Short-Range Interactions and Decision Tree-Based Protein Contact Map Predictor

Cosme E. Santiesteban-Toca[1,⋆], Gualberto Asencio-Cortés[2], Alfonso E. Márquez-Chamorro[2], and Jesús S. Aguilar-Ruiz[2]

[1] Centro de Bioplantas, University of Ciego de Ávila, Cuba
cosme@bioplantas.cu
[2] University of Pablo de Olavide, Sevilla, Spain
aguilar@upo.es

Abstract. In this paper, we focus on protein contact map prediction, one of the most important intermediate steps of the protein folding problem. The objective of this research is to know how short-range interactions can contribute to a system based on decision trees to learn about the correlation among the covalent structures of a protein residues. We propose a solution to predict protein contact maps that combines the use of decision trees with a new input codification for short-range interactions. The method's performance was very satisfactory, improving the accuracy instead using all information of the protein sequence. For a globulin data set the method can predict contacts with a maximal accuracy of 43%. The presented predictive model illustrates that short-range interactions play the predominant role in determining protein structure.

Keywords: Protein structure prediction, protein contact map prediction, short-range interactions, decision trees.

1 Introduction

The protein structure prediction still being one of the greatest challenges of bioinformatics [1]. And, inter-residual contact maps is a critical step for the inter-residue contacts prediction problem. The ability to make successful predictions involves understanding the relationship between a sequence and its protein structure [2,3,4,5].

Multiple methods to predict contact maps have been developed. Based on *ab initio* approaches, in homology methods, fold recognition, template-based methods, machine learning, neural network and others [6,7,8,9,10,11,12,13,14]. The prediction quality of these methods has not been improved to satisfactory levels, despite of years of attempts. The main reason for this is perhaps that, it is hard to learn long-range dependencies on contact maps, hence it is especially difficult to predict contacts between residues that have large sequence separations. In addition, another important drawback of these methods is the insufficient capacity to explain their knowledge model for the protein's folding process understanding.

The traditional or *ab initio* folding method employs the principle of predicting protein structure from its known amino acid sequence $(a_0, a_1, \ldots, a_n)$, in

⋆ Corresponding author.

order to derive the 3D structure of proteins. We know that a protein chain folds spontaneously and leads to a unique three dimensional structure when placed in aqueous solution. The folding process cannot occur by random conformational search for the lowest energy state. Proteins must form the structure in a time-ordered sequence of events, now called a "pathway". The nature of these events, whether they are restricted to "native contacts" (defined as contacts that are retained in the final structure) or whether they might include non-specific interactions, such as a general collapse in size at the very beginning, were left unanswered [15].

In this paper we propose a solution to predict protein contact maps based on short-range interactions. Despite of some evidences of long-range interactions in stabilizing protein folding, the objective of this research is to know how short-range interactions can contribute to a system based on decision trees to learn the correlation among the covalent structures of a protein residues. Taking into account the high degree of flexibility and the simplicity of understanding of a solution based on decision trees, the proposed algorithm employs the Quinlan C4.5 method, according to previous papers [16,17].

This article is structured as follows. A methodology section, which explains the proteins data set selection criteria, the definition of contact maps, the proposed model architecture and the measures employed for the algorithm effectiveness. A results section, we show tabular and graphical experimentation results. Finally, the conclusions of this research.

2 Materials and Methods

2.1 Data Bases

To analyse the effect of short-range interactions on prediction, we use a set of non-homologous proteins of solved 3D structure. Initially, the set counts 2485 proteins with the lowest possible homology (less than 25% of identity), extracted from the Protein Data Bank (PDB) using PDB_select tool. This set is firstly reduced by excluding those proteins which has non-standard amino acid residues. They were excluded those chains whose backbone was broken. They were chosen only the chains whose: structure does not contain redundant sequences; without ligands, to eliminate false contacts due to the presence of hetero-atoms; and, those proteins that do not belong to the same family or have a common origin. Reducing the list to 173 proteins. This data set combines maximum coverage with minimum redundancy following the Fariselli criteria [6].

