Alpha Helix Prediction Based on Evolutionary Computation

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. Multiple approaches have been developed in order to predict the protein secondary structure. In this paper, we propose an approach to such a problem based on evolutionary computation. The proposed approach considers various amino acids properties in order to predict the secondary structure of a protein. In particular, we will consider the hydrophobicity, the polarity and the charge of amino acids. In this study, we focus on predicting a particular kind of secondary structure: α-helices. The results of our proposal will be a set of rules that will identify the beginning or the end of such a structure.

Keywords: Protein Secondary Structure Prediction, α-helix, Evolutionary Computation.

1 Introduction

Bioinformatics has been described as the science of managing, mining, and interpreting information from biological sequences and structures [1]. Two important fields are considered in Bioinformatics: Genomics and Proteomics. Genomics is the study and analysis of the genomes of organisms, while Proteomics is defined as the characterization and identification of the proteins encoded in a genome.

Proteins are one of the basic components in all organisms. Proteins form the basis of cellular life since they significantly affect the structural and functional characteristics of different cells and genes. The structure of a protein is divided into four hierarchy levels. At the first level, proteins are composed of linear sequences of amino acids linked by natural peptide links. This is known as the primary structure of the protein.

The change in one amino acid in a critical area of the protein may alter the biological function, as the higher level structures of the proteins are determined by the primary structure. The secondary structure of a protein is the consequence of the polypeptide chain folding. At this level, some protein structures like α-helices, β-sheets, turns and coils are present. The tertiary structure is the three-dimensional shape of the chain, while the quaternary structure is the final three-dimensional structure composed by all polypeptides chains that form a protein [1,2].

With the success of the genome sequence projects, the amount of available proteins sequences has increased dramatically. However, the number of protein structures available is relatively small. This is due to the difficulty of predicting the structures that a protein will assume based only on its amino acid sequence. This implies that it is crucial to develop computational methods for automatically predict the 3D structure of proteins from their sequences. Knowledge of protein structure has great importance to the development of new drugs.

The problem of protein secondary structure prediction (PSSP) consists in predicting the location of α-helices, β-sheets and turns from a sequence of amino acids without any knowledge of the tertiary structure of the protein. PSSP has received much attention lately, since knowledge of the location of the elements in secondary structure could be used by approximation algorithms to obtain the tertiary structure of the protein. Being able to predict, from the amino acid sequence, how a protein will fold, is one of the main open problems in computational biology.

Several methods were applied to the PSSP problem. These methods can be divided into two categories: statistical and soft computing approaches. Statistical methods are based on the calculation of amino acid probabilities to belong to a secondary structure motif [3,4,5]. Soft computing provides processing capabilities in order to solve the problem of PSSP. The most popular soft computing paradigms for PSSP are: artificial neural networks (ANNs) [6,7,8], evolutionary computation [9], nearest neighbors [10,11] and support vector machines (SVMs)[12,13]. Some soft computing methods used in this problem are focused on determining contact maps (distances) between amino acids residues of a protein sequence. When a contact map is defined, proteins can be fold and the tertiary structure can be obtained.

In this paper, we propose a method, based on an evolutionary algorithm (EA), to predict α-helices from sequences of amino acids. We believe that EAs are good candidate form tackling this problem. In fact, PSSP can be seen as a search problem, where the search space is represented by all the possible folding rules. Such a space is very complex, and has huge size. EAs have proven to be particularly good in this kind of domains, due to their search ability and their capability of escaping from local optima.

In our proposal, prediction is made *ab initio*, i.e., without any known protein structure as a starting template for the search. The prediction model will consist in rules that predict both the beginning and the end of the regions corresponding to an α-helix. Existing methods fail in the α-helix boundaries prediction [14]. In a future development of the algorithm, we also intend to evolve rules for predicting β-sheets.

