Published online 31 July 2019 *Nucleic Acids Research, 2019, Vol. 00, No. 00* **1–4**
doi:10.1093/nar/gkn000

PconsFam: A regularly updated database of structure predictions of all Pfam families available at http://pconsfold.bioinfo.se/

John Lamb [1,*], Aleksandra Jarmoliska [2,*] Mirco Michel [1,*], David Menndez-Hurtado [1,*], Joanna Sukowska,[2,†] Arne Elofsson [1,†*]

[1]Science for Life Laboratory and Department of Biochemistry and biophysics, Stockholm Unviersity, Tomtebodav 23, 171 21 Solna, Sweden and [2] Warsaw, Poland * contributed equally †=contributing authors

Received January 1, 2019; Revised February 1, 2019; Accepted March 1, 2019

ABSTRACT

Predicting contact maps and tertiary structure of proteins continues to be important as the progress of experimentally solved structures is slow. Even though grouping by both sequence and structure homology can in many cases give an idea of the structure of protein there are still many groups of proteins that lack a representative structure. To help in this, we here present PconsFam which is an intuative and interactive webinterface for predicted contact maps and tertiary structure models of the entire PFAM-database. By modelling all families, both those with and those without a representative structure, using the PconsFold2 pipeline, and running quality assessment estimator on them we can give an estimation for how confident the contact maps and structures are. PconsFam is planned to follow each PFAM release and present predicted structures for families both with and without representative structure and present them in an easily accessible format.

INTRODUCTION

In recent years it has been shown that by the use of evolutionary information and direct coupling analysis (**?**) it is possible to obtain sufficiently accurate contact prediction of proteins from their sequence and multiple sequence alignment alone to predict accurate structures of many protein families (**?**). At first the DCA methods were limited to very large protein families, but with the use of deep learning methodologies to improve the contact predictions it is nowadays possible to accurately predict the contacts for families with only a few hundred members (**???**)

Using these predicted contacts it is then possible to model the structure of a protein using a protein folding program. Here, initially CNS was used (**?**) but Rosetta has also been used (**?**). Although it is possible that models created by Rosetta are of slighlty higher quality than models by CNS (**?**) the advantagae of CNS is clear as it is much faster.

Table 1. Number of Pfam families with unknown structure that can be modeled at 1% and 10% FPR.

	0.01	0.1
ProQ3D	36	225
PcombC	42	179
Pcons	18	218
CNS-contact	62	232
Union	114	558
All	6379	6379

Recently we developed the PconsFold2 pipeline (**?**), which uses contact predictions from PconsC3 (**?**) the CNS based CONFOLD folding algorithm (**?**) and most importantly multiple model quality estimations (**??**) to predict the structure of proteins. Here we present the related web resource PconsFam (https://pconsfam.bioinfo.se) a database with predicted structural information for most Pfam families using the PconsFold2 pipeline.

The PconsFold2 pipeline can predict accurate models (TM-score >0.5 for 51% of the large families (>1000 effective sequences). For smaller families the fraction of correct models decreases, but they still exist. Therefore, a major challenge for large scale predictions is to distinguish between correct and incorrect models. Here, we have applied a set of model quality estimation methods (**?**).

When the PconsFold2 pipeline was applied to 6379 PFAM families of unknown structure 558 models with a predicted specificity over 90% was created. Out of these, 415 had never been reported before.

With PconsFam we extend this pipeline to the full PFAM database to predict multiple contact maps and subsequently multiple structures of all representative sequences for all Pfam families.

For each family in PFAM, multiple contact maps are generated and from these a set of models are predicted and ranked with quality assessment estimators. The model quality estimation gives an indication on the reliability of the model.

[*]To whom correspondence should be addressed. Tel: +46 706951045; Email: arne@bioinfo.se

The top ranked models for each predicted family can be visualised and are available for download. The full set of contact predictions for is available for visualisation together with the predicted model in an intuitive and powerful user-interface that allows interaction between the contact maps and the predicted structure.

MATERIAL AND METHODS

PconsFold2

PconsFold2 works in three separate steps. Firstly, multiple alignments are generated using HHblits and Jackhmmer (**?**) at different E-value cutoffs. Secondly, PConsC3 is used for contact predictions. All the alignments are used to create one contact map for each alignment. Thirdly, these contact maps, together with predicted secondary structure, is used as input to CONFOLD to generate 50 models for each contact map, resulting in a total of 200 models.

