

# Simple jury predicts protein secondary structure best

Burkhard Rost <sup>1,\*</sup>, Pierre Baldi <sup>2</sup>, Geoff Barton <sup>3</sup>, James Cuff <sup>4</sup>, Volker Eyrich <sup>5,1</sup>,  
David Jones <sup>6</sup>, Kevin Karplus <sup>7</sup>, Ross King <sup>8</sup>, Gianluca Pollastri <sup>2</sup>, Dariusz  
Przybylski <sup>1</sup>

## Abstract

The field of secondary structure prediction methods has advanced again. The best methods now reach levels of 74-76% of the residues correctly predicted in one of the three states helix, strand, or other. In context of the EVA/CASP, we experimented with averaging over the best current methods. The resulting jury decision proved significantly more accurate than the best method. Although the 'jury' seemed the best choice on average, for 60% of all proteins one method was better than the jury. Furthermore, the best individual methods tended to be superior to the jury in estimating the reliability of a prediction. Hence, averaging over predictions may be the method of choice for a quick scan of large data set, while experts may profit from studying the respective method in detail.

## INTRODUCTION

*Secondary structure prediction improved again.* Accuracy increased substantially in the 90's through using evolutionary information taken from the divergence of proteins in the same structural family. Recently, the evolutionary information resulting from improved searches and larger databases has again boosted prediction accuracy by more than four percentage points to its current height around 76% of all residues predicted

correctly in one of the three states helix, strand, other. The best current methods solved most of the problems raised at earlier CASP meetings: All good methods now get segments right and perform well on strands. Could we improve prediction accuracy further by averaging over the best prediction methods?

## METHODS

**JPred2** <sup>1</sup>: Jpred2 provides multiple sequence alignments constructed automatically by both PSIBLAST searches and ClustalW alignments derived from a carefully filtered non-redundant sequence database. The default in JPred2 is to execute only the JNet <sup>1</sup> algorithm. For the results in this paper (Table 1), we used JPred2 with the option to combine JNet with PHD <sup>2,3</sup>, NNSSP <sup>4</sup>, Predator <sup>5</sup>, Mulpred (Geoff Barton, unpublished), DSC <sup>6</sup> and Zpred <sup>7</sup>. We estimated that the combination of JNet with these methods improved prediction accuracy slightly.

**PHDpsi** <sup>8</sup>: PHD <sup>9</sup> is a system of neural networks using evolutionary information as input. It was developed in 1994. The only difference between PHD and PHDpsi is that the latter uses divergent profiles as provided by PSI-BLAST <sup>10</sup>.

**Prof\_king** <sup>11, 12</sup>: The input to Prof\_king is a multiple sequence alignment produced using PSI-BLAST <sup>10</sup> and ClustalW <sup>13</sup>. The alignment is "poly-transformed" into attributes using the approaches of: GOR propensities <sup>14</sup>, PHD profiles <sup>2</sup>, and PSI-PRED profiles <sup>15</sup>. The machine learning in Prof\_king is a complicated combination of linear and quadratic discrimination, back-propagation neural networks, the use of different priors, and cascaded

---

### Affiliations

1 CUBIC, Columbia Univ., Dept. of Biochemistry and Molecular Biophysics, 650 West 168<sup>th</sup> Street, New York, NY 10032, USA

2 Univ. of California, Irvine, Dept. of Information and Computer Science, Institute of Genomics and Bioinformatics, Irvine, CA-92697, USA

3 European Bioinformatics Institute, Genome Campus, Hinxton, Cambs CB10 1SD, England

4 The Sanger Centre, Wellcome Trust Genome Campus, Hinxton, Cambs CB10 1SA, England

5 Columbia Univ., Dept. of Chemistry, 3000 Broadway MC 3167, New York, NY 10027, USA

6 Dept. of Biological Sciences, Brunel Univ., 274348, Uxbridge, Middlesex UB8 3PH, England

7 Computer Engineering, Univ. of California, Santa Cruz, Santa Cruz, CA 95064, USA

8 The Univ. of Wales Aberystwyth, Dept. of Computer Science, Penglais, Ceredigion, SY23 3DB, Wales, UK

\* Corresponding author: rost@columbia.edu, <http://cubic.bioc.columbia.edu/>

prediction. Prof\_king was designed to predict especially beta-strands well, as their prediction is essential for fold recognition. The method reaches its high level of accuracy since the use of an ensemble of predictors improves prediction because it is closer to the "Bayes-optimal" predictor.

