Progress of 1D Protein Structure Prediction at Last

Burkhart Rast and Chris Sanders
EMBL, 69012 Heidelberg, Germany

ABSTRACT Accuracy of predicting protein secondary structure and solvent accessibility from sequence information has been improved significantly by using information contained in multiple sequence alignments as input to a neural network system. For the Astrolomar meeting, predictions for 13 proteins were generated automatically using the publicly available prediction method PHD. The results confirm the estimate of 75% three-state prediction accuracy. The fairly accurate predictions of secondary structure segments made the tool useful as a starting point for modelling of higher dimensional aspects of protein structure.

© 1995 Wiley-Liss, Inc.

Keywords: automatic prediction of protein secondary structure and solvent accessibility, neural networks

SPREADING OPTIMISM BY PUBLISHING HIGH SCORES

Protein secondary structure has been predicted from sequence before the first three-dimensional (3D) structures were determined by crystallography. Two decades and dozens of methods later, the accuracy of secondary structure prediction was still not better than 60–65% (percentage of residues predicted correctly in either of the three states helix, strand, coil). In the hunt for higher scores, a growing data base of known structures and refined methods pushed the accuracy to above 60–65%.

Occasionally, even higher values were reported, but tests on representative data sets revealed that prediction accuracy was about 50% by 1992. Overoptimistic claims by predictors nourished skepticism of potential users. One major point about prediction methods became clear at the Astrolomar prediction contest: exaggerated claims are more damaging than genuine errors. Even a prediction method of limited accuracy can be useful if the user knows what to expect. For the editors of scientific journals this implies that no prediction method should be published that has not been sufficiently cross-validated.

HOW TO IMPROVE PREDICTION METHODS

A proper evaluation of prediction methods, in our view, needs to meet four requirements. (1) No significant pairwise sequence identity; the proteins used for setting up a method (training set) and those used for evaluating it should have a pairwise sequence identity of less than 25% (length dependent cut-off), otherwise homology modelling could be applied which would be much more accurate than ab initio predictions. (2) Sufficiently large data sets: all available unique proteins should be used for testing (currently more than 400), evaluations based on too small numbers are not representative. (3) Avoid comparing apples with oranges: no matter which data sets are used for a particular evaluation, results should always be reported additionally on standard sets. (4) No optimization with respect to test set: a seemingly trivial—and often violated—rule is that methods should never be optimized with respect to the data set chosen for final evaluation. For example, most methods are evaluated in n-fold cross-validation experiments (splitting the data set into a different training and test set). The exact number of n is not important provided the test set is representative and the cross-validation results are NOT misused to again change parameters.

HOW TO IMPROVE PREDICTIONS OF SECONDARY STRUCTURE AND SOLVENT ACCESSIBILITY

One key to more accurate predictions of 1D structures, such as secondary structure or solvent accessibility, has been the use of evolutionary information. This idea has long been in the literature. Most often, experts base single-case predictions on multiple alignments. The reason for the relevance of evolutionary information is that profiles of residue exchanges in naturally evolved protein families are highly specific for details of a particular protein.

Abbreviations: 3D, three-dimensional; 1D, one-dimensional; DSSP, data base containing the secondary structure and solvent accessibility derived from experimentally determined coordinates of proteins of known 3D structure; PHD, problem-based neural network prediction of secondary structure (PHDserver) and solvent accessibility (PHDserver).

© 1995 WILEY-LISS, INC.
TABLE 1. Accuracy for Blind Predictions, 0.4

<table>
<thead>
<tr>
<th>Proteins</th>
<th>N_residues</th>
<th>Overall accuracy</th>
<th>Segmented accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>chna</td>
<td>168</td>
<td>84.1</td>
<td>0.54</td>
</tr>
<tr>
<td>hau</td>
<td>769</td>
<td>71.7</td>
<td>0.30</td>
</tr>
<tr>
<td>ll4</td>
<td>122</td>
<td>90.6</td>
<td>0.4</td>
</tr>
<tr>
<td>pppk</td>
<td>865</td>
<td>70.2</td>
<td>0.24</td>
</tr>
<tr>
<td>proba</td>
<td>72</td>
<td>95.6</td>
<td>0.22</td>
</tr>
<tr>
<td>rtf</td>
<td>122</td>
<td>95.6</td>
<td>0.22</td>
</tr>
<tr>
<td>smpg</td>
<td>154</td>
<td>95.6</td>
<td>0.22</td>
</tr>
<tr>
<td>standard</td>
<td>467</td>
<td>70.2</td>
<td>0.1</td>
</tr>
<tr>
<td>xyla</td>
<td>946</td>
<td>75.7</td>
<td>0.26</td>
</tr>
<tr>
<td>all</td>
<td>2888</td>
<td>71.6</td>
<td>0.28</td>
</tr>
<tr>
<td>ref</td>
<td>9542</td>
<td>72.1</td>
<td>0.27</td>
</tr>
<tr>
<td>test</td>
<td>784</td>
<td>72.1</td>
<td>0.27</td>
</tr>
</tbody>
</table>

