PROTEIN SECONDARY STRUCTURE PREDICTION: HOW TO IMPROVE ACCURACY BY INTEGRATION

LUIGI PALOPOLI
DEIS, Università della Calabria, Italy
e-mail: palopoli@deis.unical.it

SIMONA E. ROMBO
DIMET, Università “Mediterranea” di Reggio Calabria, Italy
e-mail: rombo@unirc.it

GIORGIO TERRACINA
Dip. di Matematica, Università della Calabria, Italy
e-mail: terracina@mat.unical.it

GIUSEPPE TRADIGO
ICAR-CNR, Rende, Italy
e-mail: gtradigo@si.deis.unical.it

PIERANGELO VELTRI *
Università “Magna Græcia” di Catanzaro, Italy
e-mail: veltri@unicz.it
ph: +39 0961 3694149
fax: +39 0961 3694112

keywords: Proteomics, Protein Structure Prediction, Data Integration

In this paper a technique to improve protein secondary structure prediction is proposed. The approach is based on the idea of combining the results of a set of prediction tools, choosing the most correct parts of each prediction. The correctness of the resulting prediction is measured referring to accuracy parameters used in several editions of CASP. Experimental evaluations validating the proposed approach are also reported.

*contact author
1. Introduction

Biological functions of proteins depend on the spatial disposition of amino acids composing them. Even if new protein amino acid sequences are continuously discovered, identifying their spatial disposition requires many efforts. Experimental, or exact, methods such as X-Ray Crystallography or Solution Nuclear Magnetic Resonance (NMR-Ray), are very expensive and time consuming. Thus, computer based automatic tools have been designed to predict protein structures, and such methods received great attention in the last few years \(^1,4,5\). Recently, many tools have been proposed and are available on-line \(^6,13\), achieving good prediction accuracies. Nevertheless, quality is still not comparable with that obtained by exact methods and research for quality prediction improvements is considered an important research topic \(^3\). Moreover, most of the existing prediction tools have high accuracy only on specific groups of proteins. Thus, a challenging problem is to devise prediction methods capable of achieving high levels of accuracy independently of the input proteins they are applied to. Recently, to improve the quality of prediction and to reduce the input dependency, methods based on a joint use of available prediction tools have been proposed \(^6,7,8\).

In this paper we focus on secondary structure prediction, presenting a novel approach based on the integration of prediction results obtained by several existing prediction tools. The idea is to select and integrate the best predictions in order to obtain higher accuracy than using a single prediction tool. Such an idea is similar to what it has been done for tertiary structures prediction \(^8\), but focusing on secondary structures.

The following example shows the basic idea of the proposed approach. Given a protein \(p\), composed by \(k\) amino acids, its secondary structure can be represented as a string with cardinality \(k\) on the alphabet of three symbols \(\Sigma = \{E, H, L\}\), meaning that the corresponding amino acid stands respectively on an \(\alpha\)-helix, a \(\beta\)-strand or a non regular conformation. Let \(T_1, \ldots, T_n\) be the predictions for a protein \(p\) obtained by using \(n\) different prediction tools. The idea is to combine (integrate) \(T_1, \ldots, T_n\) to obtain a new prediction. Figure 1(a) schematically shows the foregoing when \(n = 5\). Each prediction is represented by a bar filled using three different textures, one for each possible secondary structure configuration (black for \(\alpha\)-helix (\(H\)), striped for \(\beta\)-strand (\(E\)) and white for non regular shapes (\(L\))). The bottom part of the figure reports the real (i.e., obtained by exact methods) secondary structure of \(p\), using the same notation. Figure 1 (b) shows that, combining three out of the five predictions (namely, \(T_1\), \(T_3\) and \(T_5\), the
result is closer to the real structure than the one reported in Figure 1 (a). Note, by the way, that predictions $T_2$ and $T_4$ are less accurate than $T_1$, $T_3$ and $T_5$, if compared with the real structure. The main contribution of the paper consists in the definition of a method for the integration of different predictions; this is carried out by applying an appropriate criterion to locate and combine the “best” parts of the various predictions.

2. Parameters Definition

To measure the accuracy of a prediction, some parameters have been defined in the literature\textsuperscript{12,11}. Given a prediction tool and the amino acid sequence for a protein $p$, the three-state prediction accuracy $Q_3$ represents the percentage of secondary structure configurations (i.e., states) correctly predicted by the prediction tool. The per-segment accuracy $SOV$ measures the percentage of segments of secondary structure correctly predicted, where a segment is a contiguous set of amino acids. $Q_3$ and $SOV$ can be evaluated once the real (observed) protein secondary structure is available.

