation procedure is focused on inter-residue contact predictions with linear sequence separation $\geq$ 12 and $\geq$ 24, respectively [34, 35], while in this work

Table II
PREDICTION ACCURACY OF GAMC PREDICTOR

CATH class	Classification Accuracy (Unit: %)				Protein Number
	ALL[+]	L	L/2	L/5	
Alpha	1.71	6.25	7.73	10.96	24(29)
Beta	4.56	11.07	14.53	20.11	32(37)
Alpha Beta	4.84	14.5	18.3	23.93	89(97)
Few SS[†]	10.37	11.35	15.6	24.5	1(1)[*]
Multi-domain chains	2.64	8.35	11.87	16.71	21(29)
Average	4.09	11.86	15.23	20.42	167(193)[#]

[†]'Few SS' means there are few secondary structures in this CATH domain class.
[+]'ALL' denotes the number level of original length that is used to measure the classification performance.
[*]The bracket denotes the number of protein chains except the Few SS chain, and we also calculate the average performance excluding the chain due to no statistical meaning for only one protein chain.
[#]The bracket denotes the original protein number before excluding the proteins with unknown or non-standard amino acid residues or chain breaks.

we only focus on long-range contact prediction with linear sequence separation ≥ 24 and with assessing the top $L/5$ predicted contacts, where L is the sequence length of the protein. These evaluation metrics are also similar to those used in the previous Critical Assessment of Fully Automated Structure Prediction Methods (CAFASP) [36, 37] and in the EVA contact evaluation server [38]. We use the similar procedure and the same test proteins to evaluate the accuracy and coverage for our GaMC predictor.

Table III
PERFORMANCE COMPARISON BASED ON CASP7 EVALUATION*

Methods	Separation ≥ 24	
	Accuracy	Coverage
BETApro [39]	19.7	3.2
Distill [13, 40]	13.7	1.4
GPCPRED [7]	10.5	2.0
Possum [41]	21.4	2.6
PROFcon [12]	8.1	1.6
SAM_T06_server [42]	18.5	3.9
SVMcon [11]	13.1	2.8
GaMC	18.8	3.0

*Noted that some data are extracted from Table III in literature [11]. The nine predictors are evaluated on the 13 de novo domains of CASP7. The experimental structures of the targets and the domain classification can be downloaded at the CASP7 web site (http://predictioncenter.org/casp7). The accuracy and coverage of contact predictions are evaluated at sequence separation ≥ 24 and with assessing the top $L/5$ predicted contacts.

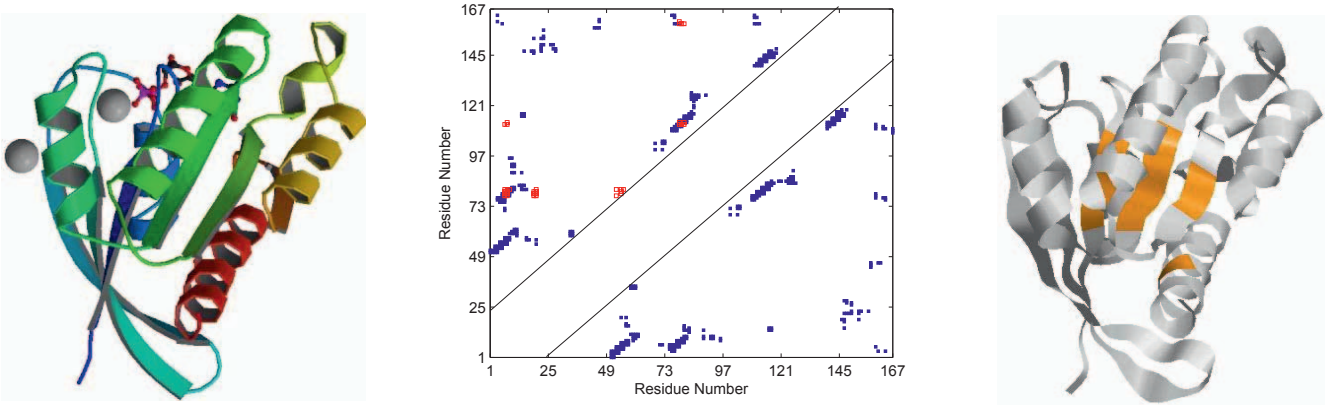
Contact map predictors participating in CASP7 include BETApro, Distill, GPCPRED, PROFcon-Rost, Possum, SAM_T06_server, SVMcon and so on. Table III reports the performance of the seven automated contact map predictors in the CASP7 experiment. The performance of GaMC predictor is also appended under the bottom of Table III. It can be seen that its accuracy is 18.8%, overall slightly behind Possum and BETApro. Its coverage at a sequence separation threshold of 24 is 3.0%, which is less than SAM-T06 and BETApro.