*Abbreviations: N_residues, overall three-state per-residue accuracy, i.e., the percentage of residues predicted correctly in either of the three states helix, strand, coil; I, information content of prediction matrix; the entropy-based measure of per-residue accuracy is less intuitive than the three-state percentage but captures well in a single value the balance between correct, false, and under-predicted, as given by a 3 × 3 matrix A, where A_{ij} gives the number of residues predicted in state i and observed in state j. Bold, percentage of residues predicted in helix and observed in strand, or predicted in strand and observed in helix;boro, overlap between observed and predicted secondary structure segments: this fuzzy score yields 100% even when the two segments observed and predicted differ slightly in length, or are slightly displaced (it is found to be the best among several alternatives of defining per-segment accuracy); Sseg, overlap between observed and predicted helical segments; Sseg, overlap between observed and predicted strand segments; helix, strand; rest, 100% two-state per-residue accuracy, i.e., percentage of residues predicted correctly in either of the two states helix or strand (exposed < 16% solvent accessible). N_residues is the number of residues.

*Abbreviations for prediction methods: MaxHom, P-loop motif, profile-based prediction; MaxE, P-loop motif, profile-based prediction; correlation coefficient between observed and predicted relative solvent accessibility; N_residues is the number of residues.

*Proteins predicted but not evaluated: scil, scil, and scil are absent from a set of 250 unique proteins used to evaluate PPhred; ref, 9542 unique proteins used to evaluate PPhred; test, 784 proteins used to evaluate PPhred.

Structure. One tool to capture the richness of such information is a neural network. For one neural network method the following levels of performance accuracy have been achieved: secondary structure prediction, per-residue accuracy of 72 ± 0.4% 0.4; solute accessibility prediction: correlation between observed and predicted accessibility of 0.54 ± 0.12 (Table 1). Other programmable methods have been shown to benefit from evolutionary information to the same extent. 0.23,0.32

**WOMAN WITH MACHINE**

Predictions of 1D structure by a system of profile-based neural networks (PHD) are publicly available (for information, send the Web help at the internet address: PredictProtein EMBL-Heidelberg DE, or see the World Wide Web site at http://www.emb. heidelberg.de/PredictProteinP/PredictProtein.html). In the Asilomar context, we submitted blind predictions for 13 proteins (Table 1, for four proteins no results are given as the experimental coordinates were not available). The accuracy of PHD predictions depends partly on parameters that can be influenced by the user, such as the sequences taken for multiple alignment, or the details of the alignment. Furthermore, experts can "filter" predictions according to additional knowledge. Such fine-tuning can improve prediction accuracy in a few cases, markedly less so than by comparing the results of PHD based on MaxHom alignments and those aligned by Tom Hubbard, as Tom Dally and Fred Cohen, this issue. For the Asilomar review, we deliberately contributed "unfiltered" automatic predictions to illustrate both the accuracy reachable in laboratories without experts and the starting point for experts to possibly improve predictions. Of course, there is no direct competition between woman and machine; intelligent experts use accurate automated methods and attempt to refine the prediction by their expertise.

**WHAT WENT RIGHT?**

Prediction accuracy within expected range. First, the secondary structure prediction accuracy using PHD on the Asilomar proteins was within the range of what had been published as expected accuracy (Table 1). The correlation between observed and predicted relative solvent accessibility was on the low side of the expected range. The deviation from expected values of accuracy indicates that a set of nine proteins is too small to derive generally valid estimates (sets of 20 or even 60 proteins—evaluation,
set published a decade ago—have been shown to be too small to derive valid estimates for prediction accuracy. For example, helices were predicted better than expected strands on the other hand were predicted worse than expected (Table 1).

Reliability of prediction correlates with accuracy. Prediction accuracy varies between different proteins. Fortunately, the reliability of PHD predictions enables one to estimate on which side of such a distribution the prediction for a given protein is to be expected. For example, the 50% of residues predicted with highest reliability are predicted at an average accuracy of over 90% (Fig. 1), and the most reliable one-third of the strands and helices were predicted at segment-based accuracy values of >60% (Fig. 1).

