Profile-Profile Alignment: A Powerful Tool for Protein Structure Prediction

N. von Öhsen, I. Sommer, R. Zimmer

Pacific Symposium on Biocomputing 8:252-263(2003)

PROFILE-PROFILE ALIGNMENT: A POWERFUL TOOL FOR PROTEIN STRUCTURE PREDICTION

NIKLAS VON ÖHSEN
FhI-SCAI - Fraunhofer Institute for Algorithms and Scientific Computing, Schloss Birlinghoven, 53754 Sankt Augustin, Germany

INGOLF SOMMER
Max-Planck-Institut für Informatik, Stuhlsatzenhausweg 85, 66123 Saarbrücken, Germany

RALF ZIMMER
Institut für Informatik, LMU München, Theresienstrasse 39, 80333 München, Germany

Abstract

The problem of computing the tertiary structure of a protein from a given amino acid sequence has been a major subject of bioinformatics research during the last decade. Many different approaches have been taken to tackle the problem, the most successful of which are based on searching databases to identify a similar amino acid sequence in the PDB and using the corresponding structure as a template for modeling the structure of the query sequence. An important advance for the evaluation of sequence similarity in this context has been the use of a frequency profile that represents a part of the protein sequence space close to the query sequence instead of the query sequence itself. In this paper, we present a further extension of this principle by using profiles instead of the template sequences, also. We show that, by using our newly developed scoring model, the profile-profile alignment approach is able to significantly outperform current state of the art methods like PSI-BLAST, HMMs, or threading methods in a fold recognition setup. This is especially interesting since we show that it holds for closely related sequences as well as for very distantly related ones.

Since the first use of alignment procedures for evaluation of similarity between two sequences, various successful advances in this concept have been proposed. One major improvement of the alignment procedure that made its way into many popular bioinformatics tools is the use of a frequency profile instead of a sequence which was first proposed by Gribskov[1] and is one of the major ingredients of the well-known PSI-BLAST program[2]. The aim of replacing a single sequence by a frequency profile representing its protein family is to discard part of the sequence information that is not conserved throughout this

family. Therefore, the profile will be a better representation of strongly conserved features like the tertiary structure of the protein than the sequence itself. While this concept proved useful when replacing one sequence in an alignment by a profile, it has been shown recently by Rychlewski et al.[3] that using profiles on both sides of the alignment is even better when trying to establish relationships between distantly related proteins. Using this profile-profile approach in their FFAS method, they managed to reach the second rank in the CAFASP2 contest of fully automated protein structure prediction servers. Despite the straight forward idea of replacing both sequences by frequency profiles, it is far from obvious how the alignment score should be calculated in this case. Rychlewski et al. used the simple dot product for computing the score but also noted that a more sophisticated method might prove advantageous.

A major new contribution to the profile-profile alignment approach has recently been made by Yona and Levitt[4] who were the first to propose a scoring formula for profile-profile alignments that was constructed on a theoretically sound basis. Their scoring system is based on an information theoretic measure of difference between the two probability distributions represented by the profiles. Since their profile-profile score is not constructed along the lines of the common similarity score for sequences, there are two major drawbacks to the approach: First, they measure only the similarity of the two probability distributions provided by the profiles and do not take into account the similarities between amino acids. Ignoring these contributions which have been crucial to sequence alignment methods for the last decades will most likely limit the sensitivity of an alignment scoring system. Second, they have to construct an ad-hoc transformation that will make their score applicable in the case of local sequence alignment. Our approach is an extension of the usual amino acid similarity score to the profile-profile situation (the sequence-sequence score is a special case of our formula) Thereby, our alignment score avoids both of these drawbacks and can directly be used for local alignment without any changes. Therefore, we believe that the proposed log average score may have significant practical and performance advantages. Yona and Levitt compare their profile-profile tool with tools from the BLAST family like Gapped BLAST, IMPALA and PSI-BLAST, showing that they can detect more protein superfamily relationships using their method than using PSI-BLAST. Due to lack of availability at the time of setting up the benchmarks performed in this paper, we have not been able to include their `prof_sim` tool in this study.

