BIOINFORMATICS ORIGINAL PAPER

Vol. 26 no. 24 2010, pages 3067–3074
doi:10.1093/bioinformatics/btq581

Structural bioinformatics

Advance Access published October 14, 2010

Model quality assessment for membrane proteins

Arjun Ray[1,2,3], Erik Lindahl[1,2,3,4] and Björn Wallner[1,2,3,*]

[1]Department of Biochemistry & Biophysics, Stockholm University, SE-106 91 Stockholm, [2]Center for Biomembrane Research, [3]Swedish eScience Research Center and [4]Department of Theoretical Physics, Royal Institute of Technology, Stockholm, Sweden

Associate Editor: Ivo Hofacker

ABSTRACT

Motivation: Learning-based model quality assessment programs have been quite successful at discriminating between high- and low-quality protein structures. Here, we show that it is possible to improve this performance significantly by restricting the learning space to a specific context, in this case membrane proteins. Since these are among the most important structures from a pharmaceutical point-of-view, it is particularly interesting to resolve local model quality for regions corresponding, e.g. to binding sites.

Results: Our new ProQM method uses a support vector machine with a combination of general and membrane protein-specific features. For the transmembrane region, ProQM clearly outperforms all methods developed for generic proteins, and it does so while maintaining performance for extra-membrane domains; in this region it is only matched by ProQres. The predictor is shown to accurately predict quality both on the global and local level when applied to GPCR models, and clearly outperforms consensus-based scoring. Finally, the combination of ProQM and the Rosetta low-resolution energy function achieve a 7-fold enrichment in selection of near-native structural models, at very limited computational cost.

Availability: ProQM is available as a server at `proqm.cbr.su.se`.

Contact: bjorn@cbr.su.se

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

Received on March 31, 2010; revised on October 8, 2010; accepted on October 11, 2010

1 INTRODUCTION

Computational modeling of 3D structure of protein molecules is an important research area that has the potential to dramatically accelerate the determination of protein structures, both in terms of purely predicted structures and as part of an otherwise experimental pipeline. An increase in the number of proteins for which high quality structural data is available would greatly enhance our ability to understand biological function, redesign proteins of interest and develop new drugs, which might even be peptides themselves (Baker and Sali, 2001; Zhang, 2009).

Membrane proteins are particularly interesting for medical applications, but structure prediction of these proteins is still in its infancy compared with globular proteins. Fortunately, many methods (in particular sequence-based ones) originally developed for water-soluble proteins can be applied directly to membrane proteins (Forrest *et al.*, 2006). However, methods using knowledge-based potentials derived from existing water-soluble proteins need to be changed and adapted to account both for the specific membrane environment and composition of membrane proteins (Barth *et al.*, 2009; Pellegrini-Calace *et al.*, 2003).

Many structure prediction methods that approaches today can generate a large number of models either by constructing models from alignments to different templates (Wallner and Elofsson, 2006) or by sampling different regions of the conformational space (Rohl *et al.*, 2004). A scoring function or model quality assessment program (MQAP) is then required to discriminate high-quality models from imperfect ones. An ideal scoring function would produce perfect correlation between the score and quality, as measured by the distance to the native structure.

The various scoring functions in use today can roughly be divided into three classes: physics, knowledge and learning based. For physical scoring functions, the goal is to describe the physics of the interaction between atoms as accurately as possible. These functions are often parameterized on much smaller systems than proteins, and the typical example is a molecular mechanics forcefield such as OPLS (Jorgensen *et al.*, 1996), CHARMM (Brooks *et al.*, 1983) or Amber (Weiner *et al.*, 1984). Knowledge-based scoring functions in contrast derive probability distributions from features extracted from native structures (Chao Zhang and Kim, 2000; Luethy *et al.*, 1992; Samudrala and Moult, 1998; Sippl, 1990). Finally, learning-based functions are trained on structural features to distinguish between correct and incorrect models and to predict the actual quality of a given model. The training methods available range from simple optimization of a few parameters to advanced machine learning methods such as neural networks or support vector machines (SVMs; Fain *et al.*, 2002; Pawlowski *et al.*, 2008; Wallner and Elofsson, 2003; Wang *et al.*, 2009).

Most scoring functions have been developed to discriminate between globally incorrect and approximately correct models. In contrast, only a handful of methods focus on the more general problem of predicting the correctness of different parts of a protein structural model (Tress *et al.*, 2003; Wallner and Elofsson, 2006). Moreover, neither the knowledge- nor learning-based methods have been developed specifically for membrane proteins. In principle, a physics-based scoring function could be used directly for membrane proteins, but in that case we would be faced with the much more complex task of accurately representing the lipid bilayer environment around the protein too. In addition, the success of physics-based scoring functions as MQAPs even for globular proteins has been quite limited, in part due to insufficient sampling. None of the top-ranking MQAPs in CASP7 or CASP8 were physics

*To whom correspondence should be addressed.

based (Cozzetto *et al.*, 2007, 2009), and there is no reason to believe the results would be any different for membrane proteins.

To address the need for accurate MQAPs for membrane proteins as well as position-specific scoring, we have instead chosen to adapt a learning-based scoring function (ProQres) to membrane proteins. The method was first developed for water-soluble proteins (Wallner and Elofsson, 2006), but we have retrained it from scratch on structural models of membrane proteins and also added several new features. These include membrane-specific properties such as topology and Z-coordinate prediction as well as new general features that were not used in the original version of ProQres, e.g. conservation and sequence profile information.

2 METHODS

2.1 Test and training data

When using machine learning method such as SVMs, it is important to have good representative test and training sets. For water-soluble proteins, there are many sets that can be used for instance the CASP datasets (Cozzetto *et al.*, 2009) (see Supplementary Material for a general discussion on training sets). However, for membrane proteins there are no standard publically available sets, so the membrane protein models used here for test and training were constructed from scratch in the following way:

(1) All pairs of the membrane protein chains available as of January 2009 with a resolution <4 Å containing more than one transmembrane helix were structurally aligned using TM-align (Zhang and Skolnick, 2005).

