Predicting the secondary structure of proteins using Machine Learning algorithms

Rui Camacho*

LIAAD & DEI-Faculdade de Engenharia da Universidade do Porto,
Rua Dr Roberto Frias s/n, 4420-465 Porto, Portugal
E-mail: rcamacho@fe.up.pt
*Corresponding author

Rita Ferreira, Natacha Rosa and Vânia Guimarães

Faculdade de Engenharia da Universidade do Porto,
Rua Dr Roberto Frias s/n, 4420-465 Porto, Portugal
E-mail: bio06027@fe.up.pt
E-mail: bio06004@fe.up.pt
E-mail: bio06018@fe.up.pt

Nuno A. Fonseca and Vítor Santos Costa

CRACS-INESC Porto L.A./FCUP,
R. Campo Alegre 1021/1055, 4169-007 Porto, Portugal
E-mail: nunofonseca@acm.org
E-mail: vsc@dcc.fc.up.pt

Miguel de Sousa and Alexandre Magalhães

REQUIMTE/Universidade do Porto,
R. Campo Alegre 687, 4169-007 Porto, Portugal
E-mail: miguel@fc.up.pt
E-mail: almagalh@fc.up.pt

Abstract: The functions of proteins in living organisms are related to their 3-D structure, which is known to be ultimately determined by their linear sequence of amino acids that together form these macromolecules. It is, therefore, of great importance to be able to understand and predict how the protein 3D-structure arises from a particular linear sequence of amino acids. In this paper we report the application of Machine Learning methods to predict, with high values of accuracy, the secondary structure of proteins, namely α-helices and β-sheets, which are intermediate levels of the local structure.

Keywords: data mining; machine learning; classification; decision trees; rule induction; instance-based learning; Bayesian algorithms; WEKA; bioinformatics; protein folding; predicting secondary structure conformations.

Reference to this paper should be made as follows: Camacho, R., Ferreira, R., Rosa, N., Guimarães, V., Fonseca, N.A., Costa, V.S., de Sousa, M. and Magalhães, A. (2012) 'Predicting the secondary structure of proteins using Machine Learning algorithms', *Int. J. Data Mining and Bioinformatics*, Vol. 6, No. 6, pp.571–584.

Biographical notes: Rui Camacho got his first degree in 1984 in Electrical Engineering and Computers from University of Porto. He got his M.Sc. degree in Electrical Engineering and Computers from Instituto Superior Technical University of Lisbon, in 1989. He got is PhD from University of Porto in 2000. He is currently Associate Professor at Faculty of Engineering University of Porto and a researcher at Laboratory of Artificial Intelligence and Decision Support (LIAAD). His research interests encompass data mining, bio-informatics, machine learning, and distributed computing.

Ana Rita Ferreira is currently a M.Sc student at Faculdade de Engenharia da Universidade do Porto, Portugal in the course of Master in Bioengineering, Biomedical Engineering. Her main interests are medical instrumentation and rehabilitation systems, medical image analysis, 3D biomodeling and biomaterials.

Natacha Rosa is currently a M.Sc student at Faculdade de Engenharia da Universidade do Porto, Portugal in the course of Master in Bioengineering. Her main interests are: biomedical instrumentation, Rehabilitation engineering, nanotechnology in drug delivery and biomimetics.

Vânia Guimarães is currently a M.Sc student at Faculdade de Engenharia da Universidade do Porto, Portugal in the course of Master in Bioengineering, Biomedical Engineering. Her main interests are medical instrumentation and rehabilitation systems, medical image analysis, computer aided diagnosis and Informatics.

Nuno A. Fonseca received in 1996 his first degree in Computer Science from the Faculty of Science of the University of Porto, later, in 2001, he obtained the M.Sc. degree in Artificial Intelligence and Computation from the Faculty of Engineering of the University of Porto, and the Ph.D. degree in Computer Science from the Faculty of Science of the University of Porto in 2006. Currently, he is a research fellow at the Center for Research on Advanced Computing Systems (CRACS) and INESC Porto LA. His research interests encompass bio-informatics, machine learning, and high performance computing.

Vítor Santos Costa is an associate professor at Faculdade de Ciencias, Universidade do Porto. He received a bachelor's degree from the University of Porto in 1984 and was granted a PhD in Computer Science from the University of Bristol in 1993. He is Visiting Professor at the University of Wisconsin-Madison. His research interests include logic programming and machine learning, namely inductive logic programming and statistical relational learning. He has published more than 100 refereed papers in Journals and International Conferences, is then main developer of YAP Prolog, has chaired two conferences, and has supervised 5 PhD students.

Miguel M. de Sousa graduated in 2002 with a degree in Biochemistry/ Biophysics. Since then has worked in the research of metallo-surfactant

properties of Iron(II) complexes and in the field of Photochemistry investigating lanthanide-based complexes and their application as molecular logic gates. He is currently working on his PhD in the field of Bioinformatics studying amino acid patterns and pattern side-chain topology in protein secondary structures, particularly alpha-helices and beta-sheets, to be applied to the development of protein structure prediction algorithms.

Alexandre L. Magalhães received in 1997 his PhD in Chemistry from the University of Porto. He is an Assistant Professor at the Faculty of Sciences of the University of Porto where he gives several courses in the area of Computational Chemistry. His scientific interests include Protein structure and supramolecular chemistry.