Model Quality assessment

The default ranking of models generated by CONFOLD is by the CNS contact energy (NOE) which is the sum of all violations of all contact restraints used in the input. To make the score comparable between models it is normalized by protein length. To easily rank the top model, the three measures Pcons, ProQ3D and PPV was combined to create the PcombC score as follows.

$$S_{PcombC}=\frac{0.3}{1.9}\cdot S_{Pcons}+\frac{0.6}{1.9}\cdot S_{ProQ3D}+\frac{1.0}{1.9}\cdot PPV$$

This was used on all families with a known structure to rank the best model.

For families without a known structure the union of all quality assesment methods were used. A score cutoff at FPR 0.01 and 0.1 was used to estimate how many unknown structures could be predicted accurately (TM-score ≥ 0.5). The result can be seen in table 2.

Updates

With our pipeline the website will be updated with every PFAM release and the currently running update on PFAM 31.0 will introduce additional quality estimators such as PconsC4 and ProQ4 scores.

Topology

The knotted topology was established using an implementation of Alexander polynomial described in the KnotProt database (**?**).

RESULTS

Comparison with other resources

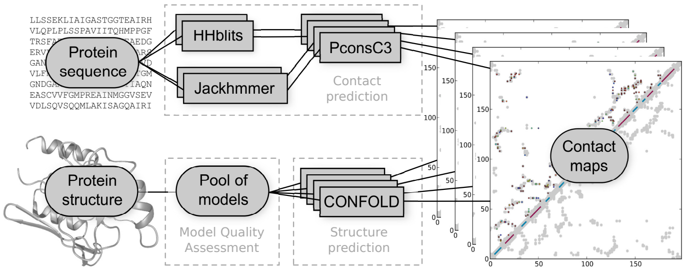

Figure 1. The PconsFold2 pipeline.

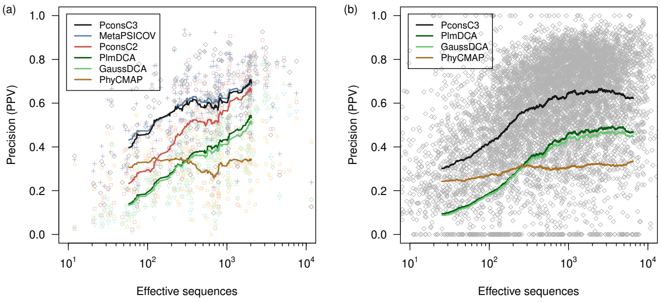

Figure 2. Comparison of contact prediction methods and available resources

	all	mainly-α	mainly-β	$\alpha-\beta$
PconsC3	0.57	0.49	0.59	0.62
MetaPSICOV	0.59	0.54	0.58	0.62
PconsC2	0.48	0.44	0.45	0.51
PlmDCA	0.36	0.34	0.32	0.38
GaussDCA	0.34	0.33	0.31	0.36
PSICOV	0.34	0.30	0.33	0.35
PhyCMAP	0.32	0.23	0.34	0.36
counts	210	55	35	110

Table 2. Average PPV of top $N/2$ predicted contacts on the benchmark dataset for different secondary structural classes.

A similar resource is Baker Labs GREMLIN database (**?**) which builds contact prediction based on PFAM 27.0 (**?**) . However, it has not been updated since 2013. In contrast we plan to update PconsFam within a few month after each Pfam release. This is possible as the pipeline uses a much less computationally costly methodology and is highly modular. Each of the three steps (generate alignments, generate contact maps and generate models) can be changed independently with faster/more accurate tools. PconsC4, which is currently being implemented, is at least one order of magnitude faster than comparative methods and by using CONFOLD the folding is much faster than Rosetta.

GREMLIN shows predicted contact maps with an option to overlay with the pdb structure if one exists. We extend on this by using a tool that can visualize the predicted contacts on the models. In our database, both contact maps and predicted structure can be investigated in detail and downloaded. We have also used a deep learning methodology for contact predictions. In general this provide the better coverage of small protein families.

Table 3. Properties of Pfam families that can be modeled accurately at FPR 0.1. Statistical significant differences from a students t-test at P-values 0.01 and 10^{-5} are marked with * and ** respectively for all columns except the last.

