Protein Secondary Structures Prediction based on Evolutionary Computation

Alfonso E. Márquez Chamorro
School of Engineering
Pablo de Olavide University
Seville, Spain
amarcha@upo.es

Federico Divina
School of Engineering
Pablo de Olavide University
Seville, Spain
fdiv@upo.es

Jesús S. Aguilar-Ruiz
School of Engineering
Pablo de Olavide University
Seville, Spain
aguilar@upo.es

ABSTRACT

In this paper we propose an approach based on evolutionary computation for the prediction of secondary protein structure motifs. The prediction model consists of a set of rules that predict both the beginning and the end of the regions corresponding to a secondary structure state conformation (α-helix or β-strand). The prediction is based on a set of specific amino acid physical-chemical properties. In addition we also propose a statistical study regarding the propensities of each pair of amino acids in capping regions of α-helix and β-strand. Experimental results confirm the validity of our proposal.[1]

Keywords

Protein Secondary Structure Prediction, α-helix, β-strand, β-sheet, Evolutionary Computation.

1. INTRODUCTION

The Protein Secondary Structure Prediction (PSSP) consists in predicting the location of α-helices, β-sheets and turns within 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, and have important applications, e.g., in the development of new drugs.

Repetitive motifs appear in a secondary structure, and the most common kind of motif is α-helices. An α-helix is a dextro-helical structure, with about 3.6 amino acids per turn. Such structure is held together by hydrogen bonds. In particular the amino group of amino acid n provides a hydrogen bond with the carbonyl group of the amino acid n + 4. Another common structure is represented by the β-sheet. β-sheets are characterized by theirs flattened and extended shape, and have a maximum number of hydrogen bonds between peptides that provide stability to the structure. Several peptide chains (β-strands), which are held together with hydrogen bonds in a zig-zag, constitute a β-sheet motif. The lamellar structure formed proportionate flexibility but no elasticity. The adjacent chains of a β-sheet can be targeted in the same direction (parallel β-sheet) or opposite direction (antiparallel β-sheet).

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 given secondary structure motif [5, 12, 17]. On the other hand, soft computing methods provide processing capabilities that can be used in order to solve the problem of PSSP. Such methods are characterized by the fact that they are tolerant of imprecision, uncertainty, partial truth, and approximation. The most popular soft computing paradigms applied to PSSP are: artificial neural networks (ANNs) [19, 18, 8], nearest neighbors [11, 21] and support vector machines (SVMs) [23, 4]. 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 determined.

In this paper, we propose a strategy based on evolutionary computation, to predict α-helices and β-sheets from sequences of amino acids. Evolutionary algorithms (EAs) are adaptive methods that can be used to solve optimization problems. We believe that EAs are good candidate for 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, the prediction is made ab initio, i.e., without any known protein structure as a starting template for the search. The prediction model will consist of rules that predict both the beginning and the end of the regions corresponding to a α-helix or a β-strand. With this we want to overcome the limitation of existing methods that typically fail at predicting motifs boundaries [25]. In particular, β-sheet determination is more difficult to predict than α-helix [9].

Other methods based on evolutionary computation have been applied to secondary structure prediction. For instance, in [6], an EA using a torsion angle representation was proposed, and [22] introduced an approach based on lattice models.

In this paper, we also propose a statistical analysis of each pair of amino acids that are located in the beginning and the end of the α-helix or β-sheets sequences. From this study, we can extract information that is useful in order to predict secondary structures.

The rest of paper is organized as follow. In the next section we propose the statistical analysis of the sequences used in this paper. In section 3, we discuss our proposal to predict protein secondary structure motifs. Section 4 provides the experimentation and the obtained results. Finally, in the last section, we draw some conclusions and analyze possible future works.

2. STATISTICAL ANALYSIS

As already mentioned, the aim of this paper is to propose a method for identifying the beginning and the end of motifs in sequence of amino acids. Figure 1 shows an example of a sub-sequence of amino acids relative to a secondary structure, e.g., an α-helix. Each amino acid in the sequence is identified by its position, and amino acids between N1 and C1 form the α-helix. It follows that amino acids in positions N-cap and C-cap are those that immediately precede or follow the beginning or the end of the structure, respectively.

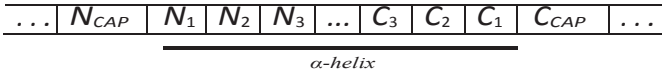

Figure 1: Relevant positions in an α-helix and a β- strand.