Previously, some evolutionary approaches have been applied to secondary structure prediction. In [15], a torsion angle representation representation was used. Torsion angles, denoted as (Φ, Ψ), represent the atom position of an amino acid chain, determining the polypeptid arquitechture chain. A possible representantion can be $[(\Phi_1, \Psi_1)...(\Phi_n, \Psi_n)]$ where n represents the total number

of residues in a protein. The values that Φ and Ψ can assume are limited, since atom colissions must be avoided according to Ramachandran chart [16]. In lattice models developed in [9], each element location can be represented as a vector $(x_1, y_1)...(x_n, y_n)$ where x and y are the coordinates of each amino acid in a 2-dimensional lattice (or three coordinates in a 3-dimensional lattice).

The rest of paper is organized as follow. In section 2, we discuss our proposal to predict protein secondary structure motifs. Section 3 provides the experimentation and the obtained results. Finally, in the last section, we draw some conclusions and analyze possible future works.

2 Our Proposal

In this section, we present our proposal for the prediction of α-helices. An α-helix corresponds to a subsequence of amino acids, as shown in figure 1. Each amino acid in the sequence is identified by its position, being amino acids in positions N-cap and C-cap those that immediately precede or follow the beginning or the end of the structure, respectively.

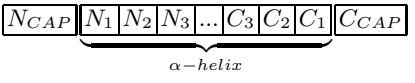

Fig. 1. Relevant positions in an α-helix

Figure 2 represents our experimental procedure to predict protein secondary structure. First, the α-helix sequences are obtained from the Protein Data Bank (PDB) [17], as described in the following sections. These data constitute the training set. Then, our EA is applied and a set of rules are generated. We generate rules for predicting the beginning and the end of an α-helix separately. At the end of the EA, a set of rules will be extracted.

In the following we discuss the various solutions we adopted for what regards the fitness, the representation and the genetic operators used.

2.1 Encoding

In our approach, each individual may represent either the beginning or the end of an α-helix. Namely, each individual represents three properties of amino acids in positions N-cap, N1 or C1, C-cap. These are the limits of an α-helix sequence. The represented properties are hydrophobicity, polarity and charge. These properties have been shown to have certain relevance in PSSP [1,2]. We use Kyte-dolitle hydropathy profile for the hydrophobicity [18]. We have selected Grantham's profile [19] for polarity and Klein's scale for net charge [20]. The values of the properties are then normalized to a range of between -1 and 1 for hydrophobicity and polarity. Three values are used to represent the net charge of a residue: -1 (negative charge), 0 (neutral charge) and 1 (positive charge).

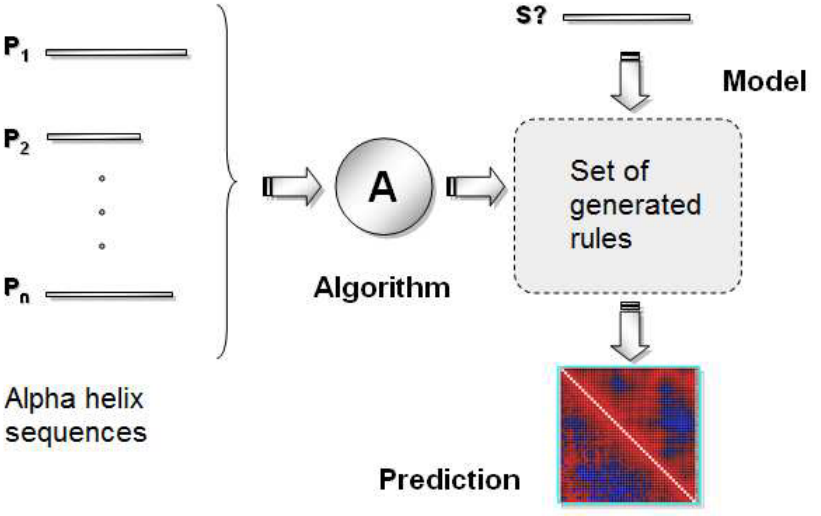

Fig. 2. Experimental and prediction procedure

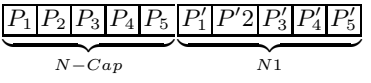

Fig. 3. Example of chromosome codification for a beginning of an α-helix

So, for instance, in figure 3, positions P_1, P_2, P'_1, P'_2 represent the hidrophobicity values of the first and second amino acid respectively. Positions P_3, P_4, P'_3, P'_4 represent the polarity values according to Grant scale of the first and second amino acid respectively. Finally, positions P_5 and P'_5 represents the net charge property values of the two amino acids.

