Residue-residue Contact Prediction based on Evolutionary Computation

Alfonso E. Márquez Chamorro, Federico Divina, Jesús S. Aguilar-Ruiz and Gualberto Asencio Cortés

Abstract In this study, a novel residue-residue contacts prediction approach based on evolutionary computation is presented. The prediction is based on four amino acids properties. In particular, we consider the hydrophobicity, the polarity, the charge and residues size. The prediction model consists of a set of rules that identifies contacts between amino acids.

1 Introduction

The problem of Protein Structure Prediction (PSP) is one of the grand challenges in Structural Bioinformatics. A protein can perform several functions, e.g., transport function, enzymatic function, structural function, etc., and its three dimensional structure determines its biological functions. The knowledge of these structures has a great importance in medical and biological areas. For instance, recent studies have demonstrated the relationship between protein missfolding and diseases such as Cystic fibrosis and Emphysema. Some methods, such as nuclear magnetic resonance (NMR) and X-ray crystallography, can determine the structure of a protein. However, such techniques are both slow and expensive. Thus, an alternative method is needed, and soft computing can provide processing capabilities in order to solve this problem.

In any computing methods, a representation of the data is needed. A particularly useful representation of the tertiary structure of a protein is provided by contact maps. A protein with an amino acid sequence of length N, can be represented by using a symmetric matrix C of size NxN. Each entry C_{ij} is equal to either 0 or 1, depending on whether or not there is a contact between amino acids i and j. Two amino acids in a protein are in contact if the distance between them is less

Alfonso Márquez Chamorro, Federico Divina, Jesús S. Aguilar-Ruiz, Gualberto Asencio Cortés
School of Engineering, Pablo de Olavide University of Sevilla, Spain, e-mail: {amarcha,fdivina,aguilar,guaasecor}@upo.es

or equal than a given threshold usually expressed in Angstroms (Å). Researching methods used in this problem are focused on determining contact maps (distances) between amino acid residues of a protein sequence. When a contact map is defined, proteins can be folded and tertiary structures are obtained. This could be done using approximation algorithms.

Several contact map prediction methods have been applied to the PSP problem, e.g., artificial neural networks (ANNs) [1], support vector machines [2], evolutionary computation [3] and template-based modelling [4]. In this paper, we propose a method to predict residue-residue contacts from sequences of amino acids based on an evolutionary algorithm (EA). The main motivation for the use of an EA, is that PSP can be seen as a search problem, where the search space is represented by all the possible folding rules. Such search space is highly complex and has huge dimensions, and in this cases, EAs have proven to perform well. The prediction model will consist of rules that predict the contact between two residues. The prediction is based on four physical-chemical properties of the amino acids described in the following. Previously, EAs have been applied to PSP, e.g., HP model and lattice model were employed in [5]. A contact map model generator was included in [3].

The rest of paper is organized as follow: in section 2, we discuss our proposal to predict protein contact maps. Section 3 provides the experimentation and the obtained results. Finally, we draw some conclusions and discuss future works.

2 Methodology

Our experimental procedure is explained as follows. We first obtain a protein data set from the Protein Data Bank (PDB) (*http://www.wwpdb.org*). This data set will be used by our EA in order to obtain a set of rules for predicting the contact between two amino acids. From these rules, we can obtain a protein contact map which will be used in order to evaluate the accuracy of the prediction.

We have selected four properties, which will be used for the prediction: hydrophobicity, polarity, charge and residue size, which have been shown to have certain relevance in PSP.We use Kyte-dolitle hydropathy profile for the hydrophobicity [6], the Grantham's profile [7] for polarity and Klein's scale for net charge [8]. The Dawson's scale [9] is employed to determine the size of the residues. A contact treshold was established at 8 Å, as in [1].

In our approach, each individual represents a rule for a residue-residue contact. Each individual represents the four properties of amino acids in two windows of size 3 that encodes the amino positions $i-1, i, i+1$ and $j-1, j, j+1$ of a protein sequence, where i and j are two possible amino acids in contact. The values of the properties are normalized to a range of between -1 and 1 for hydrophobicity and polarity, and between 0 and 1 for the residue size. Three values are used to represent the net charge of a residue: -1 (negative charge), 0 (neutral charge) and 1 (positive charge).