The analysis proposed in this section is aimed at studying the different propensities of each pair of amino acids in the studied positions, i.e, N-cap, N1, C-cap and C1. Knowing the propensities of each pair of amino acids at these positions would allow us to extract useful information about the properties of those amino acid that are located in the beginnings or ends of the different secondary structure motifs.

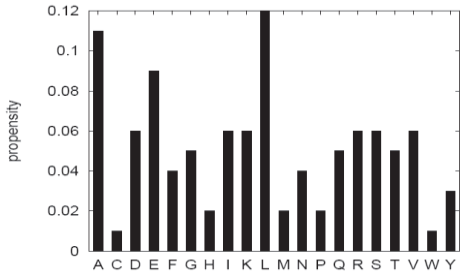

Figure 2: Amino acid propensities in helix sequences.

The data set used in this paper includes 163,461 α-helix and 216,390 β-strand sequences with a total of 2,177,854 and 1,606,246 amino acids respectively. These sequences were extracted using the DSSP program [15] from 12,860 non-redundant protein sequences taken from PDB and sharing less than 30% sequence identity. To the best of our knowledge, no other approaches have used such a high number of secondary structure states sequences for a similar study. Before the cited analysis, we have also performed several studies to analyze certain aspects of the data set. Figure 2 show a chart with the propensities for each amino acid to belong to an α-helix. It can be noticed that A (Alanine) and L (Leucine) show the higher probabilities. It is worth mentioning that these two amino acids are nonpolar and have neutral charge.

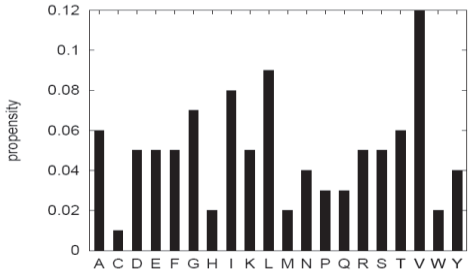

Figure 3: Amino acid propensities in beta sequences.

On the other hand, Figure 3 proposes a graph relative to the amino acid probabilities of belonging to a β-strand. In this case, V (Valine), I (Isoleucine) and L (Leucine) have a probability of 12, 9 and 8%, respectively. It is interesting to notice that also in this case these amino acids have two characteristics in common, nonpolarity and neutral charge. Figure 4 proposes a diagram that represents the ratios aimed at studying the most frequent length of an α-helix sequence. We can observe that the most common length is 6, and that after this peak the graph almost follows a normal distribution with center 13. From this study we have discovered that the average size of α-helix sequences is 13.46 residues. Figure 5 shows a diagram with the representation of the proportions for each size of β-strand sequences. Also in this case the graphs follows a normal distribution, were the average size of β-strands sequences is 7.49 residues.

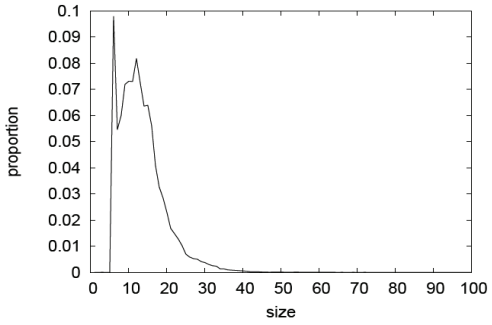

Figure 4: Representation of alpha helix sequences sizes.

Table 1 gives us more insights on propensities of each amino acid to belong to either N-cap or C-cap positions in α-helix and β-strand sequences. These probabilities are computed over the total of appearances of each amino acid. For the α-helix sequences, amino acids D, N, P and S (polar amino acids), present the higher probabilities in the case of N-Cap, while for the for C-cap position G (Glycine) is by far the most probable. As far as β-strand

sequences are concerned, amino acids D, G and P (small amino acids), have the highest probabilities to appear in N -Cap positions. Amino acids D, G and N show the higher propensities for C-cap position.

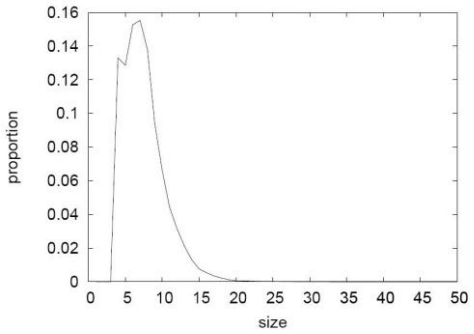

Figure 5: Representation of beta strand sequence sizes.