As previous discussed [11], something should be paid attention to understanding the meanings of the comparison among these methods. One thing should be noted that in the CASP7 experiment, methods being made predictions for part of domains, such as PROFcon, can not be directly compared with other methods. Here we include its results for completeness in Table III. Additionally, since the evaluation dataset and scheme we used may be slightly different from the official CASP7 evaluation, we only try to evaluate the current state of the art of contact predictors instead of ranking them.

Previous works may indicate that prediction accuracy of 50% for distant contacts with 8Å distance cutoff ought to suffice to reconstruct 3D protein structure, at least for proteins with less than 150 amino acids [5, 34]. Other results showed that the accuracy level of about 30% is required for deriving moderately accurate (low resolution) 3D protein structures from scratch [29-33]. Despite the lower accuracy and coverage made by protein contact predictor, it is an important step towards reaching the accurate level [5, 12, 34]. From previous CASP prediction results, it can found that in a word, these predictors tend to perform more and more better [11].

Actually, in this work, the contact prediction accuracy is related to the SP centers, the number of long-range contacts and the quality of multiple sequence alignment as well as the proportion of beta-sheets. However, it is extremely difficult to build a specific non-linear expression based on the relationship.

To clearly illustrate the prediction results of GaMC predictor, we figure the native 3D structure and the predicted contact map of protein 1kao_A (see Figure 4). Protein 1kao_A contains one protein chain with 167 residues. It consists of six α helixes and five β-sheets. In this case, $L/5$ (34) predicted contacts with sequence separation ≥ 24 are selected, and the corresponding prediction accuracy is 35.29%. It is shown that the predicted contacts are clustered



(a) 3D structure of protein chain 1kao_A

(b) Predicted (red rectangle) versus natural (blue grid) long-range contacts

(c) True predictions of long-range contacts in its 3D structure (in brown)

Figure 4. Comparison of predicted versus natural long-range contacts for protein chain 1kao_A. The two slope black lines in 4(b) denote the separation line where two members of residue pair separate 24 residues in sequence.

around the true contacts (see Figure 4(b)). It is of interest that many false positive contacts are also clustered around the true contacts. Therefore, even these contacts may be helpful to reconstructing protein structure.

IV. Conclusion

As pointed out by Baldi [43], a machine learning algorithm adopting a simple representation of a sequence space can be much more powerful and useful than using the original data containing all details. We found that most long-range contacts or non-long-range contacts are near to their SP centers[16]. Therefore, we developed the GaMC predictor to reduce the dimensionality of the input features. The purpose of the GaMC method is to transform input vectors to obtain higher classification rate. As a result, our method could be more useful for the problem of long-range contacts prediction.

Summarizing this work, we can conclude as follows: (1) Most long-range contacts or non-long-range contacts are clustered around their SP centers by using our GaMC predictor when selecting the top $L/5$ classified contacts. And it was found that about 44.1% long-range contacts are near to their SP centers and about 20.42% long-range contacts can be correctly predicted under the same condition; (2) In this work, an interesting phenomenon may give us some clues that proteins with long-range contacts to be clustered together into few clusters might make higher classification performance. Likewise, non-long-range contacts also behave the similar manner. Therefore, for contact distance-class or one non-contact distance-class, using a set of SP centers with the information of predicted secondary structures and hydrophobicity might obviously improve classification prediction; (3) A novel classifier or convincing method might be proposed to predict long-range contacts Based on our previous analysis on SP centers [16]. First, we may use techniques such as radial basis function neural network or support vector regression to predict each SP centers. Second, GaMC can be applied to transform the input feature space based on the optimization rule that the highest classification rate is obtained. Finally, fusing the outputs from multi-subclass classifiers might achieve relatively higher performance.

In conclusion, this paper proposes a promising way to predict long-range contacts based on SP centers. It can be expected that the predictor based on SP centers can make a great improvement on the prediction of contact map and even protein structures in the future research works.

Acknowledgment

This work was supported in part by the Singapore MOE ARC Tier-2 funding grant T208B2203.

References

[1] M. Niggemann and B. Steipe., *Exploring local and non-local interactions for protein stability by structural motif engineering*. J. Mol. Biol., vol. 296, pp. 181-195, 2000.

