In unison: regularization of protein secondary structure predictions that makes use of multiple sequence alignments

K.Zimmermann1 and J.-F.Gibrat

Unité de Bio-Informatique, INRA, 78352 Jouy-en-Josas, France

1To whom correspondence should be addressed

We present a method whose purpose is to post-process the fuzzy results of secondary structure prediction methods that use multiple sequence alignments, in order to obtain 'realistic' secondary structures, i.e., secondary structure elements whose length is greater than or equal to some predefined minimum length. This regularization helps with interpretation of the secondary structure prediction.

Keywords: filtering/multiple sequence alignment/prediction/
realistic secondary structure

Introduction

The prediction of a protein’s three-dimensional (3D) structure from its amino acid sequence is an important, albeit unsolved, problem of molecular biology. The development of new methods, such as threading algorithms (see Lemer et al., 1995 and references therein) might represent a crucial step towards this long sought after goal. Threading methods can be described succinctly as 3D alignments. An amino acid sequence is aligned (threaded) onto the 3D core of a protein family (the core is defined as conserved secondary structure elements, \(\alpha\)-helices and \(\beta\)-strands, in the 3D structures of the protein family) and one searches the particular alignment that minimizes an empirical energy function. In practice performing this 3D alignment is a difficult problem. Lathrop (1994) has shown that this is a NP-complete problem. A way of making this problem more tractable is to include the secondary structure prediction in the threading algorithm (J.F.Gibrat, unpublished results). This considerably reduces the alignment space that needs to be explored.

In this context it is necessary to develop methods that, obviously, provide as accurate as possible secondary structure predictions.

Current protein secondary structure prediction methods predict three residue states \(\varepsilon\): the two regular secondary structures, \(\alpha\)-helix \((\varepsilon = H)\) and \(\beta\)-strand \((\varepsilon = b)\), and the rest, random coils or loops \((\varepsilon = c)\). The predictive power of the methods is expressed in terms of the ‘precision’ or ‘quality’ index \(Q_3\) that is defined as the percentage of residues correctly predicted in the three \(\varepsilon\)-states. The precision of current methods lies between 63–67%, the use of multiple sequence alignments generally increases this figure up to about 73% (Levin, 1997).

In this paper we distinguish between the ‘precision’ and the ‘realism’ of the prediction. Due to the co-operative mechanism of the secondary structure the ‘state’ \((H, b, c)\) of a residue depends on the states of its neighbours. This results in lower limits on the lengths of the helix and strand segments in real secondary structures (an \(\alpha\)-helix must be constituted of, at least, one helix turn, corresponding to a minimum length of four residues, and for a \(\beta\)-strands the minimum length is at least two contiguous residues). However, \(Q_3\), by its very definition, is quite insensitive to this ‘realism’ of the predicted secondary structure. In most methods an \(\varepsilon\)-state is simply assigned to a residue according to the maximum value of the corresponding predicted propensity (or probability in methods such as GOR, Garnier et al., 1996) without taking into account the state of its neighbours. This can result in ‘good’ or ‘precise’ (in terms of \(Q_3\)) but ‘unrealistic’ predictions, i.e., some \(\alpha\)-helices or \(\beta\)-strands are too short to represent a real secondary structure element.

In a previous paper (Zimmermann, 1994) a regularization method, referenced here as ‘bio’ Champollion, has been proposed for solving this problem of unrealistic prediction. This method yields the best realistic secondary structure prediction, i.e., no other secondary structure provides a better compatibility with the predicted propensities while, at the same time, respecting the minimum length of the secondary structure segments.

It has been shown that the use of multiple sequence alignments improves significantly the precision \(Q_3\) of the secondary structure prediction methods (Rost and Sander, 1994). The basic idea behind this technique is that aligned residues occupy the same position in the 3D structure and thus they ought to have the same secondary structure. This is certainly an oversimplification since regular secondary structure elements can exhibit slightly different lengths in homologous proteins. However, aside from these edge effects, it remains true that most of aligned residues do have the same ‘consensus’ secondary structure. To use an analogy, if we liken aligned sequences to different voices of a partition, insertions ‘\(-\)’ representing pauses they all should sing ‘in unison’.

The power of this technique comes from the fact that, instead of having a single residue, one now has a distribution of residues characteristic of a particular position in the 3D structure. Using multiple alignment data no doubt increases the ‘signal to noise ratio’ and, on average, result in a more accurate secondary structure prediction. Therefore the additional information contained in multiple alignments, if available, is well worthwhile including in secondary structure prediction methods.

