LOGICAL ANALYSIS OF DATA APPROACH TO THE PREDICTION OF PROTEIN SECONDARY STRUCTURES

Błaziwicz J., Hammer P.L., Łukasiak P.

1 Institute of Computing Sciences, Poznan University of Technology, Poznan, POLAND
2 Institute of Bioorganic Chemistry, Polish Academy of Sciences, Poznan, POLAND

e-mail: Piotr.Lukasiak@cs.put.poznan.pl

*Corresponding author

Resume

Motivation: The reason that this problem is so important is that the structure of protein is directly dependent on its function. Experimental structure determination, or structure prediction, aids the elucidation of protein function; conversely, synthetic protein sequences might be designed so that the protein performs a desired function. The study of protein structure is therefore not only of fundamental scientific interest in terms of understanding biochemical processes, but also produces very valuable practical benefits.

Results: The obtained results over 70% for three classes of secondary structures are similar or better as compared with other methods for the protein prediction. A comparison has been made with the PHD algorithm and algorithm based on the Rough Set theory. During experiment the set of the most promising amino acids properties has been extracted for secondary structure description. LAD generated simple and strong rules which could be easily interpreted by biologists

Availability: available on request from the authors: Piotr.Lukasiak@cs.put.poznan.pl

Introduction

The first level of the protein structure, termed primary structure, refers just to the sequence of amino acids in the protein. Decades ago it was found that polypeptide chains can sometimes fold into regular structures; that is, structures which are the same in shape for different polypeptides. These structures create the second level of protein structure. When one looks at an actual polypeptide chain, its final shape is made up of secondary structures, perhaps super-secondary structural features, and some apparently random conformations. This overall structure is referred to as the tertiary structure. The three-dimensional structure of proteins is uniquely determined by its primary structure.

The widely used standard sequence search techniques like BLAST, FASTA searches of sequence databases have very good accuracy when used with care. The most widely used methods are currently the statistics-based GOR method, for its algorithmic simplicity and easy implementation, and the PHD program of Rost and Sander (Rost, Sander, 1993; Rost, 2000). The SSPAL method of Salamov and Solovyev uses multiple overlapping local alignments with sequences of known secondary structure in a nearest neighbour-like way and achieves 71% accuracy, which is uncharacteristic of a single sequence method (their multiple sequence nearest neighbour method NNSSP (Salamov, Solovyev, 1997) is 72% accurate by the same measures). The other popular solution is Monte Carlo method (Skolnick, Kolinski, 1999) trying to determine the structure which minimizes free energy. From the above overview it follows that such tools as machine learning are still needed because it is often difficult for humans to perceive patterns in data, even though strong patterns exist. The idea to create a tool helping molecular biologists was the main reason to choose the new rule-based method – Logical Analysis of Data (Boros et al., 1996).

The Method

The goal of the analysis described in this paper is to create a system which allows to receive as the output the protein secondary structure, based on its primary structure being an input, and to find rules responsible for this effect.

Let \( W = \{A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, Y, V, W\} \) be a set of all amino acids where each letter corresponds to a different amino acid.

The word \( s \) is called a protein primary structure on the condition that letters in this word are in the same order as amino acids in the protein chain are. Let the length of the word \( s \) be denoted as \( C(s) \) and \( A(s,j) \) denote an element of word \( s \), where \( j \) is an integer number from the set \( [1, C(s)] \).

In a similar way, a representation for the secondary structures as for the primary ones, can be created. Let \( F = \{H, E, X\} \) represents a set of all secondary structures, where each letter corresponds to a different secondary structure. A secondary structure is represented here by a word on the relevant alphabet of secondary structures: each kind of a secondary structure has its own unique letter. Let us denote this word by \( d \), where the length of word \( d \) is equal to the length of word \( s \).
Now, we may define the problem as the one consisting in finding a secondary structure of a protein (in a form of word $d$), based on the protein primary structure (i.e. word $s$). Moreover, for each element $A(s,j)$ one should assign an element $A(d,j)$ in the way that the obtained secondary structure $r$ is as close as possible to a real secondary structure of the considered protein.

In the paper, the Logical Analysis of Data (LAD) algorithm is used for the above problem. Examples were obtained from the Dictionary of Secondary Structures of Proteins (DSSP). DSSP contains a description of secondary structures for entries from the Brookhaven Protein Data Base.

The following three sets of secondary structures have been created for the experiments: helix (H) consisting of $\alpha$-helix (structure denoted by H in DSSP), $\beta$-strand (E) consisting of E structure in DSSP; the rest (X) consisting of structures belonging neither to set H nor to set E.

Because of a complexity of the algorithm of Logical Analysis of Data it is hard to present all aspects of this method. An interested reader is referred to (Boros et al., 1996; Boros et al., 1997) for a more detailed description of the Logical Analysis of Data method.