**PROFsec**<sup>16</sup>: This program bases upon the principle concept implemented in PHD<sup>2,3,9</sup>; it extends this concept in several ways. (1) The network system was trained on a large data set of divergent PSI-BLAST profiles. (2) PROF also uses more information for the input than PHD. (3) Finally, PROF uses a third layer of networks.

**PSIPRED**<sup>15, 17</sup>: The current version of PSIPRED (Version 2) implemented on the respective Web server and also available in source code form at the address <http://www.psipred.net> is based on the method previously published<sup>15, 17</sup>. This method, which is similar in concept to the approach previously described by Rost and Sander<sup>2</sup> employs two standard feed-forward neural networks to analyse the Position Specific Scoring Matrices (PSSMs) generated by PSI-BLAST<sup>10</sup> in order to predict protein secondary structure. Three iterations of PSI-BLAST are used, searching a large non-redundant data bank of protein sequences from which coiled-coil, transmembrane and low-complexity regions have been masked out. The first network employs 75 hidden units and 15x21 (window of 15 residue positions), and is trained using an early-stopped training strategy using 10% of the training data to detect convergence. The second level network, which is used to "filter" the predictions from the first network, comprises 15x4 inputs and 55 hidden units. Version 2 of PSIPRED uses 4 identical first level networks, each trained with a differently chosen training set, and it is the averages of these 4 sets of outputs which are now presented to the second level network. This averaging has been found to improve the prediction accuracy on average by 1.3 percentage points.

**SAM-T99sec**<sup>18</sup>: The secondary structure predictor on the SAM-T99 web site averages three different methods. All three methods use a multiple alignment of probable homologues found by the SAM-T99 method for iterative search and alignment. The methods were trained on STRIDE classification of secondary structure, not DSSP. Although STRIDE and DSSP differ in 3-state assignment for about 5% of residues, prediction accuracy is only changed by about 0.5%, as the places where they differ are hard to predict correctly anyway. The neural net method is the most accurate of the three methods, and is the only one that provides accurate estimation of probability of correctness for the prediction. The net has four layers of neurons (three hidden layers and the output layer). The input has probability vectors for the 20 amino acids (computed from the multiple alignment using

sequence weighting and a Dirichlet mixture prior) and probabilities for insertion and deletion at each position. The three hidden layers have 9, 10, and 9 units respectively, and the output layer has 3 units (corresponding to E, H and L). The inputs to each layer are a window centered on the current residue of the outputs of the previous layer. The window sizes are 5, 7, 5, and 11. Each layer uses soft-max to produce a probability vector as its output. The second method uses a general hidden Markov model which emits pairs of amino acids and secondary structure codes in each state. There are 37 "fat states" in the HMM and 162 "fat edges" (each fat state has a pair-emitting state and a deletion state, and each fat edge contains an insertion state and 9 edges). The target sequence is aligned to the HMM and the posterior probability of each label is computed by a weighted sum over all alignments. The homologues are likewise aligned, and the final prediction is the weighted average of the predictions for the individual sequences. The third method uses the target HMM used by SAM-T99 for fold recognition (one match state per residue in the target sequence), and aligns a subset of PDB sequences to it. The secondary structure codes for the PDB sequences are given weights according to how likely the given residue is to align to that position in the target HMM.

**SSpro**<sup>19</sup>: SSpro consists of 11 bi-directional recurrent neural networks (BRNNs) derived from the general theory of graphical models<sup>19, 20</sup>. BRNNs are adaptive graphical models that can learn how to translate input sequences into output sequences with variable lengths. They extend typical Markov or hidden Markov model approaches by including both a left-right and a *right-left* chain of hidden states that can carry information in both directions along the chain, and between the input and output sequences. The left-right and right-left state updates are parameterised by two recurrent networks, which can be viewed as two "wheels" that are rolled along the protein chain, starting from each terminus. A third neural networks produces the secondary structure classification output using the local left and right contexts carried by the wheels, the input, and the hidden units. BRNNs can be trained in supervised fashion using a generalised form of gradient descent. The ensemble prediction is obtained via plain average of the outputs of the single models. From our tests BRNNs are responsible for gains in the 0.5-1% range over structured feed-forward neural networks, and are capable of capturing context information over distances of about 30 amino acids. The results of the ensemble represent gains of 1.3-3.3% over single models. The profile weighting scheme provides a gain of about 0.6% with respect to plain profiles. Further improvements have recently been