	PconsC3 score	Helix	Sheet	Coil	Meff	Length	Transmembrane	TM-score
Union	0.41**	0.33*	0.18*	0.5	427*	104**	0.11**	0.56
PfamPDB	0.45**	0.34	0.21**	0.46**	1075**	126**	0.04**	0.53
NoPDB	0.32	0.36	0.15	0.50	300	187	0.25	

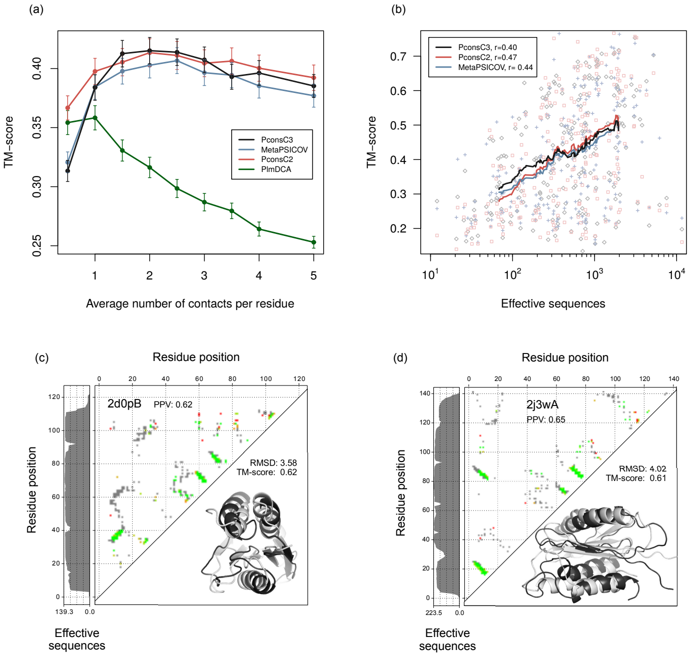

Figure 3. Comparison of contact prediction methods and available resources

Figure 4. Overview of number of structures at different qualities - from PconsFold verview of the database.

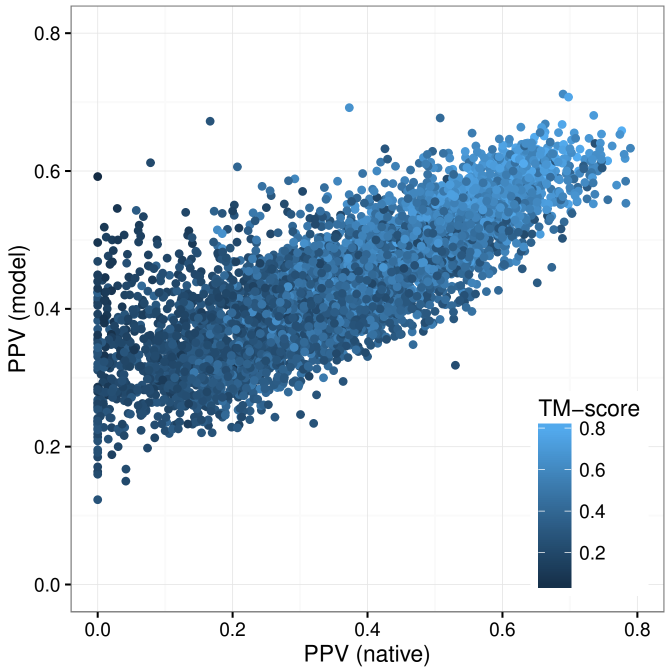

Figure 5. Links Knots - topology.

Figure 6. .

DISCUSSION

In summary the novelties of PconsFam are: An informative user interface that enables examination of multiple models and contact maps. Novel tool to help investigate and interpret contact maps. Using the same pipeline for families with and without known structures. We use model quality estimation methods (ProQ and Pcons) to evaluate FDR PconsFam is complementary to existing resources, i.e. high quality models exist for other families than in Gremlin. PconsFam will be updated at regular times to follow PFAM releases.

Modularity of pipeline As PconsFold2 is highly modular, each of the three steps in the pipeline can be changed independently. Any alignment tool can be used to generate the alignments and any contact prediction tool can be used. This opens up the possibility to run different tools for different dataset where there are known tools that works better for specific data.

Future directions

Currently the input to CONFOLD is both predicted contact maps from PConsC3 and predicted secondary structure from PSIPRED. Both of these are predicted in PConsC4 so a change to PConsC4 will streamline the process even further. PConsC4 is also much faster than the currently implemented PConsC3 and will therefore enable both more frequent updates but also higher quality contact maps.

An improved model quality estimator in ProQ4 is also being worked on to increase or quality assessment methods.

CONCLUSIONS

Here we present an intuative and interactive webinterface for contact maps and models for PFAM-families. We have used the modula PconsFold2 pipeline to predicted both multiple contact maps and multiple models for all legible families and presents both of these both visually with interactive possibilities together with quality assessment scores to highlight the confidence in both contact maps and models. The underlying pipeline is planned to follow each PFAM release and therefore show updated and relevant data.

ACKNOWLEDGEMENTS

Text. Text. Text. Text. Text. Text. Text. Text. Text. Text. Text. Text. Text. Text. Text.

Conflict of interest statement. None declared.