Table 1: Amino acid propensities for C-cap an N-cap positions in alpha and beta sequences.

	Alpha-helix			Beta-strands	
AA	***NCap***	***CCap***	***AA***	***NCap***	***CCap***
A	0,35	0,85	A	0,98	1,07
C	1,19	1,36	C	0,74	1,04
D	**2,71**	0,80	D	**1,90**	**1,87**
E	0,48	0,64	E	1,13	1,06
F	0,50	1,04	F	0,56	0,61
G	2,00	**3,46**	G	**1,87**	**1,71**
H	1,35	1,46	H	0,99	1,03
I	0,29	0,51	I	0,39	0,49
K	0,48	1,00	K	1,37	0,98
L	0,31	0,82	L	0,51	0,83
M	0,45	0,94	M	0,94	0,87
N	**2,40**	1,78	N	1,72	**1,85**
P	**2,95**	0,00	P	**2,70**	1,45
Q	0,49	0,99	Q	1,18	0,88
R	0,53	0,95	R	1,14	0,88
S	**2,68**	1,25	S	1,04	1,32
T	**2,46**	0,74	T	0,84	1,06
V	0,33	0,56	V	0,40	0,50
W	0,50	0,58	W	0,74	0,61
Y	0,62	1,01	Y	0,63	0,64

We have computed the global propensities for each pair of amino acids in N-cap, N1 positions and C1, C-cap positions (start and end of either a helix or a sheet). In this analysis, we have used the following formula:

$$P^{g}_{X_i Y_{i+1}} = \frac{n^{helix}_{X_i Y_{i+1}}}{\sum_{AB} n^{helix}_{A_i B_{i+1}}} \Big/ \frac{n^{total}_{XY}}{\sum_{AB} n^{total}_{AB}}$$

where $P_{Xi\,Yi+1}$ is the global propensity for a XY amino acid pair, $X_i\,Y_{i+1}$ represents a pair of amino acids in a helix or sheet capping position, and $A_i\,B_{i+1}$ represents a pair of amino acids in any consecutive position in a protein sequence. This equation computes the relative frequency of the amino acid pair XY at positions i, i+1 (N-cap, N1 or C1, C-cap) in helices or strands and the relative frequency of the pair in the total protein set. In order to analyze the propensity of a pair of amino acids to be in a certain position of either a helix or a strand, we used equation 1 and built four propensity matrices, shown in Figures 6, 7, 8 and 9. In these matrices, rows represent the N-cap or C-cap position and columns are relative to either the N1 or the C1 position. Each propensity is represented by a different color. A cell with blue color represents a high likelihood for this pair. On the other hand, a red cell represents a low propensity.

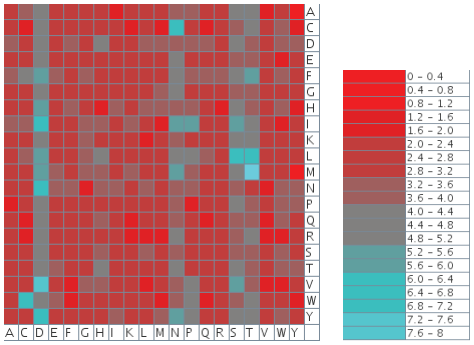

Figure 6: Propensity matrix for N-cap, N1 helix positions.

Figure 6 shows the propensity matrix for N-cap, N1 positions of the helices. From the analysis of this matrix, we can conclude that amino acids D, N ,S and T (Aspartic acid, Asparagine, Serine and Threonine, respectively) are the most likely to occupy the N1 of a helix, while there are no specific amino acids for the N-cap position. This means that every amino acid could hold such a position. By further analyzing these amino acids, we have observed that all these amino acids have a polar side chain. Moreover, the most frequent pair in an α-helix start is M, T (Methionine and Threonine).