The fitness of an individual I is given by the F-measure: $F(I) = 2 \cdot \frac{Recall \cdot Precision}{Recall + Precision}$. The higher the fitness, the better the individual. Recall represents the proportion of training examples that matches this rule. Each one of these examples represent a true contact between i and j amino acids. Precision represents the error rate. Moreover, we also consider some physical-chemical properties (hydrophobicity, polarity and charge) information of the amino acids. If two amino acids are in contact, they probably have similar conditions of hydrophobicity and polarity. On the other hand, they may have opposite charges [3]. We increase the fitness for an individual that fulfills these requirements.

Individuals are selected with a tournament of size two. One-point crossover is always applied to selected individuals, while mutation is applied with a probability of 0.5. If mutation is applied to a gene relative to the charge of the amino acid, then its value is randomly changed to one of the other two allowed possibilities. In the other cases, the values of the property is increased or decreased by 0.1. After this process, the validity of the individual is checked, and if the individual is not valid, the applied mutation is discarded. Elitism is also applied. The initial population consists of 100 individuals randomly initialized. The maximum number of generations is set to 100. However, if the fitness of the best individual does not increase over twenty generations, the algorithm is stopped. At the end, we select the best subset of rules from the final population according to their F-measure.

3 Experiments

As already stated, the data set was selected from PDB. In particular, we used the PDB Advanced Search Select. $12,830$ non-homologous and non-redundant protein sequences were extracted with a sequence identity lower than or equal to 30%. The list of PDB protein identifiers can be downloaded at *http://www.upo.es/eps/marquez/proteins.txt*. We have randomly selected a subset of 200 protein sequences from these $12,830$ proteins, with a maximum length of 318 residues. As validation method we have used a 10-fold cross-validation. Four statistical measures were calculated to evaluate the accuracy of our algorithm: Recall, Precision, Specificity and Accuracy:

- Recall represents the percentage of correctly identified positive cases. In our case, Recall indicates what percentage of contacts have been correctly identified.
- Precision is a measure to evaluate the false positive rate. Precision reflects the number of real predicted examples.
- Specificity, of True Negative Rate, measures the percentage of correctly identified negative cases. In this case, Specificity reflects what percentage of non-contacts have been correctly identified.
- Accuracy, represents the percentage of both true positives and true negatives cases over the total of the population.

Table 1 Average results and standard deviation obtained for different number of executions of the algorithm.

Runs	$Recall_{\mu\pm\sigma}$	$Spec._{\mu\pm\sigma}$	$Prec._{\mu\pm\sigma}$	$Accuracy_{\mu\pm\sigma}$
100	$0.036_{\pm 0.289}$	$0.989_{\pm 0.010}$	$0.558_{\pm 0.023}$	$0.993_{\pm 0.008}$
500	$0.181_{\pm 0.115}$	$0.992_{\pm 0.000}$	$0.522_{\pm 0.022}$	$0.994_{\pm 0.001}$
1000	$0.289_{\pm 0.092}$	$0.994_{\pm 0.000}$	$0.515_{\pm 0.031}$	$0.994_{\pm 0.001}$
2000	$0.605_{\pm 0.084}$	$0.993_{\pm 0.000}$	$0.506_{\pm 0.037}$	$0.993_{\pm 0.001}$

Results are provided in table 1. The optimal and exact number of rules is unknown. For this reason, we have varied the numbers of runs of the EA, where to a higher number of runs correponds a higher number of rules. The aim of this was to test whether or not a higher number of rules would provide better results. We show the results for 100, 500, 1,000 and 2,000 runs. For each run, a subset of rules with the best F-measure value is selected. So, for instance, for 1,000 runs we have finally obtained 2,348 rules. The set of rules provided is checked in order to eliminate repeated or redundant rules.

It can be noticed that as the number of rules increases, the recall increase. However this is reflected in a decrement of the precision. This result was quite expected, since by covering more cases, the possibility of errors increases. Therefore, we have obtained a low recall rate for 100 runs, and a maximum rate of 60% for 2,000 runs. Satisfactory levels of specificity are obtained in all cases, reaching values higher than 98%. Accuracy is also always very high, and this reflects the effectiveness of the prediction provided by the EA.

However the precision obtained always remain above 50%. Other methods for PSP, set the precision rate for a contact map prediction at about 30%. This result shows that the precision obtained by the proposed EA improves on this by more than 20%. Specificity and accuracy are always very high, and this reflects the effectiveness of the prediction provided by the EA.

