A NSGA-II Algorithm for the Residue-Residue Contact Prediction

Alfonso E. Márquez-Chamorro[1], Federico Divina[1], Jesús S. Aguilar-Ruiz[1],
Jaume Bacardit[2], Gualberto Asencio-Cortés[1],
and Cosme E. Santiesteban-Toca[3]

[1] School of Engineering, Pablo de Olavide University of Sevilla, Spain
{amarcha,fdivina,aguilar,guaasecor}@upo.es
[2] School of Computer Science, University of Nottingham, United Kingdom
jaume.bacardit@nottingham.ac.uk
[3] Centro de Bioplantas, University of Ciego de Avila, Cuba
cosme@bioplantas.cu

Abstract. We present a multi-objective evolutionary approach to predict protein contact maps. The algorithm provides a set of rules, inferring whether there is contact between a pair of residues or not. Such rules are based on a set of specific amino acid properties. These properties determine the particular features of each amino acid represented in the rules. In order to test the validity of our proposal, we have compared results obtained by our method with results obtained by other classification methods. The algorithm shows better accuracy and coverage rates than other contact map predictor algorithms. A statistical analysis of the resulting rules was also performed in order to extract conclusions of the protein folding problem.

Keywords: Protein structure prediction, contact map, multi-objective evolutionary computation.

1 Introduction

Protein Structure Prediction (PSP) is one of main challenges in Structural Bioinformatics. Since Anfinsen's experiment discovered that the amino acid sequence determines the shape of a protein [1], a huge number of computational experiments were performed with the aim of obtaining the rules of the protein folding. Knowledge of these rules would play an important role in Biomedicine for the design of new drugs. Although experimental procedures to obtain the 3D protein structure, as X-ray Crystallography and Nuclear Magnetic Resonance (NMR), have shown brilliant results [2], the cost of such techniques, both in term of time and money, is prohibitive. Besides, these techniques cannot be applied to all proteins. In fact, 25% of proteins do not crystallize and are too big for the NMR.

For these reasons, computational methods are particularly suited for this problem, since they, generally, represent a cheaper and faster way to address the protein folding problem. Some of these computational methods used a contact map representation to solve this problem. A contact map is a bi-dimensional representation of the protein structure of a protein, where if an entry i, j has value 1 then a contact between residues i and j is predicted, and a 0 indicates a no contact. We consider a contact between i and j, if the distance between them is lower than a certain threshold μ. Different approaches were developed as protein contact map predictors: artificial neural networks (ANNs) [3,4], support vector machines [5], evolutionary algorithms (EAs) [6] and template-based modeling [7]. Every two years, Critical Assessment of Protein Structure Prediction (CASP) competition [8] evaluates the most accurate computational methods for the PSP problem. One of the categories of this competition is called "Detecting residue-residue contacts in proteins (RR)". Our approach is included in this category.

Among the above mentioned methods, EAs, have become popular as robust and effective methods for solving optimization problems. In particular, they have shown the capacity of finding suboptimal solutions in search spaces when the search space is characterized by high dimensionality. This is the case for the protein folding problem, where the set of possible folding rules of a protein determine the search space. Many evolutionary approaches have been developed to tackle the PSP problem, e.g., [9] [10] [6] [11]. These methods evaluate individuals by means of a single function that provides a measure of their quality. In other words, they are evaluating a single objective function. This approach represents the classical way of addressing a problem with an EA: the objectives to optimize are combined into a single fitness function which is then used in order to guide the evolutionary search. However, there are some problem where this approach is not the most appropriate. Different solutions can produce conflicts between different objectives. A solution that is optimal with respect to one objective may not be optimal for the rest, therefore it would be improper to choose such solution as optimal solution of the problem. It becomes then necessary to establish a compromise among the objectives. The solutions that fulfill this compromise are called the Pareto set. The notion of Pareto set is based on the concept of dominance that will be explained in the next section. When an optimization problem has several objectives, the task of finding one or more suboptimal solutions is called Multi-objective optimization.

Multi-objective Evolutionary Algorithms (MOEAs) appear as an extension of EAs for single objective problems. A MOEA should be designed to achieve two purposes simultaneously: to achieve good approximations to the Pareto front and maintain the diversity of solutions, in order to adequately search the solution space and do not converge to a unique solution [12]. Some of the best known MOEAs are NSGA, SPEA, NSGA-II, SPEA-II and PAES-II [8].

