Improving the Prediction of Protein Secondary Structure in Three and Eight Classes Using Recurrent Neural Networks and Profiles

Gianluca Pollastri
Department of Information and Computer Science
Institute for Genomics and Bioinformatics
University of California, Irvine
Irvine, CA 92697-3425
gpollast@ics.uci.edu

Darisz Przybylski and Burkhard Rost
CUBIC, Columbia University
Department of Biochemistry & Molecular Biophysics
650 West, 168th Street, BB217 New York, N.Y. 10032
{*dudek,rost*}*@columbia.edu*

Pierre Baldi*
Department of Information and Computer Science
Institute for Genomics and Bioinformatics
University of California, Irvine
Irvine, CA 92697-3425
(949) 824-5809
(949) 824-4056 (FAX)
pfbaldi@ics.uci.edu

Abstract

Secondary structure predictions are increasingly becoming the workhorse for several methods aiming at predicting protein structure and function. Here we use ensembles of bidirectional recurrent neural network architectures, PSI-BLAST-derived profiles, and a large non-redundant training set to derive two new predictors: (1) the second version of the SSpro program for secondary structure classification into three categories; (2) the first version of the SSpro8 program for secondary structure classification into the eight classes produced by the DSSP program. We describe the results of three different test sets on which SSpro achieves a sustained performance of about 78% correct prediction. We report confusion matrices, compare PSI-BLAST- to BLAST-derived profiles, and assess the corresponding performance improvements. SSpro and SSpro8 are implemented as web servers, available together with other structural feature predictors at:

http://promoter.ics.uci.edu/BRNN-PRED/.

Keywords: recurrent neural networks, profiles, evolutionary information, PSI-BLAST.

1 Introduction

Secondary structure predictions are increasingly becoming the workhorse for several methods aiming at predicting protein structure and function, especially on a genomic scale [24, 35, 25, 19, 31]. Several threading techniques aiming at the identification of structural similarities between proteins with different sequences are based on predictions of secondary structure [14, 15]. Predicting contact maps from primary sequence, secondary structure, and other structural features has also emerged as a key possible strategy for predicting protein structure [7, 27, 22].

*and Department of Biological Chemistry, College of Medicine, University of California, Irvine. To whom all correspondence should be addressed.

Methods predicting protein secondary structure have improved substantially in the 90's through the use of machine learning methods and evolutionary information taken from the divergence of proteins in the same structural family [33, 9, 34, 6, 18, 13, 4]. In recent years, increases in the available training data and progress in algorithmic approaches have boosted prediction accuracy to about 76% of all residues predicted correctly in one of the three states: helix, strand, and "other" [22, 32]. On the algorithmic front, improvements of various kinds have resulted mostly from combining predictors, from using more sensitive methods for deriving evolutionary profiles, and from developing more flexible machine learning architectures.

It is well known that combining predictors usually improves prediction accuracy. Current methods for predicting secondary structure typically combine multiple neural networks, sometimes several hundreds of them [26], trained more or less independently. Combination of different systems rather than networks has also been used [12]. At the alignment level, the ability to produce profiles that include increasingly remote homologs using PSI-BLAST [18] has also contributed to performance improvement. Divergent evolutionary profiles contain not only enough information to substantially improve prediction accuracy, but even to correctly predict long stretches of identical residues observed in alternative secondary structure states depending on non-local conditions [17, 21, 37]. An example is a method automatically identifying structural switches, and thus finding a remarkable connection between predicted secondary structure and aspects of function [21, 37]. Finally, at the algorithmic level, new bidirectional recurrent neural network architectures were introduced in [6] in combination with BLAST profiles to produce a first-generation secondary structure predictor SSpro 1.0.

Here we develop the second version, SSpro 2.0, using an ensemble of bi-directional recurrent neural networks and PSI-BLAST profiles. For the purpose of comparison, we use the same training set as SSpro 1.0, but with larger validation sets that have become available since the first version. We show improved performance, to about 78% correct prediction under stringent conditions. Secondary structure classification for resolved structures is typically obtained by collapsing the eight-class output of the the DSSP program [20] into the standard three classes. Since useful information may be present in the eight classes, we also develop SSpro8, a secondary structure predictor into eight classes.

2 Materials and Methods

2.1 Data: Training and Testing

The assignment of the SS categories to the experimentally determined 3D structure is nontrivial and typically performed by the widely-used DSSP program [20]. DSSP works by assigning potential backbone hydrogen bonds (based on the 3D coordinates of the backbone atoms) and subsequently by identifying repetitive bonding patterns.