(2) All pairs with >60 aligned residues and a sequence identity between 20–90% were realigned using HHpred version 1.5.0 (Soding, 2005) in global alignment mode (Supplementary Fig. S1).

(3) Modeller (Sali and Blundell, 1993) was used to build coordinates from the alignments, using models build using other programs such as SegMod/ENCAD (Levitt, 1992) does not impact the result (data not shown).

(4) The sequences from all models were clustered with BLASTclust (Altschul *et al.*, 1997) using sequence identity of 20% and coverage of 50% (-S 20 -L 0.5), resulting in 40 clusters.

(5) A maximum of five models were randomly selected from each cluster, resulting in a final set of 103 models with a total of 33 304 residues, which is sufficient to achieve stable training performance (Supplementary Fig. S6).

The test set is available at `proqm.cbr.su.se/testset/`.

2.2 Additional test sets

An additional set of 329 031 models for 10 targets constructed *ab initio* with Rosetta-Membrane (Barth *et al.*, 2009) was used as a completely independent set to test model selection performance for the global version of ProQM.

2.3 SVM training

SVM training was performed using 5-fold cross-validation, with the restriction that models belonging the same sequence cluster (see above) had to be in the same set

The SVM$^{\text{light}}$ (Joachims, 2002) V6.01 implementation of SVM regression was used with a linear kernel function (other kernels were tried but showed no increased performance). The trade-off between training error and margin was optimized (the -c parameter) as well as the epsilon for the width of loss function in regression tube was optimized for all cross-validation sets at the same time.

2.4 Training parameters

SVMs were trained using structural features describing the local environment around each residue in the protein models and on other features that can be predicted from sequence such as membrane topology and conservation. The training was performed on full-length membrane proteins to avoid dividing the models into membrane and water facing domains. Training specifically on membrane residues did not show significant improvement (Supplementary Table S3). Combinations of the following features were used: atom–atom and residue–residue contacts, surface accessibility, secondary structure, evolutionary information and membrane topology calculated over a sequence window. Many of the features are similar to the ones used in our earlier studies (Wallner and Elofsson, 2003, 2006) but included below for clarity. The window size for the structural parameters (atom–atom contacts, residue–residue contact and surface information) was optimized to 21 by testing windows in the range 1–31.

2.4.1 Atom–atom contacts This feature describes the distribution of atom–atom contacts in the protein model. Atoms were grouped into 13 different atom types based on chemical properties (see Wallner *et al.,* 2003). Two atoms were defined to be in contact if the distance between them was within 4 Å. The 4 Å cutoff was chosen by trying different cutoffs in the range 3–7 Å. Contacts between atoms from positions adjacent in sequence were ignored. Finally, the number of contacts from each group was normalized by dividing with the total number of contacts within the window.

2.4.2 Residue–residue contacts This feature describes the distribution of residue–residue contacts. Residues were grouped in six different groups: (i) Arg, Lys; (ii) Asp, Glu; (iii) His, Phe, Trp, Tyr; (iv) Asn, Gln, Ser, Thr; (v) Ala, Ile, Leu, Met, Val, Cys; and (vi) Gly, Pro (Wallner and Elofsson, 2003). A grouping with all 20 amino acids were also tried but showed worse performance. Two residues were defined to be in contact if the distance between the Cα atoms or any of the atoms belonging to the side chain of the two residues were within 6 Å and if the residues were more than five residues apart in sequence. Many different cutoffs in the range 3–12 Å were tested and 6 Å showed the best performance. Finally, the number of contacts for each residue group was normalized with the total number of contacts within the window.

2.4.3 Solvent accessibility surfaces This features describes the exposure distribution for the same residue grouping as used for the residue–residue contacts. The surface accessibility was calculated using NACCESS (Hubbard and Thornton, 1993). The relative exposure of the side chains for each residue group was used. The exposure data were grouped into one of the four groups <25%, 25–50%, 50–75% and >75% exposed and finally normalized by the number of residues within the window.

2.4.4 Secondary structure This set of features describes the secondary structure in the model and also how it corresponds to the predicted secondary structure. STRIDE (Frishman and Argos, 1995) was used to assign three secondary structure classes, helix, sheet or coil, to each residue in the protein models based on coordinates. PSIPRED (Jones, 1999) was used to predict the probability for the same secondary structure classes.

The secondary structure information consists of three parts: (i) overall secondary structure content in the model, i.e. the fraction of helix, sheet and coil; (ii) the secondary structure in the model over the specific window; and (iii) the predicted probability for a particular secondary structure in the model for the central residue in the window.

2.4.5 Membrane-specific information These are set of features, which are specific to membrane proteins consisting of membrane spanning regions and Z-coordinate prediction. The membrane spanning regions were extracted from membrane topology predicted using TOPCONS (Bernsel *et al.*, 2009) and binary encoded over an 11 residue window, without separating inside/outside loops. The absolute Z-coordinate or the distance to the membrane center was predicted using ZPRED (Granseth *et al.*, 2006) and rescaled between 0 and 1, corresponding to the middle and outside of the membrane, respectively. In addition, a 'structural Z-coordinate' calculated from the the carbon alpha coordinates of the predicted membrane spanning regions. The optimal window size for both predicted and structural Z-coordinate was 1.

2.4.6 Evolutionary information These features consist of a window of sequence profiles and conservation scores calculated for the full-length protein sequence and windowed to get a local description. The sequence profiles were constructed by running three iterations of PSI-BLAST (Altschul *et al.*, 1997) against UniRef90, release 15.9, October 13, 2009 (Suzek *et al.*, 2007) with a 10^{-3} E-value cutoff for inclusion (-h) and all other parameters at default settings. The resulting log-odds sequence profiles were converted to values between 0 and 1 using the logistic function $1/(1+\exp^x)$. From the sequence profile, PSI-BLAST also calculates how much each position vary, i.e the conservation. A sliding window was then used over the sequence profile and conservation scores to describe the local evolutionary information. The optimal window sizes were found to be 3 for the sequence profile and 11 for the conservation score by trying window sizes in the range 1–23.