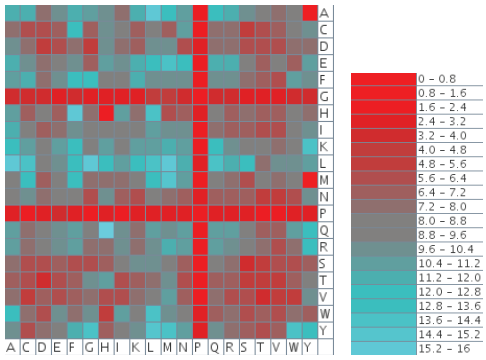

Figure 7: Propensity matrix for N-cap, N1 helix positions.

Figure 7 shows the propensity matrix for C1, C-cap positions of the helices. In this case the amino acid P (Proline) has the lowest propensity in both C1 and C-cap positions and the G amino acid (Glycine) at C1 position. The most frequent pair in an α-helix end is Q, H (Glutamine and Histidine).

Figure 8 shows the matrix for N cap, N1 positions of β-strands. From the analysis of the matrix, we can conclude that the amino acids that are most likely to appear in N1 position are G, P, N and D (Glycine, Proline, Asparagine and Aspartic acid respectively). Also in this case, these amino acids show a common characteristic: they have a small residue. Also in this case no specific conclusion can be drawn for the N-cap position. The most frequent pair in a β-strand start is P, V (Proline and Valine).

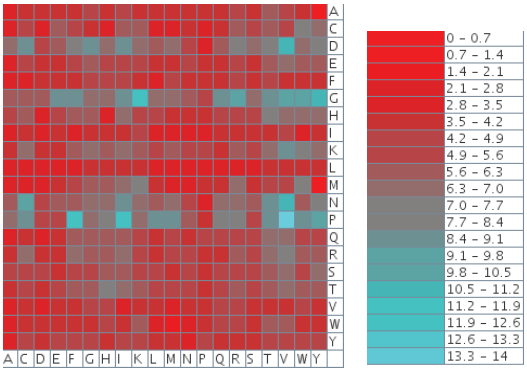

Figure 8: Propensity matrix for Ncap-N1 strand positions.

Figure 9 report the propensity matrix for C1, C-cap positions of a β-strand. In this case the amino acids N, D and P (Proline) have the lowest propensity to be in position C1, while amino acids I, V and Y (Isoleucine, Valine and Tyrosine respectively) shows the highest propensity for this position. In this case, these amino acids are all hydrophobic residues. No particular conclusion can be derived for the C-cap position. The most frequent pair in a β-strand end is Y, C (Tyrosine and Cysteine).

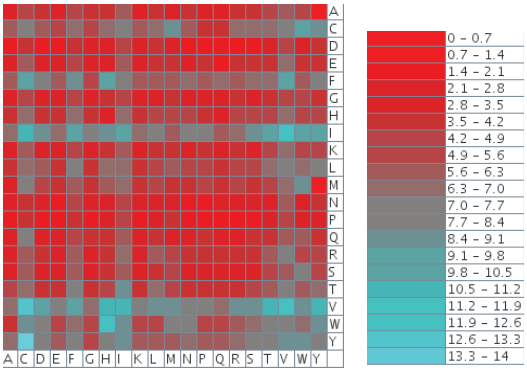

Figure 9: Propensity matrix for C1-Ccap strand positions.

3. MATERIALS AND METHODS

In this section, we present our proposal for the identification of α-helices and β-strands within a sequence of amino acids. α-helix and β-strand are a subsequence of amino acids.

The aim of this paper is to propose an EA capable of finding a set of rules that can be used in order to predict whether or not a particular amino acid lies in either position N-cap or C-cap, for both α-helices and β-sheets. With such a predicting model, we could then precisely identify the beginning and the end of the considered secondary structures.

The experimental procedure followed in this paper is shown in Figure 10. During the data acquisition stage, the α-helix and β-strand sequences are extracted from the Protein Data Bank (PDB) [3], as described in section 2. Later, these sequences are used for training our evolutionary algorithm, which will generate a set of rules representing the predictive model. As already mentioned, these rules establish if a particular amino acid is relative to either a N-cap or a C-cap position, so if the amino acid precede or follow the begin or the end of a structure. In order to test our algorithm, we apply these rules to a set of known protein sequences, thus used as test set.

