BIOINFORMATICS

Vol. 30 ECCB 2014, pages i482–i488
doi:10.1093/bioinformatics/btu458

# PconsFold: improved contact predictions improve protein models

Mirco Michel[1,2], Sikander Hayat[3], Marcin J. Skwark[4], Chris Sander[5], Debora S. Marks[3] and Arne Elofsson[1,2,*]

[1]Department of Biochemistry and Biophysics, Stockholm University, 10691 Stockholm, Sweden, [2]Science for Life Laboratory, Stockholm University, Box 1031, 17121 Solna, Sweden, [3]Department of Systems Biology, Harvard Medical School, Boston, MA, USA, [4]Department of Information and Computer Science, Aalto University, PO Box 15400, FI-00076 Aalto, Finland and [5]Computational Biology, Memorial Sloan-Kettering Cancer Center, New York, NY, USA

**ABSTRACT**

**Motivation:** Recently it has been shown that the quality of protein contact prediction from evolutionary information can be improved significantly if direct and indirect information is separated. Given sufficiently large protein families, the contact predictions contain sufficient information to predict the structure of many protein families. However, since the first studies contact prediction methods have improved. Here, we ask how much the final models are improved if improved contact predictions are used.

**Results:** In a small benchmark of 15 proteins, we show that the TM-scores of top-ranked models are improved by on average 33% using PconsFold compared with the original version of EVfold. In a larger benchmark, we find that the quality is improved with 15–30% when using PconsC in comparison with earlier contact prediction methods. Further, using Rosetta instead of CNS does not significantly improve global model accuracy, but the chemistry of models generated with Rosetta is improved.

**Availability:** PconsFold is a fully automated pipeline for *ab initio* protein structure prediction based on evolutionary information. PconsFold is based on PconsC contact prediction and uses the Rosetta folding protocol. Due to its modularity, the contact prediction tool can be easily exchanged. The source code of PconsFold is available on GitHub at https://www.github.com/ElofssonLab/pcons-fold under the MIT license. PconsC is available from http://c.pcons.net/.

**Contact:** arne@bioinfo.se

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

## 1 INTRODUCTION

The protein folding problem is one of the longest standing problems in structural biology. Although the problem of physically folding up a single protein chain in the computer remains largely unsolved, there have been continuous effort and progress, resulting in increased accuracy of predicted models (Kryshtafovych *et al.*, 2013).

The idea of using residue–residue contacts predicted from analysis of correlated mutations observed in evolution for 3D protein structure prediction is not new (Gbel *et al.*, 1994; Hatrick and Taylor, 1994; Neher, 1994; Shindyalov *et al.*, 1994; Vendruscolo *et al.*, 1997). However, until recently contacts predicted from multiple sequence alignments were not sufficiently accurate to facilitate structure prediction methods significantly (Marks *et al.*, 2012). This only became possible due to new statistical approaches to separate direct from indirect contact information (Burger and van Nimwegen, 2010; Lapedes *et al.*, 1999, 2012; Marks *et al.*, 2011; Morcos *et al.*, 2011; Weigt *et al.*, 2009) as well as a greatly increased corpus of sequence information. These efforts came to completion with the first demonstration of successful computation of correct folds with explicit atomic coordinates using maximum-entropy derived contacts (Marks *et al.*, 2011). Analysis of the relative contribution of secondary structure and co-evolutionary information pointed to potential improvements in 3D accuracy (Sulkowska *et al.*, 2012). Since then there has also been continuous effort to improve the quality of predicted contacts (Ekeberg *et al.*, 2013; Jones *et al.*, 2012; Skwark *et al.*, 2013).

In addition to the initial predictions of soluble proteins (Marks *et al.*, 2011) and protein complexes (Schug *et al.*, 2009), contacts predicted from evolutionary information have also been applied in structure prediction of membrane proteins (Hopf *et al.*, 2012; Nugent and Jones, 2012).

To optimize protein structure prediction from predicted contacts, we developed PconsFold, a pipeline for *ab initio* protein structure prediction of single-domain proteins, see Figure 1. PconsFold is based on predicted amino acid contacts from PconsC. These contacts are used within Rosetta to fold a given protein sequence from scratch. We benchmark our method on two datasets and compare it with the CNS (Brunger, 2007) protocol used in EVfold (Marks *et al.*, 2011). It was found that the improved quality of predicted contacts by PconsC (Skwark *et al.*, 2013) increases quality and native-likeliness of predicted structures by about 33% over EVfold and 16% over EVfold-PLM, that is using the improved contact predictions from plmDCA (Ekeberg *et al.*, 2013).

## 2 METHODS

### 2.1 Datasets

During the development of PconsFold, we used the proteins from Jones *et al.* (2012) (PSICOV dataset). The dataset consists of 150 single-domain proteins with sequence lengths between 52 and 266 amino acids. A list of all Protein Data Bank (PDB) (Berman *et al.*, 2000) and corresponding Uniprot (Magrane and Uniprot Consortium, 2011) IDs are given in Supplementary Table S1.

For each protein, we chose to use the PDB seqres sequence as input and not the atom sequence. This avoids internal gaps due to missing residues in the crystal structure. Uniprot sequences were not directly used to avoid including mutations and other sequential differences with the PDB sequences. This also allows direct comparisons between predicted models and native PDB structures.

As an additional comparison between PconsFold and EVfold, we used the set of 15 proteins as in Marks *et al.* (2011). PDB and Uniprot IDs and

*To whom correspondence should be addressed.

