An Evolutionary Approach for Protein Contact Map Prediction

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

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

Abstract. In this study, we present a residue-residue contact prediction approach based on evolutionary computation. Some amino acid properties are employed according to their importance in the folding process: hydrophobicity, polarity, charge and residue size. Our evolutionary algorithm provides a set of rules which determine different cases where two amino acids are in contact. A rule represents two windows of three amino acids. Each amino acid is characterized by these four properties. We also include a statistical study for the propensities of contacts between each pair of amino acids, according to their types, hydrophobicity and polarity. Different experiments were also performed to determine the best selection of properties for the structure prediction among the cited properties.

Keywords: protein structure prediction, contact map, evolutionary computation, residue-residue contact, amino acid properties.

1 Introduction

Protein Structure Prediction (PSP) is one of the main open problems in Computational Biology. The rules that determine the folding process of a protein are still unknown. Once the protein has fold, the 3D structure of the protein is revealed. Knowledge of these structures would represent a huge advance in different medical areas, like the treatment of Alzheimer and Cystic fibrosis. Existing methods which determine the protein structure are expensive and slow (e.g., NMR spectroscopy and X-ray crystallography). Therefore, computational methods are needed since they would provide a cheaper and faster way to address this problem.

In any computational method, one of the main issues to solve is how to represent the data. Contact maps are a popular representation of a protein structure, since they provide a reduced representation of the tertiary structure of a protein. A contact map is a bidimensional symmetric matrix C of size NxN, where N is the length of the amino acid sequence. Each cell represents a pair of amino acids (i, j). According to a given threshold μ, usually expressed in angstroms, we assign to each cell C_{ij} of C the value 1 (contact) if the distance between residues i and j is less than or equal to μ, or 0 otherwise.

Different approaches have been developed using protein contact maps to solve the PSP problem: artificial neural networks (ANNs) [1,2], support vector machines [3], evolutionary algorithms (EAs) [4] and template-based modeling [5]. In particular, [5] and [2] were ranked as two of the most accurate methods from CASP8 [6].

PSP methods based on EAs have used different representations. For instance, in [7] a Torsion angles representation has been used. HP model and lattice model were employed in [8]. A contact map model generator was included in [4], while [9] used a bit encoding for proteins.

In this article, we propose a residue-residue contact map predictor based on evolutionary computation (EC). Differently from previous methods based on EC, our approach will base the prediction on some specific physicochemical amino acids properties. The prediction model that will be generate, will consist of rules that can be used in order to determine whether or not there is a contact between amino acids. The generated rules will express conditions on the considered amino acids properties. PSP problem 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. Another advantage of our method is that the generated prediction model can be easily interpreted by experts, since it will consists of a set of rules.

This paper is organized as follows. In section 2, we describe our proposal to predict protein contact maps. Section 3 presents the experimentation and 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) [10]. More details on how this data set was built are given in the next section. This data will be used both as training set and test set for the EA. In particular the EA will use the training set in order to provide a set of rules that can be used for determining contacts between amino acids. This experimental procedure is represented in Figure 1.

Several amino properties are considered as significant for the protein structure prediction [11]. From this collection, we have selected hydrophobicity, polarity, charge and residue size. We have selected these four properties since they seem to have a certain relevance in the folding process of a protein [12]. Our prediction will be based on these four properties. We have performed several experiments to test the validity of the choice of the properties used in section 3. Furthermore, we have performed a statistical study of the propensities of each pair of amino acids according to these properties. This study is described in the following section.

