Bioinformatics, 31(23), 2015, 3782–3789
doi: 10.1093/bioinformatics/btv462
Advance Access Publication Date: 7 August 2015
Original Paper

Structural bioinformatics

Improving protein fold recognition with hybrid profiles combining sequence and structure evolution

Yassine Ghouzam[1,2,3,4], Guillaume Postic[1,2,3,4], Alexandre G. de Brevern[1,2,3,4] and Jean-Christophe Gelly[1,2,3,4,*]

[1]Inserm U1134, Paris, France, [2]Université Paris Diderot, Sorbonne Paris Cité, UMR_S 1134, Paris, France, [3]Institut National de la Transfusion Sanguine, Paris, France and [4]Laboratory of Excellence GR-Ex, Paris, France

*To whom correspondence should be addressed.

Associate Editor: Anna Tramontano

Received on May 4, 2015; revised on July 24, 2015; accepted on August 2, 2015

Abstract

Motivation: Template-based modeling, the most successful approach for predicting protein 3D structure, often requires detecting distant evolutionary relationships between the target sequence and proteins of known structure. Developed for this purpose, fold recognition methods use elaborate strategies to exploit evolutionary information, mainly by encoding amino acid sequence into profiles. Since protein structure is more conserved than sequence, the inclusion of structural information can improve the detection of remote homology.

Results: Here, we present ORION, a new fold recognition method based on the pairwise comparison of hybrid profiles that contain evolutionary information from both protein sequence and structure. Our method uses the 16-state structural alphabet Protein Blocks, which provides an accurate 1D description of protein structure local conformations. ORION systematically outperforms PSI-BLAST and HHsearch on several benchmarks, including target sequences from the modeling competitions CASP8, 9 and 10, and detects ~10% more templates at fold and superfamily SCOP levels.

Availability: Software freely available for download at http://www.dsimb.inserm.fr/orion/.

Contact: jean-christophe.gelly@univ-paris-diderot.fr

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

1 Introduction

Understanding how proteins adopt their 3D structure remains one of the most challenging questions in science (Kennedy and Norman, 2005). The resolution of this problem would have a great impact on various fields of biology and medicine, allowing, for example, the rational design of new protein functions and drug molecules. Despite the progress made over the last decades, our biochemical knowledge and computational power are still insufficient to accurately predict protein structure solely from the amino acid sequence. Nevertheless, on the basis of the evolutionary relationship that many proteins share, it is possible to model the structure of a target protein using, as a template, a related protein of known structure. This approach, called comparative modeling, is currently the most successful in constructing theoretical models of protein structures. Since the quality of the resulting structure prediction largely depends on the template protein selected, the initial step of homology detection is crucial in comparative modeling. When proteins share >30% sequence identity, evolutionary relationships can be found using simple pairwise sequence comparison methods (Brenner *et al.*, 1998): heuristic such as BLAST (Altschul *et al.*, 1990) or optimal such as SSEARCH (Pearson, 1991). However, in the case of non-obvious relationships below the 'twilight zone' of 20−35% sequence identity between the target and the template (Rost, 1999), fold recognition methods have been proposed to find related proteins for template-based modeling.

Early methods of remote homology detection were based on profile-to-sequence comparison (Gribskov *et al.*, 1987). A profile can be represented as a position-specific scoring matrix (PSSM) encoded from multiple sequence alignment (MSA) of related proteins and, therefore, contains evolutionary information specific to a protein family and defined by the levels of residue conservation at each sequence position. The first profile-to-sequence algorithm proposed by Henikoff and Henikoff (1994), and the widely used PSI-BLAST (Altschul *et al.*, 1997), were followed by methods derived from the probabilistic interpretation of profiles using hidden Markov models (profile HMMs) (Krogh *et al.*, 1994), like SAM-T98 (Karplus *et al.*, 1998) or HMMER (Eddy, 1998), which improved both the alignment quality and the sensitivity/specificity of detection. Compared with sequence-to-sequence alignment, profiles and profile HMMs greatly improved identification of distantly related proteins and, consequently, comparative modeling (Müller *et al.*, 1999; Park *et al.*, 1998). Fold recognition was later taken to a new level with the FFAS method (Rychlewski *et al.*, 2000), which was based on profile–profile alignment. By using profiles for both target and template, such methods fully exploit the transitivity of sequence homology, making them more sensitive and specific than profile-to-sequence comparisons (Ohlson *et al.*, 2004; Panchenko, 2003). Finally, the pairwise profile HMM comparison introduced with the HHsearch algorithm (Söding, 2005) has further pushed the boundaries of remote homology detection and provided one of the current best approach for protein fold recognition.

Besides these methods solely based on sequence, some authors have proposed structure-based fold recognition methods called ‘threading’ (Bowie *et al.*, 1991; Jones *et al.*, 1992). The principle of threading is to assess the compatibility between the target sequence (1D) and different solved protein structures (3D) by using knowledge-based scoring functions. Thus, comparative modeling may still be achieved by using templates sharing no detectable homology with the target sequence, since proteins may be structurally similar, even without detectable evolutionary relationship between their sequences. Despite interesting success, these 1D–3D methods are limited by the quality of the sequence-to-structure alignment and do not reach the accuracy of the most advanced purely sequence-based methods (McGuffin, 2008).