2.5 Target function

In this study, we have used S-score as the correctness measure for each residue in a protein model. This score was originally developed by Levitt *et al.* (1998), and is now employed in many of the functions measuring protein model quality, including MaxSub (Siew *et al.*, 2000), LGscore (Cristobal *et al.*, 2001) and TM-score (Zhang and Skolnick, 2004). The S-score is defined as:

$$S_i = \frac{1}{1+\left(\frac{d_i}{d_0}\right)^2}$$

where d_i is the distance between the i residue in the native and in the model and d_0 a distance threshold. This score ranges from 1 for a perfect prediction ($d_i = 0$) to 0 when d_i goes to infinity. The distance threshold defines the distance for which the score should be 0.5. In this way, it also monitors how fast the function should go to zero. The distance threshold was set to 3 Å and S_i was calculated from the superposition that gave the highest sum of S_i over the whole model, in the same way as in MaxSub.

3 RESULTS AND DISCUSSION

The aim of this study was to develop a membrane protein-specific scoring function that predicts local as well as global structural model correctness significantly better than general methods applied to membrane proteins.

The main idea is to featurize each structural protein model to properties that can be calculated based on its sequence (e.g. conservation and sequence profiles) or 3D coordinates (e.g atom–atom contacts, residue–residue contacts and solvent accessibility) and use these features to predict model correctness (see Section 2 for a complete description of features). To achieve a localized prediction, the environment around each residue was described by calculating the features for a sliding window around the central residue. For features involving spatial contacts, residues or atoms outside the window that are in contact with those in the window were also included. The prediction of global model correctness was then accomplished by summing up the local predictions and normalizing to sequence length (to enable comparisons between proteins). The final global score is a number in the range [0,1].

3.1 Model correctness

A wide range of different quality measures are used in the literature to compare models with the target. If the structural model is close to the native (say <3 Å), RMSD is a pretty good measure. However, when the structural model deviates further from the native structure the RMSD measure gives far too much importance to the residues which are bad (or merely less well defined) compared with the ones that are actually correct. Several measures have been developed to account for this by maximizing the sum of a normalized distance, in particular the S-score: $S_i = 1/(1+(d_i/d_0)^2)$, d_i is the deviation for residue i and d_0 a distance threshold. The S-score ranges from 0 (horrible model) to 1 (perfect). By adjusting d_0, it is possible to select which distance range to focus on. In this study, we used $d_0 = 3$ Å, making the measure most sensitive to distances in the region 3 Å±2. Other measures such as MaxSub (Siew *et al.*, 2000) and GDT_TS (Zemla *et al.*, 1999) were also used but only for benchmarking and not training since they measure global properties of the model.

3.2 Development of ProQM

From earlier studies we expect optimal performance by combining different types of input features (Wallner *et al.*, 2003). To get some understanding of which features contribute to the final performance SVMs were first trained on individual input features that then were combined into the final version of ProQM. Different kernels were tried, including linear, radial basis function and polynomial. For each case, a grid search was performed to obtain optimal training parameters for the kernel function in question. After the parameter optimization stage all kernel functions performed almost equally, which made us choose the linear kernel for speed and simplicity.

Spearman's rank correlation was used as performance measure, over other measures such as absolute error and Pearson's correlation. The benefit of a correlation measure over absolute error is that it allows comparison with other methods which do not predict exactly the same type of model correctness. Also, compared with absolute error it is less sensitive to large errors obfuscating otherwise perfect predictions. Finally, Spearman's rank correlation over Pearson's correlation coefficient to make the comparison with other methods less biased by linear relationship between predicted and correct values as implied by the Pearson's correlation.

The Spearman's rank correlation coefficient for SVMs trained with different input features is shown in Table 1. Surface accessibility contains more information than both residue–residue and atom–atom contact. Residue–residue and atom–atom contact features perform equally well, and in combination their output is on par with the surface information. For the surface information, the feature related to which types of amino acids that are exposed was more important than which are buried. This is opposite to what was observed for ProQ (Wallner *et al.*, 2003), which was trained on water-soluble proteins. This can be explained largely by the fact that other factors, such as buried polars are more important than the hydrophobic effect in membrane proteins. We chose not to separate the surface into lipid- and water-facing parts for two reasons: first, it can be difficult to define what category individual residues belong to, in particular in channels or multi-domain proteins. Second, it is quite rare for reasonable membrane protein models to expose

Table 1. Performance for different input features measured by Spearman's rank correlation

Input features	Spearman
Atom	0.30 (±0.016)
Residue	0.31 (±0.016)
Surface	0.36 (±0.016)
Atom + residue	0.38 (±0.016)
Residue + profile weighting	0.35 (±0.016)
Surface + profile weighting	0.42 (±0.015)
Atom + residue + surface [ARS]	0.40 (±0.015)
ARS + conservation	0.49 (±0.014)
ARS + profile	0.49 (±0.014)
ARS + profile weighting	0.47 (±0.014)
ARS + secondary structure	0.45 (±0.014)
ARS + predicted surface	0.45 (±0.014)
ARS + membrane topology	0.43 (±0.015)
ARS + termini	0.42 (±0.015)
All combined	0.56 (±0.013)

The error correspond to a 99.9% confidence interval.

polar residues to the lipid surface, and when they do it is usually in positions where they can 'snorkle' to interact with the headgroups instead.

To systematically evaluate which features contribute to the final predictive performance, an SVM trained on atom–atom and residue–residue contacts together with solvent accessibility surface was used as a baseline [label ARS]. The use of a baseline facilitates the comparison of different features using a realistic description of the protein structure. This baseline predictor got a correlation of 0.40, which is also close to the performance we would expect from the previous ProQres method.

Sequence conservation together with profile information yields the largest performance increase (+0.09 to 0.49). These features largely contain overlapping information since the conservation is obtained from the profile. Nevertheless, using both leads to a slight performance increase, and any overlap is handled by the SVM.