2.2 Fitness Function

The aim of the algorithm is to find both general and precise rules for identifying helices. To this aim, we have chosen as fitness of individuals the F-measure, which is given by the following formula:

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

The higher the fitness, the better the individual. Recall represents the proportion of training examples that matches this rule. Precision represents the error rate.

Moreover, we also consider some physical-chemical properties (polarity and charge) information of the amino acids in positions N-Cap, N1 or C1, C-Cap, if the rule is relative to a beginning or an end of a helix, respectively. It has been demonstrated that there are molecules with asymmetrical distributions of

charge in the limits of an α-helix [21]. This means that the residues in limits of the helix are polar, so the fitness of these individuals is increased. Moreover, in [22,23], it has been proven that many helices present a positive charge in its last turn and a negative charge at its first turn.

We increase the score of those individuals that fulfill one requirements in a 50%, and in a 100% for those individuals that present the two properties.

2.3 Genetic Operators

Individuals are selected with a roulette wheel mechanism. A roulette wheel is built, where the sector associated with each individual of the population is proportional its fitness. Individuals with higher fitnesses have more probability of being selected, having wider sectors associated to them.

Uniform crossover is used in order to generate offsprings. Crossover is applied with a 1.0 probability. All the offsprings are made by crossover except the one with best score which was copied without any change (elitism). Mutation is applied with a probability of 0.5. If mutation is applied, one gene of the individual is randomly selected, and its value is increased or decreased by 0.01. If the selected gene is relative to the charge of the amino acid, then its value is randomly changed to one of the other two allowed possibilities. After that an individual has been mutated, it is checked for validity, i.e., its values are within the ranges allowed for each properties. If the encoded rule is not valid, then the mutation is discarded.

The initial population is randomly initialized. After having evaluated the initial population, the first generation is created. If the fitness of the best individual does not increase for twenty generations, the algorithm is stopped and a solution is provided.

We evolve two populations separately: one population contains individuals that encode rules identifying the beginning of an α-helix, while the other population contains individuals representing rules for the end of the helix. At the end of the evolutionary process, the best individuals from each population are extracted, and together they form the proposed solution.

3 Experiments and Discussion

In this section, we present the experimentation performed in order to assess the validity of our proposal.

In order to test the proposed algorithm, we have used data obtained from PDB. Protein secondary structure is obtained from amino acid sequences, as well as the distances between pairs of amino acids. All this information is included in the PDB site. The Worldwide PDB [24], is an international collaboration organized by the processing and distribution of the PDB file. The online PDB file [17] is the repository that coordinates and related information on nearly $65,000$ structures ($65,378$ structures in May 18, 2010), including proteins, nucleic acids and complex macromolecules that have been obtained through

techniques of X-ray crystallography, NMR (nuclear magnetic resonance) and electron microscope.

We have obtained a set of 12,830 non-homologous different protein sequences with an homology lower than 30%, using the PDB Advanced Search [25]. We have only selected the structures which contains protein chains and not DNA or RNA chains using the Macromolecule type option. We reject the redundant sequences. The complete list of the 12,830 PDB protein identifiers can be downloaded in [26]. We parsed the required information from PDB files. At the Secondary Structure Section of PBD, different α-helix sequences of each protein can be obtained with the HELIX command. Once we have located the motifs in the protein sequence, we extract from this sequence, the amino acids from N-cap to C-cap positions of the helix (figure 1), which are relevant positions in a α-helix [21]. We have selected a subset of 5,000 α-helices sequences from a subset of proteins with length less than 150 residues from these 12,830 proteins. Each of these 5,000 sequences includes a begining and an end of helix. Thus, we have 5,000 windows of two amino acids in C-cap, C1 positions and 5,000 windows of two amino acids in N1, N-cap positions. These sequences represent our training data. The average size of the α-helix sequences is 9.86 residues.