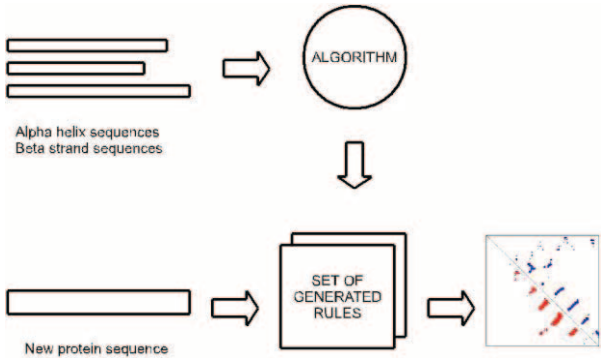

Figure 10: Experimental and prediction procedure.

3.1 Encoding

As already stated, the prediction model proposed by our algorithm will be based on a set of amino acid properties. In particular, we consider three properties:

Hydrophobicity For this property, we use Kyte-Doolitle hydropathy profile [14] for the hydrophobicity representation. In this way the hydrophobicity spectrum is discretized into a set of well defined intervals.

Polarity The Grantham's profile [13] was used for the polarity representation.

Charge Klein's scale [16] for net charge codification. We represent an amino acid with negative charge, according the Klein's scale, with 1, a positive charge with 1 and a neutral charge with 0.

Notice that the profile values of each amino acid are normalized to a range of between 1 and 1 for hydrophobicity and polarity.

Each individual of the population represents a rule, and will consist of window of two amino acids. For each amino acid, the three above mentioned properties will be represented. An individual may represent either the beginning or the end of an α-helix or a β-sheet (N-cap, N1 or C1, C-cap positions).

Figure 11 proposes an example of individual, where positions P_1, P_2, P_1' and P_2' represent the hydrophobicity ranges of the first and second amino acid of the window, respectively. P_1 and P_2 are real numbers which determine a hydrophobicity range for the amino acid in that position. Positions P_3, P_4, P_3' and P_4' represent the polarity intervals according to Grant scale of the first and second amino acid respectively. Also in these cases, these values are real numbers, which determine a polarity range for the amino acid in that position. Positions P_5 and P_5' represent the net charge property values of the two amino acids. These two positions may assume three different integer values: 1 for a negative charge, 0 for neutral charge and 1 for a positive charge.

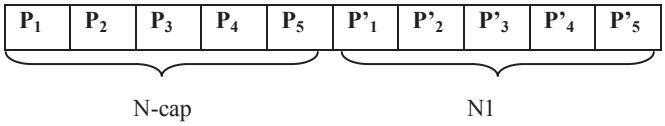

Figure 11: Example of chromosome codification for a beginning of an α-helix or a β-strand.

3.2 Fitness Function

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

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

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

In the literature, it has been proved that α-helices and β-sheets are characterized by some properties of the amino acids in positions N-cap, N1 or C1, C-cap. In order to increase the effectiveness of our proposal, we consider some of these properties. In particular, in [20] it has been demonstrated that there are molecules with asymmetrical distributions of charge in the limits of an α-helix. This means that the residues in limits of the helix are polar. Results obtained by our algorithm confirm this observation. Moreover, in [7, 10], it has been proven that many helices present a positive charge in its last turn and a negative charge at its first turn.

On the other hand, we also consider some specifications for the β β-sheet capping prediction. It has been demonstrated that hydrophobic amino acids have a high propensity to be at N1 and C1 positions (especially V, I, Y and W) in a β-sheet. In addition, many strands present a negative charge in C-cap and a positive charge in N-cap positions [9].

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

3.3 Genetic Operators

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

Elitism is also applied, i.e., the best individual always survives to the next generation.

Uniform crossover is used in order to generate offsprings. Crossover is applied with a 1.0 probability. All the offsprings are obtained 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 property: [1, 1] for hydrophobicity and polarity and 1, 0 or 1 for the net charge. If the encoded rule is not valid, then the mutation is discarded.

The initial population is randomly initiated. For the experiments proposed in this paper, the population size is set to 100. After having evaluated the initial population, the first generation is created. If the fitness of the best individual does not increase over twenty generation, the algorithm is stopped and a solution is provided.

We evolve four populations separately: one population contains individuals that encode rules identifying the beginning of an α-helix, a second population contains individuals representing rules identifying the end of the helix. A third population contains individuals that encode rules identifying the beginning of a β-sheet, and the last population contains individuals representing rules for the end of a β-sheet. At the end of the evolutionary process, the best individuals from each population are extracted, and together they form the proposed solution.

In the following we outline the main solution adopted for the EA proposed. In particular, we discuss the various solutions concerning the fitness, the representation and the genetic operators used.