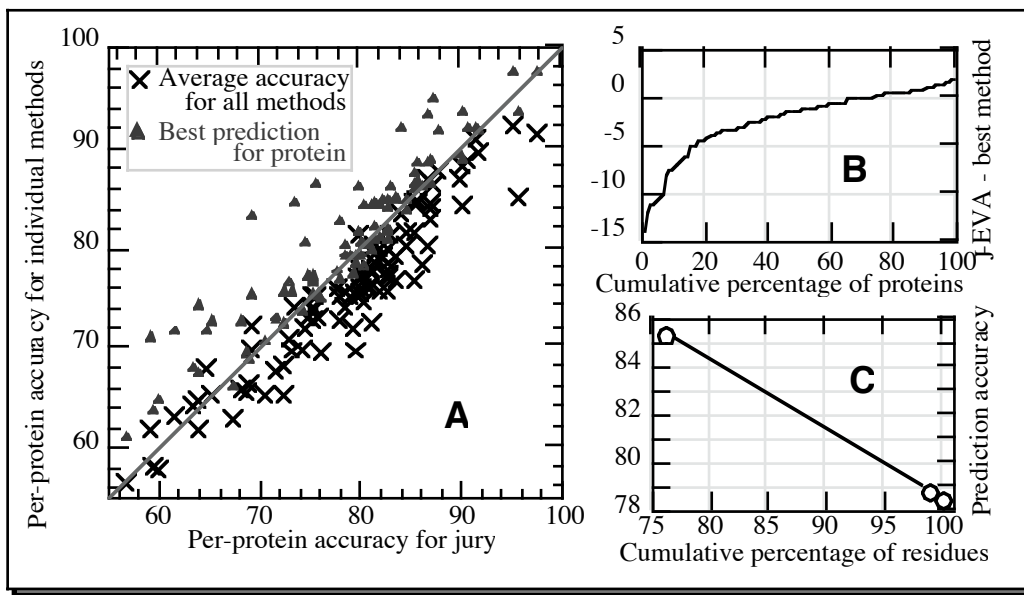
obtained using PSI-BLAST profiles<sup>10</sup> and a new improved version (SSpro 2.0) will soon be on-line. The SSpro 1.0 server and related references are currently available at <http://promoter.ics.uci.edu/BRNN-PRED/>. *J-EVA: alternatives of compiling the jury decision*. The automatic evaluation server EVA<sup>13</sup> received secondary structure prediction in CASP format. In particular, the predictions did not contain detailed values for the propensities for helix (H), strand (E), other (L). Rather, the outputs were restricted to the actual state predicted (H or E or L) and - for all methods except for SSpro - an estimate of the reliability of the prediction for each residue. We tried three alternative ways to combine all prediction methods. (1) Winner average (dubbed 'J-EVA winner'): count how many methods predicted state S and predict state with highest count. (2) Simple average (dubbed 'J-EVA simple'): weight predictions by provided reliability index. (3) Weighted average (dubbed 'J-EVA weighted'): this average normalised the weighted average by the method and protein averages:

$$out_s^i = \frac{1}{\sum_{\mu} \langle ri \rangle_{\mu}^i} \sum_{\mu} \frac{\langle ri \rangle_{\mu}^i}{\langle ri \rangle_{\mu, P}^i} \cdot out_s^i \cdot ri_s^i \quad (2)$$

where  $out_s^i = 1$ , if method  $\mu$  predicted residue in state  $s$ , and 0 otherwise;  $ri_s^i$  was the reliability (0-1) of method  $\mu$  for residue  $i$ ;  $\langle ri \rangle_{\mu}^i$  was the per-protein average of the reliability index for method  $\mu$ , and  $\langle ri \rangle_{\mu, P}^i$  the average over the reliability indices of method  $\mu$  for protein  $P$ .

## RESULTS

*Jury over best three methods significantly more accurate than best method.* All alternative strategies to compile the jury decision were successful in that the resulting predictions were on average not worse than those of the best individual method. The weighted average and the 'winner-take-all' decision compiled over the three methods PROFsec, PSIPRED, and SSpro predicted more than 78% of all residues correctly (Table 1). For the data set of 218 proteins used for evaluation this difference was significant. Interestingly, the improvement was also visible in terms of per-segment scores (SOV), and in that the number of residues



**Fig. 1. Jury prediction vs. best prediction.** For each of 98 sequence-unique proteins we averaged over the three-state per-residue accuracy of all 7 methods described in Methods. Then we compiled the jury prediction according to eq. 2 over all 7 methods. **A** relates the accuracy of the jury to the average accuracy (crosses) and to the most accurate prediction for each protein

(triangles). The jury prediction was more accurate than the average for almost all proteins. How often was the best prediction more accurate than the jury? **B** shows the cumulative difference between jury and best prediction (negative values imply that the best prediction was better than the jury). For more than 60% of all proteins, the jury decision was less accurate than the best prediction.

