Journal of Biomedical Informatics 45 (2012) 173–183

Contents lists available at SciVerse ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Protein contact map prediction using multi-stage hybrid intelligence inference systems

Anas A. Abu-Doleh [a,*], Omar M. Al-Jarrah [a], Asem Alkhateeb [b]

[a] Department of Computer Engineering, Faculty of Computer and Information Technology, Jordan University of Science and Technology, P.O. Box 3030, Irbid 22110, Jordan
[b] Biotechnology and Genetic Engineering Department, Jordan University of Science and Technology, P.O. Box 3030, Irbid 22110, Jordan

ARTICLE INFO

Article history:
Received 9 January 2011
Accepted 24 October 2011
Available online 31 October 2011

Keywords:
Protein contact map prediction
Hybrid prediction model
Fuzzy inference system

ABSTRACT

Proteins are one of the most important molecules in organisms. Protein function can be inferred from its 3D structure. The gap between the number of discovered protein sequences and the number of structures determined by the experimental methods is increasing. Accurate prediction of protein contact map is an important step toward the reconstruction of the protein's 3D structure. In spite of continuous progress in developing contact map predictors, highly accurate prediction is still unresolved problem. In this paper, we introduce a new predictor, JUSTcon, which consists of multiple parallel stages that are based on adaptive neuro-fuzzy inference System (ANFIS) and K nearest neighbors (KNNs) classifier. A smart filtering operation is performed on the final outputs to ensure normal connectivity behaviors of amino acids pairs. The window size of the filter is selected by a simple expert system. The dataset was divided into testing dataset of 50 proteins and training dataset of 450 proteins. The system produced an average accuracy of 45.2% for the sequence separation of six amino acids. In addition, JUSTcon outperformed SVMcon and PROFcon predictors in the cases of large separation distances. JUSTcon produced an average accuracy of 15% for the sequence separation of 24 amino acids after applying it on CASP9 targets.

1. Introduction

Proteins play a vital role in our life because they perform important tasks such as catalysis of biochemical reactions, transport of nutrients, and transmission of signals [1]. The importance of proteins in our life drives biologists to discover more proteins and study their biological functions. The 3D structure of a protein can be used as a good indicator of its function. The determination of 3D structure by biological methods such as X-ray [3] and NMR [4] is very cumbersome and costly [2].

Despite the improvement in experimental procedures to determine protein structure, the gap between the number of known protein sequences and their structures continues to increase [5]. Therefore, developing new machine learning approaches or improving current approaches to predict protein structure can decrease this gap.

Earlier studies indicate that developing an accurate protein contact map predictor will be very helpful in the reconstruction of protein 3D structure [6–8,13]. Accordingly, much research is implemented in this problem due to the current low accuracy [9–11].

Researchers have used several techniques to improve the accuracy of prediction. For example, Cheng and Baldi developed a new contact map predictor (SVMcon) that uses support vector machines (SVMs) to predict residues contact in the proteins [9]. Vullo et al. [14] separated the task into two stages. The first stage is devoted for the prediction of the contact map's principal eigenvector (PE) from the primary sequence using bi-directional recurrent neural networks. The second stage is dedicated for the reconstruction of the contact map from the PE and primary sequence based on DAG-RNNs that was described in [10]. The PROFcon method [11] was introduced by Punta and Rost. In this method, they combine information from sequence alignments, predictions of the secondary structure, predictions of solvent accessibility, the region between two residues, and from average properties of the entire protein. All of these inputs are fed to a simple feed-forward neural network with a back-propagation algorithm. Gobel et al. [16] used the correlated mutation approach to predict contact maps. Pollastri and Bladi proposed a new approach called GIOHMMs [10]. This method is based on 2D generalization of bi-directional input–output HMMs (BIOHMMs) and bi-directional recurrent neural networks (BRNNs).

2. CASP and protein contact map definition

Critical assessment of techniques for proteins structure prediction (CASP) is an organization that sets up the criteria for assessment

[*] Corresponding author. Address: P.O. Box 620019, Irbid 21162, Jordan.
E-mail addresses: a.abudoleh@yahoo.com (A.A. Abu-Doleh), aljarrah@just.edu.jo (O.M. Al-Jarrah), asemalkhateeb@just.edu.jo (A. Alkhateeb).

1532-0464/$ - see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2011.10.008

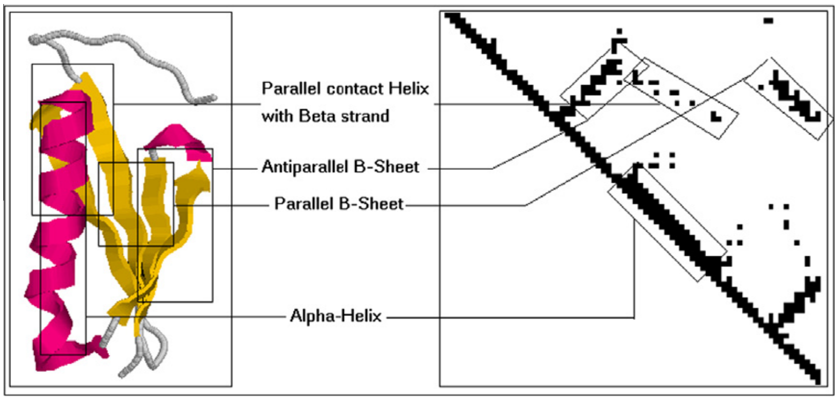


Fig. 1. Left: A view of protein with 3D representation. Right: The corresponding contact map with the threshold of contact 8 A.

of protein structure prediction in order to improve the quality of research [22]. According to CASP, two residues are considered to be in contact when the 3D physical distance between their C-beta atoms (C-alpha for GLY) is below a threshold value of 8 Å. Accordingly, the contact map for a protein sequence with N amino acids is $N \times N$ binary symmetrical matrix. The (ith,jth) component of the map is 1 if the amino acid pair at ith and jth locations fulfills the connectivity condition. Protein 2D structure can be represented as a contact map as shown in Fig. 1.

In CASP, three regions are classified according to the distance between an amino-acid contact pair: long-range (at least 24 residues along the sequence), Medium-range ($12 \leqslant \text{separation} < 24$), and Short-range ($6 \leqslant \text{separation} < 12$). Two metrics are used in CASP for measuring the performance: accuracy and coverage. Accuracy is defined as (1):

$$Accuracy = TP/(TP + FP) \tag{1}$$

where True Positives (TPs) is the number of correctly predicted contacts and False Positives (FPs) is the number of incorrectly predicted contacts. Coverage or sensitivity is defined as (2):

$$Coverage = TP/(Native\ contact) \tag{2}$$

where *Native contact* is the number of observed contacts.

According to CASP, the predicted contact values for all amino acid pairs for a specific protein will be sorted descendingly according to the predicted value. After that, the top X values in the sorted list are used only to calculate the accuracy and coverage. X is usually selected as $2L$, L, $L/2$, and $L/5$, where L is the protein length. Moreover, according to CASP, selecting $X = L/5$ is preferred over the others [23].

3. Datasets

In order to produce reliable results, proteins used in a trusted benchmarking set are used in this work. Evaluation of automatic protein structure prediction servers (EVA) supports researchers with huge dataset of proteins under specific criteria. In this work, the dataset was downloaded on February 7, 2008 from the EVA servers. Mainly, no pair in a subset has more than 33% identical residues over more than 100 aligned residues [24,25]. In addition, the preference is given to high-resolution structures. EVA lists on its servers a huge database of protein chains and their PDB files, FASTA files, PSI-Blast files, and other related useful files formats.