4. EXPERIMENTS AND DISCUSSION

In this section, we present the experimentation performed in order to assess the validity of our proposal. The prediction of protein secondary structure is obtained from amino acid sequences. For this reason, we need to obtain a set of known protein sequences. We obtain the sequences from the PDB site [3], where the information regarding secondary structures is also provided. However this information will be used only for testing our predicting model.

As explained in section 3, a set of 12,830 non-homologous and non-redundant proteins with a homology lower than 30% were obtained from PDB, using the PDB Advanced Search [2]. We have only selected the structures which contain protein chains and not DNA or RNA chains. The complete list of the 12,830 PDB protein identifiers can be downloaded in [1]. The DSSP program [15] was used in order to extract the secondary structure relative to α-helix and beta-sheet states of each protein based on the atomic coordinates in the PDB file. Once we have located the motifs in the protein sequence, we extract the amino acids from N-cap to C-cap positions of the helix or sheet (figure 1), which are the amino acids that are in relevant positions in an α-helix or beta-sheet. From these sequences, we have randomly selected a subset of 5,000 α-helix and 5,000 β-strand sequences with a minimum size of four residues. These sequences were extracted from a subset of proteins sequences with length less than 150 residues. Coils and no-motifs protein sequences are included as negative examples.

In order to validate the obtained results, 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. A model is obtained for each fold. This model consists of a set of rules that identify beginnings and ends of an α-helix or of a β-strand respectively.

For each fold, we compute the following measures:

Recall represents the percentage of correctly identified positive cases. In our case, Recall indicates what percentage of beginnings or ends of motifs 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 no beginnings or ends of motifs have been correctly identified.

Accuracy is the proportion of true results in the population.

The optimal number of rules necessary for the prediction is unknown. For this reason, we performed experiments with a different number of iterations of the algorithm, more specifically from 10 to 40. Notice that after each iteration a set of rules is provided. The more iterations of the algorithm, the more rules will be incorporated in the final prediction model.

In the experiments proposed in this paper, we used the following parameters for the EA. The population size is set to 100. Crossover and mutation probabilities are set to 1.0 and 0.5, respectively. The maximum number generations is set to 100. These parameters were established after a set of preliminary tests.

Table 2 and Table 3 show the obtained results for the helix capping prediction algorithms (starts and ends of helices). In particular, the first column provide the number of iterations of the algorithm, and the rest of the columns report the average recall, specificity, precision and accuracy. Standard deviation is also reported. It can be noticed that for the α-helix capping prediction, the algorithm obtained extremely high accuracy, with an average of 0.99. The average recall is about 0.64, in C-cap and about 0.62 in N-cap prediction. Precision shows a low rate of error in the prediction with an average of about 0.69. All the measures vary weakly depending on the number of iterations. The results become more or less stable after 20 executions for all the measures. Previous works achieve an average recall of 30 38% in N-cap prediction [24]. Our approach improved these results.

Table 2: Average results for the prediction of the beginning of α-helices obtained for different number of iterations. Standard deviation is reported between brackets.

It.	*Recall μ±σ*	*Spec. μ±σ*	*Prec. μ±σ*	*Accuracy*
10	0.604±0:103	0.993±0.001	0.693±0.024	0.992±0.002
20	0.635±0.096	0.991±0.002	0.687±0.023	0.993±0.002
30	0.638±0.066	0.993±0.000	0.692±0.012	0.992±0.001
40	0.623 ±0.055	0.995±0.000	0.732±0.010	0.992±0.001

Table 3: Average results for the prediction of the end of α-helices obtained for different number of iterations. Standard deviation is reported between brackets.

It.	*Recall μ±σ*	*Spec. μ±σ*	*Prec. μ±σ*	*Accuracy*
10	0.633±0:160	0.993±0.002	0.665±0.022	0.992±0.003
20	0.643±0:196	0.993±0.003	0.694±0.049	0.993±0.004
30	0.656±0.066	0.992±0.002	0.668±0.031	0.992±0.003
40	0.634 ±0:101	0.992±0.001	0.640±0.013	0.992±0.002

Results relative to the prediction of β-strands capping are reported in table 4 and table 5. These results are slightly less accurate than those relative to the helix prediction. The main difference can be noticed in the results obtained for the N-cap and C-cap recall, with an average recall of about 0.18 in N-cap, and about 0.52 in C-cap prediction. So, in the case of the N-cap prediction, the result is much lower than in the case of N-cap prediction of α-helices. The precision is about 0.59, in N-cap and about 0.68 in C-cap prediction. High levels of accuracy and specificity are shown in both cases. Unlike α-helices [24], to the best of our knowledge, there are not previous results reported in the literature for the β-sheet capping prediction.