Following the success of both sequence- and structure-based fold recognition methods, combined approaches have emerged. Indeed, structural information from solvent accessibility, secondary structure or backbone torsion angle can be used to improve sequence-based fold recognition. Such structural features have the advantage of being predictable, which addresses the absence of experimental model for the target protein. Thus, several methods combining discrete structural descriptors with amino acid sequence information have been proposed (Kelley *et al.*, 2000; Shi *et al.*, 2001). Given that structural cores change less rapidly, between 3 and 10 times slower than sequences (Illergård *et al.*, 2009), one can imagine the advantage of using ‘structural profiles’ that would be generated from 1D descriptions of protein structures, in the same way that sequence profiles are generated from amino acid sequences. Indeed, protein structure being more conserved than sequence through evolution, these structural profiles would be more conserved and richer in evolutionary information than sequence profiles and, therefore, better at detecting distant homologies.

In this article, we present ORION, a new method for fold recognition using evolutionary information from both amino acid sequence and protein structure, and encoded as combined sequence/structure profiles. Our method uses the accurate sequential description of protein local conformations provided by the 16-state structural alphabet protein blocks (PB) (de Brevern *et al.*, 2000), instead of the conventional three-state secondary structure, which has important drawbacks, especially for describing the loops which constitutes ~45% of the local structure of the proteins. The relevance of PB structural information has previously been demonstrated through numerous applications (Joseph *et al.*, 2010), such as structural alignment (Gelly *et al.*, 2011) and local structure prediction (Rangwala *et al.*, 2009; Zimmermann and Hansmann, 2008). The novelty of our method also lies in the use of a combination of sequence and structural profiles for both target and template. The structural profiles used for the template were derived from structural alignments of the HOMSTRAD database (Mizuguchi *et al.*, 1998). For distantly related proteins, structural alignments have been shown to be more relevant than sequence alignments (Elofsson, 2002). Thus, using as a benchmark the 1032 structural families from HOMSTRAD and 150 targets from the eighth, ninth and tenth editions of the Critical Assessment of Structure Prediction (CASP) experiments (Moult *et al.*, 2009, 2011, 2014), we show that ORION outperforms the HHsearch algorithm in recognizing proteins with same or similar folds. Our detailed analysis further indicates that this accurate homology detection can be attributed to the inclusion of the PB structural descriptors into profile–profile alignments.

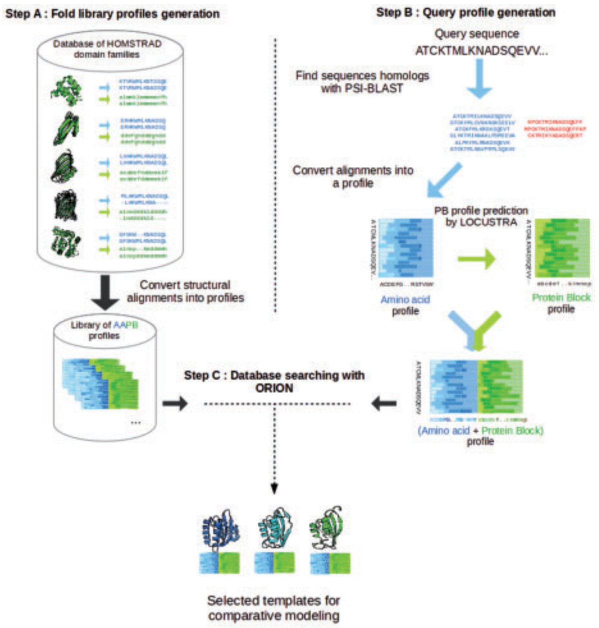
Step A : Fold library profiles generation
Database of HOMSTRAD domain families
Convert structural alignments into profiles
Library of AAPB profiles
...
Step B : Query profile generation
Query sequence
ATCKTMLKNADSQEVV...
Find sequences homologs with PSI-BLAST
Convert alignments into a profile
PB profile prediction by LOCUSTRA
Amino acid profile
Protein Block profile
(Amino acid + Protein Block) profile
Step C : Database searching with ORION
Selected templates for comparative modeling

Fig. 1. Flow chart of ORION algorithm. Template library profile is generated using structural alignments of HOMSTRAD families (Step A). Query profile is obtained from PSI-BLAST and LOCUSTRA (Step B). ORION is then used to search for related proteins in the template library (Step C)

2 Materials and methods

ORION is a fold recognition method based on the pairwise comparison of profiles combining sequence and structural information. The three major steps of ORION algorithm are presented in the following sections and illustrated in Figure 1. The first step is the generation of the template profile library. Sequence profiles are obtained from MSAs of HOMSTRAD families and represent the 20 amino acid scores for each position of the MSA. Structural profiles represent the 16 PB scores for each position of the MSA. The MSAs of templates, derived from structure alignments, were translated into PB MSAs using the atomic coordinates of the protein structures to generate template PB profiles. *A*mino *a*cid and *PB* profiles are joined to form 'AAPB profiles' containing both sequence and structural evolutionary information. From the query sequence, the second step consists in generating a query profile. The query sequence profile is obtained after three iterations of PSI-BLAST and is used to predict the query PB profile. The third and last step is the search for related proteins in the template profile library.

The construction of the AAPB profiles library is described in details in the Section 2.1 and 2.2; the benchmark datasets are detailed in the Section 2.8. ORION performs a global/local alignment by a dynamic programming approach: implementation of dynamic programming, profile scoring function and evaluation of statistical significance of score alignments are described in Sections 2.4–2.7.

2.1 Fold library

The profile library was constructed from HOMSTRAD structural alignments. This database regroups protein families through structural alignments, with 3454 protein structures regrouped into 1032 families. Each family corresponds to a domain identified in the SCOP database (Murzin *et al.*, 1995). The four major SCOP classes, α, β, α/β and $\alpha+\beta$, represent 20, 19, 25 and 22% of the database, respectively.