Several prediction methods have considered the PSP problem as a multi-objective optimization problem. For instance, [13] developed MI-PAES as a modified version of PAES using a torsion angles model. A parallel multi-objective optimization was performed by using Chemistry at HARvard Macromolecular Mechanics (CHARMM) energy function in [14]. [15] proposed a multi-objective Feature Analysis and Selection Algorithm (MOFASA) in order to solve the

Protein Fold Recognition (PFR) problem. In [16], a I-PAES algorithm is used as search procedure for exploring the space of the PSP problem. The concept of bond and non-bond energies are included in the fitness function of this approach.

In this paper, we propose a contact map predictor based on a MOEA. More specifically, it is based on a NSGA-II algorithm [17]. A NSGA-II algorithm initially creates a population (random or by a technique of initialization) of parents. The population is sorted according to levels of non-dominance (ranking Pareto fronts). Each solution is then assigned a fitness value according to their level of non-dominance (1 is the best level). Tournament selection, the crossover and mutation are used to create the offspring population of size N.

Our algorithm generates a set of rules that predicts contacts between amino acids. In particular, each rule imposes a set of conditions on some specific amino acids properties. Rules consider two windows of 3 amino acids, which are centered around the two target residues in contact.

In order to test our proposal, we obtain the training data set from the Protein Data Bank (PDB), and produce a file in arff format with the encoded information. The rules that are produced after the training phase are classified according to each specific pair of residues that they represents. For a new protein sequence, we apply the required rules for each residue pair and obtain the protein contact map. Our application also provide a graphical representation of these contact maps. The novelty of our proposal consists on the use of amino acid properties which are involved in the folding process and, to the best of our knowledge, have not been applied in similar evolutionary approaches for this problem.

The remainder of this paper is organized as follows. Section 2 introduces some basic concepts of the Multi-objective optimization. Our multi-objective evolutionary approach is described in section 3. Section 4 presents the experimentation and obtained results. Finally, section 5, includes some conclusions and possible future works.

2 Multi-objective Optimization Problem

Before describing our algorithm, this section presents a brief introduction to multi-objective optimization problems and related concepts.

A Multi-objective optimization problem is based on the optimization (minimization or maximization) of a set of objective functions, usually in conflict with each other. The existence of multiple objective poses a fundamental difference with the single objective problems: typically there will not be a single solution, but a set of solutions that can present different clashes among the values of the objectives to optimize. We can define a multi-objective optimization problem in this way: let $(f_1(x), f_2(x)...f_n(x))$ be a set of functions to be optimized, where $x = (x_1, ..., x_p)$ is a vector of decision variables belonging to a universe X and $f_i(x)$ is an arbitrary linear or non-linear function, $1 \leq i \leq n$. Therefore, the problem consists of finding the x that provides the best compromise value for all $f_i(x)$.

To solve the above problem, we should defined some criteria to determine which solutions are considered of good quality and which are not. Hence, we introduce the concept of dominance, that is used in the process of evaluating the different solutions. A solution x is said to be not dominated *iff* there is not another solution y such that: $f_i(y) <= f_i(x)$ for all $i = 1..n$ and $f_i(y) < f_i(x)$ for some i. From this, it follows that the Pareto front is formed by all the non-dominated solutions.

We have applied these concepts to the Protein Contact Prediction problem. In this article, we have considered coverage and accuracy as two different functions and are optimized separately.

In order to do so, we have implemented a MOEA based on an Elitist Non-Dominated Sorting Genetic Algorithm (NSGA-II). NSGA-II incorporates elitism and reduces the complexity of the procedure fast sorting by non-dominance of its predecessor NSGA. The algorithm performs a classification of the population using Pareto fronts. Individuals which belong to the first front are the non-dominated front, those in the second front are not dominated in the absence of previous front, and so on. Each individual is assigned a rank equal to its level of non-dominance. The best individuals are those with lower ranks. In order to maintain diversity, we use a crowding distance, which is assigned to each individual of the current population. The selection is performed by binary tournament. The tournament is won by the individual with a lower range (Pareto front level). If the two ranges are the same, the tournament is won by the individual who has lower crowding distance. This algorithm has a low time complexity of $O(NlogN)$, where N is the population size.

