BIOINFORMATICS

Vol. 26 ECCB 2010, pages i474–i481
doi:10.1093/bioinformatics/btq369

Vorescore—fold recognition improved by rescoring of protein structure models

Gergely Csaba* and Ralf Zimmer*

Practical Informatics and Bioinformatics Group, Department of Informatics, Ludwig-Maximilians-Universität München, Amalienstr. 17, D-80333 München, Germany

ABSTRACT

Summary: The identification of good protein structure models and their appropriate ranking is a crucial problem in structure prediction and fold recognition. For many alignment methods, rescoring of alignment-induced models using structural information can improve the separation of useful and less useful models as compared with the alignment score. Vorescore, a template-based protein structure model rescoring system is introduced. The method scores the model structure against the template used for the modeling using Vorolign. The method works on models from different alignment methods and incorporates both knowledge from the prediction method and the rescoring.

Results: The performance of Vorescore is evaluated in a large-scale and difficult protein structure prediction context. We use different threading methods to create models for 410 targets, in three scenarios: (i) family members are contained in the template set; (ii) superfamily members (but no family members); and (iii) only fold members (but no family or superfamily members). In all cases Vorescore improves significantly (e.g. 40% on both Gotoh and HHalign at the fold level) on the model quality, and clearly outperforms the state-of-the-art physics-based model scoring system Rosetta. Moreover, Vorescore improves on other successful rescoring approaches such as Pcons and ProQ. In an additional experiment we add high-quality models based on structural alignments to the set, which allows Vorescore to improve the fold recognition rate by another 50%.

Availability: All models of the test set (about 2 million, 44 GB gzipped) are available upon request.

Contact: csaba@bio.ifi.lmu.de; ralf.zimmer@ifi.lmu.de

1 INTRODUCTION

Protein structure prediction is one of oldest and most investigated problems in bioinformatics. Many methods have been proposed and for quite some time systematically assessed in the CASP experiment (Critical Assessment of techniques for protein Structure Prediction since 1994). Despite the hardness of the problem, homology-based methods appeared to be quite successful due to improved alignment methods and more and more available template structures. These methods produce target–template alignments and associated alignment scores. The alignments can be employed to derive structural models for the target using the 3D coordinates of the template structure according to the alignment. The alignment score can be used to rank several candidate templates and the associated alignments to identify the best suited one.

*To whom correspondence should be addressed.

A large range of methods have been proposed for both of the two related but not identical problems: the alignment problem and the model scoring problem. A consensus on the ultimate method have not been reached in both cases as can be seen from the results and discussion in the CASP experiments. Typically, different methods have their strengths for different targets. Often computing the correct alignment is the crucial step in homology-based modeling, as wrong alignments typically cannot be corrected later on during the modeling. On the other hand, the scoring system of the alignment methods is far from perfect. In some cases the method can recognize the best template with high confidence (for this usually a Z-score or P-value is used). But in the absence of highly similar templates, e.g. when only templates from the same fold but neither from the same family nor the same superfamily are available, the scoring is critical.

Even if the scoring system would be appropriate to identify the best alignment between the target and a single template this would not necessarily imply that the score can be used to compare and rank different alignments to different template structures or the resulting model structures. On the other hand, in many cases the alignment between the target and an appropriate template is good enough to build a good model, but the alignment method ranks this alignment lower than an alignment with a wrong template. This can be due to the fact that the alignment scoring is not comparable between different template structures.

Thus, many approaches have been proposed to rescore the produced models using more involved scoring functions which are not and in many cases cannot be directly optimized by the alignment method. A wide variety of those methods, so-called model quality assessment programs (or MQAPs), have been used with good success in the CASP experiments, e.g. as so-called MQAP metaservers employing some consensus of several MQAPs (Pawlowski *et al.*, 2008). In principle, the rescoring approaches can be classified into physics-based systems used for model scoring in *de novo* folding approaches such as Rosetta and heuristic scoring systems using different physico-chemical features and empirically tuned functions.

In this article, we propose a particularly simple rescoring system used in the successful structure alignment program Vorolign (Birzele *et al.*, 2007). Moreover, the Vorpsi and Vorescore methods show how to employ a structure comparison approach for structure prediction. Our results demonstrate that conservation in structural neighborhoods is an important feature for sequence–structure relationships and a determining factor for protein structures.

2 METHODS

The rescoring method we propose is based on the idea of scoring the compatibility of a (target) sequence with a proposed model structure derived

from an alignment of the target sequence with a template. The (re-)scoring of the model structure is done via comparing the two structures, the native template structure and the target model structure via a structural alignment method. We use Vorolign as it has a very high accuracy on the family (∼97%) and superfamily (∼90%) but in contrast to other methods also on the fold level (∼80%) (see Birzele *et al.*, 2007 and Section 3.1). Moreover, it is a very simple method which is targeted at exploiting sequence information in the context of structural contacts. Therefore, it is well suited for the task at hand. The rescoring approach shows how to make a successful structure comparison approach (Vorolign) available for structure prediction (Vorpsi and Vorescore).

In the following sections, we describe Voronoi contacts and the Vorolign structure alignment method. Then we discuss the quality measures we need for computing and assessing structural models (TM-align, TM-score, PPM). The rescoring method is introduced in two variants: Vorpsi using only a single Vorolign rescore for each computed model and Vorescore which takes advantage of additional Vorolign model scores, i.e. Vorpsi scores, computed for a set of similar templates. Finally, we describe the assessment setup, which allows to evaluate the performance of the method and its comparison with other prediction and rescoring methods on different levels of target difficulty (family, superfamily and fold level).

2.1 Homology-based protein structure prediction and model assessment

Homology-based protein structure prediction methods rely on an alignment of the target sequence with a template structure from which model coordinates are derived according to the alignment.