With the goal of comparing the proposed predictor with previous methods in the state of the art we employed 53 globulin protein sequences proposed by Zhang [2]. This is a set with a few homologous sequences extracted from PDB.

2.2 Contact Maps Definition

Contact maps are compactly 2D representation of 3D conformation of a protein in a symmetrical square matrix of pairwise inter-residue contacts. The calculation of the distances among the residues is determined by Euclidean distance.

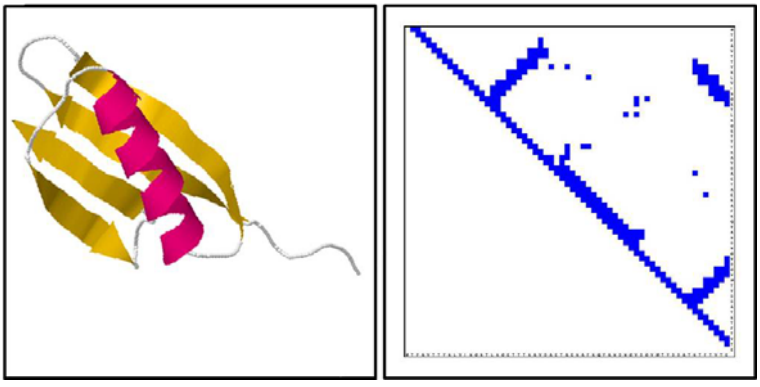

Fig. 1. Contact Map of 2igd protein, constructed with a threshold of 8Å. Left: 3D structure for protein. Right: its contact map showing parallel (top right cluster) and anti parallel sheets (top left and bottom right cluster), and helix features (thin cluster close to main diagonal).

The contact map of a protein (figure 1) is a particularly useful representation of protein structure. This representation provides useful information about the protein's structural motives and it also captures non-local interactions giving clues to its tertiary structure.

2.3 Model Architecture

Decision trees have been proved to be a successful method for prediction of contact maps of proteins [16,17]. Those classifiers make it possible to have understandable rules, which can be used to find further explanations of the data that are classified.

To predict contact map, we use an algorithm based on the Quinlan C4.5 decision tree [18], using the default setting. Our method builds decision trees for all possibles pairs of contacts, which has a total of 400 trees (20 x 20 amino acids). The prediction is treated as a classification problem, which takes into account the contacts or non-contacts between residues.

As input coding, the proposed method introduces the use of short-range interactions as a basis for training the predictor. Taking into account that oligopeptides are a few amino acids covalently joined (up to 10) and the average length of structural motives regions (up to 21), the algorithm employs vectors of length 21. This is equivalent to shift a window of length 21 by the amino acids chain. The built vector includes information of the substring formed among non adjacent amino acids. It is created a vector for each possible short-range interactions that can be formed in the protein (figure 2).

For a couple of amino A_1 A_2, the first 20 elements of the vector match the existing amino acids and contain their frequencies in the substring that is formed

Substring between the pair of amino acids				Class
A_1	A_2	...	A_{20}	Contact

Fig. 2. Scheme of input coding for decision trees. The first 20 bits in the coding, represent the frequency that appears the amino acids in the sub-chain. Where cero means that this amino acid is not present in the sub-chain. The last bit encodes by class (Contact or Not-Contact).

between the pair of amino acids analysed. To define the Class we adopt a threshold value of 8Å.

The decision tree-based predictor of protein contact maps (DTP) is shown in Figure 3. Given the distance matrix of a protein set with known structure (P_1, P_2, ..., P_n), the DTP builds a model of two-dimensional array of size NxN, where N is the number of amino acids (20). Each matrix cell contains a function $f_{(A_1,A_2,S)}$ formed by a decision tree, whose input vector is composed by the amino acids couple (A_1, A_2) and the information extracted from the substring (S) contained between them. For an unknown sequence (S?), each couples of amino acids is evaluated in the built model. The result of prediction is obtained by the occurrence of contact or non-contact.