3 Our Approach

In this section, we present the main characteristics of our proposal. As we have said before, the aim of this algorithm, called PSP-NSGAII, is the prediction of protein contact maps. In order to test our proposal, the first thing to do was to select a set of sequences. For this, we selected from PDB a set of 173 proteins that appears in [3]. We extract the required information as the amino acid sequences and distances between amino acids. To calculate the distances, we use the Euclidean distance between C_β atoms (C_α for Glycine) of each pair of residues. The formula of Euclidean distance is $d(i,j) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2}$, where (x_1, y_1, z_1) represent the atomic coordinates of the first amino acid and (x_2, y_2, z_2) are the coordinates of the second amino acid. Once the training set is prepared, we use this data to train our evolutionary algorithm. As we have said previously, we propose a Multi-Objective algorithm as a method to identify protein folding rules. These rules provide us the specified characteristics of amino acids in contact. They specify which property values and conditions must have the amino acids in contact and the ones which precede and follow them. Our proposal build the set of final rules in an incremental way. Each time the algorithm is run, a set of rules are selected and added to a final solution set. For each iteration, we select those rules which contribute to increase the F-measure of the global solution.

In the following the characteristics of the representation, the fitness function and the genetic operators used by the EA will be presented.

3.1 Encoding

Each individual is represented as follows. We have taken into account six amino acids. For two amino acid in contact i and j, we represent the amino acid $i-1$, $i+1$, $j-1$ and $j+1$, i.e., the amino acids that precede and follow i and j in the sequence. This choice was made after having performed various experiments with different window sizes, ranging from 6 to 14. Each amino acid is represented by 7 genes; two genes for the hydrophobicity (ranging from -1 to 1), two genes for polarity (ranging from -1 to 1), 1 gene for the charge (-1, 0, 1 for negative, neutral and positive charge respectively) and two genes for the volume of residue (ranging from 0 to 1). Figure 1 shows the representation for an amino acid. Our representation consists in 42 attributes in total.

$$\underbrace{\boxed{H_1}\boxed{H_2}\boxed{P_1}\boxed{P_2}\boxed{C}\boxed{V_1}\boxed{V_2}}_{i}$$

Fig. 1. Example of encoding for the amino acid i. An individual is constituted by six amino acids $i-1$, i, $i+1$, $j-1$, j and $j+1$. H_1,H_2,P_1,P_2,V_1 and V_2 are lower and upper bounds for the hydrophicity, polarity and volume values, respectively. C represents the charge value of the residue.

We selected Kyte-Doolittle hydropathy profile [18], the Grantham's profile [19] for polarity and Klein's scale for net charge [20]. The Dawson's scale [21] is employed to determine the volume of the residues. In table 1, we can appreciate the amino acid values for each property according to the cited scales and normalized between -1 and 1 for hydrophobicity and polarity, and between 0 and 1 for the residue volume.

From all the extracted data, we have built a file in arff format, with all the training data information. This file is available at *http://www.upo.es/eps/asencio/data/training_set.arff*. In this file we include protein subsequences of windows of six amino acids codified with the values of the cited four different physico-chemical properties. The positive class (contact) is represented with 1 and the negative class (no contact) is represented with 0. The total data set constitutes $123,949$ instances with $6,922$ positive and $117,027$ negative cases (contact and no contacts respectively).

3.2 Fitness Function

As already mentioned, we consider two objectives to be optimized: coverage and accuracy. Coverage represents the number of predicted contacts and accuracy evaluates the real predicted contacts rate. Therefore, $Coverage = C/C_t$ and $Accuracy = C/C_p$, where C is the number of correctly predicted contacts of a

Table 1. Values of different properties according to the cited scales for each amino acid. H represents the hydrophobicity, P the polarity, C the charge net and V is the residue volume.

Prop.	A	C	D	E	F	G	H	I	K	L
H	0.40	0.56	-0.78	-0.78	0.62	-0.09	-0.71	1.00	-0.87	0.84
P	-0.21	-0.85	1.00	0.83	-0.93	0.01	0.36	-0.93	0.58	-1.00
C	0	0	-1	-1	0	0	0	0	1	0
V	0.33	0.40	0.33	0.67	0.87	0.07	0.80	0.73	0.93	0.73

Prop.	M	N	P	Q	R	S	T	V	W	Y
H	0.42	-0.78	-0.36	-0.78	-1.00	-0.18	-0.16	0.93	-0.20	-0.30
P	-0.80	0.65	-0.23	0.38	0.38	0.06	-0.09	-0.75	-0.88	-0.68
C	0	0	0	0	1	0	0	0	0	0
V	0.80	0.67	0.73	0.80	1.00	0.40	0.67	0.67	0.93	0.93

protein, C_t is the total number of contacts of the protein and C_p is the number of predicted contacts. We aim at finding the best compromise between these two measures.