**Table 1: Accuracy for 218 identical proteins**<sup>A</sup>

Method	Q <sub>3</sub>	SOV <sub>3</sub>	BAD	C <sub>H</sub>	C <sub>E</sub>
J-EVA3 weight	<b>78.0</b>	<b>73.7</b>	<b>2.0</b>	<b>0.69</b>	<b>0.65</b>
J-EVA3 winner	<b>78.2</b>	<b>73.8</b>	<b>1.7</b>	<b>0.70</b>	<b>0.66</b>
J-EVA3 simple	77.2	72.8	2.0	0.69	0.65
PROFsec	76.8	72.8	2.2	0.66	0.65
PSIPRED	76.4	72.1	2.4	0.66	0.63
SSpro	76.1	71.2	2.5	0.67	0.64
JPred2	74.8	69.3	2.4	0.64	0.61
PHDpsi	74.7	69.7	2.9	0.64	0.61

confused between helix and strand was reduced. Furthermore, the simple averaging procedure was clearly inferior to the 'winner-take-all' and to the 'weighted average'. When compiling the jury decision over all five methods for which we had results on 218 proteins, prediction accuracy decreased slightly. Hence, averaging over the best three was better than averaging over the best five.

*Best prediction was most often better than jury.* Prediction accuracy varies strongly between proteins. Furthermore, proteins predicted at low levels of accuracy occasionally are predicted more accurately by some method<sup>14</sup>. How did the jury predictions compare to the 'field' and how to the best prediction? The jury prediction was most often more accurate than the average over all methods (Fig. 1A). However, for less than 40% of the proteins the jury prediction was better than the best prediction (Fig. 1B). Did this imply that our way of compiling the jury decision was not optimal? Obviously, the average over the best predictions on 218 proteins reached 80.8%, the average over the jury reached 78.2% (Table 1). Given that the best single method on that set was 76.8, the jury achieved only 35% (78.2-76.8 / 80.8-76.8) of the improvement we could have reached if we should have combined methods optimally.

*Individual methods defined reliability more accurately.* All 7 methods accurately estimated the reliability of the prediction for each residue<sup>14</sup>. The prediction accuracy for the jury prediction correlated less impressively with accuracy (Fig. 1C). In particular, the jury predictions

<sup>A</sup>: **Data set and sorting:** All methods have been tested on the same set of 218 different new protein structures (EVA version Feb 2001). None of these proteins was similar to any protein used to develop the respective method. Sorting and grouping reflects the following concept: if the data set is too small to distinguish between two methods, these two are grouped. For the given set of 218 proteins this yielded three groups. Inside of each group, results are sorted alphabetically. Differences < 0.6 percentage points were not significant.

<sup>B</sup>: **Method:** see abbreviations in Methods; **Scores**<sup>15</sup>: Q<sub>3</sub>: three-state per-residue accuracy, i.e., number of residues predicted correctly in either of the three states helix, strand, other; SOV<sub>3</sub>: three-state per-segment score measuring the overlap between predicted and observed segments<sup>16, 17</sup>; BAD: percentage of helical residues predicted as strand, and of strand residues predicted as helix<sup>18</sup>; C<sub>H</sub>: Matthew's correlation coefficient for state helix<sup>19</sup>; C<sub>E</sub>: Matthew's correlation for state strand<sup>19</sup>.

never reached levels above 85% accuracy for any residue. Hence, experts may still profit substantially from studying the explicit output from good methods.

*Best three methods might suffice to reach peak jury performance.* Suppose we have NB best methods, and NM methods of medium accuracy. Is it better to compile averages over only the NB best ones, or over all NB+NM methods? Contrary to the naïve expectation, including the NM methods improves as long as these differ from the NB methods (Anders Krogh, private communication). In our experiment, we did not see a significant difference between compiling the jury over the best three and over all 7 best methods. For a set of 98 proteins for which we had results for all methods, 'winner' jury over 7 methods reached 78.3%, that over 3 methods reached 78.1%. The data in Table 1 suggested that the best strategy may be the most simple one, i.e. average over the final prediction of each method discarding the individual prediction accuracy. However, this result did hide the following problem: we did not have reliability indices for SSpro. Consequently, weighting the jury according to eq. 2 was problematic. This problem was less important when compiling the jury over 7 methods. The resulting weighted jury reached 78.5% accuracy compared to 78.1% for the simple 'winner' jury. However, the set of 98 proteins was too small to conclude that the two strategies yielded significant differences.