2.4 The Pre-processing Procedure

Contact map prediction is an unbalanced problem. These maps contain, as average, a number of contacts (N_C) considerably lower than the number of non-contacts (N_{NC}) about 1/13. N_C increases almost linearly with protein sequence length (data not shown). For this reason N_{NC} increases with the square of the protein length.

The C4.5 decision trees. This algorithm is based on the data frequency and it is highly susceptible to the unbalance problem. To avoid the unbalanced effects we edit the data base applying an oversampling method. This method reproduces

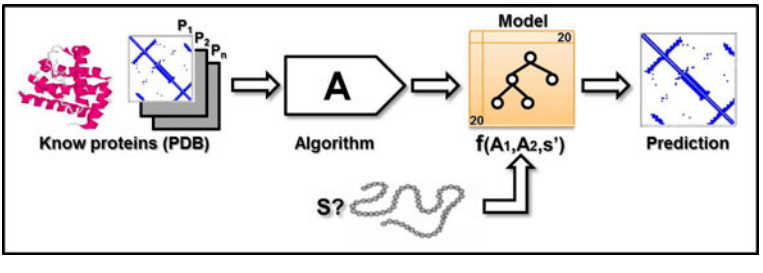

Fig. 3. Scheme of the decision tree-based predictor of protein contact maps. Where P_1 to P_n are the training proteins, A is the algorithm that creates the knowledge model and S? is the unknown sequence.

the minority class until mitigate the problem, taking into account the unbalance-ratio. This value is statistically calculated for each couple of amino acids in the protein. As result, the number of predicted contacts of a residue becomes a function of its structural environment.

2.5 Evaluation of the Efficiency

The effectiveness of prediction (Ap) is calculated as the ratio of true positives (1). This is because this equation penalizes non-contacts and prioritizes contacts.

$$Ap = TP/(TP + FN) \tag{1}$$

In order to compare the effectiveness of the predictor, an extra measure is applied: the improvement over a random predictor (2). This measure computes the ratio between Ap (1) and the accuracy of a random predictor (Nc / Np):

$$R = Ap/(Nc/Np) \tag{2}$$

where Nc is the number of real contacts in the protein of length Lp, and Np are all the possible contacts. In this paper in order to limit the prediction of local contacts (clustered along the main diagonal of the contact map) the proposed procedure does not include contacts between residues whose sequence separation is less than four residues.

3 Results

To study the influence of short-range interactions in the proteins, are analysed the distribution of protein contacts and structural motives with respect to the length of the sequence separation. We used the set of 173 proteins grouped into four classes, according to their sequences length (Ls): Ls $<$ 100 (65 proteins), 100 $\leq$ Ls $<$ 170 (57), 170 $\leq$ Ls $<$ 300 (30) and Ls $>$ 300 (21).

At first, with the aim of the study the distribution of inter-residual contacts, were analysed the frequencies of their appearance depending on the residues separation in the sequence (figure 4). It was used a thresholds range from 5Å to 12Å. It is obvious that most of contacts are concentrated in low sequences separation. Assuming a loss of 5% of contacts, the 95% is concentrated in sequence separations $\leq$ 150 and the 70% are concentrated just in residues with separation 10.

Another interesting analysis is to take into account the length by structural motives regions (helical and beta regions). We also studied the distribution of the number of residues per helical segment and per β-sheet segment (figure 5).

The fact that the length of β-regions in proteins is shorter than the helical segments is clearly shown in figure 5. Helical segment appears in regions from 3 to 20 amino acids and β-segment appears in regions from 2 to 10 amino acids. In average, the 80% of structural motives appears to be in the range of 2 to 10 amino acids.

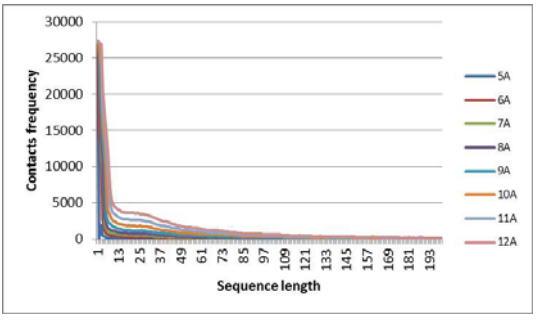

Fig. 4. Contacts distribution histogram. Plotting the contacts frequency as a function of sequence separation, for thresholds of 5Å to 12Å.