© The Author 2014. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/3.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Downloaded from http://bioinformatics.oxfordjournals.org/ at INRIA Paris on September 7, 2014

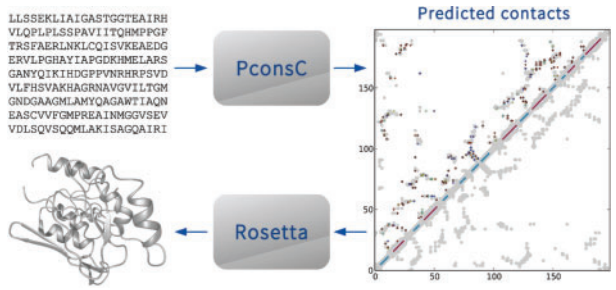



**Fig. 1.** PconsFold pipeline. Based on a given protein sequence, amino acid contacts are predicted with PconsC. These contacts then facilitate protein folding with Rosetta. In the end, PconsFold outputs a structural model for the given sequence

the sequences are given in Supplementary Table S2. According to the Uniprot IDs, we extracted the sequences from the Pfam alignments that were used in the publication. These sequences were submitted to the EVfold web server and used as input for PconsFold.

## 2.2 Contact prediction

In PconsFold, residue contacts are predicted using PconsC (Skwark *et al.*, 2013). The contact predictor plmDCA (Ekeberg *et al.*, 2013) is used by Rosetta/plmDCA and in EVfold-PLM. Rosetta/plmDCA uses the direct output from plmDCA, while in EVfold predicted contacts are further optimized according to Marks *et al.* (2011). PSICOV (Jones *et al.*, 2012) was used in Rosetta/PSICOV.

Predicted contacts were ranked according to the confidence score assigned by the respective contact prediction method. For each protein, we selected $n = f \cdot l$ top-ranked contacts, where $l$ represents the length of the protein sequence and $f$ a factor to scale $n$ relative to sequence length. Consider $f$ is set to 1.0. A protein with length 100 amino acids will thus be constrained by the first 100 predicted contacts.

## 2.3 Rosetta

In PconsFold, Rosetta/plmDCA and Rosetta/PSICOV, we apply the AbinitioRelax folding protocol (Rohl *et al.*, 2004) of Rosetta in version 2013wk42 (Leaver-Fay *et al.*, 2011). The file `abinitio.options_tmpl` in the folder `pcons-fold/folding/rosetta` of the GitHub repository lists all options we are using with AbinitioRelax.

We employ the function *FADE* to integrate predicted residue contacts into the internal scoring function of Rosetta. *FADE* calculates the energy of a given contact as a function of the distance $d$ between its residues as given by the following equation:

$$FADE(d) = \begin{cases} 0.0 & \text{for } d < lb \text{ or } d > ub \\ -2\left(\frac{d-lf}{z}\right)^3 - 3\left(\frac{d-lf}{z}\right)^2 + 1 & \text{for } d < lf \\ 2\left(\frac{d-uf}{z}\right)^3 - 3\left(\frac{d-uf}{z}\right)^2 + 1 & \text{for } d < uf \\ w & \text{otherwise.} \end{cases} \quad (1)$$

The parameters $lb$ and $ub$ are lower and upper bounds, $z$ represents the fading zone's width, $lf$ and $uf$ denote the inner boundaries for the fading zone $lb + z$ and $ub - z$, respectively. The well depth of the interval between both inner boundaries is given by $w$. We set $lb$ to –10 Å, $ub$ to 19 Å and $z$ to 10 Å. This defines a contact to be fully formed if the participating residues are within 9 Å of each other. The fading zones allow for a soft margin between formed and non-formed contacts. In terms of energy, all non-formed contacts are ignored. In our opinion, this accounts best for the fuzzy nature of predicted contacts. The fading zone at the lower bound allows Rosetta to detect and resolve overlapping residues, i.e. when there is a negative distance between two residues.

To avoid the inclusion of homologous fragments into Rosetta, the `-nohoms` flag was used during fragment picking. This means that only fragments from non-homologous protein structures were selected. This was done to simulate a real application case and not to overestimate prediction performance. If not stated otherwise, we used all 150 proteins of the PSICOV dataset as prediction targets in this section.

## 2.4 Folding with CNS using EVfold-PLM

On the PSICOV dataset, EVfold-PLM was run in a standalone version. The alignment E-value cutoff was fixed to $10^{-4}$. The parameter $m$ was set to 0.9. This defines a threshold to exclude sequences from the alignment that consist of >90% of gaps. Using just one E-value for the alignment enabled comparison across the methods, but it should be noted that the EVfold server offers optimization of the E-values based on sequence coverage and number of sequences found, which results in improved results in some cases.

EVfold-PLM was run with default parameters on the small test set using the web interface available at http://evfold.org/. We ensured that it uses pseudolikelihood maximization (plm) on all data, instead of the naïve mean field approach for contact prediction.

Folding with EVfold-PLM was performed using CNS and the same protocol as described before (Marks *et al.*, 2011). Here, the distance geometry protocol is initially used and followed by a short simulated annealing. The CNS-based folding protocol used in EVfold is approximately one order of magnitude faster than the Rosetta protocol used in PconsFold. In addition, PconsC is at least four times slower than just running the contact predictions used in EVfold-PLM.