In the following, we are presenting a formula for applying the popular similarity matrix scoring for sequences to the more general case of scoring two frequency profiles which has been introducede in more detail earlier[5] and compare the performance of the new alignment score with other popular alignment

methods in terms of fold recognition performance.

1 Introducing Log Average Scoring

The popular similarity matrix alignment scores like PAM[6] or BLOSUM[7] have a strong foundation in statistical test theory. The alignment score of an alignment of two sequences without gap penalties is a direct measure of the statistical evidence supporting the hypothesis that the two sequences are related against the alternative that they are unrelated. The definition of the term "related" in this context is part of the used substitution model. The most interesting features of alignment scores are an immediate consequence of this property. A score of zero means that there is no evidence in the alignment whether the sequences are related or unrelated whereas a positive score indicates relatedness and a negative score indicates that the sequences are rather unrelated. If p_i is the background amino acid probability distribution and p_{rel} denotes the probability distribution of "related" amino acid pairs, the substitution matrix alignment score for a pair (i, j) of amino acids is calculated by

$$M(i,j) = \log\left(\frac{p_{\mathrm{rel}}(i,j)}{p_i p_j}\right) \tag{1}$$

This is usually scaled by a factor $\frac{10}{\log 10}$ in the Dayhoff models and $\frac{2}{\log 2}$ for the BLOSUM matrices. While it seems straightforward to extend this score to two profile vectors α and β by using the formula

$$\mathrm{score}_{\mathrm{average}}(\alpha,\beta) = \sum_{i=1}^{20}\sum_{j=1}^{20} \alpha_i\beta_j \log\frac{p_{\mathrm{rel}}(i,j)}{p_i p_j} \tag{2}$$

(called *average scoring*) we have shown earlier[5] that the original meaning of the alignment score can be extended to the profile-profile case by using the *log average score*:

$$\mathrm{score}_{\mathrm{logaverage}}(\alpha,\beta) = \log\sum_{i=1}^{20}\sum_{j=1}^{20} \alpha_i\beta_j \frac{p_{\mathrm{rel}}(i,j)}{p_i p_j} \tag{3}$$

The double sum occuring here can be interpreted as a bayesian probability for the profile vectors being related according to the substitution model, thus giving a meaning to the sum of this score over all alignment positions.

2 Benchmarks

In order to evaluate whether the proposed profile-profile alignment scores are useful for measuring the relatedness of two profiles, we performed several tests measuring fold recognition performance. The SCOP-1.50 database[8] was taken as gold standard for measuring the relatedness of proteins on different levels. Using the ASTRAL server[9], a subset of this protein domain database was selected such that every two domains in the database were showing a maximum homology level of 40% sequence identity. This domain set is referred to as PDB40D[10] and its members will be called templates in the following.

2.1 Frequency Profile Construction

For each template sequence, a multiple alignment was constructed by running PSI-BLAST[2] against the KIND database[11], a non redundant protein sequence database. A frequency profile was calculated from this multiple alignment using a sequence weighting algorithm that is a slightly modified version of an algorithm by Henikoff[12]. [a] The resulting frequency profile was cut down to the original template length by throwing away positions that correspond to a gap of the template sequence in the multiple alignment.

This template database of frequency profiles was used in a fold recognition setup: The objective is to find the most closely related protein domain in the PDB40D set for a given protein chain (in the following also called *target*). This is done by first constructing a frequency profile for the target using PSI-BLAST and sequence weighting methods exactly as described for the templates. Then, all the template frequency profiles are subsequently aligned against the query frequency profile and the template with the highest score is the best guess for the structure of a domain contained in this chain.

2.2 Data Set

The performance of some state-of-the-art algorithms for fold recognition were compared with the newly developed profile-profile alignment scoring formula using a modified "leave-one-out" benchmark. All 2232 protein chains from the PDB containing one complete domain from the PDB40D set were used as benchmark set. When performing the fold recognition for each chain, all the domains belonging to the chain itself were removed from the PDB40D template

[a] The extended version tries to minimize the relative entropy regarding the background amino acid distribution rather than to maximize the absolute entropy of the profile, leading to small differences in the profile as compared to the Henikoff version. See also Krogh et. al.[13] for the connection between sequence weighting and entropy.

