BIOINFORMATICS ORIGINAL PAPER

Vol. 21 no. 2 2005, pages 152–159
doi:10.1093/bioinformatics/bth487

A simple and fast secondary structure prediction method using hidden neural networks

Kuang Lin[1,*,†], Victor A. Simossis[2,†], Willam R. Taylor[1] and Jaap Heringa[2,3]

[1]Division of Mathematical Biology, The National Institute for Medical Research, The Ridgeway, Mill Hill, London NW7 1AA, UK, [2]Bioinformatics Section, Faculty of Sciences and [3]Centre for Integrative Bioinformatics (IBIVU), Faculty of Sciences and Faculty of Earth and Life Sciences, Vrije Universiteit Amsterdam, De Boelelaan 1081A, 1081 HV Amsterdam, The Netherlands

Received on March 3, 2004; revised on August 13, 2004; accepted on August 14, 2004
Advance Access publication September 17, 2004

ABSTRACT

Motivation: In this paper, we present a secondary structure prediction method YASPIN that unlike the current state-of-the-art methods utilizes a single neural network for predicting the secondary structure elements in a 7-state local structure scheme and then optimizes the output using a hidden Markov model, which results in providing more information for the prediction.

Results: YASPIN was compared with the current top-performing secondary structure prediction methods, such as PHDpsi, PROFsec, SSPro2, JNET and PSIPRED. The overall prediction accuracy on the independent EVA5 sequence set is comparable with that of the top performers, according to the Q3, SOV and Matthew's correlations accuracy measures. YASPIN shows the highest accuracy in terms of Q3 and SOV scores for strand prediction.

Availability: YASPIN is available on-line at the Centre for Integrative Bioinformatics website (http://ibivu.cs.vu.nl/programs/yaspinwww/) at the Vrije University in Amsterdam and will soon be mirrored on the Mathematical Biology website (http://www.mathbio.nimr.mrc.ac.uk) at the NIMR in London.

Contact: kxlin@nimr.mrc.ac.uk

INTRODUCTION

The field of secondary structure prediction has a history of over 40 years and a wide range of different models has been applied to tackle the problem (for reviews see Heringa, 2000; Rost, 2001; Simossis and Heringa, 2004). Qian and Sejnowski (1988) introduced one of the earliest artificial neural network (NN)-based methods. From the 1990s up to the present time, secondary structure prediction accuracy has improved to over 70% by incorporating the evolutionary information found in multiple sequence alignments (MSAs). Among the most successful methods till date, PHD (Rost and Sander, 1993), PHDpsi (Przybylski and Rost, 2002), PROFsec (B. Rost, unpublished data), SSPro2 (Pollastri *et al*., 2002), JNET (Cuff and Barton, 2000) and PSIPRED (Jones, 1999) employ various types of NNs to perform predictions using MSAs of homologous sequences. However, the improvement in secondary structure prediction accuracy by using MSAs is also directly connected to database size and search accuracy (Przybylski and Rost, 2002). As a result, all current top-performing methods, including the ones mentioned above, employ the iterative databank-searching tool PSI-BLAST (Altschul *et al*., 1997; Altschul and Koonin, 1998) to select homologous sequences for predicting the secondary structure. The prediction performances of these programs have extensively been documented in various assessments (Jones, 1999; Jones and Swindells, 2002; Albrecht *et al*., 2003; Eyrich *et al*., 2003; Fischer *et al*., 2003; Koh *et al*., 2003; McGuffin and Jones, 2003).

Many of the current NN-based methods use feed-forward multilayer perceptron networks, which are trained with the back-propagation algorithm (Bishop, 1995). The first layer network predicts the secondary structure of the central residue of a preset window size, according to a position-specific scoring matrix (PSSM) and/or another form of an MSA encoding. This is called the sequence-to-structure network. The second layer network, called the structure-to-structure network, filters the outputs from the first one and produces the final prediction results. Additional layers of networks or other decision-making models can further complement each of these network layers. For example, in the Prof method (Ouali and King, 2000), the final prediction results were obtained using four layers, a large number of NNs, combined with linear discrimination of multiple cascaded classifiers.

In YASPIN, we apply a single NN instead of employing a complex multilayered networks of NNs. However, the

[*] To whom correspondence should be addressed.

[†] The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

problem with using a single NN is that the prediction results are often 'broken' secondary structures, even elements of only one residue. This is not desirable as most observed secondary structures are composed of more than three residues. A common way to overcome this problem is to filter the predicted secondary structure elements (SSEs) from the NN by using additional NNs. In YASPIN we apply a hidden Markov model (HMM). The forward and backward algorithms of the HMMs are also used to assign the confidence for each prediction (prediction reliability scores). Finally, the prediction results are converted into 3-state secondary structure predictions ('H'-helix, 'E'-strand and '-'-other). YASPIN can be trained in a few days and can process a prediction in a few seconds.

