Cluster Algorithm for Protein Secondary Structure Prediction

Manpreet Singh*1, Dr. Gurvinder Singh#2, Dr. Karanjeet Singh Kahlon#3

*Department of Information Technology, Guru Nanak Dev Engineering College
Ludhiana, Punjab, INDIA
impreet78@yahoo.com

#Department of CSE, Guru Nanak Dev University
Amritsar, Punjab, INDIA
gsbawa71@yahoo.com
karanvkahlon@yahoo.com

Abstract—Protein structure determination and prediction has been a focal research subject in the area of drug discovery due to the importance of protein structure in understanding the biological and chemical activities of organisms. The experimental methods used to determine the structures of proteins demand sophisticated equipment and time. A host of computational methods are developed to predict the location of secondary structure elements in proteins for complementing or creating insights into experimental results. The present paper focuses on secondary structure prediction of proteins. The chou-fasman method is implemented to predict the various parameters related to the secondary structure. These parameters include the alpha helix, beta sheets and hairpin turn. The computation is based on cluster algorithm.

Keywords—protein, chou-fasman, cluster algorithm, structure prediction

1. Cluster Analysis

Cluster analyzes the data objects without consulting a class label. The objects are clustered or grouped based on the principle of maximizing the intra-class similarity and minimizing the interclass similarity. Cluster algorithm partition the data space into a set of regions or clusters, to which the examples in the table are assigned, either experimentally or based on some hypothesis. Thus the cluster algorithm gives the different clusters of data from the given sample space [8]. The goal of clustering protein sequences is to get a biologically meaningful partitioning for knowledge discovery. Clustering protein sequences offers several advantages: Proteins can be grouped into families or particular patterns based on the sequence similarity or sequence information. Moreover, protein clustering can be used to facilitate protein structure discovery, which is very important for understanding protein’s function [1].

2. Protein Structural Organization

Proteins are large organic compounds made of amino acids arranged in a linear chain and joined together by peptide bonds between the carboxyl and amino groups of adjacent amino acid residues. The key challenges to bioinformatics relates to exponential increase of raw data arising from the study of the genome and its manifestation. The genome can be thought of as the machine code or raw instructions for creation and operation of biological molecules. The information encoded in DNA results in the creation of proteins which serve as the key building blocks for biological function [10].

The biological mechanisms include the regulation of the conversion (translation) of the raw information encoded in the DNA into the intermediate messages (mRNA) and regulation of the conversion of the mRNA into proteins, as well as modification of the proteins themselves. The protein in turn folds into different types of structures which regulate the different biological functions and processes [3].

These structures are unique in the sense that a given sequence of amino acids always folds into almost the same structure under the same environmental conditions. Structures of proteins are grouped into four categories:

- Primary Structure is the sequence of amino acids in the protein.
- Secondary Structure is the composition of common patterns in the protein. Some patterns are frequently observed in the native states of proteins. This structure class includes regions in the protein of these patterns but it does not include the coordinates of residues.
- Tertiary Structure is the native state, or folded form, of a single protein chain. This form is also called the functional form. Tertiary structure of a protein includes the coordinates of its residues in three dimensional space.
• Quaternary Structure is the structure of a protein complex. Some proteins form a large assembly to function. This form includes the position of the protein subunits of the assembly with respect to each other [7].

Different types of protein structures are shown in fig 1. Protein domains are combinations of secondary structures plus other non-secondary structures, such as loops and turns. A domain defines a modular unit folding that usually has some functional significance [4].

![Protein Structures](image)

**Figure 1.** Different representations of protein structure [11].

### 3. Secondary Structure Prediction

Given a protein sequence with amino acids a₁ a₂ . . . aₙ, the secondary structure prediction problem is to predict whether each amino acid aᵢ is in an α-helix, a β-sheet, or neither. If you know (say through structural studies), the actual secondary structure for each amino acid, then the 3-state accuracy is the percent of residues for which your prediction matches reality. It is called “3-state” because each residue can be in one of 3 “states”: α, β, or other (O). Because there are only 3 states, random guessing would yield a 3-state accuracy of about 33% assuming that all structures are equally likely [6]. There are different methods of prediction with various accuracies. Some of these methods are Chou-Fasman Method, GOR Method and PHD Method [2] [9].

#### 3.1 Chou-Fasman Method

The method takes the conformation parameters and positional frequencies for different amino acids as shown in table 1. The different computation steps are written below:

i). Assign all of the residues in the peptide the appropriate set of parameters.

ii). Scan through the peptide and identify regions where 4 out of 6 contiguous residues have P (α-helix) > 100. That region is declared an α-helix. Extend the helix in both directions until a set of four contiguous residues that have an average P (α-helix) < 100 is reached. That is declared the end of the helix. If the segment defined by this procedure is longer than 5 residues and the average P (α-helix) > P (β-sheet) for that segment, the segment can be assigned as a helix.

iii). Repeat this procedure to locate all of the helical regions in the sequence.