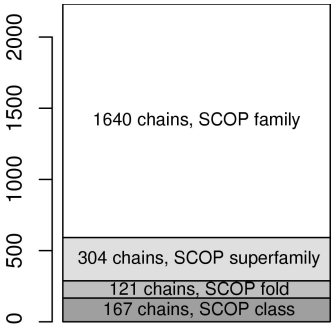

Figure 1: Test set composition: number of chains with the most closely related PDB40D domain belonging to the same SCOP level as indicated.

set. Each of the chains was aligned against the PDB40D template set except for the domains belonging to the current chain.

Figure 1 shows the composition of the benchmark set. Chains that contain a domain for which the PDB40D holds a member of the same SCOP family are likely to be quite easy fold recognition targets since SCOP family members ought to be quite closely related according to the definition of SCOP, making them tractable targets for fold recognition methods. Chains that have the most closely related domain on the SCOP superfamily level are harder to predict correctly, but according to the SCOP definition these domains probably have a probable common evolutionary origin. Hence it is possible that algorithms trying to find distant sequence similarities are successful in finding these relationships. The hardest level for fold prediction is the SCOP fold level since the templates in the PDB40D that share the same fold with a domain contained in the chain do not share more than a "major structural similarity" according to the SCOP definition. The SCOP class level finally is an impossible target for fold recognition by definition since all templates in our database have a SCOP fold different from the domains belonging to the chain. Nevertheless, these chains were taken into account when calculating the overall fold recognition percentages in order to get an unbiased estimation of the performance for a completely unknown target.

2.3 Algorithms, Implementations and Parameters

We used our Java implementation JProP of a dynamic programming engine using the two profile-profile alignment scores introduced in equations 2 and 3 to compare the performance of the profile-profile alignment with three other

successful fold prediction approaches: HMMs, threading and PSI-BLAST. In addition, we used a plain sequence alignment program to serve as a lower bound in the evaluation procedure.

Sequence Alignment: We used the 123D threading program to produce global dynamic programming sequence alignments using Dayhoff's 250 PAM matrix.

PSI-BLAST: PSI-BLAST[2] was run against the KIND database[11] augmented by the PDB40D proteins and the first hit in the PDB40D set that did not belong to the chain itself was taken as the fold prediction.

HMMer: We chose HMMer[14,15] in its latest version 2.2g as a representative for the current state of the art in profile HMMs (see Lindahl et al.[16] for a comparison of HMMer with SAM-T98 and other programs). A HMM database was built from the PDB40D list containing HMMs trained from the same multiple alignments used for constructing the frequency profiles. The search parameters of the HMM database were calibrated using the `hmmcalibrate` tool from the HMMer package. The target sequence was used to search this database using the `hmmpfam` tool. The e-value output of the found templates was used as score.

Profile 123D: 123D[17] is a fast profile threading program based on contact capacity potentials and dynamic programming which has been in use for some years already, e.g. in CAFASP2 and CASP4. It has recently been subject to parameter optimization[18] and the basis for an analysis of confidence measures[10]. We chose 123D as a representative for the class of threading algorithms tractable by dynamic programming. We used the parameters from the Zien et al. paper on parameter optimization using a machine learning approach[18]. Frequency profiles on the target side were used. Information on the tertiary structure coded in contact capacity potentials, secondary structure information and the sequence itself were used on the template side to produce a global alignment. The resulting threading score was used for further analysis.

JProP profile-profile alignment: Our program JProP is a pure Java implementation of the dynamic programming alignment algorithm and can be configured to perform various alignment scoring schemes, average scoring and log average scoring being two of them. For the benchmarks we used gap cost parameters that were optimized using a machine learning approach on a small benchmark set described by Zien et al.[18]. The substitution model used for computing the scores was the BLOSUM62 model[7] and we applied it using an affine gap cost model and global dynamic programming alignment.