METHODS AND DATASETS

The algorithm

YASPIN is a hidden neural network (HNN) secondary structure prediction method. YASPIN use a feed-forward perceptron network with one hidden layer for predicting the SSEs from the sequence. Then, these predictions are filtered using an HMM.

The YASPIN NN use the softmax transition function (Bishop, 1995) with a window of 15 residues. For each residue in this window, 20 units are used for the scores in the PSSM and 1 unit is used to mark where the window spans the terminals of protein chains. In total, the input layer has 315 units (21 × 15). For the hidden layer we use 15 units. The output layer has seven units, corresponding to seven local structure states: helix beginning (Hb), helix (H), helix end (He), strand beginning (Eb), strand (E), strand end (Ee) and coil (C). The beginnings and ends of the helix and strand elements we refer to are single residue positions.

The 7-state output of the NN is then passed through a HMM, which uses the Viterbi algorithm (Durbin, 1998) to optimally segment the 7-state predictions. The HMM defines the transition probabilities among the seven local structure states (Fig. 1). The final output is a 3-state secondary structure prediction ('H' for helix, 'E' for strand and '-' for coil).

Testing and training datasets

YASPIN was trained and tested using the SCOP1.65 database (Murzin *et al*., 1995; Hubbard *et al*., 1998). The test and training sets were built using the PDB25 set (25% maximum sequence identity) grouped together by ASTRAL (Brenner *et al*., 2000). Before using the PDB25 dataset, we removed all transmembrane entries (SCOP class f) resulting in a non-redundant set of 4256 proteins with known structures. The test set was extracted before training by random selection from the complete PDB25 set at a ratio of approximately 1:8. The 535 sequences selected for the test set were (1) at most 25% identical to the training set due to the nature of the PDB25 dataset and (2) were not part of the same superfamily as any of the remaining 3721 sequences of the training set, according to the SCOP superfamily definitions.

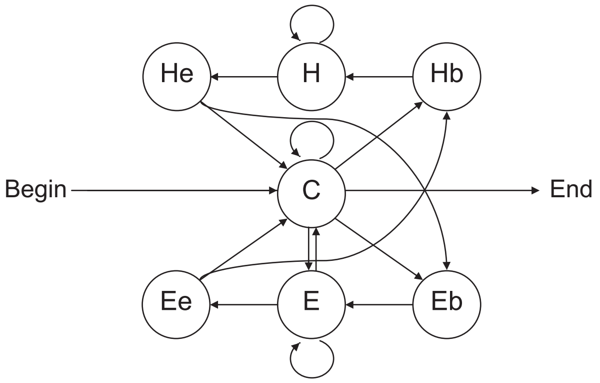

Fig. 1. The HNN state diagram. The arrows represent the allowed transitions in the HNN. H, E and C represent α-helix, β-strand and coil, respectively. The labels 'b' and 'e' indicate beginnings and ends of secondary structures.

In addition, to make a more accurate comparison between all methods, including YASPIN, we further benchmarked all methods on the independent 'common_set 5' dataset (10–2002) from EVA (Koh *et al*., 2003). To this end, we removed any sequences found in the EVA5 sequence set from the YASPIN training set. The final YASPIN training set contained 3553 sequences with known structures.

NN and HMM training

To train the YASPIN NN, we used the on-line back-propagation algorithm and 6-fold cross-validation (Bishop, 1995). In a single training iteration, each of the six subsets was successively kept apart for testing, while the remaining five were used to train the network. At the end of each training iteration, the average prediction error of the networks over all six test subsets was recorded and when the average prediction error started to increase, the training was stopped. We used a momentum term of 0.5 and a learning rate of 0.0001.

The reference secondary structure states used to train the HMM were obtained using DSSP (Kabsch and Sander, 1983). The DSSP 8-state secondary structure representation (H, G, E, B, I, S, T, -) was grouped according to the 3-state scheme proposed by Rost and Sander (1993), i.e. H and G were considered as helix (H), E and B as strand (E) and all others as coil (C). These 3-state definitions were later converted into our 7-state local structure scheme (Fig. 1). The transition probabilities of the HMM were estimated using the training set.