2.2 AAPB profile library

Amino acid profiles were generated from the MSA of family members for each family. MSAs families were enriched with homologous sequences (Mizuguchi *et al.*, 1998) and sequences with more than 90% identity were filtered out. The weight of each sequence in the profile depends on its relative diversity (i.e. compared with all the other sequences in the MSA) and is calculated using the Henikoff–Henikoff position-based sequence weights (Henikoff and Henikoff, 1994). Profiles are then weighted using 'pseudo-counts' described in the Henikoff–Henikoff weighting scheme (Henikoff and Henikoff, 1996). To avoid non-significant columns in the profile, positions with more than 70% gaps were discarded.

The PB profiles were obtained by translating the HOMSTRAD structural MSAs into PB MSAs using the atomic coordinates of the protein structures (de Brevern *et al.*, 2000). PB profiles are weighted in the same way as for AA profiles. Such profiles based on structural alphabet provide structural evolutionary information, which is expected to improve the detection of remote homologies.

2.3 Generating query AAPB profiles

MSAs were obtained by three iterations of PSI-BLAST on Uniref90 (Suzek *et al.*, 2007) (a subset of the UniProt database containing no pair of sequences with >90% identity) with an E-value threshold of 10^{-4}. At most 1000 alignments were saved at each round. Each profile was built from the last round of PSI-BLAST. MSAs were purged by removing sequences with more than 70% identity.

PB profiles are generated using the LOCUSTRA program (Zimmermann and Hansmann, 2008), which predicts a sequence of PB from a PSI-BLAST MSA, using a two-layer support vector machine (SVM) approach. The input is a PSSM matrix of the 20 residue types extracted from the checkpoint file generated by PSI-BLAST. We modified LOCUSTRA to obtain a PB profile of 16 columns corresponding to the SVM multiclass posterior probabilities for the 16 letters.

2.4 Profile–profile scoring function

The PICASSO scoring function introduced by Heger and Holm (2001) has been shown to be one of the most efficient scoring functions for comparing two amino acid profiles. Mittelman *et al.* (2003) modified it into several variants. The one we chose here is PICASSO3Q (1), which uses only the query and template frequencies *Qi* and *Qj*, from query position *i* and template position *j* of the profiles in symmetric Equations (2, 3):

$$S(i,j) = \frac{S_{AA}(i,j) + S_{PB}(i,j) - \tilde{s}p}{\sigma_{sp}} + \text{shift} \tag{1}$$

$$S_{AA}(i,j) = \sum_{aax=1}^{20} Q^i_{aax} \log_2 \frac{Q^j_{aax}}{p_{aax}} + \sum_{aax=1}^{20} Q^j_{aax} \log_2 \frac{Q^i_{aax}}{p_{aax}} \tag{2}$$

$$S_{PB}(i,j) = \sum_{pby=1}^{16} Q^i_{pby} \log_2 \frac{Q^j_{pby}}{p_{pby}} + \sum_{pby=1}^{16} Q^j_{pby} \log_2 \frac{Q^i_{pby}}{p_{pby}} \tag{3}$$

where *Qaa* and *Qpb* are the amino acid and PB frequencies, respectively. p_{aax} is the background frequency of the amino acid *x* and p_{pby} the background frequency of PB letter *y*. $S(i,j)$ is the sum of $S_{AA}(i,j)$ and $S_{PB}(i,j)$, which are the PICASSO3Q scores over the amino acid and structural alphabet profiles, respectively. The PICASSO3Q method was applied in the same way for PB and amino acid profiles. $S(i,j)$ is normalized by the median score pair *sp* and the SD $\sigma(sp)$ of $S(i,j)$ over all column pair (i.e. profile–profile) scores in the database. The median value appears to be preferable to the mean when there are a lot of extreme pair scores (Gonzales and Ottenbacher, 2001). In PICASSO3Q, the natural logarithm *ln* is used instead of $\log_2$, but we found that the latter performed slightly better (5% higher sensitity at 10% specificity, data not shown). A shift of 0.1 is added to the score to make average scores negative and consequently avoid the alignment of unrelated residues in local regions.

2.5 Dynamic programming

ORION carries out global/local dynamic programming to obtain the optimum global/local alignment. This variant of the Smith and Waterman's (1981) local/local and Needleman and Wunsch's (1970) global/global algorithm locally aligns the query against the whole template. We consider that query sequences may be composed of multiple domains that require to be locally aligned, while template family sequences are generally composed of a unique and well-delimited domain that must be globally aligned. It also established that global–local algorithm performs better for fold recognition (Fischer *et al.*, 1996). The optimal gap opening penalty *go* and the optimal gap extension penalty *ge* were determined by an iterative grid approach on a non-redundant subset of 402 sequences (i.e. the sequences having <30% sequence identity with their family members). Pairwise alignments were performed and all alignments obtained were compared with their respective structural alignments. The pair of values (*go*, *ge*) maximizing the percentage of correctly aligned positions was chosen as the optimal values. These optimal values ($go = 2$ and $ge = 0.2$) were set as default parameters. We have tested different subsets of HOMSTRAD and have observed that theses values are stable and do not depend of the subset composition.