Profile weighting is not a new feature. It actually uses the same residue–residue contact and surface information as above, but weighted by the occurrence in the sequence profile. For instance, if a position in the sequence contains 50% isoleucine and 50% leucine, contacts to this position are counted as a contact both to 50% isoleucine and 50% leucine. This effectively increases the amount of information in the feature set since data are extracted for all homologous sequences. In addition, it should also make the final predictor less sensitive to small sequence changes. This profile weighting alone resulted in a 0.07 correlation increase.

Predicted secondary structure and predicted surface information yields about the same performance increase (+0.05). Interestingly, membrane protein topology prediction only leads to a modest increase (+0.03). This could be because the SVM primarily learns features that are good both for globular and membrane regions, or because the feature is already partly present, e.g. in secondary structure information.

Finally, a small increase was also observed by adding the distance to the termini in the sequence.

Table 2. Description of the MQAPs that were compared in the benchmark

Method	Description
ProQM	SVM trained to predict S-score
ProQ	Neural network trained to predict LGscore (only global) (Wallner and Elofsson, 2003)
ProQres	Neural network trained to predict S-score (only local) (Wallner and Elofsson, 2006)
Errat	Statistics potential based on different atom types (Colovos and Yeates, 1993)
Verify3D	3D–1D profiles (Luethy *et al.*, 1992)
ProSa2003	Statistical potential of mean force for atom pair and protein–solvent interactions (Sippl, 1993)

All methods except ProQM are optimized on water-soluble proteins.

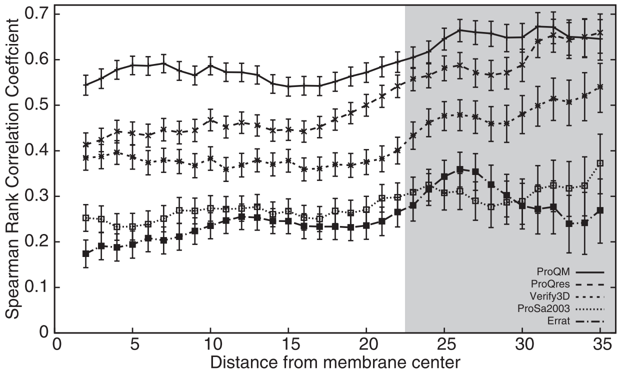

Fig. 1. Spearman's rank correlation coefficient for different distance to the membrane center. The error bars correspond to a 99.9% confidence interval.

3.3 Benchmarking local model correctness

To benchmark both local and global model correctness, the initial set of models (Section 2) was extended from 103 to 287 by including models created using different alignment techniques, filtering out identical alignments. Unfortunately, the current lack of dedicated membrane model quality predictors forced us to benchmark ProQM against four methods primarily developed for water-soluble proteins (Table 2). This is not as bad as it might seem as a majority of the residues (55%) in the test set are actually not in the membrane, and all these methods (including ProQres) have been used to assess quality of membrane protein models in applications.

We focused the evaluation on the ability to rank residues correctly, as measured by Spearman's rank correlation coefficient (Fig. 1), and on the ability to identify *correct* and *incorrect* residues (Supplementary Figs S2 and S3). The analysis of the correct and incorrect regions were performed using receiver operating characteristic (ROC) plots and a cutoff of 3 Å for correct versus incorrect resides. In addition, we also analyzed the accuracy for correct and incorrect residues that were predicted to be in the 10% highest and lowest rank from each method (Fig. 2). This type of measure has proven to summarize the results from ROC plots quite efficiently in previous studies (Wallner and Elofsson, 2006).

To investigate whether the performance differs between residues in the membrane and extra-membrane, the Spearman's rank correlation between local S-score and predictions by the different

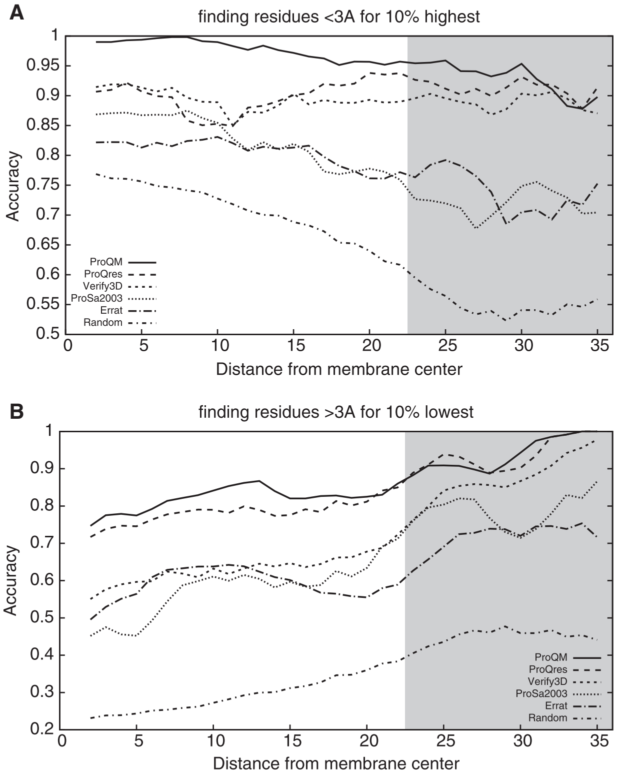

Fig. 2. Fraction correct and incorrect residues among the 10% highest (**A**) and lowest (**B**) ranked residues from each method for different distance to the membrane center.

methods were calculated for different distances to the membrane center (Fig. 1). ProQM has the highest average correlation (0.60) over *all* distances (Supplementary Table S1), and in comparison with other methods the performance difference is most pronounced in the membrane region (<15 Å), while the performance outside the membrane is almost equal to that of ProQres. There is a slight increase in correlation from Z> 15 Å for all predictors. In particular, ProQres and Verify3D perform much better in water than in the membrane. The root mean square prediction error for ProQM is 1.40 Å in the membrane, 1.54 Å outside and 1.46 Å overall.