Here, we use simple pairwise alignment of sequences with a *Dayhoff*-like substitution matrix and affine gap costs. Optimal alignments can be efficiently computed with the Gotoh algorithm (Gotoh, 1982). The Gotoh algorithm with the *Dayhoff* matrix (Dayhoff *et al.*, 1978) and gap open costs of -12 and gap extension costs of -1 computing global alignments is called GOTOH in the following.

In addition, we use a sensitive alignment method HHalign from the HHpred toolbox (Söding, 2005) based on the alignment of hidden Markov models, which was very successful in the latest CASP experiments, with standard parameters and call it HHALIGN.

For the generation of candidate models we additionally used 123D (Alexandrov *et al.*, 1996) and profile–profile alignment (PPA) (von Öhsen and Zimmer, 2001) also with standard parameters given in the papers.

As model quality assessment method we use the Rosetta scoring function (called ROSETTA) from the Rosetta package (Raman *et al.*, 2009) with standard parameters, and the ProQ method. For ProQ, we used two versions, one based on C_α atoms from the Pcons package (Wallner and Elofsson, 2003) and the standalone method ProQ(combined) = 5*$\text{MaxSub}_{\text{ProQ}}$ + 1*LG_{ProQ}, i.e. the combined version uses a weighted combination (B.Wallner, ProQ, parameters, personal communication) of the two predicted scores of the structure comparison methods MaxSub (Siew *et al.*, 2000) and LGA (Local-Global-Alignment, Zemla *et al*., 2003), see ProQ web page http://www.sbc.su.se/bjornw/ProQ/ProQ.html.

2.2 Protein structure comparison and alignment via Vorolign

Vorolign compares two protein structures and produces a structural alignment of the two structures and an associated alignment score. The idea behind Vorolign is very simple: it uses the (sequence) conservation of structural neighborhoods as a measure for structural similarity of residue positions in the two structures.

The method works as follows: for any residue in a protein structure, its structurally contacting residues are computed (structural neighborhoods). This is done via a Voronoi decomposition of the (typically C_β) atoms of the structure (therefore the name Voro-lign), where contacts between residues are defined by shared faces of the Voronoi tesselation. The similarity of residue positions in the two structures is measured by aligning the contacts of the respective positions and by scoring the aligned neighbors via a substitution matrix (e.g. the Dayhoff matrix). The alignment and the overall similarity of the two structures is then computed via dynamic programming optimizing this score over all alignments of structure positions.

Thereby, Vorolign exploits the sequence similarity of the two proteins but does so in considering the residue contacts in the two structures. The more similar the structural neighborhoods—both in terms of the number and the residue similarity of the contacting positions—the higher the score. Effectively, the two structures determine via their contacts which positions are contributing (via their sequence similarity) to the Vorolign similarity score.

Interestingly, this simple and very fast (double dynamic programming) approach performs very well in recognizing similar protein structures on the family, superfamily and the fold level. This means that for a query protein, Vorolign is able to predict the correct (SCOP-) fold via taking the best scoring fold from pairwise structural alignments of the query with a set of template structures. The typical fold recognition rates are 97% if family members of the respective query proteins are available in the template set, 90% if no family but superfamily members are available and about 80% if only fold members but not superfamily members are used in the comparison. With these fold recognition accuracies, Vorolign is among the best structural alignment approaches. Vorolign with standard parameters (Birzele *et al.*, 2007) is called VOROLIGN in the following.

As due to its construction VOROLIGN accounts for structural flexibility and as it relies heavily on sequence–structure compatibility to measure structural similarity it appears natural to exploit the VOROLIGN measure for structure prediction as well. How this can be done is presented in this article with the proposed Vorpsi and Vorescore methods.

A final remark on VOROLIGN: other more involved structural alignment methods, e.g. PPM (Csaba *et al.*, 2008) or TM-align (Zhang and Skolnick, 2004) are better suited for producing highly accurate structural alignments, but it is less clear how to use them for sequence–structure prediction.

2.3 Protein structure comparison measures

There is a wide range of methods to measure the structural similarity of protein structure and, related to that, the similarity of a model structure to a native structure. The latter problem is used to assess the quality of structure predictions (e.g. in the CASP experiment) and is somewhat easier as the involved protein sequences are the same and, thus, the alignment is given.

Well-known methods are Vorolign, PPM (Csaba *et al.*, 2008), TM-align optimizing the TM-score (Zhang and Skolnick, 2004), GDT_TS (Global-Distance-Test, Zemla, 2003), LG-score and MaxSub. TM-score and GDT_TS are often used by assessors and predictors to evaluate predicted models in the CASP experiment.

Due to the importance of the problem, e.g. for protein structure classification, structure analysis and structure prediction many more approaches have been proposed, classical ones such as CE (Shindyalov and Bourne, 1998) and also recent ones [Mustang (Konagurthu *et al.*, 2006)], STACCATO (Shatsky *et al.*, 2006), FATCAT (Ye and Godzik, 2005), both pairwise and multiple alignment methods.

Interestingly, the problem of protein structure comparison has not yet reached overall consensus and there is still room for discussion on both the best alignment and the best score. This is also true for the best model assessment method as all methods arguably have some difficulties.

As GDT is kind of 'official' CASP assessment protocol we focus on the TM-score as an independent measure for similarity of model and native structure and use it in the following as our main quality measure. TM-score is derived from the maximum number of matched atoms in a rigid superposition of two structures and is supposed to be protein-length independent. In the following, we evaluate our methods, Vorpsi and Vorescore, with respect to the TM-score and the increase of TM-score in comparison with models predicted by competing methods.

