Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Manuscript Category

OXFORD

Subject Section

Large-scale structure prediction enabled by reliable model quality assessment and improved contact predictions for small families.

Mirco Michel [1], David Menéndez Hurtado [1], Karolis Uziela [1] and Arne Elofsson [1,*]

[1] Science for Life Laboratory and Department of Biochemistry and Biophysics, Stockholm University, Stockholm 10691, Sweden

* To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Accurate contact predictions can be used for predicting the structure of proteins. Until recently these methods were limited to very big protein families, decreasing their utility. However, progress in contact prediction has made it possible to predict accurate contact maps for many small families. Here, we ask the question if it is possible to model these families and importantly if we can identify the correct models.

Results: We do find that we can with an estimated 90% (99%) accuracy predict the structure of 445 (97) Pfam families of unkown structure. Out of these 340 (73) have not been reported before in any large scale study of ab-initio predictions, even the one using meta-genomics data indicating that the PconsC3 contact predicion approach used here is complementary to just using more sequence data. We do also report surprisingly many large Pfam families where the structures generated are unlikely to be correct indicating that the properties of the "dark proteome"

Availability: Datasets as well as models of all the 445 Pfam families are available at c3.pcons.net. All programs used here are freely available.

Contact: arne@bioinfo.se

Supplementary information: No supplementary data

1 Introduction

A few years ago maximum entropy methods revolutionized the accuracy of contact predictions in proteins (Weigt *et al*., 2009; Burger and van Nimwegen, 2010; Aurell, 2016). This enabled the prediction of accurate protein models using no information from homologous protein structures (Marks *et al*., 2011; Morcos *et al*., 2011). It has been shown that accurate protein structures can be obtained for soluble proteins (Marks *et al*., 2011), membrane proteins (Nugent and Jones, 2012; Hopf *et al*., 2012; Hayat *et al*., 2015) and even disordered proteins (Toth-Petroczy *et al*., 2016). These methods have also been used to predict interactions between proteins (Weigt *et al*., 2009; Ovchinnikov *et al*., 2014; Hopf *et al*., 2014).

Until recently such methods have been limited to very large protein families (Kamisetty *et al*., 2013; Skwark *et al*., 2014). However, by the inclusion of additional information and improved machine learning methods it is now often possible to obtain accurate contact maps for families as small as a few hundred effective sequences (Michel *et al*., 2017; Jones *et al*., 2015; Wang *et al*., 2017).

Pfam contains today approximately 16,000 protein families that vary in size between a few tens to hundreds of thousands effective sequences. About half (46%) of these protein families contain no representative structure, i.e. there is more than 7,500 protein families without a structure. The families with structure are on average larger than the ones without, median size 680 vs. 134 effective sequences, i.e. most of the families without a structure are too small for maximum entropy contact prediction but might be within reach for methods that combine these with advanced machine learning.

Now, we ask the question how many of these roughly 7,500 protein families without a structure can be modeled reliably by using state of the art contact prediction methods. To the best of our knowledge the largest effort to model protein families was performed by the Baker group who modeled structures for 614 families by including a very large set

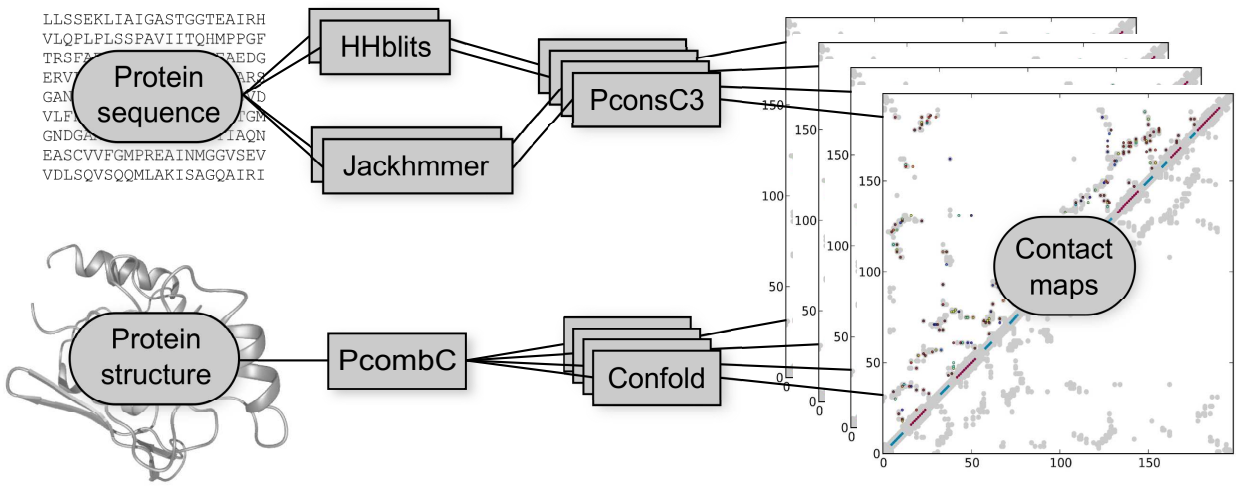

Fig. 1. PconsFold2 workflow. Given an input sequence four alignments are created using HHblits and Jackhmmer at two different E-value thresholds of 1 and 10^{-4}. These alignments are then used input to PconsC3 to generate four contact maps. The 2.5L (L=length of sequence) top ranked contacts are then used by CONFOLD to generate 50 models for each alignment, resulting in 200 models for each query sequence. These models are then ranked by a model quality assessment program.

of sequences from meta-genomics (Ovchinnikov *et al*., 2017). However, their approach for contact prediction was based on a maximum entropy method (Antala *et al*., 2015) and not newer methodsusing deep learning. Here, we propose a method to expand this by thousands of families without using meta-genomics sequences.