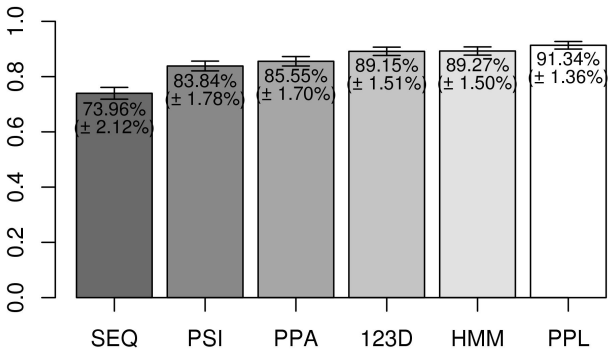

Figure 2: Fold recognition performance for the 1640 chains from the SCOP family level. Numbers indicating percentage of correctly predicted targets from this difficulty level, line segments indicating estimated 95% confidence intervals. SEQ: plain sequence-sequence alignment, PSI: PSI-BLAST, PPA: profile-profile alignment using average scoring, 123D: 123D profile threading, HMM: HMMer, PPL: JProP profile-profile alignment using log average scoring

3 Fold Recognition Results

3.1 SCOP Family Level

We performed the modified leave-one-out fold recognition benchmarks and analysed the results separately for the difficulty of the targets as described above. Figure 2 shows the results for the 1640 chains from the SCOP family level. The 95% confidence intervals indicated in the plot were estimated by using a normal approximation. It is remarkable that even on this level of close relationship (SCOP definition is "clear evolutionary relationship" with a general level of pairwise sequence identity greater than 30%) the sequence alignment is clearly outperformed by all other methods. PSI-BLAST is very good at detecting these close relationships but is already outperformed by the simple profile-profile average scoring approach and clearly left behind by the threading program 123D. As expected, the HMM is very good at detecting and precisely evaluating these close relationships with 89.27% correctly assigned targets. Thus it is very interesting to see that the newly introduced profile-profile alignment with log average scoring can still add more than 2%, yielding a total of 91.34% of fold recognition performance which is a quite significant lead due to the high performance level and the large sample size.

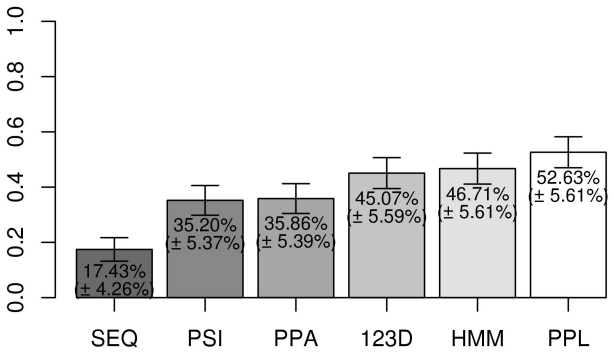

Figure 3: Fold recognition performance for the 304 chains from the SCOP superfamily level. See text or caption of figure 2 for details.

3.2 SCOP Superfamily Level

Since profile-profile alignment was originally designed to detect remote homology relationships we expect to see the largest performance gain on the SCOP superfamily level shown in Figure 3. The plain sequence alignment is clearly the worst when compared to the more sophisticated methods, getting less than half the performance of the next candidate PSI-BLAST. While PSI-BLAST is on par with the simple profile-profile approach, the performance gap to the threading program is already widening. The HMM is better than the threading approach on this level of relationship, predicting 46.71% of the targets correctly. The log average scoring profile-profile alignment clearly shows its strength in detecting weak sequence homology relationships here by outperforming the HMM approach by almost 6% getting a total of 52.63% correct recognition results.

3.3 SCOP Fold Level

On the fold level the relationships between the proteins to be recognized are fairly weak. Since the relations are weaker than the SCOP superfamily level it is not likely that the most closely related domain from the PDB40D set shares the same evolutionary origin with part of the chain. Only a major structural similarity is present. This is the setting for which threading approaches are designed, since they make use of tertiary and secondary structure information instead of relying on the sequence information alone.