1 Introduction

Proteins are complex structures synthesised by living organisms. They are a fundamental type of biomolecules that perform a large number of functions in cell biology. Proteins can assume catalytic roles and accelerate or inhibit chemical reactions in our body. They can assume roles of transportation of smaller molecules, storage, movement, mechanical support, immunity and control of cell growth and differentiation (Alberts et al., 2002). All of these functions rely on the 3D-structure of the protein. The process of going from a linear sequence of amino acids, that together compose a protein, to the protein's 3D shape is named *protein folding*. Anfinsen's work (Sela et al., 1957) has proven that primary structure determines the way protein folds. Protein folding is so important that whenever it does not occur correctly it may produce diseases such as Alzheimer's, Bovine Spongiform Encephalopathy (BSE), usually known as *mad cows disease*, Creutzfeldt-Jakob (CJD) disease, a Amyotrophic Lateral Sclerosis (ALS), Huntingtons syndrome, Parkinson disease, and other diseases related to cancer.

A major challenge in Molecular Biology is to unveil the process of protein folding. Several projects have been set up with that purpose. Although protein function is ultimately determined by their 3D structure there have been identified a set of other intermediate structures that can help in the formation of the 3D structure. We refer the reader to Section 2 for a more detailed description of protein structure. To understand the high complexity of protein folding it is usual to follow a sequence of steps. One starts by identifying the sequence of amino acids (or residues) that compose the protein, the so-called *primary structure*; then we identify the *secondary structures conformations*, mainly α-helices and β-sheet; and then we predict the *tertiary structure* or 3D shape.

In this paper we address the step of predicting α-helices and β-strands based on the sequence of amino acids that compose a protein. More specifically, in this study models based on Machine Learning algorithms were built to predict the start, inner points and end of secondary structures. A total of 1499 protein sequences were selected from the PDB and data sets were appropriately assembled to be used by Machine Learning algorithms and thus construct the models. In this context rule induction algorithms, decision trees, functional trees, Bayesian methods, and ensemble methods

were applied. The models achieved an accuracy between 84.9% (in the prediction of α-helices) and 99.6% (in the prediction of the inner points of β-strands). The results show also that small and intelligible models can be constructed.

The rest of the paper is organised as follows. Section 2 gives basic definitions on proteins required to understand the reported work. Related work is reported in Section 3. Our experiments, together with the results obtained, are presented in Section 4. Conclusions are presented in Section 5.

2 Proteins

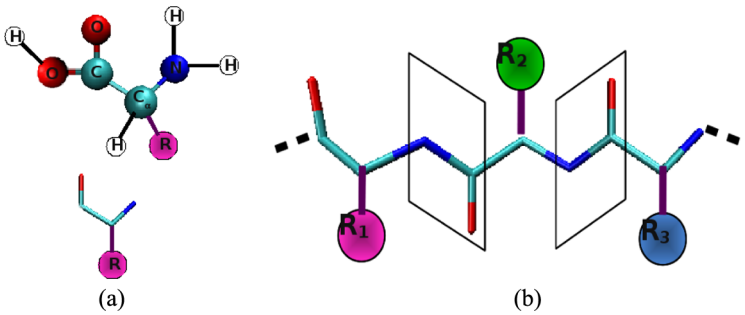
Proteins are build up of amino acids, connected by peptide bonds between the carboxyl and amino groups of adjacent amino acid residues as shown in Figure 1(b) (Petsko and Petsko, 2007). All amino acids have common structural characteristics that include an α carbon to which are connected an amino group and a carboxyl group, an hydrogen and a variable side chain as shown in Figure 1(a). It is the nature of side chain that determines the identity of a specific amino acid. There are 20 different amino acids that integrate proteins in cells. Once the amino acids are connected in the protein chain they are designated as residues.

In order to function in an organism a protein has to assume a certain 3D conformation. To achieve those conformations apart from the peptide bonds there have to be extra types of weaker bonds between residues. These extra bonds are responsible for the secondary and tertiary structure of a protein (Gsponer et al., 2003).

One can identify four types of structures in a protein. The primary structure of a protein corresponds to the linear sequence of residues. The secondary structure is composed by subsets of residues arranged mainly as α-helices and β-sheets, as seen in Figure 2. The tertiary structure results for the folding of α-helices or β-sheets. The quaternary structure results from the interaction of two or more polypeptide chains.

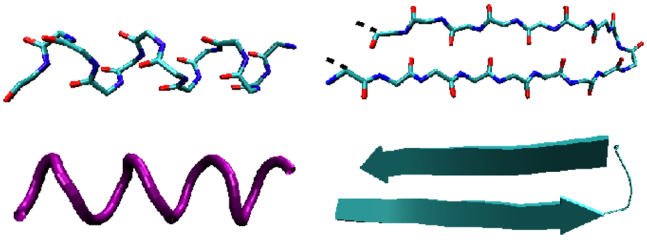
Secondary structure's conformations, α-helices and β-sheets, were discovered in 1951 by Linus Carl Pauling. These secondary structure's conformations are obtained due to the flexibility of the peptide chain that can rotate over three different chemical bonds. Most of the existing proteins have approximately 70% of their structure as helices that is the most common type of secondary structure.

Figure 1 (a) General structure of an amino acid; side chain is represented by the letter R and (b) a fraction of a proteic chain, showing the peptide bounds (see online version for colours)



(a) (b)

Figure 2 Secondary structure conformations of a protein: α-helices (left); β-sheet (right) (see online version for colours)



3 Related work

Arguably, protein structure prediction is a fundamental problem in Bioinformatics. Early work by Chou and Fasman (1978), based on single residue statistics, looked for contiguous regions of residues that have an high probability of belonging to a secondary structure. The protein sample used was very small, resulting in overestimating the accuracy of the reported study.