The first step one has to do, is to prepare a set of observations (based on a protein sequence) to be acceptable by the LAD. Below an example is presented, that illustrates the way a protein chain is changed into a set of observations. Let us consider a protein chain called $4gr1$ (in PDB). The first and the last fifteen amino acids in the sequence are shown below:

\[ \text{VASYDYLVIGGGSGG} \ldots \text{VAIHPTSSEELVTLR} \]

For every amino acid the corresponding secondary structure in DSSP is given as follows:

\[ \_EE\_EEEE\_SHHH \ldots \_SS\_SGGGGGS \]

One may change this structure into secondary structures involving three main secondary structures only in the way depicted below:

\[ XEEXXXEEEEXXXHHH \ldots XXXXXXXHHHHHXXX \]

At the end of a chain consisting of $n$ amino acids one obtains a set consisting of $n$ observations as shown in Table 1.

A window of length 6 generates an observation with 6 attributes ($a_{-3}, a_{-2}, a_{-1}, a_0, a_1, a_2$) representing a secondary structure corresponding to the amino acid located in place $a_0$. Of course, at this moment all values of attributes are symbols of amino acids. Secondary structures on the boundaries have been omitted from the consideration.

### Table 1. An example transformation from a sequence to a set of observations.

<table>
<thead>
<tr>
<th>#</th>
<th>Condition attributes</th>
<th>Code in DSSP</th>
<th>Codes of the three secondary structure</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>* * V A S Y</td>
<td>E</td>
<td>E</td>
</tr>
<tr>
<td>2</td>
<td>* * V A S Y D</td>
<td>S</td>
<td>X</td>
</tr>
<tr>
<td>3</td>
<td>V A S Y D V</td>
<td></td>
<td>X</td>
</tr>
</tbody>
</table>

The last step of the preprocessing is to replace in each observation symbols of amino acids (treated as attributes) with numbers representing relevant properties of amino acids. During experiment only the physical and chemical properties of the amino acids offered by ProtScale have been taken into account. Originally we considered 54 properties, but, after a discussion with domain experts, 28 of them have been chosen for the first experiments. For a detailed description of all properties see (Blazewicz et al., 2001). At the end from the set of 54 properties, 6 of them have been extracted which had the most important influence on the created secondary structures.

### Results and Discussion

85 protein chains have been chosen into consideration. Using FASTA algorithm we checked annotated alignment of related sequences in the considered set. FASTA recognized 50 sequences with no alignment, 9 groups consisted of 2 aligning sequences, 2 groups consisted of 3 aligning sequences and one group consisted of 11 aligning sequences. Based on these 85 protein chains about 20 000 observations have been created using the algorithm described. As one mentioned above, originally we considered 54 properties, but after a discussion with domain experts 28 of them have been chosen for the experiment. For the next part of experiments one decided to create a set consisting of 2000 observations and apply 5-fold cross validation test. These 2000 observations were selected randomly from the set of 20000 observations described above. The number of observations were smaller because we decided to enlarge the number of attributes. Now, each observation consisted of 12 attributes, where the first 6 corresponded to one property, and the last six corresponded to another property.

The mix of the properties increased the accuracy by 10% as compared to the results obtained using each property separately (Blazewicz et al, 2001).
In Figure one can see results obtained using LAD with results obtained using PHD method. For PHD only results for class H and E were shown, because only the results for these two structures were presented in (Rost, Sander, 1993). Average results are similar and none of the methods proved its superiority, but LAD gave also rules which could explain for biologists some features of the phenomenon like protein prediction problem. Some rules explored during experiments are presented in Table 2.

<table>
<thead>
<tr>
<th>#</th>
<th>(a_0)</th>
<th>(a_{-1})</th>
<th>(a_{-2})</th>
<th>(a_{-3})</th>
<th>Property</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>&gt;-0.705</td>
<td>&gt;0.285</td>
<td>&lt;0.065</td>
<td>------</td>
<td>Hydrophobicity scale ((\pi-r))</td>
</tr>
<tr>
<td>2</td>
<td>&lt;0.620</td>
<td>&lt;0.130</td>
<td>&gt;1.795</td>
<td>------</td>
<td>&gt;-0.020</td>
</tr>
<tr>
<td>3</td>
<td>------</td>
<td>&gt;1.745</td>
<td>&lt;0.195</td>
<td>&gt;1.225</td>
<td>&gt;1.795</td>
</tr>
</tbody>
</table>

In the context of machine learning algorithms LAD gave results similar and better to the best standing alone methods used for protein prediction problem. In the molecular biology context LAD generates simple and strong rules which could be easily interpreted by biologists. It has been shown above that not only certain global features, such as the presence of helix or strand structure can be predicted with a usable accuracy and reliability using the method like LAD but also the description of the secondary structure of each residue can be predicted.

References