2.6 Metafold: a supplementary level for remote protein structure similarities

The concept of metafold was introduced to characterize fold similarities (Day *et al.*, 2003). For example, authors clustered the TIM barrels and the beta propellers into the same fold family (Söding and Remmert, 2011), while these structures are classified in two different folds in SCOP. Metafold classification is obtained by a consensus of fold classifications in SCOP, CATH and Dali databases, and thus provides a more unified view of fold space. Here, we propose to use the metafold as an additional hierarchical level above the fold level of SCOP database. There are 422 metafolds groups in the HOMSTRAD database. As expected, immunoglobulin-like and Rossmann fold groups are the most populous ones.

2.7 Statistical significance of score alignments

The raw alignment score depends on the composition of the two profiles compared but also on the alignment length. For each family template, a score distribution has been precalculated. For a given family, the template family profile was aligned to a set of 422 profiles of unrelated proteins. Proteins used to obtain score distributions of unrelated targets must be structurally independent. Therefore, scores from protein pairs belonging to the same SCOP fold level or within the same metafold class were removed. Scores of amino acid sequence alignments usually follow an extreme value distribution, a Gumbel law in most cases. Here, score distributions show a good fitness to the Weibull distribution law (Supplementary Fig. S1). In fact, the addition of structural information into profiles modifies the score distributions that are close to distributions of threading scores (Fayyaz Movaghar *et al.*, 2011). ORION computes a *P* value from the raw score and the Weibull law parameters that are specific to the template family.

2.8 Benchmark datasets

ORION was compared with different programs to evaluate the distant homology detection rate. Therefore, a query test set has been constituted from the 1032 HOMSTRAD families. For each family, the representative sequence is selected based on the sequence identity between the family members, knowing that the representative sequence share the highest average sequence identity with the rest of the family. The SCOP annotation is hierarchical with the higher levels (family and superfamily levels) describing near evolutionary relationships; the fold level describes distant relationships and/or structural similarities. The metafold level is a consensus level that describes very distant protein similarities. To assess the programs for different degrees of protein relationship, three query datasets have been constituted, one for each level (superfamily, fold and metafold) by selecting sequences having at least one structure in HOMSTRAD with the same SCOP level. This selection resulted in 484, 413 and 434 sequences for the superfamily, fold and metafold levels, respectively. All the sets are non-redundant since 95% of sequences pairs have less than 18, 17 and 16% sequence identity (superfamily, fold and metafold level, respectively). The methods were independently tested for these different levels. HHsearch and PSI-BLAST were tested in the same way as for ORION. For HHsearch, we generated 'hhm' profiles for the queries and the database from MSAs of PSI-BLAST and HOMSTRAD famillies, respectively, using the hhmake program available in the HH-suite package (ftp://toolkit.genzentrum.lmu.de/pub/HH-suite/). The database of hhm profiles has been generated following the HH-suite user guide. HHsearch was run using the query hhm as input file and the HOMSTRAD database of hhm profiles as the input database. For PSI-BLAST, a three-iteration search was run in the HOMSTRAD enriched sequence database with the default E-value threshold for inclusion (0.002). For each query, the methods give a ranked list of proteins from the database. Proteins are ranked by E-value for PSI-BLAST, probability for HHsearch and raw score for ORION. The true positives (TPs) and false positives (FPs) were counted considering the classification level of interest. A TP was denoted when the two proteins compared (which define a 'hit') belong to the same class level (e.g. b.1.18.2 and b.1.10.1 at fold SCOP level) and a FP is counted for hits within different class levels (e.g. b.1.18.2 and b.2.6.1 at fold SCOP level). The hits with the same superior class levels were ignored (e.g. b.1.18.11 and b.1.18.2 at fold SCOP level). Each protein is then labeled as TP or FP in the ranked list. Similar benchmark procedures are employed by authors for assessing the template detection performance of their methods (Söding, 2005; Söding and Remmert, 2011; Xu *et al.*, 2014; Yang *et al.*, 2011).

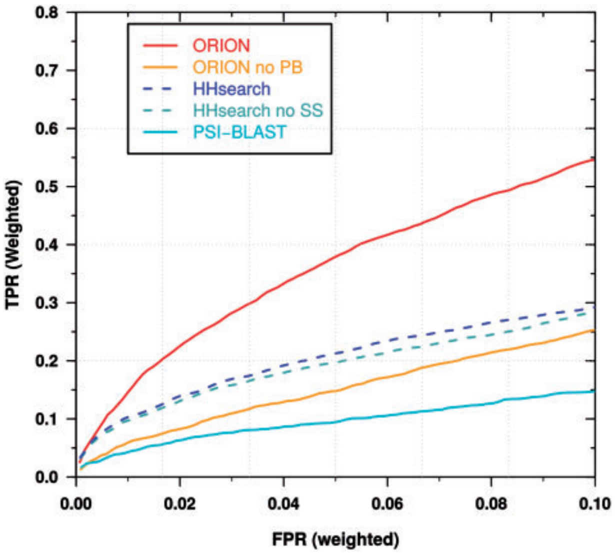

Fig. 2. Performance of ORION, HHsearch and PSI-BLAST at detecting related proteins within the same scop level (family, superfamily and fold levels are combined), for all pairs of the HOMSTRAD dataset. Performances of ORION and HHsearch without structural information ('ORION no PB' and 'HHsearch no SS', respectively) are also shown

Some SCOP classes in query test sets and template database may be overrepresented, and the results could suffer from this compositional bias. For example, the immunoglobulin fold and the Rossmann fold are the most represented in the HOMSTRAD database, which could bias the results for these families. To prevent these compositional biases from dominating the benchmarks, each template and query is weighted with the number of member belonging to the same level. Recently, Remmert *et al.* (2012) used a similar approach to weigh the folds that constitute their dataset.