A 10-fold cross-validation has been applied. The data set is divided into 10 subsets, and the holdout method is repeated 10 times. Each time, one of the 10 subsets is used as the test set and the other 9 subsets are put together to form a training set. Then the average result across all 10 trials is computed.

For each fold, we obtained the confusion matrix. Each column of the matrix represents the number of true or false predictions of a class, and each row represents the number of real instances. More specifically, the matrix contains information about the True Negatives (TN), False Positives (FP), False Negatives (FN), and True Positives (TP). TN is the number of correct predictions for a negative case (no ends or beginnings), FP is the number of incorrect predictions for a positive case (ends or beginnings of an helix), FN is the number of incorrect predictions for a negative case (no ends or beginnings) and TP is the number of correct predictions for a positive case (ends or beginnings of an helix).

For each fold, we compute the following results:

- Recall represents the percentage of correctly identified positive cases. In our case, Recall indicates what percentage of motifs has been correctly identified.

$$Recall = \frac{TP}{TP + FN}.$$

- Precision is a measure of false positive rate. Precision reflects the number of real predicted examples.

$$Precision = \frac{TP}{TP + FP}.$$

- Specificity, or True Negative Rate, measures the percentage of correctly identified negative cases. In this case, Specificity reflects what percentage of no motifs has been correctly identified.

$$Specificity = \frac{TN}{TN + FP}.$$

– Accuracy is also calculated.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}.$$

At each execution, a model is obtained. This model consists of two rules, one that identifies the beginning of an α-helix, and the other that identifies the end of such a structure. Since the number of rules needed to provide the best results is unknown, we have performed different experiments with different number of runs of the algorithm, namely from 10 to 40. So, for instance, after the experiments with 10 runs, a model with twenty rules is obtained, where half of the rules represent the beggining of an α-helix and the other half represent the end.

Table 1 and 2 show the obtained results relative to the N-cap and C-cap prediction, respectively. The first column specifies the number of execution of the algorithm, the second column gives the average recall obtained. The third and fourth columns provide the average specificity and precision, respectively. The last column is relative to the average accuracy obtained. For each measure, the standard deviation is also provided.

From tables 1 and 2, it can be noticed that the model provided by the algorithm is always very accurate, in fact, the average accuracy obtained is very high in all the cases, being the average 0.99. The precision of the model is also satisfactory, with an average of 0.70. This means that model obtained commits few classification errors. The average recall is about 0.60 for beginnings and 0.58 for ends of helix, which represents a good result es well, and it means that on average, 60% of the α-helices are recognized as such. We can also notice

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

Executions	$Recall_{\mu\pm\sigma}$	$Spec._{\mu\pm\sigma}$	$Prec._{\mu\pm\sigma}$	$Accuracy_{\mu\pm\sigma}$
10	$0.5525_{\pm 0.0437}$	$0.9895_{\pm 0.0005}$	$0.6553_{\pm 0.0232}$	$0.9935_{\pm 0.0008}$
20	$0.6212_{\pm 0.1156}$	$0.9924_{\pm 0.0007}$	$0.6857_{\pm 0.0220}$	$0.9948_{\pm 0.0015}$
30	$0.6275_{\pm 0.0922}$	$0.9948_{\pm 0.0005}$	$0.7368_{\pm 0.0315}$	$0.9940_{\pm 0.0016}$
40	$0.6025_{\pm 0.0848}$	$0.9937_{\pm 0.0006}$	$0.7320_{\pm 0.0372}$	$0.9937_{\pm 0.0013}$

Table 2. Average results and standard deviation obtained for different number of executions of the algorithm for C-cap prediction