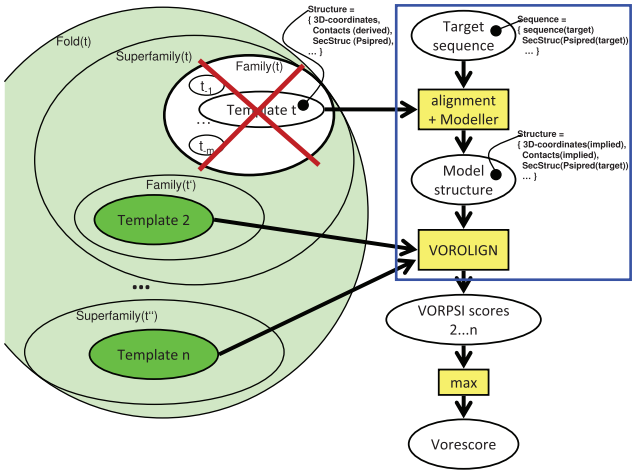

Fig. 1. Overview of the VORSPI (box) and the VORESCORE methods. Proposed models are rescored with VOROLIGN against the used template structure (VORPSI) or against all members of the template's fold except for the template's own family (VORESCORE), respectively.

2.4 Rescoring of homology-based models

Rescoring methods predict 3D structural models for protein target sequences. They consist of the following steps (see box in Fig. 1): (i) a prediction method is used to propose model structures; (ii) if the prediction/alignment method produces a highly confident model and alignment with a template structure, we take this template-induced model as the prediction (i.e. Z-score > 7.0 for HHALIGN); and (iii) Otherwise, the models are rescored to determine the best scoring model as the final structure prediction.

For the prediction step, typically, alignment or threading methods are used to align the target sequence against a library of protein templates. This yields a ranked list of alignments with associated scores (*prediction method scores* or *alignment scores*), which are used to rank the alignments and to select the best one. As prediction method many methods have been proposed such as pairwise and multiple sequence alignment (Gotoh, 1982, 1996), profile (Luthy *et al.*, 1992) and hidden Markov alignments (Eddy, 1996, 1998), profile–profile (von Öhsen and Zimmer, 2001) and HMM–HMM (Söding, 2005) alignments and structure-based alignments [123D (Alexandrov *et al.*, 1996), RDP (Thiele *et al.*, 1999)].

The alignment is used to derive structure models for the target protein from the template coordinates. For this we use the program modeller (Eswar *et al.*, 2008).

2.5 Scoring of a model based on the template: VORPSI

The rescoring method VORPSI needs for the second (rescoring) step only the model and template structures. Vorpsi tries to evaluate the compatibility of the target sequence with the proposed model structure. For this, it employs the VOROLIGN structural comparison between the model structure for the target with the native template structure used for the model. If alignment approaches are used to produce the model structures, also the alignments and in particular also the used templates are known and can directly be used. Otherwise templates need to be identified via structural search with the model structure against a template library.

The rescoring method computes new model scores by comparing the model structure with native template structures via VOROLIGN. The resulting model scores are sorted and the best scoring model is selected as the VORPSI structure prediction for the target. Thus, the rescore score replaces the original prediction method (alignment) score to select the best model. This selection can be better or worse than the original one. This is assessed by comparing the respective models using the TM-score between the model and the native structure. The evaluation estimates how much and how often the TM-score is increased by the rescored model.

The rationale of this simple rescoring approach is as follows. Even simple prediction methods might be able to propose good models, but maybe are not able to score and rank them appropriately.

2.6 Scoring of models using knowledge of the structure space: VORESCORE

In order to exploit knowledge on the sequence–structure space for the model selection problem, we use the following simple approach. Instead of using the template structure of the model for the rescoring we take *all* structures in the fold of the template, apply VORPSI, and take the best score among these templates. This requires to compute additional VOROLIGN alignments for the selected templates in the fold.

In more detail, VORESCORE works as follows (Fig. 1): (i) apply VORPSI for the target; (ii) if the best rescored template does not have additional structures beyond its family in the same fold, VORESCORE takes the best scored model; and (iii) otherwise, for every model from a template t, new alignments/models are computed for the target using all templates in the fold of t except for the family of t. The best of these templates according to the VOROLIGN score is selected as the VORESCORE prediction.

The rationale behind VORESCORE is the following: if the prediction method mistakenly ranks a certain template best, e.g. because of chance similarities with the wrong and missing sequence similarities with the correct template, chances are high that the VORPSI rescoring will also be mistaken. In these cases, we resort to alternative models taken from the same fold. Therefore, we explicitly remove all family members of the original template and take all its fold members as alternative candidates.

3 RESULTS AND DISCUSSION

We propose to employ our structure comparison method VOROLIGN for structure prediction by using it for rescoring structural models produced by some alignment method.

The assessment works for any alignment method, which aligns target sequences against structural templates. The rescoring then VOROLIGNs the model structure with the template structure used for the model building.

We have comprehensively evaluated the VORESCORE method for a range of alignment and threading methods. Here we focus on results for representative methods, a simple method and a highly sensitive and accurate method producing poor and high-quality alignments and models, respectively. We compare our new method Vorescore against a range of quality assessment and rescoring approaches and present results on a physics-based (ROSETTA) and an empirical PROQ method as examples.

Our results on VORESCORE from this assessment are as follows:

First (Section 3.1), we define an appropriate test set of 410 target proteins for the performance evaluation. This set is difficult for sequence-based alignment methods, but allows for good structural template-based models with a TM-score >0.3 (up to high scores of >0.9). The test set is derived from the CATHSCOP consensus set and allows for an unbiased, large-scale evaluation of the performance in different scenarios and different levels of difficulty, e.g. family, superfamily and fold recognition.

Second (Section 3.2), we apply different alignment methods such as pairwise sequence alignment (GOTOH) and hidden Markov model alignment (HHALIGN). From the alignments we build model structures using modeller (Eswar *et al.*, 2008) and rescore the resulting models with VOROLIGN and other rescoring methods such as ROSETTA and PROQ. The results show improvements

of VORESCORE rescoring over the original alignments in about 40% of the cases, significant improvements of VOROLIGN over ROSETTA, which cannot really be used for successful rescoring of models and some improvement over the best state-of-the-art rescoring methods (such as PROQ) using involved scoring methods and meta-/consensus approaches.