A filtering process is applied to eliminate unwanted protein PDB files [26] that have unusable data. These files may give misleading results during training stage. This process was adopted in many previous works [9–11]. It starts by removing the corrupted PDB files, which may contain erroneous or incomplete data. Then short and long proteins are removed. Short sequences with less than 30 amino acids are removed because most likely they do not have actual structure and may disrupt the system during the training stage, which may result in unreliable outputs for the testing

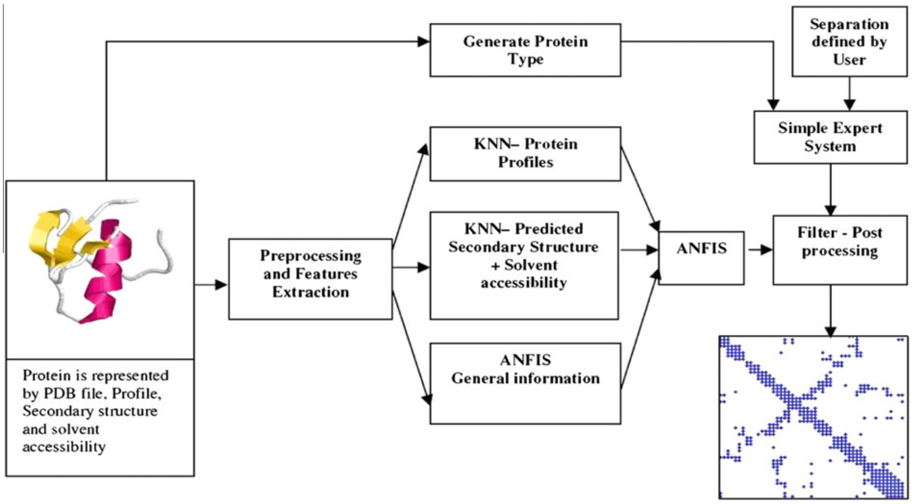

Generate Protein Type
Separation defined by User
Simple Expert System
KNN– Protein Profiles
KNN– Predicted Secondary Structure + Solvent accessibility
ANFIS General information
Preprocessing and Features Extraction
ANFIS
Filter - Post processing
Protein is represented by PDB file. Profile, Secondary structure and solvent accessibility

Fig. 2. JUSTcon model preview.

Table 1
Group-one training dataset and sample numbers.

Dataset	Contact samples	Non-contact samples
Group-one full dataset	182,014	8,954,458
Group-one training dataset	182,014	448,828
Profile KNN model dataset	165,123	407,883
Secondary structure and solvent accessibility KNN model dataset	165,132	406,125
General features ANFIS model	164,866	407,054

Table 2
Group-two training dataset.

Dataset	Contact samples	Non-contact samples
Group-two full training dataset	6182	246,371
Group-two training dataset	6182	12,346

dataset. Long proteins with sequences more 400 amino acids are also removed because they could compromise the computation process in term of processing time. All structures that were extracted using NMR method were removed from the list [27]. Furthermore, proteins that have broken chains were removed.

3.1. Generating training and testing datasets

The final dataset used in this work contains 500 proteins after eliminating all unwanted proteins from the original dataset. These proteins were divided into three groups: group-one training proteins, group-two training proteins, and test proteins. The size of each group dataset is 430, 20, and 50 proteins, respectively. For the full dataset and according to CASP criteria, the samples of amino acid pairs are defined as in contact or non-contact. The majority of samples were non-contact. To balance between contact and non-contact samples in the training datasets, all contact samples were used and only 5% of the non-contact samples were randomly selected and used.

3.1.1. Group-one training proteins

Group-one dataset is used to train the first layer of the system. Fig. 2 shows that the first layer contains three parts: Profile-KNN, the secondary structure and solvent accessibility (SSSA) KNN, and the first ANFIS (general information). In order to obtain more accurate results and to increase the generalization capability of the system during the training phase, each part in the first-layer of the system is trained by a dataset, which is 90% randomly, selected from group-one training dataset. Table 1 provides more details about each dataset and the associated number of contact and non-contact samples.

3.1.2. Group-two training proteins

Group-two training dataset is formed from 20 separate proteins. This dataset is generated with the same features and criteria that have been used for group-one training dataset. A 5% randomly selected non-contact samples are used with full contact samples as shown in Table 2. This dataset is used to select the best *K* values for the KNN models in the first-layer of the model In addition, it is used as a validation dataset while training the ANFIS in first layer. Furthermore, this dataset is used to select the best model parameters like window size, window count, and factorization numbers (NMF). The full dataset in group two training proteins is used to select the best parameters for the expert system and a suitable filter window size. Finally, it is used to select the best output pattern. For this training group, the accuracy is calculated in a similar way to that for a full protein. After sorting the predicted values, the accuracy is calculated by checking the best 6182 highly predicted items since the number of contact samples in the original set was 6182.

3.1.3. Testing proteins

The testing dataset consists of 50 proteins. All samples should be presented to the model to get full outputs of the testing proteins. The distribution of samples was 22,024 for contact and 1,187,738 for non-contact.

4. JUSTcon model

4.1. Model preview

Fig. 2 shows the architecture of the new protein contact map prediction model (JUSTcon). It consists of multiple parallel stages of neuro-fuzzy inference system [17,18] and KNN classifier [19].

Before using this model, one has to select the protein dataset, eliminate unwanted data, and preprocess it as mentioned previously. After the features are extracted, the KNN models are built, which are related to profiles, secondary structure [20,21], and solvent accessibility (SSSA). In parallel, An ANFIS model is trained by proteins and samples of amino acid pairs' general information. In the second layer, an ANFIS model works as a combiner that takes the outputs of KNNs and the previous ANFIS to provide an initial prediction of the contact map. Finally, a filtering operation is performed on the initial prediction, which is supervised by a simple expert system that determines the filter window size. This expert system uses the protein type and the needed sequence separation distance to ensure that the amino acid pairs are consistent and have a normal profile.

The difficulty of building a predictor system comes from the complexity of the protein structure, the variety of features sources, and the huge dataset. These factors drive us to adopt a modular approach that divides the system into parallel-multi-stage systems. Selection of a specific machine leaning system is related to the type of the features; the features that have fuzziness behavior like the general information of amino acid pair should be fed to ANFIS system. On the other hand, the features that could be recognized by the similarity are fed to a KNN system. Furthermore, the second stage is an ANFIS system that works as a merger of the first layer outputs because of the uncertainty in their behavior.

4.2. Features selection

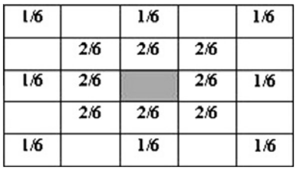
Some features have been identified and used by other researchers [9–11]. The same features are adopted in this work after some manipulation to fit with the machine-learning model. Some of these features are predicated by other servers like secondary structure and solvent accessibility, while the others are obtained from huge databases like PSI-Blast outputs [28,29]. Using external prediction servers to prepare the features was used in previous research like SVMcon [9] and PROFcon [11]. The features can be categorized as local window features and general information about amino acid pair.

4.2.1. Local window features

This feature set includes information from the protein profile, secondary structure predictions, and solvent accessibility predictions. Five windows of size nine amino acids are used for studying every amino acid pair. Two windows are centered at each amino acid of the studied pair and three windows are selected at equidistance from the segment between the amino acid pair.

Table 3
The affinity score matrix.

