A decision tree-based method for protein contact map prediction

Cosme E. Santiesteban Toca[1], Alfonso Márquez Chamorro[2], Gualberto Asencio Cortes[2], and Jesus 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. We describe a method where contact maps are predicted using decision tree-based model. The algorithm includes the subsequence information between the couple of analyzed amino acids. In order to evaluate the method generalization capabilities, we carry out an experiment using 173 non-homologous proteins of known structures. Our results indicate that the method can assign protein contacts with an average accuracy of 0.34, superior to the 0.25 obtained by the FNETCSS method. This shows that our algorithm improves the accuracy with respect to the methods compared, especially with the increase of protein length.
Keywords: protein structure prediction, protein contact map prediction, decision trees.

1 Introduction

One of the greatest challenges of bioinformatics is the protein structure prediction [1], and inter-residual contact maps is a critical step in this problem. The solution of inter-residue contacts prediction in proteins may be useful in protein-folding recognition. The secondary structure, fold topology and others patterns can be highlighted easily from a contact map. Similarly, the contact map information is used to predict unknown structures and proteins functions. The ability to make successful predictions involves understanding the relationship between a sequence and its protein structure [2–5].
In the last 20 years, multiple methods to predicting contact maps have been developed. Based on ab initio approaches, in homology methods, fold recognition, machine learning and others. [6–8].
In this paper we propose a solution based on decision trees, taking into account the high degree of flexibility and ease of understanding. Our algorithm uses the Quinlan C4.5 method [9], with the objective of know how a system based on decision trees can learn the correlation between residue covalent structure of a protein and its contact map, since it is calculated from their known 3D structure.
This article is structured as follows. A methodology section to explain the data

selection criteria, the proposed algorithm and its effectiveness measuring. A section for the experimentation results. Finally, the conclusions of this work.

2 Materials and methods

2.1 Data bases

With the goal of completing the training and validation process, we choose 173 proteins from protein data bank (PDB). This proteins are grouped into four classes, according to their length sequences (Ls): Ls $<$ 100 (65 proteins), 100 $\leq$ Ls $<$ 170 (57), 170 $\leq$ Ls $<$ 300 (30) and Ls $>$ 300 (21). This data set combines maximum coverage with minimum redundancy following the Fariselli criteria[6]. Were chosen only the chains: with lowest homology possible (less than 25% of identity); 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, to evaluate the generalization capability of the predictor. Was excluded those chains whose backbone was broken. The proposed procedure does not include contacts between residues whose sequence separation is less than four residues, to avoid small ranges of false contacts.

2.2 Contact maps definition

An alternative view of the protein uses of a distance matrix (Figure 1), a symmetric square *NxN* matrix whose elements are the Euclidean distances among the atoms in the protein. This representation is obviously redundant, it requires only *N(N-1)/2* degrees of freedom.

2.3 Model architecture

In the proposed method we use distances matrix as a basis for training the predictor. It is to avoid the lost of information in the binaries contact maps. In our method, the prediction is treated as a classification problem, which takes into account the contacts, quasi-contacts or non-contacts between residues.
Moreover, decision trees are classifiers which make it possible to have understandable rules, which can be used to find further explanations of the data that are classified. The proposed algorithm is based on the C4.5 decision tree, using the default setting [9]. We build decision trees for all possibles pairs of contacts, which has a total of 400 trees (20 x 20 amino acids).
As input coding, the algorithm uses vectors of length 23, which includes information of the substring formed between non adjacent amino acids, the distance and the sequence separation between the couple of analyzed amino acids (Figure

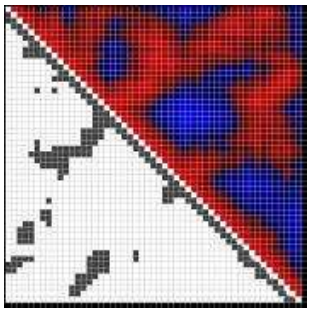

Fig. 1. Contact Map of 4sgb-I protein. In the lower left, the binary contact map constructed with a threshold of 8Å. In the upper right, the distance matrix, which is independent of the selected threshold.

Sub sequence information							A_1, A_2 information		
A	C	G	F	P	- - -	Y	Ls	D	Class
0	3	1	0	1	- - -	0	15	6,3	Contact

Fig. 2. Scheme of coding input for decision trees. It is formed by the sub sequence and the amino acids couple information.