Reliability scores

The YASPIN prediction algorithm provides four different position-specific prediction confidence scores (reliability scores). These scores are generated based on the NN-predicted probabilities of each residue being in one of the defined seven states. The first three scores are secondary structure-specific scores, representing helix, strand and coil prediction confidence, and are generated as the sums of the probabilities of each respective secondary structure type. For example, let a residue X have a probability of being in any of the seven states. Its helix confidence score would be the sum of the Hb, H and He scores for that position. These three scores are normalized to always add up to 9.

The fourth score is the position-specific prediction confidence number, which represents the score of the state the Viterbi algorithm has chosen in its optimal segmentation path. All four scores are estimated using the HMM forward and backward algorithms.

PSSMs

All sequences in the test set were sequentially used as queries in a PSI-BLAST search against the non-redundant (NR) database. All involved secondary structure prediction methods were tested on the same PSI-BLAST results to make the comparison as unbiased as possible. The search parameters were set to satisfy the formatting and output needs of all the involved methods according to the suggestions of their corresponding authors. We used a cut-off of 0.001 (-h 0.001) as suggested by the PSIPRED parameter settings, a maximum of three iterations (-j 3), output formatting of type 6 that is needed by JNET, and also finally generated PSSM and Check files for each sequence. The actual command line was ‘blastpgp -i [query sequence] -h 0.001 -m 6 -j 3 -d nr -Q [PSSM] -C [CHECKFILE] > [BLAST OUTPUT]’.

Benchmarking

Benchmarking of YASPIN was performed using locally installed versions of the PHDpsi, PROFsec, SSPro2, JNET and PSIPRED programs. PHDpsi and PROFsec predictions were performed using the extracted alignments of the PSI-BLAST run. JNET was run using the extracted PSI-BLAST alignments, the PSI-BLAST PSSM files and the generated frequency profile files according to the authors’ instructions. The HMM profiles were included only in the prediction when available.

YASPIN’s prediction accuracy was compared with that of PHDpsi, PROFsec, SSPro2, JNET and PSIPRED by using the corresponding DSSP-derived secondary structures as a standard of truth. The translation from 8-state to 3-state secondary structure classification was performed according to the EVA (Koh *et al*., 2003) conversion scheme. The prediction accuracy of all methods was measured using the standard formulas for the Q3, SOV (Zemla *et al*., 1999) and Matthew’s correlation coefficients (MCCs) (for a review see Simossis and Heringa, 2004) as given on the EVA server (Koh *et al*., 2003).

Calculating prediction errors

We separated the prediction errors for helix and strand into four classes in accordance with the classification used by McGuffin and Jones (2003): (1) wrong prediction (w), (2) overprediction (o), (3) under-prediction (u) and (4) length (l) errors. The length errors were also recorded separately as overprediction and underprediction for comparison purposes between the methods. The four error types are illustrated for clarity as follows:

```
AA     MDYFTLFGLPARYQLDTQALSLRFQQLAAVQTINQ...
SS            HHHH    EEEEE HHH HHHHH     ...
DSSP     HHH  HHHH HHHHHHHHHH             ...
Errors   uuu       lllwwwwwl ll ooooo     ...
```

IMPLEMENTATION

YASPIN was trained on a non-redundant set of 3553 proteins with known structure from the PDB25 SCOP1.65 database. Its performance was tested using 535 proteins with known structure from the PDB25 dataset that were neither present in the training set nor part of the same SCOP-defined superfamily as any structure in the training set.

The PDB25 test set was also used to compare YASPIN to current top-performing methods, such as PHDpsi, PROFsec, SSPro2, JNET and PSIPRED. From the 535 sequences in the test set, 409 were found to be common to all methods, i.e. all methods returned a prediction for these proteins. This comparison was relatively unfair for YASPIN since many of these state-of-the-art methods have used sequences from this test set for their training. Nonetheless, the Q3 and SOV score results in Table 1 show that YASPIN is the best in strand prediction and also outperforms most methods in helix prediction, except SSPro2 and PSIPRED, which are clearly superior to YASPIN in this respect.

In addition, these methods were also benchmarked against the independent EVA5 sequence set (cumulative 10/2002). Since the EVA5 sequences were removed from the YASPIN training set (see Methods and dataset section) and all the other methods did not include these cases in their training, this dataset allows us to accurately compare YASPIN to these methods as well as the methods between themselves. From the 217 sequences in the EVA5 test set, 188 were found to be common for all methods. The prediction accuracies were assessed in three ways:(1) the 3-state per-residue prediction accuracy measure (Q3) (Fig. 2a), (2) the segment overlap measure (SOV) (Fig. 2b), both calculated using the SOV software (Zemla *et al*., 1999) and (3) the MCCs (Table 2).