The PconsFold pipeline is described in Figure Fig. 1. Given an input protein sequence PconsFold2 generates four alignments. These alignments are then used by PconsC3 (Michel *et al*., 2017) to predict four different contact maps. The 2.5L (L=length of sequence) top ranked contacts are then used to fold the protein. In contrast to PconsFold (Michel *et al*., 2014) we do use CONFOLD (Adhikari *et al*., 2015), i.e. the NMR protocol of CNS (Brunger, 2007) and not ROSETTA (Leaver-Fay *et al*., 2011). This makes the pipeline much faster but possibly slightly less accurate.

CONFOLD is then used to fold 50 models for each of these four contact maps. The final step in the pipeline is the model quality assessment. In addition to earlier methods we here introduce PcombC representing a linear combination of three separate assessment scores: Pcons, ProQ3D, and the agreement between contacts in the model and the underlying predicted contact map (PPV).

First, we show that many protein families can be modeled using our most recent contact prediction method, PconsC3 (Michel *et al*., 2017) and the CONFOLD CNS based pipeline. However, we do find that using this analysis it is difficult to know if a given protein family is modeled correctly or not. Therefore, we utilize our experience in model quality assessment to examine if the accuracy can be improved. We do find that using the best single model quality assessment protocol (ProQ3D (Uziela *et al*., 2017)) we can model about 235 Pfam families at a FPR of 10%.

2 Methods

2.1 Datasets

There are 16,295 protein domain families in Pfam 29.0. Out of these 7733 domains have a known structure with a HHsearch hit in PDB with an E-value of less than 10^{-3} that covers at least 75% of it's representative sequence. The representative sequence of a Pfam domain with known structure is set to be the protein sequence ranked first by the HHsearch (Söding, 2005) run against the PDB database bundled with HHsuite (Meier and Söding, 2015) (date: 2016-09-07)

The test dataset was generated from 626 Pfam domains that were randomly selected from 6925 domains with known structure that are longer than 50 residues. We used a subset of this dataset to optimize parameters for the linear combination of model quality assessment scores and folding (number of models to generate, and whether to use one or multiple alignments).

From the remaining Pfam domains we further excluded all Pfam domains that can be found in the `pdbmap` file from Pfam release 29.0. For this dataset of 7537 Pfam domains with unknown structure, analogously we defined the sequence highest ranked in the HHblits alignment against uniref20 (date: 2016-02-26) to be the reference sequence of the family. We refer to the length of a Pfam family by the length of its representative sequence.

2.2 Alignments

The input to direct coupling analysis (DCA)-based contact prediction methods are multiple sequence alignments. These alignments were generated using both HHblits and Jackhmmer, each at E-value thresholds of 1 and 10^{-4}. HHblits was run against the uniprot20 database from HHsuite (date: 2016-02-26). The parameter `-all` has been used and `-maxfilt` and `-realign_max` were set to 999999 as in (Michel *et al*., 2017). Jackhmmer searches were performed against Uniprot90 (Magrane and Consortium, 2011) and were run for five iterations with both -E and -incE set to the respective E-value cutoffs.

2.3 Contact prediction

To overcome the limit of DCA methods requiring large alignments, PconsC3 combines the results of such methods with contacts predicted by a machine-learning based method (Michel *et al*., 2017). It then uses a similar pattern recognition aproach than PconsC2 to iteratively increase the quality of the predicted contact map. PconsC3 was run as described earlier (Michel *et al*., 2017), however we used all four alignments as inputs predicting one contact map for each alignment. The searches were started from the Pfam representative sequence, i.e. the Pfam alignments were ignored. Contact map quality is measured in positive predictive value (PPV) over the same number of top-ranked contacts that were used during folding (2.5 · sequence length (L)). The average contact score refers to the mean PconsC3 score of these contacts.