Qian and Sejnowski (1988) used neural networks to predict secondary structures but achieved an accuracy of only 64.3%. They used a window technique (of size 13) where the secondary structure of the central residues was predicted on the base of its 12 neighbours.

Neural Networks were also used in the work by Rost and Sander (1993). They used a database of 130 representative protein chains of known structure and achieved an accuracy of 69.7%. Later, Rost and Sanderwith used the PHD (Rost, 1996) method on the RS126 data set and achieved an accuracy of 73.5%. JPRED (Cuff et al., 1998) exploited multiple sequence alignments to obtain an accuracy of 72.9%. NNSSP (Salamov and Solovyev, 1995) is a scored nearest neighbour method by considering the position of N and C terminal in α-helices and β-strands. Its prediction accuracy on the RS126 data set was 72.7%. PREDATOR (Frishman and Argos, 1997) used propensity values for seven secondary structures and local sequence alignment. The prediction accuracy of this method for RS126 data set achieved 70.3%. PSIPRED (Jones, 1999) used a position-specific scoring matrix generated by PSI-BLAST to predict protein secondary structure and achieved 78.3.

DSC (King and Sternberg, 1996) achieved 71.1% prediction accuracy in the RS126 data set by exploring amino acid profiles, conservation weights, indels, and hydrophobicity.

Using a Inductive Logic Programming (ILP) another series of studies improved the secondary structure prediction score. In 1990 Muggleton et al. (Muggleton, 1992) used only 16 proteins (in contrast with 1499 used in our study) and the GOLEM (Muggleton and Feng, 1990) ILP system to predict if a given residue in a given position belongs or not to an α-helix. They achieved an accuracy of 81%. Previous results had been reported by Kneller and Langridge (1990) using Neural Networks, achieving only 75% accuracy. The propositional learner PROMIS (King and Sternberg, 1990; Sternberg et al., 1992) achieved 73% accuracy on the GOLEM data set.

It has been shown that the helical occurrence of the 20 type of residues is highly dependent on the location, with a clear distinction between N-terminal, C-terminal and interior positions (Richardson and Richardson, 1988). The computation of amino

acid propensities may be a valuable information both for pre-processing the data and for assessing the quality of the constructed models (Fonseca et al., 2008). According to Blader et al. (1993) an important influencing factor in the propensity to form α-helices is the hydrophobicity of the side-chain. Hydrophobic surfaces turn into the inside of the chain giving a strong contribution to the formation of α-helices. It is also known that the protein surrounding environment has influence in the formation of α-helices. Modelling the influence of the environment in the formation of α-helices, although important, is very complex from a data analysis point of view (Krittanai and Johnson, 2000).

4 Experiments

4.1 Experimental settings

To construct models to predict the remarkable points of secondary structures we have proceeded as follows. We first downloaded a list of proteins with low structure identity from the Dunbrak's website (Wang and Dunbrack, 2003).[1] The list contains 1499 proteins with structure identity less than 20%. We then downloaded the PDB[2] for each of the protein in the list. Each PDB was processed in order to extract secondary structure information and the linear sequence of residues of the protein. We have used a data set much larger than the standard RS126 dataset of Rost et al. (Rost, 1996). We have also used a data set of proteins with structure identity (20%) lesser than the one used in RS126 (25%).

In our data sets an example is a sequence of a fixed number of residues (window) before and after the remarkable points[3] of secondary structures. We have produced 24 data sets using 4 different window sizes (2, 3, 4 and 5), 3 types of remarkable points (start, inner and end points) and 2 types of structures (α-helices and β-sheets). The size of the data sets, for the different window sizes, is shown in Table 1. To obtain the example sequences to use we have selected sequences that are:

1 at the start of a α-helix

2 at the end of a α-helix

3 in the interior of a α-helix

4 at the start of a β-strand

5 at the end of a β-strand

6 in the interior of a β-strand.

To do so, we identify the 'special' point where the secondary structures start or end, and then add W residues before and after that point. Therefore the sequences are of size $2 \times W + 1$, where $W \in [2, 3, 4, 5]$. In the interior of a secondary structure we just pick sequences of $2 \times W + 1$ residues that do not overlap. With these sequences we envisage to study the start, interior and end points of secondary structures.

Table 1 Characterisation of the data sets according to the number of examples (**E**) and number of attributes (**A**). The number of examples and attributes depends only on the window size (W)

Window size (W)							
$W = 2$		$W = 3$		$W = 4$		$W = 5$	
E	*A*	*E*	*A*	*E*	*A*	*E*	*A*
62,050	270	49,242	451	40,528	632	34,336	813

The attributes used to characterise the examples are of three main types: whole structure attributes; window-based attributes; and attributes based on differences between the 'before' and 'after windows'.

The whole structure attributes include: the size of the structure; the percentage of hydrophobic residues in the structure; the percentage of polar residues in the structure; the average value of the hydrophobic degree; the average value of the hydrophilic degree; the average volume of the residues; the average area of the residues in the structure; the average mass of the residues in the structure; the average isoelectric point of the residues; and, the average topological polar surface area.

For the window-based attributes we have computed a set of attributes based on the properties of residues shown in Table 2. For each amino acid of the window and amino acid property we computed the following attributes: the value of the property of each residue in the window; either if the property 'increases' or decreases the value along the window; the number of residues in the window with a specified value and; whether a residue at each position of the window belongs to a pre-computed set of values.