The overall Q3 and SOV prediction accuracies of PHDpsi, PROFsec, SSPro2, YASPIN, JNET and PSIPRED on the 188 sequences in the EVA5 common test set are listed in Table 1 and plotted in Figure 2 with their significant error margins. The Q3 prediction accuracy results for separate SSEs (H, E and C) showed that PSIPRED and SSPro2 were the best in the prediction of helix with no significant differences between themselves, while YASPIN was significantly better than all the remaining methods. In addition, YASPIN was the

Table 1. The average Q3 and SOV scores for the predictions of 409 PDB25 common sequences from the testing set and 188 common sequences from the EVA5 set, with respect to the DSSP reference databases

	Q3	Q3H	Q3E	Q3C	SOV	SOVH	SOVE	SOVC
PDB25								
PSIPRED	67.63	69.36	55.17	71.15	63.14	67.39	57.36	61.99
Errsig	0.96	1.41	1.61	0.94	1.08	1.47	1.70	1.01
SSPRO2	67.39	67.22	52.91	72.78	62.33	65.30	55.75	62.28
Errsig	0.94	1.47	1.59	0.93	1.06	1.53	1.67	0.96
PROFsec	66.64	62.92	55.91	71.89	62.70	63.02	57.92	62.24
Errsig	0.91	1.51	1.56	0.91	1.04	1.58	1.64	0.95
YASPIN	66.41	64.34	58.40	69.92	62.15	63.82	58.87	60.60
Errsig	0.95	1.49	1.60	0.94	1.05	1.54	1.66	0.97
JNET	65.41	61.84	54.58	70.85	61.02	60.72	57.10	60.59
Errsig	0.90	1.51	1.56	0.94	1.01	1.57	1.64	0.93
PHDpsi	65.03	63.49	54.92	67.76	60.27	61.74	55.93	59.19
Errsig	0.90	1.51	1.57	0.96	0.99	1.53	1.61	0.93
EVA5								
PSIPRED	79.20	80.69	73.77	77.56	75.62	78.68	75.53	71.29
Errsig	0.68	1.60	1.93	0.93	1.09	1.74	2.02	1.14
SSPRO2	78.77	79.58	72.53	78.23	74.17	77.39	75.20	70.75
Errsig	0.71	1.71	1.86	0.92	1.13	1.85	1.98	1.11
PROFsec	77.56	71.73	73.52	77.56	73.61	71.17	75.18	69.74
Errsig	0.70	2.10	1.95	0.91	1.08	2.18	2.04	1.13
YASPIN	77.06	73.35	77.05	76.72	73.88	74.19	77.64	69.67
Errsig	0.74	1.83	1.87	0.90	1.11	1.90	1.95	1.19
JNET	75.72	72.55	70.89	76.75	72.94	71.78	74.21	69.54
Errsig	0.73	1.97	1.90	1.01	1.07	2.03	1.96	1.16
PHDpsi	75.44	72.47	70.79	73.39	71.19	70.19	72.78	66.23
Errsig	0.75	2.12	1.98	1.07	1.08	2.15	2.03	1.16

Q3H/E/C and SOVH/E/C values are the specific Q3 and SOV scores of the predicted helical, strand and coil regions, respectively. Errsig is the significant difference margin for each score and is defined as the SD (σ) over the square root of the number of proteins ($\sqrt{N}$). All the values are averaged over all $\alpha, \beta, \alpha + \beta$ and α/β proteins.

best at strand prediction with a significant difference from all the other methods. The above observations were also confirmed using the SOV scores. However, the MCCs showed that YASPIN and PROFsec are equivalent in prediction quality (Table 2), which suggests that the prediction error types made by each method are not accurately reflected in the Q3 and SOV scores.

Closer investigation of the types of errors made by each method on the EVA5 test set (Fig. 3) showed that all the methods are more or less missing out strand segments at the same rate (EU). On the other hand, PSIPRED and SSPro2 more frequently overpredict, while the rest underpredict, helix segments (Hlo/Hlu). YASPIN's prediction scores mainly suffer from relatively frequently mistaking helices for strand (HW), over-elongating strand segments (Elo) and keeping helices too short (Hlu).

YASPIN position-specific reliability measures

The reliability-scoring scheme applied in YASPIN correlated well with the average secondary structure prediction accuracy (Q3). The relationship between the assigned reliability scores and their corresponding average prediction accuracy was almost linear. This means that the YASPIN confidence-scoring scheme accurately describes the reliability of each prediction. In ~48% of the predicted residues showing a confidence value of 5 or greater, 90% were accurately predicted (Fig. 4).

