Evolutionary Protein Contact Maps Prediction Based on Amino Acid Properties

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

School of Engineering, Pablo de Olavide University of Sevilla, Spain
{amarcha,fdivina,aguilar}@upo.es

Abstract. Protein structure prediction is one of the main challenges in Bioinformatics. An useful representation for protein 3D structure is the protein contact map. In this work, we propose an evolutionary approach for contact map prediction based on amino acids physicochemical properties. The evolutionary algorithm produces a set of rules that identifies contacts between amino acids. The rules obtained by the algorithm imposes a set of conditions on four amino acid properties in order to predict contacts. Results obtained confirm the validity of the proposal.

Keywords: Protein Structure Prediction, Contact Map, Evolutionary Computation, Residue-residue Contact.

1 Introduction

Proteins are compounds composed of carbon, hydrogen, oxygen, and nitrogen, which are arranged as strands of amino acids. They are essential components of every live form, as they play an essential function, e.g., transport function, enzymatic function, structural function, etc. The specific biochemical function of a protein is consequence of its structure complexity, which is determined by the specific sequence of amino acids, as demonstrated by Anfinsen [1]. Being able to predict the three-dimensional structure of a protein from its sequence of amino acids, is one of the main open problems in Bioinformatics. Such a problem is known as Protein Structure Prediction (PSP). Solving this problem would allow to acquire the possibility of knowing the function of a protein directly from its amino acids sequence, and would represent a huge advance in various field, e.g., medical or biological areas.

The structure of a protein can be experimentally determined using techniques such as X-ray crystallography or resonance scans (NMR techniques). However, these techniques are expensive and time consuming.

In addition to this, more and more amino acids sequences of proteins are available by the day, but their three-dimensional structures remain often unknown, and so their functions cannot be determined. Consequently the gap between protein sequence information and protein structural information is increasing rapidly. It follows that computational methods are needed in order to reduce this gap, as they would provide a cheaper and faster way to solve the PSP problem.

We can identify three computational approaches for the prediction of 3D structures of proteins: homology-based models, threading models and *ab initio* models. As its name suggests, homology-based models predict protein structures based on sequence homology with known structures, e.g., [2,3]. The principle behind this is that if two proteins share a high degree of similarity in their sequences, then they should have similar 3D structures. Threading, or sequence-structure alignment methods try to determine the structure of a new protein sequence by finding its best "fit" to some fold in a library of structures, see, for instance [4]. *Ab initio* methods attempt to generate models of proteins solely based on sequence information and without the aid of known protein structures, e.g., [5,6]. Our proposal lies in this last category.

When a computational method is used, the first thing one has to do is selecting a way to encode the data. A common way for representing the three dimensional structure of a protein is provided by contact maps. A contact map (CM) is a two-dimensional representation of the tertiary structure of a protein. Contact maps are an useful and interesting approach for the representation of the structure of proteins since they capture all important feature of the folding process. A protein with an amino acid sequence of length N, can be represented by using a NxN symmetric matrix C. Each entry C_{ij} of C is either equal to 1 or 0, depending on whether or not there is a contact between amino acids i and j of the protein. In order to determine if two amino acids are in contact, a threshold μ, expressed in angstroms, is used. If the distance between amino acids i and j is less than μ, then we say that i and j are in contact.

In this paper, we propose a residue-residue contact map predictor based on an evolutionary algorithm (EA). The EA uses a set of representative amino acid properties in order to predict contacts between amino acids. We believe that EAs are suitable for solving the PSP problem, since PSP can be seen as a search problem through the space determined by all the possible foldings. Such a space is highly complex and has a huge size. Finding the optimal solution in such space is very hard. In these cases, EAs have proven to be effective methods that can provide sub-optimal solutions. The EA we propose, will produce a set of rules that express conditions on the particular biochemical properties of the amino acids. Such a model can then be used in order to determine whether or not there is a contact between amino acids. An advantage of such an approach is that the generated rules can easily be interpreted by experts in the field in order to extract further insight of the folding process of proteins.

Previously, EAs have been applied to PSP, e.g., [6], where Torsion angles representation has been used. Hydrophobic-Polar (HP) model and lattice model were employed in [5]. A contact map model generator was included in [4], while [7] used a bit encoding for proteins. Our evolutionary proposal is based on the incorporation of new amino acid properties which have not yet be considered for the prediction of protein structure.