Table 2 List of amino acid properties used in the study

Polarity	Hydrophobicity	Size	Isoelectric
Charge	h-bonddonor	xlogp3	Side chain polarity
Acidity	Rotatable bond count	h-bondacceptor	Side chain charge

For the attributes capturing the differences we have computed the average values of numerical properties of amino acids in each window and then define the a new attribute as the difference in the values of those two averages. For non numerical properties we first performed some countings in the windows before and after the critical point. Based on those countings we have defined new attribute as the difference between those countings. For example we have counted the number of hydrofobic amino acids in the before and after window and the define a new attribute as the difference between those countings.

Altogether there are between 253 (window size of 2) to 745 (window size of 5) attributes. In each of the 6 types of data sets, the sequences with one chosen remarkable point is taken as belonging to one class and all other sequences are assumed to belong to the other class. For example when predicting the end-point of β-strands the sequences with a β-strands end-point are in one class and all other sequences (start-point of both helices and strands, inner points of both helices and strands and the end-points of

helices) are in the other class (thus transforming the problem to a binary classification problem).

Table 3 Performance measures (b) used in the experiments computed after the confusion matrix (a)

	actual	
p	n	
TP	FP	p **predicted**
FN	TN	n

(a)

Accuracy	**True Positive Rate**	**True Negative Rate**
$\frac{TP+TN}{TP+TN+FP+FN} \times 100\%$	$\frac{TP}{(TP+FN} \times 100\%$	$\frac{TN}{TN+FP} \times 100\%$

(b)

The quality of the constructed models was estimated using measures computed after the Confusion Matrix[4] (Table 3(a)). From the Confusion Matrix we compute the Accuracy measure, the True Positive Rate (TPR) and the True Negative Rate (TNR) of the model (Table 3(b)). The Accuracy captures the global performance of the model whereas the TPR and the TNR provide information on the performance of predicting the individual classes.

The experiments were done in a machine with 2 quad-core Xeon 2.4 GHz and 32 GB of RAM, running Ubuntu 8.10. We used machine learning algorithms from the Weka 3.6.0 toolkit (Witten and Frank, 2005) and a 10-fold cross validation procedure to estimate the quality of constructed models. We have used rule induction algorithms (Ridor), decision trees (J48 Quinlan, 1993 and ADTree Freund and Mason, 1999), functional trees (FT Gama, 2004; Landwehr et al., 2005), instance-based learning (IBk Aha and Kibler, 1991), bayesian algorithms (NaiveBayes and BayesNet John and Langley, 1995) and one ensemble method (RandomForest Breiman, 2001).[5]

4.2 Experimental results

The results obtained with the Machine Learning algorithms are shown in Tables 4–9.

The results presented show high values of accuracy with a minimum of 84.9%, in the prediction of the starting point of helices using Functional Trees, and a maximum of 99.6%, in the prediction of inner positions of β-strands also using Functional Trees. In each table the best accuracy value was quite above the base line value. The base line value was taken as the ZeroR prediction that is actually the majority class prediction. Good values of TPR were also achieved with a maximum of 72.6% in the prediction of β-strands end point. Functional Tree algorithm produce the best results for α-helices whereas in the prediction of β-strands Bayesian Networks achieves the best TPRs, while Random Forest and IBk the best accuracy values. Overall Functional Trees have a very good performance both in Accuracy and TPR in almost all prediction problems. The Bayesian Networks performed quite well in terms of TPR in the β-sheet prediction problems.

Looking at Tables 4–6 we see that the best TPR is obtained by functional Trees using a window size of five. That is a reasonable result since α-helices have 3.6 amino

Table 4 Results of predicting α-helices starting point. The accuracy results (%) are shown in (a). The improvement over the majority class prediction is shown in (b). The true positive rate (%) results are shown in (c). The true negative rate (%) results are shown in (d). Results were obtained for windows of size 2, 3, 4 and 5 residues. RF stands for RandomForest and FT for Functional Tree

	Window size			
Algorithm	**2**	**3**	**4**	**5**
Ridor	83.4	80.6	79.2	78.0
J48	84.0	81.6	79.6	77.7
RF	84.2	80.1	77.7	74.1
FT	**84.9**	82.8	81.7	79.7
ADTree	83.2	80.4	78.3	76.3
IBk	81.9	76.6	72.5	69.2
NaiveBayes	70.4	65.8	63.9	63.4
BayesNet	69.7	65.9	64.6	64.2
ZeroR	82.5	76.9	72.4	67.7

(a)

	Window size			
Algorithm	**2**	**3**	**4**	**5**
Ridor	+0.9	+3.7	+6.8	+10,3
J48	+1.5	+4.7	+7.2	+10
RF	+1.7	+3.2	+5.3	+6.4
FT	+2.9	+5.9	+9.3	**+12.0**
ADTree	+0.7	+3.5	+5.9	+8.6
IBk	-0,6	-0.3	+0.1	+1.5
NaiveBayes	-12.1	-11.1	-8.5	-4.3
BayesNet	-12.8	-11.0	-7.8	-3.5

(b)

	Window size			
Algorithm	**2**	**3**	**4**	**5**
Ridor	16.5	24.4	36.3	45.4
J48	28.0	34.7	44.1	49.3
RF	31.0	38.5	45.3	49.5
FT	36.2	45.8	53.5	**60.5**
ADTree	22.4	33.0	41.6	48.8
IBk	14.6	22.6	29.2	37.0
NaiveBayes	50.3	53.3	57.5	59.1
BayesNet	54.8	55.5	58.9	60.0

(c)