The use of multiple alignment data, while increasing the accuracy of the prediction, also makes the problem of realism of the prediction more subtle. Usually predictions obtained using a simple consensus do not yield realistic secondary structures. Conversely, regularized (realistic) individual predictions do not result in a consensus secondary structure. Both types of constraint, minimum length constraints as well as constraints imposed by the consensus must be considered at the same time.

In this paper we present a method that yields the best consensus realistic secondary structure prediction, i.e., there is no secondary structure which provides a better compatibility with the set of predicted propensities of the whole alignment while, at the same time, respecting the consensus and the
minimum secondary structure segment lengths. The results show that the overall precision of the method, measured by the index $Q_3$, is not affected by this regularization procedure.

**Method**

**Formulation of the problem**

Let us consider an alignment of $M$ sequences, the length of the alignment being $N$ (i.e., $N$ is the number of columns of the alignment) and a secondary structure prediction method (such as GOR, Garnier et al., 1996; or SIMPA96, Levin, 1997) that yields propensities $p_{ij}^ε$ ($i = 1, \ldots, M; j = 1, \ldots, N; ε = H, b, c$). If insertions in the alignment, ‘–’, are treated as pseudo-residues, the table of propensities can be completed by assigning them some constant value. Then the problem of finding the best realistic consensus secondary structure can be formulated as:

$$
\{ε_j, j = 1 \ldots N\} = \text{Arg}(\max (\sum_{i}^M \sum_{j}^N \log p_{ij}^ε, ε \in \{H, b, c\}))
$$

subject to the ‘realism’ constraints:

$$
L_ε \geq L_{\text{min}}
$$

where $L_ε$ is the length of a secondary structure $ε$-segment.

Constraints (Equation 2) insure that secondary structure segments will have a length greater than or equal to the minimum length (Note: the lengths $L$ are calculated without taking into account insertions ‘–’).

The ‘consensus’ constraints are respected implicitly by the fact that the assigned states $ε_j$ ($j = 1, \ldots, N$) do not depend on the ‘line’ index $i$ ($i = 1, \ldots, M$) and so all the residues in any column $j$ of the alignment will have the same secondary structure state.

Notice, here, that:

(i) The standard ‘winner-takes-all’ consensus interpretation of the propensities represents the solution of Equation 1 without the ‘realism’ constraints (Equation 2).

(ii) Our previous paper (Zimmermann, 1994) dealt with the simpler problem

$$
\{ε_j, j = 1 \ldots N\} = \text{Arg}(\max (\sum_{j}^N \log p_{j}^ε, ε \in \{H, b, c\}))
$$

subject to ‘realism’ constraints (Equation 2).

**Algorithm**

In a previous paper (Zimmermann, 1994) two variants of the regularization algorithm to solve Equation 3 have been presented. The algorithm that was used would lead here to a ‘combinatorial explosion’ when used to solve Equation 1. In

---

**Fig. 1.** Prediction of the secondary structure of a prion (hamster prion protein) using multiple sequence alignments. (a) Multiple sequence alignment. The multiple alignment was performed by CLUSTAL (Thompson et al., 1994). (b) Standard GOR (‘winner takes all’) prediction for individual sequences. Standard interpretation of the propensities, in which each residue is assigned the secondary structure corresponding to the maximal propensity (‘winner-takes-all’). The minimum length constraints (Equation 2), $L_{\text{min}} = 6$, $L_{\text{min}} = 6$, as well as the consensus is violated several times. (c) GOR consensus (‘winner-takes-all’). A consensus secondary structure is assigned to each column of the alignment. The consensus is, in this way, respected, but the ‘realism’ constraints (Equation 2) are violated. (d) ‘Bio’ Champollion: individually regularized predictions. (e) ‘In unison’ regularized consensus. The best secondary structure prediction. Consensus as well as ‘realism’ constraints are all respected. (f) Comparison of the results obtained for the first sequence of the alignment by the various methods. Observed, observed secondary structure deduced from the NMR three-dimensional structure (Riek et al., 1996). Note that only part of the protein has been solved by NMR from residue 30 to 140 in the numbering scheme presented here.

863
this paper we use the other one which searches the optimal (highest value) path from the beginning to the end of the alignment. A path is composed of oriented (from left to right) valued ε-arcs (ε ∈ {H, b, c}). An ε-arc which starts before the ith residue must end at least after the jth (j = i + Lminε(ε) - 1) residue and its value is the sum of the logarithms of the ε-propensities spanned by that arc. To find the highest value path of such a system is an elementary problem of the graph theory (Winston, 1993).

Using the properties of the optimal path, it can be shown that the secondary structure found is globally optimal, i.e., no other path (secondary structure) respecting the constraints can have a value greater than the one found.