The distribution of contacts and structural motives, indicates that contacts in proteins are not randomly distributed and occur, predominantly, among residues with a low sequence separation.

To solve our specific problem, three methods are implemented:

- **DTP**: employs all information included in the protein sequences. The length of sub-sequences is not limited.
- **DTPsi**: method variation that employs as input coding only the short-interactions present. The input vector will be formed by the information of amino acids with maximal sequence separation up to 20.
- **DTPsi_ed**: it is the DTPsi method but we apply a pre-processing algorithm to the input data. Taking into account the unbalanced nature of present classes in this problem, we used an oversampling method to balance the database.

The implemented methods are tested on the selected database using a 10 folds cross-validation procedure. With the intention of highlighting the relationship

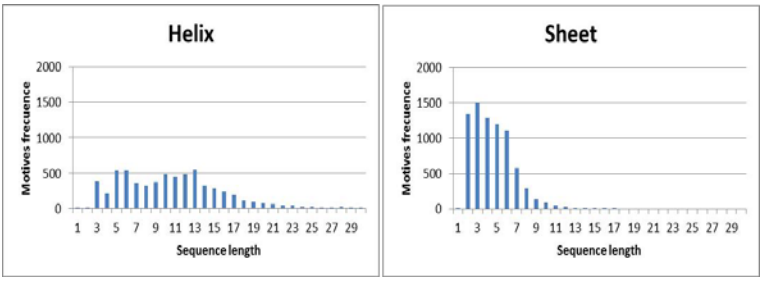

Fig. 5. Distribution of the number of residues per helical segment and per β-sheet segment

Table 1. Comparison of the performance of the different methods used to predict contact maps

	$Ls < 100_{(65)}$		$100 \leq Ls < 170_{(57)}$		$170 \leq Ls < 300_{(30)}$		$Ls \geq 300_{(21)}$	
	Ap	R	Ap	R	Ap	R	Ap	R
DTP	0,12	2,33	0,05	2,75	0,03	4,26	0,03	7,52
DTPsi	0,13	2,10	0,07	2,14	0,06	1,18	0,04	3,47
DTPsi_ed	0,18	1,71	0,14	1,61	0,13	1,38	0,12	1,49

between the results and the proteins size, the values of effectiveness were calculated after grouping proteins according to their sequence length (table 1).

The results show that, in general, for all proteins, the algorithm trained with short-range interactions (DTPsi) show a good behaviour. DTPsi not only improves the minimum efficiency threshold proposed by the DTP algorithm, when is applied an algorithm to balance the class (DTPsi_ed), it improves drastically the prediction effectiveness.

Figure 7 shows the effectiveness of predictions based on the proteins length, using different methods (DTP, DTPsi and DTPsi_ed). This graph shows that the effectiveness of the algorithm is dependent on the length of the protein. However, like the rest of algorithms, DTPsi_ed is more efficient to predict contacts in short sequences and it's efficiency decreases when the sequence length is incremented.

3.1 Comparison with the Previous Methods

To compare the accuracy of our algorithm with respect to the previous methods we used the set of 53 proteins. Here the protein sequences are grouped into four classes: Ls<100, 100≤Ls<200, 200≤Ls<300, Ls>300, according to their sequences' length (Ls). The proteins 1TTF, 1E88, 1NAR, 1BTJ_B and 1J7E_A, were used to test the trained algorithm. The proposed procedure does not include contacts between residues whose sequence separation is less than four, to avoid small ranges of false contacts.

The table 2 shows the comparative results for the algorithms: Occ (Occupancy method) [19], based on a filtered procedure, reached an accuracy about 26%; Net_75 method [20], it uses multiple sequence alignment as input for a classical feed-forward neural network trained with a standard back-propagation algorithm, reached the accuracy of about 28%; RBFNN method [2] uses a binary input encoding scheme with a radial-based function neural network optimized by a genetic algorithm, reached an accuracy of 32%; and DTPsi_ed, achieved the best accuracy: 43%.