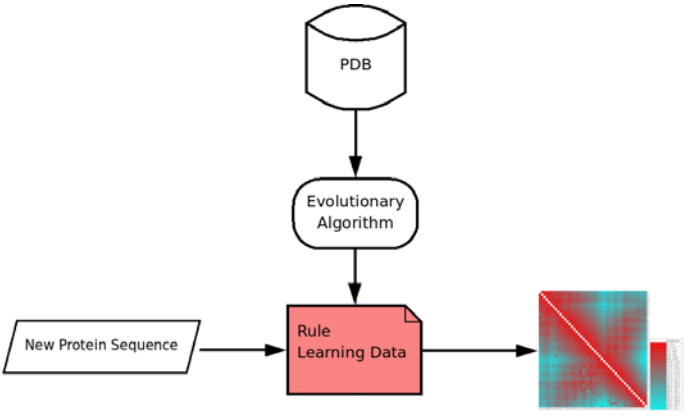


Fig. 1. Experimental and prediction procedure

2.1 Statistical Analysis

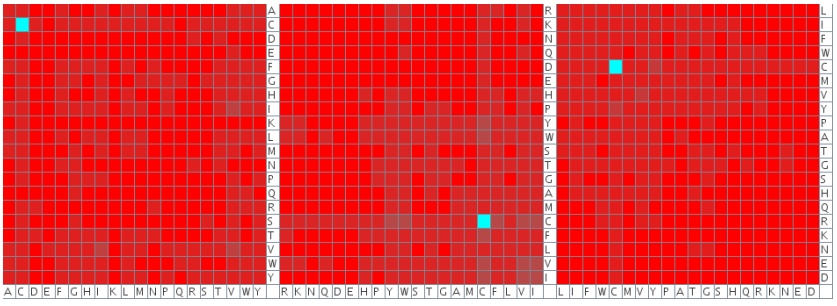
The goal of this study is to extract some information that could be used in order to aid the EA in its search. To this aim, we conducted a statistical study in order to compute the contact propensities of each amino acid. This study was performed with 12,830 non-homologous and non-redundant protein sequences extracted from PDB with 30% homology cutoff. This data set was obtained using the PDB Advanced Search Select [13]. The complete list of the 12,830 protein identifiers can be downloaded from [14].

In order to compute the global propensity of any given pair of amino acids (A, B) the following equation was used:

$$P_{AB} = \frac{n_{AB}}{\sum_{XY} n_{XY}} / \frac{T_{AB}}{\sum_{XY} T_{XY}} \tag{1}$$

In equation 1, n_{AB} represents the number of (A, B) contacts, $\sum_{XY} n_{XY}$ stands for the sum of the total number of contacts for any possible pair of amino acids in a protein sequence, T_{AB} represents the total number of occurrences (contacts or not) for (A, B) in all the whole database, and $\sum_{XY} T_{XY}$ is the sum of the total number of occurrences for any pair of amino acids. The idea behind this formula is to calculate the contact frequency of each pair of amino acids over the total pairs of amino acids. Using equation 1, we can extract some conclusions regarding the physicochemical properties of those pairs of amino acids characterized by a high contact propensity.

In order to do so, we have built a matrix according to equation 1. Such a matrix is reported in Figure 2a, where each cell represents a pair of amino acids and their propensity to be in contact. The different propensity levels are represented by a different color. A cell with blue or dark red color represents a high likelihood for this pair. On the other hand, a light red cell represents a low propensity. In matrices 2b and 2c, amino acids are ordered, in an increasing order, according to their hydrophobicity and polarity values [15][16], i.e., amino



(a) Residue type (b) Hydrophobicity (c) Polarity

Fig. 2. Propensity symmetrical matrices for residue-residue contacts ordered by a specified criteria

acid R is the least hydrophobic and amino acid I is the most hydrophobic in Figure 2b.

The contact threshold μ is established in 8 angstroms (Å), as in [2]. We have discarded local range contacts, which have a sequence separation lower than 7 amino acids [1].

By analyzing these matrices, we can observe that the amino acid pair with the higher contact propensity is (C, C), and in general, amino acid C (Cysteine) is the amino acid with higher contact propensity. From Figures 2b and 2c, we can conclude that the higher the value of hydrophobicity for a residue pair, the higher the contact probability for this pair. In fact, in Figure 2b, we can individuate a region, located in the lower right part of the matrix, characterized by a high contact propensity, and this region corresponds to amino acids that are characterized by high hydrophobicity values. In the same way, we can conclude that the lower the value of polarity for a residue pair, the higher the contact probability of this pair. A similar study was performed for net charge and residue size, but no clear conclusion was extracted, and for this reason, matrices relative to these properties were not included in this paper.

These results will be incorporated in the fitness function, as it will be explain in the following sections. In the following we discuss the various solutions we adopted for what regards the fitness, the representation and the genetic operators used by the EA.

2.2 Encoding

In our approach, each individual encodes a rule for a residue-residue contact. Each individual represents the four selected properties of amino acids in two windows of size 3. One window is relative to amino acids $i-1, i, i+1$ and the other window is associated with the amino acids $j-1, j, j+1$, where i and j are two possible amino acids in contact.

