BIOINFORMATICS APPLICATIONS NOTE Vol. 29 no. 16 2013, pages 2056–2058
doi:10.1093/bioinformatics/btt344

Structural bioinformatics

Advance Access publication June 14, 2013

Porter, PaleAle 4.0: high-accuracy prediction of protein secondary structure and relative solvent accessibility

Claudio Mirabello[1,2] and Gianluca Pollastri[1,2,*]

[1]School of Computer Science and Informatics and [2]Complex and Adaptive Systems Laboratory, University College Dublin, Dublin, Ireland

Associate Editor: Anna Tramontano

ABSTRACT

Summary: Protein secondary structure and solvent accessibility predictions are a fundamental intermediate step towards protein structure and function prediction. We present new systems for the *ab initio* prediction of protein secondary structure and solvent accessibility, Porter 4.0 and PaleAle 4.0. Porter 4.0 predicts secondary structure correctly for 82.2% of residues. PaleAle 4.0's accuracy is 80.0% for prediction in **two** classes with a 25% accessibility threshold. We show that the increasing training set sizes that come with the continuing growth of the Protein Data Bank keep yielding prediction quality improvements and examine the impact of protein resolution on prediction performances.

Availability: Porter 4.0 and PaleAle 4.0 are freely available for academic users at http://distillf.ucd.ie/porterpaleale/. Up to 64 kb of input in FASTA format can be processed in a single submission, with predictions now being returned to the user within a single web page and, optionally, a single email.

Contact: gianluca.pollastri@ucd.ie

Received on April 2, 2013; revised on June 1, 2013; accepted on June 8, 2013

1 INTRODUCTION

As the Critical Assessment of Techniques for Protein Structure Prediction enters its third decade of life after CASP10, one-dimensional protein structural features such as secondary structure (SS) and relative solvent accessibility (RSA) still remain a fundamental stage towards structure prediction, as they are indispensable for remote homology detection (Fischer and Eisenberg, 1996), contribute to contact map predictions (Vullo *et al*., 2006b), help the selection of local structural fragments for model reconstruction (Simons *et al*., 1999) and, ultimately, when known, considerably reduce the degrees of freedom available to a protein. SS and SA, alongside other structural features, have also proven useful in many other tasks, e.g. the prediction of protein subcellular localization (Nair and Rost, 2005), intrinsic disorder (Vullo *et al*., 2006a), bioactive peptides within proteins (Mooney *et al*., 2013) and protein–protein interactions (Porollo and Meller, 2007). Reflecting the usefulness of SS and RSA, our public predictors of these features, Porter and PaleAle have served a combined 275 000 queries to date from 40 000 distinct users, and Porter's SS predictions are used for fragment selection within the protein structure prediction package Rosetta (Simons *et al*., 1999).

We have trained larger, more expressive ensembles of cascaded bidirectional recurrent neural networks (BRNN) on recent redundancy-reduced subsets of the Protein Data Bank to predict SS and RSA. The new systems, Porter 4.0 and PaleAle 4.0, outperform our older public web servers Porter and PaleAle, with SS now at >82% correct prediction and 2-class (25% threshold) RSA predictions 80% correct. The new systems, Porter 4.0 and PaleAle 4.0, are freely available for academic users through a new public web server.

2 METHODS AND PERFORMANCES

Porter 4.0 and PaleAle 4.0 are based on ensembles of BRNN (Baldi *et al*., 1999). In both servers, we use a cascaded architecture as in (Pollastri and McLysaght, 2005; Pollastri *et al*., 2007; Mooney and Pollastri, 2009) with a first BRNN predicting SS from the primary sequence and multiple sequence alignments (MSA), and a second BRNN filtering the predictions of the first stage. MSA are generated with three rounds of PSI-BLAST (Altschul *et al*., 1997) against a 90% redundancy-reduced UniProt (Suzek *et al*., 2007). The main differences between Porter 4.0 and PaleAle 4.0 and our previous servers Porter and PaleAle are the greatly expanded size of the training sets, the rough doubling of the number of free parameters of the models to accommodate a greater number of examples and a deeper training procedure.