3.3 Genetic Operators

We use two mutation operators. The first operator follows a Gaussian distribution for a randomly selected individual and increase or decrease a gene value with a probability of 0.5. A second operator randomly selects a gene that is related to a given property, and moves the bounds to the maximum or minimum of the domain, making the property irrelevant in this rule. For example, if the property is the hydrophobicity, we change the range to -1, 1 so the rule does not take into account this property in this case. This type of mutation is applied with a 0.1 probability. For each individual, we test that the mutated value was between the allowed ranges.

A 2-point crossover operation was used with a 0.5 probability. A binary tournament selection is applied with a probability of 0.5. In each tournament, we select the individual which is located in the better Pareto front. If the two individuals are on the same front, we use the crowding distance to determine the winning configuration. The crowding distance is a measure of the diversity of the population. This process is called Stacking tournament selection.

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. At the end of the execution, repeated or redundant rules are discarded from the solution set.

All the parameters were set after having performed several trial runs of the algorithm.

4 Experiments and Results

As mentioned in the previous section, in order to test our proposal, we have selected a protein data set specified in [3]. This data set consists of 173 proteins with percentage of sequence identity lower than 25%. Four subsets have been classified according to the sequence length; lower than 100 residues (DS1), between 100 and 170 (DS2), between 170 and 300 (DS3), and higher than 300 residues (DS4). The minimum and maximum size of the proteins are 31 and 753 amino acids respectively. A threshold of 8 angstroms (Å) was established to determine a contact. In order to avoid the effect of learning local contacts, we set a minimum sequence separation of 7 residues between each pair of amino acids to establish a contact. A 3-fold cross-validation were performed during all the experimentations. All these requirements were also found in [3]. In order to validate our experimentations, accuracy and coverage rates were calculated. These two measures are also employed to validate the prediction algorithms in CASP competitions.

We have performed several experiments with three Weka classifiers [22]: Näive Bayes (NB), C4.5 classifier tree (J48), Nearest Neighbor approach with $k = 1$ (IB1). The obtained results can be seen in Table 2 for a 3-fold cross-validation. We appreciate low coverage and accuracy values in all the cases. The training data used contained all the possible subsequences of size 6 of the DS1 protein data set with a minimum separation between contact residues of 7 amino acids. This experiment was performed with the aim of validating our representation and confirms that this representation provides enough information for a good performance of a learning classifier. Moreover, we can also notice that PSP-NSGAII achieved the best results for this experiment.

Table 2. Average results obtained for different classification Weka algorithms for the DS1 protein data set

Algorithm	Data Set	$Coverage_\mu$	$Accuracy_\mu$
J48	DS1	0.03	0.31
IB1	DS1	0.09	0.09
NB	DS1	0.20	0.13
PSP-NSGAII	DS1	0.21	0.33

The optimal number of rules for the prediction is unknown. In order to establish the optimal number of executions, we have run several preliminary experiments and compared the obtained results. From these, we have concluded that the best results were obtained when the algorithm was run for 1,000 executions.

Table 3 shows the average results obtained using the dataset. Our results were compared with the ones showed in [3]. We can observe as main conclusion, how the coverage and accuracy rates decrease if the size of the proteins increases. This is due to the fact that, generally, *ab initio* methods only work well with peptides lower than 150 amino acids [23]. We obtain better results for proteins whose

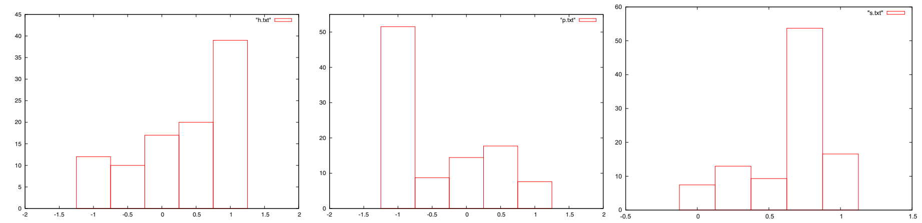
Table 3. Average results and standard deviation obtained for 1,000 executions of the algorithm for the different protein data subsets

