This is the accepted version of a paper published in *Bioinformatics*. This paper has been peer-reviewed but does not include the final publisher proof-corrections or journal pagination.

**Citation for the original published paper (version of record):**

PconsC: combination of direct information methods and alignments improves contact prediction.
*Bioinformatics*, 29(14): 1815-1816
http://dx.doi.org/10.1093/bioinformatics/btt259

Access to the published version may require subscription.

N.B. When citing this work, cite the original published paper.

**Permanent link to this version:**
http://urn.kb.se/resolve?urn=urn:nbn:se:diva-92624
PconsC: Combination of direct information methods and alignments improves contact prediction.

Marcin J. Skwark¹,², Abbi Abdel-Rehim¹ and Arne Elofsson¹,²,∗∗

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

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Summary: Recently, several new contact prediction methods have been published. They use (i) large sets of multiple aligned sequences (ii) and 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 cutoffs, provides a relative improvement of 20% in comparison to 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

1 INTRODUCTION

Protein structure prediction ab-initio is one of the longest standing challenges in structural biology. Initial methods for prediction showed very little success, when tested blindly, due to 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 utilizing a large number of sequences and applying a “global” model describing direct and indirect interactions (Giraud et al., 1999; Weigt et al., 2009; Burger and van Nimwegen, 2010). 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 cutoffs (10⁻⁴⁰, 10⁻¹⁰, 10⁻⁴, 1) and five iterations were used.

Including alignments from psiblast or Pfam 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 due to great similarity in approach to plmDCA were not included.

The evaluation was conducted on the 150 proteins used in the development of PSICOV —‘test set’. In order 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 others or to any member of the test sets (no hits with e-value <0.1 with a jackhammer 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 5 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 extensive data sets, but they tend to be prone to over-fitting. In order 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 scikit learn 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 sixteen (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 cutoff chosen. Both methods used are equally suitable for contact prediction. Generally, using more permissive e-value cutoff results in greater prediction accuracy with jackhmmer, but for HHblits the greatest precision is achieved with a cutoff of $10^{-4}$. Nevertheless, PconsC incorporates four thresholds, in order 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 cutoffs does improve the precision (positive predictive value) by a few percentage points, see Table 1, while combining alignments from jackhmmer and HHblits provides an additional improvement, in particular at the higher ranked predictions (see Supplementary Material). Finally, combining plmDCA and PSICOV into the PconsC method shows a quite substantial improvement, both when using only one set of alignments, see Table 1 or in particular when using all eight alignments, thick line in Figure 1.

### Concluding remarks

In conclusion, the benchmark data show that PconsC, a random forest approach, combining 4 jackhammer alignments and 4 HHblits alignments at e-value cutoffs of $10^{-40}$, $10^{-10}$, $10^{-4}$ and 1 each and both PSICOV and plmDCA, provides a relative improvement of over 20% in prediction precision (approx. 12 percentage points in absolute terms). For PconsC on average nearly three quarters of the top L predictions appear to be approximately correct. This result is robust and holds also for other definitions of true and false contacts (e.g. $\alpha/\beta$ distances, different distance cutoffs etc, see the supplementary material). The question of feasibility of predicted contacts for structure modeling is out of the scope of this paper, as is the discussion on potential 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 an astounding accuracy.

### ACKNOWLEDGEMENT

**Funding** This work was supported by grants from the 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 7Th Framework Program by support to the EDICT project, contract No: FP7-HEALTH-F4-2007-201924. MJS has been funded by TranSys, a Marie Curie ITN (No FPT-PEOPLE-2007-ITN-215524). We also thank Debora Marks and Johannes Söding for valuable comments when reviewing this paper.

### REFERENCES