## 2.5 Identification of top-ranked model

In addition to internal Rosetta scoring, we assessed the quality of all predicted models with the model quality assessment programs (MQAPs) Pcons (Lundström *et al.*, 2001), ProQ2 (Ray *et al.*, 2012) and DOPE from the Modeller software package (Eswar *et al.*, 2006). Pcons uses a comparative approach and ranks all decoys according to pairwise structural similarity between them. ProQ2 and DOPE assess single proteins and evaluate structural features, such as side-chain placement and overall shape. For our ranking, we use the global score each MQAP assigns to a predicted structural model (decoy). Residue-wise information about local model quality is thus not used.

For each structure prediction method, all decoys were re-ranked with each MQAP. The top-ranked model was then selected and compared with its native structure. For this comparison, we used the TM-score (Zhang and Skolnick, 2004) as a measure of structural similarity. As native structure, we used the PDB structure of each protein without further loop closing or other refinement. Residues present in seqres but not observed in the crystal structure, although modelled, are therefore ignored in the structural comparison.

The positive predictive value (PPV) was calculated to assess the quality of predicted models. It indicates how well a given contact map fits a given structure. All PPV values were calculated using reference contacts $C^{\beta}$ ($C^{\alpha}$ in case of glycine) distances in the structure (native or model) with a cutoff at 8 Å.

## 2.6 Model quality assessment

Accurate stereochemistry becomes important when the predicted models are of correct overall fold. Therefore, we used MolProbity (Chen *et al.*, 2010) to assess the chemical model quality. The MolProbity tool `oneline-analysis` was used to detect clashes and to evaluate backbone dihedrals as well as side-chain rotamers. To prepare all structures for this analysis, we first relaxed them with fixed backbone atoms. The Rosetta

Downloaded from http://bioinformatics.oxfordjournals.org/ at INRIA Paris on September 7, 2014

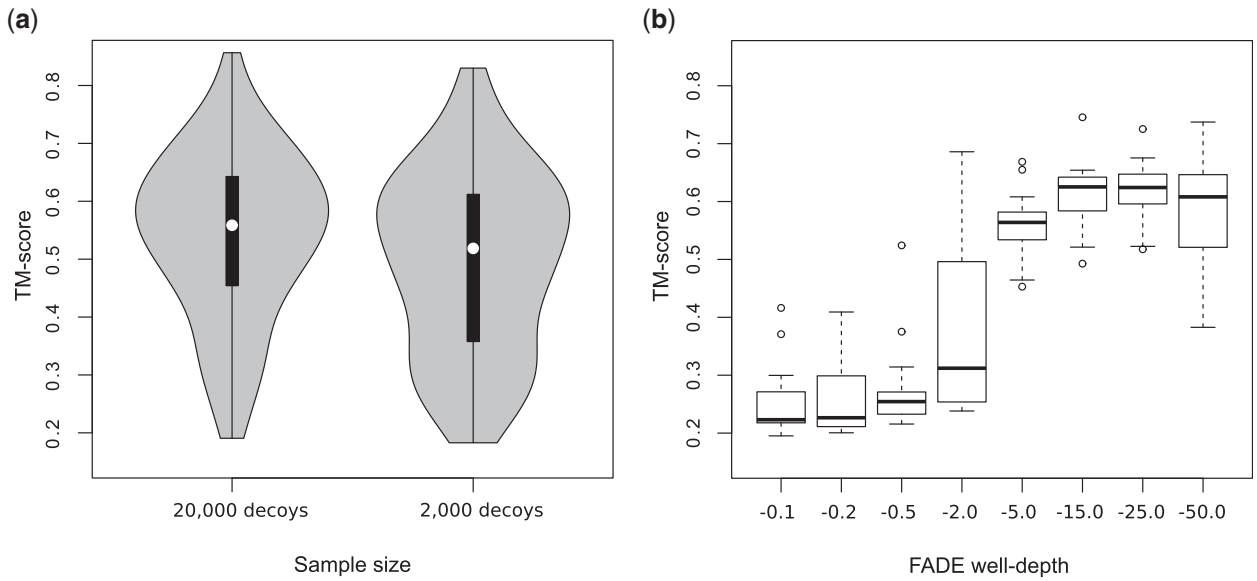



**Fig. 2.** Model quality in TM-score for adjustments of two different Rosetta parameters. (**a**) Performance distributions for two different sample sizes of 20 000 (left) and 2000 (right) decoy structures. The black boxes indicate upper and lower quartile with white dots at the median of the distributions. For each protein in the full PSICOV dataset the top-ranked model was selected from the decoys by its Rosetta score and compared with the native structure. (**b**) Effects of adjustments to the well-depth parameter of the FADE function. A low absolute well-depth (left side) puts low weight on predicted constraints. Constraints are stronger weighted by higher absolute values of well-depth (right side). A subset of 14 proteins of the PSICOV dataset was used here

protocol `relax.linuxgccrelease` was used with the flags `-relax:quick -in:file:fullatom -constrain_relax_to_start_coords`. Hydrogen atoms were then added with the MolProbity tool `reduce-build`. All resulting structures are provided in the Supplementary Material.

### 2.7 Running time

A reduction from 20 000 to 2000 decoy structures corresponds to a 10-fold decrease of runtime during the Rosetta folding step but also leads to an expected decline in average model quality, see Figure 2a. On one core of an Intel E5-2660 (2.2 GHz Sandybridge), the calculation of one decoy takes between 1 and 10 min for the shortest and longest protein in the PSICOV dataset, respectively. To compute 2000 decoys for each protein, we ran 16 threads in parallel. Each of these threads then generates 125 decoys, starting with independent random seeds. With this setting it takes between two hours and nearly one day to generate all decoys for one protein. This simple scaling is possible, as Rosetta runs are independent of each other as long as they are based on independent random seeds. Increasing the number of threads can then be used to reduce overall runtime or to generate more decoys within a given time span.