Data Set	#proteins	$Coverage_{\mu\pm\sigma}$	$Accuracy_{\mu\pm\sigma}$	$Accuracy_{\mu}$[3]
DS1	65	$0.21_{\pm 0.02}$	$0.33_{\pm 0.01}$	0.26
DS2	57	$0.10_{\pm 0.01}$	$0.21_{\pm 0.02}$	0.21
DS3	30	$0.08_{\pm 0.03}$	$0.13_{\pm 0.02}$	0.15
DS4	21	$0.06_{\pm 0.03}$	$0.09_{\pm 0.03}$	0.11

sequence length is lower than 100 (DS1), 0.33 against 0.26. We have obtained the same accuracy rate for the second subset DS2, and similar rates for the third and fourth group. We could not compare the coverage rates, because they are not included in the cited paper [3].

We have analyzed the set of resulting rules, and they show that a vast majority of amino acids in contact have high values of hydrophobicity. On the other hand, a high percentage of contacts have non-polar residues. These conclusions were expected, because hydrophobic and non-polar amino acids tend to be located in the inner of the protein. Therefore, these type of residues have more probabilities to be in contact [6]. According to the residue volume, residues with values between 0.5 and 0.75 are the most representative. We have not observed any clear conclusion according to the net charge. Although the amino acids with opposite charges are supposed to be in contact [6], this condition seems to be irrelevant in our rule set and does not appear as a clear conclusion. Figure 2 shows the graphical representation of the probability of appearance of each property in our whole set of resulting rules for the amino acid i. The properties values have been discretized in five groups in intervals of 0.5 from -1 to 1 for the hydrophobicity and polarity and from 0 to 1 in intervals of 0.25 for the residue volume. The rest of amino acid positions in the rules presents similar behaviors.

In figure 3, we show a graph which represents the different Pareto fronts for five generations (from generation 0 to 80 with an interval of 20) of an execution in order to test the correct performance of our multi-objective evolutionary algorithm. Each different symbol represents an individual of the Pareto front in different generations. The X-axis represents the coverage and the Y-axis shows the accuracy rate. These two measures are the two parameters which should be optimized during the executions. We can notice how the quality of individuals improve with the generations.



(a) Hydrophobicity **(b)** Polarity **(c)** Residue volume

Fig. 2. Representation of appearance percentages for the different properties at the i-position of the rules

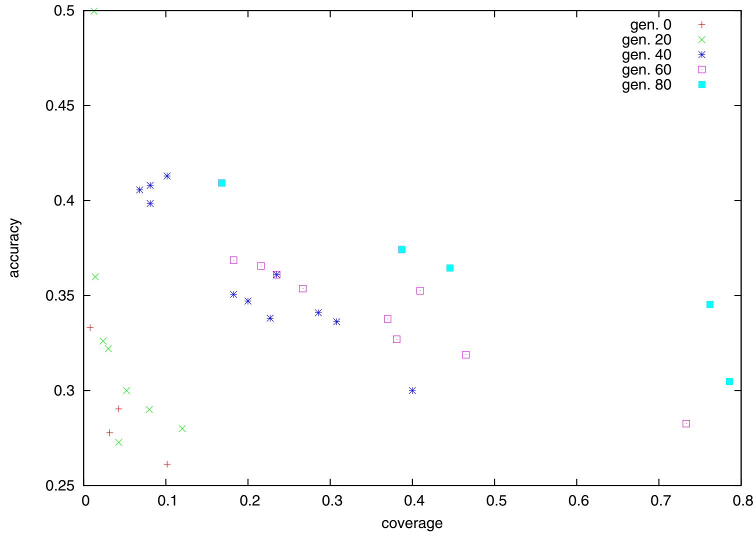


Fig. 3. First Pareto fronts for an execution in different generations

5 Conclusions and Future Work

In this work, we presented a multi-objective optimization algorithm for the residue-residue contact prediction. This algorithm generates rules that predict the necessary conditions for the contact between two amino acids based on their physico-chemical properties. The algorithm was tested on a set of protein sequences that had been previously used in the literature and achieve similar coverage and accuracy rates than other contact map predictor algorithm. We have analyzed the resulting rules set and drawn some conclusions about the folding prediction problem. From the obtained results, we can conclude that our algorithm, as other *ab initio* methods, obtains lower accuracy if the size of the protein is increased. Although these methods are computationally expensive, they have a main advantage; by only taking the sequence as baseline information, it is possible to obtain a folding model for an unknown protein.