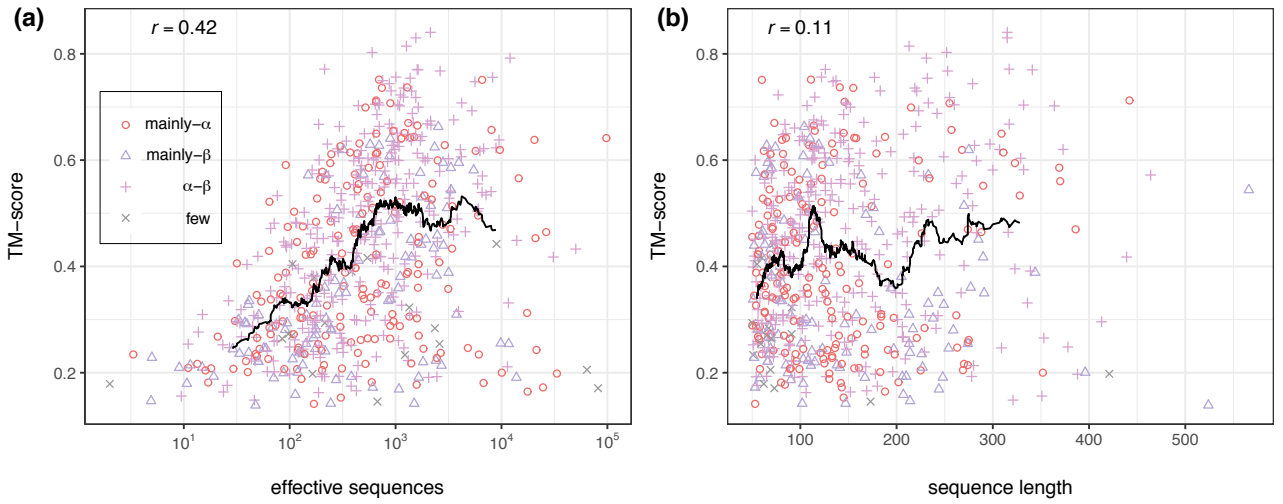

Fig. 2. PconsFold2 model quality (TM-score) on the benchmark dataset of the top-ranked models against (a) effective sequences in the underlying alignment and (b) against the length of the input sequence. Pearson correlation r is shown in the upper left corner, black lines represent moving averages with a window of 40. Colors indicate whether the model has been predicted to be above (blue) or below (red) 0.5 TM-score.

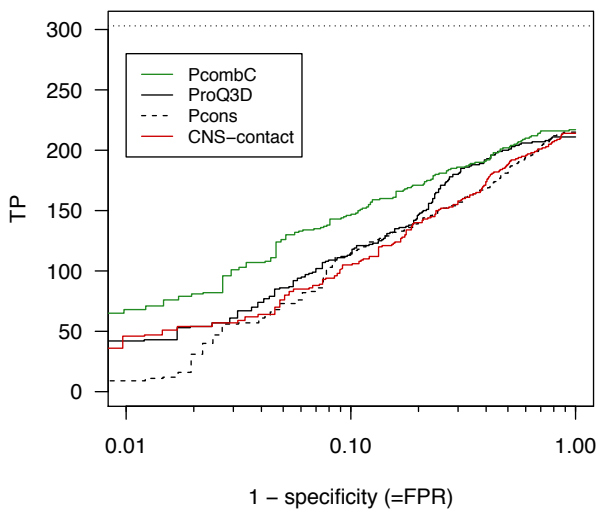

Fig. 3. ROC-like plot for different ways of evaluating and ranking the models in the benchmark dataset. While the x-axis shows true positive rate on a logarithmic scale, the y-axis shows the number of proteins with TM-score > 0.5. The horizontal line indicates the best possible outcome, i.e. the number of families with TM-score> 0.5 when ranking the models by TM-score score.

2.4 Model generation

Contacts predicted by PconsC3 are then applied as distance restraints between the C_β-atoms (C_α in the case of glycine) during protein structure prediction. We use CONFOLD (Adhikari *et al*., 2015) for this task. When folding a protein using CONFOLD a fixed number of contacts are used. Here, contacts are sorted by their PconsC3 score and a threshold is set on the number of top-ranked contacts to use. This threshold is based on the length of the input sequence as an input to CONFOLD, which folds the protein using CNS (Brunger, 2007). For each alignment we generated 50 models resulting in a total of 200 models per Pfam family.

2.5 PcombC

PcombC is a linear combination between the scores of these three methods, similar to what we used in CASP4 (Wallner *et al*., 2003) and CASP5 (Wallner and Elofsson, 2005a).

$$S_{PcombC} = a \cdot S_{Pcons} + b \cdot S_{ProQ3D} + c \cdot PPV$$

Coefficients a, b, and c have been determined using a grid-search on a 10x10x10 grid with values ranging from 0 to 1 and a step size of 0.1, optimizing the area under the ROC-curve for determining whether a model is correct or not (TM-score threshold of 0.5). This gridsearch has been performed on the training dataset. Optimal weights are $a = 0.5$, $b = 0.2$, and $b = 0.9$. In order for S_{PcombC} to remain within the same scale as the input scores, the coefficients have been normalized to:

$$S_{PcombC} = \frac{0.5}{1.6} \cdot S_{Pcons} + \frac{0.2}{1.6} \cdot S_{ProQ3D} + \frac{0.9}{1.6} \cdot PPV$$

PcombC has thus been optimized for discriminating between accurate and inaccurate models. This has the advantage of being able to interpred the predicted score in terms of absolute model quality, enabling statements about the confidence of a predicted model being correct.

2.6 Evaluation

Model quality is measured in template modeling score (TM-score) scores (Zhang and Skolnick, 2004). For the ROC-analysis we set a TM-score threshold of 0.5 to distinguish between correct and incorrect models. Area under the curve (AUC) has been calculated taking the best possible ranking as reference.

2.7 Runtime

The running time of the folding step was measured on a single core of an Intel Xeon E5-2690 v4 processor. For the test dataset it takes around 30 seconds on average to generate one model with a minimum of 2s per model for the shortest family (32 residues) and 245s for the longest (524 residues).

3 Results

3.1 Utilization of predicted contacts

Preliminary data indicated that using a threshold for the number of contacts utilized during folding of 2.5 times the length of the sequence has been optimal. Using this threshold, we investigated the effect of the number of generated models on the quality of the best and top-ranked model, Table 1. For this experiment models were ranked by their CNS-contact score. The quality of top-ranked models does not increase as much as that of the best possible models. However, there is a flattening after around 50 models. For the best possible models there is slight increase in average performance against the default of 20 models from 0.43 to 0.45 TM-score. We thus decided that generating 50 models for a given contact map is a good tradeoff between the quality of the best possible model and running time.