The YASPIN server

YASPIN is freely available on-line at the Bioinformatics Unit website (http://ibivu.cs.vu.nl/programs/yaspinwww/) at the Vrije University in Amsterdam and will also be mirrored on the Division of Mathematical Biology website (http://mathbio.nimr.mrc.ac.uk/) at the NIMR in London. The YASPIN server can perform predictions using a protein sequence or an already existing PSSM.

In addition, YASPIN has been integrated into the automated secondary structure prediction initiative of the EVA server (Koh *et al*., 2003) for continual assessment of its prediction capabilities.

DISCUSSION

The difference between YASPIN and classical NN-based programs, such as JNET (JPRED), PSIPRED, SSPro2, PROFsec and PHDpsi, is the HNN model (Krogh and Riis, 1999). It is worth noting that in the original HNN paper (Krogh and Riis,

Table 2. The different error types and the MCCs for the YASPIN predictions in comparison to the other methods on the 409 PDB25 and 188 EVA5 common test sets

	Different error types								MCCs			
	HW	HO	HU	HL	EW	EO	EU	EL	MCC	MCC_H	MCC_E	MCC_C
PDB25												
PHDpsi	1670	336	2153	9165	2281	1193	2277	5594	0.45	0.52	0.43	0.40
PROFsec	1548	344	2241	8506	1964	1043	2347	5472	0.48	0.54	0.46	0.43
SSPro2	1181	793	1701	8495	2150	896	2378	5051	0.49	0.56	0.47	0.45
YASPIN	2037	441	2087	8332	1908	1058	2165	5857	0.47	0.54	0.45	0.42
JNET	1569	368	2406	8887	2249	1052	2204	5453	0.46	0.52	0.45	0.42
PSIPRED	1297	604	1757	8311	2182	856	2221	4992	0.50	0.57	0.49	0.46
EVA5												
PHDpsi	369	196	888	2843	362	283	751	2330	0.61	0.69	0.62	0.54
PROFsec	320	210	873	2529	253	259	732	2111	0.65	0.72	0.66	0.58
SSPro2	236	381	608	2457	340	206	698	1959	0.67	0.73	0.67	0.60
YASPIN	515	242	827	2316	186	287	712	2079	0.65	0.72	0.66	0.59
JNET	397	168	940	2732	527	263	634	2160	0.62	0.68	0.62	0.57
PSIPRED	225	291	650	2339	343	167	725	1879	0.68	0.74	0.68	0.61

H/EW, wrong prediction (H → E, E → H); H/EO, helix or strand structure overpredicted; H/EU, helix or strand structure underprediction; and H/EL, helix or strand structure length error.

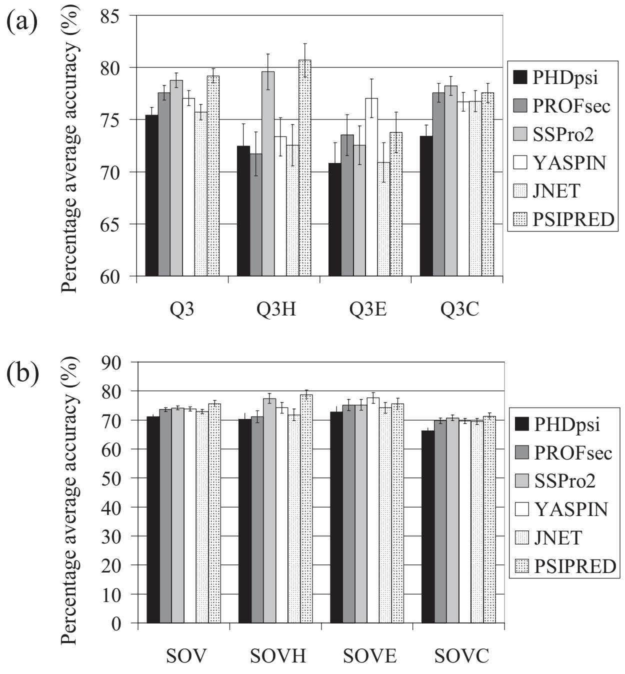

Fig. 2. The (a) Q3 and (b) SOV scores for PHDpsi, PROFsec, SSPro2, YASPIN, JNET and PSIPRED on the independent EVA5 common dataset (188 sequences). Q3H/E/C and SOVH/E/C values are the specific Q3 and SOV scores of the predicted helical, strand and coil regions, respectively.

1999), the NN and HMM components of the HNN model were trained in combination, while in later approaches including YASPIN, the NN and HMM have been trained separately. The latter training mode has also recently been applied to an HNN model for the prediction of protein residue contacts (Martelli *et al.*, 2002).