Using 2D predictions to guide modeling of 3D structure. For xylanase the relative solvent accessibility was predicted more accurately than the expected average (Table 1). Liisa Helm used this prediction of accessibility and the corresponding prediction of secondary structure to infer a (correct) ab initio model of the 3D structure. However, the topology of the model corrected by one of us (C.S.) was incorrect in part. For some proteins in the context Tim Hubbard used PHD predictions as a baseline for correct predictions of topology (Hubbard and Park and Defay and Cohen, this issue).

WHAT WENT WRONG?

Helices predicted as too long. Helices were predicted at a higher than average length (13.5 residues instead of 10). This may partly be explained by the unusual content of secondary structure in the nine proteins evaluated here: helix, 52% vs. about 32% in a set of representative proteins $^{22,23}$, strand 18 vs. 21%. These deviations basically indicate that a data set of nine proteins is too small to be representative.
prosub: propeptide of subtilisin BPN', 71 residues, Q = 55%

RTP: replication termination protein, 116 residues, Q = 56%

strauf3: domain 3 of straufen, 57 residues, Q = 46%

Fig. 2. Three examples of errors in secondary structure prediction. Abbreviations used: AA, amino acid; I, in one-letter code. The secondary structure assignment is based on 3D structure by DSSP.113 PHD prediction by neural network system; R, reliability of prediction; A, low, B, high, 2 as high. The names reflect the strength of the prediction, i.e., the difference between the output units (note, the P-O networks have three output units coding for helix, sheet, and coil) with the highest value (winner unit) and the output unit with the next highest value. Symbols for secondary structure assignments: H, helix; E, extended; P, pleated sheet; T, turn.

The prediction strength depends on the diversity of the sequence family aligned. For some of the predictions, the alignments did not contain many sequences, e.g., only two sequences were aligned to the replication termination protein (RTP). This resulted in unusually high levels of prediction strength (reliability index). In general, the correctness and diversity of the multiple alignment used for prediction are crucial factors influencing prediction accuracy.

Unpredictable accuracy for nonconserved proteins. The neural networks used for 3D predictions were trained on globular water-soluble proteins; predictions tend to be wrong for other proteins.5 The replication termination protein illustrated this point as an interaction between dimers is crucial.24

Free text
the secondary structure segments were correctly predicted, but the segment is a strand rather than a helix (two strands predicted as helix in R7P and the strand at position 24-25 of preseq in Fig. 2). How can a segment be placed correctly if the type is confused? Strata, some 9-17 adjacent residues (input to neural networks) have preferences for forming regular arrays of backbone hydrogen bonds. However, some short sequence motifs can occur in both strands and helices. A region may have a higher preference for forming a helix than a strand, but interactions nonlocal in sequence may result in that the formation of a 3-sheet is energetically more favorable. Thus, the confusion between helices and strands indicates that the prediction is focused on preferences for formation of regular secondary structure, rather than on preferences for forming certain secondary structure types. A hypothesis that would have to be considered is that these segments are confused which are formed due to nonlocal interactions.

WHAT DID WE LEARN?

First, prediction of 3D protein structure is now sufficiently accurate to be useful as a starting point for, e.g., threading techniques (single example: Tim Hubbard, Craig Livingston et al., Geoffrey Baron, Stephen Brenner and Halldin Gerloff in this issue; for an automated threading method). Second, in some cases accuracy is sufficient to base correct modeling of 3D structure on 2D predictions. Third, even false predictions can contribute useful information. For example, for the least accurately predicted replication termination protein (Fig. 2) the 3D structure for one region was modeled accurately by mean-force potential-based threading (Manfred Stipl, this issue). However, the threading-based modeling excluded the N-terminal helix. This helix was predicted correctly by PHD. Thus, reliable segment predictions are useful as a starting point for 3D modeling, and PHD constitutes a tool for focusing on the most reliably predicted segments.

ACKNOWLEDGMENTS

We should like to express our gratitude to John Moult (CABF, Washington, D.C.) and to his colleagues for having organized an extremely interesting meeting that emphasized the importance of objective evaluation in a research field of increasing importance. Thanks to Tim Hubbard (MRC, Cambridge, U.K.) for encouragement that paved the way to California. Furthermore, thanks to colleagues from EMBL, Heidelberg; Carl Fried, for valuable ideas in developing PHD; Reinhard Schneider, for Maxioform and important help and discussions; Antoine de Daruvar, for managing the painful struggle with wrong formats and CPU crashes when running the PredictProtein server; and all other members of the Protein Design Group for many discussions.

Thanks also to the referees for very helpful comments. Last, but not least, thanks to all those who contributed protein structure coordinates from NMR and X-ray crystallography.

REFERENCES