Considering the relationship between the residue length and the average accuracy, our algorithm can improve the prediction performance dramatically. Except for sequence length less than 100 where there are not differences respect to RBFNN method.

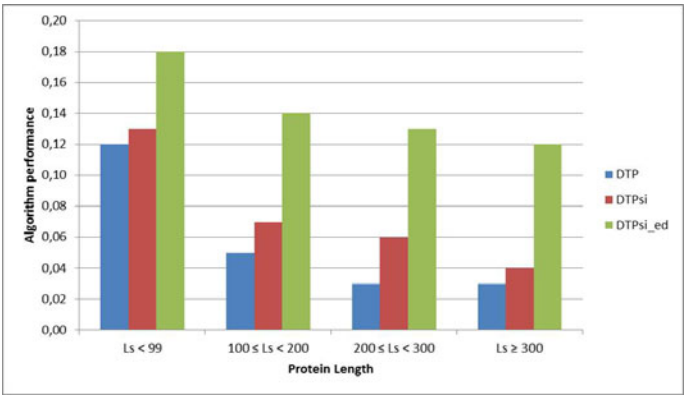

Fig. 6. This graph shows the efficiency of the contacts prediction based on the sequence lengths of proteins. In the x-axis values are represented the effectiveness achieved by the predictors, depending on the length of the sequences. The vertical axis represents the effectiveness values.

Table 2. Comparison of the predictors accuracy: Occ, Net_75, RBFNN and DTPsi_ed (our method). Ls is the length of the protein sequence. For this comparison it was employed the experimental results reported by Zhan[2].

Methods	Ls<100	100≤Ls<200	200≤Ls<300	Ls>300
Occ	0,26	0,21	0,15	0,10
Net75	0,26	0,28	0,21	0,20
RBFNN	0,30	0,31	0,32	0,28
DTPsi_ed	0,30	0,43	0,35	0,29

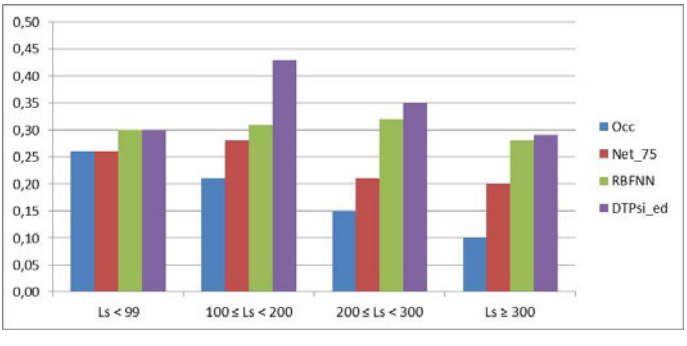

Fig. 7. This graph shows the comparative results in the prediction of contacts considering the sequence lengths of proteins. In the x-axis are represented the values effectively achieved by the predictors, depending on the length of the sequences. The vertical axis represents the effectiveness.

4 Conclusions

The presented predictive model illustrates how short-range interactions play a predominant role in determining protein structure. The proposed method combines the use of decision trees with a new input encoding for short-range interactions. The method performance was very satisfactory. It improves the accuracy with respect to the obtained by the DTP method. In a comparison with reported algorithm for a globulin data set, DTPsi_ed can predict contacts with a maximal accuracy of 43%.

Acknowledgements. This research is inserted in the doctoral program in Soft Computing, developed by the University of Las Villas in Cuba and the Andalusian Universities, under the sponsorship of the AUIP, which has promoted and apported the financial support to the entire program and research visits. Special thanks to Lic Nataniel Giménez Velázquez and Ernesto Estrada Cruz by their contributions.