It has previously been observed that the quality of predicted contacts depends on the underlying alignment (Skwark *et al.*, 2013). This in turn results in varying quality of the predicted models based on these contact maps. We therefore investigated whether model quality can be improved by using a set of alignments with varying methods and E-value thresholds instead of a single fixed alignment. Table 1 further compares using either 200 models of an HHblits alignment with an E-value threshold of 1 or when using four alignments with 50 models each (see Methods). The average maximum TM-score improves by 6.5% when using four different alignments and the average top-ranked TM-score by 5%.

Table 1. Best possible (max) and top-ranked average TM-scores on the training dataset for varying number of models on a single alignment (HHblits E-value 1) and for 50 models on all four alignments each (4 x 50).

No. models	TM-score (max)	TM-score (top-ranked)
1	0.36	0.36
5	0.41	0.39
20	0.43	0.40
50	0.45	0.40
100	0.45	0.40
200	0.46	0.40
4 x 50	0.49	0.42

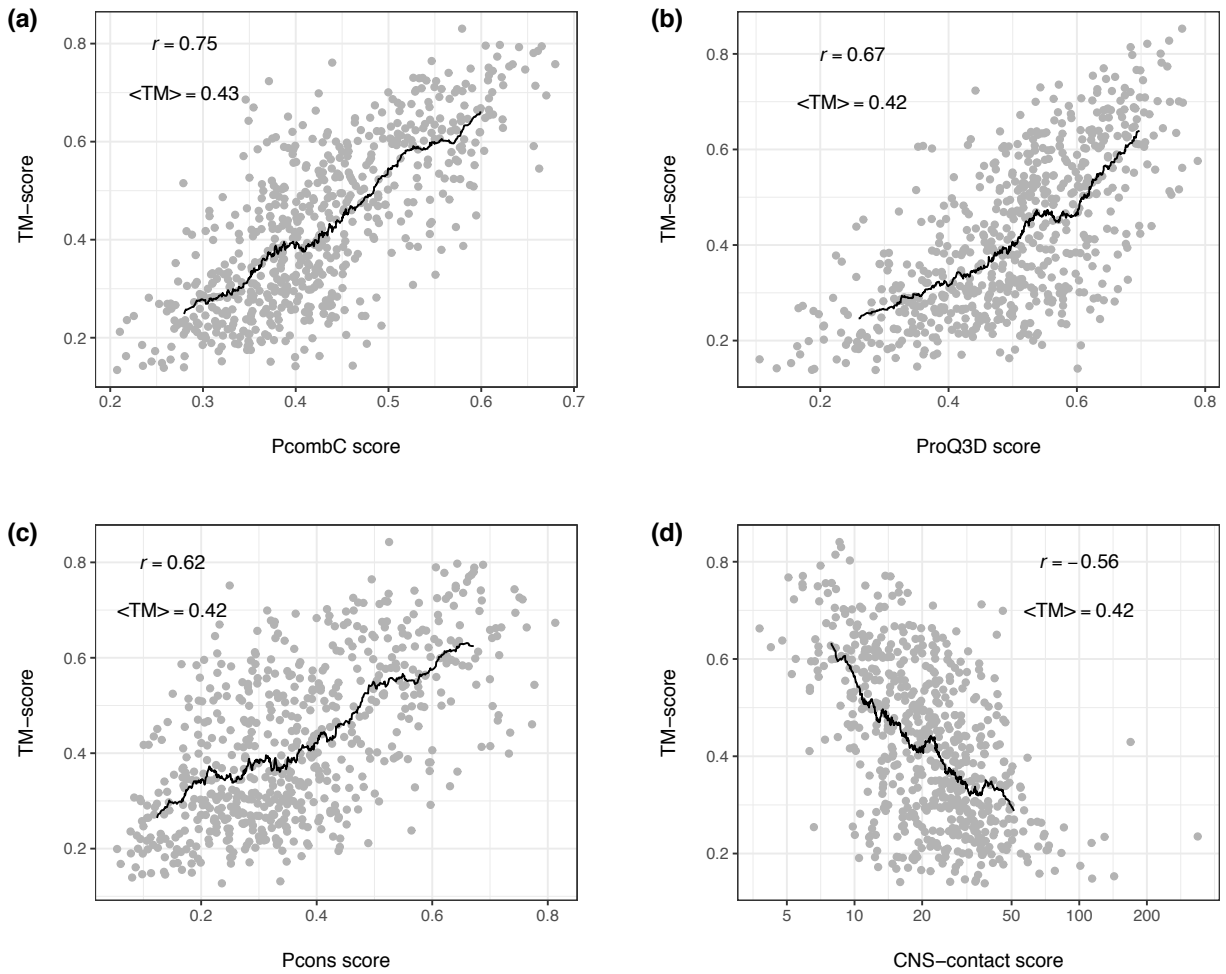

Fig. 4. Model quality as measured by TM-score on the benchmark dataset of the top-ranked models against (a) PcombC score (scoring function of the PconsFold2 pipeline), (b) Pcons score, (c) ProQ3D score, (d) and (d) CNS contact energy normalized by sequence length. Pearson correlations r and average TM-scores (<TM>) are shown, black lines represent moving averages with a window of 40. For CNS-contact the pearson correlation has been calculated on log10(CNS-contact)