	A	*C*	*D*	*E*	*F*	*G*	*H*	*I*	*K*	*L*	*M*	*N*	*P*	*Q*	*R*	*S*	*T*	*V*	*W*	*Y*
A	0.16	0.24	0.06	0.06	0.21	0.12	0.11	0.23	0.07	0.20	0.18	0.08	0.10	0.09	0.08	0.10	0.13	0.23	0.16	0.17
C	0.24	1.00	0.13	0.10	0.43	0.24	0.27	0.37	0.14	0.38	0.37	0.17	0.14	0.19	0.17	0.25	0.23	0.40	0.55	0.38
D	0.06	0.13	0.03	0.01	0.08	0.08	0.13	0.06	0.08	0.05	0.05	0.09	0.07	0.04	0.10	0.08	0.07	0.07	0.06	0.09
E	0.06	0.10	0.01	0.00	0.06	0.05	0.07	0.08	0.07	0.05	0.06	0.04	0.05	0.03	0.09	0.06	0.08	0.00	0.08	0.08
F	0.21	0.43	0.08	0.06	0.30	0.14	0.17	0.32	0.08	0.28	0.28	0.10	0.13	0.11	0.12	0.16	0.17	0.31	0.25	0.24
G	0.12	0.24	0.08	0.05	0.14	0.15	0.12	0.12	0.06	0.11	0.12	0.10	0.10	0.10	0.09	0.11	0.13	0.14	0.13	0.15
H	0.11	0.27	0.13	0.07	0.17	0.12	0.21	0.16	0.04	0.13	0.16	0.10	0.11	0.08	0.10	0.12	0.13	0.16	0.18	0.17
I	0.23	0.37	0.06	0.08	0.32	0.12	0.16	0.44	0.09	0.36	0.29	0.09	0.11	0.10	0.13	0.12	0.19	0.43	0.23	0.27
K	0.07	0.14	0.08	0.07	0.08	0.06	0.04	0.09	0.01	0.07	0.05	0.05	0.04	0.04	0.02	0.07	0.07	0.09	0.08	0.11
L	0.20	0.38	0.05	0.05	0.23	0.11	0.13	0.36	0.07	0.33	0.24	0.07	0.10	0.09	0.10	0.11	0.15	0.35	0.27	0.26
M	0.18	0.37	0.05	0.06	0.28	0.12	0.16	0.29	0.05	0.24	0.21	0.10	0.12	0.10	0.09	0.12	0.16	0.25	0.23	0.23
N	0.08	0.17	0.09	0.04	0.10	0.10	0.10	0.09	0.05	0.07	0.10	0.11	0.09	0.07	0.08	0.10	0.12	0.09	0.12	0.11
P	0.10	0.14	0.07	0.05	0.13	0.10	0.11	0.11	0.04	0.10	0.12	0.09	0.10	0.09	0.08	0.09	0.10	0.12	0.15	0.16
Q	0.09	0.19	0.04	0.03	0.11	0.10	0.08	0.10	0.04	0.09	0.10	0.07	0.09	0.05	0.07	0.07	0.09	0.11	0.12	0.15
R	0.08	0.17	0.10	0.09	0.12	0.09	0.10	0.13	0.02	0.10	0.09	0.08	0.08	0.07	0.05	0.10	0.10	0.12	0.14	0.16
S	0.10	0.25	0.08	0.06	0.16	0.11	0.12	0.12	0.07	0.11	0.12	0.10	0.09	0.07	0.10	0.11	0.12	0.12	0.11	0.13
T	0.13	0.23	0.07	0.08	0.17	0.13	0.13	0.19	0.07	0.15	0.16	0.12	0.10	0.09	0.10	0.12	0.13	0.18	0.12	0.16
V	0.23	0.40	0.07	0.09	0.31	0.14	0.16	0.43	0.09	0.35	0.25	0.09	0.12	0.11	0.12	0.12	0.18	0.42	0.24	0.27
W	0.16	0.55	0.06	0.08	0.25	0.13	0.18	0.23	0.08	0.27	0.23	0.12	0.15	0.12	0.14	0.11	0.12	0.24	0.23	0.31
Y	0.17	0.38	0.09	0.08	0.24	0.15	0.17	0.27	0.11	0.26	0.23	0.11	0.16	0.15	0.16	0.13	0.16	0.27	0.31	0.26



1/6		1/6		1/6
	2/6	2/6	2/6	
1/6	2/6	▨	2/6	1/6
	2/6	2/6	2/6	
1/6		1/6		1/6

Fig. 3. Window filter with window size 2 and the original contact value between amino acid pair is in gray.

4.2.1.1. Protein profile. A protein profile is extracted from the PSI-Blast file. This file is based on the output of sequence alignments on sequences stored in multi-databases. EVA server provided us with the PSI-Blast files for every protein included on the dataset. The profile comprises a matrix of 21 rows (20 amino acids and 1 for gab) and the length of the protein as columns. Each cell in the profile represents the frequency of specific amino acid in specific location of the profile.

4.2.1.2. Secondary structure predictions. The secondary structures are predicted using PROFsec tool [30]. This tool uses the PSI-Blast output file as input. According to EVA, the accuracy of the PROFsec is about 76%, which can make a good improvement in contact map prediction. Every amino acid position in a protein chain is represented by three values: Alpha-helix, Beta sheet, and coil. It represents the probability of being at that amino acid.

4.2.1.3. Solvent accessibility predictions. Solvent accessibility is used to increase the prediction accuracy of the contact map using PROFacc [24], which can achieve 78% accuracy. Two values are used from the output of PROFacc: predicted relative solvent (range 0–90) and reliability index (range 0–9).

4.2.2. Amino acid pairs general information

For specific amino acid pair, the general information is represented by six inputs. The length of the protein plays an important role as well as the distance between amino acid pair. Affinity score, which is a statistical value, represents how much specific amino acid is most likely to be in contact with other amino acids. Finally, the average of probabilities of Alpha-helix, Beta-sheet, and coil between the amino acid pair are used as three inputs.

4.3. Features vectors preprocessing

In order to reduce the size of features vectors and extract important features, Non-negative Matrix Factorization (NMF) [31] is used. In the case of profile features, the size of features vector is 21 times the window size. For that, the profile features vector is reduced using NMF with a 25% factorization factor. On the other hand, the size of the features vector in the case of the secondary structure is three times the window size, while it is twice the window size in the case of the solvent accessibility. In both cases, NMF will represent the features vector by a new one with a same size and extract hidden and important features.

One of the most known tools for NMF algorithm is bioNMF. This tool contains a user-friendly graphical interface to interactively explore results and facilitate the data analysis process. Standard NMF implements the classical Lee and Seung [32] NMF algorithm, which was used in this work.

4.4. KNN model

KNN model is a simple approach that selects the nearest *K* samples from the training dataset to the testing sample vector. After selecting the closest *K* training samples, their outputs are summed then divided by *K*. Selecting the best *K* value is obtained by applying a different dataset that is not included in the training set of KNN model. After that, the *K* value that has the largest accuracy is used on testing dataset.

In both KNN models, after selecting the main parameters including the window size, window count, and factorization factor, a simple operation is performed to select all possible windows from the first group of training dataset. These windows are used to train the NMF model to produce the H matrix (Encoding vector), which is used to convert any sample window to a new one.

In the training stage of the NMF model, some preprocessing was done on the NMF training windows. First, each vector's mean is mapped onto 0 and deviation onto 1. Then, an offset value is added to avoid negative values, which is a precondition of the NMF. The same operations are performed on all samples. Therefore, the same mapping parameters and the same offsets will be used in any window sample from the training or testing datasets to give correct results when it is multiplied by the encoding vector.

For example, in the case of profile KNN, assuming the window size is five, and the factorization factor is 25%. The original window

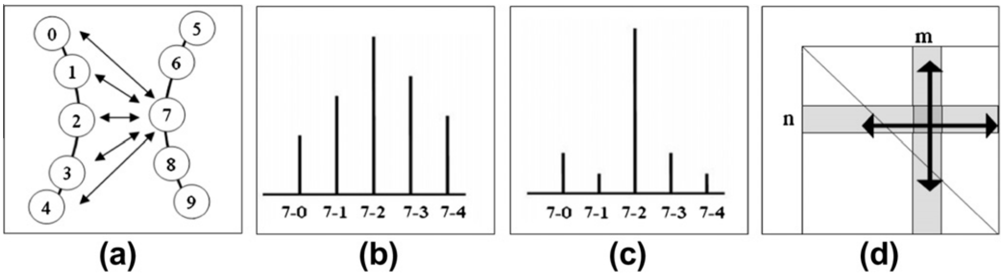


Fig. 4. (a) Behavior of connectivity between amino acid (7) and other amino acids (0, 1, 2, 3, and 4). Note that the numbers inside the cells represent the order in the protein sequence; the sequence starts from 0 then 1, etc. (b) Normal connectivity behavior between pairs; (7-2) has the largest connectivity, followed by (7-1) and (7-3), and so on. (c) Irregular connectivity behavior, which most likely does not represent a real contact, and therefore, it will be filtered out as a noisy output. (d) The windowing process, where for the case shown in figure (a) n = 0,1,2,3 and 4 (called vertical) and m = 5, 6, 7, 8 and 9 (called horizontal).