The DSSP program classifies each residue into eight classes (H = alpha helix; B = residue in isolated beta-bridge; E = extended strand, participates in beta ladder; G = 3-helix [3/10 helix]; I = 5 helix [pi helix]; T = hydrogen bonded turn; S = bend; and "."). These are typically collapsed into the three standard classes associated with helices, beta-strand, and coils. In the CASP experiments [24, 11, 22] α contains H and G, β contains E and B, and γ contains everything else. This assignment is known to be somewhat "harder" to predict than the other ones used in the literature where, for instance, α is formed by DSSP class H, β by E, and γ by everything else (including DSSP classes G, S, T, B, I, and "."). Other assignments used in the literature include [30], where α contains DSSP classes H, G, and I. A study of the effect of various assignments on prediction performance can be found in [12]. It is clear, however, that it may also be of interest to build a finer-grained predictor for the eight classes produced by DSSP. Because some classes are fairly rare, lack of data may have been an obstacle in the past. But with the steady stream of new structures deposited in the PDB [10] every week, the time may have come to revisit this issue.

Four main data sets are used to develop and test the approach: a training set (TRAIN), and three test sets (R126, EVA, and CASP4) to assess algorithmic performance in the most objective way. The distribution of the eight and three classes in these sets are summarized in Table 2.1.

TRAIN: In order to ensure fair comparison with SSpro 1.0, we use the same training test, originally developed at the end of 1999. We constructed a data set containing all proteins in PDB which are (1) at least 30 amino acids long; (2) have no chain breaks (defined as neighboring amino acids in the sequence having C^{α}-distances exceeding 4.0Å); (3) produce a DSSP output; and (4) are obtained by X-ray diffraction methods with a resolution of at least 2.5 Å. Internal homology is reduced using an all-against-all alignment approach ([16]), keeping the PDB sequences with the best resolution. A 50% threshold curve is used for homology reduction. Furthermore, the proteins in the set have less than 25% identity with the sequences in the set R126. The resulting training set consists of 1180 sequences corresponding to 282,303 amino acids.

R126: As a first independent test set, we use the original set of 126 sequences of Rost and Sander, currently corresponding to a total of 23,363 amino acid positions [this number has varied slightly over the years due to changes and corrections in the PDB].

EVA: A novel test set is provided by all the sequences available from the real-time evaluation experiment EVA (http://cubic.bioc.columbia.edu/ eva/), which compares a number of prediction servers on a regular basis using the sequences deposited in the PDB every week. In particular, we use the set labeled "common3" published on 3/3/2001, the largest EVA data set on which SSpro 1.0 had been tested together with the other main servers at that date. The set consists of 223 proteins with a total of 47,370 residues and contains sequences with no homology to proteins previously stored in PDB. This set was uploaded in PDB between 3/2000 and 3/2001; thus it has no homology to the TRAIN set extracted from PDB in September 1999.

CASP4: The last novel test set is provided by the 40 CASP4 sequences available at http://predictioncenter.llnl.gov/casp4/ and corresponding to 9,047 residues. Some of the sequences show homology to PDB structures. In this case we could not obtain the PDB file of three of the proteins [T91, T92, T93], hence were unable to run DSSP on them and did not use this set for testing classification into eight classes. The secondary structure assignment into three classes for this set was downloaded directly from the CASP4 web site.

Table 1: Eight- and three-class assignment statistics for the four sets adopted. For each set, the first column gives the total number of residues and the second the corresponding percentages. Three-class: H = helix; E = strand;

and C = coil. Eight-class: H = alpha helix; B = residue in isolated beta-bridge; E = extended strand, participates in beta ladder; G = 3-helix [3/10 helix]; I = 5 helix [pi helix]; T = hydrogen bonded turn; S = bend; and ".")

3class	8class	TRAIN		R126		EVA		CASP4	
	H	91911	32.56	6573	28.13	16421	34.67	-	-
	G	11173	3.96	862	3.69	1751	3.70	-	-
	I	67	0.02	5	0.02	8	0.02	-	-
H		103151	36.54	7440	31.85	18180	38.38	3600	39.79
	E	59302	21.01	5068	21.69	8940	18.87	-	-
	B	3634	1.29	353	1.51	489	1.03	-	-
E		62936	22.29	5421	23.20	9429	19.91	2048	22.64
	S	25036	8.87	2539	10.87	4536	9.58	-	-
	T	33452	11.85	2718	11.63	5581	11.78	-	-
	.	57728	20.45	5245	22.45	9644	20.36	-	-
C		116216	41.17	10502	44.95	19769	41.73	3399	37.57
All		282303	100.00	23363	100.00	47370	100.00	9047	100.00

2.2 Profiles

To improve prediction, we use both BLAST and PSI-BLAST input profiles. Using profiles at the input level generally has been shown to yield better results than using profiles at the output level [6].