Table 4: Average results for the prediction of the beginning of β-strands obtained for different number of iterations. Standard deviation is reported between brackets.

It.	*Recall μ±σ*	*Spec. μ±σ*	*Prec. μ±σ*	*Accuracy*
10	0.151±0.083	0.995±0.001	0.671±0.039	0.978±0.002
20	0.163±0.040	0.988±0.001	0.596±0.018	0.996±0.001
30	0.198±0.015	0.994±0.000	0.551±0.019	0.975±0.000
40	0.187 ±0.055	0.995±0.001	0.565±0.044	0.975±0.002

Table 5: Average results for the prediction of the end of β-strands obtained for different number of iterations. Standard deviation is reported between brackets.

It.	*Recall μ±σ*	*Spec. μ±σ*	*Prec. μ±σ*	*Accuracy*
10	0.489±0:101	0.990±0.001	0.615±0.014	0.984±0.003
20	0.524±0.040	0.991±0.001	0.663±0.008	0.988±0.001
30	0.516±0.014	0.994±0.000	0.760±0.019	0.985±0.000
40	0.586 ±0.059	0.993±0.001	0.728±0.019	0.987±0.002

Figure 12 shows an example of rule discovered by our algorithm. In particular this rule is relative to the beginning of an α-helix. If we inspect this rule, we can see that the hydrophobicity value for the amino acid in the N-cap position is between 0.52 and 0.92, the polarity value lies between 1.0 and 0.93 and neutral charge (0.0). Therefore, this amino acid could be L (Lysine) or F (Phenylalanine), which fulfills these features according to the cited scales. As it can be noticed, the rules that compose the prediction model provided by our algorithm are easily interpretable. We believe that this is an important factor, since this would facilitate the interpretation of results by an expert in the field.

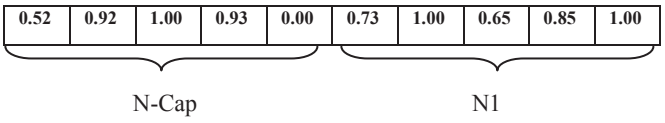

Figure 12: Example of a resulting rule for a beginning of an α-helix.

In conclusion, we can say that the proposed algorithm obtained satisfactory results. Moreover, the algorithm has been tested using a high number of sequences (5,000 helix and 5,000 strand sequences). We believe that this represent an important factor. In fact, the number of protein sequences available increase by the day, and thus, having a method that is scalable would be very important.

5. CONCLUSIONS AND FUTURE WORK

In this paper, we have proposed an evolutionary algorithm for α-helix and β-strand capping prediction from sequences of amino acid. The prediction is based on three amino acids properties, i.e., hydrophobicity, polarity and net charge. Moreover, some particular characteristic of these motifs are considered in order to improve the search process performed by the algorithm.

We have performed a statistical analysis aimed at discovering the amino acid propensities in capping positions in 163,461 α-helices and 216,390 β-sheets extracted from PDB using the DSSP program. We have computed the probability, for each pair of possible amino acids, to appear in both N-cap and N1 positions and C1, C-cap positions. This study provided us with useful information for the prediction of secondary structure. In fact, this information could be used for modifying the fitness function, improving in this way the evolutionary search. A study of each single amino acid has been also developed in each position. From this study, we could individuate which amino acid is more probable to appear in one of the positions taken into consideration.

In order to test the validity of the proposed algorithm, we performed a set of experiments using 5,000 α-helix and 5,000 β-strand sequences. These sequences were extracted from a protein data set from Protein Data Bank. In particular, we considered 12; 830 non-redundant and non-homologous protein with a homology rate lower than 30%. To the best of our knowledge, no other approaches have used such a high number of sequences in α-helix capping regions prediction. Results obtained on the prediction of α-helices are very encouraging and in particular, the accuracy characterizing the prediction models obtained is very high independently from the number of rules generated. As far as the experiments on the prediction of β-sheets, we have not found other results in the literature to contrast our results. However, also in this case, the accuracy obtained is satisfactory, even if the results are slightly worse than those obtained for the α-helices.

