ORFeus: detection of distant homology using sequence profiles and predicted secondary structure

Krzysztof Ginalski, Jakub Pas, Lucjan S. Wyrwicz, Marcin von Grotthuss, Janusz M. Bujnicki and Leszek Rychlewski*

Bioinformatics Laboratory, BioInfoBank Institute, ul. Limanowskiego 24A, 60-744 Poznan, Poland

Received January 21, 2003; Revised and Accepted March 3, 2003

ABSTRACT

ORFeus is a fully automated, sensitive protein sequence similarity search server available to the academic community via the Structure Prediction Meta Server (http://BioInfo.PL/Meta/). The goal of the development of ORFeus was to increase the sensitivity of the detection of distantly related protein families. Predicted secondary structure information was added to the information about sequence conservation and variability, a technique known from hybrid threading approaches. The accuracy of the meta profiles created this way is compared with profiles containing only sequence information and with the standard approach of aligning a single sequence with a profile. Additionally, the alignment of meta profiles is more sensitive in detecting remote homology between protein families than if aligning two sequence-only profiles or if aligning a profile with a sequence. The specificity of the alignment score is improved in the lower specificity range compared with the robust sequence-only profiles.

INTRODUCTION

Detection of homology between proteins based on similarity of their sequences can provide a basis for functional predictions for not annotated protein families. However, protein sequences diverge rapidly due to accumulation of amino acid substitutions, hampering the detection of similarity based on pairwise comparisons between many remote homologs. The effectiveness of identification of distant protein relationships has been greatly increased since the introduction of a database search strategy utilizing sequence alignments/profiles as queries instead of simple sequences, as implemented in the broadly used Position-Specific Iterated BLAST (PSI-BLAST) program (1–3). Further significant improvement was made possible owing to the use of sequence profiles for both the query protein and every protein from the database, as implemented in FFAS (4) and in a recently published tool based on information theory (5). ORFeus, the method presented here, incorporates a technique known from fold recognition algorithms. Predicted secondary structure is added to the scoring function, which compares sequence profiles representing potentially homologous protein families. Fold recognition methods (6–9) compare the predicted secondary structure of the query protein with an experimentally determined secondary structure of the protein with known fold. In contrast to this procedure, the predicted secondary structure is used by ORFeus for both the query and the template. This symmetric approach introduced an expected error of over 20% in the description of the secondary structure of the template but has one crucial advantage: protein families with unknown tertiary structure can also be included in the template database. This results in an over 10-fold expansion of the applicability of the algorithm, because most of the known protein sequences lack an experimentally determined structure assignment.

Algorithm

The secondary structure prediction is stored in the form of a profile of probabilities. ORFeus can utilize any secondary structure prediction method that produces estimated probabilities for local structure described using three states, that is, the helix, the beta sheet and the loop. Currently the values produced by PSIPRED (10) are used. The sequence profiles are generated as in FFAS (4). The main difference is that all the vectors of probabilities for the occurrence of all amino acids at each position are normalized using the \( p = 1 \) norm (the sum of all 20 values is equal to 1). The similarity between two positions (elements of the dynamic programming matrix) equals the shifted dot product of the sequence profile vector plus the shifted dot product of the secondary structure probability vector multiplied by the secondary structure weight.

\[
S_{ij} = \sum_{k=1}^{20} (seq_{ik} \cdot seq_{jk}) - Z_{seq} + w \cdot \left[ \sum_{k=1}^{3} (str_{ik} \cdot str_{jk}) - Z_{str} \right]
\]

where \( S_{ij} \) = score of aligning two positions, \( seq_{ik} \) = value for the amino acid \( k \) in the sequence profile in position \( i \) (first protein), \( seq_{jk} \) = value for the amino acid \( k \) in the sequence

*To whom correspondence should be addressed. Tel: +48 618653520; Fax: +48 618132606; Email: leszek@bioinfo.pl
residues). All the secondary structure and sequence vectors below zero and the template protein alignment score. FFAS also transforms the alignment score expressed in normalized scores, but represents only a raw normalization of the dynamic programming matrix is constrained to two positions is below zero. In contrast with FFAS, no penalty, a weight for the contribution of the secondary structure improves the accuracy problem known from earlier FFAS versions, proteins was used for the development of parameters to reduce classes) into 118 targets. The benchmark of two-domain proteins was used for the development of parameters to reduce the accuracy problem known from earlier FFAS versions, indicating that the incorporation of the predicted secondary structure improvements the accuracy of the prediction method even though the secondary structure profiles are calculated.