## CONCLUSIONS

Is it a good idea to combine the best method for secondary structure prediction? We found that a simple jury prediction using the best three methods improved accuracy on average (Table 1). However, for most proteins, at least one of the seven best methods was more accurate than the jury. Furthermore, the best individual methods were more successful in identifying reliably predicted regions than was the jury prediction. Hence, is it a good idea to consider regions as more reliably predicted if all methods agree? Supposedly not. If we know the structure of a protein, we can simply always select the best prediction. This would yield an accuracy above 80%. Can we reach this level without knowing the structure? Not yet.

## REFERENCES

1. Cuff, J. A. and Barton, G. J. Application of multiple sequence alignment profiles to improve protein secondary structure prediction. *Proteins* 40:502-511, 2000.
2. Rost, B. and Sander, C. Prediction of protein secondary structure at better than 70% accuracy. *J. Mol. Biol.* 232:584-599, 1993.
3. Rost, B. and Sander, C. Combining evolutionary information and neural networks to predict protein secondary structure. *Proteins* 19:55-72, 1994.
4. Salamov, A. A. and Solovyev, V. V. Prediction of protein secondary structure by combining nearest-neighbor algorithms and multiple sequence alignment. *J. Mol. Biol.* 247:11-15, 1995.
5. Frishman, D. and Argos, P. Incorporation of non-local interactions in protein secondary structure prediction from the amino acid sequence. *Prot. Engin.* 9:133-142, 1996.
6. King, R. D. and Sternberg, M. J. Identification and application of the concepts important for accurate and reliable protein secondary structure prediction. *Prot. Sci.* 5:2298-2310, 1996.
7. Zvelebil, M. J., Barton, G. J., Taylor, W. R. and Sternberg, M. J. E. Prediction of protein secondary structure and active sites using alignment of homologous sequences. *J. Mol. Biol.* 195:957-961, 1987.
8. Przybylski, D. and Rost, B. Alignments grow, secondary structure prediction improves. Columbia University, 2001.
9. Rost, B. PHD: predicting one-dimensional protein structure by profile based neural networks. *Meth. Enzymol.* 266:525-539, 1996.
10. Altschul, S., et al. Gapped Blast and PSI-Blast: a new generation of protein database search programs. *Nucl. Acids Res.* 25:3389-3402, 1997.
11. King, R. D., et al. Is it better to combine predictions? *Prot. Engin.* 13:15-19, 2000.
12. Ouali, M. and King, R. D. Cascaded multiple classifiers for secondary structure prediction. *Prot. Sci.* 9:1162-1176, 2000.
13. Higgins, D. G., Thompson, J. D. and Gibson, T. J. Using CLUSTAL for multiple sequence alignments. *Meth. Enzymol.* 266:383-402, 1996.
14. Garnier, J., Osguthorpe, D. J. and Robson, B. Analysis of the accuracy and Implications of simple methods for predicting the secondary structure of globular proteins. *J. Mol. Biol.* 120:97-120, 1978.
15. Jones, D. T. Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* 292:195-202, 1999.
16. Rost, B. Predicting protein structure: better data, better results! *J. Mol. Biol.* in submission, 2001.
17. McGuffin, L. J., Bryson, K. and Jones, D. T. The PSIPRED protein structure prediction server. *Bioinformatics* 16:404-405, 2000.
18. Karplus, K., et al. Predicting protein structure using only sequence information. *Proteins* S3:121-125, 1999.
19. Baldi, P., Brunak, S., Frasconi, P., Soda, G. and Pollastri, G. Exploiting the past and the future in protein secondary structure prediction. *Bioinformatics* 15:937-946, 1999.
20. Baldi, P. and Brunak, S. *Bioinformatics: the machine learning approach.* Cambridge: MIT Press, 2001.
21. Eyrich, V., et al. EVA: continuous automatic evaluation of protein structure prediction servers. *Bioinformatics* submitted, 2001.
22. Rost, B. and Eyrich, V. EVA: large-scale analysis of secondary structure prediction. *Proteins* submitted, 2001.
23. Rost WWW, B. EVA measures of secondary structure prediction accuracy. WWW document ([http://cubic.bioc.columbia.edu/eva/doc/measure\\_sec.html](http://cubic.bioc.columbia.edu/eva/doc/measure_sec.html)): EMBL, 2001.
24. Rost, B., Sander, C. and Schneider, R. Redefining the goals of protein secondary structure prediction. *J. Mol. Biol.* 235:13-26, 1994.
25. Zemla, A., Venclovas, C., Fidelis, K. and Rost, B. A modified definition of SOV, a segment-based measure for protein secondary structure prediction assessment. *Proteins* 34:220-223, 1999.
26. Defay, T. and Cohen, F. E. Evaluation of current techniques for ab initio protein structure prediction. *Proteins* 23:431-445, 1995.
27. Matthews, B. W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Ac.* 405:442-451, 1975.