Protein structure prediction. Implications for the biologist

G Deléage, C Blanchet, C Geourjon

Institute of Biology and Chemistry of Proteins, 7, passage du Vercors, 69367 Lyon cedex 07, France

(Received 16 July 1997; accepted 18 September 1997)

Summary — Recent improvements in the prediction of protein secondary structure are described, particularly those methods using the information contained into multiple alignments. In this respect, the prediction accuracy has been checked and methods that take into account multiple alignments are 70% correct for a three-state description of secondary structure. This quality is obtained by a 'leave-one out' procedure on a reference database of proteins sharing less than 25% identity. Biological applications such as 'protein domain design' and structural phylogeny are given. The biologist's point of view is also considered and joint predictions are encouraged in order to derive an amino acid based accuracy. All the tools described in this paper are available for biologists on the Web (http://www.ibcp.fr/predict.htm).

protein structure prediction / sequence analysis / protein domain design / biocomputing / SOPMA

Introduction

Secondary structure prediction from protein sequences is perhaps the most studied problem in computational molecular biology. In addition to its intrinsic interest to biology, it has also become a starting point for tertiary structure prediction. Historically, the numerous methods yet available (more than 20 have been developed) were classified in four different categories: i) the statistical methods; ii) the similarity based methods; iii) learning methods; and iv) empirical methods. More recently, the incorporation of multiple alignment of related protein sequences into predictive methods makes this classification obsolete and today we should consider methods that are based either on single sequence or on multiple alignment. In this paper, we present these predictive methods with their extensions and the way the biologist (obviously not familiar with these computing methods) should integrate the information of secondary structure into his experiments.

Single sequence based methods

The first developed methods were statistical methods even if the size and the sampling representativity criteria were obviously not met. The pioneering work of Chou and Fasman [1–3] consisted in the derivation of the frequency with which individual residues or short sequences are found in given conformational states (helix, sheet, turn and a-periodic). Further improvements in this method have been achieved by distinguishing internal and external β-sheets [4], the integration of structural class prediction [5], a profile consideration in the special case of β-α-β proteins [6]. Another approach [7] consisted in the application of information theory to protein sequence and secondary structure. In this latter method the predictive parameters have been revisited and the pair information has been added [8]. However, statistical methods fail to improve despite the increasing size of the reference database [9]. About 1C years ago similarity based prediction methods appeared [10, 11]. Briefly, each polypeptide of the query sequence is compared with all peptides with the same length in a reference database. If both peptides are similar (by using a secondary structure substitution matrix), the known observed conformation is given to the query peptide with the similarity value. The process is allowed to continue until all peptides have been compared and the scores are summed up. The final predicted state is the one with the highest conformational score. This strategy has been used in several methods with a 17 amino acid long peptide [12] and combined with an automated optimization procedure [13]. More recently several groups have used neural networks [14–16] with slight, significant improvements over previously described methods.

Incorporation of multiple alignments

A promising starting point for predicting a structure for a given amino acid sequence is to determine whether that sequence is sufficiently similar to any other sequence for which biophysical data (ideally a 3D structure) are available. In this context, several authors have investigated the possibility of improving the prediction accuracy by taking into account the information contained into a multiple alignment of related protein sequences [17–19]. This approach has been applied to both statistical and similarity based methods for the prediction of α-helices [19].
optimized prediction methods [18] and to neural network methods [20]. One interesting feature in the Rost and Sander method is a quality index given at each amino acid position [20] and that considering only the residues with highest scores may yield as much as more than 86% on about 30% of residues [21]. It seems now that no significant improvements in secondary structure prediction will be obtained without taking into account the effect of tertiary structure onto local folding.

Comparison of predictive success

The objective comparison of predictive accuracy has been addressed for a long time. It requires at least three problems to be solved: i) the reference database; ii) calculating the prediction accuracy; and iii) the reproducibility of methods developed onto different sets. The first problem to address is the definition of a reference database [22]. In this context, the cut-off value of identity between protein sequence is critical and the cut-off yet admitted is 25% [16] (a value below which homology modeling is rather hazardous). Another problem is to derive the secondary structure from atomic coordinates. It has been shown that using different criteria (geometric or energetic) may result in an overall agreement as low as 70% [23]. The first accuracy index is the global percentage of correctly predicted residues Q% and has been used in an objective comparison of different methods [9]. This parameter can also be used on a state-by-state basis, but the percentage of correct prediction does not give any information about the correlation between prediction and observation. That is why a correlation parameter has been introduced and applied to secondary structure prediction [24]. This C parameter varies from -1 (total anti-correlation) to 1 (total correlation) and is nil if there is no correlation between prediction and observation. More recently, the segment overlap [25] is a measure of how the predicted segments correspond to observed segments. This is the single method that is not based on a residue-by-residue comparison. The last but not least bias that must be avoided is the independency between learning and testing sets for methods based on neural networks. Indeed, even if a jackknife procedure could be applied (removing the protein from learning set prior its own prediction), this procedure seems to be too time consuming to be extensively used. A comparison of different methods is given in table I. Methods based on a single sequence have a global success rate of about 60% when checked on the 126 protein chains sharing less than 25% identity [16]. On the other hand, methods that exploit the information contained into the