The rest of the paper is organized as follows. In section 2, we describe our proposal to predict protein contact maps. Section 3 presents the experimentation performed in order to assess the validity of our proposal and a discussion of the obtained results. Finally, in section 4, we draw some conclusions and analyze possible future work.

2 Methodology

In order to test our proposal, we have obtained a data set of protein structures from the Protein Data Bank (PDB) [8]. From this data set, we have then extracted physicochemical information about each protein. This data will be divided into a training set and a test set. The EA will use the training set in order to produce a set of contact prediction rules. From these rules, a contact map is built for each test protein sequence, which will be used to establish the accuracy of the obtained prediction model. This experimental procedure is illustrated in Figure 1.

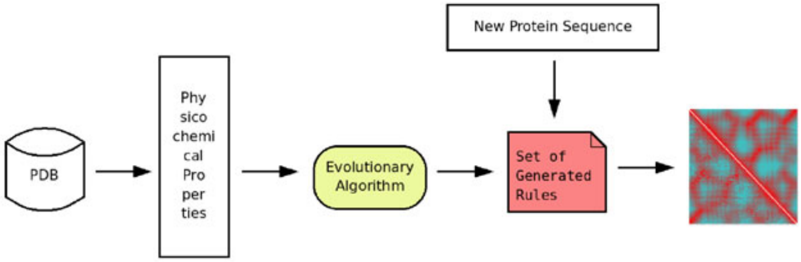

Fig. 1. Experimental and prediction procedure

As already mentioned, our EA will exploit a set of amino acids properties. A set of conditions on the selected properties will establish if a contact between two amino acids is predicted. It is known that amino acid properties have an important role in the PSP problem [9]. Several PSP methods were proposed that relied on amino acids properties, e.g., hydrophobicity and polarity were employed in HP models [5]. The prediction performed by our proposal will be based on four amino acids properties, in particular, hydrophobicity, polarity, charge and residue size.

In the following we address the various solutions adopted for what regards the representation, the genetic operators and the fitness function used by the EA.

2.1 Encoding

In our approach, an individual will encode a rule that determines if there is a contact between amino acid i and j. Such a rule will impose a set of conditions on the four properties of the amino acids. In order to do so, two windows of three amino acids are maintained, where one window is relative to amino acids $i-1, i, i+1$ and the other is relative to amino acids $j-1, j, j+1$.

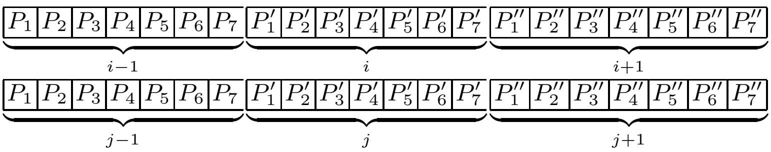
P_1 P_2 P_3 P_4 P_5 P_6 P_7 P'_1 P'_2 P'_3 P'_4 P'_5 P'_6 P'_7 P''_1 P''_2 P''_3 P''_4 P''_5 P''_6 P''_7
$i-1$
i
$i+1$
P_1 P_2 P_3 P_4 P_5 P_6 P_7 P'_1 P'_2 P'_3 P'_4 P'_5 P'_6 P'_7 P''_1 P''_2 P''_3 P''_4 P''_5 P''_6 P''_7
$j-1$
j
$j+1$

Fig. 2. Example of chromosome encoding for the $i-1$,i,$i+1$,$j-1$,j,$j+1$ residues

The encoding of the individuals is illustrated in figure 2. In the figure, positions P_1, P_2 represent the range of the hydrophobicity values for amino acid $i-1$. In the same way, positions P_3, P_4 represent the range of the polarity values, position P_5 encodes the net charge property value of the amino acid. Finally, positions P_6 and P_7 represent the range for the residue size of amino acid $i-1$. All these positions are encoded as real values.

We selected Kyte-Doolittle hydropathy profile for the hydrophobicity [10], the Grantham's profile [11] for polarity and Klein's scale for net charge [12]. The Dawson's scale [13] is employed to determine the size of the residues. The values of these properties are then normalized to a range between -1 and 1 for hydrophobicity, 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).

2.2 Genetic Operators and Fitness Function

The algorithm starts with a randomly initialized population and the algorithm is run for a maximum of 100 generations. If the fitness of the best individual does not increase over twenty generations, the algorithm is stopped and a solution is provided. The population size is set to 100. In order to obtain the next generation, individuals are selected with a tournament selection mechanism of size two. Crossover and mutation are then applied in order to generate offspring. Elitism is also applied, therefore the fittest individual is always preserved in the next generation.