BLAST: Input profiles for SSpro 1.0 were constructed primarily by running the BLAST program [1] against the NR (non-redundant) database [3, 10], with standard default parameters (E=10.0, BLOSUM62 matrix). The version used was available online in October 1999 and contained approximately 420,000 protein sequences. For redundancy reduction, instead of applying a hard threshold that requires an arbitrary cutoff choice, we used a continuous weighting scheme. In this scheme, the weight of a sequence measures how different the sequence is from the profile in terms of the Kullback-Leibler divergence [4]. More specifically, for any given sequence, the weight is the sum over all columns in the profile of the Kullback-Leibler divergence between the delta distribution associated with the composition of the sequence in the column and the corresponding profile distribution. This is also a measure of the Shannon information in the sequence, given the current profile. Formally, the weight of sequence s is

$$W(s) = -\sum_c \log P[s(c)] = -logP[s] \tag{1}$$

where $P[s(c)]$ is the probability according to the profile of the letter s has in column c. Highly redundant sequences have higher probabilities and therefore are assigned a lower weight. In summary, every sequence in a given alignment is assigned a weight proportional to the information the sequence carries with respect to the unweighted profile. A weighted profile matrix is then recompiled and used as input to the prediction algorithm (see also [6]).

PSI-BLAST: Here we derive new profiles by aligning all proteins against the NR database using PSI-BLAST [2] with the following four-step protocol [29]. First, we filter and remove all database sequences with COILS to mark coiled-coil regions [23] and SEG to mark regions of low complexity [36]. Second, we align the query protein against this filtered database with an E-value threshold for the iteration of 10^{-10} (PSI-BLAST 'h' threshold) and a final threshold of E$\leq 10^{-3}$ to accept hits. The number of iterations is restricted to three to avoid drift [18, 29]. Third, we align the query against the unfiltered NR database using the previously-found, position-specific profile. Finally, we use the same weighting scheme as in the case of BLAST profiles to balance the profile and remove redundancy.

2.3 Recurrent Neural Network Architectures

In [6] BRNNs (Bidirectional Recurrent Neural Networks) were proposed as a class of recurrent neural network architectures that can address some of the limitations of simple feed-forward networks associated with small fixed-length input windows. A typical BRNN architecture is represented in Figure 1. In this architecture, the output decision or classification is determined by three components. First, there is a central component associated with the local window at the location t of the current prediction, as in standard feed-forward neural networks for secondary structure prediction. The main

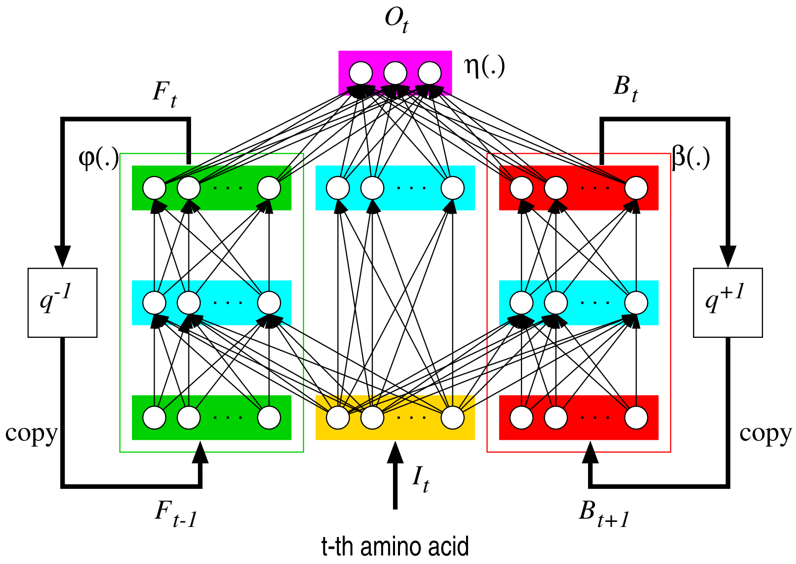

Figure 1: A BRNN architecture with a left (forward) and right (backward) context, F_t and B_t, associated with two recurrent networks (wheels). The output layer O_t has three normalized exponential units associated with membership in each one of the three secondary structure classes for the current residue at position t. The functions ϕ, β and η are implemented by feed-forward neural networks.

difference between the BRNN and the standard approach is the contribution of the left and right "contexts." These are produced by two similar recurrent networks which, intuitively, can be thought of in terms of two "wheels" being rolled along the protein chain, from the N and the C terminus all the way to the point of prediction.