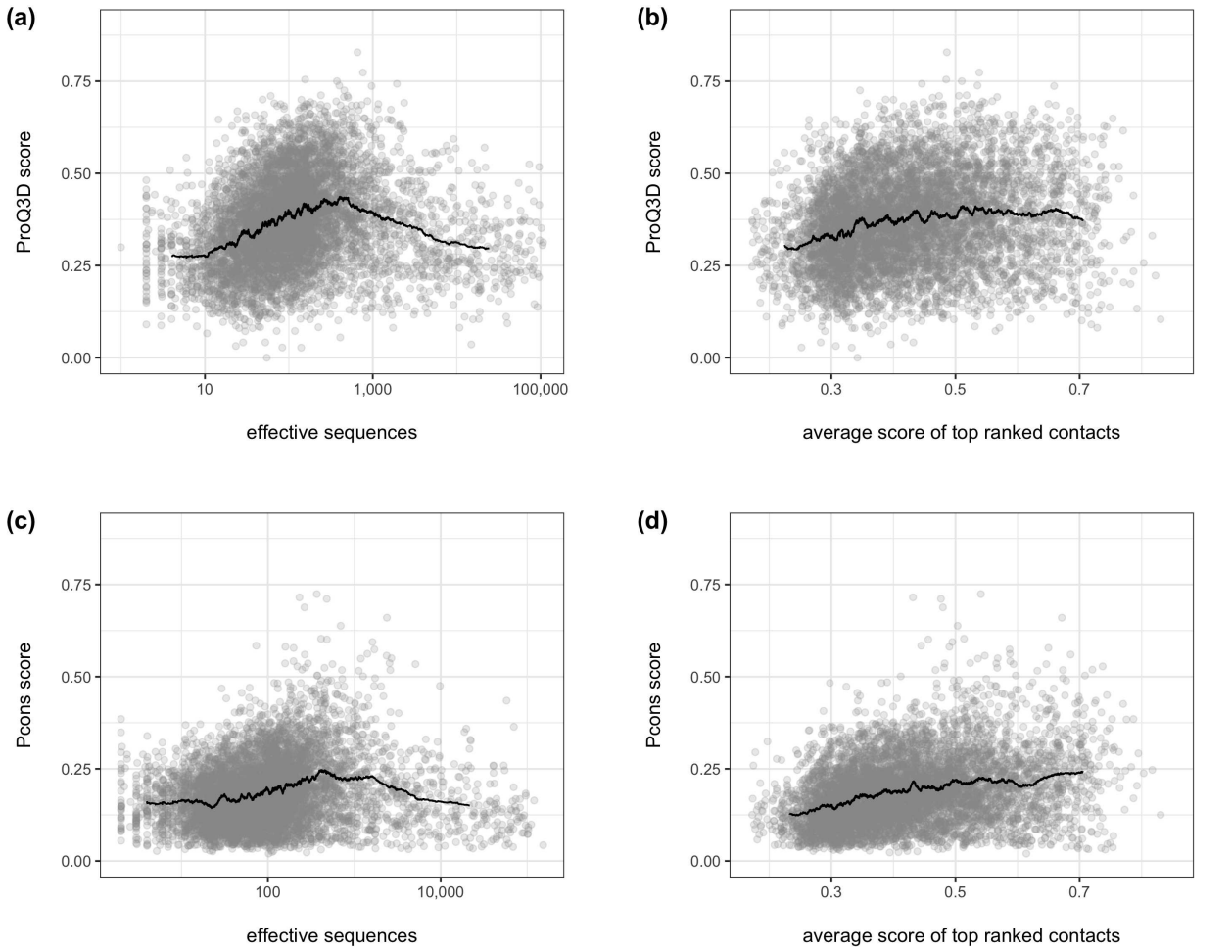

Fig. 5. Pfam .

3.2 PcombC

In order to blindly score and rank that pool of 200 models as accurately as possible, three different model quality assessment methods are run. Pcons predicts a quality score for each model based on a comparison of all models against each other (consensus method) (Wallner and Elofsson, 2005b). ProQ3D is a single model quality assessment method, that looks at one model at a time and assesses its quality using deep learing on various features of the given model (Uziela *et al*., 2017). The third score, PPV, is the contact agreement between the model and the underlying contact map, using the same number of top-ranked contacts as used during folding.

IT CAN BE SEEN THAT PCOMBC is better than XXX

PcombC is a linear combination between the scores of these three methods, similar to what we used in CASP4 (Wallner *et al*., 2003) and CASP5 (Wallner and Elofsson, 2005a). PcombC has thus been optimized for discriminating between accurate and inaccurate models. This has the advantage of being able to interpred the predicted score in terms of absolute model quality, enabling statements about the confidence of a predicted model being correct.

Table 2. ROC analysis when classifying whether a model is correct (TM-score ≥ 0.5) or not.