Various crossover operators have been tested. In particular, we have tested the performances of one-point, two-points, uniform and BLX-α crossovers. These crossover operators act at the level of the amino acid properties. So, for instance, the one-point crossover randomly select a point inside two parents and then builds the offspring using one part of each parent. For example, in figure 2, a possible point selected could be P_3 or P'_6. It follows that the resulting rule has to be tested for validity, since the so produced rule could contained incorrect ranges.

BLX-α crossover creates a new offspring $C = C_1, C_2 ... C_4 2$, where C_i is randomly selected in the interval $[C_{min} - I\alpha,\ C_{max} + I\alpha]$. C_{min} and C_{max} are the lower and higher values of the two parents at position i, and I is equal to $max_i - min_i$. An α value is also selected. In our case, we set the α value for the crossover in 0.1. This parameter must be higher or equal than 0. This crossover operator can be seen as a linear combination of the two parents. After having performed several runs of the algorithm, the best results were obtained when the two-points crossover was used, which was then used as standard crossover in the algorithm.

Crossover operator and mutation are applied with a probability of 0.5. If mutation is applied, one gene of the individual is randomly selected, and its value is changed. If the selected gene is relative to the hydrophobicity, polarity, or residue size, its value is increased or decreased by 0.1. If the selected gene is relative to the net charge, its value is randomly changed to another allowed value $(-1, 0, 1)$. If the values of a mutated individual are not within the allowed ranges for each properties, the mutation is discarded.

The aim of the algorithm is to find both general and precise rules for identifying residue-residue contacts. Therefore, we have chosen as fitness function the F-measure, which is given by equation 1.

$$F = 2 \cdot \frac{Recall \cdot Precision}{Recall + Precision}. \qquad (1)$$

Recall represents the proportion of identified contacts, while precision represents the error rate. The algorithm aims at maximizing the fitness.

At the end of the EA, the best rules are selected and they represent the solution. This is done in an incremental way: first, the best individual, according to its F-measure, is selected and added to the solution S. Then the next best individual is added to S, and the F-measure of S is calculated. This process is repeated until the addition of a rule causes the F-measure of S to decrease.

3 Experiments and Results

The experimentation has been performed with a data set described in [14]. This protein data set (56PDB) consists of 56 proteins with a sequence identity value lower than 25% and a sequence length lower than 100. As validation method we have used a 10-fold cross-validation.

Three statistical measures were computed to evaluate the accuracy of our algorithm:

- 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 correctly predicted examples.
- Specificity, or 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.

As already mentioned, an execution of the algorithm provides as a result a set of rules. If the algorithm is run several times, the final prediction model will consist of all the rules obtained at each execution. In other words, each time the algorithm is run, a number of rules are added to the final solution. Repeated or redundant rules are not included in the final solution. As optimal and exact number of rules is unknown, we have performed various experiments varying the numbers of runs of the EA, where to a higher number of runs corresponds on higher number of rules. The aim of this was to test whether or not a higher number of rules would yield better results.

Table 1 shows the results for 100, 500, 1,000 and 2,000 executions. For instance, for 1,000 runs we have obtained 2,190 rules and for 2,000 runs we have finally obtained 4,866 rules.

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

Runs	$Recall_{\mu\pm\sigma}$	$Spec._{\mu\pm\sigma}$	$Prec._{\mu\pm\sigma}$	#rules
100	$0.144_{\pm 0.025}$	$0.993_{\pm 0.010}$	$0.502_{\pm 0.115}$	227
500	$0.539_{\pm 0.080}$	$0.985_{\pm 0.020}$	$0.462_{\pm 0.118}$	1,153
1000	$0.658_{\pm 0.130}$	$0.973_{\pm 0.025}$	$0.453_{\pm 0.125}$	2,190
2000	$0.683_{\pm 0.185}$	$0.965_{\pm 0.028}$	$0.438_{\pm 0.112}$	4,866

By examining table 1, we can notice that a low recall rate is achieved for 100 runs. However this result improves substantially when the number of rules in the final solution increases, reaching a maximum of 68% of correctly identified contacts for 2,000 runs. This is a consequence of having more rules in the final solution. In fact, the more rules in the final solution, the more cases are covered. The opposite holds for precision. In fact, it can be noticed that the precision obtained decreases with the increase of the number of rules in the final solution. This result was quite expected, since by covering more cases, the possibility of prediction errors increases due to the higher number of rules. Even in this case, the obtained precision is satisfactory in all the cases, as it never goes below 40%. We can also notice, that specificity is very satisfactory in all the settings, reaching values higher than 95% in all the cases. Notice that as for the precision, the specificity decreases when the number of rules increases. As far as the optimal number of rules is concerned, we can conclude that it is difficult to establish the optimal number of rules; with more rules, the true positive rate increases, however the false positive rate is also increased.