An example of individual is reported in Figure 3. We can see that seven genes are associated with each amino acid. All the genes are represented by real

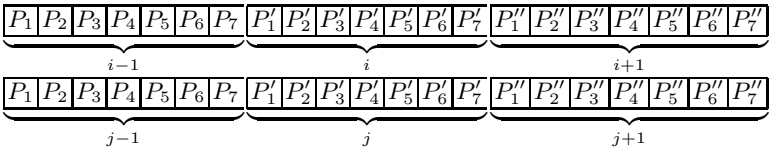


Fig. 3. Example of chromosome encoding for the $i-1$,i,$i+1$,$j-1$,j,$j+1$ residues. An individual is constituted by these six amino acids.

values. For instances, genes in positions P_1, P_2 represent the range relative to the hydrophobicity values for amino acid $i-1$, while genes P_3, P_4 represent its polarity value range. Gene P_5 represents the net charge property values of the amino acid, and finally, genes P_6 and P_7 represent the range relative to the residue size.

We selected Kyte-Doolittle hydropathy profile for the hydrophobicity [15], the Grantham's profile [16] for polarity and Klein's scale for net charge [17]. The Dawson's scale [18] 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). The encoding of the individuals is illustrated in Figure 3.

2.3 Fitness Function

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 the equation 2:

$$F = 2 \cdot \frac{Recall \cdot Precision}{Recall + Precision}. \tag{2}$$

The higher the fitness, the better the individual, so the aim of the EA is to maximize this value.

The analysis proposed in section 2.1 have shown that if two amino acids are in contact, they have more probabilities to have high hydrophobicity values and low polarity values. Based on this, we increase the fitness of an individual if amino acids i and j fulfill these properties.

2.4 Genetic Operators

Individuals are selected with a tournament mechanism of size two. Offsprings are generated by using a one-point crossover, which is applied with 1.0 probability. New generated individuals undergo mutation with a probability of 0.5. These values were determinated after having performed several runs of the algorithm, and were the ones yielding the best results.

If mutation is applied, one gene of the individual is randomly selected, and its value is increased or decreased by 0.1. If the selected gene is relative to the

charge of the amino acid, then its value is randomly changed to another allowed value. After that mutation is applied to an individual, the rule it encodes is always checked for validity, i.e., if its values are within the ranges allowed for each properties. If the encoded rule is not valid, then mutation is discarded. Elitism is also always applied, therefore the fittest individual is always preserved in the next generation.

The population size is set to 100, and the initial population is randomly initialized. The maximum number of generations that can be performed is set to 100. However, if the fitness of the best individual does not increase over twenty generations, the algorithm is stopped and a solution is provided.

The final solution is built in an incremental way. In particular, the solution is extracted by first selecting the fittest individual. After that, the following best rule, according to the F-measure, is added to the solution, and the global F-measure of the solution is computed. If the F-measure of the solution decrease, then the added rule is discarded and the process stops. Otherwise the rule is added to the solution and the next best rule is considered. Repeated or redundant rules are not considered in this process.

3 Experiments and Results

In order to test the effectiveness of our proposal, we have used two datasets. The first one, called $300PDB$ from now on, consists of a subset of 300 protein sequences randomly extracted from the dataset described in section 2.1 with a maximum length of 403 residues. The second dataset, called $56PDB$ was taken from [1]. This protein dataset consists of 56 proteins with an identity value $<25\%$ 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, Precision and Specificity:

- 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, 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.

In order to test the effect of varying the total number of rules that a solution consists of, we have performed experiments by varying the number of executions of the EA. To this aim, the rules provided by each run of the EA are added to a final solution. Repeated or redundant rules are not inserted in the final solution. Thus, the more runs of the EA, the more rules will be added to the final solution.

Before analyzing the global performances of our proposal, we report results of an experiment that was conducted in order to test the effect of basing the

Table 1. Average precision rate results and standard deviation for different type and number of amino acid properties. H represents the hydrophobicity, P the polarity, C the charge net and S is the residue size.

$Properties$	$Precision_{\mu\pm\sigma}$
H, P, C, S	$0.562_{\pm 0.128}$
H, P, C	$0.461_{\pm 0.121}$
H, P, S	$0.483_{\pm 0.117}$
H, C, S	$0.504_{\pm 0.148}$
P, C, S	$0.453_{\pm 0.108}$
H, P	$0.415_{\pm 0.099}$
C, S	$0.473_{\pm 0.110}$
H, S	$0.502_{\pm 0.158}$
H, C	$0.437_{\pm 0.123}$
P, S	$0.422_{\pm 0.144}$
P, C	$0.456_{\pm 0.147}$
H	$0.364_{\pm 0.093}$
P	$0.413_{\pm 0.127}$
S	$0.379_{\pm 0.108}$
C	$0.215_{\pm 0.060}$