	Window size			
Algorithm	**2**	**3**	**4**	**5**
Ridor	**98.6**	97.4	95.6	93.6
J48	96.8	95.6	93.2	91.2
RF	96.2	92.6	90.1	85.9
FT	96.0	93.9	92.5	88.9
ADTree	97.0	94.6	92.3	89.4
IBk	97.2	92.8	89.0	84.5
NaiveBayes	75.0	69.5	66.4	65.5
BayesNet	73.1	69.0	66.8	66.2

(d)

Table 5 Results of predicting α-helices inner points. The accuracy results (%) are shown in (a). The improvement over the majority class prediction is shown in (b). The true positive rate (%) results are shown in (c). The true negative rate (%) results are shown in (d). Results were obtained for windows of size 2, 3, 4 and 5 residues. RF stands for RandomForest and FT for Functional Tree

	Window size			
Algorithm	**2**	**3**	**4**	**5**
Ridor	84.6	87.5	90.6	92.2
J48	85.7	88.0	90.5	92.3
RF	84.9	86.5	88.6	90.2
FT	86.7	89.5	92.4	**93.8**
ADTree	85.2	88.2	91.3	93.1
IBk	77.8	83.0	85.9	88.5
NaiveBayes	73.2	73.5	74.4	80.4
BayesNet	76.3	77.3	77.7	80.8
ZeroR	75.9	81.2	84.5	87.3

(a)

	Window size			
Algorithm	**2**	**3**	**4**	**5**
Ridor	+8.7	+6.3	+6.1	+4.9
J48	+9.8	+6.8	+6.0	+5.0
RF	+9.0	+5.3	+4.1	+2.9
FT	**+10.8**	+8.3	+7.9	+6.5
ADTree	+9.3	+7.0	+6.8	+5.8
IBk	+1.9	+1.8	+1.4	+1.2
NaiveBayes	-2.7	-7-7	-10.1	-6.9
BayesNet	+0.4	-3.9	-6.9	-6.5

(b)

	Window size			
Algorithm	**2**	**3**	**4**	**5**
Ridor	57.3	52.0	56.0	65.6
J48	65.8	60.0	60.8	64.3
RF	54.6	39.3	31.8	26.5
FT	69.0	67.3	72.5	**73.9**
ADTree	65.2	63.6	66.9	68.7
IBk	23.1	18.6	14.4	18.9
NaiveBayes	63.1	64.2	63.2	66.7
BayesNet	68.1	66.9	66.0	69.9

(c)

	Window size			
Algorithm	**2**	**3**	**4**	**5**
Ridor	93.2	95.8	96.9	96.1
J48	91.9	94.5	96.0	96.3
RF	94.6	97.4	99.0	**99.5**
FT	92.3	94.7	96.0	96.7
ADTree	91.6	93.9	95.7	96.7
IBk	95.2	98.0	99.0	98.7
NaiveBayes	76.5	75.7	76.4	82.3
BayesNet	78.9	79.7	79.8	82.4

(d)

Table 6 Results of predicting α-helices end point. The accuracy results (%) are shown in (a). The improvement over the majority class prediction is shown in (b). The true positive rate (%) results are shown in (c). The true negative rate (%) results are shown in (d). Results were obtained for windows of size 2, 3, 4 and 5 residues. RF stands for RandomForest and FT for Functional Tree

	Window size			
Algorithm	**2**	**3**	**4**	**5**
Ridor	83.4	80.1	79.3	78.2
J48	83.9	81.1	79.5	77.6
RF	83.3	78.6	76.9	72.0
FT	**85.0**	83.2	81.6	80.4
ADTree	82.7	80.4	77.6	75.9
IBk	82.2	76.5	72.6	68.8
NaiveBayes	69.6	66.8	65.3	63.8
BayesNet	68.6	67.1	66.1	64.8
ZeroR	81.8	77.4	72.8	68.2

(a)

	Window size			
Algorithm	**2**	**3**	**4**	**5**
Ridor	+1.6	+2.7	+6.6	+10.0
J48	+2.1	+3.7	+6.7	+9.4
RF	+1.3	+1.2	+4.1	+3.8
FT	+3.2	+5.8	+8.8	**+12.2**
ADTree	+0.9	+3.0	+4.8	+7.7
IBk	+0.4	-0.9	-0.2	+0.6
NaiveBayes	-12.2	-10.6	-7.5	-4.4
BayesNet	-13.2	-10.3	-6.7	-3.4

(b)

	Window size			
Algorithm	**2**	**3**	**4**	**5**
Ridor	14.0	31.0	55.3	55.3
J48	22.7	32.1	45.5	54.4
RF	32.4	33.2	43.8	43.5
FT	36.3	48.3	57.7	**64.1**
ADTree	15.6	29.1	41.8	58.9
IBk	6.7	22.9	28.0	34.3
NaiveBayes	48.3	54.8	58.4	58.9
BayesNet	55.3	58.9	61.3	61.0

(c)

	Window size			
Algorithm	**2**	**3**	**4**	**5**
Ridor	98.9	94.5	88.3	88.8
J48	97.5	95.5	92.1	88.4
RF	94.5	91.9	89.3	85.2
FT	95.8	93.4	90.6	88.0
ADTree	97.6	95.3	90.9	84.7
IBk	**98.9**	92.1	89.3	84.9
NaiveBayes	74.3	70.3	67.8	66.1
BayesNet	71.6	69.5	67.8	66.5

(d)

Table 7 Results of predicting β-strand start point. The accuracy results (%) are shown in (a). The improvement over the majority class prediction is shown in (b). The true positive rate (%) results are shown in (c). The true negative rate (%) results are shown in (d). Results were obtained for windows of size 2, 3, 4 and 5 residues. RF stands for RandomForest and FT for Functional Tree