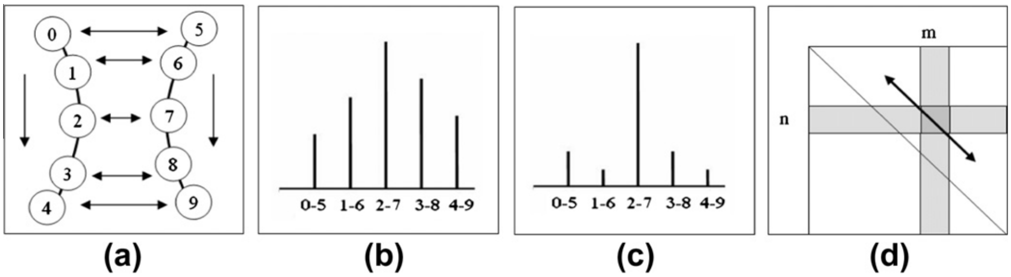


Fig. 5. Behavior of connectivity between parallel amino acid pairs. Sequence passes through amino acids in the same as their order. (b) and (c) Show the regular and irregular connectivity behaviors between amino acids, respectively. (d) The region of the window filer assuming n = 0, 1, 2, 3, and 4 and m = 5, 6, 7, 8, and 9.

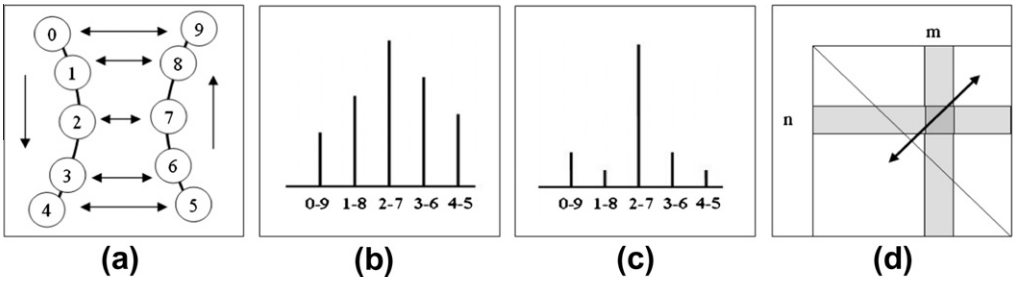


Fig. 6. Behavior of connectivity between anti-parallel amino acid pairs. (b) and (c) Show the regular and irregular connectivity behaviors between amino acids, respectively. (d) The region of window filer assuming n = 0, 1, 2, and 4 and m = 9, 8, 7, 6, and 5.

Table 4
Results for selecting the best K value for the Profile KNN and SSSA-KNN models using the best parameters.

	Model	Best (K) number	Accuracy (%)
Output Pattern 1	Profile KNN	4	70.55
Output Pattern 2		3	70.74
Output Pattern 3		5	70.25
Output Pattern 1	SSSA KNN	8	76.73
Output Pattern 2		6	76.70
Output Pattern 3		6	76.46

size (profile size) is 105 (21 × 5 = 105). After preprocessing, the encoding matrix of the profile NMF model will have a size of 105 × 26 (25% × 105 = 26). As a result, the profile NMF model will be trained using the all possible windows in the group-one training dataset to generate that Encoding matrix. Then, a protein with N × 105 matrix, where N is the number of samples in that protein and 105 is the original window size is extracted. After that, the generated matrix $N \times 105$ will be multiplied by the encoding matrix (105 × 26), a new reduced matrix will be generated with N × 26 size. In the SSSA-KNN model, the feature vector consists of predicted secondary structure and solvent accessibility as one combined vector.

4.5. General information ANFIS

In this stage, an ANFIS model is used to predict the contact for given samples. One of the feature factors is affinity score. This feature was calculated statistically from Group-one training dataset by building a 20 × 20 symmetrical matrix. This matrix represents how much amino acid X is most likely in contact with amino acid Y. The values of the cells in the matrix are calculated as follows:

$$Affinity\ (A,B) = Contact\ (A,B)/(Contact\ (A,B) + NonContact\ (A,B)) \quad (3)$$

where $Contact\ (A,B)$ is the total number of contacts between (A) and (B) in the group-one full training dataset and $NonContact\ (A,B)$ is the number of non-contacts between (A) and (B). The amino acid pair was calculated for pairs with separation distance ⩾6. Table 3 shows

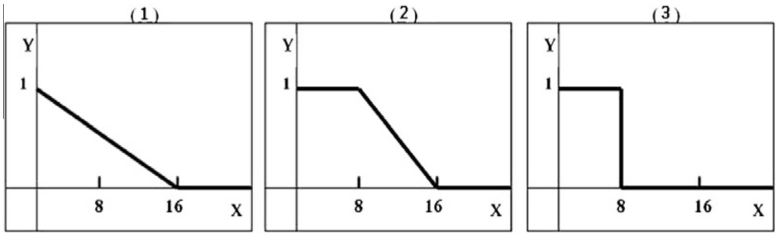


Fig. 7. The contact output patterns, (1) the contact value in pattern 1 is calculated proportional to separation physical distance between amino acid pair. (2) The contact value in pattern 2 is 1 for all distances from 0 to 8 A and it then linearly decreases to zero at 16 A distance. (3) The contact values in pattern 3 are Boolean with a separation point at 8 A; the contact for distances <8 A is 1 and 0 otherwise.

Table 5
The accuracy ($L/5$) when changing the filter window size with separation ⩾6, ⩾12 and ⩾24.

	Separation	FWS = 0 (%)	FWS = 1 (%)	FWS = 2 (%)	FWS = 3 (%)	FWS = 4 (%)	FWS = 5 (%)	Avg. (%)
Pattern 1	⩾6	42.83	43.35	41.69	43.94	44.58	43.79	43.36
Pattern 2		35.85	44.42	39.95	39.54	41.00	41.49	40.38
Pattern 3		44.98	41.90	40.07	40.82	41.71	41.36	41.81
Pattern 1	⩾12	37.89	42.69	41.72	41.78	40.68	41.12	40.98
Pattern 2		29.21	39.32	39.24	38.23	39.51	39.13	37.44
Pattern 3		36.63	40.99	41.28	39.92	40.34	39.74	39.82
Pattern 1	⩾24	26.10	31.16	33.35	32.26	33.70	32.93	31.58
Pattern 2		20.62	29.10	30.87	30.36	30.63	30.39	28.66
Pattern 3		27.34	30.36	31.00	31.31	31.44	31.43	30.48

Table 6
Shows the expert system rules.

Separation	Type 1: more alpha	Type 2: more beta	Type 3: other
⩾6	4	1	4
⩾ 12	4	2	1
⩾24	4	5	2

the generated affinity score from the group-one full training dataset after scaling it in the range from 0 to 1.

After generating the dataset for ANFIS-Layer one, fuzzy subtractive clustering with a cluster's center range of influence of 0.25 was used to determine the number of rules and antecedents membership functions [33,34]. GenFis2 function in Matlab@Mathworks was used for this purpose. Using this function, six fuzzy rules were generated. After that, the ANFIS model is trained using the ANFIS training dataset and group-two training dataset for validation.

4.6. The initial contact prediction ANFIS model

In this step, we train the ANFIS in layer-two, which works as a merger of the outputs from the previous layers and provides the initial prediction of the contacts. The training is done by using group-two training dataset and 20 randomly selected proteins from group-one training dataset. The contact samples were 11,597 and non-contact samples were 25,314. Similar to the previous ANFIS model, GenFis2 was used to generate suitable fuzzy rules that comprise three rules. Then, the group-two training dataset was used to train the ANFIS model.

4.7. Post processing – filter

The final step after generating the initial contact image is to filter the output. The filtering is performed using $N \times N$ special averaging windows. Fig. 3 shows an example of that filter with window size 2. The filtering operation is based on the idea that the contacting behavior of amino acid pairs, which are close to each other in a small window, could be generally consistent and have a normal profile. That is, if amino acid i is in contact with amino acid j, the connectivity should fall down or rise up gradually as we go away from j as can be seen in Figs. 4–6. In all cases, the filter output is calculated as in Eq. (4). In general, the filter role is to give extra weight to the predicted contact value of amino acid pair that is in contact and has regular connectivity behavior around it. On the contrary, it will decrease the weight of the predicted contact value of amino acid pair that has irregular connectivity pattern. Consequently, the sort operation will set the predictions of amino acid pairs that have regular pattern on the top of the list to enter accuracy and coverage analysis. The output of the filter is given by:

$$Output(Wj) = \sum_{i=-1*WS}^{i=WS\&i\neq0} (Contact\ (k,l) * ((WS - Abs(i) + 1)/(WS * (WS + 1))) \quad (4)$$

where Wj is the type of the window (1 horizontal, 2 vertical, 3 parallel connection, and 4 anti-parallel connection), WS is the window size, $Contact\,(k,l)$ is the value of contact map at specific index (k,l), k and l are calculated according to Wj and i, and Abs is the absolute value.