Architectural variants can be obtained by changing the size of the input windows, the size of the window of hidden states used to determine the output, the number of hidden layers, the number of hidden units in each layer, and so forth. As in standard secondary-structure and other protein prediction architectures, we use sparse encoding for the 20 amino acids.

In what follows, we use the following notation:

Ct = size of semi-window of context states considered by the output network;

NFB = number of output units in the left (forward) and right (backward) context networks (wheels);

NHO = number of hidden units in the output network;

NHC = number of hidden units in the context networks.

In the 3 (resp. 8) class-prediction applications considered here, there are 3 (resp. 8) normalized exponential output units that enable us to estimate the class-membership probability for the residue being considered. The output error is the relative entropy between the output and target probability distributions [4]. All the weights of the BRNN architecture, including the weights in the recurrent wheels, can be trained from examples in a supervised fashion using a generalized form of gradient descent or backpropagation through time, or rather space, because of the forward and backward nature of the chains.

BRNNs were used to develop the first version of the SSpro predictor [6]. They have also been used for the prediction of amino acid partners in beta sheets [8], contact numbers [27], and relative solvent accessibility. In [6] evidence is provided that in the case of secondary structure prediction, these architectures extend the range over which information can be effectively integrated with respect to a feed-forward neural network, up to an effective window size of about 30 amino acids.

2.4 Experiments

In terms of architectures, both the three- and eight-class predictors use the same ensemble of 11 BRNNs, as in the on-line version of SSpro 1.0. Parameters of the 11 networks are given in Table 2. The total number of parameters varies over one order of magnitude from 1,899 to 18,107. Here we train SSpro 2.0 using PSI-BLAST profiles, SSpro8 1.0 using BLAST profiles, and SSpro8 2.0 using PSI-BLAST profiles. SSpro predicts the CASP three-class assignment, whereas SSpro8 predicts the eight classes that are produced by the DSSP program.

The training strategy adopted is the same for all the systems, and derives from the preliminary studies reported in [6]. In a typical case, we use a hybrid learning approach that combines online and batch training, with 430 batch blocks (2-3 proteins each) per training set. Thus, weights are updated 430 times per epoch after each block. The learning rate per block is initially set at about 1.5×10^{-4}, corresponding to the number of blocks divided by 10 times the number of residues ($0.1 \times 430/280000$), and is progressively decreased. The training set is also shuffled at each epoch. When the error

Table 2: Parameters and total weights of the 11 BRNN models. Ct = size of semi-window of context states; NFB = number of output units in the forward and backward context networks; NHO = number of hidden units in the output network; NHC = number of hidden units in the context networks.

Model	Ct	NFB	NHO	NHC	Weights
0	3	8	11	9	2181
1	2	9	11	8	1899
2	3	12	11	9	2949
3	3	12	9	9	2565
4	3	15	12	13	4167
5	3	17	12	15	4831
6	3	17	14	15	5355
7	3	15	14	15	4839
8	4	15	14	15	5679
9	4	25	30	27	18107
10	3	25	30	27	15107

does not decrease for 50 consecutive epochs, the learning rate is divided by 2, and training is restarted from the lowest error model. Training stops after 8 or more reductions, corresponding to a learning rate that is 256 times smaller than the initial one, which usually happens after 1,500-2,500 epochs.

Several indices can be used to score the efficiency of the algorithm [5]. Here we use Q3 or Q8, the number of correctly predicted residues divided by the total number of residues, as well as the corresponding per-class version Qclass (percentage of residues in a given structural class that are correctly predicted) and the Matthew's correlation coefficient.

3 Results and Discussion

3.1 SSpro

Performance for each individual model and for the ensemble average is given in Table 3 for the TRAIN and R126 sets. In all cases, PSI-BLAST profiles provide a Q3 improvement of at least 1.5%, and often more. On the EVA set, SSpro 2.0 achieves 77.7%, better than all the other evaluated systems. Incidentally, training on BLAST profiles and testing on PSI-BLAST profiles also leads to some performance improvement, although not as much.

Table 3: Performances of all the 11 models and the Ensemble, for SSpro 1.0 and SSpro 2.0, on the R126 test set and on the training set using the percentage Q3 of correctly classified residues.