iv). Scan through the peptide and identify a region where 3 out of 5 of the residues have a value of P (β-sheet) > 100. That region is declared as a β-sheet. Extend the sheet in both directions until a set of four contiguous residues that have an average P (β-sheet) < 100 is reached. That is declared the end of the beta-sheet. Any segment of the region located by this procedure is assigned as a beta-sheet if the average P (β-sheet) > 105 and the average P (β-sheet) > P (α-helix) for that region.

v). Any region containing overlapping alpha-helical and beta-sheet assignments are taken to be helical if the average P (α-helix) > P (β-sheet) for that region. It is a beta sheet if the average P (β-sheet) > P (α-helix) for that region.

vi). To identify a bend at residue number j, calculate the following value:

\[
P(t) = f(j) + f(j+1) + f(j+2) + f(j+3)
\]

Where the \( f(j+1) \) value for the \( j+1 \) residue is used, the \( f(j+2) \) value for the \( j+2 \) residue is used and the \( f(j+3) \) value for the j+3 residue is used. If:

- \( P(t) > 0.000075 \).
- The average value for P (turn) > 1.00 in the tetra peptide.
- The averages for the tetra peptide obey the inequality P (α-helix) < P (turn) > P (β-sheet), then a beta-turn is predicted at that location [5].

### 4. Cluster Algorithm

It is a statistical approach to predict the secondary
structure of protein sequence. The input sequence is given and the alpha helices, beta sheets and beta turns are determined for the input sequence as depicted in the figure 2, figure 3 and figure 4 respectively.

Table 1. Conformational parameters and positional frequencies [5]

<table>
<thead>
<tr>
<th>S.No.</th>
<th>Amino Acids</th>
<th>P(a)</th>
<th>P(b)</th>
<th>P(turn)</th>
<th>f(i)</th>
<th>f(i+2)</th>
<th>f(i+2)</th>
<th>f(i+3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Alanine</td>
<td>142</td>
<td>83</td>
<td>66</td>
<td>0.060</td>
<td>0.076</td>
<td>0.035</td>
<td>0.058</td>
</tr>
<tr>
<td>2</td>
<td>Arginine</td>
<td>98</td>
<td>93</td>
<td>95</td>
<td>0.070</td>
<td>0.106</td>
<td>0.099</td>
<td>0.085</td>
</tr>
<tr>
<td>3</td>
<td>Aspartic acid</td>
<td>101</td>
<td>54</td>
<td>146</td>
<td>0.147</td>
<td>0.110</td>
<td>0.179</td>
<td>0.081</td>
</tr>
<tr>
<td>4</td>
<td>Asparagine</td>
<td>67</td>
<td>89</td>
<td>156</td>
<td>0.161</td>
<td>0.083</td>
<td>0.191</td>
<td>0.091</td>
</tr>
<tr>
<td>5</td>
<td>Cysteine</td>
<td>70</td>
<td>119</td>
<td>119</td>
<td>0.149</td>
<td>0.050</td>
<td>0.117</td>
<td>0.128</td>
</tr>
<tr>
<td>6</td>
<td>Glumatic acid</td>
<td>151</td>
<td>37</td>
<td>74</td>
<td>0.056</td>
<td>0.060</td>
<td>0.077</td>
<td>0.064</td>
</tr>
<tr>
<td>7</td>
<td>Glutamine</td>
<td>111</td>
<td>110</td>
<td>98</td>
<td>0.074</td>
<td>0.098</td>
<td>0.037</td>
<td>0.098</td>
</tr>
<tr>
<td>8</td>
<td>Glycine</td>
<td>57</td>
<td>75</td>
<td>156</td>
<td>0.102</td>
<td>0.085</td>
<td>0.190</td>
<td>0.152</td>
</tr>
<tr>
<td>9</td>
<td>Histidine</td>
<td>100</td>
<td>87</td>
<td>95</td>
<td>0.140</td>
<td>0.047</td>
<td>0.093</td>
<td>0.054</td>
</tr>
<tr>
<td>10</td>
<td>Isoleucine</td>
<td>108</td>
<td>160</td>
<td>47</td>
<td>0.043</td>
<td>0.034</td>
<td>0.013</td>
<td>0.056</td>
</tr>
<tr>
<td>11</td>
<td>Leucine</td>
<td>121</td>
<td>130</td>
<td>59</td>
<td>0.061</td>
<td>0.025</td>
<td>0.036</td>
<td>0.070</td>
</tr>
<tr>
<td>12</td>
<td>Lysine</td>
<td>114</td>
<td>74</td>
<td>101</td>
<td>0.055</td>
<td>0.115</td>
<td>0.072</td>
<td>0.095</td>
</tr>
<tr>
<td>13</td>
<td>Methionine</td>
<td>145</td>
<td>105</td>
<td>60</td>
<td>0.068</td>
<td>0.082</td>
<td>0.014</td>
<td>0.055</td>
</tr>
<tr>
<td>14</td>
<td>Phenylalanine</td>
<td>113</td>
<td>138</td>
<td>60</td>
<td>0.059</td>
<td>0.041</td>
<td>0.065</td>
<td>0.065</td>
</tr>
<tr>
<td>15</td>
<td>Proline</td>
<td>57</td>
<td>55</td>
<td>152</td>
<td>0.102</td>
<td>0.301</td>
<td>0.034</td>
<td>0.068</td>
</tr>
<tr>
<td>16</td>
<td>Serine</td>
<td>77</td>
<td>75</td>
<td>143</td>
<td>0.120</td>
<td>0.139</td>
<td>0.125</td>
<td>0.106</td>
</tr>
<tr>
<td>17</td>
<td>Threonine</td>
<td>83</td>
<td>119</td>
<td>96</td>
<td>0.086</td>
<td>0.108</td>
<td>0.065</td>
<td>0.079</td>
</tr>
<tr>
<td>18</td>
<td>Tryptophan</td>
<td>108</td>
<td>137</td>
<td>96</td>
<td>0.077</td>
<td>0.013</td>
<td>0.064</td>
<td>0.167</td>
</tr>
<tr>
<td>19</td>
<td>Tyrosine</td>
<td>69</td>
<td>147</td>
<td>114</td>
<td>0.082</td>
<td>0.065</td>
<td>0.114</td>
<td>0.125</td>
</tr>
<tr>
<td>20</td>
<td>Valine</td>
<td>106</td>
<td>170</td>
<td>50</td>
<td>0.062</td>
<td>0.048</td>
<td>0.028</td>
<td>0.053</td>
</tr>
</tbody>
</table>
Start
Accept amino acid sequence = tline