Using such parameters, we define new ones in order to evaluate the accuracy of a prediction tool w.r.t. a set of proteins. In particular, given a secondary structure prediction tool $T_i$, and a set $P$ of $m$ proteins whose observed secondary structure is known, we define the average per-segment accuracy coefficient $SOV_{(i)}$, as follows:

$$SOV_{(i)} = \frac{\sum_{j=1}^{m} SOV_{(i,j)}}{m} \times 100 \quad (1)$$

where $SOV_{(i,j)}$ indicates the value of $SOV$ corresponding to the prediction of $T_i$ for the j-th protein in $P$. $SOV$ indicates the ability of a prediction tool
to correctly predict entire sections of secondary structures. Such information is necessary to evaluate how much the “opinion” of such prediction tool is to be considered accurate, whenever a situation of disagreement among prediction tools occurs.

Given a set $T$ of $n$ prediction tools, and given a protein $p$ with $k$ amino acids, we define a consensus parameter to measure the agreement among prediction tools in $T$ while predicting the secondary structure of $p$. In particular let $T_i \in T$ be a prediction tool, and $k_j$ the $j$-th amino acid in $p$, we define the consensus percentage $C_{(i,j)}$ as follows:

$$C_{(i,j)} = \frac{N_{c_{(i,j)}}}{n} \times 100,$$

(2)

where $N_{c_{(i,j)}}$ is the number of prediction tools in $T$ that have predicted the same result as $T_i$ for the $j$-th amino acid of $p$.

The consensus percentage indicates how much a prediction tool agrees with the remaining $n-1$ ones in predicting a single amino acid state. Similarly, given a prediction tool $T_i$ and a segment $s_j$ in the predicted structure for $p$, in order to evaluate the consensus of the prediction tool $T_i$ with the remaining $n-1$ prediction tools in $T$ w.r.t. the segment $s_j$, we define the superposition mutual coefficient for segment $SOV_{\text{mutual}}$ as:

$$SOV_{\text{mutual}(i)} = \frac{\sum_{l=1, l \neq i}^{n} SOV^{T_l T_i}}{n-1},$$

(3)

where $SOV^{T_l T_i}$ is analogous to $SOV$ with the difference that $SOV$ is evaluated by considering a predicted and an observed secondary structure, whereas $SOV^{T_l T_i}$ measures the segment overlap existing between two predicted structures.

3. Integrating Prediction Results Approach

The integration approach exploits a set $T$ of $n$ prediction tools. The inputs are the amino acid sequence (primary structure) of a protein $p$ whose secondary structure is unknown, and a set $P$ of $m$ proteins whose structures are known (observed), and such that they are related to the protein $p$. More precisely, the proteins in $P$ are required to be homologous to $p$ in their structures, and in their biological functions. We notice that several tools and databases that classify proteins in families, referring to their biological functions, mutations or protein structures, are available on-line.9,10.
3.1. Assigning a Vote to Each Prediction Tool

Let $p$ be a protein with $k$ amino acids and let $T$ be a set of $n$ prediction tools that have expressed $n$ distinct predictions for the secondary structure of $p$. The proposed approach combines the $n$ predictions in order to obtain a predicted secondary structure for $p$ with a better accuracy than each single prediction. It consists in selecting, for each amino acid of $p$, the most probable state (helix, strand or irregular state) from the $n$ predictions. To integrate predictions we define a voting matrix $M(n \times k)$ where $M[i, j]$ represents a vote for the prediction of the $i$-th prediction tool w.r.t. the $j$-th amino acid of $p$. $M$ is defined by using the parameters defined in Section 2, evaluated running the prediction tools in $T$ on a set $P$ of proteins whose secondary structure is known and such that proteins in $P$ are homologous to $p$. The voting matrix $M$ is then defined as follows:

$$M[i, j] = \overline{SOV}(i) + C_{i,j} + SOV_{\text{mutual}}(i)$$ (4)

In particular, $\overline{SOV}(i)$ can be considered as a reliability score for the $i$-th prediction tool to predict the secondary structures of the proteins in $P$. $C_{i,j}$ represents a punctual agreement among the $i$-th prediction tool and the other ones in predicting the $j$-th amino acid of $p$, whereas $SOV_{\text{mutual}}(i)$ represents a structural agreement index comparing the $i$-th prediction w.r.t. the remaining ones. The following section reports the integration algorithm exploiting the voting matrix.