	Window size			
Algorithm	**2**	**3**	**4**	**5**
Ridor	89.2	90.0	91.3	92.6
J48	89.5	90.3	91.4	92.5
RF	92.8	93.4	93.7	**94.9**
FT	90.6	91.7	92.7	93.8
ADTree	88.9	88.4	90.3	91.6
IBk	86.4	85.7	87.4	89.4
NaiveBayes	72.2	72.6	71.8	72.7
BayesNet	73.6	73.6	72.9	73.0
ZeroR	84.1	84.2	86.1	88.8

(a)

	Window size			
Algorithm	**2**	**3**	**4**	**5**
Ridor	+5.1	+5.8	+5.2	+3.8
J48	+5.4	+6.1	+5.3	+3.7
RF	+8,7	**+9.2**	+7.6	+6.1
FT	+6.5	+7.5	+6.6	+5.0
ADTree	+4.8	+4.2	+4.2	+2.8
IBk	+2.3	+1.5	+1.3	+0.6
NaiveBayes	-11.9	-11.6	-14.3	-16.1
BayesNet	-10.5	-10.6	-13.2	-15.8

(b)

	Window size			
Algorithm	**2**	**3**	**4**	**5**
Ridor	41.9	46.4	47.6	46.3
J48	50.2	60.7	56.8	51.1
RF	64.0	69.5	61.5	60.3
FT	58.0	64.7	66.2	71.1
ADTree	41.4	46.1	50.3	47.3
IBk	45.6	49.3	47.9	44.8
NaiveBayes	66.0	70.6	70.1	70.3
BayesNet	70.0	**72.4**	71.7	71.7

(c)

	Window size			
Algorithm	**2**	**3**	**4**	**5**
Ridor	98.1	98.1	98.3	98.4
J48	96.9	95.9	97.0	97.8
RF	98.2	97.8	98.9	**99.3**
FT	96.8	96.7	97.0	96.7
ADTree	96.9	96.3	96.8	97.3
IBk	94.1	92.5	93.8	95.1
NaiveBayes	73.3	73.0	72.1	73.0
BayesNet	74.2	73.8	73.1	73.1

(d)

Table 8 Results of predicting β-strand inner points. The accuracy results (%) are shown in (a). The improvement over the majority class prediction is shown in (b). The true positive rate (%) results are shown in (c). The true negative rate (%) results are shown in (d). Results were obtained for windows of size 2, 3, 4 and 5 residues. RF stands for RandomForest and FT for Functional Tree

Algorithm	Window size			
	2	**3**	**4**	**5**
Ridor	92.9	96.1	98.0	99.1
J48	93.0	96.1	98.0	99.1
RF	85.5	86.5	88.6	90.2
FT	93.1	96.3	98.3	99.3
ADTree	92.8	96.0	98.0	99.1
IBk	94.8	97.5	99.0	**99.6**
NaiveBayes	72.2	75.2	77.0	85.0
BayesNet	75.8	76.1	77.8	78.5
ZeroR	92.4	95.9	90.0	99.1

(a)

Algorithm	Window size			
	2	**3**	**4**	**5**
Ridor	+0.5	+0.1	+8.0	0.0
J48	+0.6	+0.1	+8.0	0.0
RF	-6.9	-9.4	-1.4	-8.9
FT	+0.7	+0.4	+8.3	+0.2
ADTree	+0.4	+0.1	+8.0	0.0
IBk	+2.4	+1.6	**+9.0**	+0.5
NaiveBayes	-20.2	-20.7	-13.0	-14.1
BayesNet	-16.6	-19.8	-12.2	-20.6

(b)

Algorithm	Window size			
	2	**3**	**4**	**5**
Ridor	11.0	8.1	0.7	0.0
J48	28.2	13.8	0.0	0.0
RF	53.1	39.3	31.8	26.5
FT	47.2	47.6	47.9	56.7
ADTree	13.6	6.7	0.6	1.5
IBk	35.2	38.8	48.6	58.0
NaiveBayes	57.4	59.9	54.7	46.1
BayesNet	63.7	**64.4**	61.0	55.8

(c)

Algorithm	Window size			
	2	**3**	**4**	**5**
Ridor	99.6	99.8	**100.0**	**100.0**
J48	98.3	99.6	**100.0**	**100.0**
RF	95.8	97.4	99.0	99.5
FT	96.9	98.4	99.3	99.7
ADTree	99.3	99.8	**100.0**	**100.0**
IBk	99.7	**100.0**	**100.0**	**100.0**
NaiveBayes	73.5	75.8	77.5	85.4
BayesNet	76.8	76.6	78.1	78.7

(d)

Table 9 Results of predicting β-strand end point. The accuracy results (%) are shown in (a). The improvement over the majority class prediction is shown in (b). The true positive rate (%) results are shown in (c). The true negative rate (%) results are shown in (d). Results were obtained for windows of size 2, 3, 4 and 5 residues. RF stands for RandomForest and FT for Functional Tree

Algorithm	Window size			
	2	**3**	**4**	**5**
Ridor	89.3	89.7	91.0	92.4
J48	89.2	89.8	91.0	92.4
RF	92.7	93.1	93.9	**94.7**
FT	89.4	90.4	92.0	93.5
ADTree	89.0	90.1	89.6	92.5
IBk	88.1	88.6	90.3	92.3
NaiveBayes	73.1	72.3	72.2	72.8
BayesNet	73.9	73.0	73.3	73.0
ZeroR	84.3	84.3	86.2	88.9