It is clear that the cases of parallel (Wj = 3) and anti-parallel (Wj = 4) will not occur in the same sample. Therefore, the filter selects the higher value in calculating the final output for that specific amino acid pair. As a result, the final output is given by (5):

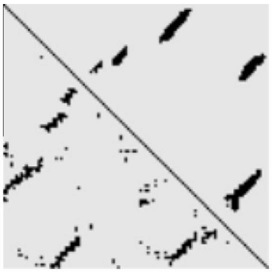
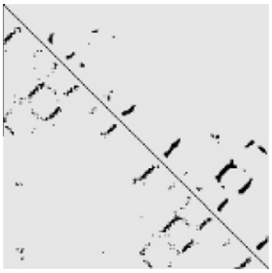
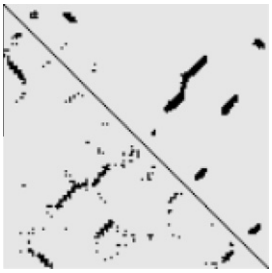
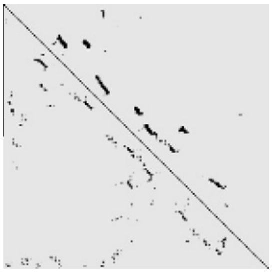
$$\begin{aligned} Filter\ final\ output\ (k0,l0) = (&Contact\ (k0,l0) + Output\ (W0) \\ &+ Output\ (W1) \\ &+ Max\ (Output\ (W2), Output\ (W3))/4. \end{aligned} \quad (5)$$

where $k0$, $l0$ are the indexes in the initial contact map.

4.8. Selecting the best parameters for the KNN Models

The parameters that play a major role in generating the KNN model are the window size, the window count, the NFM

Table 7
Samples of JUSTcon prediction outputs.

Protein info	Generated contact map image	Protein info	Generated contact map image
Name: 1c9h Chain: A Length: 107 Class: A+		Name: 1dn2 Chain: A Length: 207 Class: B	
Name: 1cuo Chain: A Length: 129 Class: B		Name: 1dxh Chain: A Length: 335 Class: A/B	

This table shows samples of JUSTcon prediction contact map images. The first column shows protein information and the second column shows the output images. It shows the top 2*L* contacts and the upper right side represents the predicted image and the lower left side represents the actual contact image. The class means the type of protein: A means all alpha helix, B means beta sheet, *A*/*B* means alpha helix alternating beta sheet, and *A* + *B* means alpha helix and beta sheet.

factorization percentage, and the *K* value. In order to select the best parameters, group-two training dataset was applied to both systems by changing these parameters. In addition, the value of *K* in the KNN model is varied from 1 to 46 to select the value that produces the best accuracy.

After studying the effect of changing the main parameters, the Profile KNN parameters were: window size = 9, window count = 5, and NMF factorization percentage = 25% and the parameters for the SSSA KNN model were: window size = 9 and window count = 5. The next step was selecting the best K values for both models as shown in Table 4.

The first column shows the contact value pattern according to Fig. 6, the second column shows the KNN model either for Profile or SSSA, the third column shows the best *K* value should be used, and the forth column shows the accuracy of testing that on the group-two data set

4.9. Contact output patterns

In this study, three output patterns were studied which represent the contact behavior between amino acid pairs. Fig. 7 shows the output patterns where *X*-axis represents the physical separation distance between amino acid pair in Angstrom and the *Y*-axis shows the contact value.

4.10. Selecting the best output pattern

In order to select the best output pattern to be used for the testing dataset, the full model is used on a protein dataset. This protein dataset is the same 20 proteins, which were used for generating the group-two training dataset. Using the best parameters that were found in last sections for each pattern, the accuracy of $(L/5)$ is calculated for each pattern with changing the filter window size to study the full behavior of each pattern. Changing the filter window size will result in different accuracies that are averaged for each pattern. Table 5 shows the results after applying the full model on the 20 proteins using different filter window sizes and output patterns for different separations.

This table shows the accuracy after applying the system with preferred parameters on the full mode of the group-two dataset. The first column shows the pattern different types, the second column shows the separation distance between amino acids pair, and the other columns show the accuracy when changing the filter window size.

It is clear from the results that the average accuracy of the first pattern is better than the other patterns. Therefore, pattern one will be used for the rest of this study for generating final proteins contact map images. There is no conflict between JUSTcon and CASP or other prediction systems since they use pattern three and we use pattern one. We use pattern one as a representation of the data to feed the JUSTcon model so we do not break the rule by using different pattern. However, in this work we use pattern three to calculate the accuracy and coverage of the final output according to CASP criteria.

4.11. Building the expert system

In order to select the best window size for the filter, an expert system is added in the final stage, which selects the suitable window size based on two factors: the protein general class and the minimum sequence separation distance.

To build an unbiased expert system, we will use the same 20 proteins dataset that were used before to generate group-two dataset. This protein dataset was applied on the full model in order to study the effects of the protein type and the minimum separation distance on selecting the best filter window size.

Table 6 shows the results after testing on group-two proteins. The rows show the window size according to a specific separation and the columns show the type of protein extracted from its prediction file, which was generated from PROFsec. In this experiment, the accuracy of $(L/5)$ was used to decide the best window size.

First column shows the separation distance, the first row shows the type of the protein, and the internal cells shows the preferred window size for the filter.

The protein type is determined by calculating the predicted values of secondary structure from the output of PROFsec. The

Table 8
Detailed prediction results for 50 testing proteins using different sequence separations and different number of selected residues pairs.