3 Results and discussion

The distant homology detection rate of ORION has been assessed and compared with those of the widely used PSI-BLAST and the state-of-the-art method HHsearch. The results are described in the Fig. 2 by receiver operating characteristic (ROC) curves, which illustrate the performance of a binary classifier, here, the evolution of the true positive rate (TPR) in function of the false positive rate (FPR) ranging from 0 to 10%. As expected, a basic profile–profile method like ORION without PB outperforms a profile-sequence method like PSI-BLAST, with a TPR > 10% higher for 10% FPR. Not surprising either is the fact that the HMM-HMM profile alignments of HHsearch demonstrate higher performance than our amino acid profile–profile method (2.7% higher TPR for 10% FPR). However, the inclusion of PBs into ORION profiles leads to a significant increase of the homology detection rate, with 2.1 times more TP detected for 10% FPR, and makes our method outperform HHsearch and PSI-BLAST. Indeed, for 10% FPR, ORION is 1.8 times (0.55/0.30) and 3.6 times (0.55/0.15) more sensitive than

Table 1. Success rate (%) of ORION, HHsearch and PSI-BLAST at recognizing HOMSTRAD proteins within the same superfamily, fold or metafold levels

Method	Superfamily only			Fold only			Metafold only		
	1st	T5	T10	1st	T5	T10	1st	T5	T10
ORION	31.1	48.4	56.6	21.8	47.8	60.2	18.2	50.1	62.6
HHsearch	33.2	51.6	58.2	11.4	29.5	37.0	12.5	36.7	51.3
PSI-BLAST	13.5	22.2	37.6	4.1	12.7	28.2	10.9	29.8	52.0

ORION, ORION with PBs profiles; HHsearch, HHsearch with predicted secondary structure information.

HHsearch and PSI-BLAST, respectively. Thus, the structural features encoded into PBs improve particularly fold recognition and remote homology detection. These results could be explained by the fact that the 16 structural patterns catch almost every local conformation. Additionally, using structural information in the form of structural profiles, which take into account the frequency of each state at each position, gives to our fold recognition method a significant advantage.

Table 1 shows the proportion of proteins for which at least one correct template is found among the first five (Top 5) and first ten (Top 10) results, for proteins belonging to the same superfamily, fold and metafold levels. This kind of benchmark is particularly relevant for the evaluation of template selection for comparative modeling. ORION detects on average 17.3 and 10.3% more correct templates than PSI-BLAST and HHsearch respectively, in the top 1, 5 and 10. ORION is very efficient for detecting distant protein relationships (fold and metafold levels), with 20.4% and 13.7% more templates detected than PSI-BLAST and HHsearch, respectively. At superfamily level, HHsearch performs slightly (2.1%) better than ORION. Proteins within the same superfamily but in different families share a probable common evolutionary origin (Murzin *et al.*, 1995) with a higher sequence similarity than at fold level. Therefore, the sensibility of sequence methods is still sufficient to detect a relationship between two homolog proteins at this level of divergence.

Similar results are observed with the ROC curves at superfamily, fold and metafold levels (Fig. 3). At superfamily level, HHsearch performs slightly better than ORION for very low FPR (0–1%), but for higher FPR (1–10%) ORION clearly outperforms HHsearch. As for top ranking, ORION outperforms HHsearch and PSI-BLAST at fold and metafold levels and the difference becomes greater with the inclusion of more templates. All methods have the lowest performances at metafold level, since this level is based on a higher structural diversity among metafold classes, resulting in harder detection. These results suggest that ORION is effective for very remote homolog detection with a high success rate. At fold and metafold levels, true hits (i.e. correct detection of two related proteins) are very distant, with <15% sequence identity and, hence, cannot be easily identified solely from their amino acid sequences. Therefore, methods including additional structural information, such as ORION, allow better detection at these levels.

The CASP targets dataset was constituted by selecting the targets of CASP8, 9 and 10 belonging to the ‘Template-based modeling’ category. We kept targets with at least one template in the HOMSTRAD database within the same fold level. Thus, our CASP dataset contains 150 targets: 49 (out of 120) from CASP8, 72 (out of 111) from CASP9 and 29 (out of 67) from CASP10. This dataset is composed of ‘Hard’ difficulty targets, since 95% of the pairs share <15% sequence identity and the true positive pair maximum sequence identity is 21.1%. The related ROC curves (Fig. 4) show that ORION performs better than PSI-BLAST and HHsearch at detecting template structures. For 10% FPR, ORION detects 2.1 times more related proteins than HHsearch and 2.5 times more than PSI-BLAST.

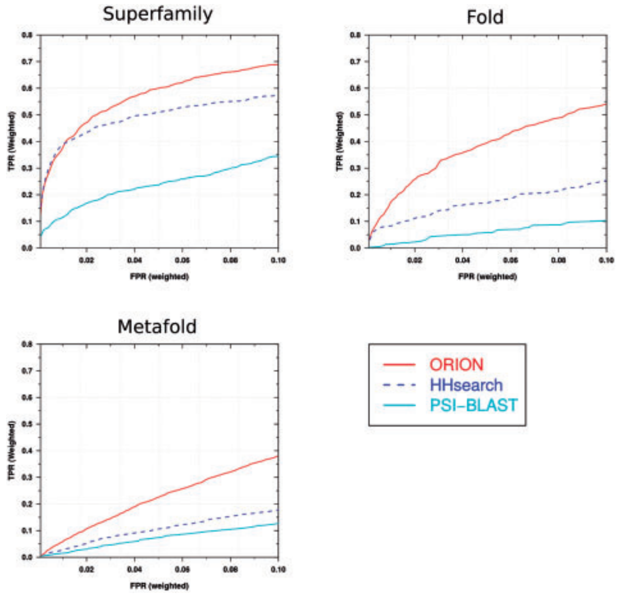

Fig. 3. Performance of ORION, HHsearch and PSI-BLAST at detecting related proteins within the same superfamily, fold and metafold levels for all pairs of the HOMSTRAD dataset