	# models at FPR		
Method	0.01	0.1	1.0
PcombC	68 (22%)	146 (48%)	217 (72%)
Pcons	9 (3%)	115 (38%)	215 (71%)
ProQ3D	42 (14%)	114 (38%)	211 (70%)
CNS-contact	46 (15%)	105 (35%)	214 (71%)
best possible	303	303	303

3.3 Model quality assessment

Figure 4 shows how the scores of different QA tools predict TM-score. As before, for each family in the test set we ranked all models by each QA score and selected the top ranked model. Pearson correlation (r) is highest for PcombC, followed by Pcons and ProQ3D. PPV correlates equally well with TM-score as CNS-contact does (data not shown)

In order to estimate model quality when there is no known structure available it needs to be predicted as accurately as possible. The goal is not only to select the best models from a set of predictions but also to predict how much these models can be trusted. Figure 3 shows the false positive rate (FPR) for different quality assessment (QA) tools when classifying

predictions into correct (TM-score $\geq$ 0.5) or incorrect (TM-score $\leq$ 0.5) models. For this analysis all models have first been assessed and ranked by the respective QA tool. Then the top-ranked model has been selected and classified whether it is correct or not.

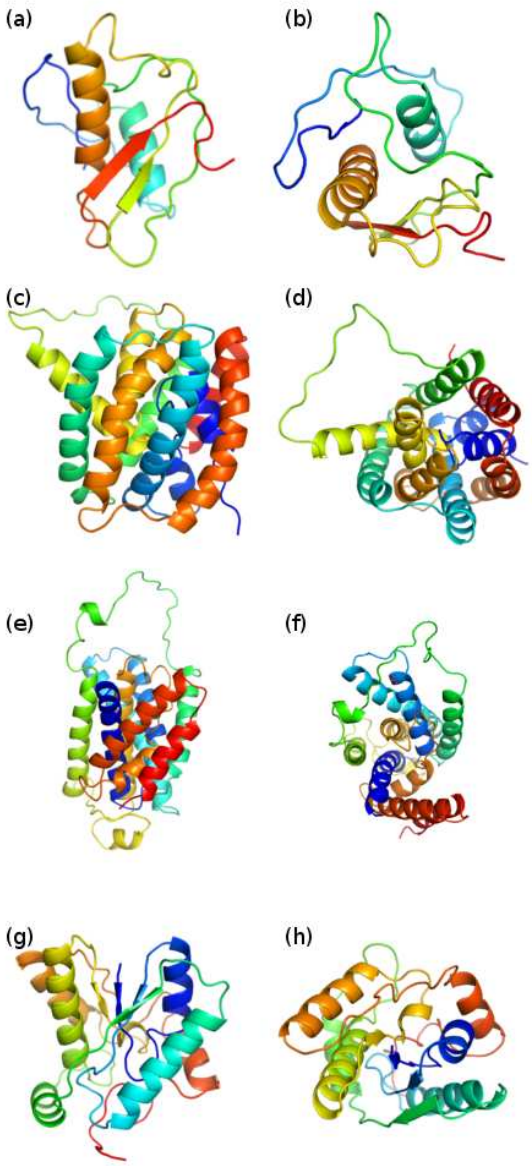

Fig. 6. Models of some Pfam families. (a) and (b) front and top-view of PF10646, (c) and (d) PF01925, (e) and (f) PF02653, and (g) and (h) PF02677

Although the overall number of correct top-ranked models does not change much (number of true positives at FPR = 1.0), there are substantial differences in the ability of the different scores to classify the models. As Table 2 shows, the default ranking from CONFOLD (CNS-contact score) normalized by length performs not very well in assessing the models, although it ranks them comparatively well (total of 108 models above 0.5 TM-score). At an FPR of 0.1 (specificity of 90%) only 46 models are picked up from the 149 correct models (dashed horizontal line in Figure 3). In this scenario ProQ3D selects 58 and Pcons is able to select 64 models correctly. By a margin the largest number of detected models has PcombC with more than 51% of the total correct models at 0.1 FPR.

The results shown here indicate that a combination of QA methods along with contact map agreement provides a significant improvement in detecting correct models over the best single methods. It can be used to reliably predict model accuracy while at the same time being more sensitive than previous methods.

3.4 Overall performance

Figure 2(a) shows the performance of PconsFold2 in TM-score of the top-ranked model against the family size, measured in effective sequences. Generally, model accuracy depends on logarithmic family size with a correlation of 0.54. The color indicates whether a protein was predicted to be correct or not, i.e. if the PcombC score used to rank the models in PconsFold2 was above or below the 0.1 FPR cutoff. In the region of TM-score 0.5-0.6 the top ranked model for about half of the proteins is corretly predicted to be accurate. Above TM-score 0.6 nearly all proteins are correctly identified.

There are however 5 cases where PconsFold2 predicts the model to be correct whereas the TM-score is below 0.4. All of these proteins are relatively short with lengths ranging from 56 to 109 residues. Visual inspection shows that they contain beta sheets where some of the strands are swapped. Whether this causes the quality assessment tools to overestimate its quality remains to be determined.

Fig. 2(a) further reveals that for families smaller than 100 effective sequences, structure prediction has a success rate (percentage of models that are correctly identified to be above 0.5 TM-score) of only 2%. As soon as the family becomes larger than 100 effective sequences the number of correct models increases rapidly. The success rate is 23% for families between 100 and 1000 effective sequences and 53% for large families with more than 1000 effective members.

PconsFold2 model quality seems largely independent of the length of the input sequence as Fig 2(b) shows. The fraction of correctly identified models decreases slightly towards longer proteins.