	SSpro 1.0		SSpro 2.0	
Model	Q3	Q3train	Q3	Q3train
0	74.91	77.24	76.85	78.27
1	74.82	76.93	76.53	78.36
2	75.20	78.03	76.62	79.02
3	74.98	77.62	76.35	78.94
4	75.15	79.02	76.30	80.39
5	74.54	79.80	75.90	80.79
6	74.50	78.93	76.08	80.87
7	74.80	78.66	76.61	80.91
8	74.57	78.55	76.32	80.91
9	74.40	81.00	74.84	85.21
10	73.37	83.26	75.16	84.21
Ensemble	76.62	81.01	78.13	83.02

In Table 4, we give the performance of SSpro 1.0 and SSpro 2.0 on the three test sets measured by Q3 and the percentage per class (Qclass). Again, the PSI-BLAST profiles lead to significant improvement in all categories. In some cases (E in EVA and CASP4) the improvement exceeds 3%. Similar results using Matthew's correlation coefficients are given in Table 5.

SSpro 2.0 achieves Q3 of 78.13% on R126 and 77.67% on EVA, which, to the best of our knowledge, at the time of this writing is second to none. [The results reported on EVA as of 3/3/01 at http://cubic.bioc.columbia.edu/ eva/sec/bup_common/2001_03_03/common1.html for the other tested predictors are: PROF1 76.8%, PSIpred 76.5%, JPRED 74.7%, PHDpsi 74.7%, and PHD 71.5%]. The error margin on the performance for single residues range from 0.19% for the EVA set to 0.42% for the CASP4 set, so the second decimal for Q3 is not particularly significant. The results currently reported on the EVA web site correspond to SSpro 1.0 only, since SSpro 2.0 came on line in April 2001 and EVA does not offer an automated procedure to evaluate a newly entered predictor using all sequences up to the time of entry. Although SSpro 2.0 is derived after the CASP4 experiment, it is trained using the same training set as SSpro 1.0 prior to CASP4. SSpro achieves a Q3 performance of 80.65% on the CASP4 set, which is known to contain sequences with a wide difficulty range [22]. For comparison, the predictor PSIPRED achieves 79.9% correct prediction per residue, computed using the official predictions reported at CASP4 for 39 sequences and submitting the remaining sequence (T0106) directly to the PSIPRED server. An overall performance of 80% also has been reported in [26], but using an easier mapping of the eight DSSP classes into three rather than the one used for the CASP experiments. With this easier assignment, SSpro 2.0 achieves performances above 80%.

Table 4: Performances of SSpro 1.0 and SSpro 2.0, on the R126, EVA, and CASP4 test sets. Q3 and Qclass percentages.

		SSpro 1.0	SSpro 2.0
126	H	80.79	82.38
	E	63.23	66.19
	C	80.56	81.26
	Q3	76.62	78.13
EVA	H	80.76	82.48
	E	62.50	65.56
	C	78.05	79.03
	Q3	76.00	77.67
CASP4	H	83.86	86.08
	E	61.87	68.51
	C	80.99	82.20
	Q3	77.80	80.65

Table 6 provides the confusion matrices of SSpro 1.0 and SSpro 2.0 mea-

Table 5: Performances of SSpro 1.0 and SSpro 2.0, on the R126, EVA, and CASP4 test sets. Matthews' correlation coefficients.

		SSpro 1.0	SSpro 2.0
R126	H	0.732	0.752
	E	0.598	0.634
	C	0.571	0.59
EVA	H	0.695	0.722
	E	0.594	0.632
	C	0.568	0.588
CASP4	H	0.749	0.788
	E	0.609	0.674
	C	0.599	0.634

sured on the R126 test set. Perhaps not surprisingly, beta-strands continue to remain the most difficult class probably du to a number of factors including the involvement of long-ranged interactions with respect to the primary sequence and the fact that they are underrepresented in the data (roughly: 20% strands and 35% helices).

3.2 SSpro8

For the eight-class prediction, it is first worth noting that class I is almost irrelevant since it represents 0.02% of cases. Class B is small (1-1.5%), and the number of training examples is not yet large enough to yield any reliable prediction. Class G is also relatively small (approximately 4%), but some generalization appears to be possible. The same is true for class S which represents approximately 9% of cases. Each one of the other four classes contains at least 8% of the total residues.

The performances of SSpro8 1.0 and 2.0 are reported in Tables 7 and 8. For both versions of SSpro8, no residue is predicted as being in either class I or class B. Residues observed in class I are predicted as being in class H, and residues in class B are predicted somewhat evenly as being in classes in "." or B, as shown by the confusion matrix in Table 9. Classes G and S are under-predicted. Roughly 30% of the residues classified in G by

Table 6: Confusion matrices for SSpro 1.0 and SSpro 2.0 on the set R126. Xpred = structure X is predicted. Yobs = structure Y is observed. Rows sum to 100%. The number in row Xpred and column Yobs represents the percentage of times structure Y is observed, given that structure X has been predicted.