References

1. Ouzounis, C.A., Valencia, A.: Early bioinformatics: the birth of a discipline a personal view. Bioinformatics 19(17), 2176–2190 (2003)
2. Quan, Z.H., Zhang, G.-Z., Huang, D.S.: Combining a binary input encoding scheme with RBFNN for globulin protein inter-residue contact map prediction. Pattern Recognition Letters 26, 1543–1553 (2005)
3. Glasgow, J., Kuo, T., Davies, J.: Protein structure from contact maps: A case-based reasoning approach. Inf. Sys. Front 8, 29–36 (2006)
4. Ramanathan, A.: Using Tensor Analysis to characterize Contact-map Dynamics of Proteins. PhD thesis, Carnegie Mellon University Pittsburgh, PA (2008)
5. Zhou, J., Arndt, D., Wishart, D.S., Lin, G., Shi, Y., Zhou, J., Arndt, D., Wishart, D.S., Lin, G.: Protein contact order prediction from primari sequences. BMC Bioinformatics 9(255), 1–21 (2008)
6. Fariselli, P., Olmea, O., Valencia, A., Casadio, R.: Prediction of contact maps with neural networks and correlated mutations. Protein Engineering 14(11), 835–843 (2001)
7. Pollastri, G., Baldi, P.: Prediction of contact maps by GIOHMMs and recurrent neural networks using lateral propagation from all four cardinal corners. Bioinformatics 18, 1–9 (2002)
8. Kim, H.: Computational analysis of hydrogen bonds in protein-RNA complexes for interaction patterns. FEBS Letters 552, 231–239 (2003)
9. Martin, A.J.M., Walsh, I., Bau, D.: Ab initio and template-based prediction of multi-class distance maps by two-dimensional recursive neural networks. BMC Structural Biology 9(5), 1–38 (2009)
10. Ahmad, M., Mathkour, H.: An integrated approach for protein structure prediction using artificial neural network. In: 2010 Second International Conference on Computer Engineering and Applications, pp. 484–488. IEEE (2010)
11. Sinha, S., Durga Bhavani, S., Suvarnavani, K.: Mining of protein contact maps for protein fold prediction. WIREs Data Mining Knowl. Discov. 1(4), 362–368 (2011)

12. Saraee, M., Korbekandi, H., Habibi, N.: Protein contact map prediction using committee machine approach. International Journal of Data Mining and Bioinformatics 2, 205–209 (2011)
13. Hossein, M., Narjes, S., Habibi, K.: Protein contact map prediction based on an ensemble learning method. In: 2009 International Conference on Computer Engineering and Technology 2009, vol. 2, pp. 205–209. IEEE (2009)
14. Min, H., Yoon, S., Kim, J., Kim, H.: Constructing accurate contact maps for hydroxyl-radical-cleavage-based high-throughput rna structure inference. IEEE Transactions on Biomedical Engineering 58(5), 1347–1355 (2011)
15. Shao, Y., Bystroff, C., Zaki, M.J., Hu, J., Shen, X.: Mining Protein Contact Maps. In: BIOKDD 2002: Workshop on Data Mining in Bioinformatics (with SIGKDD 2002 Conference), pp. 3–10 (2002)
16. Toca, C.E.S., Márquez Chamorro, A.E., Asencio Cortes, G., Aguilar Ruiz, J.S.: A Decision Tree-Based Method for Protein Contact Map Prediction. In: Giacobini, M. (ed.) EvoBIO 2011. LNCS, vol. 6623, pp. 153–158. Springer, Heidelberg (2011)
17. Santiesteban-Toca, C.E., Aguilar-Ruiz, J.S.: DTP: Decision Tree-Based Predictor of Protein Contact Map. In: Mehrotra, K.G., Mohan, C.K., Oh, J.C., Varshney, P.K., Ali, M. (eds.) IEA/AIE 2011, Part II. LNCS, vol. 6704, pp. 367–375. Springer, Heidelberg (2011)
18. Quinlan, J.R.: C4.5: Programs for Machine Learning. Morgan Kaufmann (1993)
19. Valencia, A., Olmea, O.: Improving contact predictions by the combination of correlated mutations and other sources of sequence information. Protein Engineering 2, S25–S32 (1997)
20. Casadio, R., Fariselli, P.: A neural network based predictor of residue contacts in proteins. Protein Engineering 12(1), 15–21 (1999)