BIOINFORMATICS APPLICATIONS NOTE Vol. 21 no. 8 2005, pages 1719–1720
doi:10.1093/bioinformatics/bti203

Structural bioinformatics

Porter: a new, accurate server for protein secondary structure prediction

Gianluca Pollastri[1,*] and Aoife McLysaght[2]

[1]Computer Science Department, University College Dublin, Belfield, Dublin 4, Ireland and [2]Genetics Department, Trinity College Dublin, Dublin 2, Ireland

Received on October 9, 2004; revised on November 29, 2004; accepted on December 2, 2004
Advance Access publication December 7, 2004

ABSTRACT

Summary: Porter is a new system for protein secondary structure prediction in three classes. Porter relies on bidirectional recurrent neural networks with shortcut connections, accurate coding of input profiles obtained from multiple sequence alignments, second stage filtering by recurrent neural networks, incorporation of long range information and large-scale ensembles of predictors. Porter's accuracy, tested by rigorous 5-fold cross-validation on a large set of proteins, exceeds 79%, significantly above a copy of the state-of-the-art SSpro server, better than any system published to date.

Availability: Porter is available as a public web server at http://distill.ucd.ie/porter/

Contact: gianluca.pollastri@ucd.ie

Protein secondary structure (SS) prediction is an important stage for the prediction of protein structure and function. Accurate SS information has been shown to improve the sensitivity of threading methods (e.g. Jones, 1999b) and is at the core of most *ab initio* methods (e.g. see Bradley *et al*., 2003) for the prediction of protein structure. Virtually all modern methods for protein SS prediction are based on machine learning techniques (Jones, 1999a; Pollastri *et al*., 2002), and exploit evolutionary information in the form of profiles extracted from alignments of multiple homologous sequences (MSAs). The progress of these methods over the last 10 years has been slow, but steady, and is due to numerous factors: the ever-increasing size of training sets; more sensitive methods for the detection of homologues, such as PSI-BLAST (Altschul *et al*., 1997); the use of ensembles of multiple predictors trained independently, sometimes tens of them (Petersen *et al*., 2000); more sophisticated machine learning techniques (e.g. Pollastri *et al*., 2002).

We have developed Porter, a new server for protein SS prediction. Porter is based on two layers of Bidirectional Recurrent Neural Networks (BRNN) and is an evolution of SSpro (Pollastri *et al*., 2002), one of the most accurate public servers to date (Rost and Eyrich, 2001; Lesk *et al*., 2001). The novel elements of Porter are accurate coding of input profiles obtained from MSA, second stage filtering by recurrent neural networks, incorporation of long-range information, large-scale ensembles of predictors and larger training sets.

Datasets. Porter is trained on the December 2003 25% pdb_select list. After processing by DSSP (Kabsch and Sander, 1983) the set contains 2171 proteins and 344 653 amino acids. We assign eight DSSP classes as follows: H, G, I → Helix; E, B → Strand; S, T, . → Coil. This assignment is known to be 'hard' and had been adopted at CASP (Lesk *et al*., 2001). More lenient assignments generally lead to higher performances. Profiles obtained from MSA have been shown to improve significantly SS prediction performances (starting from Rost and Sander, 1993). In Porter, we use MSA extracted from the NR database as available on March 3, 2004, containing over 1.4 million sequences. Redundancy in the database was first reduced at a 98% threshold, leading to 1.05 million sequences finally. The alignments were generated by three runs of PSI-BLAST (Altschul *et al*., 1997).

Input coding. }In Porter, the input at each residue is coded as a letter out of an alphabet of 25. Beside the 20 standard amino acids, B, U, X, Z and . (gap) are considered. The input presented to the networks is the frequency of each of the 24 non-gap symbols, plus the total frequency of gaps in each column of the alignment. This input coding scheme is richer than the 20-letter scheme adopted in SSpro (Pollastri *et al*., 2002).

Output filtering, incorporation of long-range information. We adopt a filtering network as for example in Rost and Sander (1993), but we augment the input to this network by the predictions of the first-stage network averaged over multiple contiguous windows, i.e. if $\sigma_j = (\alpha_j, \beta_j, \gamma_j)$ are the outputs in position j of the first stage network corresponding to the estimated probabilities of helix, strand and coil given the inputs, the input to the second stage network in position j is the array I_j:

$$I_j = \left(\sigma_j, \sum_{h=k_{-p}-w}^{k_{-p}+w} \sigma_h, \ldots, \sum_{h=k_p-w}^{k_p+w} \sigma_h \right),$$

where $k_f = j + f(2w+1)$, $2w+1$ is the size of the window over which first-stage predictions are averaged and $2p+1$ is the number of windows considered. In Porter $w = 7$ and $p = 7$, i.e. predictions at 225 contiguous residues are considered by the filtering network.