Table I. Comparison of secondary structure methods on the Rost and Sander database.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>α-helix</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cα</td>
<td>0.42</td>
<td>0.53</td>
<td>0.45</td>
<td>0.61</td>
<td>0.60</td>
<td>0.65</td>
<td></td>
</tr>
<tr>
<td>σα</td>
<td>10.92</td>
<td>9.06</td>
<td>10.95</td>
<td>7.9</td>
<td>8.5</td>
<td>–</td>
<td></td>
</tr>
<tr>
<td>SOV</td>
<td>0.609</td>
<td>0.601</td>
<td>0.603</td>
<td>0.737</td>
<td>–</td>
<td>0.706</td>
<td></td>
</tr>
<tr>
<td>Lp</td>
<td>8.2</td>
<td>7.9</td>
<td>6.3</td>
<td>7.3</td>
<td>9.3</td>
<td>7.5</td>
<td></td>
</tr>
<tr>
<td>Lo</td>
<td>9.1</td>
<td>9.1</td>
<td>9.1</td>
<td>9.1</td>
<td>9.1</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>Q3 (%)</td>
<td>53.5</td>
<td>62</td>
<td>53.9</td>
<td>69.9</td>
<td>72</td>
<td>86.9</td>
<td></td>
</tr>
<tr>
<td>β-sheet</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cβ</td>
<td>0.39</td>
<td>0.54</td>
<td>0.38</td>
<td>0.60</td>
<td>0.52</td>
<td>0.61</td>
<td></td>
</tr>
<tr>
<td>σβ</td>
<td>8.99</td>
<td>8.34</td>
<td>8.49</td>
<td>8</td>
<td>8.1</td>
<td>–</td>
<td></td>
</tr>
<tr>
<td>SOV</td>
<td>0.578</td>
<td>0.533</td>
<td>0.536</td>
<td>0.621</td>
<td>–</td>
<td>0.74</td>
<td></td>
</tr>
<tr>
<td>Lp</td>
<td>3.9</td>
<td>3.6</td>
<td>4.1</td>
<td>3.9</td>
<td>5.0</td>
<td>4.0</td>
<td></td>
</tr>
<tr>
<td>Lo</td>
<td>5.1</td>
<td>5.1</td>
<td>5.1</td>
<td>5.1</td>
<td>5.1</td>
<td>5.3</td>
<td></td>
</tr>
<tr>
<td>Q3 (%)</td>
<td>38.8</td>
<td>40.5</td>
<td>42.2</td>
<td>53.3</td>
<td>66</td>
<td>81.3</td>
<td></td>
</tr>
<tr>
<td>Coil</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cc</td>
<td>0.49</td>
<td>0.41</td>
<td>0.53</td>
<td>0.52</td>
<td>–</td>
<td>0.55</td>
<td></td>
</tr>
<tr>
<td>σc</td>
<td>8.49</td>
<td>8.75</td>
<td>9.5</td>
<td>8.1</td>
<td>–</td>
<td>–</td>
<td></td>
</tr>
<tr>
<td>SOV</td>
<td>0.528</td>
<td>0.592</td>
<td>0.585</td>
<td>0.702</td>
<td>–</td>
<td>0.642</td>
<td></td>
</tr>
<tr>
<td>Lp</td>
<td>8.2</td>
<td>7.8</td>
<td>7.6</td>
<td>6.9</td>
<td>–</td>
<td>4.8</td>
<td></td>
</tr>
<tr>
<td>Lo</td>
<td>5.8</td>
<td>5.8</td>
<td>5.8</td>
<td>5.8</td>
<td>–</td>
<td>6.0</td>
<td></td>
</tr>
<tr>
<td>Q3 (%)</td>
<td>69.1</td>
<td>74.2</td>
<td>72.1</td>
<td>74.6</td>
<td>74</td>
<td>82.8</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>57.6</td>
<td>63.1</td>
<td>59.9</td>
<td>68.6</td>
<td>70.8</td>
<td>83.7 (76)</td>
<td></td>
</tr>
</tbody>
</table>

The predictive success is calculated on the Rost and Sander database [16] by using the classical index reported [25]. Briefly, Q is the percentage of correctly predicted residues, C is the correlation coefficient between prediction and observation (C = -1 for negative correlation, C = 0 for absence of correlation and C = +1 for positive correlation), and σ is the standard deviation between predicted secondary content. Lp is the mean predicted length of predicted segments. Lo is the averaged length for observed secondary structure and SOV is the segment overlap index.
Fig 1. The multiple protein sequence analysis (MPSA) program interface. The alignment of sequences that belong to the Fos family of transcription factors was performed from within the MPSA program (version 0.1 for Silicon Graphics) by using the ClustalW (1.6) package and secondary structure predictions come either from local implementations of well-known methods [7, 8, 11] or from the Network Protein Sequence Analysis (http://www.ibcp.fr/predict.html) [18].