Finally, in (Section 3.3), we evaluate the potential of the rescoring if very good structural alignments and models are available. On one hand, VORESCORE can improve by another 50% of the cases as compared with the GOTOH or HHALIGN models, which again shows the ability of the simple VOROLIGN scoring to select good models. On the other hand, there is still much room for improvement: even if very good models are available the selection process misses the best models in many cases, and this margin is even larger for the rescoring of actual models predicted by sequence methods, which reinforces the necessity of producing good alignments (and models) in the first place.

3.1 Definition of an appropriate test set

CATH (Orengo *et al.*, 1997) and SCOP (Murzin *et al.*, 1995) are the two most prominent hierarchical classifications of protein structure domains. Unfortunately, they are not completely consistent w.r.t. similarities and dissimilarities. Therefore, we defined a comprehensive and consistent set of similar (and maybe more importantly dissimilar) pairs of domains, the CATHSCOP set (Csaba *et al.*, 2009). The CATHSCOP set contains lists of templates of consistently classified domains for 4859 target domains with pairwise maximal sequence similarities of 50%. As the CATHSCOP set defines both similar and dissimilar pairs, we can perform tests on more distant similarity recognition. The CATHSCOP set contains 3919 and 2197 target domains, where superfamily (beyond the family similarities) or fold similarities (beyond the superfamily similarities) can be found, respectively.

As our tests are computationally costly due to the model building runs, we create a smaller, but similarly challenging test set using the 'hard' targets for the simple sequence alignment method GOTOH.

The *test set* is defined as the set of the domains, where GOTOH scores a dissimilar domain (neither in the same SCOP fold nor in the same CATH architecture) better than a domain from its own SCOP family and CATH superfamily. We found 410 such targets, inducing 321 641 target–template pairs. In the test set 338 and 181 targets can be used to perform superfamily and fold tests, respectively.

To investigate the properties of the consistent sets, we checked the performances of different sequence alignment methods on both the whole CATHSCOP set (4891 targets, 3 739 824 pairs) and the test set (410 targets covering 224 families, 171 superfamilies and 134 folds, 321 641 pairs). The results are summarized in Table 1 and 2, respectively. We perform three *fold recognition tests*: (i) *family-level test*: all similarities are available, (ii) *superfamily-level test*: members of the target family are excluded; and (iii) *fold-level test*: members of the target superfamily are excluded.

Due to its construction, GOTOH does not recognize the correct family for the targets in the test set. This does not necessarily mean that GOTOH does not produce any meaningful alignments. In fact, as shown in Figure 2 a large number of acceptable models can be derived from the GOTOH alignments. Figure 2 also shows that the test set allows for lots of improvements between the favored models of GOTOH ('+') and the best models possible ['best(all models)']. The optimum among the GOTOH alignments would be 'best(GOTOH)' and the actual improvement achieved by the VORESCORE rescoring is 'VORESCORE'.

Table 1. Fold recognition rates for the CATHSCOP set

	Maximum similarity in the CATHSCOP set		
	Family (4859)	Superfamily (3919)	Fold (2197)
GOTOH	84.07% (4085)	40.50% (1587)	23.81% (523)
123D	77.94% (3784)	30.75% (1205)	21.94% (482)
PPA	93.72% (4554)	73.92% (2897)	50.71% (1114)
HHALIGN	94.11% (4573)	76.24% (2988)	47.06% (1034)
VOROLIGN	**97.53% (4739)**	**90.25% (3537)**	**77.70% (1707)**

We show for every test the number of targets involved in the given set in parentheses. Recognizing the correct fold having all similarity levels (family column) available is an easy task for the current best alignment methods (PPA and HHALIGN), which achieve almost the performance of VOROLIGN. For the superfamily and fold level only more distant similarities are available. Thus, the recognition rates for sequence-based methods are much lower. On the fold level, even the best methods fail on the fold-recognition task on every second target. Also for this case, the use of structural neighborhoods exploited by VOROLIGN improves by >50% on the best sequence method (PPA) to over 75% fold recognition rate. The maximum values in each column, i.e. maximum fold recognition rate for each level, are indicated in bold.

Table 2. Fold recognition rates on the test set

	Maximum similarity in the test set		
	Family (410)	Superfamily (338)	Fold (181)
GOTOH	22.68% (93)	27.51% (93)	16.57% (30)
123D	22.93% (94)	22.19% (75)	19.34% (35)
PPA	81.46% (334)	50.07% (176)	20.99% (38)
HHALIGN	89.76% (368)	66.57% (225)	37.02% (67)
VOROLIGN	**96.10% (394)**	**87.87% (297)**	**76.80% (139)**

The test set is a subset of the CATHSCOP set with 410 query proteins. It is somewhat harder for the sequence methods, but the fold recognition performance of VOROLIGN on the test set is about the same as on the comprehensive CATHSCOP set. The maximum values in each column, i.e. maximum fold recognition rate for each level, are indicated in bold.

The need for a smaller test set results from the expensive tests that we perform. We create models for each alignment computed with several sequence and structural methods resulting in about 2 000 000 models for the smaller test set. On our machines this takes almost a year single-CPU time (modeler takes about 15 s pro-alignment on average).

The defined test set is representative and only slightly more challenging than the CATHSCOP set and it allows for improvement via rescoring. We use the test set for the evaluation in the following.