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 between the pair of amino acids analyzed. Ls represents the length of the substring or the separation between the pair of amino acids. D, the Euclidean distance between the amino acids couple and Class is the discretization of the distance (D) in contact, quasi-contact and non-contact, depending on the established thresholds in Angstroms (Å). For this model we classified the contacts using this criteria: contact $\leq$ 8Å$<$ quasi-contact $\leq$ 12Å$<$ non-contact.

The decision tree-based predictor of protein contact maps (DTP) is shown in Figure 3. Given the distance matrix of a proteins 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 $\boldsymbol{f}(A_1,A_2,S)$ 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 builded model. The result of prediction is obtained by the occurrence of contact, quasi-contact or non-contact.

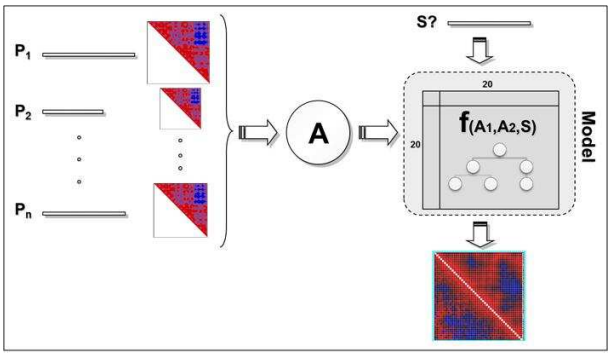

Fig. 3. Scheme of the decision tree-based predictor of protein contact maps, where P_1 to P_n are the training proteins, A is the algorithm that creates the knowledge model and S? is the unknow sequence. The knoledged model is formed by a matrix of functions f(A_1,A_2,S), where it is introduced the sub sequence S for the amino acids A_1 and A_2. As result, this algorithm returns the predicted distances matrix for the sequence S?.

2.4 Evaluation of the efficiency

To evaluate the effectiveness of the predictor, the problem was reduced to contact and non-contacts classes. Taking into account the unbalanced nature of the problem, was employed the precision *Ap=TP/(TP+FP)* as performance measure. In order to compare the effectiveness of the predictor, two extra measures are implemented. "The improvement over a random predictor (R)" and "the distribution of the predicted distances (Xd)".

R, computes the ratio between Ap and the accuracy of a random predictor *Ar=Nc/Np*. Where *Nc* is the number of real contacts in the protein of length *Lp*, and *Np* are all the possible contacts. In order to limit the prediction of local contacts, we considering 4 as the minimum length of sequence separation between residues. Computing *Np* as *(Lp-4)(Lp-3)/2*.

Xd, measures the difference in the distribution of the inter-residue distances in the 3D structure for predicted pairs compared with all pair distances in the structure. This index is defined by the equation $Xd=\Sigma_{i=1}^{n}(Pic\text{-}Pia)/n \cdot di$. Where n is the number of bins of the distance distribution (15 equally distributed bins from 4 to 60 cluster all the possible distances of residue pairs observed in the protein structure); di is the upper limit (normalized to 60) for each bin; Pic and Pia are the percentage of predicted contact pairs (with distance between di and *di-1*) and that of all possible pairs, respectively. By definition, values of Xd=0 indicate no separation between the two distance populations, meaning that the predicted contacts are randomly distributed; values of Xd>0 indicate positive cases, when the population of the distances between predicted contact pairs is shifted to smaller values with respect to the population of the distances of all residue pairs in the protein. For contact distances with an upper limit of 8Å, the larger and positive Xd is, more efficient the prediction of contacts is. Similarly

Algorithms	$All_{(173)}$			$Ls < 100_{(65)}$			$100 \leq Ls < 170_{(57)}$			$170 \leq Ls < 300_{(30)}$			$Ls \geq 300_{(21)}$		
	Ap	R	Xd	Ap	R	Xd	Ap	R	Xd	Ap	R	Xd	Ap	R	Xd
DTP	0,34	10,96	64,07	0,27	4,71	12,3	0,31	8,99	40,55	0,31	13,72	92,22	0,36	24,86	174,32
FNETCSS	0,25	8,05	11,87	0,33	6,28	12,91	0,25	7,33	12,14	0,19	9,47	10,78	0,15	12,71	9,77
MDS	0,23	5,17	-	0,17	2,61	-	0,19	5,28	-	0,15	6,82	-	0,15	9,54	-

Table 1. Protein data set. The proteins identity value is less than 25% and the sequence length (L_s) is equivalent to the number of structure covalent residues.

to the other two indexes, Xd is also averaged on the protein sets [6].