In YASPIN, the initial predictions from the sequence-to-structure network are 7-state predictions of protein local structures, instead of the commonly used 3-state. The importance of this is that termini of SSEs, especially helices, have statistically significant different composition from other parts of the protein sequence (Richardson and Richardson, 1988; Serrano and Fersht, 1989). The network used in YASPIN is trained to capture these differences and provide the additional information via producing these 7-state predictions. Furthermore, the HMM that optimizes these predictions before they are transformed into the secondary structure is much simpler than the layers of networks in other programs. YASPIN is capable of modelling higher order relationships between SSEs, since it finds a global best solution for the segmentation of the sequence into SSEs. The prediction accuracy of YASPIN can be compared with the existing top performing methods and the program is much faster.

The classic approach of defining protein local structures as 3-state secondary structures has been questioned recently (Pollastri *et al.*, 2002; Karchin *et al.*, 2003). One problem is that ∼50% of residues are regarded as parts of random coil, except for some that are found in distinct local structures. In addition, the amino acid composition of alpha helices and strands varies enormously. Efforts have been made to obtain finer classifications of local structures. For example, the I-sites library defines some of these sequence-structure motifs by clustering sequence segments from an NR database of known structures (Han and Baker, 1996; Bystroff and Baker, 1998). In this approach, an HMM (HMMSTR) was implemented to describe the transitions between these motifs

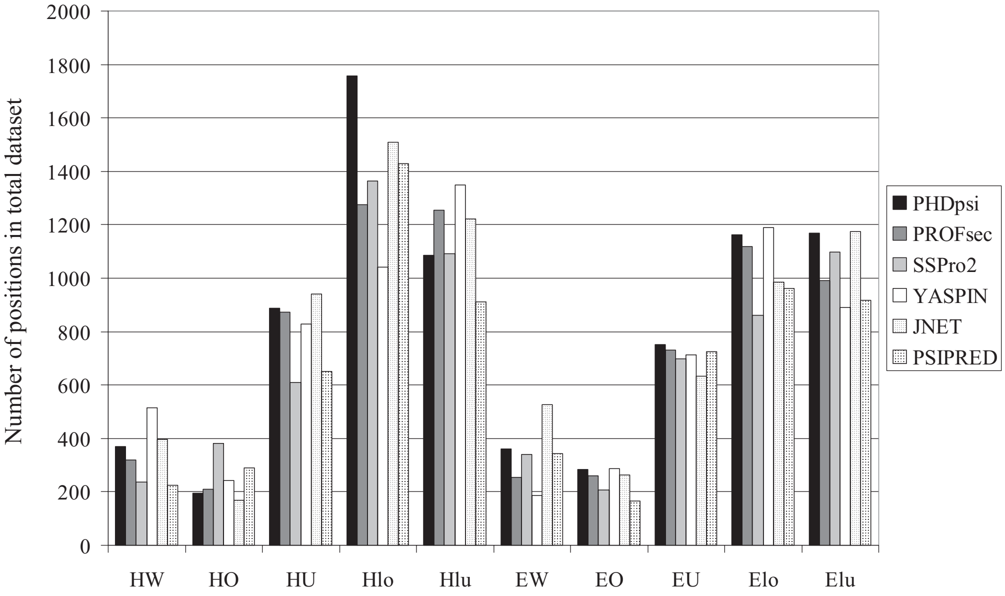

Fig. 3. The extent of errors made by each of PHDpsi, PROFsec, SSPro2, YASPIN, JNET and PSIPRED on the independent EVA5 common dataset (188 sequences). H/EW, wrong prediction (H → E, E → H); H/EO, helix or strand structure overpredicted: H/EU, helix or strand structure underprediction; H/Elo, helix or strand structure length errors due to over-prediction; and H/Elu, helix or strand structure length errors due to underprediction.