			Separation ⩾ 6					Separation ⩾ 12					Separation ⩾ 24				
Protein name + chain ID	SCOP type	Length	Total contact numbers	(%) Accuracy (2L)	% Coverage (2*L*)	% Accuracy (*L*/5)	% Coverage (*L*/5)	Total contact numbers	% Accuracy (2*L*)	% Coverage (2*L*)	% Accuracy (*L*/5)	% Coverage (*L*/5)	Total contact numbers	% Accuracy (2*L*)	% Coverage (2*L*)	% Accuracy (*L*/5)	% Coverage (*L*/5)
1dt0_a	Alpha	197	382	41.6	42.9	71.8	7.3	316	37.8	47.2	66.7	8.2	239	32.5	53.6	51.3	8.4
1efy_A	Alpha	350	776	5.9	5.3	11.4	1.0	723	5.1	5.0	2.9	0.3	617	6.0	6.8	1.4	0.2
1ew6_A	Alpha	137	164	12.8	21.3	14.8	2.4	138	14.6	29.0	7.4	1.5	120	14.2	32.5	44.4	10.0
1eyh_A	Alpha	144	217	20.5	27.2	41.4	5.5	181	13.2	21.0	31.0	5.0	145	9.7	19.3	17.2	3.5
1f06_A	Alpha	320	718	16.3	14.5	60.9	5.4	639	13.1	13.2	54.7	5.5	494	10.3	13.4	37.5	4.9
1f2u_B	Alpha	145	222	29.3	38.3	69.0	9.0	172	23.8	40.1	65.5	11.1	116	21.4	53.5	55.2	13.8
1f4o_A	Alpha	165	221	26.4	39.4	57.6	8.6	175	19.4	36.6	54.6	10.3	123	14.9	39.8	39.4	10.6
1fbv_A	Alpha	388	691	7.6	8.5	7.7	0.9	559	6.1	8.4	9.0	1.3	431	3.1	5.6	9.0	1.6
1fk0_A	Alpha	93	144	10.2	13.2	26.3	3.5	126	8.1	11.9	15.8	2.4	106	4.8	8.5	5.3	0.9
1ft5_A	Alpha	210	356	8.8	10.4	16.7	2.0	284	3.6	5.3	2.4	0.4	216	2.6	5.1	4.8	0.9
1fyz_E	Alpha	168	190	6.6	11.6	8.8	1.6	151	2.1	4.6	2.9	0.7	132	3.6	9.1	2.9	0.8
1hm6_A	Alpha	345	602	27.1	31.1	63.8	7.3	512	28.4	38.3	62.3	8.4	451	25.9	39.7	62.3	9.5
1hnw_D	Alpha	208	343	7.2	8.8	9.5	1.2	276	3.4	5.1	7.1	1.1	189	2.4	5.3	0.0	0.0
1i1r_A	Alpha	302	805	16.6	12.4	23.3	1.7	696	12.9	11.2	15.0	1.3	513	5.5	6.4	6.7	0.8
1i94_D	Alpha	208	351	8.2	9.7	14.3	1.7	285	6.0	8.8	7.1	1.1	195	2.4	5.1	4.8	1.0
1cvi_a	Alpha/Beta	342	716	12.4	11.9	32.4	3.1	631	8.3	9.0	20.6	2.2	535	4.4	5.6	4.4	0.6
1dxe_A	Alpha/Beta	253	571	30.0	26.6	47.1	4.2	486	29.1	30.3	58.8	6.2	358	15.2	21.5	49.0	7.0
1dxh_a	Alpha/Beta	335	791	46.0	38.9	85.1	7.2	700	44.9	43.0	77.6	7.4	607	39.7	43.8	77.6	8.6
1e6k_A	Alpha/Beta	130	244	40.0	42.6	88.5	9.4	203	40.4	51.7	88.5	11.3	136	33.1	63.2	76.9	14.7
1e6l_A	Alpha/Beta	127	235	47.2	51.1	100.0	10.6	203	46.1	57.6	100.0	12.3	144	39.4	69.4	100.0	17.4
1evi_A	Alpha/Beta	340	869	19.1	15.0	42.7	3.3	789	16.0	13.8	35.3	3.0	680	12.8	12.8	35.3	3.5
1f1m_A	Alpha/Beta	162	251	10.5	13.6	28.1	3.6	212	8.0	12.3	12.5	1.9	180	8.0	14.4	15.6	2.8
1f5o_A	Alpha/Beta	149	223	20.1	26.9	50.0	6.7	194	19.5	29.9	50.0	7.7	168	16.8	29.8	50.0	8.9
1faa_A	Alpha/Beta	120	232	52.1	53.9	95.8	9.9	201	49.2	58.7	95.8	11.4	158	45.0	68.4	95.8	14.6
1i0z_A	Alpha/Beta	333	759	50.6	44.4	83.6	7.4	655	50.0	50.8	74.6	7.6	532	44.0	55.1	89.6	11.3
1c9h_A	Alpha+Beta	107	267	66.4	53.2	95.2	7.5	230	63.6	59.1	95.2	8.7	180	58.9	70.0	95.2	11.1
1e8i_A	Alpha+Beta	117	276	49.2	41.7	73.9	6.2	216	42.7	46.3	65.2	6.9	171	38.9	53.2	65.2	8.8
1eoe_A	Alpha+Beta	100	152	40.0	52.6	70.0	9.2	109	33.0	60.6	70.0	12.8	81	27.5	67.9	70.0	17.3
1euv_A	Alpha+Beta	220	448	19.3	19.0	63.6	6.3	373	15.5	18.2	50.0	5.9	278	13.0	20.5	13.6	2.2
1ezv_F	Alpha+Beta	125	104	5.2	12.5	4.0	1.0	82	0.4	1.2	0.0	0.0	81	0.0	0.0	0.0	0.0
1fnt_K	Alpha+Beta	198	451	26.5	23.3	25.0	2.2	367	24.5	26.4	25.0	2.7	284	21.5	29.9	27.5	3.9
1fwk_A	Alpha+Beta	296	793	15.7	11.7	22.0	1.6	690	11.2	9.6	13.6	1.2	559	7.9	8.4	1.7	0.2
1g24_A	Alpha + Beta	211	457	13.5	12.5	31.0	2.8	402	10.9	11.4	26.2	2.7	351	7.8	9.4	7.1	0.9
1g62_A	Alpha + Beta	224	617	25.2	18.3	42.2	3.1	502	17.9	15.9	35.6	3.2	357	14.5	18.2	33.3	4.2
1hqm_E	Alpha + Beta	91	88	6.0	12.5	16.7	3.4	61	2.8	8.2	0.0	0.0	56	9.9	32.1	11.1	3.6
1io1_A	Alpha + Beta	395	855	4.8	4.4	6.3	0.6	720	2.7	2.9	1.3	0.1	560	0.6	0.9	0.0	0.0
1qi7_A	Alpha + Beta	253	572	35.2	31.1	58.8	5.2	477	32.4	34.4	64.7	6.9	359	26.5	37.3	64.7	9.2
1c3g_A	Beta	170	364	18.2	17.0	35.3	3.3	300	18.2	20.7	17.7	2.0	200	4.7	8.0	8.8	1.5
1cuo_A	Beta	129	332	62.0	48.2	92.3	7.2	290	60.9	54.1	92.3	8.3	241	58.9	63.1	92.3	10.0
1dn2_a	Beta	207	509	48.1	39.1	90.2	7.3	440	47.8	45.0	95.1	8.9	314	36.7	48.4	95.1	12.4
1dqi_A	Beta	124	303	22.2	18.2	32.0	2.6	248	16.9	16.9	24.0	2.4	174	4.8	6.9	24.0	3.5
1ds0_a	Beta	323	817	5.7	4.5	16.9	1.4	744	4.3	3.8	10.8	0.9	689	5.4	5.1	7.7	0.7
1eo2_B	Beta	238	547	10.3	9.0	31.3	2.7	473	7.8	7.8	14.6	1.5	399	7.1	8.5	10.4	1.3
1eqd_A	Beta	184	464	22.0	17.5	27.0	2.2	395	13.3	12.4	32.4	3.0	254	1.1	1.6	0.0	0.0
1ff5_A	Beta	219	563	20.3	15.8	31.8	2.5	486	15.1	13.6	18.2	1.7	364	5.9	7.1	0.0	0.0
1g43_A	Beta	160	451	11.9	8.4	3.1	0.2	407	11.3	8.9	3.1	0.3	335	5.9	5.7	3.1	0.3
1hwm_B	Beta	264	746	29.4	20.8	39.6	2.8	589	15.7	14.1	30.2	2.7	395	3.6	4.8	1.9	0.3
1i5i_A	Beta	174	457	20.4	15.5	31.4	2.4	392	14.7	13.0	11.4	1.0	299	5.8	6.7	0.0	0.0
1i94_Q	Beta	104	188	55.8	61.7	100.0	11.2	159	50.5	66.0	100.0	13.2	101	35.1	72.3	100.0	20.8
1df9_C	SMALL	70	130	51.4	55.4	85.7	9.2	106	45.0	59.4	85.7	11.3	66	30.7	65.2	64.3	13.6
AVG		**207.48**	**440.48**	**25.2**	**24.9**	**45.5**	**4.6**	**375.3**	**21.8**	**25.5**	**40.0**	**4.8**	**294.66**	**17.0**	**26.8**	**34.7**	**5.6**

This table shows the full prediction output of the testing dataset. First column shows the protein name as PDB code and its chain, the second column shows the SCOP type (Alpha, Beta, Alpha/Beta and Alpha + Beta), the third column shows the length of the protein, the other columns show the accuracy and the coverage of the prediction for different separation sequence (⩾6, ⩾12 and ⩾24) and different number of select contact pairs (2*L* and *L*/5).

Table 9
Accuracy and coverage comparison between JUSTcon, SVMcon and PROFcon with different separation sequences and different number of selected amino acids pairs.