3 Results

In order to determine the efficiency of the predictor implemented, a 10 fold cross-validation method was performed. The results were compared with those obtained by Fariselli and Casadio (FNETCSS) [6] and a simple predictor based on the average of the distances and the sequences separation length (MDS) (Table 1).

With the aim of highlighting the results dependence to the proteins size, the effectiveness values were calculated after grouping proteins according to their sequence length.

The results show that, in general, for all proteins, the proposed algorithm (DTP) shows good behavior. DTP not only improves the minimum efficiency threshold proposed by the MDS algorithm, but except for less than 100 amino acid protein that has a lower performance than FNETCSS, DTP is visibly superior.

Figure 4 shows the effectiveness of predictions based on the proteins length, using different methods (DTP, FNETCSS and MDS). This graph shows that the effectiveness of the algorithm is dependent on the length of the protein. However, unlike FNETCSS, which is more efficient to predict contacts in short sequences, the proposed DTP method is more efficient to predict in large sequences, even when the density of contacts is much lower.

4 Conclusions

This work clearly demonstrates that the decision trees are efficient tools for protein contact maps prediction. The proposed method combines the use of decision trees with a newest input codification for all possible pairs of amino acids that were formed in the training dataset. The method performance was very satisfactory (0.34), especially with the increase of protein length, greatly enhancing the accuracy with respect to the 0.25 obtained by the FNETCSS method.

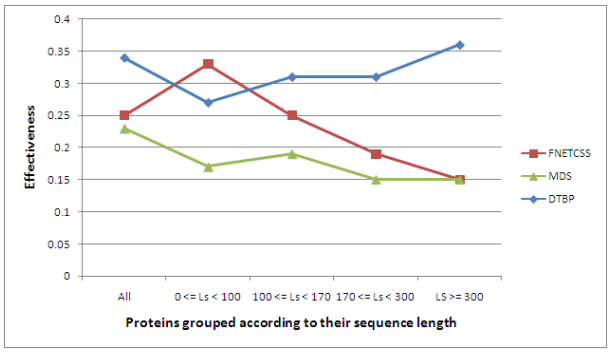

Fig. 4. This graph shows the efficiency of the prediction of contacts based on the sequence lengths of proteins.

5 Acknowledgements

This research is inserted in the doctoral program in Soft Computing, under the sponsorship of the AUIP.

References

1. Christos A Ouzounis and Alfonso Valencia. Early bioinformatics : the birth of a discipline a personal view. *Bioinformatics*, 19 (17):2176–2190, 2003.
2. Arvind Ramanathan. *Using Tensor Analysis to characterize Contact-map Dynamics of Proteins.* PhD thesis, Carnegie Mellon University Pittsburgh, PA, 2008.
3. Jianjun Zhou, David Arndt, David S Wishart, Guohui Lin, Yi Shi, Jianjun Zhou, David Arndt, David S Wishart, and Guohui Lin. Protein contact order prediction from primari sequences. *BMC Bioinformatics*, 9(255):1–21, 2008.
4. Guang-zheng Zhang and Kyungsook Han. Hepatitis C virus contact map prediction based on binary encoding strategy. *Computational Boiology and Chemistry*, 31:233–238, 2007.
5. Janice Glasgow, Tony Kuo, and Jim Davies. Protein structure from contact maps : A case-based reasoning approach. *Inf Sys Front*, 8:29–36, 2006.
6. Piero Fariselli, Osvaldo Olmea, Alfonso Valencia, and Rita Casadio. Prediction of contact maps with neural networks and correlated mutations. *Protein Engineering*, 14(11):835–843, 2001.
7. G Pollastri and P Baldi. Prediction of contact maps by GIOHMMs and recurrent neural networks using lateral propagation from all four cardinal corners. *Bioinformatics*, 18:1–9, 2002.
8. H Kim. Computational analysis of hydrogen bonds in protein-RNA complexes for interaction patterns. *FEBS Letters*, 552:231–239, 2003.
9. J. R Quinlan. C4.5: Programs for Machine Learning. *Morgan Kaufmann*, 1993.