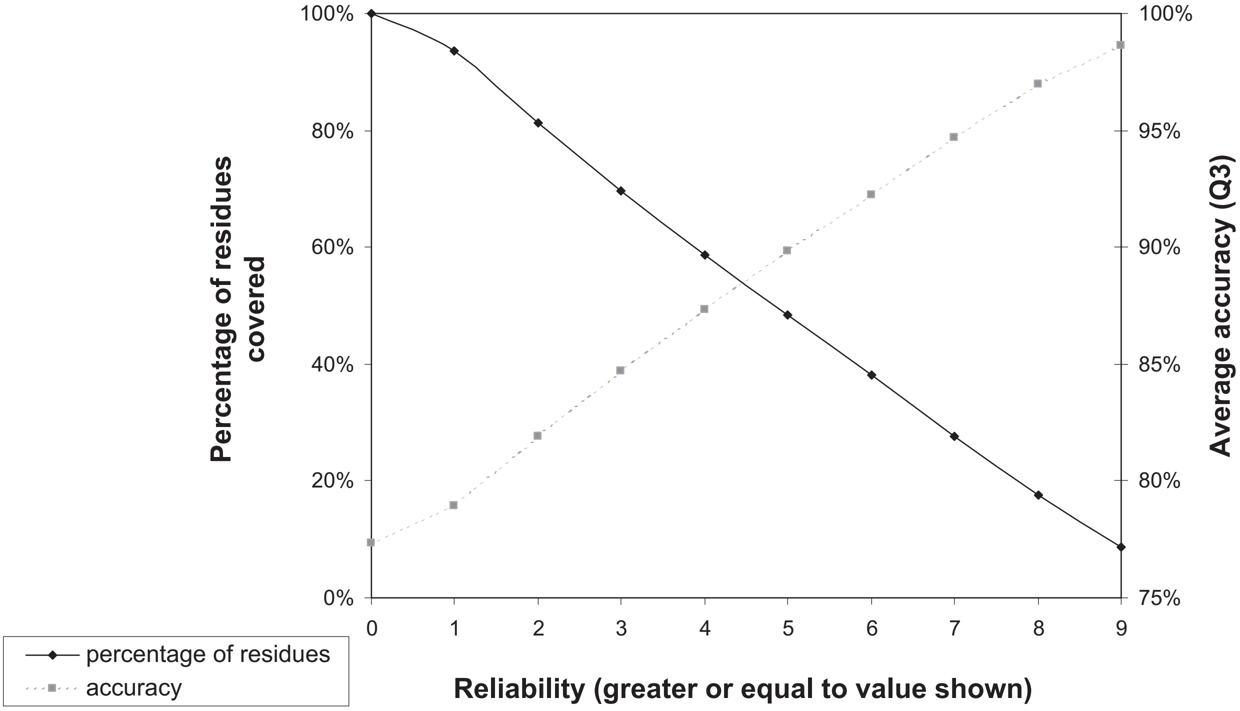

Fig. 4. Average secondary structure prediction accuracy (Q3) and the percentage of residues against cumulative reliability index from the YASPIN method. For example, for residues with reliability index of ≥6, the average accuracy is 92% and the percentage of residues with this index is 38%.

(Bystroff *et al*., 2000). This Markov model was also used for the prediction of protein secondary structures. However, its performance was not as good as some of the NN-based programs. HMMSTR tried to capture the recurrent local features of both protein sequences and protein structures in a single model. Sequence information was mostly represented as the amino acid preferences at different sites of motifs, rather than being memorized in NNs. This model was much more complex than the Markov model employed in YASPIN, which records transition probabilities of local structures only.

YASPIN does not use an alignment algorithm directly, but uses the information as encoded in the PSSM that can be generated using PSI-BLAST (Altschul *et al*., 1997; Altschul and Koonin, 1998) or any other alignment program. Prediction of local structure is performed using an NN, like many NN-based programs. However, the targets of our NN prediction are 7-state local structures, rather than the common 3-state secondary structures targeted in most NN-based programs. In this manner, more structural information can be obtained via the sequence-to-structure network. A problem with our model (Fig. 1) is that strands predicted by YASPIN must be of at least three residues as well. According to the DSSP definition, β-bridges can often have only one residue. To overcome this problem, two different Markov models were designed, each having fewer states of strand structures than those currently used (Eb, E and Ee), but there is a decrease in the prediction accuracy (data not shown). This suggests that the sequence signals of the strand termini are important for the prediction.

The current YASPIN implementation is a predictor designed for the traditional 3-state secondary structure definitions. However, the architecture of the HNN model makes it very easy to adopt the program to predict local structures with different classifications.

REFERENCES

Albrecht,M., Tosatto,S.C., Lengauer,T. and Valle,G. (2003) Simple consensus procedures are effective and sufficient in secondary structure prediction. *Protein Eng.*, **16**, 459–462.

Altschul,S.F. and Koonin,E.V. (1998) Iterated profile searches with PSI-BLAST—a tool for discovery in protein databases. *Trends Biochem. Sci.*, **23**, 444–447.

Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Bishop,C.M. (1995) *Neural Networks for Pattern Recognition*. Clarendon Press, Oxford University Press, Oxford.

Brenner,S.E., Koehl,P. and Levitt,M. (2000) The ASTRAL compendium for protein structure and sequence analysis. *Nucleic Acids Res.*, **28**, 254–256.

Bystroff,C. and Baker,D. (1998) Prediction of local structure in proteins using a library of sequence-structure motifs. *J. Mol. Biol.*, **281**, 565–577.