	SSpro 1.0			SSpro 2.0		
	Hpred	Epred	Cpred	Hpred	Epred	Cpred
Hobs	80.79	2.64	16.57	82.38	1.83	15.79
Eobs	4.37	63.23	32.40	3.30	66.19	30.51
Cobs	9.85	9.60	80.55	9.64	9.10	81.26
	Hobs	Eobs	Cobs	Hobs	Eobs	Cobs
Hpred	82.50	3.25	14.25	83.69	2.44	13.87
Epred	4.24	73.90	21.86	2.91	76.57	20.52
Cpred	10.74	15.27	73.99	10.31	14.49	75.20

DSSP are classified as H by SSpro8, and 21% are classified as being in a turn (T). When the system predicts G or S, however, there is about a 40-50% chance it is correct, which is considerably better than a random prediction. Although class T represents only 11% of the residues, it tends to be slightly over-predicted. Approximately 43-45% of the observed turns are predicted correctly, and 40% of the predicted turns are actual turns. More precisely, bends (S) tend to be confused with turns (T). Out of the observed bends, more than 20% are predicted as turns (less than 10% as bends). If a T is predicted, the probability of the actual residue being in either a T or an S structure is approximately 2/3. A class obtained by merging turns and bends would be classified by SSpro8 2.0 with a 65% correct percentage of predicted residues. The overall performance shows a gain of 1.4-1.8% with PSI-BLAST profiles, reaching the 62.6-63.3% range. The error margin on the performance for single residues range from 0.22% for the EVA set to 0.32% for the R126 set.

Table 8 compares the two versions of SSpro8 on the other two test sets using Q8 and Qclass. The confusion matrix of SSpro8 2.0 is reported in Table 9.

Table 7: Performances of SSpro8 1.0 and SSpro8 2.0, on the R126 and EVA test sets. Q8 percentage and Qclass percentage of observed residues.

		SSpro8 1.0	SSpro8 2.0
R126	H	89.21	89.93
	G	6.38	8.70
	I	0.00	0.00
	E	76.85	78.77
	B	0.00	0.00
	S	6.58	7.48
	T	43.34	45.44
	.	57.79	61.33
	Q8	60.74	62.58
EVA	H	88.38	89.49
	G	4.63	6.85
	I	0.00	0.00
	E	74.49	76.10
	B	0.00	0.00
	S	4.43	5.78
	T	39.67	40.75
	.	58.59	60.51
	Q8	61.89	63.31

If we combine the predictions of SSpro8 using the CASP assignment, the performance obtained is inferior to SSPro by more than a percentage point. This is perhaps not too surprising, since SSpro8 is trained for a different task. Performance very close to SSpro (within 0.2%) can be achieved by cascading SSpro8 with a small neural network trained with threefold cross-validation on the test set. Overall, prediction into eight classes does not seem to improve prediction of secondary structure into three classes. However current results are encouraging, especially for turns, and are bound to improve as more data becomes available.

Table 8: Performances of SSpro8 1.0 and SSpro8 2.0, on the R126 and EVA test sets. Q8 percentage and Qclass percentage of predicted residues (percentage of correctly classified X, given that X is predicted). Note that no numbers are available for classes I and B because they are never predicted.

		SSpro8 1.0	SSpro8 2.0
R126	H	75.55	78.38
	G	36.91	41.44
	I	-	-
	E	62.39	66.02
	B	-	-
	S	49.41	48.47
	T	42.04	43.89
	.	49.94	50.36
	Q8	60.74	62.58
EVA	H	76.06	77.11
	G	33.61	38.46
	I	-	-
	E	61.67	64.70
	B	-	-
	S	38.58	44.33
	T	42.92	43.70
	.	48.83	49.91
	Q8	61.89	63.31

4 Conclusion

We have developed two state-of-the-art predictors for secondary structure in three and eight categories, using ensembles of bidirectional recurrent neural networks and PSI-BLAST profiles. We have assessed the gains attributable to the use of PSI-BLAST profiles over BLAST profiles and have implemented the programs in the form of two web servers, SSpro and SSpro8, available over the Internet at http://promoter.ics.uci.edu/BRNN-PRED/. Users can enter primary amino acid sequences, and predictions are emailed back to them after

Table 9: Confusion matrices for SSpro8 2.0 on the set R126. Xpred = structure X is predicted. Yobs = structure Y is observed. Rows sum to 100%. For instance, the number on Row Xpred, Column Yobs represents the percentage of times structure Y is observed, given that structure X has been predicted. Note that no numbers are available in the rows for classes I and B predicted, because they are never predicted.

