BIOINFORMATICS APPLICATIONS NOTE Vol. 29 no. 14 2013, pages 1815–1816
doi:10.1093/bioinformatics/btt259

Structural bioinformatics

Advance Access publication May 8, 2013

PconsC: combination of direct information methods and alignments improves contact prediction

Marcin J. Skwark[1,2], Abbi Abdel-Rehim[1] and Arne Elofsson[1,2,*]

[1]Department of Biochemistry and Biophysics, Stockholm University, 10691 Stockholm and [2]Science for Life Laboratory, Stockholm University, Box 1031, 17121 Solna, Sweden

Associate Editor: Anna Tramontano

ABSTRACT

Summary: Recently, several new contact prediction methods have been published. They use (i) large sets of multiple aligned sequences and (ii) assume that correlations between columns in these alignments can be the results of indirect interaction. These methods are clearly superior to earlier methods when it comes to predicting contacts in proteins. Here, we demonstrate that combining predictions from two prediction methods, PSICOV and plmDCA, and two alignment methods, HHblits and jackhmmer at four different e-value cut-offs, provides a relative improvement of 20% in comparison with the best single method, exceeding 70% correct predictions for one contact prediction per residue.

Availability: The source code for PconsC along with supplementary data is freely available at http://c.pcons.net/

Contact: arne@bioinfo.se

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on January 30, 2013; revised on March 25, 2013; accepted on May 2, 2013

1 INTRODUCTION

Protein structure prediction *ab initio* is one of the longest standing challenges in structural biology. Initial methods for prediction showed little success, when tested blindly, because of insurmountable dimensionality of the unrestrained search space. One approach to enhance the structure prediction of a protein is to predict interacting residues for use as constraints during the folding process. However, until recently, the accuracy of contact predictors has been limited. This was overcome by using a large number of sequences and applying a 'global' model describing direct and indirect interactions (Burger and van Nimwegen, 2010; Giraud *et al.*, 1999; Weigt *et al.*, 2009). It has been shown that this information is sufficient for structure predictions (Marks *et al.*, 2011). Here, we examine whether it is possible to combine alternative alignment and prediction methods to improve the contact predictions further.

2 METHODS

PconsC uses predictions from two methods of inferring direct information: PSICOV (Jones *et al.*, 2012) (inverse covariance matrix estimation) and plmDCA (Ekeberg *et al.*, 2013) (pseudolikelihood with Potts models), using default settings and alignments from jackhmmer (Johnson *et al.*, 2010) against UniRef100 and HHblits (Remmert *et al.*, 2012) against its bundled nr20 database. For the different alignment methods, four different e-value cut-offs ($10^{-40}, 10^{-10}, 10^{-4}, 1$) and five iterations were used. Including alignments from PSI-BLAST or PfamA and/or using only mutual information performed significantly worse. Other direct information methods—mfDCA (Morcos *et al.*, 2011) and EVC (Hopf *et al.*, 2012)—perform on par with PSICOV, but because of great similarity in approach to plmDCA, they were not included.

The evaluation was conducted on the 150 proteins used in the development of PSICOV—'test set'. To avoid bias, we used an independent set ('training set') of 48 proteins (12 globular and 36 membrane ones) that are not homologous to each other or to any member of the test sets (no hits with e-value <0.1 with an jackhmmer search).

The training set contains only predicted contacts that appear within the top L (length of protein)-ranked contacts in any of the 16 input method-alignment combinations. For each of the combinations, the contacts (training samples) in the training set were identical. The input to the classifier consisted only of the raw prediction scores from the selected methods.

Both in training and benchmarking, a true contact was defined as two residues with at least one non-hydrogen atom not further than 5 Å away from any of the atoms of the other residue. Two residues are defined not to be in contact, if all of their non-hydrogen atoms are at least 8 Å apart from the atoms of the other residue. Only residue pairs with sequence separation of at least five amino acids space were considered.

Direct information-based contact prediction is inherently a classification problem, based on a noisy input. Moreover, predicted contact scores in both input methods are not directly comparable between different prediction targets. Therefore, we decided to fit an ensemble classifier—random forest—to the training data.

Random forests are known to be highly accurate, even on noisy datasets, but they may be prone to overfitting. To avoid it, we trained a 100-tree random forest with an early stopping condition, requiring at least 500 samples in newly created leaves. The number of trees in the forest and minimum leaf size were chosen based on the 5-fold cross-validation on the training set. PconsC optimization used the implementation of random forests available in the Python sklearn package (Pedregosa *et al.*, 2011). Using alternative classifiers, such as support vector machines, provided similar results.

3 RESULTS AND DISCUSSION

First, we analyzed the performance of the 16 (four different e-values, two different alignment methods and two different prediction methods) individual methods. The prediction performance of plmDCA is clearly superior to the other methods, regardless of the alignment method and cut-off chosen. Both methods used are equally suitable for contact prediction.

*To whom correspondence should be addressed.