**Program**

The program requires two input files: a multiple alignment file (CLUSTAL output file; Thompson et al., 1994) and a file containing the set of predicted secondary structure propensities p_j^ε (i = 1, . . . , M; j = 1, . . . , N; ε = H, b, c). The user must also specify the minimal lengths Lmin_H, Lmin_b, Lmin_c for the secondary structure segments. The program is available for academics on request.

**Results and discussion**

The method has been tested on a set of 268 families of aligned sequences (Garnier et al., 1996). The first sequence in each family is the sequence ‘to be predicted’ (and whose secondary structure is known). The propensities were calculated by the GOR method (Garnier et al., 1996) using a Jack-knife. Two test runs were carried out: first with the length constraints (Equation 2) Lmin_H = 4, Lmin_b = 2, Lmin_c = 1; second with the length constraints Lmin_H = 6, Lmin_b = 4, Lmin_c = 1. The results are summarized in Tables I–III. An example of an application (using the sequences of prion proteins) is shown in Figure 1.

**Table I. Percentage of residues belonging to secondary structure elements ε that are at least Lminε residues long (Lmin_H = 4, Lmin_b = 2) in the sequence to be predicted**

<table>
<thead>
<tr>
<th></th>
<th>H</th>
<th>b</th>
<th>H&amp;b</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard GOR (wta)</td>
<td>60.5</td>
<td>79.7</td>
<td>69.4</td>
</tr>
<tr>
<td>Consensus GOR (wta)</td>
<td>61.8</td>
<td>79.7</td>
<td>69.7</td>
</tr>
</tbody>
</table>

Standard GOR (wta), standard interpretation of the propensities for a single sequence (‘winner-takes-all’). Consensus GOR (wta), standard interpretation using the alignment to obtain a consensus prediction.

Table I shows that using the standard interpretation of the propensities, even using the minimum length for an α helix (four residues) and for a β strand (two residues), there are about 40% of the helices and 20% of the β-strands that are too short and should thus be either extended or suppressed.

Table II shows various indices describing the secondary structure prediction results for the GOR IV method (here nicknamed wta for winner takes all), the GOR method that makes use of the alignment to obtain a consensus prediction (consensus GOR), the regularized prediction as described in Zimmermann (1994) (‘bio’ Champollion) and the method presented here (in unison). Table III presents the segment statistics, i.e., the number of segments and their mean length for the above methods compared with what is observed in the database (first row).

The standard GOR method has a Q_3 of 64.1% when only the sequence to be predicted is used. Using multiple alignment and a simple consensus improves Q_3 by 3.7% (Table II), but does not improve the realism of the prediction (see Table I).

On the other hand, the regularized prediction (Table II) yields secondary structures that respect the minimum lengths with a similar accuracy Q_3. Here let us emphasise again that the regularization procedure does not result, on average, in a less accurate prediction.

This clearly shows that the Q indices (Q_1 and other similar indices defined in Table II) are quite insensitive to the realism of the prediction. However, when the secondary structure prediction is regularized, Q indices measure the correlation in the segment lengths and positions and therefore no other specific ‘segment indices’ are necessary to describe the prediction quality (Rost and Sander, 1994).

The regularization improves considerably the total number of segments with the secondary structure H. In parentheses the mean length of the segments. Observed, segments observed in the data base. Other definitions are similar to those in Table I and II.

**Table II. Prediction accuracy Q (%)**

<table>
<thead>
<tr>
<th></th>
<th>Q_1</th>
<th>Q_1^b</th>
<th>Q_1^d</th>
<th>Q_2^b</th>
<th>Q_2^d</th>
<th>Q_3^b</th>
<th>Q_3^d</th>
<th>Q_3^c</th>
<th>Q_3^d</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard GOR (wta)</td>
<td>64.1</td>
<td>67.0</td>
<td>64.2</td>
<td>38.4</td>
<td>58.7</td>
<td>56.2</td>
<td>62.7</td>
<td>74.0</td>
<td>65.5</td>
</tr>
<tr>
<td>Consensus GOR (wta)</td>
<td>67.8</td>
<td>70.2</td>
<td>69.7</td>
<td>38.5</td>
<td>66.5</td>
<td>58.2</td>
<td>68.9</td>
<td>79.8</td>
<td>66.8</td>
</tr>
<tr>
<td>‘bio’ Champollion^a</td>
<td>64.3</td>
<td>67.4</td>
<td>64.6</td>
<td>38.7</td>
<td>59.1</td>
<td>56.6</td>
<td>63.1</td>
<td>73.9</td>
<td>65.4</td>
</tr>
<tr>
<td>In unison^a</td>
<td>67.8</td>
<td>70.1</td>
<td>70.1</td>
<td>38.4</td>
<td>66.7</td>
<td>58.1</td>
<td>69.2</td>
<td>79.9</td>
<td>66.5</td>
</tr>
<tr>
<td>‘bio’ Champollion^b</td>
<td>64.0</td>
<td>67.4</td>
<td>64.3</td>
<td>38.6</td>
<td>59.0</td>
<td>56.5</td>
<td>62.8</td>
<td>73.4</td>
<td>65.1</td>
</tr>
<tr>
<td>In unison^b</td>
<td>67.8</td>
<td>70.5</td>
<td>70.1</td>
<td>38.3</td>
<td>66.9</td>
<td>58.3</td>
<td>69.3</td>
<td>79.6</td>
<td>66.4</td>
</tr>
</tbody>
</table>

