Porter 5: fast, state-of-the-art ab initio prediction of protein secondary structure in 3 and 8 classes

Mirko Torrisi, Manaz Kaleel and Gianluca Pollastri
University College Dublin

Abstract

Motivation: Although secondary structure predictors have been developed for decades, current ab initio methods have still some way to go to reach their theoretical limits. Moreover, the continuous effort towards harnessing ever expanding data sets and more sophisticated, deeper Machine Learning techniques, has not come to an end.
Results: Here we present Porter 5, the latest release of one of the best performing ab initio secondary structure predictors. Version 5 achieves 84% accuracy (84% SOV) when tested on 3 classes, and 73% accuracy (77% SOV) on 8 classes, on a large independent set, significantly outperforming all the most recent ab initio predictors we have tested.
Availability: The web and standalone versions of Porter5 are available at http://distilldeep.ucd.ie/porter/.
Contact: gianluca.pollastri@ucd.ie

1 Introduction

The prediction of protein Secondary Structure (SS) has been a central topic of research in Bioinformatics for many decades ([18]). In spite of this, even the most recent and sophisticated ab initio SS predictors are not able to reach the theoretical limit of three-state prediction accuracy (88-90%), while only a few predictors are currently able to generate eight-state SS predictions.

2 Approach

Similarly to its previous versions ([9]), Porter 5 is based on ensembles of Bidirectional Recurrent Neural Networks (BRNN). Our implementation consists of two similar cascaded stages, both of which contain a classic BRNN layer followed by a convolutional layer. We have determined the hyperparameters of the system through extensive testing in five-fold cross-validation. The resulting optimal hyperparameters have been used to train the final predictor on the full training set, while the results we report are obtained on a completely independent set. The novel elements of Porter

5 are the use of HHblits ([11]) in its pipeline alongside PSI-BLAST ([13]), a more informative encoding of the inputs, larger models, deeper training and the reliance on a larger training set than in the previous versions.

3 Methods

3.1 Datasets

The selection and preparation of datasets to adopt has a central role in any machine learning method ([15]). To build the training set we started from the Protein Data Bank (PDB) ([2]) from December 11, 2014 and redundancy-reduced it at a 25% sequence identity threshold. For testing purposes we started from the PDB released on June 14, 2017, redundancy-reduced it at a 25% sequence identity threshold against the training set, and then against itself to remove internal redundancy. Finally, we removed all proteins with at least 10 undetermined amino acids (AA) from both datasets. The training set contains 15,753 proteins and the test set 3,154 proteins, among the largest ever used to either train or test a SS predictor to the best of our knowledge.

3.2 Evolutionary information

The second key aspect of any SS predictor developed since the early 90's is harnessing evolutionary information ([12]) in the form of profiles or position-specific substitution matrices extracted from multiple aligned sequences. PSI-BLAST ([1]) has been widely used to find remote homologues, and a key component of Porter since its first release ([9]). More recently, HHblits ([11]) has distinguished itself for fast iterations and high quality results. Porter 5 relies on both algorithms, iterating them 3 times with an e-value of 0.001. PSI-BLAST is run on the Jun 3, 2016 version of UniRef90 ([14]). HHblits is run on the February, 2016 version of UniProt20. Our empirical tests show similar results when a model is trained with either PSI-BLAST or HHblits, but significantly improved results when both are used, either via an ensemble or by concatenating the resulting profiles.

3.3 Input Encoding

Although the size of the training set is large, our experiments confirm a positive correlation between number of alignments found, and quality of the prediction ([10]). We do not limit the number of alignment reported by either PSI-BLAST or HHblits, resulting in the generation of alignments with an average of ˜14,000, and ˜1,300 proteins with PSI-BLAST, and HHblits, respectively. We encode the alignments as frequency profiles. In particular we have 20 frequencies for the standard AA, 1 frequency for unknown or non-standard and one for gaps. Frequencies for AA and gaps are computed separately, that is: AA frequencies are computed ignoring gaps, while the gap frequency is equal to the total number of gaps divided by the total number of sequences aligned. Aligned sequences are weighed

by the information they carry with respect to the unweighed frequency profile ([6])), which in our test results in a significant performance improvement. Moreover, the frequency for the AA appearing in the query sequence is artificially "clipped" to 1. This does not reduce the information of the profile, as any one of the values in the initial (unclipped) profile is equal to 1 minus the sum of the others, while encoding into the input the identity of the query sequence. We have found this simple trick to be very beneficial in our tests, leading to significant performance improvements. We also build a version of the predictor which outputs the 8 SS classes produced by the DSSP program [5]. In this case three more numbers, representing the output of the three-state Porter 5, are added to the input.