An example of a resulting rule is showed in Figure 1. Each position represents a value for a different property as explained before and encodes a feature of a possible amino acid. For instance, the hydrophobicity value for the amino acid *i* is between 0.52 and 0.92, the polarity value between -1.0 and -0.93, neutral charge (0.0), and a residue size between 0.77 and 0.97. Therefore, the amino acid *i* could be L (Lysine) or F (Phenylalanine) which fulfills all these features according to the cited scales.

$$\underbrace{\begin{array}{|c|c|c|c|c|c|c|}\hline -0.39 & -0.19 & -0.78 & -0.68 & 0.00 & 0.83 & 1.03 \\ \hline\end{array}}_{i-1}\underbrace{\begin{array}{|c|c|c|c|c|c|c|}\hline 0.52 & 0.92 & -1.00 & -0.93 & 0.00 & 0.77 & 0.97 \\ \hline\end{array}}_{i}$$

$$\underbrace{\begin{array}{|c|c|c|c|c|c|c|}\hline -1.00 & -0.64 & -1.00 & -0.90 & 0.00 & 0.63 & 0.83 \\ \hline\end{array}}_{i+1}\underbrace{\begin{array}{|c|c|c|c|c|c|c|}\hline 0.74 & 0.84 & -1.00 & -0.90 & 0.00 & 0.73 & 0.83 \\ \hline\end{array}}_{j-1}$$

$$\underbrace{\begin{array}{|c|c|c|c|c|c|c|}\hline -1.00 & -0.93 & -0.95 & -0.65 & 0.00 & 0.57 & 0.87 \\ \hline\end{array}}_{j}\underbrace{\begin{array}{|c|c|c|c|c|c|c|}\hline 0.73 & 1.00 & -0.85 & -0.65 & 1.00 & 0.57 & 0.77 \\ \hline\end{array}}_{j+1}$$

Fig. 1 Example of a resulting prediction rule.

4 Conclusions

In this paper, we have developed a novel approach based on evolutionary computation for residue-residue contact prediction. The contribution of our study is to provide a possible approach for the contact map prediction using four amino acids properties: hydrophobicity, polarity, net charge and size of residue. These properties helped to improve the search process performed by the algorithm. The resulting rules of our algorithm determine a contact between amino acids and can be easily interpreted and analyzed for experts in the field. As future work, we intend to test other amino acid properties, and to expand the window size of a rule, ideally by having a variable lenght windows, were the optimal length would be found by the evolutionary search performed.

Acknowledgements

This research was supported by the Project of Excellence P07-TIC-02611, by Spanish Ministry of Science and Technology under grants TIN2007-68084-C02-00 and by the Junta de Andalucia, Project P07-TIC-02611.

References

1. Wang Z. Eickholt J. Cheng J. Tegge, AN. Nncon: Improved protein contact map prediction using 2d-recursive neural networks. *Nucleic Acids Research*, 37(2):515–518, 2009.
2. J. Cheng and P. Baldi. Improved residue contact prediction using support vector machines and a large feature set. *Bioinformatics*, 8:113, 2007.
3. Mangal N. Biswas S. Gupta, N. Evolution and similarity evaluation of protein structures in contact map space. *Proteins: Structure, Function, and Bioinformatics*, 59:196–204, 2005.
4. Y. Zhang. I-tasser: fully automated protein structure prediction in casp8. *Proteins: Structure, Function, and Bioinformatics*, 77:100–113, 2009.
5. R. Unger and J. Moult. Genetic algorithms for protein folding simulations. *Biochim. Biophys.*, 231:75–81, 1993.
6. Kyte J. and R.F. Doolittle. A simple method for displaying the hydropathic character of a protein. *J. J. Mol. Bio.*, 157:105–132, 1982.
7. R. Grantham. Amino acid difference formula to help explain protein evolution. *J. J. Mol. Bio.*, 185:862–864, 1974.
8. P. Klein, M. Kanehisa, and C. DeLisi. Prediction of protein function from sequence properties: Discriminant analysis of a data base. *Biochim. Biophys.*, 787:221–226, 1984.
9. DM. Dawson. *The Biochemical Genetics of Man*. Brock, DJH., Mayo, O. eds., 1972.