Figure 4 shows the results for the 121 chains from this category. A slightly

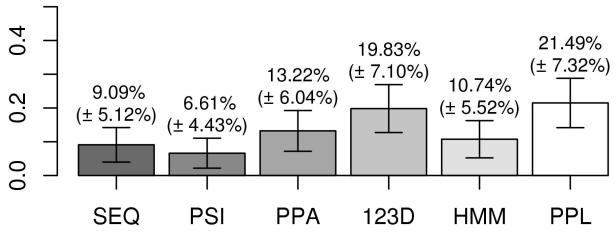

Figure 4: Fold recognition performance for the 121 chains from the SCOP fold level. See text or caption of figure 2 for details.

different picture shows up here. The worst performer is PSI-BLAST with only 6.6% followed by sequence alignment and HMMs with 9.09% and 10.74%, respectively. It should be noted that the results of the PSI-BLAST and the HMMer program on this level are probably hampered by the fact that these programs are the only ones to use significance cutoffs. Thus, sometimes no prediction at all is produced by these two programs, lowering their chance of producing "random" hits. The 123D profile threading programs performs competitively on this level, but again, even the threading approach on these hard targets at 19.81% is outperformed by the profile-profile alignment with the log average score leading with 21.49%. The confidence intervals indicate that these differences are not very significant due to the small sample size, but it is still intriguing to see the completely sequence homology based profile-profile alignment outperform the threading program which makes use of additional structural information. A closer look at the composition of the $\approx 20\%$ shares for 123D and log average profile-profile alignment revealed here that only about 10% of the recognized targets for these two candidates were identical. This stresses the usefulness of trying different algorithms when predicting folds in this very hard category. It also leaves room for speculations on an algorithm combining the strengths of these candidates being possibly capable of reaching 30% fold recognition performance on this SCOP level.

3.4 Overall Fold Recognition

Figure 5 shows a weighted average of the previous results combined with the 167 chains from the SCOP class level that cannot be correctly predicted by homology search. The theoretical maximum performance that can be reached in this plot is thus 92.52%. The results obey the pattern from the previous results. The threading approach and the HMM both at about 72.5% outper-

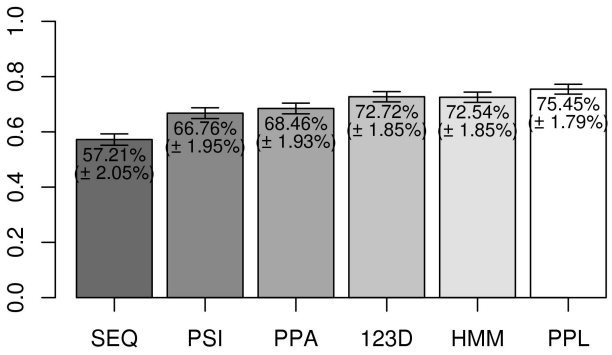

Figure 5: Fold recognition performance for all 2232 chains. See text or caption of figure 2 for details.

form PSI-BLAST and the average scoring profile-profile approach leaving the plain sequence alignment well behind. The log average scoring profile-profile alignment manages to increase the recognition rate by another 3%.

4 Discussion

The profile-profile alignment approach to fold recognition is basically a method for detecting very remote sequence similarity relationships. Sequence information that is not conserved in the most closely related sequences is thrown away by using the frequency profiles constructed by PSI-BLAST instead of the sequences themselves. Thus, a frequency profile is representing a part of the sequence space around the sequence rather than a single point in sequence space. In principle, this should allow for an improved detection of remote sequence homologies.

Nevertheless it is crucial to use a meaningful and sensitive approach to calculate the alignment score in order to receive best results. The effect of this can clearly be seen in the differing results between the mediocre performance of the average score and the superior performance of the log average score.

Our results show, that even simple profile-profile approaches like the average scoring perform competitively to PSI-BLAST. HMM and threading methods are already capable of outperforming PSI-BLAST in terms of fold recognition performance. Choosing the log average scoring, our profile-profile alignment tool outperforms these more advanced tools. On the superfamily level, which is by design the most suitable application scenario for profile-profile

alignment, the log average profile-profile alignment leads the competition by 6% fold recognition performance. Perhaps even more interesting is the fact that it can outperform the applications HMMer, fine tuned on the family level, and 123D threading, fine tuned on the fold level, as well. Thus profile-profile alignment proves to be a useful tool for judging the similarity of two proteins by the alignment score for a broad range of similarity relations from very close to very remote.'

5 Further developments

We are currently working on an extension of the profile-profile alignment score to incorporate a secondary structure component into the scoring system. Furthermore, it will be interesting to see whether the promising results of the alignment score when used for fold recognition also translate into a gain of alignment quality and reliability.