(a)

Algorithm	Window size			
	2	**3**	**4**	**5**
Ridor	+5.0	+5.4	+4.8	+3.5
J48	+4.9	+5.5	+4.8	+3.5
RF	+8.4	**+8.8**	+7.7	+5.8
FT	+5.1	+6.1	+5.8	+4.6
ADTree	+4.7	+5.8	+3.4	+3.6
IBk	+3.8	+4.3	+4.1	+3.4
NaiveBayes	-11.2	-12.0	-14.0	-16.1
BayesNet	-10.4	-11.3	-12.9	-15.9

(b)

Algorithm	Window size			
	2	**3**	**4**	**5**
Ridor	42.6	45.2	48.1	42.8
J48	50.9	52.3	54.6	48.5
RF	64.0	65.6	64.8	57.4
FT	65.0	68.8	70.1	70.0
ADTree	48.0	54.4	41.8	53.2
IBk	36.1	37.2	35.4	33.8
NaiveBayes	66.0	69.6	70.7	70.1
BayesNet	69.8	71.2	**72.6**	71.8

(c)

Algorithm	Window size			
	2	**3**	**4**	**5**
Ridor	97.9	98.0	97.2	98.6
J48	96.4	96.7	96.9	97.9
RF	98.0	98.2	98.5	99.3
FT	94.0	94.4	95.6	96.6
ADTree	96.6	96.7	96.8	97.4
IBk	97.8	98.2	99.1	**99.7**
NaiveBayes	74.4	72.8	72.5	73.1
BayesNet	74.7	73.3	73.4	73.2

(d)

Figure 3 Attributes tested near the root of a 139 node tree constructed by J48

```
criticalPointSize = tiny
| nHydroHydrophilicWb2 ≤ 1
|      | xlogp3AtPositionA2 ≤ -1.5: noStart (3246.0/816.0)
|      | xlogp3AtPositionA2 > -1.5: helixStart (51.0/24.0)
| nHydroHydrophilicWb2 > 1
|      | rotatablebondcountAtPositionB1 ≤ 1
...
|      | rotatablebondcountAtPositionB1 > 1
...
criticalPointSize = small
| criticalPtGroup = polarweak
|      | chargeAtPositionGroupA2 = negativeneutral: helixStart (1778.0/390.0)
|      | chargeAtPositionGroupA2 = neutralpositive
...
| criticalPointGroup = nonpolarweak: helixStart (1042.0/35.0)
criticalPointSize = large
| chargeAtPositionGroupA2 = negativeneutral
|      | sizeAtPositionGroupB1 = tinysmall
...
|      | sizeAtPositionGroupB1 = smalllarge
...
```

acids per turn of the helix, which places the C=O group of amino acid in position P exactly in line with the H-N group of amino acid P + 4. This happens for all algorithms in the prediction of the start of the helix and for most algorithms in the prediction of inner and end points of helices. Since β-strands do not have a periodic structure the window size with the best TPR are 3 and 4 suggesting that close neighbour residues are sufficient for making good predictions.

We have also investigated the use of the different types of attributes. We have inspected the models constructed by Functional Tree and by J48. There is no significant difference in the percentage of the different types of attributes between the terminal points of β-strands and the inner points. There is, however, in α-helices a significant difference in the use of attributes that differentiate properties of the window before the remarkable point and properties of the window after the remarkable point. The number of such attributes are much higher in the trees predicting the start or end-point of an helix than the trees predicting inner positions.

For some data mining applications having a very high accuracy is not enough. In some applications it would be very helpful if one can extract knowledge that helps in the understanding of the underlying phenomena that produced the data. That is very true for most of Biological problems addressed using data mining techniques. Some of the algorithms used in this study can produce models that are intelligible to experts, such as J48 and Ridor. Using J48 we manage to produce a small size decision tree (shown in Figure 3) that uses very informative attributes near the root of the tree.

5 Conclusions and future work

In this paper we have addressed a very relevant problem in Molecular Biology, namely that of predicting the occurrence of a secondary structure. To study these problems we have collected sequences of amino acids from proteins described in the PDB. For each problem of predicting a 'remarkable' point is a specific structure we have defined two class values: a class of sequences were the 'remarkable' point in study occurs and; all other types of sequences where other remarkable points not in study occur.

We have applied a set of Machine Learning algorithms and almost all of them made predictions above the naive procedure of predicting the majority class. We have achieved a maximum score of 99.6% accuracy and 72.6% True Positive Rate with an algorithm called Functional Tree. We have also managed to construct a small decision tree that has accuracy under 80%, but that is an intelligible model that can help in unveiling the chemical justification of the formation of α-helices.

Acknowledgements

This work has been partially supported by the projects ILP-Web-Service (PTDC/EIA/70841/2006), HORUS (PTDC/EIA-EIA/100897/2008), STAMPA (PTDC/EIA/67738/2006), and by the Fundação para a Ciência e Tecnologia.

References

Aha, D. and Kibler, D. (1991) 'Instance-based learning algorithms', *Machine Learning*, Vol. 6, pp.37–66.

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K. and Walter, P. (2002) *Molecular Biology of the Cell*, 4th ed., Garland Publishing, New York.

Blader, M., Zhang, X. and Matthews, B. (1993) 'Structural basis of aminoacid alpha helix propensity', *Science*, Vol. 11, pp.1637–1640.

Breiman, L. (2001) 'Random forests', *Machine Learning*, Vol. 45, No. 2, pp.5–32.