3.2. The Integration Algorithm

Let $p$ be a protein, $P$ a set of known proteins homologous to $p$ and $T$ a set of $n$ prediction tools. Suppose to know the $n$ secondary structure predictions for $p$ given by the prediction tools in $T$, and let $M$ be the voting matrix computed as in equation 4 w.r.t. $p$, $P$ and $T$.

```plaintext
procedure Integrate(p, T, M, s) begin
    for each amino acid $j$ of $p$ do begin
        if all the prediction tools in $T$ agree on the amino acid $j$ then
            $s[j]$ = the $j$-th symbol of the prediction of one prediction tool in $T$;
        else begin
            select $i$ such that $M[i, j]$ is the maximum of the column $j$ of $M$;
            $s[j]$ = the $j$-th symbol of the prediction of the $i$-th prediction tool;
        end else
    end for
end procedure
```
The integration algorithm obtains a secondary structure prediction \( s \) for \( p \) as follows. For each amino acid \( j \) in \( p \), the prediction of the \( i \)-th tool is chosen, where \( i \) is obtained determining the maximum value \( M[i, j] \) for the column \( j \). Finally, the prediction \( s \) is the sequence obtained by concatenating the predictions chosen for each amino acid in \( p \).

### Table 1. Comparison between accuracy measures of the prediction tool scoring the maximum values of SOV (resp. Q3) and the integration tool. \( T^S \) (resp., \( T^Q \)) indicates that the tool \( T \) scored the best SOV (resp., Q3) for that protein.

<table>
<thead>
<tr>
<th>Protein</th>
<th>Tools</th>
<th>Max SOV</th>
<th>Max Q3</th>
<th>Integration</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Tool</td>
<td>SOV</td>
<td>Q3</td>
<td>Tool</td>
</tr>
<tr>
<td>1dw</td>
<td>porter, psipred(^S), rosetta(^S)</td>
<td>83.62</td>
<td>86.21</td>
<td>83.21</td>
</tr>
<tr>
<td>1mwb</td>
<td>prof, julio, rosetta(^S), yaspin</td>
<td>91.87</td>
<td>96.00</td>
<td>91.87</td>
</tr>
<tr>
<td>1dr</td>
<td>porter, julio, psipred(^S), yaspin</td>
<td>98.77</td>
<td>91.91</td>
<td>96.95</td>
</tr>
<tr>
<td>1d78</td>
<td>psipred(^S), rosetta(^S), sam</td>
<td>68.80</td>
<td>78.45</td>
<td>67.10</td>
</tr>
<tr>
<td>1k24</td>
<td>rosetta(^S), porter</td>
<td>67.11</td>
<td>78.66</td>
<td>67.11</td>
</tr>
<tr>
<td>1lz</td>
<td>psipred, prof, porter, psipred(^S), rosetta(^S), yaspin</td>
<td>84.20</td>
<td>79.64</td>
<td>71.89</td>
</tr>
<tr>
<td>livs</td>
<td>prof, porter, psipred(^S), rosetta, sam, yaspin</td>
<td>72.94</td>
<td>77.87</td>
<td>72.94</td>
</tr>
<tr>
<td>1la</td>
<td>porter, psipred(^S), rosetta, sam(^S)</td>
<td>82.28</td>
<td>80.75</td>
<td>81.02</td>
</tr>
<tr>
<td>1k1y</td>
<td>prof, porter, psipred, sam(^S), yaspin</td>
<td>62.48</td>
<td>73.43</td>
<td>62.48</td>
</tr>
<tr>
<td>1set</td>
<td>prof, porter, sam(^S)</td>
<td>79.57</td>
<td>76.96</td>
<td>75.78</td>
</tr>
<tr>
<td>1haf</td>
<td>psipred(^S), julio, yaspin, rosetta(^S)</td>
<td>85.42</td>
<td>83.54</td>
<td>82.44</td>
</tr>
<tr>
<td>1bq0</td>
<td>porter, julio(^S), yaspin, rosetta, sam(^S)</td>
<td>56.98</td>
<td>71.84</td>
<td>52.55</td>
</tr>
<tr>
<td>1xbl</td>
<td>porter, julio(^S), rosetta</td>
<td>76.85</td>
<td>89.72</td>
<td>76.85</td>
</tr>
<tr>
<td>1hdj</td>
<td>porter(^S), yaspin, rosetta</td>
<td>92.21</td>
<td>90.91</td>
<td>92.21</td>
</tr>
<tr>
<td>1po</td>
<td>porter(^S), julio, prof, yaspin, sam(^S)</td>
<td>92.73</td>
<td>91.8</td>
<td>91.71</td>
</tr>
<tr>
<td>1mm4</td>
<td>porter, psipred(^S), sam(^S)</td>
<td>81.43</td>
<td>82.35</td>
<td>81.43</td>
</tr>
<tr>
<td>1p4t</td>
<td>porter, yaspin(^S), sam</td>
<td>81.02</td>
<td>79.35</td>
<td>81.02</td>
</tr>
<tr>
<td>1gq8</td>
<td>porter, psipred(^S), sam(^S)</td>
<td>76.08</td>
<td>76.35</td>
<td>67.15</td>
</tr>
<tr>
<td>1g50</td>
<td>prof, psipred(^S), sam(^S)</td>
<td>79.98</td>
<td>82.95</td>
<td>77.67</td>
</tr>
<tr>
<td>1bxw</td>
<td>prof, psipred(^S), sam</td>
<td>89.68</td>
<td>87.79</td>
<td>89.68</td>
</tr>
<tr>
<td>1qip</td>
<td>prof, psipred(^S), sam</td>
<td>88.08</td>
<td>85.96</td>
<td>88.08</td>
</tr>
</tbody>
</table>