As future work, we are planning to include more useful information based on amino acid properties in our rules representation as secondary structure prediction and solvent accessibility. The variability of the window size must be taken into account for the next version of the algorithm. Furthermore, our algorithm must be validated with a higher number of proteins data set.

Acknowledgements. This research was supported by by the Junta de Andalucia, 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”.

References

1. Anfinsen, C.: The formation and stabilization of protein structure. The Biochemical Journal 128, 737–749 (1972)
2. Bashan, A., Yonath, A.: Ribosome crystallography: From early evolution to contemporary medical. Ribosomes Structure, Function, and Dynamics, 3–18 (2011)
3. Fariselli, P., Olmea, O., Valencia, A., Casadio, R.: Prediction of contact map with neural networks and correlated mutations. Protein Engineering 14, 133–154 (2001)
4. Tegge, A.N., 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)
5. Cheng, J., Baldi, P.: Improved residue contact prediction using support vector machines and a large feature set. Bioinformatics 8, 113 (2007)
6. 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)
7. Zhang, Y.: I-tasser: fully automated protein structure prediction in casp8. Proteins: Structure, Function, and Bioinformatics 77, 100–113 (2009)
8. Kinch, L.N., Shi, S., Cheng, H., Qian Cong, Q., Pei, J., Mariani, V., Schwede, T., Grishin, N.V.: Casp9 target classification. Proteins: Structure, Function, and Bioinformatics 79, 21–36 (2011)
9. 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)
10. Unger, R., Moult, J.: Genetic algorithms for protein folding simulations. Biochim. Biophys. 231, 75–81 (1993)
11. Zhang, G., Han, K.: Hepatitis c virus contact map prediction based on binary strategy. Comp. Biol. and Chem. 31, 233–238 (2007)
12. Konak, A., Coit, D.W., Smith, A.E.: Multi-objective optimization using genetic algorithms: A tutorial. Reliability Engineering and System Safety 91(9), 992–1007 (2006)
13. Judya, M.V., Ravichandrana, K.S., Murugesan, K.: A multi-objective evolutionary algorithm for protein structure prediction with immune operators. Comp. Methods in Biomechanics and Biomedical Engineering 12(4), 407–413 (2009)

14. Calvo, J.C., Ortega, J.: Parallel protein structure prediction by multiobjective optimization. Parallel, Distributed and Network-based Processing 12(4), 407–413 (2009)
15. Shi, S., Suganthan, N.: Parallel protein structure prediction by multiobjective optimization. KanGAL Report 7, 1–7 (2004)
16. Cutello, V., Narzisi, G., Nicosia, G.: A multi-objective evolutionary approach to the protein structure prediction problem. J. R. Soc. Interface 3, 139–151 (2006)
17. Deb, K., Agrawal, S., Pratap, A., Meyarivan, T.: A Fast Elitist Non-Dominated Sorting Genetic Algorithm for Multi-Objective Optimization: NSGA-II. In: Deb, K., Rudolph, G., Lutton, E., Merelo, J.J., Schoenauer, M., Schwefel, H.-P., Yao, X. (eds.) PPSN VI 2000. LNCS, vol. 1917, pp. 849–858. Springer, Heidelberg (2000)
18. Kyte, J., Doolittle, R.F.: A simple method for displaying the hydropathic character of a protein. J. J. Mol. Bio. 157, 105–132 (1982)
19. Grantham, R.: Amino acid difference formula to help explain protein evolution. J. Mol. Bio. 185, 862–864 (1974)
20. Klein, P., Kanehisa, M., DeLisi, C.: Prediction of protein function from sequence properties: Discriminant analysis of a data base. Bioch. Bioph. 787, 221–226 (1984)
21. Dawson, D.M.: The Biochemical Genetics of Man. Brock, D.J.H., Mayo, O., eds. (1972)
22. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The weka data mining software: An update. SIGKDD Explorations 11 (2009)
23. Fernandez, M.A., Paredes, A.B., Ortiz, L.R., Rosas, J.L.: Sistema predictor de estructuras de proteínas utilizando dinámica molecular (modypp). Revista Internacional de Sistemas Computacionales y Electrónicos, 6–16 (2009)