Predicting the start of protein α-helices using Machine Learning algorithms

Rui Camacho[1], Rita Ferreira[1], Natacha Rosa[1], Vânia Guimarães[1], Nuno A. Fonseca[2], Vítor Santos Costa[2,3], Miguel de Sousa[4] and Alexandre Magalhães[4]

[1] LIAAD & Faculdade de Engenharia da Universidade do Porto, Portugal
[2] CRACS-INESC Porto LA, Portugal
[3] DCC-Faculdade de Ciências da Universidade do Porto, Portugal
[4] REQUIMTE/Faculdade de Ciências da Universidade do Porto, Portugal

Abstract. Proteins are molecules that play a fundamental role in the functioning of living organisms. They actively participate in more than 90% of chemical activity of our body. Protein function is related to their 3-D structure, which is known to be determined by their primary structure, i.e. the linear sequence of amino acids. It is, therefore, of great importance to be able to understand and predict how the 3D-structure is achieved from the linear sequence of amino acids that compose the protein. Predicting the 3D-structure from the linear sequence of amino acids (primary structure) is a major step and it usually breaks into two phases. First, we predict the secondary structure (α-helices and β-sheet); from the secondary structure we then predict the 3D-structure.
In this paper we report on the application of Machine Learning methods to predict the secondary structure of proteins, specifically the prediction of the starting position of α-helices, from sequences of residues around the starting point and also based on a set of properties of the amino acids. We have used information of the proteins collected in the Protein Data Bank (PDB) and applied Machine Leaning algorithms encoded in the Weka software package. We achieved 84.4% accuracy on the prediction of the starting point of the α-helices.

1 Introduction

Proteins are complex structures synthesised by living organisms. They are actually a fundamental type of molecules and can 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 [1]. 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 [2] 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* made of 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 (parts of the secondary structure) based on the sequence of amino acids that compose a protein. More specifically, in this study we have built models to predict the start of α-helices. We have applied Machine Learning to construct such models. We have collected the sequences of 1499 proteins from the PDB and have assembled data sets that were further used by Machine Learning algorithms to construct the models. We have applied rule induction algorithms, decision trees, functional trees, Bayesian methods, and ensemble methods. We have achieved a 84.4% accuracy and were able to construct some small and intelligible models.

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, connect by peptide bonds between the carboxyl and amino groups of adjacent amino acid residues as shown in Figure 1b) [3]. 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 side chain that determines the identity 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 side chains. These extra bonds are responsible for the secondary and tertiary structure of a protein [4].

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 as α-helices and β-sheets, as seen in Figure 2. The tertiary structure results for the folding of α-helices or β-sheets.

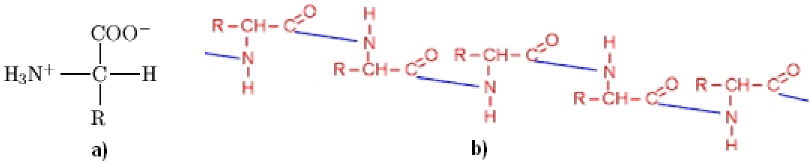

Fig. 1. a) General Structure of an amino acid; side chain is represented by the letter R. b) A fraction of a proteic chain, showing the peptide bounds.

The quaternary structure results from the interaction of two or more polypeptide chains.

Secondary structures, α-helices and β-sheets, were discovered in 1951 by Linus Carl Pauling. These secondary structures 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.

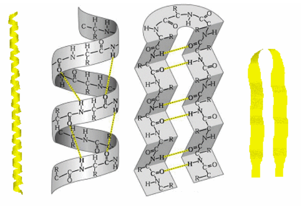

Fig. 2. Secondary structure conformations of a protein: α-helices (left); β-sheet (right).

3 Related Work

Arguably, protein structure prediction is a fundamental problem in Bioinformatics. Early work by Chou et al. [5], based on single residue statistics, looked for contiguous regions of residues that have an high probability of belonging to a secondary structure. The protein samples used was very small which resulted in an overestimation in accuracy of the reported study.

Qian et al [6] used neural networks to predict secondary structures but achieved an accuracy of only 64.3%. They used a window (of size 13) technique where the secondary structure of the central residues was predicted on the base of its 12 neighbours.

Rost and Sanderwith used the PHD [7] method on the RS126 data set and achieved an accuracy of 73.5%. JPRED [8], exploiting multiple sequence alignments, got an accuracy of 72.9%. NNSSP [9] is a scored nearest neighbour method by considering position of N and C terminal in α-helices and β-strands. Its prediction accuracy for RS126 data set achieved 72.7%. PREDATOR [10] 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 [11] used a position-specific scoring matrix generated by PSI-BLAST to predict protein secondary structure and achieved 78.3. DSC [12] used amino acid profile, conservation weights, indels, hydrophobicity were exploited to achieve 71.1% prediction accuracy in the RS126 data set.

