Protein Structure Analysis

Iosif Vaisman

BINF 731

2006

Protein Structure Hierarchy

• Primary - the sequence of amino acid residues
• Secondary - ordered regions of primary sequence (helices, beta-sheets, turns)
• Tertiary - the three-dimensional fold of a protein subunit
• Quaternary - the arrangement of subunits in oligomers.

Secondary Structure Conformations

The Ramachandran Plot

The Ramachandran Plot

Secondary Structure Conformations

<table>
<thead>
<tr>
<th></th>
<th>$\phi$</th>
<th>$\psi$</th>
</tr>
</thead>
<tbody>
<tr>
<td>alpha helix</td>
<td>-57</td>
<td>-47</td>
</tr>
<tr>
<td>alpha-L</td>
<td>57</td>
<td>47</td>
</tr>
<tr>
<td>3-10 helix</td>
<td>-49</td>
<td>-26</td>
</tr>
<tr>
<td>$\pi$ helix</td>
<td>-57</td>
<td>-80</td>
</tr>
<tr>
<td>type II helix</td>
<td>-79</td>
<td>150</td>
</tr>
<tr>
<td>$\beta$-sheet parallel</td>
<td>-119</td>
<td>113</td>
</tr>
<tr>
<td>$\beta$-sheet antiparallel</td>
<td>-139</td>
<td>135</td>
</tr>
</tbody>
</table>

Secondary Structure Prediction

Three-state model: helix, strand, coil
Given a protein sequence:
- NWVLSTAADMQGVTDGMASGLDKD...
Predict a secondary structure sequence:
- LEEEELLLLHHHHHHHHHHL... 

Methods:
• statistical
• stereochemical

Accuracy: 50-85%
Statistical Methods

Residue conformational preferences:
- Glu, Ala, Leu, Met, Gln, Lys, Arg - helix
- Val, Ile, Tyr, Cys, Trp, Phe, Thr - strand
- Gly, Asn, Pro, Ser, Asp - turn

Chou-Fasman algorithm:
- Identification of helix and sheet "nuclei"
- Propagation until termination criteria met

Chou-Fasman Parameters

<table>
<thead>
<tr>
<th>Name</th>
<th>P(a)</th>
<th>P(b)</th>
<th>P(turn)</th>
<th>f(1)</th>
<th>f(1+2)</th>
<th>f(1+4)</th>
<th>f(1+7)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alanine</td>
<td>142</td>
<td>83</td>
<td>66</td>
<td>0.06</td>
<td>0.076</td>
<td>0.035</td>
<td>0.058</td>
</tr>
<tr>
<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>Aspartic Acid</td>
<td>101</td>
<td>58</td>
<td>146</td>
<td>0.147</td>
<td>0.110</td>
<td>0.179</td>
<td>0.081</td>
</tr>
<tr>
<td>Asparagine</td>
<td>67</td>
<td>80</td>
<td>156</td>
<td>0.161</td>
<td>0.083</td>
<td>0.191</td>
<td>0.091</td>
</tr>
<tr>
<td>Cysteine</td>
<td>70</td>
<td>119</td>
<td>119</td>
<td>0.149</td>
<td>0.050</td>
<td>0.137</td>
<td>0.128</td>
</tr>
<tr>
<td>Glutamic 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>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>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>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>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>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>Lysine</td>
<td>114</td>
<td>74</td>
<td>101</td>
<td>0.055</td>
<td>0.115</td>
<td>0.072</td>
<td>0.093</td>
</tr>
<tr>
<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>Phenylalanine</td>
<td>113</td>
<td>129</td>
<td>60</td>
<td>0.059</td>
<td>0.061</td>
<td>0.065</td>
<td>0.065</td>
</tr>
<tr>
<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>Serine</td>
<td>77</td>
<td>76</td>
<td>143</td>
<td>0.120</td>
<td>0.139</td>
<td>0.125</td>
<td>0.104</td>
</tr>
<tr>
<td>Threonine</td>
<td>83</td>
<td>110</td>
<td>96</td>
<td>0.086</td>
<td>0.108</td>
<td>0.065</td>
<td>0.070</td>
</tr>
<tr>
<td>Tyrosine</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>Tyrosine</td>
<td>69</td>
<td>147</td>
<td>114</td>
<td>0.082</td>
<td>0.055</td>
<td>0.114</td>
<td>0.125</td>
</tr>
<tr>
<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>

Chou-Fasman Algorithm

Identification of helix and sheet "nuclei"
- helix - 4 out of 6 residues with high helix propensity (P > 100)
- sheet - 3 out of 5 residues with high sheet propensity (P > 100)
- Propagation until termination criteria met

Turn prediction
1) \( p(t) > 0.000075 \)
2) \( P(\text{turn}) > 1.00 \)
3) \( P(a) < P(\text{turn}) < P(b) \)

where \( p(t) = f(j)f(j+1)f(j+2)f(j+3) \)

Garnier - Osguthorpe - Robson (GOR) Algorithm

Likelihood of a secondary structure state depends on the neighboring residues:

\[
L(S_j) = \sum (S_j; R_{j+m})
\]

Window size - [j-8; j+8] residues
- Accuracy for a single sequence - 60%
- Accuracy for an alignment - 65%

Evolutionary Methods

Taking into account related sequences helps in identification of “structurally important” residues.

Algorithm:
- find similar sequences
- construct multiple alignment
- use alignment profile for secondary structure prediction

Additional information used for prediction
- mutation statistics
- residue position in sequence
- sequence length

Neural Networks

Perceptron

\[
Y = \begin{cases} 
1 \text{ if } \sum w_i x_i > \Theta \\
0 \text{ otherwise }
\end{cases}
\]

Learning process: \( \Delta w_i = (T_p - Y_p) x_i \)
Neural Networks Methods

Helix Sheet

Output layer (2 units)

Hidden layer (2 units)

Input layer (7x21 units)

MKFGNFLTYQP [ PELSQTE ] VMKRLVNLGKASEGC...

Stereochemical Methods

Patterns of hydrophobic and hydrophilic residues in secondary structure elements:

• segregation of hydrophobic and hydrophilic residues
• hydrophobic residues in the positions 1-2-5 and 1-4-5
• oppositely charged polar residues in the positions 1-5 and 1-4 (e.g. Glu (i), Lys (i+4))

Definitions of hydrophobic and hydrophilic residues (hydrophobicity scales) are ambiguous

Stereochemical Methods

Hydropathic correlations in helices and sheets

<table>
<thead>
<tr>
<th></th>
<th>F-F</th>
<th>F-L</th>
<th>L-F</th>
<th>L-L</th>
</tr>
</thead>
<tbody>
<tr>
<td>α</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>-</td>
</tr>
</tbody>
</table>

β

Polar face

Hydrophobic face

1st helix in Myoglobin

Second strand in CD8

Polar face

Hydrophobic face

Chemotaxis protein CheY

Residues 81-88

Hydrophobic face
Accuracy of Prediction

$$Q_1 = \frac{PH + PE + PC}{N}$$

$$W = \log\frac{TP \times TN}{FP \times FN}$$

Range: 50-85%