4 Discussion

Table 3. Number of Pfam families with unknown structure that can be modelled at 1% and 10% FPR of which the overlap with the Baker studies are given by number in parenthesis. TM-score columns show the average TM-score between the models of this study and Ovchinnikov et al. (2017) for all overlapping models at 0.1 and 0.01 FPR, respecively.

	0.01	TM-score	0.1	TM-score
ProQ3D	36 (10)	0.53	235 (77)	0.53
PcombC	13 (8)	0.70	44 (23)	0.62
Pcons	2 (1)	0.69	44 (25)	0.61
CNS-contact	62 (13)	0.54	231 (37)	0.51
Union	97 (24)	0.52	445 (104)	0.50
All	6383	558	6383	558

We are currently running all evaluations. Unfortunately all estimates of the PcombC results are not done at this deadline, therefore we can only estimate the number of correctly modelled Pfam domains from the small test-set. Using this estimate of the 8562 protein families in Pfam without a known structure we estimate that we can model about 2000, see Table **??**. This corresponds to 25% of all Pfam families without a known structure.

Alternatively we can estimate the number of accurate models using Pcons. A Pcons cutoff of 0.36 that should correspond to a FPR of 10% (TM-score$<$ 0.5. This corresponds to 388 pfam families out of 6754 for

which we ran Pcons, see Figure 5. Out of the 388 families 124 are in common with the recent study by Ovchinnikov *et al*. (2017), that leaves 264 novel families.

Figure 6 shows a few examples of the models top-ranked by Pcons where Pcons predicts the model to be accurate. PF10646 is the GerMN domain of the Bacillus GerM protein involved in sporulation and spore germination. PF01925 represents a sulfite exporter TauE/SafE transmembrane protein. PF02653 contains branched-chain amino acid transport system proteins and PF02677 is still completely uncharacterized and marked in Pfam as a domain of unknown function. Out of these examples only PF10646 has been predicted in Ovchinnikov *et al*. (2017). All predictions including their quality scores will be provided on c3.pcons.net.

5 Conclusion

Here, we estimate that between 400 and 2000 Pfam families can be modeled using contacts predicted from PconsC3 at a false positive rate of 10%. This is made possible by using a combination of improved contact predictions in PconsC3 and accurate model quality assessment methods. This number is obtained without the use of meta-genomic data, and might increase significantly if such sequences were included. Our approach provides a significant increase in the structural coverage available today. By visual inspections the models appear feasible, see Figure 6.

Acknowledgements

We thank Björn Wallner, Christian Blau, and the entire CASP community for valuable discussions.

Funding

This work has been supported by the grants from the Swedish Research Council (VR-NT 2012-5046) and Swedish e-Science Research Center. Computational resources were provided by the Swedish National Infrastructure for Computing (SNIC) at NSC, HPC2N and Uppmax.

References

Adhikari, B., Bhattacharya, D., Cao, R., and Cheng, J. (2015). Confold: Residue-residue contact-guided ab initio protein folding. *Proteins: Structure, Function, and Bioinformatics*, **83**(8), 1436–1449.

Antala, S., Ovchinnikov, S., Kamisetty, H., Baker, D., and Dempski, R. (2015). Computation and functional studies provide a model for the structure of the zinc transporter hZIP4. *J Biol Chem*, **290**(29), 17796–17805.

Aurell, E. (2016). The maximum entropy fallacy redux? *PLoS Comput Biol*, **12**(5), e1004777.

Brunger, A. (2007). Version 1.2 of the crystallography and NMR system. *Nat Protoc*, **2**(11), 2728–2733.

Burger, L. and van Nimwegen, E. (2010). Disentangling direct from indirect co-evolution of residues in protein alignments. *PLoS computational biology*, **6**(1), e1000633.

Hayat, S., Sander, C., Marks, D. S., and Elofsson, A. (2015). All-atom 3D structure prediction of transmembrane β-barrel proteins from sequences. *Proceedings of the National Academy of Sciences of the United States of America*, **112**(17), 5413–8.

Hopf, T., Scharfe, C., Rodrigues, J., Green, A., Kohlbacher, O., Sander, C., Bonvin, A., and Marks, D. (2014). Sequence co-evolution gives 3d contacts and structures of protein complexes. *Elife*, **3**.

Hopf, T. A., Colwell, L. J., Sheridan, R., Rost, B., Sander, C., and Marks, D. S. (2012). Three-dimensional structures of membrane proteins from genomic sequencing. *Cell*, **149**(7), 1607–1621.

Jones, D. T., Singh, T., Kosciolek, T., and Tetchner, S. (2015). MetaPSICOV: Combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins. *Bioinformatics*, **31**(7).

Kamisetty, H., Ovchinnikov, S., and Baker, D. (2013). Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era. *Proceedings of the National Academy of Sciences of the United States of America*, **110**(39), 15674–15679.

Leaver-Fay, A., Tyka, M., Lewis, S. M., Lange, O. F., Thompson, J., Jacak, R., Kaufman, K., Renfrew, P. D., Smith, C. A., Sheffler, W., Davis, I. W., Cooper, S., Treuille, A., Mandell, D. J., Richter, F., Ban, Y.-E. A., Fleishman, S. J., Corn, J. E., Kim, D. E., Lyskov, S., Berrondo, M., Mentzer, S., Popović, Z., Havranek, J. J., Karanicolas, J., Das, R., Meiler, J., Kortemme, T., Gray, J. J., Kuhlman, B., Baker, D., and Bradley, P. (2011). ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. *Methods in enzymology*, **487**, 545–574.