Bystroff,C., Thorsson,V. and Baker,D. (2000) HMMSTR: a hidden Markov model for local sequence-structure correlations in proteins. *J. Mol. Biol.*, **301**, 173–190.

Cuff,J.A. and Barton,G.J. (2000) Application of multiple sequence alignment profiles to improve protein secondary structure prediction. *Proteins*, **40**, 502–511.

Durbin,R. (1998) *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, Cambridge, UK.

Eyrich,V.A., Przybylski,D., Koh,I.Y., Grana,O., Pazos,F., Valencia,A. and Rost,B. (2003) CAFASP3 in the spotlight of EVA. *Proteins*, **53** (Suppl. 6), 548–560.

Fischer,D., Rychlewski,L., Dunbrack,R.L., Jr, Ortiz,A.R. and Elofsson,A. (2003) CAFASP3: the third critical assessment of fully automated structure prediction methods. *Proteins*, **53** (Suppl. 6), 503–516.

Han,K.F. and Baker,D. (1996) Global properties of the mapping between local amino acid sequence and local structure in proteins. *Proc. Natl Acad. Sci., USA*, **93**, 5814–5818.

Heringa,J. (2000) Computational methods for protein secondary structure prediction using multiple sequence alignments. *Curr. Protein Pept. Sci.*, **1**, 273–301.

Hubbard,T.J., Ailey,B., Brenner,S.E., Murzin,A.G. and Chothia,C. (1998) SCOP, Structural Classification of Proteins database: applications to evaluation of the effectiveness of sequence alignment methods and statistics of protein structural data. *Acta Crystallogr. D Biol. Crystallogr.*, **54**, 1147–1154.

Jones,D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.*, **292**, 195–202.

Jones,D.T. and Swindells,M.B. (2002) Getting the most from PSI-BLAST. *Trends Biochem Sci.*, **27**, 161–164.

Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.

Karchin,R., Cline,M., Mandel-Gutfreund,Y. and Karplus,K. (2003) Hidden Markov models that use predicted local structure for fold recognition: alphabets of backbone geometry. *Proteins*, **51**, 504–514.

Koh,I.Y., Eyrich,V.A., Marti-Renom,M.A., Przybylski,D., Madhusudhan,M.S., Eswar,N., Grana,O., Pazos,F., Valencia,A., Sali,A. and Rost,B. (2003) EVA: evaluation of protein structure prediction servers. *Nucleic Acids Res.*, **31**, 3311–3315.

Krogh,A. and Riis,S.K. (1999) Hidden neural networks. *Neural Comput.*, **11**, 541–563.

Martelli,P.L., Fariselli,P., Malaguti,L. and Casadio,R. (2002) Prediction of the disulfide bonding state of cysteines in proteins with hidden neural networks. *Protein Eng.*, **15**, 951–953.

McGuffin,L.J. and Jones,D.T. (2003) Benchmarking secondary structure prediction for fold recognition. *Proteins*, **52**, 166–175.

Murzin,A.G., Brenner,S.E., Hubbard,T. and Chothia,C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536–540.

Ouali,M. and King,R.D. (2000) Cascaded multiple classifiers for secondary structure prediction. *Protein Sci.*, **9**, 1162–1176.

Pollastri,G., Przybylski,D., Rost,B. and Baldi,P. (2002) Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles. *Proteins*, **47**, 228–235.

Przybylski,D. and Rost,B. (2002) Alignments grow, secondary structure prediction improves. *Proteins*, **46**, 197–205.

Qian,N. and Sejnowski,T.J. (1988) Predicting the secondary structure of globular proteins using neural network models. *J. Mol. Biol.*, **202**, 865–884.

Richardson,J.S. and Richardson,D.C. (1988) Amino acid preferences for specific locations at the ends of alpha helices. *Science*, **240**, 1648–1652.

Rost,B. (2001) Review: protein secondary structure prediction continues to rise. *J. Struct. Biol.*, **134**, 204–218.

Rost,B. and Sander,C. (1993) Prediction of protein secondary structure at better than 70% accuracy. *J. Mol. Biol.*, **232**, 584–599.

Serrano,L. and Fersht,A.R. (1989) Capping and alpha-helix stability. *Nature*, **342**, 296–299.

Simossis,V.A. and Heringa,J. (2004) Integrating secondary structure prediction and multiple sequence alignment. *Curr. Protein Pept. Sci.*, **5**, 1–15.

Zemla,A., Venclovas,C., Fidelis,K. and Rost,B. (1999) A modified definition of SOV, a segment-based measure for protein secondary structure prediction assessment. *Proteins*, **34**, 220–223.