[2] M. M. Gromiha and S. Selvaraj, *Inter-residue interactions in proteins folding and stabiligy*. Progress in Biophysics & Molecular Biology, vol. 86, pp. 235-277, 2004.

[3] P. Fariselli and R. Casadio, *A neural network based predictor of residue contacts in proteins*. Protein Engineering, vol. 12, pp. 15-21, 1999.

[4] M. Vendruscolo, E. Kussell, and E. Domany, *Recovery of protein structure from contact maps*. Fold Des., vol. 2, pp. 295-306, 1997.

[5] G. Pollastri and P. Baldi, *Prediction of contact maps by GIOHMMs and recurrent neural networks using lateral propagation from all four cardinal corners*. Bioinformatics, vol. 18, pp. S62-S70, 2002.

[6] S. Vicatos, V. B. R. Boojala, and Y. Kaznessis, *Prediction of distant residue contacts with the use of evolutionary information*. PROTEINS: Structure, Function, and Bioinformatics, vol. 58, pp. 935-949, 2005.

[7] R. M. MacCallum, *Striped sheets and protein contact prediction*. Bioinformatics, vol. 20, pp. 224-231, 2004.

[8] P. Fariselli, O. Olmea, A. Valencia, and R. Casadio, *Prediction of contact maps with neural networks and correlated mutations*. Protein Engineering, vol. 14, pp. 835-843, 2001.

[9] P. Chen, D.-S. Huang, X.-M. Zhao, and X. Li, *Predicting Contact Map Using Radial Basis Function Neural Network with Conformational Energy Function*. International Journal of Bioinformatics Research and Applications, vol. 4, pp. 123-136, 2008.

[10] P. Chen, K. Han, X. Li, and D. S. Huang, *Predicting Key Long-Range Interaction Sites by B-Factors*. Protein & Peptide Letters, vol. 15, pp. 478-483, 2008.

[11] J. L. Cheng and P. Baldi, *Improved residue contact prediction using support vector machines and a large feature set*. BMC Bioinformatics, vol. 8, p. 113, 2007.

[12] M. Punta and B. Rost, *PROFcon: novel prediction of long-range contacts*. Bioinformatics, vol. 21, pp. 2960-2968, 2005.

[13] A. Vullo, L. Walsh, and G. Pollastri, *A two-stage approach for improved prediction of residue contact maps*. BMC Bioinformatics, vol. 7, p. 180, 2006.

[14] S. Wu and Y. Zhang, *A comprehensive assessment of sequence-based and template-based methods for protein contact prediction*. Bioinformatics, vol. 24, pp. 924-931, 2008.

[15] A. N. Tegge, Z. Wang, J. Eickholt, and J. Cheng, *NNcon: improved protein contact map prediction using 2D-recursive neural networks*. Nucleic Acids Research, vol. 37, pp. W515-W518, 2009.

[16] P. Chen, B. Wang, H.-S. Wong, and D.-S. Huang, *Prediction of Long-range Contacts from Sequence Profile*. in Neural Networks, 2007. IJCNN 2007. International Joint Conference on, Orlando, FL, USA, 2007, pp. 938-943.

[17] T. Noguchi and Y. Akiyama, *PDB-REPRDB: a database of representative protein chains from the Protein Data Bank (PDB) in 2003*. Nucleic Acids Res., vol. 31, pp. 492-493, 2003.

[18] F. Glaser, Y. Rosenberg, A. T. P. Kessel, and N. Ben-Tal, *The ConSurf-HSSP Database: The Mapping of Evolutionary Conservation Among Homologs Onto PDB Structures*. PROTEINS: Structure, Function, and Bioinformatics, vol. 58, pp. 610-617, 2005.

[19] J. Gorodkin, *Using sequence motifs for enhanced neural network prediction of protein distance constraints*. Int. Conf. Intell. Syst. Mol. Biol., pp. 95-105, 1999.

[20] O. Lund, K. Frimand, J. Gorodkin, H. Bohr, J. Bohr, J. Hansen, and S. Brunak, *Protein distance constraints predicted by neural networks and probability density functions*. Prot. Eng., vol. 10, pp. 1241-1248, 1997.

[21] C. Dodge, R. Schneider, and C. Sander, *The HSSP database of protein structure-sequence alignments and family profiles*. Nucleic Acids Res., vol. 26, pp. 313-315, 1998.

[22] R. H. Leary, J. B. Rosen, and P. Jambeckz, *An Optimal Structure-Discriminative Amino Acid Index for Protein Fold Recognition*. Biophysical Journal, vol. 86, pp. 411-419, 2004.