3.2 Model quality improvement over alignment methods

In this section, we present the improvements achieved with VORESCORE when applied to various prediction methods. Due to space constraints, we restrict the results presented here to two alignment methods, GOTOH and HHALIGN and to two

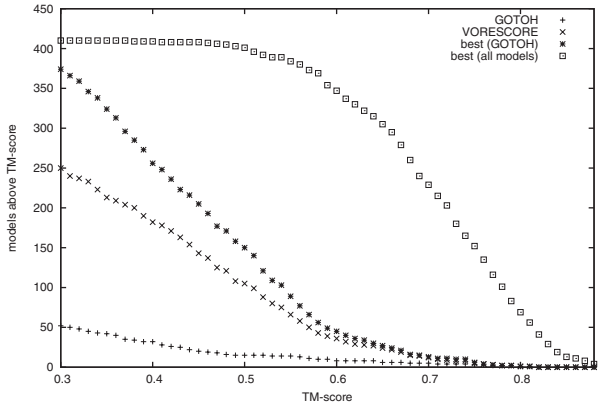

Fig. 2. Model quality of GOTOH alignments. The figure shows the quality (measured with the TM-score) for the models build from GOTOH alignments for the family-level recognition test. The *y*-axis shows the number of targets having a selected model with TM-score larger than the value on the *x*-axis. Due to the construction of the test set the rates for the best model of GOTOH are rather low, but as shown by the 'best(GOTOH)' rates, there are a large number of high-quality models which could be predicted via perfect rescoring of the GOTOH models. In fact, 'VORESCORE' can improve the quality of the selected models significantly. Of course, much better models beyond the GOTOH alignments are possible ['best(all models)'].

rescoring methods, ROSETTA and PROQ. We also focus only on the performance of VORESCORE, which performs slightly better than the simpler VORPSI method. The assessment is based on a comprehensive evaluation of a large number of difficult targets (410). It involves building and scoring more than two million models.

Significant improvements can be observed for both simple (GOTOH) and the most sensitive alignment methods (HHALIGN). The largest improvements of about 40% of the cases are found on the most difficult (fold) level, as should be expected for the sequence-based prediction methods. Surprisingly, the simple VORESCORE method outperforms both the involved ROSETTA [4-fold, the net improvement of ROSETTA on the fold level is $-8\% = (6.63 - 14.92)$ for GOTOH, +10% for HHALIGN, for VORESCORE +40% for both GOTOH and HHALIGN] and PROQ (by about 30%) scoring systems with respect to the net improvements of selecting better models (Table 3).

Table 3 and Figure 3 show different aspects of the same data. The figures show that for all three levels VORESCORE achieves improvements over the whole TM-score range, i.e. for all levels of target difficulty. The more pronounced improvements are observed for the more difficult cases (TM-scores between 0.3 and 0.5). For high TM-scores the performances converge, as due to the high similarities the alignment models often are already the best models possible and, thus, cannot be improved via rescoring.

Table 3 summarizes our results on the overall rescore success rates. VORESCORE selects worse models in only very few cases. For GOTOH it is able to find better models in 60% of the cases for family level and 40% of the cases for superfamily and fold levels. The test set is not easy: ROSETTA does much worse here, it selects worse models in 20–30% of the cases and better models only in rare cases 3–7% (best performance for the fold level). Results are interesting in that they show that improvements are possible even

Table 3. Rescore success rate for ROSETTA, PROQ and VORESCORE on two alignment methods GOTOH and HHALIGN

	Family	Superfamily	Fold
GOTOH			
Rescored model worse			
ROSETTA	21.95% (90)	24.56% (83)	14.92% (27)
PROQ	9.02% (37)	11.83% (40)	6.08% (11)
VORESCORE	**3.17% (13)**	**5.03% (17)**	**2.76% (5)**
Rescored model better			
ROSETTA	4.88% (20)	4.14% (14)	6.63% (12)
PROQ	47.80% (196)	34.02% (115)	35.91% (65)
VORESCORE	**62.44% (256)**	**43.49% (147)**	**43.65% (79)**
HHALIGN			
Rescored model worse			
ROSETTA	11.22% (46)	17.46% (59)	20.44% (37)
PROQ	6.14% (25)	7.25% (24)	10.23% (18)
VORESCORE	**4.88% (20)**	**4.73% (16)**	**5.52% (10)**
Rescored model better			
ROSETTA	5.12% (21)	17.75% (60)	30.39% (55)
PROQ	8.35% (34)	**26.59% (88)**	40.91% (72)
VORESCORE	**9.02% (37)**	23.67% (80)	**44.20% (80)**

If the alignment method predicts with high confidence, all three rescoring methods simply accept this prediction. Otherwise, the rescoring with ROSETTA, PROQ and VORESCORE is based only on the models predicted by the respective alignment (GOTOH and HHALIGN) method. We call the rescored model worse or better if the TM-score difference between the rescored model and the methods first model is smaller than −0.05 or larger than 0.05, respectively, and neutral otherwise. The net improvement of a method is given by the respective difference between the number of better and worse models. The best performance among the three methods ROSETTA, PROQ and VORESCORE for both GOTOH and HHALIGN alignments and the three levels Family, Superfamily and Fold is highlighted as bold.

for very simple alignment methods. It appears that due to the low specificity of the method wrong models can be scored much better than good models and that good models nevertheless are actually be produced despite the overall low performance of the method. And VORESCORE is able to detect these models via a simple structure comparison of model and template structure.