The ROC plots for finding correct residues agree well with the correlation data presented above. ProQM clearly outperforms the other methods in the membrane region, finding more than twice as many correct residues for the same number of incorrect as the best alternative. In contrast, in the water region ProQM is only slightly better with 25% more correct residue for the same number of incorrect as the best of the other methods. The difference in picking up incorrect residues is not as large as for detecting correct ones: for the membrane region ProQM detects only 25% more incorrect residues and even less in the water region. This is even more evident from the accuracy for the 10% highest and lowest scoring residues (Fig. 2), with highest corresponding to 'correct' and lowest to 'incorrect'. There is a relatively large performance difference in the membrane region for the top-scoring 10%, while the there is virtually no difference to the best of the other methods for the lowest scoring 10%. This indicates that what is considered as 'unfavorable' in the membrane is also likely to be so in the water and vice versa. For instance, an exposed residue without any contacts in coil conformation is probably incorrect in both environments, while the quality of a buried residue in a helix with many contacts will depend on the contacts as well as the environment.

Table 3. Spearman's rank correlation coefficient between predictions from the different methods and different quality measures

Method	*S*-score	MX	GDT
ProQM	**0.76**	**0.74**	**0.74**
ProQ	0.64	0.65	0.65
Errat	0.62	0.59	0.60
Verify3D	0.51	0.53	0.53
ProSa2003	0.37	0.39	0.39

Numbers in bold are significantly better with $P < 0.05$.

3.4 Benchmarking global model correctness

ProQM predicts global model correctness by summing up the local predicted S_i scores and normalizing by the target length. The Spearman's rank correlation coefficient was used to assess performance against three different measure of correctness: *S*-score, MaxSub and GDT_TS (Table 3). ProQM consistently achieve the highest correlation for all quality measures, followed by ProQ, Errat, Verify3D and ProSa2003. However, due to the small sample size (287 models), the *P*-value for this correlation difference using Fisher's r-to-z transform is only <0.05.

3.5 GPCR tests

As a practical test of the usefulness of ProQM, it was tested on two sets of models of G-protein coupled receptor (GPCR) structures. All models used were constructed before their structures were determined experimentally. The first set consists of 907 models of all human GPCRs constructed using TASSER (Zhang *et al.*, 2006). Unfortunately, only three of these GPCRs have a known structure and could be used in the assessment (Skolnik set). The second set consists of all submission to the community-wide assessment of GPCR structure modeling and ligand docking: GPCR Dock 2008 (Michino *et al.*, 2009). In this experiment, before the release of the experimental structure, predictors were asked to submit models of the A_{2A} adenosine receptor with a bound ligand.

3.6 Skolnick set

This set consists of five models for each of the three GPCRs that now have known structures: β_2-adrenergic receptor (B2), A_{2A} adenosine receptor (A2A) and β_1-adrenergic (B1) receptor. The five B1 models all have similar quality with long variable loops, but since these loops are unresolved in the crystal structure they could not be used for evaluating model quality. For B2 and A2A, three of the five models are of much higher and almost equal quality (Supplementary Table S2). It turns out that the higher quality models are helical bundles, fairly similar to the fold of rhodopsin, which one would expect, while the others are completely wrong fold of α/β class. Still, it is an interesting test whether ProQM is able to pick out the best models and distinguish between the good and the bad parts. To avoid any bias, the version of ProQM used here was trained on models of structures without any homology to GPCRs.

As a first test, the local S-score was predicted using ProQM and compared with the true S-scores. The S-score can be transformed to a direct prediction of residue-local distance deviation by solving the S-score equation for d ($d = 3\sqrt{1/S - 1}$). For Spearman's rank correlation, it does not matter if S-score or d is used since the transformation does not change the ranking. The rank correlation coefficient for the correct distance deviation from native to the predicted was 0.69 over all 15 models (4210 residues). The correlation was much better for the 'correct' models as compared with the 'wrong' models: 0.57 versus 0.28. This is no surprise, since it is impossible to rank insignificantly matching residues. It would also be rather useless, since all residues >5 Å off are equally bad. A typical local deviation prediction result for β_2-adrenergic receptor model 1 is shown in Figure 3, all predictions are shown in Supplementary Figure S4. The general agreement between predicted and true deviation is good ($R = 0.78$). There is a slight tendency for ProQM to predict larger deviations for the loops and smaller for the helices, except for the loop between TM1 and TM2 which is correctly predicted by ProQM to be around 1 Å off. The predictions also tend to be lower than the true values. This is even more evident for the prediction on models, which are completely wrong like A2A model 5 (Supplementary Fig. S4). The predicted deviations are seldom >4 Å even though the true deviations are much higher. This is natural because the S-score focus the prediction on good regions rather than bad. However, despite these imperfections we believe the result clearly shows significant added value of local over global quality assessment. The most impressive result is the prediction for the loop around position 150, which is a beta hairpin in the model that superimposes poorly with the helix in the native structure (Fig. 3).

As a second test, the average predicted S-score was used to rank the models globally. The ranking produced in this way easily separates between 'correct' and 'wrong' models. The three good models for B2 are of roughly equal quality and also predicted to be equally good, and the same holds for the two bad ones. For A2A, the correct ranking is not exactly reproduced, as ProQM ranks models 4 and 3 before model 1, when the correct order should be 1, 4 and 3. For B1, the ranking is rather useless since all models are of almost equal quality, but the predicted quality agrees rather well with the correct (Supplementary Table S2).