Magrane, M. and Consortium, U. (2011). UniProt Knowledgebase: a hub of integrated protein data. *Database*, **2011**(0), bar009.

Marks, D. S., Colwell, L. J., Sheridan, R., Hopf, T. A., Pagnani, A., Zecchina, R., and Sander, C. (2011). Protein 3D structure computed from evolutionary sequence variation. *PloS one*, **6**(12), e28766.

Meier, A. and Söding, J. (2015). Automatic prediction of protein 3d structures by probabilistic multi-template homology modeling. *PLoS Comput Biol*, **11**(10), e1004343.

Michel, M., Hayat, S., Skwark, M., Sander, C., Marks, D., and Elofsson, A. (2014). Pconsfold: improved contact predictions improve protein models. *Bioinformatics*, **30**(17), i482–8.

Michel, M., Skark, M. J., Menéndez-Hurtado, D., Ekeberg, M., and Elofsson, A. (2017). Predicting accurate contacts in thousands of pfam domain families using pconsc3. *Submitted*.

Morcos, F., Pagnani, A., Lunt, B., Bertolino, A., Marks, D., Sander, C., Zecchina, R., Onuchic, J., Hwa, T., and Weigt, M. (2011). Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc Natl Acad Sci U S A*, **108**(49), 1293–301.

Nugent, T. and Jones, D. T. (2012). Accurate de novo structure prediction of large transmembrane protein domains using fragment-assembly and correlated mutation analysis. *Proceedings of the National Academy of Sciences of the United States of America*, **109**(24), 1540–1547.

Ovchinnikov, S., Kamisetty, H., and Baker, D. (2014). Robust and accurate prediction of residue-residue interactions across protein interfaces using evolutionary information. *Elife*, **3**, e02030.

Ovchinnikov, S., Park, H., Varghese, N., Huang, P.-S., Pavlopoulos, G. A., Kim, D. E., Kamisetty, H., Kyrpides, N. C., and Baker, D. (2017). Protein structure determination using metagenome sequence data. *Science*, **355**(6322), 294–298.

Skwark, M. J., Abdel-Rehim, A., and Elofsson, A. (2013). PconsC: Combination of direct information methods and alignments improves contact prediction. *Bioinformatics*, **29**(14), 1815–1816.

Table 4. Properties of Pfam families that can be modelled accurately at FPR 0.1. Score is the average quality assessment score of the respective method, PconsC3 score the average score of the contacts used to fold the model; helix, sheet, and coil represent percentages of predicted secondary structural elements; transmembrane denotes the fraction of predicted transmembrane proteins. The union combines the non-overlapping families that are predicted by the four quality assessment methods. 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 first.

	score	PconsC3 score	helix	sheet	coil	Meff	L	transmembrane
ProQ3D	0.64	0.5**	0.47**	0.1**	0.42**	375.72**	102.6**	0.07**
PcombC	0.55	0.62**	0.35	0.2	0.46	2218.11	99.27**	0.09*
Pcons	0.55	0.55**	0.51**	0.11	0.38**	877.15	107.55**	0.18
CNS-contact	10.47	0.59**	0.08**	0.27**	0.65**	3960.06*	101.29**	0.01**
Union	0.29	0.54**	0.29**	0.18*	0.53*	2203.32	103.65**	0.06**
All	–	0.42	0.36	0.15	0.5	1289.9	186.85	0.25

Skwark, M. J., Raimondi, D., Michel, M., and Elofsson, A. (2014). Improved Contact Predictions Using the Recognition of Protein Like Contact Patterns. *PLoS Computational Biology*, **10**(11).

Söding, J. (2005). Protein homology detection by HMM-HMM comparison. *Bioinformatics*, **21**(7), 951–960.

Toth-Petroczy, A., Palmedo, P., Ingraham, J., Hopf, T., Berger, B., Sander, C., and Marks, D. (2016). Structured States of Disordered Proteins from Genomic Sequences. *Cell*, **167**(1), 158–170.e12.

Uziela, K., Menéndez Hurtado, D., Shu, N., Wallner, B., and Elofsson, A. (2017). ProQ3D: Improved model quality assessments using Deep Learning. *Bioinformatics*, page btw819.

Wallner, B. and Elofsson, A. (2005a). Pcons5: combining consensus, structural evaluation and fold recognition scores. *Bioinformatics*, **21**(23), 4248–4254.

Wallner, B. and Elofsson, A. (2005b). Pcons5: combining consensus, structural evaluation and fold recognition scores. *Bioinformatics (Oxford, England)*, **21**(23), 4248–4254.

Wallner, B., Fang, H., and Elofsson, A. (2003). Automatic consensus-based fold recognition using pcons, proq, and pmodeller. *Proteins*, **53 Suppl 6**, 534–541.

Wang, S., Sun, S., Li, Z., Zhang, R., and Xu, J. (2017). Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model. *PLOS Computational Biology*, **13**(1), e1005324.

Weigt, M., White, R. A., Szurmant, H., Hoch, J. A., and Hwa, T. (2009). Identification of direct residue contacts in protein-protein interaction by message passing. *Proceedings of the National Academy of Sciences of the United States of America*, **106**(1), 67–72.

Zhang, Y. and Skolnick, J. (2004). Scoring function for automated assessment of protein structure template quality. *Proteins*, **57**(4), 702–710.