The combined local alignment of two sequence profiles and two secondary structure profiles conducted by ORFeus requires five parameters: gap initiation penalty, gap extension penalty, a weight for the contribution of the secondary structure profiles and two values, which shift the expected dot product of the secondary structure and sequence vectors below zero (expected score of aligning two vectors representing two residues). All five parameters were selected using brute-force optimization on a test set of artificially constructed two-domain families.

The set was based on sequence families extracted from SCOP (11), version 1.55. A set of 472 domains was chosen. The set was divided into two equal groups, so that no fold was represented in both groups (representatives of a fold are either in one or the other group, not in both). In all, 236 proteins in each group were used to create artificial two-domain proteins by concatenating two members (always from different fold classes) into 118 targets. The benchmark of two-domain proteins was used for the development of parameters to reduce the accuracy problem known from earlier FFAS versions, where two-domain proteins had the tendency to be predicted as similar to other unrelated two-domain proteins. The optimization was conducted on one set and the other set was used for the evaluation. A genetic algorithm was used to evolve and improve the parameters. To increase the speed of the optimization the dynamic programming matrices of all 6903 pairs of targets were kept in memory, using a total of 4 Gb RAM on eight dual Pentium®III computers. The new parameters were used only to find the optimal local alignment on a pre-calculated set of dynamic programming matrices.

Two types of scoring functions were used for the optimization of parameters aimed at improving the sensitivity and the specificity of the prediction, respectively. The total sensitivity score for the test set was measured as the sum of prediction scores over all 118 targets. Each prediction score, calculated for each target, is the sum of all correct hits scaled by the number of wrong hits with higher alignment score: 1. if targets t and i share same fold 0, else

$$\text{Score}_t = \sum_{i=1}^{117} \begin{cases} 1, & \text{if targets } t \text{ and } i \text{ share same fold} \\ 0, & \text{else} \end{cases} \frac{\text{number of false hits to this target with higher alignment score}}{\text{number of false hits with higher alignment score}}$$

The specificity score was calculated in a similar manner but with all 118-117 alignment scores evaluated simultaneously:

$$\text{Specificity} = \sum_{i,j} \begin{cases} 1, & \text{if targets } i \text{ and } j \text{ share same fold} \\ 0, & \text{else} \end{cases} \frac{\text{number of false hits with higher alignment score}}{\text{number of false hits with higher alignment score}}$$

Table 1. Optimization results using sensitivity or specificity as scoring function

<table>
<thead>
<tr>
<th></th>
<th>Sequence and 2D profiles&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Sequence profiles only&lt;sup&gt;b&lt;/sup&gt;</th>
<th>2D weight&lt;sup&gt;c&lt;/sup&gt;</th>
<th>Other parameters&lt;sup&gt;d&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Specificity&lt;sup&gt;e&lt;/sup&gt;</td>
<td>721</td>
<td>705</td>
<td>0.06</td>
<td>0.08 0.60 0.02 0.63</td>
</tr>
<tr>
<td>Sensitivity&lt;sup&gt;f&lt;/sup&gt;</td>
<td>806</td>
<td>749</td>
<td>0.30</td>
<td>0.08 0.60 0.05 0.70</td>
</tr>
<tr>
<td>Limit&lt;sup&gt;g&lt;/sup&gt;</td>
<td>1038</td>
<td>1038</td>
<td>—</td>
<td></td>
</tr>
</tbody>
</table>

<sup>a</sup>Scores obtained by the profiles used in ORFeus.
<sup>b</sup>Results obtained when the predicted secondary structure profiles were not used (2D structure weight equal to zero).
<sup>c</sup>2D weight: weight of the secondary structure profile when computing the match score between two positions during dynamic programming. The values indicate a 5-fold increase for the importance of the predicted secondary structure when optimized for sensitivity.
<sup>d</sup>Other parameters: zero shift of the sequence profile score, initiation and extension gap penalty and zero shift of the secondary structure profile score.
<sup>e</sup>Results obtained using the parameter set optimized for specificity.
<sup>f</sup>Results obtained using the parameter set optimized for sensitivity.
<sup>g</sup>The best theoretically possible results.
only on the sequence profiles and do not utilize the experimental data.