3.7 GPCR Dock 2008 set

The second GPCR test set consists of 197 models of A_{2A} adenosine receptor (A2A) that were submitted to the Critical Assessment of GPCR Structure Modeling and Docking 2008. Many of the submitted models are rather similar, since most groups built models using the structure of β_2-adrenergic receptor (B2) as template, ending up with around 66% of the residues in correct position (Fig. 4). Still, there are a handful of models that achieve >70% of the residues in correct position and the two best models have 72% correct, which is 17 more correct (<3 Å) residues compared with the B2 template. This makes this test more of a high-resolution local quality prediction compared with the test on the Skolnick set. ProQM was applied to predict the global quality of all models in the GPCR Dock 2008 set. Even though most of models are very similar, ProQM was still able to rank the two single best models among the top 10 (ranks 4 and 8) out of 197. This is significantly better than random ($P < 0.002$) and it is also significantly better than a selection by a consensus method such as Pcons (Wallner and Elofsson, 2007), which would simply select a model with the most structural neighbors, i.e. from the 0.66 region in Figure 4A. Superpositions to the native structure for the best model selected by ProQM and the best model selected by Pcons is shown in Figure 4B. The Pcons model represented the majority of all submitted models and, not surprising, it is also the model that most resembles the B2 template. In contrast, the highest scoring model selected by ProQM contains clear improvements over the template. The most significant are observed at the N-terminus and the loop between transmembrane helices 6 and 7. We believe that this demonstrates the power of using a separate structural assessment as opposed to a pure consensus-based approach. This is particularly true in a close homology modeling case like this, where the best models only constitute 2% of the total set.

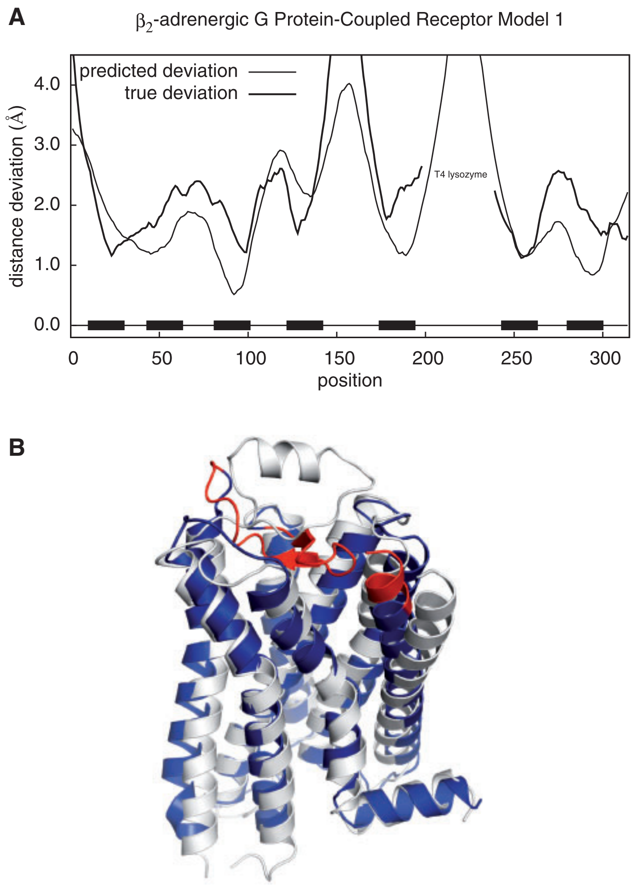

Fig. 3. (**A**) Predicted deviation by ProQM for the first-ranked model of the β_2-adrenergic G protein-coupled receptor made by Skolnick (Zhang *et al.*, 2006) before the experimental structure was solved together with true deviation. (**B**) Superposition of the native structure in white to the model in blue with regions predicted to be >3 Å from native in red.

3.8 Rosetta model selection test

In is important that new MQAPs are benchmarked on models created using different methods. As a final test, we applied ProQM to a large set of low-resolution membrane protein models created *ab initio* using the membrane version of Rosetta (Barth *et al.*, 2009). These models have not yet been subjected to the time consuming process of all-atom refinement. The idea with this

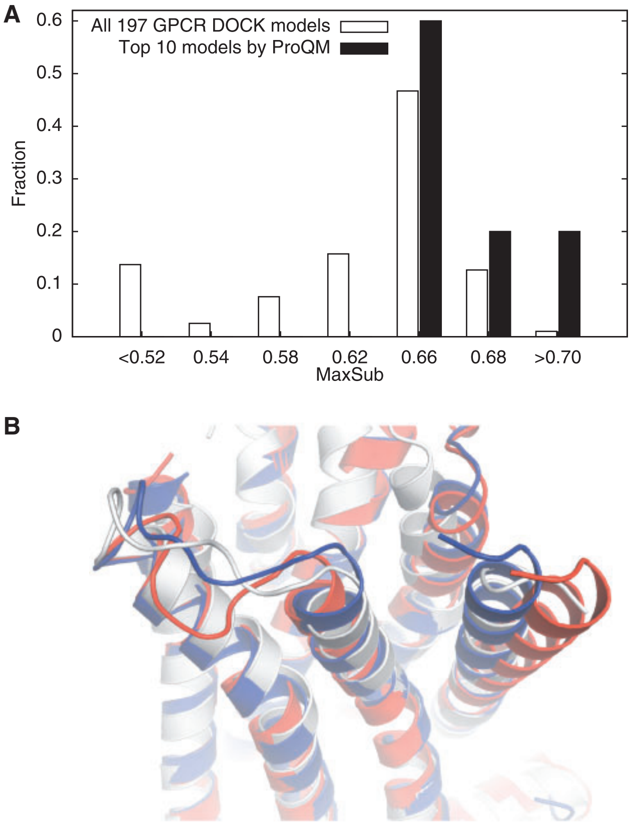

Fig. 4. (**A**) Distribution of model quality as measured by MaxSub for (i) all 197 models of the A_{2A} adenosine receptor that participated in GPCR Dock 2008 and (ii) the top 10 ranked prediction by ProQM (filled). (**B**) Superposition of the native structure in white with best model selected by ProQM in blue and a representative model for the largest cluster in red.