## 3 RESULTS AND DISCUSSION

### 3.1 PconsFold

For all results in this section, we used the internal Rosetta score to rank all predicted decoys. The top-ranked decoy structure was then selected as the final model.

Previous studies have shown that 20 000–200 000 decoys are necessary to sample native-like conformations without using spatial constraints (Simons *et al.*, 1999). We set out using 20 000 models and then reduced the number of decoy structures to 2000. The bulk at around 0.3 TM-score observed for when only 2000 models are generated represents an increased number of low quality models, see Figure 2a. In general, the practical advantages of massively shorter Rosetta runs outweigh slightly worse predicted models. This parameter is therefore set to a default value of 2000 but can be specified by the user via a command line argument. All further results in this section are generated with this default setting, i.e. using 2000 models.

We selected the FADE function to incorporate predicted residue contacts into Rosetta's native energy function. We then optimized the parameter $w$ (well-depth) of FADE (Equation 1). A subset of 14 proteins of the PSICOV dataset, see Supplementary Table S1, was used to reduce CPU hours of this step. The resulting TM-scores are shown in Figure 2b. The energy term from spatial constraints diminishes with well-depth values of –0.5 and above. This leads to a significant decrease in model quality. Stronger weights than –5.0 put a larger absolute weight on the constraints resulting in higher model qualities. This trend can also be observed when looking at the full dataset. The average TM-score decreases to 0.28 using a well-depth of –1.0. At the end, we choose to use a value of –15.0 to achieve high model quality without outweighing Rosetta's internal energy function completely.

With any number of contacts used during structure prediction, the resulting model quality is on average higher than it would be without contact information. Figure 3a shows that predicted contacts generally improve model quality, regardless of method or amount of contacts.

In addition, improvements in contact prediction methods further increase the quality of predicted structures. With an average TM-score of 0.55, PconsFold, i.e. using PconsC contacts, provides a 10% improvement over using PSICOV or plmDCA contacts. This observation is consistent with a direct comparison of contact prediction methods as in Skwark *et al.* (2013).



Downloaded from http://bioinformatics.oxfordjournals.org/ at INRIA Paris on September 7, 2014

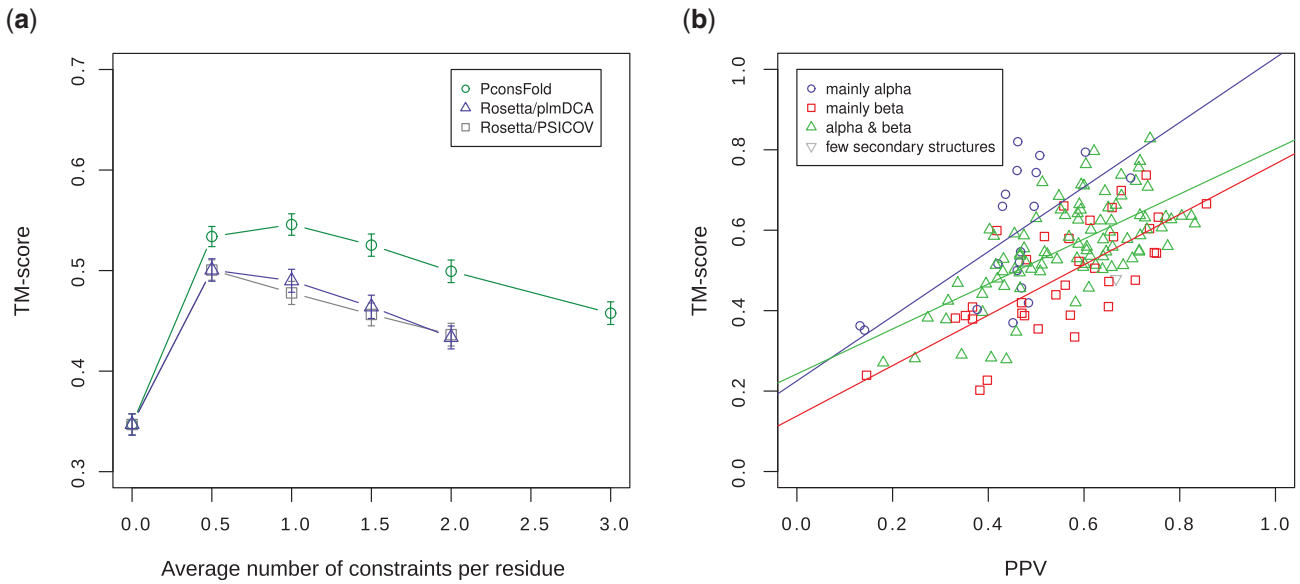



**Fig. 3.** Folding performance on the full PSICOV dataset. (**a**) The number of contacts used in structure prediction is plotted against average TM-score for three different methods: PconsFold (green circles), Rosetta/plmDCA (blue triangles) and Rosetta/PSICOV (black squares). For each protein, the number of top-ranked contacts was selected relative to its sequence length. A value of 1.0 on the *x*-axis represents one contact per residue on average. Error bars indicate standard errors. (**b**) TM-scores are compared with the PPV of underlying contact maps for PconsFold (using PconsC). The colours represent all four CATH fold classes. Lines are fitted to the data to illustrate performance differences between the fold classes

However, the performance difference between PSICOV and plmDCA diminishes when their contacts are used in structure prediction.