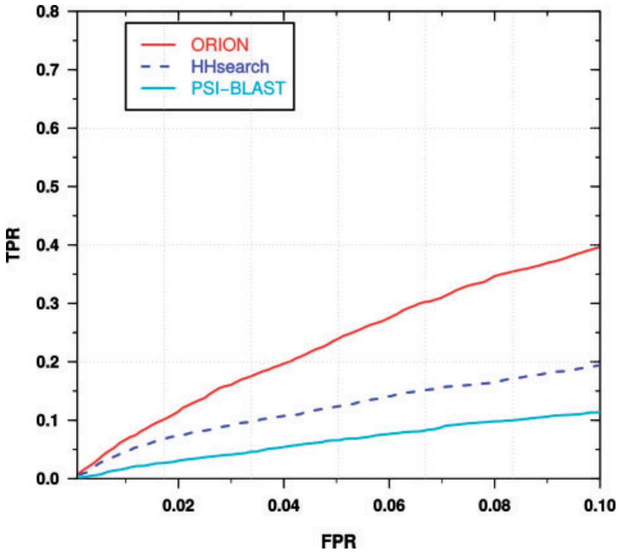

Fig. 4. Performance of ORION, HHsearch and PSI-BLAST at detecting related proteins within the same fold, for all pairs of the CASP dataset

This holds true for the top ranking comparisons (Table 2), with 8.6 and 22.9% more homologs detected at the top 10, compared with HHsearch and PSI-BLAST, respectively. Results of ROC curves and top rankings on CASP dataset are similar to those obtained from the HOMSTRAD dataset, except that lower performances are observed due to the overall difficulty of the CASP targets.

Six examples of CASP11 targets are presented to underline the importance of the template selection in the protein structure prediction (Table 3). Models were generated with MODELLER (Eswar et al., 2006) based on templates and alignments provided by ORION and HHsearch and were evaluated by the GDT-HA (Read and Chavali, 2007), GDT_TS (Zemla et al., 1999) and TM-scores (Zhang and Skolnick, 2004). The results show that the templates selected by ORION and HHsearch are different. Indeed, ORION selects better templates than HHsearch, which significantly increases the GDT_TS (improvement ranging from 0.7% for T0829-D1 up to 18.7% for T0855-D1). Thus, ORION provides better templates for generating models that are closer to the target structure. For T0784-D1 and T0847-D1 models, ORION and HHsearch converged to identical templates (PDB codes 1byrA and 2lrgA) but the additional templates selected by ORION lead to better models than those obtained from HHsearch (4.4 and 2.6% of GDT_TS increases). The prediction of the 'Easy' target T0773-D1, based on the three templates found by ORION (PDB codes 1mlaA, 2jsxA and 1in0A), is close to the target structure, with the same orientation of beta sheets and alpha helices (Fig. 5A). For the 'Medium' T0829-D1 and 'Hard' T0855-D1 targets, the prediction of α helices positioning is more difficult but the pattern of beta strands is correct (Fig. 5B and C).

4 Conclusion

We have developed ORION, a profile–profile alignment method for fold recognition based on a structural alphabet that improves remote homology detection. ORION is able to detect between 2 and 10% more homologs than the state-of-the-art method HHsearch on a balanced test set derived from the HOMSTRAD database. On another test set of target sequences from CASP8, 9 and 10, ORION systematically outperforms HHsearch and PSI-BLAST, detecting, respectively, 4.6 and 16.0% more templates at fold level (Table 2, mean difference in success rate for top 1, 5 and 10). Our method works particularly well for distantly related proteins due to the addition of accurate predictions of local structural information in the form of PBs.

Table 2. Success rate (%) of ORION, HHsearch and PSI-BLAST at recognizing proteins from the CASP dataset within the same fold level

Method	Fold only		
	First	Top 5	Top 10
ORION	10.1	26.2	40.3
HHsearch	7.6	23.5	31.7
PSI-BLAST	2.8	8.3	17.4

ORION, ORION with PBs profiles; HHsearch, HHsearch with predicted secondary structure information.

A possible improvement of ORION would be to incorporate other accurately predicted structural features, such as solvent accessibility. The combination of such structural descriptors with the PBs should provide a good approximation of the environment