Standard GOR and consensus GOR, see Table I for definitions; ‘bio’ Champollion, results obtained applying constraints (Equation 2) to the first sequence only; in unison, results obtained with the method presented here.

^aConstraints Lmin_H = 4 and Lmin_b = 2 were used.
^bConstraints Lmin_H = 6 and Lmin_b = 4 were used.
Q_1, the overall prediction accuracy (percentage of correctly predicted residues in all three states H, b and c).
Q_1^b, Q_2^b, Q_3^b percentage of correctly predicted residues in ε-state with respect to the total number of residues observed in this ε-state.
Q_1^d, Q_2^d, Q_3^d, Q_3^c, Q_3^d percentage of correctly predicted residues in ε-state with respect to the total number of residues predicted in ε-state.
of helices and their mean length (results for ‘bio’ Champollion and in unison Table III(b) are very close to the observed number of H-segments and their mean length). The regularization improves also the mean length of the β-strand segments, the down side of this being an under prediction of the number of β-strand segments. As discussed in Zimmermann (1994) ‘bio’ Champollion just reinterpret the information contained in the propensities and do not create new information (this is also true for in unison). The known weakness of the GOR method consists in its difficulty to correctly predict β-strands. As can be seen in Table II—not only β-strands are underpredicted but in addition they constitute the secondary structure that is the least well predicted. In unison has to cope with this situation. It results in the increase of the mean segment length at the expense of the number of segments predicted.

The example shown in Figure 1 illustrates the typical problems of the prediction realism. The standard GOR, ‘winner takes all’ interpretation of the propensities (Figure 1b) provides a secondary structure prediction that is not realistic. The consensus GOR secondary structure (Figure 1c) gives a similar result. Individual regularization of each sequence (Figure 1d) yields realistic secondary structures but not a consensus. It is only the solution of the complete problem (Equations 1 and 2) which guarantees a realistic consensus secondary structure (Figure 1e).

As shown in Table II, on average, for the database, the percentage of correctly predicted residues is about 68%. Here the result (see Figure 1f) is below that expected, with about 62% correctly predicted residues. Notice that the regularization suppresses the two short β-strands (sequence positions [37–40], [70–73]) that are only marginally perceived by the GOR algorithm. On the other hand the regularization rightly suppresses the three-residue-long strand at position [47–49] that corresponds to coils in the observed structure. In fact the main weakness of the prediction is not so much that the two short β-strands are ill predicted but the fact that the second helix (sequence position [88–102]) is consistently and strongly predicted as a β-strand. Huang et al. (1994) using several secondary structure prediction algorithms found that all the algorithms fail to predict a helix at this location. They argued that the normal prion protein (mainly helical) is marginally stable and can, under certain circumstances, switch to a mainly β-strand protein (the scrapie isoform). In our experience, it does happen that a segment strongly predicted to be in a particular secondary structure is observed in another. It is certainly due to the fact that, although there is clearly some local information about the secondary structure adopted by the protein (otherwise it would be impossible to predict the secondary structure), the observed secondary structure is ultimately defined during the folding process. While most of the initial segments are left more or less untouched, some are forced to adopt a new secondary structure type. We thus are rather sceptical about a protein switching between two isoforms, one rich in α-helices, the other rich in β-strands. In our opinion the high β-strand content of the scrapie isoform might be better explained by a partial denaturation followed by an aggregation of the protein as it occurs with amyloidogenic proteins (Kelly, 1996).

**Conclusion**

We have presented here a method which yields the best realistic consensus secondary structure prediction when a multiple alignment is used. This method is a filter that only interprets the propensities provided by a front-end secondary structure prediction method.

This filter meets its intended goal, i.e., it does not affect the overall precision $Q_\beta$ while facilitating the interpretability of the prediction. This is the best and objective alternative to other heuristic methods that intend to ‘smooth’ predicted secondary structures.

**References**


Received November 3, 1997; revised April 21, 1998; accepted April 28, 1998