There is an optimal number of contacts, specific to the contact prediction method. In Figure 3a, the average TM-score maximum is reached at $x = 1.0$ for PconsFold, while for PSICOV and plmDCA fewer contacts provide better models. This can be explained by the quality of predicted contacts, i.e. there exists a larger number of correct contacts in PconsC compared with plmDCA or PSICOV (Skwark *et al.*, 2013). Using more of these contacts during structure prediction leads to improved models.

This relation between contact and model quality can also be observed in Figure 3b. The overall Pearson correlation between PPV and TM-score in this dataset is 0.59, i.e. proteins with contact maps of lower quality tend to be modelled less accurate.

Proteins belonging to the *mainly alpha* CATH fold class seem to be easier to fold than proteins from the *alpha & beta* fold class, and proteins from the *mainly beta* fold class seem to be hardest to fold, see Figure 3b and Supplementary Figure S1. On average, predicted contact maps in mainly $\alpha$-helical proteins (blue) have similar or lower PPV values than those of $\beta$-sheet containing proteins (green and yellow), but the resulting models are more accurate in terms of TM-score. This might be explained by higher contact-order (Plaxco *et al.*, 1998) in beta-sheet proteins, which renders such proteins more difficult to fold (Bradley and Baker, 2006).

### 3.2 Identification of top-ranked model

Table 1 summarizes the evaluation of different MQAPs on the predictions for the PSICOV dataset. It shows average TM-scores for top-ranked models after re-ranking all models with each MQAP. The difference between two TM-score values is significant with >95% confidence if its absolute value is 0.02 or above. Due to the different ways constraints are used within CNS and Rosetta, we did not apply Rosetta's internal scoring function to the models generated by CNS. To provide a TM-score baseline, we ran the Rosetta folding step without constraints and using the internal scoring only. This resulted in an average TM-score of 0.34. For PconsFold and Rosetta/plmDCA, the internal scoring performed best, which is indicated by the highest average TM-score in Table 1. The internal scoring function takes into account the number of satisfied predicted contacts as a part of the energy function, i.e. models that satisfy more predicted contacts are assigned low energies and thus ranked on top.

**Table 1.** Average TM-scores for top-ranked models

| Method | EVfold-PLM | Rosetta/plmDCA | PconsFold |
|---|---|---|---|
| Rosetta | – | 0.50 | 0.55 |
| Pcons | 0.47 | 0.47 | 0.53 |
| ProQ2 | 0.36 | 0.46 | 0.51 |
| DOPE | 0.46 | 0.32 | 0.36 |

Models were ranked by different MQAPs.

In EVfold-PLM the best method to identify top-ranked models is by using Pcons (Lundström *et al.*, 2001). Further, Pcons performed only slightly worse than the internal scoring function of Rosetta on the models from PconsFold and Rosetta/plmDCA, showing the advantage of simple clustering methods.

Downloaded from http://bioinformatics.oxfordjournals.org/ at INRIA Paris on September 7, 2014

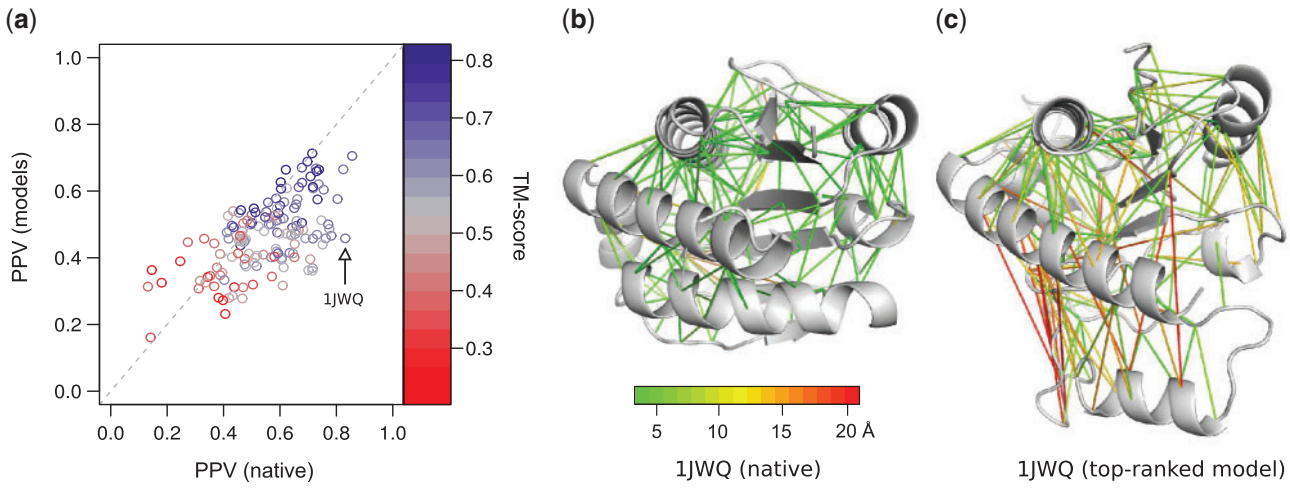



**Fig. 4.** Analysis of contact maps in native structures and top-ranked models. PPVs were calculated for the sets of contacts that were used during folding ($1.0 \cdot l$ top-ranked contacts) with a $C^{\beta}$ distance cutoff of 8 Å in the structures. (**a**) PPV values for PconsC contacts on native structures (*x*-axis) against PPVs on the top-ranked models from PconsFold (*y*-axis). The colours represent TM-scores of models against native structures. (**b**) Native structure of 1JWQ. Lines represent all predicted contacts. The colour scheme indicates spatial distances of residue pairs in the structure. The PPV is 0.83. (**c**) Predicted contacts in the top-ranked model for 1JWQ with the same color scheme. This model has a TM-score of 0.62 and a PPV of 0.46