Executions	$Recall_{\mu\pm\sigma}$	$Spec._{\mu\pm\sigma}$	$Prec._{\mu\pm\sigma}$	$Accuracy_{\mu\pm\sigma}$
10	$0.5933_{\pm 0.0565}$	$0.9889_{\pm 0.0005}$	$0.6338_{\pm 0.0218}$	$0.9955_{\pm 0.0007}$
20	$0.5728_{\pm 0.1185}$	$0.9943_{\pm 0.0006}$	$0.6589_{\pm 0.0250}$	$0.9952_{\pm 0.0018}$
30	$0.5936_{\pm 0.0933}$	$0.9935_{\pm 0.0006}$	$0.6859_{\pm 0.0302}$	$0.9972_{\pm 0.0020}$
40	$0.5870_{\pm 0.0848}$	$0.9925_{\pm 0.0006}$	$0.7005_{\pm 0.0450}$	$0.9966_{\pm 0.0015}$

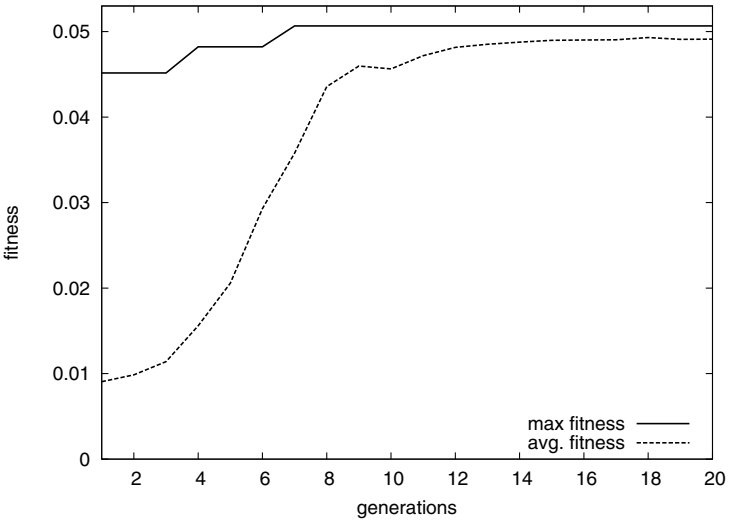

Fig. 4. Maximum Fitness vs. Average Fitness

that producing a model with more rules (the more executions the more rules will be part of the model produced) does not neccessarily help in increasing the precision. For the rest of the measures, the results become more or less stable after 20 executions of the algorithm.

Our algorithm is capable of producing satisfactory results using an elevated number of sequences (5,000 beginnings and 5,000 ends of helix sequences). This is, in our opinion, an important result, since the number of protein sequences available increase by the day, and thus, having a method that is scalable would be very important.

Other approaches were developed to predict starts of helix. The start position are correctly predicted for approximately 30% of all predicted helices in [14]. The number of correctly predicted alpha-helix start positions was improved from 30% to 38% in [27]. These results are widely exceeded by our approach, as our algorithm predicts about 60% of the start positions correctly. We have not found references for the C-cap helix prediction in literature.

It is also interesting to inspect the behavior of our EA. Figure 4 shows a graphical representation of the maximum and average fitness values at different generations relative to a typical run. We can notice that the maximum fitness is achieved very early, at about generation seven, and then it is stable. This may suggest that we should try to increase the mutation probability, or apply a mutation operator that introduces more changes in an individual, in order to increase diversity in the population. Another estrategy, could be to apply some local search method with a given probability. Such local search would help in improving the fitness of the individuals.

On the other hand, the average fitness increases constantly, and tends to converge to the maximum fitness toward the end of the run.

4 Conclusions and Future Work

In this paper, we have proposed an evolutionary algorithm for the prediction of α-helix motifs in protein sequences. The algorithm incorporates in the fitness three amino acids properties: hydrophocity, polarity and net charge. These properties have been shown to be relevant in the determination of the beginning and end of helices, and thus they helped to improve the search process performed by the algorithm.

We have performed experiments using a set of 5,000 α-helix sequences extracted from a protein data set from Protein Data Bank composed by 12,830 non-redundant and non-homologous protein with an homology rate lower than 30%. To the best of our knowledge, no other approaches have used such an high number of sequences in α-helix capping regions prediction. Results obtained on this data set are encouraging and in particular, the accuracy characterizing the prediction models obtained is very high independently from the number of rules generated.