Future works will be focused on the analysis of different properties to be included in the fitness function, with the aim of increasing the quality of the prediction model. For example we are planning to incorporate the residue size, which has a significant relevance according to our statistical study. We will also expand the number of residues in the window of amino acids.

Furthermore, we are studying the possibility of incorporating a local search phase in the algorithm that will help to improve individuals. We also intend to extend our experimentation to other dataset of protein sequences.

6. REFERENCES

[1] Complete list of pdb protein identifiers used in this article. http://www.upo.es/eps/marquez/proteins.txt.

[2] Protein data bank advanced search http://www.pdb.org/pdb/search/advSearch.do.

[3] Protein data bank web. http://www.wwpdb.org.

[4] J. Cheng and P. Baldi. Improved residue contact prediction using support vector machines and a large feature set. Bioinformatics, 8, 113, 2007.

[5] P. Chou and G. Fasman. Prediction of protein conformation. Biochemistry, 13(2), 222–245, 1974.

[6] Y. Cui, R. Chen, and W. Hung. Protein folding simulation with genetic algorithm and supersecondary structure constraints. Proteins: Structure, Function and Genetics, 31, 247–257, 1998.

[7] Doig and B. R.L. N- and c-capping preferences for all 20 amino acids in alpha-helical peptides. Protein Science, 4(7), 1325–1336, 1995.

[8] P. Fariselli and R. Casadio. A neural network based predictor of residue contacts in proteins. Protein Engineering, 12, 15–21, 1999.

[9] F. Farzadfard, N. Gharaei, H. Pezeshk, and S. Marashi. Beta-sheet capping: Signals that initiate and terminate beta-sheet formation. J. Structural Biology, 161, 101–110, 2008.

[10] N. Fonseca, R. Camacho, and A. Magalhaes. Amino acid pairing at the n- and c-termini of helical segments in proteins. Proteins, 70, 188–196, 2007.

[11] D. Frishman and P. Argos. Incorporation of non-local interactions in protein secondary structure prediction from the amino acid sequence. Protein Engineering, 9, 133–142, 1996.

[12] J. Garnier, D. Osguthorpe, and B. Robson. Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. J. Mol. Biol., 120, 97–120, 1978.

[13] R. Grantham. Amino acid difference formula to help explain protein evolution. J. J. Mol. Bio., 185, 862–864, 1974.

[14] K. J. and R. Doolittle. A simple method for displaying the hydropathic character of a protein. J. J. Mol. Bio., 157, 105–132, 1982.

[15] W. Kabsch and C. Sander. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers, 22(12), 2577–2637, 1983.

[16] P. Klein, M. Kanehisa, and C. DeLisi. Prediction of protein function from sequence properties: Discriminant analysis of a data base. Biochim. Biophys., 787, 221–226, 1984.

[17] V. Lim. Algorithms for prediction of a-helical and b-structural regions in globular proteins. J. Mol. Biol., 88, 857–872, 1974.

[18] L. McGullun, K. Bryson, and D. Jones. The psipred protein structure prediction server. Bioinformatics, 16, 404–405, 2000.

[19] N. Qian and T. Sejnowski. Predicting the secondary structure of globular proteins using neural network models. J. Mol. Biol., 202, 865–884, 1988.

[20] J. Richardson and D. Richardson. Amino acid preferences for specific locations at the ends of alpha helices. Science, 240, 1648–1652, 1998.

[21] Salamov and V. Solovyev. Protein secondary structure prediction using local alignments. J. Mol. Biol., 268, 31–36, 1997.

[22] R. Unger and J. Moult. Genetic algorithms for protein folding simulations. Biochim. Biophys., 231, 75–81, 1993.

[23] J Ward, L. McGullun, B. Buxton, and D. Jone. Secondary structure prediction with support vector machines. Bioinformatics, 13, 1650–1655, 2003.

[24] C. Wilson, P. Boardman, A. Doig, and S. Hubbard. Improved prediction for n-termini of alpha-helices using empirical information. Proteins, 57(2), 322–330, 2004.

[25] C. Wilson, S. Hubbard, and A. Doig. A critical assessment of the secondary structure prediction of alpha-helices and their n-termini in proteins. Protein Eng., 15, 545–554, 2002.