Using a Inductive Logic Programming (ILP) another series of studies improved the secondary structure prediction score. In 1990 Muggleton et al. [13] used only 16 proteins (in contrast with 1499 used in our study) and the GOLEM [14] 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 have been reported by [15] using Neural Networks but achieving only 75% accuracy. The propositional learner PROMIS[16, 17] achieved 73% accuracy on the GOLEM's 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 [18]. 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 [19]. According to Blader et al. [20] 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 [21].

4 Experiments

4.1 Experimental Settings

To construct models to predict the start of α-helices we have proceeded as follows. We first downloaded a list of low homology proteins from the Dunbrak web site [22][1]. The downloaded list contained 1499 low homology proteins. We then downloaded the PDBs[2] for each of the protein in the list. Each PDB was then processed in order to extract secondary structure information and the linear sequence of residues of the protein.

Each example of a data set is a sequence of a fixed number of residues (window) before and after the start or end of a secondary structure. The window size is fixed for each data set and we have produced 4 data sets using 4 different window sizes. To obtain the example sequences to use we selected sequences such that they 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-helix.

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 of

[1] http://dunbrack.fccc.edu/Guoli/PISCES.php

[2] `http://www.rcsb.org/pdb/home/home.do`

secondary structures. In this paper, however, we just address the first step of the study, namely, we focus on the start of α-helices.

The size of the data sets, for the different window sizes, are shown in Table 1.

Table 1. Characterisation of the four data sets according to the window size.

Window size	**2**	**3**	**4**	**5**
Data set size	62053	49243	40529	34337
Number of attributes	253	417	581	745

The attributes used to characterise the examples are of two main types: whole structure attributes; and, window-based attributes. 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 used a window of size W before the "special" point (start or end of either a helix or strand), the "special" point and a window of size W after the "special" point. For each of these regions, whenever appropriate, we have computed a set of properties based on the set of individual properties of residues listed in Table 2.

Table 2. List of amino acid properties used in the study.

polarity	hydrophobicity	size	isoelectricpt
charge	h-bonddonor	xlogp3	side chain polarity
acidity	rotatable bond count	h-bondacceptor	side chain charge

For each amino acid of the window and amino acid property we compute other attributes, namely: 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. Altogether there are between 253 (window size of 2) to 745 (window size of 5) attributes. We have used boolean values: a sequence includes the start of an helix; the sequence does not contain a start of an helix. All collected sequences where an helix does not start were included in the "nonStartHelix" class. These later sequences include interior of α-helices, end points of α-helices, start, interior and end points of beta strands.

The experiments were done in a machine with 2 quad-core Xeon 2.4GHz and 32 GB of RAM, running Ubuntu 8.10. We used machine learning algorithms from the Weka 3.6.0 toolkit [23]. We used a 10-fold cross validation procedure to estimate the quality of constructed models. We have used rule induction algorithms (Ridor), decision trees (J48 [24] and ADTree [25]), functional trees

(FT [26][27]), instance-based learning (IBk [28]), bayesian algorithms (NaiveBayes and BayesNet [29]) and an ensemble method (RandomForest [30]).

4.2 Experimental Results

Table 3. Accuracy results (%) of the different algorithms on data sets with windows of size 2, 3, 4 and 5 residues before and after helix start.

	Window size			
Algorithm	**2**	**3**	**4**	**5**
Ridor	83.4	80.6	76.1	77.3
J48	83.9	81.1	79.4	77.0
RandomForest	84.4	81.6	78.4	77.1
FT	79.9	80.5	80.2	75.5
ADTree	83.4	80.3	75.1	76.1
IBk	81.5	76.1	75.2	70.4
NaiveBayes	71.1	66.1	63.2	62.9
BayesNet	70.3	66.2	64.2	64.0
ZeroR	81.5	76.9	72.4	67.8

The results obtained with the Machine Learning algorithms are resumed in Table 3. Apart from the Bayesian methods, most algorithms achieved an accuracy value above the ZeroR predictions. The ZeroR algorithm is used here as the baseline predictor, it just predicts the majority class. The algorithm that achieved the best accuracy values was RandomForest, that is an ensemble method. Basically RandomForest constructs several CART-like trees [31] and produces its prediction by combining the prediction of the constructed trees.

For some data mining applications having a very high accuracy is not enough. In some applications it would be very helpful 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. In the problem at hands in this paper we have algorithms that can produce models that are intelligible to experts. J48 and Ridor are examples of such algorithms. Using J48 we mange 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 when, in a sequence of amino acids, an α-helix will start forming. To study this problem we have collected sequences of amino acids from proteins described in the PDB. We have defined two class values: a class of sequences were an α-helix starts forming and; all other types of sequences where an α-helix does not start.