The dataset Porter 4.0 and PaleAle 4.0 are trained on is derived from the Protein Data Bank as available on June 23, 2012. We sorted the set by quality (measured as resolution + r_value/20 or fixed at 10 for NMR structures) and redundancy, and reduced it at a 25% sequence identity threshold, resulting in 9152 proteins. We further selected the 7522 proteins with a quality better than 4 (Full_set), but reserved the remaining 1630 proteins of lower quality for testing purposes (LowRes_set). A total of 2128 proteins of Full_set have a quality better than 2.5 (HiRes_set).

Porter 4.0 is trained to predict SS in three classes: Helix = H,G,I from DSSP (Kabsch and Sander, 1983); Strand = E,B from DSSP; and Coil = the rest. PaleAle 4.0 predicts RSA in four classes, with accessibility ranges: [0%,4%]; [4%,25%]; [25%,50%]; [50%,∞]. Both systems were trained and benchmarked in 5-fold crossvalidation on Full_set. Five architecturally different BRNN were trained for each fold and then ensembled. The number of free parameters in each BRNN ranges between

*To whom correspondence should be addressed.

Table 1. Porter and PaleAle performance improvements

Set size	2179	3129	4818	7522
Date	December, 2003	January, 2007	November, 2009	June, 2012
Porter (SS)	79.1%	80.5%	81.8%	82.2%
PaleAle (RSA)	–	54.4% (79.1%)	54.9% (79.5%)	55.3% (80%)

Note: Improvements for Porter and PaleAle over time. Two-class performances in brackets for PaleAle.

Table 2. Performances versus quality of structures

Training set	Test set	Porter (SS) (%)	PaleAle (RSA) (%)
Full_set	HiRes_set	82.3	56.8 (81)
	Full_set	82.2	55.3 (80)
	LowRes_set	78.6	49.6 (75.2)
HiRes_set	HiRes_set	81.3	55.9 (80.7)
	Full_set	80.8	54.2 (79.4)
	LowRes_set	77.5	48.4 (74.7)

Note: Performances of Porter and PaleAle on sets of diverse quality. When the training set is Full_set, performances on Full_set and HiRes_set are in 5-fold cross validation, whereas LowRes_set performances are measured as the average obtained by models from all 5-folds of Porter and PaleAle. When the training set is HiRes_set, performances on HiRes_set are in 5-fold cross validation, whereas Full_set and LowRes_set performances are averages from the 5 training folds. See text for more details on how the sets were obtained. Two-class performances in brackets for PaleAle.

13 000 and 18 000, compared with 5000–8000 parameters for the old Porter and PaleAle. The overall prediction performances are reported in Table 1, alongside the performances of past versions of the servers. Porter 4.0 predicts 82.2% of all residues in the correct class, an improvement of 3.1% over Porter (Pollastri and McLysaght, 2005). PaleAle's improvement is 0.9% over the 2007 version, for a 4-class correct classification of 55.3%, or 80.0% when the 4 classes are recast into 2 with an RSA threshold of 25%. Although PaleAle 4.0 is not optimized for this task, real valued solvent accessibility obtained from its class outputs by simple linear regression yields a mean average error of 0.14.

We also evaluated how the quality of a protein structure relates to the servers' accuracy (Table 2). Porter 4.0 is only marginally more accurate on the HiRes_set than on the Full_set (82.3% versus 82.2%), whereas PaleAle 4.0 is significantly more accurate (56.8% versus 55.3%). Similarly, when we test the methods on LowRes_set, which contains poor resolution and NMR proteins, PaleAle's performances decrease more markedly from Full_set levels (−5.7%) than those of Porter (−3.6%). We also trained the full Porter and PaleAle systems on all 9152 proteins (Full_set + LowRes_set). The overall results (not reported) were only marginally different, with a fractional improvement on LowRes_set and a loss of <0.1% on the rest of the set. Finally, we trained Porter and PaleAle only on HiRes_set and tested them on all three sets (Table 2). In this case, the performances are significantly lower on all sets for both predictors. The 5-fold cross validation performance of Porter on HiRes_set is 81.3%, which is higher than the 2004 results (79.1%), which were obtained on a similarly sized training set but with older MSA, but lower than what we obtain on Full_set (82.2%), which is larger, but based on the same MSA. This seems to suggest that at least some of the gains of the more recent predictors (Table 1) may come from training set sizes, rather than from the size of the sequence databases MSA are obtained from.