Table 2. Average results and standard deviation obtained for different number of executions of the algorithm for the 300PDB protein data set

Runs	$Recall_{\mu\pm\sigma}$	$Spec._{\mu\pm\sigma}$	$Prec._{\mu\pm\sigma}$	#rules
100	$0.056_{\pm 0.038}$	$0.915_{\pm 0.010}$	$0.565_{\pm 0.130}$	223
500	$0.251_{\pm 0.120}$	$0.990_{\pm 0.022}$	$0.484_{\pm 0.110}$	1,147
1000	$0.501_{\pm 0.160}$	$0.988_{\pm 0.035}$	$0.445_{\pm 0.116}$	2,075
2000	$0.628_{\pm 0.235}$	$0.972_{\pm 0.020}$	$0.434_{\pm 0.105}$	4,984

prediction on different amino acids properties. To this aim, we have varied the properties encoded in an individual. This experiment was performed on the $56PDB$. For each setting, 500 executions of the EA were performed and the precision rates were calculated. We have considered the precision rate as accuracy measure for this experiment as in [1]. It corresponds to the number of correctly assigned contacts divided by the total of predicted contacts. Table 1 reports the results of this experiment for each possible combination of the four physicochemical amino acid properties. For instance, when all the properties were considered, setting (H,P,C,S), a precision of 56,2% was obtained.

From these results, we can conclude that the best results are achieved by encoding all the properties. In order to support this conclusion, we have performed two non-parametric statistical tests (Friedman Test and Willcoxon Test) on the results. The results obtained from this test sustain our conclusions, since the differences of the results obtained by the various settings are statistically significant.

Table 3. Average results and standard deviation obtained for different number of executions of the algorithm for the 56PDB protein data set

Runs	$Recall_{\mu\pm\sigma}$	$Spec._{\mu\pm\sigma}$	$Prec._{\mu\pm\sigma}$	#rules
100	$0.064_{\pm 0.036}$	$0.991_{\pm 0.007}$	$0.601_{\pm 0.152}$	221
500	$0.257_{\pm 0.091}$	$0.962_{\pm 0.022}$	$0.562_{\pm 0.128}$	1,057
1000	$0.515_{\pm 0.161}$	$0.929_{\pm 0.036}$	$0.556_{\pm 0.139}$	2,383
2000	$0.653_{\pm 0.175}$	$0.903_{\pm 0.039}$	$0.550_{\pm 0.132}$	4,573

Tables 2 and 3 report the results for 100, 500, 1,000 and 2,000 executions of the EA on the two data sets. In particular, in the two tables, in the first column, the number of runs performed is shown, the second, the third and fourth columns report the average recall, specificity and precision, respectively. Standard deviation is also shown. The last column reports the total number of rules contained in the final solution.

The first conclusion that can be drawn by a first analysis of the two tables is that if the number of runs increases, the recall rate will also increase. On both datasets, a low recall rate was obtained for 100 runs. However, when the number of runs are increased to 2,000, the recall obtained increases, reaching a value of about 62% on the 300PDB dataset and of about 65% on 56PDB dataset. This is due the higher number of rules contained in the solution in the case of 2,000 runs. However, the precision rate will decrease. This result was quite expected, since by covering more cases, the possibility of errors increases.

As far as the specificity is concerned, good results were obtained in all the settings. This means that the algorithm is very accurate in predicting non-contacts between amino acids. Even if the precision rate decreases as the number of runs increases, the average results relative to this properties are satisfactory.

We cannot report a direct comparison between our method and other existing methods, since each method uses different structural data bases, different sets of proteins and different measures to evaluate the accuracy of their algorithms. However, we can report that other methods for PSP, e.g., [19], set the precision rate for a contact map prediction at about 30%. Our approach clearly outperformed this result, since even the minimum precision rate obtained was of 43%.

We can also conclude that it is difficult to determine the optimal number of rules. In fact, the more rules added to the final solution, the higher the recall obtained. However, this has a negative effect on the precision.