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

```
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
...
```

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

have achieved a maximum score of 84.4% accuracy with an ensemble algorithm called Random Forest. We have also managed to construct a small decision tree that has smaller accuracy than 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 project ILP-Web-Service (PTDC-/EIA/70841/2006), project STAMPA (PTDC/EIA/67738/2006), and by the Fundação para a Ciência e Tecnologia.

References

1. Pietzsch, J.: The importance of protein folding. Horizon Symposia (2009)
2. Sela, M., White, F.H., Anfinsen, C.B.: Reductive cleavage of disulfide bridges in ribonuclease. Science **125** (1957) 691–692
3. Petsko, G.A., Petsko, G.A.: Protein Stucture and Function (Primers in Biology). New Science Press Ltd (2007)
4. Saraiva, L., Lopes, L., Universidade Nova de Lisboa, Instituto de Tecnologia Química e Biológica (2007)
5. P.Y., C., G.D., F.: Prediction of secondary structure of proteins from their amino acid sequence. Advances in Enzymology and Related Areas of Molecular Biology **47** (1978) 45–148
6. N., Q., J., S.T.: Predicting the secondary structure of globular proteins using neural network models. Journal of Molecular Biology **202** (1988) 865–884
7. B., R.: Phd: predicting 1d protein structure by profile based neural networks. Meth.in Enzym **266** (1996) 525–539
8. J.A., C., M.E., C., A.S., S., Finlay.M., J.G., B., MJE., S.: Jpred : a consensus secondary structure prediction server. J. Bioinformatics **14** (1998) 892–893
9. AA, S., VV., S.: Prediction of protein structure by combining nearest-neighbor algorithms and multiple sequence alignments. J.Mol Biol **247** (1995) 11–15
10. D., F., P., A.: Seventy-five percent accuracy in protein secondary structure prediction. Proteins **27** (1997) 329–335
11. D., J.T.: Protein secondary structure prediction based on position-specific scoring matrices. Journal of Molecular Biology **292** (1999) 195–202

12. King, R., Sternberg, M.: Identification and application of the concepts important for accurate and reliable protein secondary structure prediction. Protein Sci. **5** (1996) 2298–2310
13. Muggleton, S., ed.: Inductive Logic Programming. Academic Press (1992)
14. Muggleton, S., Feng, C.: Efficient induction of logic programs. In: Proceedings of the First Conference on Algorithmic Learning Theory, Tokyo, Ohmsha (1990)
15. D. Kneller, F.C., Langridge, R.: Improvements in protein secondary structure prediction by an enhanced neural network. Journal of Molecular Biology **216** (1990) 441–457
16. King, R., Sternberg, M.: A machine learning approach for the protein secondary structure. Journal of Molecular Biology **214** (1990) 171–182
17. Sternberg, M., Lewis, R., King, R., Muggleton, S.: Modelling the structure and function of enzymes by machine learning. Proceedings of the Royal Society of Chemistry: Faraday Discussions **93** (1992) 269–280
18. JS, R., DC, R.: Amino acid preferences for specific locations at the ends of α-helices. Science **240** (1988) 1648–1652
19. Fonseca, N., Camacho, R., aes, A.M.: A study on amino acid pairing at the n- and c-termini of helical segments in proteins. PROTEINS: Structure, Function, and Bioinformatics **70** (2008) 188–196
20. Blader, M., Zhang, X., Matthews, B.: Structural basis of aminoacid alpha helix propensity. Science **11** (1993) 1637–40
21. Krittanai, C., Johnson, W.C.: The relative order of helical propensity of amino acids changes with solvent environment. Proteins: Structure, Function, and Genetics **39** (2000) 132–141
22. G, W., RL, D.J.: Pisces: a protein sequence culling server. Bioinformatics **19** (2003) 1589–1591
23. Witten, I.H., Frank, E.: Data Mining: Practical machine learning tools and techniques. 2nd edn. Morgan Kaufmann (2005)
24. Quinlan, R.: C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, San Mateo, CA (1993)
25. Freund, Y., Mason, L.: The alternating decision tree learning algorithm. In: Proceeding of the Sixteenth International Conference on Machine Learning, Bled, Slovenia (1999) 124–133
26. Gama, J.: Functional trees. Machine Learning **55** (2004) 219–250
27. Landwehr, N., Hall, M., Frank, E.: Logistic model trees. Machine Learning **95** (2005) 161–205
28. Aha, D., Kibler, D.: Instance-based learning algorithms. Machine Learning **6** (1991) 37–66
29. John, G.H., Langley, P.: Estimating continuous distributions in bayesian classifiers. In: Eleventh Conference on Uncertainty in Artificial Intelligence, San Mateo, Morgan Kaufmann (1995) 338–345
30. Breiman, L.: Random forests. Machine Learning **45** (2001) 5–32
31. Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J.: Classification and Regression Trees. Wadsworth International Group, Belmont, California (1984)