Server	Separation	Accuracy				Coverage			
		2L (%)	L (%)	$L/2$ (%)	$L/5$ (%)	2L (%)	L (%)	$L/2$ (%)	$L/5$ (%)
JUSTCon	⩾6	25.2	32.1	38.6	45.5	24.9	15.9	9.6	4.6
SVMcon		23.1	30.9	38.1	46.8	22.9	15.5	9.6	4.8
PROFcon		20.1	25.5	30.5	38.0	20.2	13.0	7.9	4.0
JUSTCon	⩾12	21.8	29.2	34.8	40.0	25.5	17.2	10.4	4.8
SVMcon		17.9	24.0	30.0	36.6	20.6	14.1	8.9	4.4
PROFcon		15.8	20.2	24.5	30.7	18.9	12.1	7.5	3.8
JUSTCon	⩾24	17.0	23.0	28.7	34.7	26.8	18.3	11.7	5.6
SVMcon		13.6	18.4	23.6	32.0	20.2	14.0	9.3	5.1
PROFcon		11.9	15.3	18.7	23.1	18.8	12.0	7.4	3.6

Table 10
Accuracy comparison between JUSTcon, SVMcon, and PROFcon with different protein lengths.

Server	Length	Separation ⩾ 6		⩾12		⩾24	
		2L	$L/5$	2L	$L/5$	2L	$L/5$
JUSTCon	Short	34.7	62.1	31.2	58.0	26.4	56.9
SVMcon		24.4	48.9	18.5	40.9	15.6	35.3
PROFcon		19.9	37.0	15.9	29.2	14.7	28.5
JUSTCon	Middle	20.9	36.4	16.9	30.3	11.2	20.6
SVMcon		23.4	41.4	18.1	30.7	12.3	29.0
PROFcon		21.1	41.4	16.2	32.3	9.9	17.8
JUSTCon	Long	19.0	38.0	16.9	31.5	13.8	27.8
SVMcon		20.6	53.3	16.9	41.1	13.2	32.7
PROFcon		18.5	33.5	14.8	30.0	11.5	24.9

Table 11
Accuracy comparison between JUSTcon, SVM, and PROFcon with different protein types.

Server	Type	Separation ⩾ 6		⩾12		⩾24	
		2L	$L/5$	2L	$L/5$	2L	$L/5$
JUSTCon	Alpha	15.3	29.7	12.5	23.1	10.2	21.0
SVMcon		18.0	41.8	13.5	30.4	10.0	32.9
PROFcon		17.0	36.1	13.4	27.9	10.1	19.5
JUSTCon	Beta	27.2	44.3	23.0	37.5	14.6	28.6
SVMcon		28.5	41.7	23.5	33.7	16.9	28.1
PROFcon		22.6	37.8	17.6	33.2	11.0	19.7
JUSTCon	Alpha+Beta	25.6	42.4	21.4	37.2	18.9	32.5
SVMcon		21.9	47.7	15.1	32.9	12.4	26.5
PROFcon		18.9	34.9	14.1	22.4	11.8	15.8
JUSTCon	Alpha/Beta	34.3	70.5	32.1	67.1	26.9	62.3
SVMcon		22.8	45.8	22.0	56.3	17.0	45.4
PROFcon		22.9	46.1	18.9	40.8	16.3	41.7

PROFsec predicts the secondary structure type on specific amino acid including more alpha, more beta, and others. Therefore, for each protein, a summation was done for each type in all amino acid sequence, and then the most frequent type was used to represent the type of the protein.

5. Results and comparisons

5.1. Testing JUSTcon and comparison with other predictors

After tuning the JUSTcon full model, it was tested and compared with other predictor servers. The testing dataset consists of 50 proteins. All samples should be presented to the model because of the need to get full outputs of the testing proteins. The outputs of the test dataset after presentation to the model are used to construct the contact map of the protein as shown in Table 7. Table 8 shows all testing protein chains, protein length, SCOP type [5], the number of contact pairs, the accuracy, and the coverage for 2L and $L/5$ under sequence separations ⩾6, ⩾12, and ⩾24, respectively.

The performance of JUSTcon was compared with that of SVMcon [9] and PROFcon [11] in term of accuracy because they produced very good accuracy in CASP5 and CASP7. The comparisons will consider many factors to study the points of strengths and weaknesses of JUSTcon including the sequence separation distances (Separation ⩾6, 12, 24), the length of the proteins for three categories: Short (Length < 153), Middle (l53 < length < 276) and Long (276 < length < 400), and the SCOP type (Alpha, Beta, Alpha + Beta, Alpha/Beta).

Table 9 shows the accuracy and the coverage of the three predictors using different separation distance ⩾6, ⩾12 and ⩾24 and different number of select contact pairs (2L, L, $L/2$ and $L/5$). These results prove that JUSTcon outperforms SVMcon on the average by 10% and PROFcon on the average by 35%, especially, when the separation distance ⩾12 and 24. This is becoming more significant because CASP is focusing now on getting good accuracy for separation distance ⩾24.

The other set of experiments were conducted to study the behavior of JUSTcon while changing the type and the length of testing proteins. Table 10 shows the accuracy of the three predictors for separation distance ⩾6, ⩾12, and ⩾24 using different length types.

In addition, the effect of protein SCOP types is studied. Table 11 shows the accuracy of the three predictors for separation distances ⩾6, ⩾12, and ⩾24 using different protein SCOP types.

These results also prove that JUSTcon outperforms SVMcon and PROFcon on the average by 48.7% and 89.4%, respectively in term of predictions for short proteins and it has an acceptable accuracy in the cases of middle and long proteins. Furthermore, the comparison shows that JUSTcon outperforms SVMcon and PROFcon on the average by 34.8% and 58.5%, respectively for Alpha + Beta type and Alpha/Beta protein SCOP types and it has an acceptable accuracy for Alph and Beta types. According to some previous studies [10], generating a contact map using top L contacts with an accuracy >30% can be used to produce a 3D view of the protein with low resolution, which already has been achieved by JUSTcon.

5.2. Applying JUSTcon on CASP9 targets

JUSTcon was applied on the targets of CASP9 [12] to study its performance in comparison with the other state of the art predictors. The CASP9 evaluation criteria focuses on the contact between residues with sequence separation ⩾24. Similar to CASP9, the analysis was performed on FM and TBM/FM domains only so 28 domains were tested. According to [15] the average accuracy of the participant groups where 16.5%, 18% and 20% for $L/5$, $L/10$ and Top 5 respectively. The average accuracy of JUSTcon was 15%, 19% and 23% for $L/5$, $L/10$ and Top 5 respectively.

In the Fig. 8, the accuracy of the prediction per target is shown. We included the average prediction accuracy of CASP9 groups for ($L/5$). From the Fig. 8, JUSTcon exceeds the average accuracy of CASP9 on several targets.

The previous analysis shows that JUSTcon is in a good level from the other competitor servers and located on the top third. So we believe that some minor improvements will make JUSTcon a high competitor.

In the Fig. 9 the top two accurately predicated targets by JUSTcon (T0604-D1 and T0553-D1) were represented as contact map images. The upper right side represents the predicated contact image and the left down side represents the actual contact map.

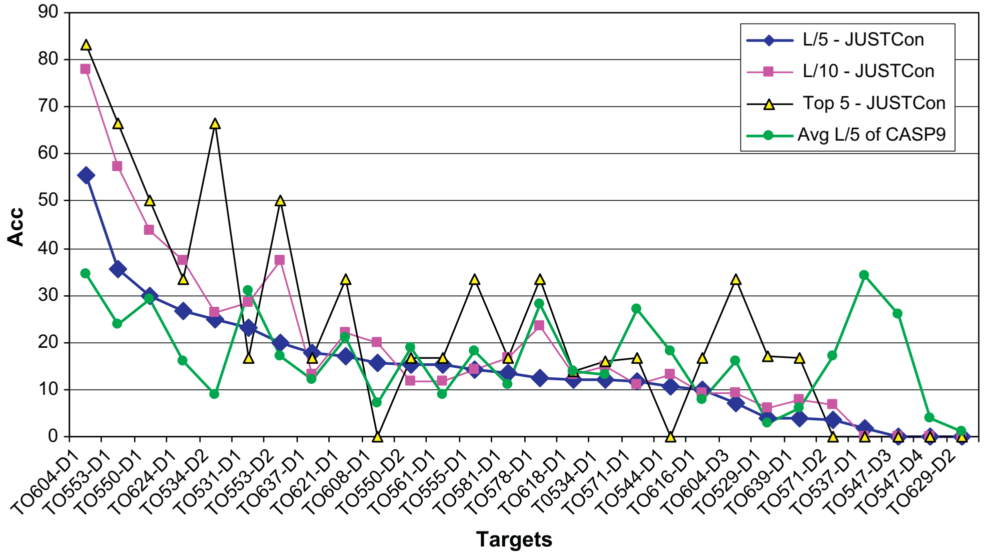


Fig. 8. Shows the accuracy per target of $L/5$, $L/10$ and top 5 of JUSTcon predictions and the $L/5$ average accuracy prediction of the participant groups in CASP9.