Breiman, L., Friedman, J.H., Olshen, R.A. and Stone, C.J. (1984) *Classification and Regression Trees*, Wadsworth International Group, Belmont, California.

Chou, P. and Fasman, G. (1978) 'Prediction of secondary structure of proteins from their amino acid sequence', *Advances in Enzymology and Related Areas of Molecular Biology*, Vol. 47, pp.45–148.

Cuff, J., Clamp, M., Siddiqui, A., Finlay, M., Barton, J. and Sternberg, M. (1998) 'JPRED: a consensus secondary structure prediction server', *J. Bioinformatics*, Vol. 14, No. 10, pp.892–893.

Fonseca, N.A., Camacho, R. and Magalhaes, A. (2008) 'A study on amino acid pairing at the n- and c-termini of helical segments in proteins', *PROTEINS: Structure, Function, and Bioinformatics*, Vol. 70, No. 1, pp.188–196.

Freund, Y. and Mason, L. (1999) 'The alternating decision tree learning algorithm', *Proceeding of the Sixteenth International Conference on Machine Learning*, Bled, Slovenia, pp.124–133.

Frishman, D. and Argos, P. (1997) 'Seventy-five percent accuracy in protein secondary structure prediction', *Proteins*, Vol. 27, pp.329–335.

Gama, J. (2004) 'Functional trees', *Machine Learning*, Vol. 55, No. 3, pp.219–250.

Gsponer, J., Haberthur, U. and Caflisch, A. (2003) 'The role of side-chain interactions in the early steps of aggregation:Molecular dynamics simulations of an amyloid-forming peptide from the yeast prion Sup35', *Proceedings of the National Academy of Sciences-USA*, Vol. 100, No. 9, pp.5154–5159.

John, G.H. and Langley, P. (1995) 'Estimating continuous distributions in Bayesian classifiers', *Eleventh Conference on Uncertainty in Artificial Intelligence*, San Mateo, Morgan Kaufmann, pp.338–345.

Jones, T.D. (1999) 'Protein secondary structure prediction based on position-specific scoring matrices', *Journal of Molecular Biology*, Vol. 292, pp.195–202.

King, R. and Sternberg, M. (1990) 'A machine learning approach for the protein secondary structure', *Journal of Molecular Biology*, Vol. 214, pp.171–182.

King, R. and Sternberg, M. (1996) 'Identification and application of the concepts important for accurate and reliable protein secondary structure prediction', *Protein Sci.*, Vol. 5, pp.2298–2310.

Kneller, F.C.D. and Langridge, R. (1990) 'Improvements in protein secondary structure prediction by an enhanced neural network', *Journal of Molecular Biology*, Vol. 216, pp.441–457.

Krittanai, C. and Johnson, W.C. (2000) 'The relative order of helical propensity of amino acids changes with solvent environment', *Proteins: Structure, Function, and Genetics*, Vol. 39, No. 2, pp.132–141.

Landwehr, N., Hall, M. and Frank, E. (2005) 'Logistic model trees', *Machine Learning*, Vol. 95, Nos. 1–2, pp.161–205.

Muggleton, S. (Ed.) (1992) *Inductive Logic Programming*, Academic Press.

Muggleton, S. and Feng, C. (1990) 'Efficient induction of logic programs', *Proceedings of the First Conference on Algorithmic Learning Theory*, Tokyo, Ohmsha.

Petsko, G.A. and Petsko, G.A. (2007) *Protein Stucture and Function (Primers in Biology)*, New Science Press Ltd.

Qian, N. and Sejnowski, T.J. (1988) 'Predicting the secondary structure of globular proteins using neural network models', *Journal of Molecular Biology*, Vol. 202, pp.865–884.

Quinlan, R. (1993) *C4.5: Programs for Machine Learning*, Morgan Kaufmann Publishers, San Mateo, CA.

Richardson, J. and Richardson, D. (1988) 'Amino acid preferences for specific locations at the ends of α-helices', *Science*, Vol. 240, pp.1648–1652.

Rost, B. (1996) 'Phd: predicting 1d protein structure by profile based neural networks', *Meth.in Enzym*, Vol. 266, pp.525–539.

Rost, B. and Sander, C. (1993) 'Prediction of protein secondary structure at better than 70% accuracy', *Journal of Molecular Biology*, Vol. 232,No. 2, pp.584–599.

Salamov, A. and Solovyev, V. (1995) 'Prediction of protein structure by combining nearest-neighbor algorithms and multiple sequence alignments', *J. Mol. Biol.*, Vol. 247, pp.11–15.

Sela, M., White, F.H. and Anfinsen, C.B. (1957) 'Reductive cleavage of disulfide bridges in ribonuclease', *Science*, Vol. 125, pp.691–692.

Sternberg, M., Lewis, R., King, R. and Muggleton, S. (1992) 'Modelling the structure and function of enzymes by machine learning', *Proceedings of the Royal Society of Chemistry: Faraday Discussions*, Vol. 93,pp.269–280.

Wang, G. and Dunbrack Jr., R. (2003) 'PISCES: a protein sequence culling server', *Bioinformatics*, Vol. 19, pp.1589–1591.

Witten, I.H. and Frank, E. (2005) *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd ed., Morgan Kaufmann.

Notes

[1]http://dunbrack.fccc.edu/Guoli/PISCES.php
[2]http://www.rcsb.org/pdb/home/home.do
[3]Start, inner position and end of a secondary structure.
[4]Also known as Contingency Table.
[5]Basically RandomForest constructs several CART-like trees (Breiman *et al.*, 1984) and produces its prediction by combining the prediction of the constructed trees.