It is not possible to make an exact comparison between our method and the other existing methods. Each method uses a different structural data bases, different sets of proteins and different measures to evaluate the accuracy of their algorithms. However, other methods for PSP, e.g., [15], set the precision rate for a contact map prediction at about 30%. The precision obtained by our method clearly performed this result.

Figure 3 shows an example of a resulting rule. It can be seen that the rule has been divided into two windows of size three, as explained in section 2. If we inspect this rule, we can infer that, for example, the hydrophobicity value for the amino acid i lies 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. As it can be noticed the produced rules are easily interpretable by experts in the field.

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

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

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

Fig. 3. Example of a resulting prediction rule

4 Conclusions and Future Work

In this article, we proposed an evolutionary algorithm approach for protein contact map prediction. Our algorithm generates a set of rules for residue-residue contact prediction using a representation based on four amino acid properties. Our approach can be used in an incremental way, in fact, the algorithm can be executed several times, and at each run, a set of rules is added to the final solution. This allows us to tune the results according to what measure we want to optimize. For example, by increasing the number of rules the recall increases. The rules forming the final solution express a set of conditions on the four phsysicochemical properties of amino acids. As a consequence, such rules can easily be interpreted and analyzed by experts in the field in order to obtain more insight on the protein folding process.

As for future work, we intend to expand this study to other significant amino acid properties, e.g., isoelectric point and steric parameter. Moreover, the length of window size of each individual could be variable, where the estimation of an adequate length would be determined by the evolutionary process. Another future development is the application of the algorithm to larger proteins data set, in order to test the validity of our proposal in these cases.

References

1. Anfinsen, C.: The formation and stabilization of protein structure. Biochem. J. 128, 737–749 (1972)
2. Zhang, Y.: I-tasser: fully automated protein structure prediction in casp8. Proteins: Structure, Function, and Bioinformatics 77, 100–113 (2009)
3. Tegge, A., Wang, Z., Eickholt, J., Cheng, J.: Nncon: Improved protein contact map prediction using 2d-recursive neural networks. Nucleic Acids Research 37(2), 515–518 (2009)
4. Gupta, N., Mangal, N., Biswas, S.: Evolution and similarity evaluation of protein structures in contact map space. Proteins: Structure, Function, and Bioinformatics 59, 196–204 (2005)

5. Unger, R., Moult, J.: Genetic algorithms for protein folding simulations. Biochim. Biophys. 231, 75–81 (1993)
6. Cui, Y., Chen, R.S., Hung, W.: Protein folding simulation with genetic algorithm and supersecondary structure constraints. Proteins: Structure, Function and Genetics 31, 247–257 (1998)
7. Zhang, G., Han, K.: Hepatitis c virus contact map prediction based on binary strategy. Comp. Biol. and Chem. 31, 233–238 (2007)
8. Protein data bank web, `http://www.pdb.org`
9. Gu, J., Bourne, P.E.: Structural Bioinformatics (Methods of Biochemical Analysis). Wiley-Blackwell, Chichester (2003)
10. Kyte, J., Doolittle, R.F.: A simple method for displaying the hydropathic character of a protein. J. J. Mol. Bio. 157, 105–132 (1982)
11. Grantham, R.: Amino acid difference formula to help explain protein evolution. J. J. Mol. Bio. 185, 862–864 (1974)
12. Klein, P., Kanehisa, M., DeLisi, C.: Prediction of protein function from sequence properties: Discriminant analysis of a data base. Biochim. Biophys. 787, 221–226 (1984)
13. Dawson, D.M.: The Biochemical Genetics of Man. In: Brock, D.J.H., Mayo, O. (eds.) (1972)
14. Fariselli, P., Olmea, O., Valencia, A., Casadio, R.: Prediction of contact map with neural networks and correlated mutations. Protein Engineering 14, 133–154 (2001)
15. Zhang, G.Z., Huang, D.S., Quan, Z.H.: Combining a binary input encoding scheme with rbfnn for globulin protein inter-residue contact map prediction. Pattern Recognition Letters 16(10), 1543–1553 (2005)