Table 1. Prediction precision for combinations of prediction and alignment methods

Prediction method	plmDCA	PSICOV	plmDCA + PSICOV
PfamA	0.49	0.38	—
jackhmmer $e = 10^{-40}$	0.39	0.34	0.45
jackhmmer $e = 10^{-10}$	0.59	0.48	0.66
jackhmmer $e = 10^{-4}$	0.60	0.49	0.67
jackhmmer $e = 1$	0.60	0.50	0.67
HHblits $e = 10^{-40}$	0.34	0.33	0.40
HHblits $e = 10^{-10}$	0.58	0.48	0.66
HHblits $e = 10^{-4}$	0.60	0.51	0.69
HHblits $e = 1$	0.57	0.50	0.68
All jackhmmer	0.66	0.54	0.70
All HHblits	0.67	0.58	0.72
All HHblits and jackhmmer	0.69	0.61	0.73

Note: True positive: distance <5 Å, False positive: distance >8 Å.

Generally, using more permissive e-value cut-off results in greater prediction accuracy with jackhmmer, but for HHblits, the greatest precision is achieved with a cut-off of 10^{-4}. Nevertheless, PconsC incorporates four thresholds to capture evolutionary couplings in variably conserved areas of proteins. PfamA alignments render in general slightly worse results than the other alignment methods.

The combination of different e-value cut-offs does improve the precision (positive predictive value) by a few percentage points, see Table 1, whereas combining alignments from jackhmmer and HHblits provides an additional improvement, in particular at the higher ranked predictions (Supplementary Material). Finally, combining plmDCA and PSICOV into the PconsC method shows a substantial improvement, both when using only one set of alignments, see Table 1, or in particular when using all eight alignments, red line in Figure 1.

4 CONCLUSION

In conclusion, the benchmark data show that PconsC, a random forest approach, combining four jackhmmer alignments and four HHblits alignments at e-value cut-offs of 10^{-40}, 10^{-10}, 10^{-4} and 1 each and both PSICOV and plmDCA, provides a relative improvement of >20% in prediction precision (~12 percentage points in absolute terms). For PconsC, on average nearly three quarters of the top L predictions seem to be correct. This result is robust and holds also for other definitions of true and false contacts (e.g. Cα/β distances, different distance cut-offs etc, Supplementary Material). The question of feasibility of predicted contacts for structure modeling is out of the scope of this article, as is the discussion on potential other causes of strong evolutionary couplings, such as alternative conformers, multimeric contacts and functional implications.

It has not escaped our notice that the progress in the field of evolution-based contact prediction has been absolutely remarkable. We have moved from being able to predict a handful of contacts per protein, with a relatively low precision (with Mutual Information methods), to predicting hundreds, with astounding accuracy.

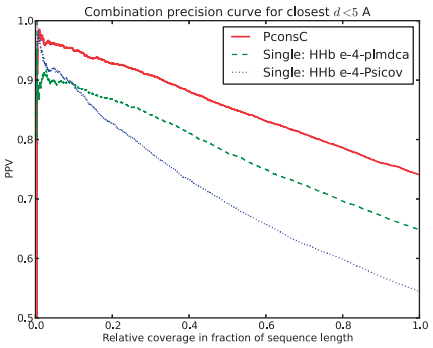

Fig. 1. Running average of precision (PPV) of PconsC predictions in comparison to individual methods. *X* axis: the relative rank of predicted contact (in terms of fraction of protein length), *Y* axis: average PPV of contacts with such rank or higher. Results on the 'test set' (150 sequences)

ACKNOWLEDGEMENT

The authors also thank Debora Marks and Johannes Söding for valuable comments while reviewing this article.

Funding: Swedish Research Council (VR-NT 2009-5072, VR-M 2010-3555), SSF (the Foundation for Strategic Research) and Vinnova through the Vinnova-JSP program, the EU 7'Th Framework Program by support to the EDICT project, contract No: FP7-HEALTH-F4-2007-201924. M.J.S. has been funded by TranSys, a Marie Curie ITN (No FP7-PEOPLE-2007-ITN-215524). Computing resources were provided by SNIC.

Conflict of Interest: none declared.

REFERENCES

Burger,L. and van Nimwegen,E. (2010) Disentangling direct from indirect co-evolution of residues in protein alignments. *PLoS Comput. Biol.*, **6**, e1000633.

Ekeberg,M. *et al.* (2013) Improved contact prediction in proteins: using pseudolikelihoods to infer Potts models. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.*, **87**, 012707.

Giraud,B. *et al.* (1999) Superadditive correlation. *Phys. Rev. E Stat. Phys. Plasmas Fluids Relat. Interdiscip. Topics*, **59(5 Pt. A)**, 4983–4991.

Hopf,T. *et al.* (2012) Three-dimensional structures of membrane proteins from genomic sequencing. *Cell*, **149**, 1607–1621.

Johnson,L. *et al.* (2010) Hidden markov model speed heuristic and iterative HMM search procedure. *BMC Bioinformatics*, **11**, 431.

Jones,D. *et al.* (2012) PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics*, **28**, 184–190.

Marks,D. *et al.* (2011) Protein 3D structure computed from evolutionary sequence variation. *PLoS One*, **6**, e28766.

Morcos,F. *et al.* (2011) Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc. Natl Acad. Sci. USA*, **108**, E1293–E1301.

Pedregosa,F. *et al.* (2011) Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.*, **12**, 2825–2830.

Remmert,M. *et al.* (2012) HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat. Methods*, **9**, 173–175.

Weigt,M. *et al.* (2009) Identification of direct residue contacts in protein-protein interaction by message passing. *Proc. Natl Acad. Sci. USA*, **106**, 67–72.