For HHALIGN, the situation is different as HHALIGN belongs to the most sensitive and most accurate sequence-based methods. HHALIGN is supposed to produced better alignment scores and better alignments for the best scoring one—but also for the other candidates. So again the situation is not easy for a rescoring method. Again VORESCORE selects better models in many cases (10% on the family up to 50% on the fold level) and worse models in very few cases (5–6%). For every level the net improvement (better-worse) is significant (5, 15, and 40% for the different levels, respectively), with an overall improvement for 189 of the 410 targets as compared with 41 targets, where the models are actually worse than the original ones. ROSETTA again performs very differently here: rescoring selects more worse than better models depending on the level resulting in a net improvement only on the fold level, resulting in an overall worse performance (127 worse, 114 better) as compared with the original HHALIGN. Thus, ROSETTA cannot be used as a rescoring method in, e.g. the CASP competition. PROQ works quite well with similar performance as VORESCORE for HHALIGN on the family and superfamily levels, but 30% lower rates on the fold level [$+30\% = (40.91 - 10.23)$ as compared with +40% for

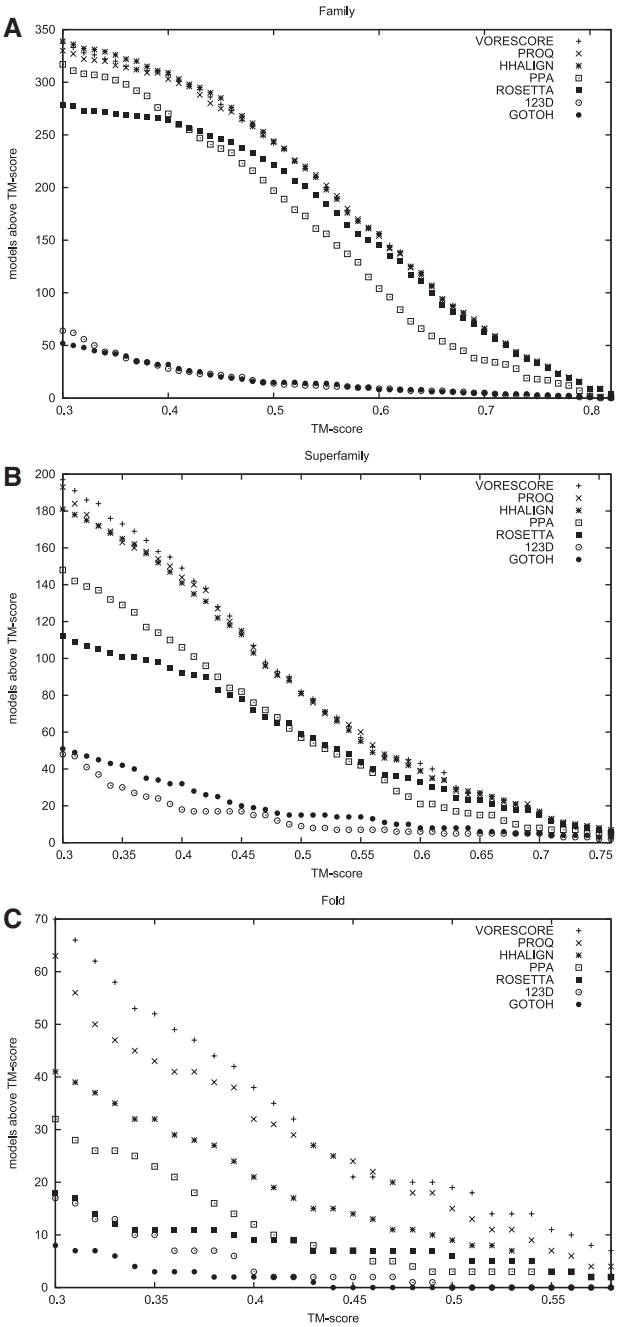

Fig. 3. Comparative recognition performances for alignment and rescoring methods. The three figures show the recognition rates for the family (**A**) superfamily (**B**) and fold (**C**) levels for the relevant range of model qualities (TM-score >0.3 up to convergence). The relative performance is similar for all levels and all model quality ranges: the pairwise alignments are clearly outperformed by profile alignment and these by rescoring methods.

VORESCORE]. For the GOTOH alignments PROQ performs much worse as compared with VORESCORE on all three levels (39% versus 59%, 22% versus 38%, and 30% versus 41%).

The same data of Table 3 is presented in more detail in Figure 3. In the figure the results are plotted against the TM-score of the models to show the dependency on the similarity of the target and template structure (as measured by the TM-score between the predicted model and the native structure). The results are shown in three plots for the family, superfamily and fold levels, respectively.

For various methods, each figure shows the number of models produced by the method above a certain TM-score. Of course this number is highest for smaller TM-scores and is decreased by increasing the required TM-score. Overall, there are 410 targets (family level, 338 for the superfamily and 181 for the fold level) in the used test set above the TM-score threshold of 0.3 (that is why the plot starts at TM-score 0.3).

The figure shows seven plots for seven 'methods'. The methods shown are the four alignment methods: HHALIGN, PPA, 123D and GOTOH and the three rescoring methods: VORESCORE, PROQ and ROSETTA applied to all models proposed by the alignment methods.

For example, 'HHALIGN' shows the number of original HHALIGN ranked models (i.e. selecting the best scoring HHALIGN alignment) above the respective TM-score (as compared with the native structure of the target). Thus, this number approximates the HHALIGN performance in a CASP assessment for the 410 targets in our test set. The other plots show the respective numbers for the other alignment and rescoring methods. The selected figures are just a small portion of the data that we produced on the test set.

The plots show several things. The same trend (with different amounts) and the same ranking is observed for all family, superfamily and fold levels. First, the curves are clearly sorted for all TM-score levels, i.e. there is a consistent ranking for all TM-score values. The highest (i.e. best performing) curve are the two rescoring methods, VORESCORE followed by PROQ. This is followed by HHALIGN, which on the family and superfamily level performs equally well but worse on the fold level. HHALIGN clearly outperforms PPA on all levels. The ROSETTA rescoring only works for the family and superfamily level where it is better than PPA for the easier targets but worse for the harder ones. Thus, ROSETTA appears to work for good models only, on the fold level its performance is almost down to the pairwise alignment methods, 123D and GOTOH, which performs worst on all the levels on the hard targets of the test set.