### 4. Experiments

To validate our approach we tested the algorithm in Section 3.2 on proteins whose secondary structures are published in the Protein Data Bank (PDB)\(^2\), using a set of 9 available prediction tools, namely \( \text{porter, psipred, psapredict, julio, prof, rosetta, sam and, yaspin} \).

Results of experimental evaluations are reported in Table 1. Here, each
row corresponds to a test on a protein \( p \) (assumed unknown as far as the test was concerned) whose PDB identifier is reported in the first column. For each \( p \) the table shows: (i) the prediction tools considered in the integration process; in this column we use the notation \( T^S \) (resp., \( T^Q \)) to highlight the tool \( T \) scoring the best SOV (resp., \( Q^3 \)) on \( p \) among the considered ones; (ii) values of SOV and \( Q^3 \) for \( T^S \); (iii) values of SOV and \( Q^3 \) for \( T^Q \); (iv) values of SOV and \( Q^3 \) for our integration tool. Integration improves SOV parameter in 85.7% of cases, and \( Q^3 \) parameter in 66.6% of cases. Note that, usually the prediction tool scoring the maximum SOV value does not obtain the maximum \( Q^3 \), and vice versa (see, e.g., rows 1, 3 and 4). On the contrary, the integration tool scores the best accuracy for both SOV and \( Q^3 \) in many cases and it shows accuracy values that are very close to the maximum ones in the other cases. As a consequence, it is possible to conclude that our integration tool tends to improve the overall accuracy of the prediction, considering both measures.

The selection of prediction tools used for the integration process is currently semi-automatic. We are working to define an appropriate technique for the automatic selection of the best prediction tools set, to provide a fully automated tool for the prediction of protein secondary structures. Integration procedure (described in Section 3.2), and voting matrix evaluation are fully implemented in Java. Automatic querying of available prediction tools and results normalization are currently under development. All these modules are part of a more complex architecture for the automatic prediction of secondary structures whose prototyping will be soon available.

Finally we plan to face a further challenge, that is, the use of the presented tool in combination with tertiary structures prediction ones to further improve overall accuracy.

5. Related Work

In an interactive protein secondary structure prediction Internet server is presented. The server allows a single sequence or multiple alignment to be submitted, and returns predictions from six secondary structure prediction algorithms that exploit evolutionary information from multiple sequences. The main difference w.r.t. our approach is that they aim at individuating the best results among the available predictions exploiting a consensus technique, whereas our system integrates only the subset of available predictions allowing to improve prediction accuracy.

In a method based on the cooperative exploitation of different tertiary structure prediction tools is proposed. The tool is based on the selection
of models predicted by a number of independent fold recognition servers, by confidence assignment. In an approach for tertiary structures prediction is proposed. Such approach considers the characterization of the performances of a team of prediction tools jointly applied over a prediction problem, choosing the best team for a prediction problem and integrating prediction results of the tools in the team in order to obtain a unique prediction. Differently from our approach, 7,8 face the problem of protein tertiary structures prediction, thus the technique to combine predictions, the voting matrices and the reference measures of precision are completely different from our own.

References