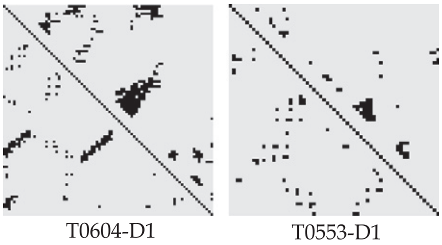


Fig. 9. Samples of JUSTcon prediction outputs for CASP9. The images show the real and predicted contact map of domains T0604-D1 (left) and T0553-D1 (right). The upper right side represents the predicted contact image and the lower left side represents the real contact image.

6. Conclusions

In this work, we have proposed a new machine-learning model (JUSTcon) for protein contact map prediction. The model is based on adaptive neuro-fuzzy inference system (ANFIS) and K nearest neighbors (KNN) algorithm. Our proposed model is novel in the domain for protein contact prediction in terms of its architecture and accuracy. The model has the ability to produce a set of amino-acid pairs predictions, which are more likely to be in contact.

The model uses many features that are extracted from the protein PDB files and other protein structure predictor. All of the proteins data are preprocessed and filtered to ensure correctness before the features are extracted. The features are then divided into three groups: profiles, secondary structure and solvent accessibility (SSSA), and general information. Each group is processed by its own phase that runs in parallel with the others. The first two groups, namely profiles and SSSA, are fed into their KNN models, while the amino acid pairs'general information features are processed by their own ANFIS model. The outputs from the KNN and ANFIS models are then passed to another ANFIS model in the second stage that works as a combiner and provides initial predictions. Finally, a filtering operation is applied on the final outputs and the window size of the filter is selected by a simple expert system. This expert system selects a suitable Filter window size based on the type of the protein and the targeted sequence separation distance.

Generated contact map using top L contacts with an accuracy >30% can be used to produce a 3D view of the protein with low resolution, which already has been already achieved by JUSTcon since the average accuracy is 32% for L and can reach 45.5% for $L/5$. On the other side, the comparison with other competitor servers proves that JUSTcon outperforms the other predictor on the average by 10% for SVMcon and 35% for PROFcon. In addition, these results also prove that JUSTcon outperforms SVMcon and PROFcon on the average by 48.7% and 89.4%, respectively in term of predictions for short proteins and it has an acceptable accuracy in the cases of middle and long proteins. Furthermore, the comparison shows that JUSTcon outperforms SVMcon and PROFcon on the average by 34.8% and 58.8%, respectively for Alpha + Beta type and Alpha/Beta protein SCOP types and it has an acceptable accuracy for Alpha and Beta types. In other side, JUSTcon produced an average accuracy of 15% for the sequence separation of 24 amino acids after applying it on CASP9 targets.

References

[1] Rost B. Protein structure prediction in 1D, 2D, and 3D. The Encyclopedia of Computational Chemistry; 1998. p. 2242–55.
[2] Gupta N, Mangal N, Biswas S. Evolution and similarity evaluation of protein structures in contact map space. Proteins: Struct, Funct, Bioinform 2005;59:196–204. doi:10.1002/prot.20415.
[3] Geerlof A et al. The impact of protein characterization in structural proteomics. Acta Crystallogr 2006. ISSN 0907-4449.
[4] M Tyszka J, E Fraser S, E Jacobs R. Magnetic resonance microscopy: recent advances and applications. Curr Opin Biotechnol 2005;16(1):93–9.
[5] Murzin Alexey G, Brenner Steven E, Hubbard Tim, Chothia Cyrus. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 1995:536–40.
[6] Vendruscolo M, Kussell E, Domany E. Recovery of protein structure from contact maps. Fold Des 1997;2:295–306.
[7] Aszodi A, Gradwell M, Taylor W. Global fold determination from a small number of distance restraints. J Mol Biol 1995;251:308–26.
[8] Skolnick J, Kolinski A, Ortiz. A: MONSSTER: a method for folding globular proteins with a small number of distance restraints. J Mol Biol 1997;265:217–41.
[9] Jianlin Cheng, Pierre Baldi. Improved residue contact prediction using support vector machines and a large feature set. BMC Bioinform 2007;8:113.
[10] Pollastri G, Baldi P. Prediction of contact maps by recurrent neural network architectures and hidden context propagation from all four cardinal corners. Bioinformatics 2002;18(Suppl. 1):S62–70.
[11] Marco Punta, Burkhard Rost. PROFcon: novel prediction of long-range contacts. Bioinformatics 2005:2960–8.
[12] http://predictioncenter.org/casp9/domain_definitions.cgi.
[13] Zhang Y. Progress and challenges in protein structure prediction. Curr Opin Struct Biol 2008;18(3):342–8.
[14] Alessandro Vullo, Ian Walsh, Gianluca Pollastri. A two-stage approach for improved prediction of residue contact maps. BMC Bioinform 2006;7:180.
[15] http://predictioncenter.org/casp9/doc/presentations/CASP9_RR.pdf.
[16] Gobel U, Sander C, Schneider R, Valencia A. Correlated mutations and residue contacts in proteins. Proteins 1994;18:309–17.

[17] Klir George J, Yuan Bo. Fuzzy sets and fuzzy logic: theory and applications. 1st ed. Prentice Hall; 1995.

[18] Jang J-SR, Sun C-T, Mizutani E. Neuro-Fuzzy Soft Comput 1998.

[19] Shakhnarovish, Darrell, Indyk. Nearest-neighbor methods in learning and vision: theory and practice. The MIT Press; 2005.

[20] Frishman D, Argos P. Knowledge-based protein secondary structure assignment. Proteins 1995;23(4):566–79.

[21] Pauling L, Corey RB, Branson HR. The structure of proteins: two hydrogen-bonded helical configurations of the polypeptide chain. Proc Natl Acad Sci Wash 1951;37:205.

[22] http://predictioncenter.org/.

[23] http://predictioncenter.org/casp7/meeting/presentations/Presentations_assessors/CASP7_RR_Clarke.pdf.

[24] http://cubic.bioc.columbia.edu/eva/

[25] Burkhard R, Eyrich VA. EVA: large-scale analysis of secondary structure prediction. Proteins: Struct, Funct, Genet 2001(Suppl. 5):192–9.

[26] Atomic Coordinate Entry Format Version 3.2. wwPDB, October 2008. <http://www.wwpdb.org/documentation/format32/v3.2.html>.

[27] Gorodkin J et al. Using sequence motifs for enhanced neural network prediction of protein distance constraints. Int Conf Intell Syst Mol Biol 1999:95–105.

[28] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol 1990;215(3):403–10.

[29] Altschul SF et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389–402.

[30] Rost B. PROF: predicting one-dimensional protein structure by profile based neural networks. <http://cubic.bioc.columbia.edu/pp/doc/methodsPP.html>.

[31] Pascual-Montano A, Carmona-Saez P, Chagoyen M, Tirado F, Carazo JM, Pascual-Marqui RD. bioNMF: a versatile tool for non-negative matrix factorization in biology. BMC Bioinform 2006;7:366.

[32] Lee DD, Seung HS. Learning the parts of objects by non-negative matrix factorization. Nature 1999;401:788–91.

[33] Chiu S. Fuzzy model identification based on cluster estimation. J Intell Fuzzy Syst 1994.

[34] Yager R, Filev D. Generation of fuzzy rules by mountain clustering. J Intell Fuzzy Syst 1994.