3.4 Model and Training

We conducted preliminary experiments to identify optimal hyperparameters using only the training set in 5-fold cross-validation. In this phase we also extensively tested many types of neural network, including convolutional networks of increasing depths and simple windowed neural networks with up to 17 hidden layers, but our two-stage BRNN/CNN model ([9]) performed over 1% better than any of the alternatives we tried. The hyperparameters selected for the BRNNs trained on PSI-BLAST appeared close to optimal for both the BRNN trained on HHblits, and the BRNN trained on the concatenation of the inputs created with HHblits and PSI-BLAST. Only a slight increase of the hyperparameters was needed when switching from three-state to eight-state SS. Purely based on the training set we settled on an ensemble of 7 BRNN: 3 trained on PSI-BLAST, 3 trained on HHblits, and 1 trained on both (44 concatenated inputs rather than 22). These 7 BRNN were then trained on the full training set (on either three-state, or eight-state targets). The relatively small scale of the ensemble and the modest size of the individual models (on average 39k free parameters for the 3-class networks, 58k for 8 classes) ensure high speed at prediction time without affecting the accuracy or SOV score (Table 1). The test set was only used at the final stage, to obtain unbiased performance estimates for Porter 5 and other recent SS predictors ([15]).

4 Results

We tested Porter 5 against Porter 4 ([8]), SPIDER3 ([3]), SSpro 5.2 ([7]), PSIPRED 4.01 ([4]), RaptorX-Property ([16]) and DeepCNF-SS ([17]) on the test set we created, containing 3,154 proteins. However we were limited in this by the fact that SPIDER3 rejects proteins containing undetermined (X) amino acids (562 in the test set contain at least one), and that when we use the parameters required by SPIDER3, either PSI-BLAST or HHblits do not return a valid result for 129 proteins. Because of this we report performance results on two sets: one where we exclude the proteins on which we could not obtain a valid response from SPIDER3, containing 2,463 entries in total (Table 1); the full set of 3,154 proteins (Table 2) on which SPIDER3 is not assessed.

Method	Q3 per AA	SOV per AA	Q3 per protein	SOV per protein
Porter 5	**83.81%**	**83.29%**	**84.32%**	**84.57%**
SPIDER3	83.15%	82.04%	83.42%	83.17%
Porter5 (HHblits only)	83.06%	82.21%	83.68%	83.71%
SSpro 5 with templates	82.58%	80.13%	83.94%	82.49%
PSIPRED 4.01	81.88%	80.17%	82.48%	81.70%
RaptorX-Property	81.86%	81.86%	82.57%	83.13%
Porter 4	81.66%	81.11%	82.29%	82.51%
SSpro 5 ab initio	81.17%	78.54%	81.10%	79.45%
DeepCNF	81.04%	80.84%	81.16%	81.46%

Table 1: Performances on the set for which SPIDER3 generates predictions, sorted by Q3 accuracy.

Method	Q3	SOV	Q8	SOV
Porter 5	**84.19%**	**83.99%**	**73.02%**	**76.83%**
Porter 5 (HHblits only)	83.39%	82.86%	71.80%	75.45%
SSpro 5 with templates	82.62%	80.65%	71.91%	78.55%
PSIPRED 4.01	82.06%	80.55%	N.A.	N.A.
RaptorX-Property	82.04%	82.30%	70.74%	74.88%
Porter 4	82.00%	81.85%	N.A.	N.A.
DeepCNF	81.00%	81.02%	69.76%	73.60%
SSpro 5 ab initio	80.70%	78.68%	68.85%	72.26%

Table 2: Q3/Q8 accuracy and SOV score per AA on the full test set

Porter 5 is the most accurate predictor in our tests with 3-class accuracy just shy of 84% on the smaller testing set and 84.2% on the larger set, 0.7% better than SPIDER3, 1.2% and 1.6% better than SSpro with templates, and at least 2% more accurate than all the other predictors. Porter 5 is also very fast compared to the alternatives given the relatively small size of its models and the fact that it is built on in-house, heavily profiled code. For instance, once the alignments by PSI-BLAST and HHblits are present, Porter 5.0 runs 2 orders of magnitude faster than SPIDER3, although SPIDER3 is able to predict backbone angles, contact numbers, and solvent accessibility at the same time as secondary structure. To fully exploit its speed, we also assessed, and make available, a version of Porter 5 which is roughly three times faster depending on HHblits only.

It should be noted that our test set is redundancy reduced only against our training set, so our assessment of other recent SS predictors might be somewhat optimistic, as the test set may contain proteins similar to those they were trained on. Porter 5 is the best three-state and eight-state SS predictor according to all the measures observed on this set. Porter 5 with HHblits only performs better than SSpro 5 with templates, at least 1.2% better than RaptorX-Property (the only other predictor based on HHblits only), and has similar performances to SPIDER3.