Overall, the effects are TM-score dependent in that the differences are largest for smaller TM-score thresholds and are converging for the higher TM-score thresholds as the models tend to be very similar (and thus the problem easier) for large TM-scores. The improvements observed for VORESCORE are quite small on the family level and they increase for the superfamily level to the quite drastic effects on the fold level (as hoped). On the fold level, VORESCORE is able to produce almost twice as many good models based on the HHALIGN alignments (as compared with HHALIGN) for lower TM scores (say <0.5). For example, on the fold level, GOTOH finds two, 123D three, ROSETTA nine, PPA 12, HHALIGN 21, PROQ 32 and VORESCORE 38 models with a TM-score of ≥ 0.4.

3.3 Model quality improvement using high-quality alignments

The above results show the performance of the rescoring method in a realistic structure prediction setup similar to, e.g. the CASP assessment. In this section, we evaluate the performance of the rescoring methods in a somewhat artificial setting where the best

Table 4. Rescore success rate of VORESCORE over GOTOH and HHALIGN on all models

Model	Family	Superfamily	Fold
GOTOH			
Worse	1.46% (6)	4.73% (16)	1.66% (3)
Neutral	17.80% (73)	30.47% (103)	31.49% (57)
Better	**80.73% (331)**	**64.79% (219)**	**66.85% (121)**
HHALIGN			
Worse	5.85% (24)	5.62% (19)	3.87% (7)
Neutral	**79.27% (325)**	**55.62% (188)**	27.62% (50)
Better	14.88% (61)	38.76% (131)	**68.51% (124)**

Rescoring is based on both GOTOH and HHALIGN predictions and, additionally, TM-align- and PPM-based models. We call the rescoring worse or better if the TM-score difference between the rescored model and the methods first model is smaller than -0.05 or larger than 0.05, respectively, and neutral otherwise. The largest of the respective three values is highlighted as bold.

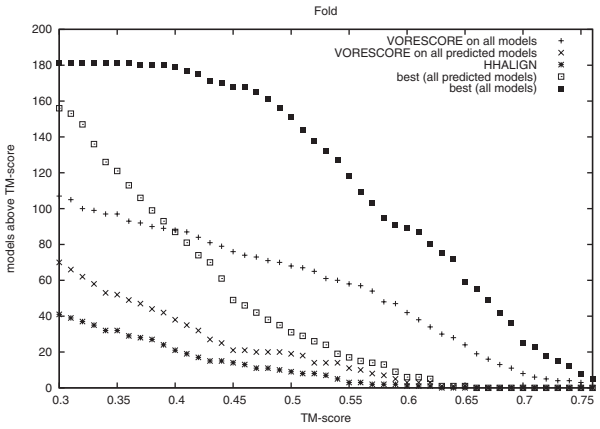

Fig. 4. Theoretical model quality for the test set (fold level). The figure shows the number of models above a certain TM-score threshold for several methods as compared with the theoretical optimum. Here, structure alignment-based models are also available for the rescoring. The number for the best possible template-based model, the best of all predicted models, and the best VORESCORE models (rescoring of *all* models) are compared with the actual VORESCORE (on *all predicted* models) and HHALIGN performance.

models are actually given and also subject to the rescoring. For this, we compute structural alignments with TM-align and PPM and add them to the models under investigation. Again VORESCORE is used to rescore the model structures against the native template structures and rank all the models. Table 4 and Figure 4 show the results.

In Table 4 three numbers are given: how often a worse model is selected; how often a better model is selected; and as an intermediate neutral case, how often the original and rescored models are of the same quality (TM-score difference less than 0.05). First Table 4 shows these numbers for the family, superfamily and fold levels for GOTOH alignments. VORESCORE selects worse models in very rare singular cases and better models in >80% of the cases (family level). The number and percentage of better models is somewhat smaller at 65% and 67% of the cases at the superfamily and fold levels. This reflects the fact that models for the family level are much better as compared with the more distant models on the fold level which appear to be more difficult also for rescoring.

For HHALIGN (Table 4), the situation is similar: again rescoring performs extremely well and improves on the HHALIGN models in 15, 38 and 68% of the cases for the respective levels (family, superfamily and fold). Of course, HHALIGN produces much better rankings and models in the first place, such that it is more difficult to select better models if at all (e.g. on the family level). So, on the family and superfamily level it is often not possible to improve on the HHALIGN alignment (in 80 and 55% of the cases). On the fold level, however, VORESCORE comes up with better models in 60% and with better or neutral models in >96% of the cases.

The Figure 4 is similar to Figure 3. The difference is that in the latter only predicted models, i.e. by sequence-based alignments, are used, whereas the former applies to the rescoring also to structure alignments from TM-align and PPM computed between the template and the native target structure. Moreover, the figure contains the model qualities for the best and best predicted models. Thus, the 'best(all models)' (1) figure shows the number of the overall best possible model for the target at the given TM-score value—this is the theoretical optimum for the test set given the template library independent from any alignment and rescoring method. The 'best(all predicted models)' (2) gives this figure for all actually proposed models, i.e. the performance of a perfect rescoring method. 'VORESCORE on all predicted models' (3) presents the actual performance achieved by VORESCORE on the proposed models and the difference to the former figure (2) indicates the shortcomings of the VORESCORE rescoring. 'VORESCORE on all models' (4) shows what the rescoring could do if the best (structure-based) models would be available, again the difference to the theoretical optimum (1) shows that the rescoring is not perfect. On the other hand, it shows that the rescoring could perform very well on good models [as the difference to the actual performance (3) is quite large]. Finally, 'HHALIGN' (5) shows the HHALIGN performance in comparison. The difference of what is possible with the best alignment methods (5) to what can be done with rescoring (3) is large. The difference to what could be done if better models (4) and/or better rescoring (1) and (2) would be available is even larger. As an example, at the TM-score threshold 0.4, HHALIGN finds 21 models, VORESCORE on all predicted models 38, VORESCORE on all models (also structural alignments) 88, whereas the best of all predicted models yields 87 and the best possible 179 models above TM-score 0.4.