multiple alignment of related proteins sequences such as SOPMA [18] and PHD [16] have global success rates of about 70% in the Rost and Sander database. Thus multiple alignment and consensus prediction using profiles significantly improve the prediction accuracy of protein secondary structure. What are the future directions for secondary structure? From the biologist’s point of view, it seems more important to know the correctness of prediction on some given stretches rather than to get a global prediction accuracy for the whole sequence. That is why some authors have investigated the possibility to give an accuracy index for each residue of a sequence. Very often, this index is based on the difference between the conformational state that has the highest value and the second value [8, 20]. Other authors have chosen to investigate the possibility of joint prediction [18, 27] or a combination of several methods based on different principles [13]. This stresses the necessity for the biologist to compare the results obtained by several methods (especially based on different principles), to make consensus prediction and to have the methods available either integrated into a single software or through Internet.

Integration into packages and Web servers availability

In order to give the biologist access to different secondary structure methods coupled with protein sequence analysis methods, a package called ANTHEPROT that integrates several prediction methods has been being developed for several years [28, 29]. This program runs on PC computers (DOS, Windows or IBM rs6000) and is available for academic users (ftp.ibcp.fr). The program is being adapted for other Unix machines (SUN, IBM) and Macintosh. This new program is called MPSA (multiple protein sequence analysis) program. The main window of MPSA showing the combination of multiple alignment with different secondary structure prediction is given in figure 1. In order to investigate the conformational scores for each state (helix, sheet, turn and coil), a graphical view can be displayed. The graphics are fully interactive with cursors, scrolling, zoom, resizing and selection functions. The main window can be sent to a file in either Postscript Adobe or rich text format (RTF)-Microsoft formatted files. Great efforts have been made to allow large alignments (200 sequences of 5000
Applications to biological projects

In this chapter we present three different examples of secondary structure prediction: some have been performed in the absence of 3D data, others have been combined with 3D structural data.

The secondary structure is more often conserved than the sequence since the secondary structure code is degenerated starting from 20 amino acids to finish with three or four different conformational states. That is why some interesting information can be masked in the sequence and become visible after secondary structure prediction. A typical example of this feature is given with the nucleotide binding domains of multidrug resistance proteins (P-glycoprotein).

In the nucleotide binding domains of different proteins, the amino acid conservation (fig 3) is very low since only three amino acids out of 38 are identical into the six sequences. These amino acids correspond to the requisite positions G-x (4)-G-K for nucleotide binding. When the observed secondary structures (first four sequences) are compared with predicted structures (last two sequences), the structural conservation is obvious. This type of investigation permits to validate the multiple alignment and the location of secondary structure elements and to postulate a typical Rossman fold for both nucleotide binding domains of P-glycoprotein.

In the AIDS field, secondary structure prediction coupled with multiple alignments has been used to predict the 'most likely' sequence of HIV gp120 [31] and to propose an explanation at the secondary structure level for the difference in the dimerization rate of HIV-1 and HIV-2 reverse transcriptases (HIV-RT) enzymes [32]. When predicted secondary structures for HIV-1 RT and HIV-2 RT are compared, one can clearly identify the main difference, located in the 312–340 region (fig 4). Thus we formulated the hypothesis that region 312–340 could explain the difference in the dimerization rate of HIV-RT subunits [32]. Then peptides in
these regions were synthesized. The synthetic peptide from HIV-2 clearly folds into an α-helix whereas the HIV-1 peptide is rather unfolded as observed by circular dichroism (Divita, personal communication). More striking, these synthetic peptides have been found to inhibit the subunit dimerization.

Another example is given by the design of the functional protein domain. This recent concept comes with the possibility to determine the tertiary structure of molecules of limited size by NMR. Due to the modular organization of proteins, one should be able to predict the limits of biologically active domains. In our laboratory, this approach has been successfully applied to the fructose repressor DNA binding domain and to heat shock factor. These domains (mainly designed from sequence analysis) have kept DNA binding capabilities and, as shown in figure 5, their structure has been solved by NMR [33].

The last application is the use of predicted secondary structure as a starting point for tertiary structure prediction or folds recognition. We are addressing this problem by a combination of simulated annealing, database sampling, Monte-Carlo and distance geometry methods. In some particular cases (protein mostly helical), structures that differ by less than 3.6 Å from actual structure have been obtained and these results have to be generalized to a representative subset of the PDB. One of the most promising challenges for the next decades is probably the understanding of the folding process and the prediction of protein tertiary structures from sequences.

Fig 4. Predicted secondary structure of HIV-1 and HIV-2 RT. The secondary structure of HIV-1 and HIV-2 RT p66 subunits was predicted by using the SOPM method [13]. Closed and open boxes correspond to the α-helices and β-strands, respectively.

Fig 5. An example of protein domain design. The fructose repressor DNA-binding domain has been designed from both sequence analysis and proteolytic experiments [33].
Acknowledgments

The work presented in this review has been carried out with the support of IMABIO program (CNRS) and MENESR grants (ACC-SV13).

References

10. Sweet R (1986) Evolutionary similarity among peptide segment is a basis for prediction of protein folding. Biopolymers 25, 1565–1577