FELLS: Fast Estimator of Latent Local Structure
Damiano Piovesan\textsuperscript{1}, Ian Walsh\textsuperscript{1,2}, Giovanni Minervini\textsuperscript{1} and Silvio C.E. Tosatto\textsuperscript{1,3,*}
\textsuperscript{1}Department of Biomedical Sciences, University of Padua, Viale G. Colombo 3, 35121 Padova, Italy. \textsuperscript{2}Bioprocessing Technology Institute, Agency for Science, Technology and Research (A*STAR), Singapore, Singapore. \textsuperscript{3}CNR Institute of Neuroscience, Viale G. Colombo 3, 35121 Padova, Italy.

Supplementary Material

FAST ESTIMATOR OF SECONDARY STRUCTURE (FESS)
The accuracy of ab initio secondary structure prediction has reached an impressive level above 80\% (Mirabello and Pollastri, 2013; Magnan and Baldi, 2014). In the last decade this accuracy seems to have plateaued and in the same period next generation sequencing has reached a state which can release vast quantities of sequences (Baker, 2010). UniProt at the time of writing contains over 65 million protein sequences. Comparing this number to the Protein Data Banks ca. 120,000 entries it is clear that experimental annotation (e.g. X-ray crystallography) and even homology based inference is not enough to understand complete organisms. Optimal prediction algorithms produce an output for an average sequence of 300 amino acids in approximately five minutes. These slow and accurate methods are perfect for analyzing single proteins of interest but to understand large proteomes on a single standard CPU takes months. The consequence of this is large scale predictions become inaccessible to the average scientist.

Evolutionary information in the form of multiple sequence alignments (MSA) improve the prediction of secondary structure significantly (Jones, 1999). This is well established for a long time and state-of-the-art predictors incorporate them in different ways (Rost and Sander, 1993). MSAs are not only important for secondary structure but function (Piovesan et al., 2015), residue-residue contacts (Piovesan et al., 2016), intrinsic disorder (Walsh et al., 2011) and basically any conserved property of the protein. However, their calculation is the main bottleneck for run-time and in addition installing third party alignment software is far from user friendly.

Here, we investigate this MSA bottleneck and argue that as sequence databases grow, the accuracy improves (Przybylski and Rost, 2002) but the run-time becomes extremely slow. A simple machine learning algorithm is described which is optimized for no alignment input. As a result the software allows genomic scale predictions on basic CPU and RAM resources. The accuracy, as expected, degraded but to an extent that is acceptable to make proteomic scale observations.

FESS predicts secondary structure from the primary sequence only. It exploits the Atchley amino acid propensities (Atchley et al., 2005) which can be calculated very rapidly and without the need of external tools. The FESS implementation is similar to Porter 4.0 (Mirabello and Pollastri, 2013) which is based on ensembles of BRNNs (Baldi et al., 1999) that work in cascade as described in (Pollastri and McLysaght, 2005). Also the training set is the same as Porter 4.0, it includes 7,522 high quality protein structures with less than 25\% sequence identity. The main difference with Porter is that the FESS parameters are optimized for no alignment input. In Figure 1 the Q3 (% correct helix, strand and coil residues) is reported for FESS and it is compared with Porter 4.0 and Psipred (Jones, 1999) as a function of computational time. Accuracy correlates with the quality of MSA profiles that in turn depends on the size of the searching database (ca. 500K sequences in SwissProt\textsuperscript{90} and 42 million in UniRef\textsuperscript{90} - all sequence pairs share no more than 90\% sequence identity). When predictions are generated without alignments FESS overall accuracy is ca. 2\% higher than Porter and ca. 4\% higher than Psipred (Supplementary Figure S1). The three-state accuracy (Q3) and the segment overlap score (SOV) (Zemla et al., 1999) for the three SS states are reported in Supplementary Table S1. FESS outperforms the fast version (no alignments) of Porter 4.0 and Psipred for helix and strand predictions.

SEQUENCE AMPHIPATHICITY
For standard alpha helices and considering a window of 19 residues, the central amino acid (position 10) is in the same surface as residues 3, 6, 7, 13, 14 and 16, whereas residues 1, 4, 5, 8, 12, 15, 16 and 19 lie on the other side. For beta sheet the amphipathicity is calculated over a window of 11 residues with even positions lying on the same side of the central residue and odd position in the opposite surface. The propensity of opposite residues is inverted, thus contributing positively to the global score of the window.
Supplementary Figure S1. Three-states accuracy (Q3) versus time plot. Time estimated on one CPU for 1,000 Homo sapiens proteins. Notice the logarithmic time x-axis scale and the cut y-axis range. ‘Psipred Fast’ and ‘Porter Fast’ use no alignments. ‘Psipred fast’ was optimized for no alignments but ‘Porter Fast’ was not. Alignments calculated using 3 rounds of PsiBlast.

Supplementary Table S1. Performances of fast methods for each secondary structure state.

<table>
<thead>
<tr>
<th>Metric</th>
<th>Method</th>
<th>All</th>
<th>Helix</th>
<th>Strand</th>
<th>Coil</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q3</td>
<td>FESS</td>
<td>72.06</td>
<td>69.92</td>
<td>59.76</td>
<td>72.44</td>
</tr>
<tr>
<td></td>
<td>Porter 4.0 f</td>
<td>70.27</td>
<td>66.79</td>
<td>57.15</td>
<td>73.75</td>
</tr>
<tr>
<td></td>
<td>Psipred fast</td>
<td>68.15</td>
<td>61.45</td>
<td>56.61</td>
<td>77.90</td>
</tr>
<tr>
<td>SOV</td>
<td>FESS</td>
<td>68.87</td>
<td>67.49</td>
<td>63.37</td>
<td>66.74</td>
</tr>
<tr>
<td></td>
<td>Porter 4.0 f</td>
<td>66.10</td>
<td>64.49</td>
<td>60.49</td>
<td>64.94</td>
</tr>
<tr>
<td></td>
<td>Psipred fast</td>
<td>63.97</td>
<td>62.40</td>
<td>60.78</td>
<td>63.80</td>
</tr>
</tbody>
</table>

Performances are calculated by averaging over single protein prediction scores. The best performance in each category is highlighted in bold.

REFERENCES