We have also analysed the performance of Porter 5 separately on proteins resolved by Nuclear Magnetic Resonance (NMR) and by X-ray

Method	Q3	SOV	Q8	SOV
NMR	81.61%	78.23%	67.52%	70.71%
X-ray	**84.65%**	**85.08%**	**74%**	**77.93%**

Table 3: Porter 5 on NMR vs X-ray crystallography proteins

crystallography. NMR proteins are predicted at a significantly lower Q3 (81.6%), possibly because of their different statistics (e.g. average length and composition) or less certain determination of SS. The X-ray only section of the test set, which is roughly 90% of the total, is predicted at an average Q3 of 84.65% (Table 3).

5 Conclusion

We have built a state-of-the-art 3- and 8-state predictor of protein SS, Porter 5. In our tests it improves by roughly 2% accuracy on its previous version, and outperforms all the most recent predictors of secondary structure, including a template based one. Porter is freely available as a web server at http://distilldeep.ucd.ie/porter/ and also as a standalone program at the same address, alongside with all the datasets, and alignments generated.

Acknowledgements

We acknowledge the key role of the UCD cluster, and the UCD Research IT staff.

Funding

The work of M.T., and M.K. is supported by the Irish Research Council [GOIPG/2015/3717, and GOIPG/2014/603].

References

[1] S F Altschul, T L Madden, A A Schäffer, J Zhang, Z Zhang, W Miller, and D J Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 25(17):3389–3402, September 1997.

[2] Helen M. Berman, John Westbrook, Zukang Feng, Gary Gilliland, T. N. Bhat, Helge Weissig, Ilya N. Shindyalov, and Philip E. Bourne. The Protein Data Bank. *Nucleic Acids Research*, 28(1):235–242, January 2000.

[3] Rhys Heffernan, Yuedong Yang, Kuldip Paliwal, and Yaoqi Zhou. Capturing non-local interactions by long short-term memory bidirectional recurrent neural networks for improving prediction of protein secondary structure, backbone angles, contact numbers and solvent accessibility. *Bioinformatics*, 33(18):2842–2849, September 2017.

[4] David T Jones. Protein secondary structure prediction based on position-specific scoring matrices. *Journal of Molecular Biology*, 292(2):195–202, September 1999.

[5] Wolfgang Kabsch and Christian Sander. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12):2577–2637, December 1983.

[6] A. Krogh and G. Mitchison. Maximum entropy weighting of aligned sequences of proteins or DNA. *Proceedings. International Conference on Intelligent Systems for Molecular Biology*, 3:215–221, 1995.

[7] Christophe N. Magnan and Pierre Baldi. SSpro/ACCpro 5: almost perfect prediction of protein secondary structure and relative solvent accessibility using profiles, machine learning and structural similarity. *Bioinformatics*, 30(18):2592–2597, May 2014.

[8] Claudio Mirabello and Gianluca Pollastri. Porter, PaleAle 4.0: high-accuracy prediction of protein secondary structure and relative solvent accessibility. *Bioinformatics*, 29(16):2056–2058, June 2013.

[9] Gianluca Pollastri and Aoife McLysaght. Porter: a new, accurate server for protein secondary structure prediction. *Bioinformatics*, 21(8):1719–1720, December 2005.

[10] Dariusz Przybylski and Burkhard Rost. Alignments grow, secondary structure prediction improves. *Proteins*, 46(2):197–205, February 2002.

[11] Michael Remmert, Andreas Biegert, Andreas Hauser, and Johannes Söding. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nature Methods*, 9(2):173–175, February 2012.

[12] Burkhard Rost and Chris Sander. Combining evolutionary information and neural networks to predict protein secondary structure. *Proteins: Structure, Function, and Bioinformatics*, 19(1):55–72, May 1994.

[13] Alejandro A. Schäffer, L. Aravind, Thomas L. Madden, Sergei Shavirin, John L. Spouge, Yuri I. Wolf, Eugene V. Koonin, and Stephen F. Altschul. Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic Acids Research*, 29(14):2994–3005, July 2001.

[14] The UniProt Consortium. UniProt: the universal protein knowledgebase. *Nucleic Acids Research*, 45(D1):D158–D169, November 2016.

[15] Ian Walsh, Gianluca Pollastri, and Silvio C. E. Tosatto. Correct machine learning on protein sequences: a peer-reviewing perspective. *Briefings in Bioinformatics*, 17(5):831–840, September 2016.

[16] Sheng Wang, Wei Li, Shiwang Liu, and Jinbo Xu. RaptorX-Property: a web server for protein structure property prediction. *Nucleic Acids Research*, 44(W1):W430–W435, April 2016.

[17] Sheng Wang, Jian Peng, Jianzhu Ma, and Jinbo Xu. Protein Secondary Structure Prediction Using Deep Convolutional Neural Fields. *Scientific Reports*, 6, January 2016.

[18] Yuedong Yang, Jianzhao Gao, Jihua Wang, Rhys Heffernan, Jack Hanson, Kuldip Paliwal, and Yaoqi Zhou. Sixty-five years of the long march in protein secondary structure prediction: the final stretch? *Briefings in Bioinformatics*, December 2016.