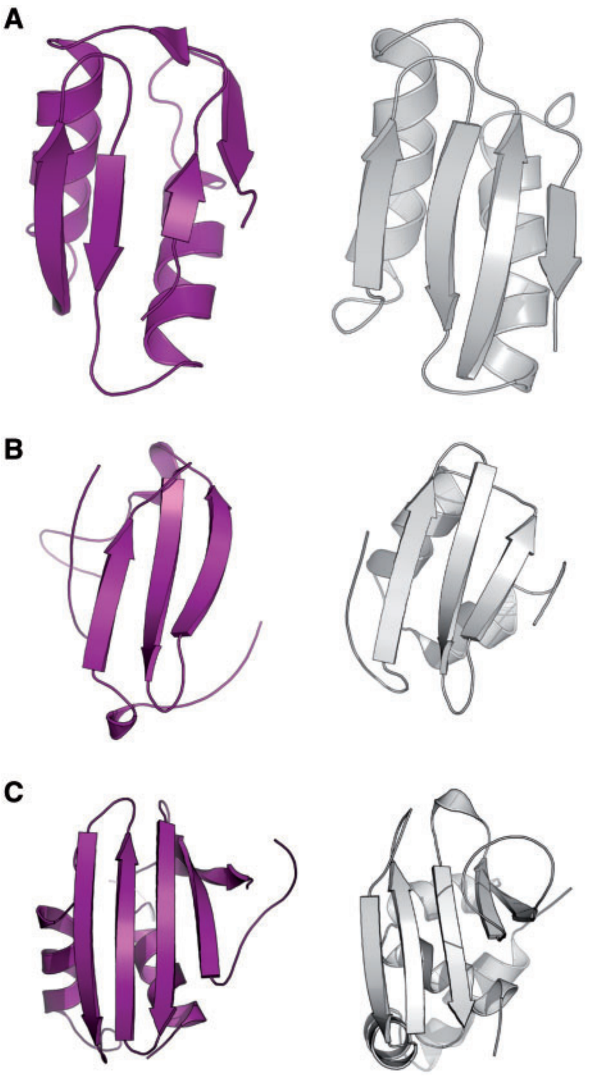

Fig. 5. Examples of protein structure predictions based on templates detected by ORION. The structural models were obtained with MODELLER, using ORION sequence alignments. CASP targets (on the right) are colored in grey and models are on the left. Target and model structures were aligned with the TM-align program (Zhang and Skolnick, 2005). The three CASP11 targets are T0773-D1 (**A**), T0829-D1 (**B**) and T0855-D1 (**C**)

Table 3. Structural alignment scores of six targets from CASP11 with models generated from the templates ('Templ.') found by ORION and HHsearch

	ORION				HHsearch			
Target	Templ.	GDT-TS	GDT-HA	TM-score	Templ.	GDT-TS	GDT-HA	TM-score
T0829	3f5rA	49.63	35.45	0.46	4ifsA, 4khbD3k8rA	48.88	33.79	0.45
T0855	2vt8A	40.22	23.91	0.46	1cv8A3bbzA2ipqX	21.52	16.55	0.30
T0784	2lrgA 3u6gA	91.60	78.00	0.93	2lrgA 4hzuS	87.20	69.00	0.90
T0769	3e8oA1n5sA	71.91	51.29	0.75	3udcA1xhjA	60.83	38.92	0.67
T0773	1mlaA2jsxA 1in0A	72.39	51.49	0.73	2k49A2inpL 3hluA	66.79	47.02	0.64
T0847	1byrA 1bysA2f5tx	73.52	55.77	0.80	1byrA 4gelA 4ggiA	70.86	52.52	0.78

preferences for each amino acid of the target protein sequence. This information is helpful to predict structural core positions and, therefore, identify more remote homologs. For example, a series of central buried alpha helices could be identified from the target sequence with a combination of the predicted PBs and solvent accessibility descriptors, these latter restricting the number of possible template hits.

Another possible improvement would touch on the weights of the different terms in the scoring function. Optimal weights were obtained from an evaluation of the successful detection rate on a balanced HOMSTRAD test set. This assessment indicates that the sequence and structural scoring terms should be equally weighted for the highest success rate (Supplementary Fig. S2). However, the weights could be dynamically adjusted according to the target sequence. Thus, while close homologs detection is based on the sequence, very remote homologs detection requires using more structural information, since the structure is more conserved than the sequence throughout the evolution. To address the issue of the weighting of the scoring terms, ORION may evaluate the difficulty of the detection in the initial search and then adjust the different parameters and perform another search in the template library. Finally, ORION and HHsearch show significant differences in template detection suggesting that they are complementary. Thus, ORION would also contribute to the improvement of the quality of meta-servers and other consensus-based prediction algorithms.

Acknowledgement

We thank Dr Olav Zimmermann and Dr Ulrich H.E. Hansmann for providing the LOCUSTRA software.

Funding

This work was supported by grants from INSERM, Paris Diderot University, Sorbonne Paris Cité, Institut National de la Transfusion Sanguine and Laboratory of Excellence GR-Ex.

Conflict of Interest: none declared.

References

Altschul,S.F. *et al*. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

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

Bowie,J.U. *et al*. (1991) A method to identify protein sequences that fold into a known three-dimensional structure. *Science*, **253**, 164–170.

Brenner,S.E. *et al*. (1998) Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. *PNAS*, **95**, 6073–6078.

de Brevern,A.G. *et al*. (2000) Bayesian probabilistic approach for predicting backbone structures in terms of protein blocks. *Proteins*, **41**, 271–287.

Day,R. *et al*. (2003) A consensus view of fold space: combining SCOP, CATH, and the Dali domain dictionary. *Protein Sci.*, **12**, 2150–2160.

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

Elofsson,A. (2002) A study on protein sequence alignment quality. *Proteins*, **46**, 330–339.

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

Fayyaz Movaghar,A. *et al*. (2011) Statistical significance of threading scores. *J. Comput. Biol.*, **19**, 13–29.

Fischer,D. *et al*. (1996) Assessing the performance of fold recognition methods by means of a comprehensive benchmark. *Pac. Symp. Biocomput.*, 300–318.