Further, we examined two quality assessments that have been reported to show good ability to identify accurate protein models, ProQ2 (Ray *et al.*, 2012) and Dope (Shen and Sali, 2006). The ProQ2 quality did not perform well on EVfold-PLM models but only slightly worse than Pcons on models from PconsFold and Rosetta/plmDCA. We also tested a combination of ProQ2 and Pcons, as in Wallner *et al.* (2007). However, the results are omitted, as they were not significantly different from those of Pcons alone. DOPE scoring of EVfold-PLM models worked almost as good as Pcons, but we observe a strong decline in average model quality for PconsFold and Rosetta/plmDCA.

### 3.3 Does PconsFold optimally use the contact information?

Next, we examined if the folding protocol has fully used the available contact information. This was done, by comparing the number of satisfied constraints (PPV) in the top-ranked model and the native structure. On average, the PPV of predicted contacts is higher in native structures than in the models (0.55 versus 0.47). This shows that using a more efficient folding protocol would improve the models. All points in the lower right triangle in Figure 4a indicate models that could be improved. Clearly, for predicted contacts below PPV = 0.4, many models satisfy the constraints better than the native structures, i.e. we would not expect that a better folding protocol would improve the models. It is possible that these low-quality contact maps mislead the structure prediction process and that this results in models that diverge from native structures with a better fit to the predicted contacts. Possibly, in these cases, it would be better to use fewer contacts. However, for most proteins with a PPV higher than 0.4, we are clearly not able to satisfy all constraints. Here, an improved folding protocol would improve the models.

In Figure 4, we also study one example of non-optimal folding more closely. We selected 1JWQ, as it represents the worst-case scenario where the difference between native and model PPV is largest. With a PPV of 0.83, most of the contacts were predicted correctly for this protein. The overall PPV decreased to 0.46 for this model. Clearly, the folding protocol has not been able to produce a compact model. Generating more decoys might solve this problem but a more efficient folding protocol would be preferable.

### 3.4 PconsFold versus EVfold

When looking at Figure 5, the majority of alpha-helical proteins (blue) achieved higher TM-scores with PconsFold than with EVfold-PLM. On average, we see an improvement of 14% of PconsFold over EVfold-PLM on alpha-helical proteins. The same improvement can also be observed for Rosetta/plmDCA. Together with Figure 3, this indicates that Rosetta performs better on alpha-helical proteins. The average performance of PconsFold on alpha & beta proteins (green) is 15% higher than for EVfold-PLM. However, Rosetta/plmDCA performs equally good as EVfold-PLM on average on such targets. The improvement might thus be due to PconsC contact maps, as it is only observable for PconsFold and not for Rosetta/plmDCA. At a level of single proteins, we see quite some divergence between the results of EVfold-PLM and PconsFold in Figure 5. This is especially true for *mainly beta* proteins and *alpha & beta* proteins.

The analysis with MolProbity reveals that the backbone dihedral quality of Rosetta models is higher than for EVfold-PLM models. The percentage of Ramachandran outliers is 0.62% for PconsFold and 12.16% for EVfold-PLM. MolProbity further reports an average clash score of 10.48 for PconsFold and 9.90 for EVfold-PLM. The EVfold-PLM models have slightly less clashes than the models from PconsFold. The average final MolProbity score is with 1.85 better for PconsFold then 2.38 for EVfold-PLM. This corresponds to a better average MolProbity percentile rank of 80.69 for PconsFold than the average rank of 54.21 for EVfold-PLM. Additional analysis



Downloaded from http://bioinformatics.oxfordjournals.org/ at INRIA Paris on September 7, 2014