The parameters optimized for highest sensitivity were chosen in the final ORFeus implementation, because the improvement of sensitivity over sequence-only profiles was more profound.

IMPLEMENTATION

An independent test was conducted on a set containing 1713 family representatives extracted from the current SCOP version 1.57 (representative sequences longer than 600 residues or shorter than 50 residues were removed). Figure 1 shows the number of correctly predicted superfamily relationships as a function of the number of false predictions with higher alignment score. Only one top-scoring prediction for each family is taken into account. This corresponds to the common procedure of specificity evaluation conducted in the LiveBench program (where the evaluated prediction methods use different fold libraries). The performance of ORFeus optimized for sensitivity is compared with a version of the program where only the sequence part of the profile is utilized (optimized with the secondary structure weight equal to zero) and with the PSI-Blast program used also to create the sequence alignments utilized in the profile building procedure. The results show that the more complex meta profiles that utilize predicted secondary structure preferences are more specific than the simple sequence-only profiles in the low specificity range where more than 10% of errors are expected.

However, the main advantage of meta profiles is their sensitivity. This is demonstrated in Figure 2. The number of correct predictions (with top rank) is plotted for all three prediction procedures. The data confirms again the superiority of aligning sequence profiles over the alignment of a profile with a sequence (PSI-Blast) (12). The alignment of meta profiles conducted by ORFeus is able to boost the sensitivity even further, providing an additional 50% improvement compared with the difference in sensitivity between the other two methods.

The initial comparison with other prediction methods was conducted using the ToolShop service (13). On the ToolShop-2 set ORFeus ranked second in the total sensitivity evaluation of difficult targets and second in total specificity for all targets. Both results were obtained using parameters optimized for sensitivity in November 2001. The only method showing better performance was a consensus predictor Pcons, which combines results obtained from several fold recognition servers.
A detailed analysis of the performance is available on the ToolShop pages (13). Owing to the nature of the ToolShop, the result obtained by a server at a later point in time may be overly-optimistic. The quick growth of sequence databases provides an artificial advantage for predictions that are conducted later. A more rigorous evaluation was conducted in the LiveBench program (15,16). The completed fourth session confirmed the utility of the presented method (http://BioInfo.PL/LiveBench/4). In the sensitivity evaluation the ORFeus server was only ranked behind novel consensus methods, which utilize several servers or server components to create a jury prediction. The consensus approach is known to result in increased accuracy compared with individual (single-template) prediction methods. From the individual servers only INBGU (8) showed better results than ORFeus on the difficult (HARD) targets, while no individual server was more sensitive than ORFeus on the easy targets. FFAS (4), the sequence-only profile alignment method, was the only individual server that showed higher specificity than ORFeus (a result expected in agreement with prior test results).

The performance of ORFeus was also confirmed in the last CAFASP-3 (17) blind prediction experiment. In the evaluation of autonomous servers ORFeus (the low PSI-Blast iteration version, ORFeus-B) obtained the first rank in the homology modeling sensitivity category and the third rank in the fold recognition category where the second rank was obtained by a server [SHGU (18)] that utilizes consensus building and fragment splicing technology.

CONCLUSIONS

The addition of predicted secondary structure to conventional sequence profiles is able to boost the sensitivity of profile–profile comparison methods substantially. This addition is, however, accompanied by a serious distortion of the alignment score distribution. The increase of sensitivity should result in an increase of specificity in our benchmarks since more correct predictions are expected in total. This has not happened and the specificity of sequence-only profiles remains on a similar level as that of the meta profiles. In particular, in high specificity ranges the conventional sequence-only profiles remain more robust.

The currently best way to boost the specificity of predictions is the application of consensus methods. ORFeus will become a valuable component of such methods providing a high number of correct family assignments despite limited specificity of the alignment score. ORFeus has already been incorporated in newer Pcons/Modeller versions (14).

ACCESS

ORFeus is available to the academic community via the convenient Structure Prediction Meta Server (http://BioInfo.PL/Meta) or through the experimental higher throughput GRDB system (http://grdb.bioinfo.pl). A commercial standalone version of the program is available upon request. ORFeus is also coupled to the continuous online server evaluation program, LiveBench (http://BioInfo.PL/LiveBench/).

ACKNOWLEDGEMENTS

This work was supported in part by the 5th Framework Programme grant QLK3-CT-2000-00170.

REFERENCES