[23] D. E. Goldberg, *Genetic Algorithms in Search Optimization and Machine Learning*. Reading, MA: Addison-Wesley, 1989.

[24] J. Kittler and F. M. Alkoot, *Sum versus vote fusion in multiple classifier systems*. IEEE Trans. Pattern Anal. Mach. Intell., vol. 25 pp. 110-115, 2003.

[25] L. I. Kuncheva, *Combing pattern classifiers: methods and algorithms*. U.S.: Wiley, 2004.

[26] L. K. Hansen and P. Salamon, *Neural network ensembles*. IEEE Trans. Pattern Anal. Mach. Intell., vol. 12, pp. 993-1001, 1990.

[27] D. Fischer, L. Rychlewski, R. L. Dunbrack, O. Jr., A. R., and A. Elofsson, *CAFASP3: the third critical assessment of fully automated structure prediction methods*. Proteins: Structure,Function, and Genetics, vol. 53, pp. 503-516, 2003.

[28] C. A. Orengo, A. D. Machie, S. Jones, D. T. Jones, M. B. Swindells, and S. M. Thornton, *CATH - a hierarchic classification of protein domain structures*. Structure, vol. 5, pp. 1093-1108, 1997.

[29] A. Aszodi, M. Gradwell, and W. Taylor, *Global fold determination from a small number of distance restraints*. J Mol Biol, vol. 251, pp. 308-326, 1995.

[30] Y. Zhang and J. Skolnick, *Automated structure prediction of weakly homologous proteins on a genomic scal*. P.N.A.S., vol. 101, pp. 7594-7599, 2004.

[31] J. Skolnick, A. Kolinski, and A. Ortiz, *MONSTER: A method for folding globular Proteins with a small number of distance restraints*. J Mol Biol, vol. 265, pp. 217-241, 1997.

[32] A. Ortiz, A. Kolinski, P. Rotkiewicz, B. Ilkowski, and J. Skolnick, *Ab initio folding of proteins using restraints derived from evolutionary information*. Proteins Suppl, vol. 3, pp. 177-185, 1999.

[33] Y. Zhang, A. Kolinski, and J. Skolnick, *TOUCHSTONE II: a new approach to ab initio protein structure prediction*. Biophysical Journal, vol. 85, pp. 1145-1164, 2003.

[34] O. Grana, D. Baker, R. MacCallum, J. Meiler, M. Punta, B. Rost, M. Tress, and A. Valencia, *CASP6 assessment of contact prediction*. Proteins, vol. 61, pp. 214-224, 2005.

[35] J. Moult, K. Fidelis, A. Tramontano, B. Rost, and T. Hubbard, *Critical assessment of methods of protein structure prediction (CASP) - round VI*. Proteins, vol. 61, pp. 3-7, 2005.

[36] A. Lesk, L. L. Conte, and T. Hubbard, *Assessment of novel fold targets in CASP4: predictions of three-dimensional structures, secondary structures, and interresidue contacts*. Proteins, vol. 45, pp. 98-118, 2001.

[37] D. Fischer, A. Elofsson, L. Rychlewski, F. Pazos, A. Valencia, A. Godzik, B. Rost, A. Ortiz, and R. Dunbrack, *CAFASP-2: the second critical assessment of fully automated structure prediction methods*. Proteins, vol. 45, pp. 171-183, 2001.

[38] O. Grana, V. Eyrich, F. Pazos, B. Rost, and A. Valencia, *EVAcon: a protein contact prediction evaluaton*. Nucleic Acid Res., vol. 33, pp. W347-W351, 2005.

[39] J. Cheng and P. Baldi, *Three-Stage Prediction of Protein Beta-Sheets by Neural Networks, Alignments, and Graph Algorithms*. Bioinformatics, vol. 21, pp. i75-i84, 2005.

[40] D. Bau, A. Martin, C. Mooney, A. Vullo, I. Walsh, and G. Pollastri, *Distill: a suite of web servers for the prediction of one-, two- and three-dimensional structural features of proteins*. BMC Bioinformatics, vol. 7, p. 402, 2006.

[41] N. Hamilton, K. Burrage, M. Ragan, and T. Huber, *Protein contact prediction using patterns of correlation*. Proteins, vol. 56, pp. 679-684, 2004.

[42] K. Karplus, C. Barrett, and R. Hughey, *Hidden Markov models for detecting remote protein homologies*. Bioinformatics, vol. 14, pp. 846-856, 1998.

[43] P. Baldi and S. Brunak, *Bioinformatics: The machine learning approach*. London, England: The MIT Press, 2001.