With respect to what is theoretically possible on the current test set the results are somewhat disappointing: while on the family and superfamily levels VORESCORE as the best method can find about 75 and 60% of the best models (data not shown) this is not possible on the fold level, where VORESCORE only finds about 45% of the models and fails in well >50% of the cases (Fig. 4). Also in comparison to the best structural models (TM-align and PPM models) the difference is very large, which indicates that sequence-based alignments fail to produce any reasonable model for many cases (at least at the fold level). Unfortunately, also on the superfamily level there is still a large margin (data not shown).

Last but not least, there is also quite some margin between the TM-align and PPM models and the native structures, which indicates more intricate problems of scoring and assessing structures and structural models, again especially on the fold level as structures become more diverse.

4 CONCLUSIONS

Selecting the correct model is of great importance for homology-based modeling. The alignment score is not the best measure for identifying the best template and alignment. Even if the alignment scoring is far from perfect often good models are proposed and can be selected via appropriate rescoring, e.g. VORESCORE, which performs best among the methods evaluated here.

From our results we draw the following conclusions:

(1) We have shown that template-based protein structure prediction can be significantly improved by using structure information via rescoring of alignment-induced models. Many alignment methods often produce reasonable alignments which score worse than alignments with alternative templates. Rescoring the produced alignments using appropriate sequence–structure information helps to select those models which improve GDT- or TM-score, measures used in the CASP predictions to assess the model quality and to rank the predictions.

(2) As might be clear beforehand, scoring functions which do not directly take sequence–structure information into account such as ROSETTA do perform badly on the task of selecting the best alignment-induced templates. This is especially the case in comparison with VORESCORE, which drastically outperforms these methods.

(3) Maybe not very surprisingly, there is still much room for improvement of those methods. In particular, the method can only select the best of the proposed alignment-induced models. Thus, it is important that good models are actually proposed. If models from the best structural alignment methods are available, rescoring can improve by another large margin (on rescoring only actual sequence-based alignment models). Unfortunately, rescoring is not perfect even in these cases. It was and remains crucial to determine the best alignments with the best models.

It remains to be seen how far one can go with purely sequence–based methods, but rescoring is certainly one way to incorporate additional sequence–structure information into the structure prediction process. The particularly simple rescoring system VORESCORE shows how to use structure comparison for structure prediction and the presented results demonstrate that conservation in structural neighborhoods is an important feature for sequence–structure relationships and a determining factor for protein structures.

Conflict of Interest: none declared.

REFERENCES

Alexandrov,N.N. *et al*. (1996) Fast protein fold recognition via sequence to structure alignment and contact capacity potentials. *Pac. Symp. Biocomput.*, 53–72.

Birzele,F. *et al*. (2007) Vorolign–fast structural alignment using voronoi contacts. *Bioinformatics*, **23**, e205–e211.

Csaba,G. *et al*. (2008). Protein structure alignment considering phenotypic plasticity. *Bioinformatics*, **24**, i98–i104.

Csaba,G. *et al*. (2009) Systematic comparison of SCOP and CATH: a new gold standard for protein structure analysis. *BMC Struct. Biol.*, **9**, 23.

Dayhoff,M.O. *et al*. (1978) A model of evolutionary change in proteins. *Atlas Prot. Seq. Struct.*, **5**, 345–352.

Eddy,S.R. (1996) Hidden Markov models. *Curr. Opin. Struct. Biol.*, **6**, 361–365.

Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.

Eswar,N. *et al*. (2008) Protein structure modeling with modeller. *Methods Mol. Biol.*, **426**, 145–159.

Gotoh,O. (1982) An improved algorithm for matching biological sequences. *J. Mol. Biol.*, **162**, 705–708.

Gotoh,O. (1996) Significant improvement in accuracy of multiple protein sequence alignments by iterative refinement as assessed by reference to structural alignments. *J. Mol. Biol.*, **264**, 823–838.

Konagurthu,A.S. *et al*. (2006) Mustang: a multiple structural alignment algorithm. *Proteins*, **64**, 559–574.

Luthy,R. *et al*. (1992) Assessment of protein models with three-dimensional profiles. *Nature*, **356**, 83–85.

Murzin,A.G. *et al*. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536–540.

Orengo,C.A. *et al*. (1997) CATH–a hierarchic classification of protein domain structures. *Structure*, **5**, 1093–1108.

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

Raman,S. *et al*. (2009) Structure prediction for casp8 with all-atom refinement using rosetta. *Proteins*, **77**, (Suppl. 9), 89–99.

Shatsky,M. *et al*. (2006) Optimization of multiple-sequence alignment based on multiple-structure alignment. *Proteins*, **62**, 209–217.

Shindyalov,I.N. and Bourne,P.E. (1998) Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng.*, **11**, 739–747.

Siew,N. *et al*. (2000) Maxsub: an automated measure for the assessment of protein structure prediction quality. *Bioinformatics*, **16**, 776–785.

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

Thiele,R. *et al*. (1999) Protein threading by recursive dynamic programming. *J. Mol. Biol.*, **290**, 757–79.

von Öhsen,N. and Zimmer,R. (2001) Improving profile-profile alignments via log average scoring. In Gascuel and Moret (eds), *WABI'01*, Vol. 2149 of *LNCS*, Springer, Berlin/Heidelberg, pp. 11–26.

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

Ye,Y. and Godzik,A. (2005) Multiple flexible structure alignment using partial order graphs. *Bioinformatics*, **21**, 2362–2369.

Zemla,A. (2003) LGA - a method for finding 3D similarities in protein structures. *Nucleic Acids Res.*, **31**, 3370–3374.

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