As mentioned before, the rules provided by our methods can be easily interpreted. In order to show this, an example of a resulting rule is shown in Figure 4. As explained in section 2, a rule imposes a set of conditions on the represented physicochemical properties of the amino acids. For example, the proposed rule states that the hydrophobicity value of the amino acid i lies in the range $[0.52, 0.92]$, the polarity value between -1.0 and -0.93, the charge should be neutral, i.e., 0.0, and the residue size of the amino acids should be contained in the interval $[0.77, 0.97]$. From this rule, we could conclude that amino acid i could be L (Lysine)

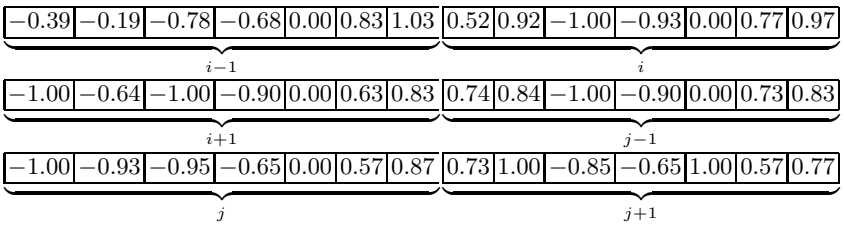


Fig. 4. Example of a prediction rule

or F (Phenylalanine), since these amino acids fulfill all these conditions imposed by the rule. It is easy to conclude that by inspecting the rules provided by the algorithm, experts in the field could extract useful information regarding the four properties represented by the individuals, getting more insight regarding the possible connections between the properties and the folding of the protein.

4 Conclusions and Future Work

In this article, we proposed an evolutionary approach for solving the protein contact map prediction problem. Our EA generates a set of rules for residue-residue contact prediction using an encoding based on four amino acid properties. A statistical study has been performed, analyzing the probabilities of contact according to several amino acids properties. From this study, we have extracted useful information that is used in order to aid the EA in the search it performs. We have performed several experiments with different encodings in order to determine the best combination of amino acid properties to represent. After applying two non-parametrical statistical tests, we have concluded that by encoding all four properties obtains the best results would be obtained. In general, the achieved results indicate that our method is successful in provide a good prediction of contacts among amino acids.

Another important aspect of the process of extracting knowledge from datas, is the interpretability of the results. We believe that the results provided by our methods are highly interpretable, since they consists of rules can be easily interpreted and analyzed by experts in the field.

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 could be determined by the evolutionary process.

Acknowledgements

This research was supported by the Project of Excellence P07-TIC-02611 and by Spanish Ministry of Science and Technology under grants TIN2007-68084-C02-00 “Sistemas Inteligentes para descubrir patrones de comportamiento. Aplicación a base de datos biológicas” and by the Junta de Andalucia, Project P07-TIC-02611.

References

1. Fariselli, P., Olmea, O., Valencia, A., Casadio, R.: Prediction of contact map with neural networks and correlated mutations. Protein Engineering 14, 133–154 (2001)
2. 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)
3. Cheng, J., Baldi, P.: Improved residue contact prediction using support vector machines and a large feature set. Bioinformatics 8, 113 (2007)
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. Zhang, Y.: I-tasser: fully automated protein structure prediction in casp8. Proteins: Structure, Function, and Bioinformatics 77, 100–113 (2009)
6. Casp8 competition official web, `http://predictioncenter.org/casp8`
7. 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)
8. Unger, R., Moult, J.: Genetic algorithms for protein folding simulations. Biochim. Biophys. 231, 75–81 (1993)
9. Zhang, G., Han, K.: Hepatitis c virus contact map prediction based on binary strategy. Comp. Biol. and Chem. 31, 233–238 (2007)
10. Protein data bank web, `http://www.pdb.org`
11. Russell, R.B., Betts, M.J., Barnes, M.R.: Amino acid properties and consequences of subsitutions. Bioinformatics for Geneticists. Wiley, Chichester (2003)
12. Gu, J., Bourne, P.E.: Structural Bioinformatics (Methods of Biochemical Analysis). Wiley-Blackwell, Chichester (2003)
13. Protein data bank advanced search, `http://www.pdb.org/pdb/search/advSearch.do`
14. Complete list of 12,830 pdb protein identifiers used in this article, `http://www.upo.es/eps/marquez/proteins.txt`
15. Kyte, J., Doolittle, R.F.: A simple method for displaying the hydropathic character of a protein. J. J. Mol. Bio. 157, 105–132 (1982)
16. Grantham, R.: Amino acid difference formula to help explain protein evolution. J. J. Mol. Bio. 185, 862–864 (1974)
17. 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)
18. Dawson, D.M.: The Biochemical Genetics of Man. In: Brock, D.J.H., Mayo, O. (eds.) (1972)
19. 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)