Gelly‘,J.-C. *et al*. (2011) iPBA: a tool for protein structure comparison using sequence alignment strategies. *Nucleic Acids Res.*, **39**, W18–W23.

Gonzales,V.A. and Ottenbacher,K.J. (2001) Measures of central tendency in rehabilitation research: what do they mean? *Am. J. Phys. Med. Rehabil.*, **80**, 141–146.

Gribskov,M. *et al*. (1987) Profile analysis: detection of distantly related proteins. *Proc. Natl Acad. Sci. USA*, **84**, 4355–4358.

Heger,A. and Holm,L. (2001) PICASSO: generating a covering set of protein family profiles. *Bioinformatics*, **17**, 272–279.

Henikoff,J.G. and Henikoff,S. (1996) Using substitution probabilities to improve position-specific scoring matrices. *Comput. Appl. Biosci.*, **12**, 135–143.

Henikoff,S. and Henikoff,J.G. (1994) Position-based sequence weights. *J. Mol. Biol.*, **243**, 574–578.

Illergård,K. *et al*. (2009) Structure is three to ten times more conserved than sequence—a study of structural response in protein cores. *Proteins*, **77**, 499–508.

Jones,D.T. *et al*. (1992) A new approach to protein fold recognition. *Nature*, **358**, 86–89.

Joseph,A.P. *et al*. (2010) A short survey on protein blocks. *Biophys. Rev.*, **2**, 137–145.

Karplus,K. *et al*. (1998) Hidden Markov models for detecting remote protein homologies. *Bioinformatics*, **14**, 846–856.

Kelley,L.A. *et al*. (2000) Enhanced genome annotation using structural profiles in the program 3D-PSSM1. *J. Mol. Biol.*, **299**, 501–522.

Kennedy,D. and Norman,C. (2005) What don’t we know? *Science*, **309**, 75.

Krogh,A. *et al*. (1994) Hidden Markov models in computational biology: applications to protein modeling. *J. Mol. Biol.*, **235**, 1501–1531.

McGuffin,L.J. (2008) Protein fold recognition and threading. In: Schwede,T. and Peitsch,M.C. (eds) *Computational Structural Biology: Methods and Applications 1st Edition*. World Scientific Publishing Co Pte Ltd, pp. 37–60.

Mittelman,D. *et al*. (2003) Probabilistic scoring measures for profile–profile comparison yield more accurate short seed alignments. *Bioinformatics*, **19**, 1531–1539.

Mizuguchi,K. *et al*. (1998) HOMSTRAD: a database of protein structure alignments for homologous families. *Protein Sci.*, **7**, 2469–2471.

Moult,J. *et al*. (2009) Critical assessment of methods of protein structure prediction Round VIII. *Proteins*, **9**, 1–4.

Moult,J. *et al*. (2011) Critical assessment of methods of protein structure prediction (CASP)—round IX. *Proteins*, **79**, 1–5.

Moult,J. *et al*. (2014) Critical assessment of methods of protein structure prediction (CASP) — round x. *Proteins*, **82**, 1–6.

Müller,A. *et al*. (1999) Benchmarking PSI-BLAST in genome annotation. *J. Mol. Biol.*, **293**, 1257–1271.

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.

Ohlson,T. *et al*. (2004) Profile–profile methods provide improved fold-recognition: a study of different profile–profile alignment methods. *Proteins*, **57**, 188–197.

Panchenko,A.R. (2003) Finding weak similarities between proteins by sequence profile comparison. *Nucleic Acids Res.*, **31**, 683–689.

Park,J. *et al*. (1998) Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. *J. Mol. Biol.*, **284**, 1201–1210.

Pearson,W.R. (1991) Searching protein sequence libraries: comparison of the sensitivity and selectivity of the Smith-Waterman and FASTA algorithms. *Genomics*, **11**, 635–650.

Rangwala,H. *et al*. (2009) svmPRAT: SVM-based protein residue annotation toolkit. *BMC Bioinformatics*, **10**, 439.

Read,R.J. and Chavali,G. (2007) Assessment of CASP7 predictions in the high accuracy template-based modeling category. *Proteins*, **69**(Suppl. 8), 27–37.

Remmert,M. *et al*. (2012) HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat. Methods*, **9**, 173–175.

Rost,B. (1999) Twilight zone of protein sequence alignments. *Protein Eng.*, **12**, 85–94.

Rychlewski,L. *et al*. (2000) Comparison of sequence profiles. Strategies for structural predictions using sequence information. *Protein Sci.*, **9**, 232–241.

Shi,J. *et al.* (2001) FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties1. *J. Mol. Biol.*, **310**, 243–257.

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

Söding,J. and Remmert,M. (2011) Protein sequence comparison and fold recognition: progress and good-practice benchmarking. *Curr. Opin. Struct. Biol.*, **21**, 404–411.

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

Xu,D. *et al.* (2014) FFAS-3D: improving fold recognition by including optimized structural features and template re-ranking. *Bioinformatics*, **30**, 660–667.

Yang,Y. *et al.* (2011) Improving protein fold recognition and template-based modeling by employing probabilistic-based matching between predicted one-dimensional structural properties of the query and corresponding native properties of templates. *Bioinformatics*, **27**, 2076–2082.

Zemla,A. *et al.* (1999) Processing and analysis of CASP3 protein structure predictions. *Proteins*, **37**, 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.

Zimmermann,O. and Hansmann,U.H.E. (2008) LOCUSTRA: accurate prediction of local protein structure using a two-layer support vector machine approach. *J. Chem. Inf. Model.*, **48**, 1903–1908.