As for future development, we are analyzing different properties to be included in the fitness function in order to increase the quality of the prediction model. Moreover, we are studying the possibility of incorporating a local search phase that will help to improve individuals. We also intend to extend our experimentation to other datasets of protein sequences and we want to expand the number of residues in the window of amino acids. Finally, we also want to produce a model for the prediction of both α-helices and β-sheets.

Acknowledgements

This research was supported by the Project of Excellence P07-TIC-02611 "Sistemas Inteligentes para descubrir patrones de comportamiento. Aplicación a base de datos biológicas" and 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. Gu, J., Bourne, P.E.: Structural Bioinformatics (Methods of Biochemical Analysis). Wiley-Blackwell, Chichester (2003)
2. Berg, J.M., Stryer, L.: Biochemistry. Reverte (2008)
3. Chou, P.Y., Fasman, G.D.: Prediction of protein conformation. Biochemistry 13(2), 222–245 (1974)
4. Garnier, J., Osguthorpe, D.J., Robson, B.: Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. J. Mol. Biol. 120, 97–120 (1978)
5. Lim, V.I.: Algorithms for prediction of a-helical and b-structural regions in globular proteins. J. Mol. Biol. 88, 857–872 (1974)
6. Qian, N., Sejnowski, T.J.: Predicting the secondary structure of globular proteins using neural network models. J. Mol. Biol. 202, 865–884 (1988)

7. McGuffin, L.J., Bryson, K., Jones, D.T.: The psipred protein structure prediction server. Bioinformatics 16, 404–405 (2000)
8. Fariselli, P., Casadio, R.: A neural network based predictor of residue contacts in proteins. Protein Engineering 12, 15–21 (1999)
9. Unger, R., Moult, J.: Genetic algorithms for protein folding simulations. Biochim. Biophys. 231, 75–81 (1993)
10. Frishman, D., Argos, P.: Incorporation of non-local interactions in protein secondary structure prediction from the amino acid sequence. Protein Engineering 9, 133–142 (1996)
11. Salamov, A.A., Solovyev, V.V.: Protein secondary structure prediction using local alignments. J. Mol. Biol. 268, 31–36 (1997)
12. Ward, J.J., McGuffin, L.J., Buxton, B.F., Jone, D.T.: Secondary structure prediction with support vector machines. Bioinformatics 13, 1650–1655 (2003)
13. Cheng, J., Baldi, P.: Improved residue contact prediction using support vector machines and a large feature set. Bioinformatics 8, 113 (2007)
14. Wilson, C.L., Hubbard, S.J., Doig: A critical assessment of the secondary structure prediction of alpha-helices and their n-termini in proteins. Protein Eng. 15, 545–554 (2002)
15. 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)
16. Ramakrishnan, C., Ramachandran, G.N.: Stereochemical criteria for polypeptide and protein chain conformation. Byophys Journal 5, 909–933 (1965)
17. Protein data bank online repository, `ftp://ftp.wwpdb.org`
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. 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. Biochim. Biophys. 787, 221–226 (1984)
21. Richardson, J.S., Richardson, D.C.: Amino acid preferences for specific locations at the ends of alpha helices. Science 240, 1648–1652 (1998)
22. Doig, A.J.: Baldwin R.L. N- and c-capping preferences for all 20 amino acids in alpha-helical peptides. Protein Science 4(7), 1325–1336 (1995)
23. Fonseca, N.A., Camacho, R., Magalhaes, A.L.: Amino acid pairing at the n- and c-termini of helical segments in proteins. Proteins 70, 188–196 (2007)
24. Protein data bank web, `http://www.wwpdb.org`
25. Protein data bank advanced search, `http://www.pdb.org/pdb/search/advSearch.do`
26. Complete list of pdb protein identifiers used in this article, `http://www.upo.es/eps/marquez/proteins.txt`
27. Wilson, C.L., Boardman, P.E., Doig, A.J., Hubbard, S.J.: Improved prediction for n-termini of alpha-helices using empirical information. Proteins 57(2), 322–330 (2004)