Large-scale ensembles. Five two-stage BRNN models are trained independently and ensemble averaged to build Porter. Differences among models are introduced by two factors: stochastic elements in the training protocol, such as different initial weights of the networks and different shuffling of the examples; different architecture and size of the models. In particular, we resorted to BRNN architectures with shortcuts (Baldi *et al*., 1999). In these, connections along the forward and backward hidden chains span more than

*To whom correspondence should be addressed.

Table 1. Overall Q_3%, Q_α%, Q_β%, C_α% and C_β% for SSpro 2.0 in Pollastri *et al*. (2002), SSpro retrained (SSproR) and incremental improvements leading to Porter

Method	Q_3 (%)	Q_h (%)	Q_e (%)	C_h (%)	C_e (%)
SSpro 2.0	78.13	82.4	66.2	75.2	63.4
SSproR	78.33	81.7	68.8	74.6	64.7
SSproRs	78.48	82.0	68.4	74.7	65.1
+25 sym	78.54	81.9	68.6	74.8	65.4
+Filter	78.89	82.4	69.2	75.3	66.2
Porter	79.01	82.2	69.4	75.6	66.4

Results for all systems except SSpro 2.0 are measured in 5-fold cross-validation. Differences >0.07% are statistically significant. SSproRs, shortucut models; 25 sym, 25 input symbols.

one-residue intervals, creating shorter paths between inputs and outputs. Averaging the five models' outputs leads to improvements in the range of 1–1.5% over single models. In Petersen *et al*. (2000), a slight improvement in the prediction accuracy was obtained by 'brute ensembling' of several tens of different models trained independently. Here, we adopted a less expensive technique: a copy of each of the five models is saved at regular intervals during training. The training protocol (similar to that described by Pollastri *et al*., 2002) guarantees that differences during training are non-trivial. In Porter we build an ensemble of 45 such copies (9 for each of the 5 models).

Results and conclusions. We measured the performances of each incremental improvement separately, by a 5-fold cross-validation procedure. The percentages of correctly classified residues (Q_3), helices and strands (Q_α, Q_β), and Matthews' correlation coefficients for helices and strands (C_α, C_β) by all systems are shown in Table 1. Q_3 differences >0.07% are statistically significant. An exact copy of SSpro, retrained on the new sets, obtains $Q_3 = 78.33\%$. An ensemble of five models with shortcuts achieves $Q_3 = 78.48\%$. When 25 input symbols are adopted, an improvement at the margin of statistical significance is observed ($Q_3 = 78.54\%$). The most sizeable gain (+0.35%) is obtained when two-layer BRNNs with long-range filtering are adopted. Large-scale ensembles lead to a further improvement, comparable to that reported by Petersen *et al*. (2000). The overall performance of Porter is $Q_3 = 79.01\%$ (SOV = 75.0%). Tested on the more lenient class assignment by Petersen *et al*. (2000), Porter surpasses 81% correct classification. Performance indices for the single classes indicate that most of Porter's gains come from more accurate prediction of strands.

We also tested Porter on the EVA (Rost and Eyrich, 2001) common2 set, as available in November 2004, containing 134 proteins. To ensure a fair comparison, we retrained Porter from scratch, after having excluded from its training set all sequences with >25% similarity to any sequence in common2. On this set, Porter achieves SOV = 72.0% and $Q_3 = 76.8\%$, better by at least 1.2 and 1.9%, respectively, than all the other servers evaluated.

ACKNOWLEDGEMENTS

The work of G.P. is supported by an SFI BRG 2004 and a UCD President's Award 2004.

REFERENCES

Altschul,S., Madden,T. and Schaffer,A. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

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

Bradley,P., Chivian,D., Meiler,J., Misura,K., Rohl,C., Schief,W., Wedemeyer,W., Schueler-Furman,O., Murphy,P., Schonbrun,J., Strauss,C. and Baker,D. (2003) Rosetta predictions in casp5: successes, failures, and prospects for complete automation. *Proteins*, **53**, 457–468.

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

Jones,D. (1999b) Genthreader: an efficient and reliable protein fold recognition method for genomic sequences. *J. Mol. Biol.*, **287**, 797–815.

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

Lesk,A., Lo Conte,L. and Hubbard,T. (2001) Assessment of novel fold targets in CASP4: predictions of three-dimensional structures, secondary structures, function and genetics. *Proteins*, **Suppl. 5**, 98–118.

Petersen,T., Lundegaard,C., Nielsen,M., Bohr,H., Bohr,J., Brunak,S., Gippert,G. and Lund,O. (2000) Prediction of protein secondary structure at 80% accuracy. *Proteins*, **41**, 17–20.

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.

Rost,B. and Eyrich,V. (2001) EVA: large-scale analysis of secondary structure prediction. *Proteins*, **Suppl. 5**, 192–199.

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