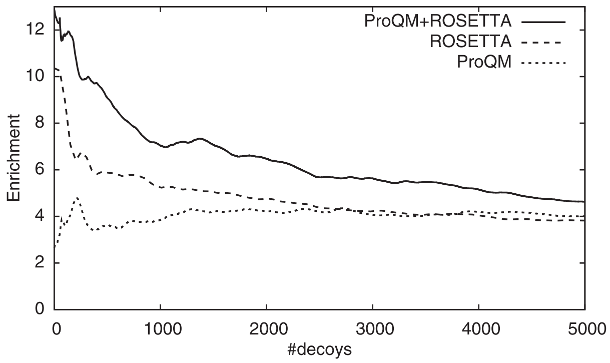

Fig. 5. The enrichment in models that are close to native (top 1% from the initial population) for ProQM, ROSETTA low-resolution energy function and a combination of ProQM and ROSETTA.

test is to investigate if it would be possible to use ProQM as a filter to enrich the selection of near-native models before all-atom refinement. This is important since the models need to be sufficiently close to the native to enable successful refinement (Bradley *et al.*, 2005) and the refinement process cannot correct models to far from the native. Given equal sampling time any enrichment will increase the chance of successful refinement by the same degree. In the past, various clustering techniques have been used to select representative structures and enrich the population of near-native models before all-atom refinement. However, with the total number of models for a given target reaching 100 000 or even 500 000 clustering has become practically impossible. This is especially true when using a distributed computing system like Rosetta@Home or Folding@Home.

Before running ProQM, SCWRL (Canutescu *et al.*, 2003) was used to add complete side chains to each model, followed by standard calculation of all input parameters and prediction using an SVM not trained on any homologs to the target sequence in question. Performance was measured by calculating the enrichment in the best 1% of models from the initial population for different selections based on ProQM, the Rosetta scoring function and a linear combination of the two (Fig. 5). For instance, by selecting the top 1000 ranked models for each method, ProQM has about four, Rosetta five and the combination seven times as many near-native models as the starting population of models. It is clear that even though ProQM performs on par or slightly worse compared with the Rosetta scoring function, they contain different type of information since the perform clearly better in combination. This combination is simply the sum of the Z-score normalized ProQM and Rosetta scores, and as such it will select models that score well using both scores and filter models that score badly in either. By analyzing the per-target results, it is evident that the combination is seldom the best method (2/10) but always better than the worst (Supplementary Fig. S5). Thus, the main reason for the improved performance of the combination is its filtering of high-scoring bad models rather than selection of really good models.

4 CONCLUSIONS

The term 'model quality assessment prediction' can be a bit misleading. While that is indeed the end result, the underlying algorithms are simply scoring functions for protein structures. The advantage of physics- and knowledge-based such functions is that they are differentiable and can be used in combination with minimization methods. However, the more general learning-based scoring functions (that are not differentiable) has the one key advantage that they frequently perform better. In addition, as we have shown here, it is possible to restrict the learning space to a specific context—such as membrane proteins—to further improve performance for certain types of structures. The learning-based functions also appear to be remarkably good at predicting local quality of models along the sequence. There is nothing fundamental that prevents local scoring with the other types of functions, but the lower sensitivity to specific atom positions might give the learning-based methods an important advantage.

To the best of our knowledge, the ProQM method described here is the first model quality assessment program developed specifically for membrane proteins. ProQM predicts the local correctness significantly better than the other methods in the membrane region, and it beats all methods except ProQres (which it matches) for the extra-membrane regions of membrane proteins. The last fact is particularly encouraging for modeling, e.g. of large receptors, since the other methods actually should have an advantage for globular parts.

There are a couple of reasons for the improved performance. First, ProQM is optimized for membrane proteins including membrane-specific features and should therefore perform better in the membrane region than methods optimized for water-soluble proteins. Second, ProQM is apparently able to find a good balance or average between membrane and extra-membrane features as the performance is maintained over all regions. Third, as observed for the earlier globular protein predictor ProQ (Wallner and Elofsson, 2003), the combination of several structural as well as predicted features is crucial for good performance.

If anything, we believe that prediction of local model quality is even more important for membrane proteins than globular ones. Even when there are close homologs known, there can be important differences in loops, and for many membrane proteins this is the region where binding sites are located. As seen in the GPCR tests, ProQM is able to separate globally correct from incorrect models, and the local prediction correctly identifies good/bad regions in the model of the β_2-adrenergic receptor. The GPCR DOCK set had two models that were significantly better than all others and ProQM ranked these in positions 4 and 8 out of 197, which clearly outperforms consensus scoring.

The one drawback of ProQM is that it is still somewhat slower than physics- and knowledge-based functions, which makes it impossible to use on-the-fly in combination with a Monte Carlo scheme to sample structures. However, as shown in the results, it appears to be quite efficient to first generate a large number of structures and then use a combination of scoring functions. When ProQM is paired with the Rosetta low-resolution energy function, we achieve a 7-fold enrichment in selection of near-native structural models at very limited computational cost.

In summary, we believe ProQM provides a highly useful improvement for quality assessment of membrane protein structures, in particular due to its local scoring. The predictor can be accessed over the Internet as `proqm.cbr.su.se`.

Funding: Swedish Research Council and the Swedish Foundation for Strategic Research grants (to E.L.); Carl Trygger Foundation grant as well as a Marie Curie Fellowship (to B.W.).

Conflict of Interest: none declared.

REFERENCES

Altschul,S. *et al.* (1997) Gapped BLAST and PSI–BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Baker,D. and Sali,A. (2001) Protein structure prediction and structural genomics. *Science*, **294**, 93–96.

Barth,P. *et al.* (2009) Prediction of membrane protein structures with complex topologies using limited constraints. *Proc. Natl Acad. Sci. USA*, **106**, 1409–1414.

Bernsel,A. *et al.* (2009) TOPCONS: consensus prediction of membrane protein topology. *Nucleic Acids Res.*, **37**, W465–W468.

Bradley,P. *et al.* (2005) Toward high-resolution de novo structure prediction for small proteins. *Science*, **309**, 1868–1871.

Brooks,B. *et al.* (1983) CHARMM: a program for macromolecular energy minimization and dynamics calculations. *J. Comput. Chem.*, **4**, 187–217.