Count length of tline, store it in lenseq

For i=1 && j=1
i==lenseq
Yes, i++
No, Display "Enter correct sequence"

tline(i)=amino acid sequence(j)
Yes, j=20
No, Display "Alpha helix not possible"

j==20
Yes, Display "Alpha helix not possible"
No, Display "Enter another sequence"
P(a) > 100
Yes, Assign tline(i)=0
No, Assign tline(i)=1, j++

i==lenseq
6
Yes, tot=0
No, Display "Alpha helix not possible"

j==6
Yes, tot=tot+tlinen(i+j)
No, Display "Alpha helix not possible"

tot>4
Yes, Display "Alpha helix not possible"
No, Display "Enter another sequence"

Strtemp1=concatenate(tline(i),tline(i+1),tline(i+2),tline(i+3),tline(i+4),tline(i+5),i++,j++)

Display "Strtemp1"
Startextend=i+j, startalpha=i
Figure 2. Prediction of alpha helices
Figure 3. Prediction of beta sheets
Accept amino acid sequence = tline

Count length of tline, store it in lenseq

For i=1 & j=1

i==lenseq, No → Display "Enter another sequence"

j==20, No → Display "Enter correct sequence"

cline(i)==amino acid sequence(j), No → Display "beta turns not possible"

Tlinen(1)=data(j,4), Tlinen(2)=data(j,5)
Tlinen(3)=data(j,6), Tlinen(4)=data(j,7)
Tlinen(5)=data(j,3), Tlinen(6)=data(j,1)
Tlinen(7)=data(j,2)

j++
i++

For i=1

i==lenseq, No → Display "beta turns not identified"

P(t)=tlinen(i,1)+tlinen(i+1,2)+tlinen(i+2,3)+tlinen(i+3,4)
Avg=(tlinen(i,5)+tlinen(i+1,5)+tlinen(i+2,5)+tlinen(i+3,5))/4

If p(t)>0.000075 & avg >100?

No → Display "beta turns not identified"

Yes → Spa=tlinen(i,6)+tlinen(i+1,6)+tlinen(i+2,6)+tlinen(i+3,6)
Spt=tlinen(i,5)+tlinen(i+1,5)+tlinen(i+2,5)+tlinen(i+3,5)
Sph=tlinen(i,7)+tlinen(i+1,7)+tlinen(i+2,7)+tlinen(i+3,7)

j++
i++
5. Results and Discussion

If the author using LATEAmino acid Sequence given as an input to Cluster Algorithm is as follows:
CAENKLHVRGPTCILFMTWYNDGP
The output produced is as follows:
Potential helix at: CAENKL
Encountered four contiguous P (a) < 0 at: RGPT
Alpha helix predicted in region: CAENKLDHV
Potential beta strand at: GPTCIL
Encountered four contiguous P (b) < 0 at: NDGP Beta strand predicted in region: GPTCILFMTWY
Betas turn at: PTCI
Beta turns at: TWYN

6. Results and Discussion

Secondary structure predictions can be useful in analyzing the relationships between sequences and folding and would also seem to be a logical step towards predicting tertiary structure. The present work applies cluster analysis to the structure prediction. By applying data mining to the input data an optimized result is produced i.e. the secondary structure predicted is of high conformation. This research work will also act as a base for protein function prediction.

We assume that the protein folding and structure formation are independent of the factors like hydrophobic forces, Wander wall forces etc. The future scope of the existing work is written below:

- The system can be extended to predict tertiary structure of the protein.
- Protein structure is largely responsible for its function and hence the existing system can be further extended to even predict the function of the protein from the amino acid sequence.
- Choice of various formats of amino acid sequences can be utilized.

References