	SSpro8 2.0							
	Hpred	Gpred	Ipred	Epred	Bpred	Spred	Tpred	.pred
Hobs	89.93	0.36	0.00	2.58	0.00	0.14	3.06	3.93
Gobs	28.19	8.70	0.00	11.49	0.00	0.58	20.88	30.16
Iobs	100.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Eobs	3.99	0.08	0.00	78.77	0.00	0.45	2.82	13.89
Bobs	7.37	0.00	0.00	30.59	0.00	0.85	7.37	53.82
Sobs	9.89	0.94	0.00	13.59	0.00	7.48	23.12	44.98
Tobs	20.46	1.36	0.00	7.69	0.00	2.35	45.44	22.70
.obs	6.62	0.32	0.00	21.43	0.00	1.87	8.43	61.33
	Hobs	Gobs	Iobs	Eobs	Bobs	Sobs	Tobs	.obs
Hpred	78.38	3.23	0.07	2.68	0.34	3.33	7.37	4.60
Gpred	13.26	41.44	0.00	2.21	0.00	13.26	20.44	9.39
Ipred	-	-	-	-	-	-	-	-
Epred	2.81	1.63	0.00	66.02	1.79	5.71	3.46	18.58
Bpred	-	-	-	-	-	-	-	-
Spred	2.29	1.28	0.00	5.87	0.76	48.47	16.33	25.00
Tpred	7.14	6.40	0.00	5.08	0.92	20.86	43.89	15.71
.pred	4.04	4.07	0.00	11.02	2.97	17.88	9.66	50.36

a short period of time, depending on server load. SSpro and SSpro8 are also part of a broader suite of programs aimed at predicting protein 3D structure via contact map prediction, and contact map prediction via prediction of structural features, such as secondary structure, relative solvent accessibility (ACCpro), and contact numbers (CONpro) [28].

Improvements of a few percentage points are significant, especially in the context of the Human Genome Project and other genome sequencing projects as well as high-throughput structural-proteomics projects. It is encouraging to witness gradual performance improvements year after year, as a result of algorithmic improvements and data growth. Perfect prediction cannot be expected for a number of reasons, including (1) dynamic properties of protein chains; (2) quaternary structures; (3) existence of proteins that do not fold spontaneously; (4) errors and variability in databases, as well as in DSSP program output; and (5) effects of external variables, such as pH, that currently are not taken into consideration. Thus, some degree of prediction saturation is likely to emerge in the coming years, although the exact level at which it will occur remains unclear. For the time being, efforts towards exhaustive prediction and exhaustive taxonomy of protein folds should continue to advance in synergy.

Acknowledgements

The work of PB and GP is supported by a Laurel Wilkening Faculty Innovation award and a Sun Microsystems award to PB at UCI. The work of DP and BR is supported by NIH grant R01-GM63029-01.

References

[1] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, and D.J. Lipman. Basic local alignment search tool. *J. Mol. Biol.*, 215:403–410, 1990.

[2] S.F. Altschul, T.L. Madden, and A.A. Schaffer. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucl. Acids. Res.*, 25:3389–3402, 1997.

[3] A. Bairoch and R. Apweiler. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucl. Acids Res.*, 28:45–48, 2000.

[4] P. Baldi and S. Brunak. *Bioinformatics: the machine learning approach.* MIT Press, Cambridge, MA, 2001. Second edition.

[5] P. Baldi, S. Brunak, Y. Chauvin, C. A. F. Andersen, and H. Nielsen. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, 16:412–424, 2000.

[6] P. Baldi, S. Brunak, P. Frasconi, G. Pollastri, and G. Soda. Exploiting the past and the future in protein secondary structure prediction. *Bioinformatics*, 15:937–946, 1999.

[7] P. Baldi and G. Pollastri. Machine learning structural and functional proteomics. *IEEE Intelligent Systems. Special Issue on Intelligent Systems in Biology*, 2001. In press.

[8] P. Baldi, G. Pollastri, C. A. F. Andersen, and S. Brunak. Matching protein β-sheet partners by feedforward and recurrent neural networks. In *Proceedings of the 2000 Conference on Intelligent Systems for Molecular Biology (ISMB00), La Jolla, CA*, pages 25–36. AAAI Press, Menlo Park, CA, 2000.

[9] G. J. Barton. Protein secondary structure prediction. *Curr. Opin. Str. Biol.*, 5:372–376, 1995.

[10] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The Protein Data Bank. *Nucl. Acids Res.*, 28:235–242, 2000.

[11] CASP3. Third community wide experiment on the critical assessment of techniques for protein structure prediction. Unpublished results available in `http://predictioncenter.llnl.gov/casp3`, December 1998.

[12] J. A. Cuff and G. J. Barton. Evaluation and improvement of multiple sequence methods for protein secondary structure prediction. *Proteins*, 34:508–519, 1999.