Canutescu,A.A. *et al.* (2003) A graph-theory algorithm for rapid protein side-chain prediction. *Protein Sci.*, **12**, 2001–2014.

Chao Zhang,C. and Kim,S.-H. (2000) Environment-dependent residue contact energies for proteins. *Proc. Natl Acad. Sci. USA*, **97**, 2550–2555.

Colovos,C. and Yeates,T. (1993) Verification of protein structures: patterns of nonbonded atomic interactions. *Protein Sci.*, **2**, 1511–1519.

Cozzetto,D. *et al.* (2007) Assessment of predictions in the model quality assessment category. *Proteins*, **69** (Suppl. 8), 175–183.

Cozzetto,D. *et al.* (2009) Evaluation of CASP8 model quality predictions. *Proteins*, **77** (Suppl. 9), 157–166.

Cristobal,S. *et al.* (2001) A study of quality measures for protein threading models. *BMC Bioinformatics*, **2**, 5.

Fain,B. *et al.* (2002) Design of an optimal chebyshev-expanded discrimination function for globular proteins. *Protein Sci.*, **11**, 2010–2021.

Forrest,L.R. *et al.* (2006) On the accuracy of homology modeling and sequence alignment methods applied to membrane proteins. *Biophys. J.*, **91**, 508–517.

Frishman,D. and Argos,P. (1995) Knowledge-based protein secondary structure assignment. *Proteins*, **23**, 566–579.

Granseth,E. *et al.* (2006) ZPRED: predicting the distance to the membrane center for residues in alpha-helical membrane proteins. *Bioinformatics*, **22**, e191–e196.

Hubbard,S. and Thornton,J. (1993) *Naccess - Computer Program*. Department of Biochemistry and Molecular Biology, University College London.

Joachims,T. (2002) *Learning to Classify Text Using Support Vector Machines*. Kluwer, MA.

Jones,D. (1999) Protein secondary structure prediction based on position–specific scoring matrices. *J. Mol. Biol.*, **292**, 195–202.

Jorgensen,W. *et al.* (1996) Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *J. Am. Chem. Soc.*, **118**, 11225–11236.

Levitt,M. (1992) Accurate modeling of protein conformation by automatic segment matching. *J. Mol. Biol.*, **226**, 507–533.

Levitt,M. and Gerstein,M. (1998) A unified statistical framework for sequence comparison and structure comparison. *Proc. Natl Acad. Sci. USA*, **97**, 5913–5920.

Luethy,R. *et al.* (1992) Assessment of protein models with three–dimensional profiles. *Nature*, **356**, 283–285.

Michino,M. *et al.* (2009) Community-wide assessment of GPCR structure modelling and ligand docking: GPCR dock 2008. *Nat. Rev. Drug Discov.*, **8**, 455–463.

Pawlowski,M. *et al.* (2008) MetaMQAP: a meta-server for the quality assessment of protein models. *BMC Bioinformatics*, **9**, 403.

Pellegrini-Calace,M. *et al.* (2003) Folding in lipid membranes (FILM): a novel method for the prediction of small membrane protein 3d structures. *Proteins*, **50**, 537–545.

Rohl,C.A. *et al.* (2004) Protein structure prediction using Rosetta. *Methods Enzymol.*, **383**, 66–93.

Sali,A. and Blundell,T. (1993) Comparative modelling by statisfaction of spatial restraints. *J. Mol. Biol.*, **234**, 779–815.

Samudrala,R. and Moult,J. (1998) An all–atom distance-dependent conditional probability discriminatory function for protein structure prediction. *J. Mol. Biol.*, **275**, 895–916.

Siew,N. *et al.* (2000) Maxsub: an automated measure to assess the quality of protein structure predictions. *Bionformatics*, **16**, 776–785.

Sippl,M. (1990) Calculation of conformational ensembles from potentials of mean force. an approach to the knowledge-based prediction of local structures in globular proteins. *J. Mol. Biol.*, **213**, 859–883.

Sippl,M. (1993) Recognition of errors in three–dimensional structures of proteins. *Proteins*, **17**, 355–362.

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

Suzek,B.E. *et al.* (2007) Uniref: comprehensive and non-redundant uniprot reference clusters. *Bioinformatics*, **23**, 1282–1288.

Tress,M.L. *et al.* (2003) Predicting reliable regions in protein alignments from sequence profiles. *J. Mol. Biol.*, **330**, 705–718.

Wallner,B. and Elofsson,A. (2003) Can correct protein models be identified? *Protein Sci.*, **12**, 1073–1086.

Wallner,B. and Elofsson,A. (2006) Identification of correct regions in protein models using structural, alignment, and consensus information. *Protein Sci.*, **15**, 900–913.

Wallner,B. and Elofsson,A. (2007) Prediction of global and local model quality in CASP7 using Pcons and ProQ. *Proteins*, **69** (Suppl. 8), 184–193.

Wallner,B. *et al.* (2003) Automatic consensus-based fold recognition using Pcons, ProQ, and Pmodeller. *Proteins*, **53** (Suppl. 6), 534–541.

Wang,Z. *et al.* (2009) Evaluating the absolute quality of a single protein model using structural features and support vector machines. *Proteins*, **75**, 638–647.

Weiner,S. *et al.* (1984) A new force field for molecular mechanical simulation of nucleic acids and proteins. *J. Am. Chem. Soc.*, **106**, 765–784.

Zemla,A. *et al.* (1999) Processing and analysis of CASP3 protein structure predictions. *Proteins*, (Suppl. 3), 22–29.

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

Zhang,Y. and Skolnick,J. (2005) TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.*, **33**, 2302–2309.

Zhang,Y. *et al.* (2006) Structure modeling of all identified g protein-coupled receptors in the human genome. *PLoS Comput. Biol.*, **2**, e13.

Zhang,Y. (2009) Protein structure prediction: when is it useful? *Curr. Opin. Struct. Biol.*, **19**, 145–155.