6 Acknowledgements

Part of this work was supported by the DFG priority programme "Informatikmethoden zur Analyse und Interpretation großer genomischer Datenmengen", grant ZI616/1-1. The authors thank Alexander Zien for the help with using his optimization tools and Daniel Hanisch for the collaboration on the program library underlying the JProP implementation.

1. Michael Gribskov and Stella Veretnik. Identification of sequence patterns with profile analysis. In *Methods in Enzymology*, volume 266, chapter 13, pages 198–212. Academic Press, Inc., 1996.
2. Stephen F. Altschul, Thomas L. Madden, Alejandro A. Schäffer, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 25(17):3389–3402, 1997.
3. Leszek Rychlewski, Lukasz Jaroszewski, Weizhong Li, and Adam Godzik. Comparison of sequence profiles. Strategies for structural predictions using sequence information. *Protein Science*, 9:232–241, 2000.
4. Golan Yona and Michael Levitt. Within the twilight zone: A sensitive profile-profile comparison tool based on information theory. *J. Mol. Biol.*, 315:1257–1275, 2002.
5. Niklas von Öhsen and Ralf Zimmer. Improving profile–profile alignment via log average scoring. In Olivier Gascuel and Bernard M. E. Moret, editors, *Algorithms in Bioinformatics, First International Workshop, WABI 2001, Aarhus, Denmark, August 2001, Proceedings*, volume 2149 of *Lec-*

ture Notes in Computer Science, pages 11–26. Springer-Verlag Berlin Heidelberg New York, 2001.

6. Margaret O. Dayhoff, R.M. Schwartz, and B.C. Orcutt. A model of evolutionary change in proteins. In *Atlas of Protein Sequence and Structure*, volume 5, chapter 22, pages 345–352. National Biochemical Research Foundation, Washington DC, 1978.
7. Steven Henikoff and Jorja G. Henikoff. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. USA*, 89(22):10915–9, 1992.
8. L. Lo Conte, B. Ailey, T. J. Hubbard, S. E. Brenner, A. G. Murzin, and C. Chothia. SCOP: a structural classification of proteins database. *Nucleic Acids Res*, 28(1):257–9., 2000.
9. S. E. Brenner, P. Koehl, and M. Levitt. The ASTRAL compendium for protein structure and sequence analysis. *Nucleic Acids Res*, 28(1):254–6, 2000.
10. Ingolf Sommer, Alexander Zien, Niklas von Öhsen, Ralf Zimmer, and Thomas Lengauer. Confidence measures for protein fold recognition. *Bioinformatics*, 18(6):802–812, 2002.
11. Yvonne Kallberg and Bengt Persson. KIND – a non-redundant protein database. *Bioinformatics*, 15(3):260–261, March 1999.
12. Steven Henikoff and Jorja G. Henikoff. Position–based sequence weights. *J. Mol. Biol.*, 243(4):574–578, 1994. 4. November.
13. Anders Krogh and Graeme Mitchison. Maximum entropy weighting of aligned sequences of protein or DNA. In C. Rawlings, D. Clark, R. Altman, L. Hunter, T. Lengauer, and S. Wodak, editors, *Proceedings of ISMB 95*, pages 215–221. AAAI Press, 1995.
14. S. R. Eddy. Profile hidden markov models. *Bioinformatics*, 14:755–763, 1998.
15. S. R. Eddy. Hmmer: Profile hidden markov models for biological sequence analysis. (http://hmmer.wustl.edu/), 2001.
16. Erik Lindahl and Arne Elofsson. Identification of Related Proteins on Family, Superfamily and Fold Level. *J. Mol. Biol.*, 295(3):613–625, January 2000.
17. Nick Alexandrov, Ruth Nussinov, and Ralf Zimmer. Fast protein fold recognition via sequence to structure alignment and contact capacity potentials. In Lawrence Hunter and Teri E. Klein, editors, *Pacific Symposium on Biocomputing'96*, pages 53–72. World Scientific Publishing Co., 1996.
18. Alexander Zien, Ralf Zimmer, and Thomas Lengauer. A simple iterative approach to parameter optimization. *Journal of Computational Biology*, 7(3):483–501, 2000.