[13] J. A. Cuff and G. J. Barton. Application of multiple sequence alignments profiles to improve protein secondary structure prediction. *Proteins*, 40:502–511, 2000.

[14] D. Fischer, C. Barret, K. Bryson, A. Elofsson, A. Godzik, D. Jones, K. J. Karplus, L. A. Kelley, R. M. MacCallum, K. Pawowski, B. Rost,

L. Rychlewski, and M. Sternberg. CAFASP-1: critical assessment of fully automated structure prediction methods. *Proteins Suppl.*, 3:209–217, 1999.

[15] D. Fischer, A. Elofsson, L. Rychlewski, F. Pazos, A. Valencia, A. Godzik, B. Rost, A. R. Ortiz, and R. L. Dunbrack. CAFASP-2: the second critical assessment of fully automated structure prediction methods. *Proteins*, 2001. Submitted.

[16] U. Hobohm, M. Scharf, R. Schneider, and C. Sander. Selection of representative data sets. *Prot. Sci.*, 1:409–417, 1992.

[17] I. Jacoboni, P. L. Martelli, P. Fariselli, M. Compiani, and R. Casadio. Predictions of protein segments with the same aminoacid sequence and different secondary structure: a benchmark for predictive methods. *Proteins*, 41:535–544, 2000.

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

[19] D. T. Jones. Protein structure prediction in the postgenomic era. *Curr. Opin. Struct. Biol.*, 10:371–379, 2000.

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

[21] K. Kirshenbaum, M. Young, and S. Highsmith. Predicting allosteric switches in myosins. *Prot. Sci.*, 8:1806–1815, 1999.

[22] A. M. Lesk, L. Lo Conte, and T. J. P. Hubbard. Assessment of novel fold targets in CASP4: predictions of three-dimensional structures, secondary structures, function and genetics. *Proteins*, 2001. Submitted.

[23] A. Lupas, M. Van Dyke, and J. Stock. Predicting coiled coils from protein sequences. *Science*, 252:1162–1164, 1991.

[24] J. Moult, T. Hubbard, S. H. Bryant, K. Fidelis, and J. T. Pedersen. Critical assessment of methods of protein structure prediction (CASP): Round II. *Proteins*, 29:2–6, 1997. Supplement 1.

[25] J. Moult, T. Hubbard, S. H. Bryant, K. Fidelis, and J. T. Pedersen. Critical assessment of methods of protein structure prediction (CASP): round III. *Proteins Suppl.*, 3:22–29, 1999.

[26] T. N. Pedersen, C. Lundegaard, M. Nielsen, H. Bohr, J. Bohr, S. Brunak, G. P. Gippert, and O. Lund. Prediction of protein secondary structure at 80% accuracy. *Proteins*, 41(1):17–20, 2000.

[27] G. Pollastri, P. Baldi, P. Fariselli, and R. Casadio. Improved prediction of the number of residue contacts in proteins by recurrent neural networks. *Bioinformatics*, 17:S234–S242, 2001. Proceedings of the ISMB 2001 Conference.

[28] G. Pollastri, P. Baldi, P. Fariselli, and R. Casadio. Prediction of coordination number and relative solvent accessibility in proteins. *Proteins*, 2001. In press.

[29] D. Przybylski and B. Rost. Alignments grow, secondary structure prediction improves. *Proteins*, 2001. In press.

[30] S. K. Riis and A. Krogh. Improving prediction of protein secondary structure using structured neural networks and multiple sequence alignments. *J. Comput. Biol.*, 3:163–183, 1996.

[31] B. Rost. Review: protein secondary structure prediction continue to rise. *J. Struct. Biol.*, 134:204–218, 2001.

[32] B. Rost and V. Eyrich. EVA: large-scale analysis of secondary structure prediction. *Proteins*, 2001. In press.

[33] B. Rost and C. Sander. Combining evolutionary information and neural networks to predict protein secondary structure. *Proteins*, 19:55–72, 1994.

[34] B. Rost and C. Sander. Bridging the protein sequence-structure gap by structure predictions. *Annu. Rev. Biophys. Biomol. Struct.*, 25:113–136, 1996.

[35] R. Sanchez and A. Sali. Large-scale protein structure modeling of the Saccharomyces cerevisiae genome. *PNAS*, 954:13597–13602, 1998.

[36] J. C. Wootton and S. Federhen. Analysis of compositionally biased regions in sequence databases. *Meth. Enzymol.*, 266:554–571, 1996.

[37] M. Young, K. Kirshenbaum, K. A. Dill, and S. Highsmith. Predicting conformational switches in proteins. *Prot. Sci.*, 8:1752–1764, 1999.