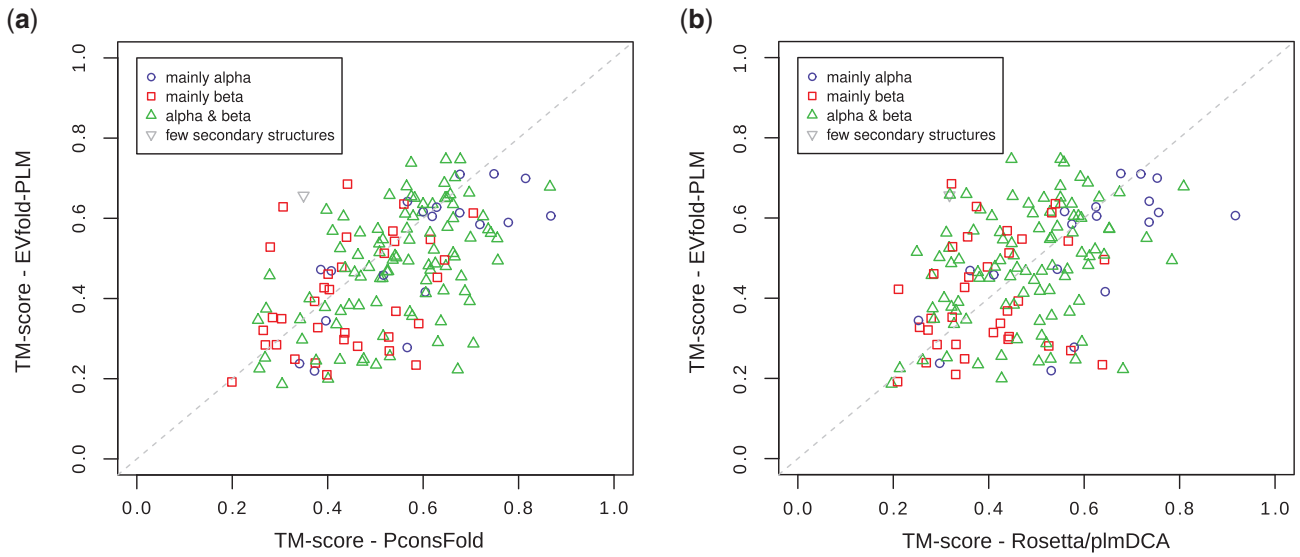



**Fig. 5.** TM-score comparison for top-ranked models of the proteins in the PSICOV dataset. The decoys for each method were re-ranked using Pcons to assess the performance of the structure prediction process independent of the model ranking scheme. The colours represent all four CATH fold classes. (**a**) PconsFold compared with EVfold-PLM. (**b**) Rosetta/plmDCA compared with EVfold-PLM

with PROCHECK confirmed this trend with 0.1% Ramachandran outliers for PconsFold and 8.6% for EVfold.

Table 2 contains TM-scores for the top-ranked models of each protein in the small test dataset. We compare the EVfold results, as published in Marks *et al.* (2011), with results from the current EVfold-PLM and PconsFold with 20 000 generated decoys. With an average TM-score of 0.55, EVfold-PLM performs 15% better than EVfold with mean field within the 90% significance interval, showing the importance of improved contact prediction. When using PconsFold, there is an additional improvement of 16% to an average TM-score of 0.64 within the 95% significance interval. This further supports our observation that improved contact maps improve structure prediction.

Generally, co-evolution based methods are applicable to larger proteins, as earlier studies on membrane proteins have shown (Hopf *et al.*, 2012; Nugent and Jones, 2012). Limitations are given by the number of available sequences per residue (Hopf *et al.*, 2012; Kamisetty *et al.*, 2013) and runtime.

**Table 2.** TM-scores for top-ranked models comparing EVfold with mean field, EVfold-PLM and PconsFold with 20 000 decoys

| Protein | EVfold | EVfold-PLM | PconsFold |
|---|---|---|---|
| BPT1_BOVIN | 0.49 | 0.25 | 0.57 |
| CADH1_HUMAN | 0.55 | 0.54 | 0.53 |
| CD209_HUMAN[a] | 0.39 | 0.64 | 0.54 |
| CHEY_ECOLI | 0.65 | 0.66 | 0.82 |
| ELAV4_HUMAN | 0.57 | 0.61 | 0.80 |
| O45418_CAEEL | 0.48 | 0.62 | 0.65 |
| OMPR_ECOLI | 0.35 | 0.44 | 0.59 |
| OPSD_BOVIN | 0.50 | 0.55 | 0.56 |
| PCBP1_HUMAN | 0.25 | 0.43 | 0.60 |
| RASH_HUMAN | 0.70 | 0.62 | 0.67 |
| RNH_ECOLI | 0.54 | 0.66 | 0.61 |
| SPTB2_HUMAN | 0.37 | 0.51 | 0.74 |
| THIO_ALIAC | 0.55 | 0.56 | 0.83 |
| TRY2_RAT | 0.53[b] | 0.78 | 0.54 |
| YES_HUMAN | 0.35 | 0.31 | 0.57 |
| Mean | 0.48 | 0.55 (0.09*) | 0.64 (0.04**) |

[a]The Uniprot entry A8MVQ9_HUMAN of the EVfold publication was renamed into CD209_HUMAN.
[b]This value was corrected, as in the original publication it showed the value for the best possible model.
**P*-value for a one-sample t-test of TM-score difference to EVfold TM-scores.
***P*-value for a one-sample t-test of TM-score difference to EVfold-PLM TM-scores.

## 4 CONCLUSION

Here, we show that improved contact predictions from PconsC (Skwark *et al.*, 2013) actually lead to improvements in protein structure prediction. Further, it is clear that for proteins with better-predicted contacts, the generated models are of higher quality, i.e. future improvements in contact predictions will result in higher quality models. According to Kamisetty *et al.* (2013), there are 422 protein families without known structure that are suitable for co-evolution-based contact prediction methods. This number will surely increase due to improved contact prediction methods and growing sequence databases at increasing rates. A comparison between using Rosetta and CNS indicates that using similar contact predictions generates models of similar quality. However, Rosetta models are chemically more correct. Finally, it is also clear that in many top-ranked models, the contacts are less well satisfied than in the native structures, i.e. an improved folding protocol would improve the models. One option would be to use model-PPV as a measure of model quality, or as a stopping criterion for decoy generation, another option is to use both CNS and Rosetta and then try to identify the optimal model. Our results on different fold classes show that model quality, especially in beta-sheet containing proteins, could

Downloaded from http://bioinformatics.oxfordjournals.org/ at INRIA Paris on September 7, 2014

be even further enhanced by focusing future work on long-range contacts.