3 THE PORTER 4.0 AND PALEALE 4.0 WEB SERVERS

Porter 4.0 and PaleAle 4.0 are freely available for academic users as web servers at the address: http://distillf.ucd.ie/porterpaleale/. The servers implement an ensemble of all the models from all folds, that is 5 models by 5 folds or 25 models in total per server. As this effectively means that the servers are trained on 25% more proteins than their benchmark in 5-fold cross validation, their performances are expected to at least match those reported, although this will have to be validated when new independent sets will become available in the future.

The interface of the servers requires the user to input the queries in FASTA format and, optionally, an email address. Up to 64 kb of input (~200 average proteins) can be processed in a single submission. Larger predictions can be processed via subsequent submissions of up to 64 kb each. An exemption from the 64 kb limit can be obtained on a one-off basis by email request to the authors. The predictions are returned to the user as a web page and, optionally, via email if an address is provided. Links on the response web page also give access to confidence values for the predictions, and to PaleAle's outputs mapped onto real valued RSA. Depending on the load on the queue, a protein of 300 residues takes as little as 4 min to be processed and a maximal submission of 64 kb as little as 2 h. The sets used for training the servers are available for download on the web server.

Funding: Science Foundation Ireland [10/RFP/GEN2749]; Irish Research Council for Science, Engineering and Technology [postgraduate fellowship to C.M.].

Conflict of Interest: none declared.

REFERENCES

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

Baldi,P. *et al.* (1999) Exploiting the past and the future in protein secondary structure prediction. *Bioinformatics*, **15**, 937–946.
Fischer,D. and Eisenberg,D. (1996) Protein fold recognition using sequence-derived predictions. *Protein Sci.*, **5**, 947–955.
Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.
Mooney,C. and Pollastri,G. (2009) Beyond the Twilight Zone: automated prediction of structural properties of proteins by recursive neural networks and remote homology information. *Proteins*, **77**, 181–190.
Mooney,C. *et al.* (2013) PeptideLocator: prediction of bioactive peptides in protein sequences. *Bioinformatics*, **29**, 1120–1126.
Nair,R. and Rost,B. (2005) Mimicking cellular sorting improves prediction of subcellular localization. *J. Mol. Biol.*, **348**, 85–100.
Pollastri,G. and McLysaght,A. (2005) Porter: a new, accurate server for protein secondary structure prediction. *Bioinformatics*, **21**, 1719–1720.
Pollastri,G. *et al.* (2007) Accurate prediction of protein secondary structure and solvent accessibility by consensus combiners of sequence and structure information. *BMC Bioinformatics*, **8**, 12.
Porollo,A. and Meller,J. (2007) Prediction-based fingerprints of protein–protein interactions. *Proteins*, **66**, 630–645.
Simons,K. *et al.* (1999) Ab initio protein structure prediction of CASP III targets using ROSETTA. *Proteins*, **36** (**Suppl. 3**), 171–176.
Suzek,B.E. *et al.* (2007) UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics*, **23**, 1282–1288.
Vullo,A. *et al.* (2006a) Spritz: a server for the prediction of intrinsically disordered regions in protein sequences using kernel machines. *Nucleic Acids Res.*, **34** (**Suppl. 2**), W164–W168.
Vullo,A. *et al.* (2006b) A two-stage approach for improved prediction of residue contact maps. *BMC Bioinformatics*, **7**, 180.