## ACKNOWLEDGEMENT

Joel Hedlund is acknowledged for technical advice and valuable contributions to the code. Computational resources: The computations were performed on resources at PDC Centre for High Performance Computing (PDC-HPC) and on resources provided by SNIC through Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX) and National Supercomputer (NSC) Centre in Linköping, Sweden.

*Funding*: This work was supported by grants from the Swedish Research Council (VR-NT 2012-5046, VR-M 2010-3555), SSF (the Foundation for Strategic Research) and Vinnova through the Vinnova-JSP program, and SeRC the Swedish E-science Research Center. M.J.S. is supported by the Academy of Finland Center of Excellence COIN. D.S.M., C.S. and S.H. are supported by NIH award R01 GM106303.

*Conflict of Interest*: none declared.

## REFERENCES

Berman,H.M. *et al.* (2000) The protein data bank. *Nucleac Acids Res.*, **28**, 235–242.

Bradley,P. and Baker,D. (2006) Improved beta-protein structure prediction by multilevel optimization of nonlocal strand pairings and local backbone conformation. *Proteins*, **65**, 922–929.

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

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

Chen,V.B. *et al.* (2010) MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr. D. Biol. Crystallogr.*, **66** (Pt. 1), 12–21.

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.

Eswar,N. *et al.* (2006) Comparative protein structure modeling using modeller. *Curr. Protoc. Bioinformatics*, **Chapter 5**, Unit 5.6.

Gbel,U. *et al.* (1994) Correlated mutations and residue contacts in proteins. *Proteins*, **18**, 309–317.

Hatrick,K. and Taylor,W. (1994) Sequence conservation and correlation measures in protein structure prediction. *Comput. Chem.*, **18**, 245–249.

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

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

Kamisetty,H. *et al.* (2013) Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era. *Proc. Natl Acad. Sci. USA*, **110**, 15674–15679.

Kryshtafovych,A. *et al.* (2013) CASP10 results compared to those of previous CASP experiments. *Proteins*, **2**, 164–174.

Lapedes,A. *et al.* (2012) Using sequence alignments to predict protein structure and stability with high accuracy. *ArXiv e-prints.*

Lapedes,A.S. *et al.* (1999) Correlated mutations in models of protein sequences:phylogenetic and structural effects. In: *Proceedings of the IMS/AMS International Conference on Statistics in Molecular Biology and Genetics*. Monograph Series of the Inst. for Mathematical Statistics, Hayward, CA, pp. 236–256.

Leaver-Fay,A. *et al.* (2011) ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. *Meth. Enzymol.*, **487**, 545–574.

Lundström,J. *et al.* (2001) Pcons: a neural-network-based consensus predictor that improves fold recognition. *Protein Sci.*, **10**, 2354–2362.

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

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

Marks,D.S. *et al.* (2012) Protein structure prediction from sequence variation. *Nat. Biotechnol.*, **30**, 1072–1080.

Morcos,F. *et al.* (2011) Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc. Natl Acad. Sci. USA*, **108**, 1293–1301.

Neher,E. (1994) How frequent are correlated changes in families of protein sequences? *Proc. Natl Acad. Sci. USA*, **91**, 98–102.

Nugent,T. and Jones,D.T. (2012) Accurate de novo structure prediction of large transmembrane protein domains using fragment-assembly and correlated mutation analysis. *Proc. Natl Acad. Sci. USA*, **109**, 1540–1547.

Plaxco,K.W. *et al.* (1998) Contact order, transition state placement and the refolding rates of single domain proteins. *J. Mol. Biol.*, **277**, 985–994.

Ray,A. *et al.* (2012) Improved model quality assessment using ProQ2. *BMC Bioinformatics*, **13**, 224.

Rohl,C.A. *et al.* (2004) Protein structure prediction using rosetta. In: Brand,L. and Johnson,M.L. (eds) *Methods in Enzymology,* Volume 383 of *Numerical Computer Methods, Part D*. Academic Press, Waltham, Massachusetts, USA. pp. 66–93.

Schug,A. *et al.* (2009) High-resolution protein complexes from integrating genomic information with molecular simulation. *Proc. Natl Acad. Sci. USA*, **106**, 22124–22129.

Shen,M. and Sali,A. (2006) Statistical potential for assessment and prediction of protein structures. *Protein Sci.*, **15**, 2507–2524.

Shindyalov,I.N. *et al.* (1994) Can three-dimensional contacts in protein structures be predicted by analysis of correlated mutations? *Protein Eng.*, **7**, 349–358.

Simons,K. *et al.* (1999) Ab initio protein structure prediction of CASP III targets using ROSETTA. *Proteins*, **34**, 82–95.

Skwark,M.J. *et al.* (2013) PconsC: combination of direct information methods and alignments improves contact prediction. *Bioinformatics*, **29**, 1815–1816.

Sulkowska,J. *et al.* (2012) Genomics-aided structure prediction. *Proc. Natl Acad. Sci. USA*, **109**, 10340–10345.

Vendruscolo,M. *et al.* (1997) Recovery of protein structure from contact maps. *Fold Des.*, **2**, 295–306.

Wallner,B. *et al.* (2007) Pcons.net: protein structure prediction meta server. *Nucleic Acids Res.*, **35**, W369–W374.

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.

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

Downloaded from http://bioinformatics.oxfordjournals.org/ at INRIA Paris on September 7, 2014