Combining Secondary Structure Element Alignment and Profile-Profile Alignment for Fold Recognition

Jan E. Gewehr[1], Niklas von Öhsen[2] and Ralf Zimmer[1]
[1]Research Unit for Practical Informatics and Bioinformatics, Institute for Informatics, Ludwig-Maximilians-University Munich, Amalienstr. 17, D-80333 Munich, Germany,
[2] Institute for Algorithms and Scientific Computing (SCAI), Fraunhofer Gesellschaft, Schloss Birlinghofen, 53754 Sankt Augustin, Germany,
EMail: jan.gewehr@bio.ifi.lmu.de

Abstract: One of the most intensely studied problems of bioinformatics is the prediction of a protein structure from an amino acid sequence. In fold recognition, one reduces this problem to assigning a protein of unknown structure to one of the known fold classes as defined in the SCOP or CATH classifications. Here, we combine two alignment methods, secondary structure element alignment and log average profile-profile alignment that have been proven to perform well on this task. Our results show that the combination yields remarkably better fold recognition accuracy on well-known benchmark sets obtained from the literature. Especially on a difficult set built by McGuffin and Jones this new approach significantly outperforms other recently proposed fold recognition methods.

1 Introduction

One of the most intensely studied problems of bioinformatics is the prediction of a protein structure from an amino acid sequence. While there are nearly three millions of known protein sequences in the NR database provided by the National Center for Biotechnology Information (NCBI), the protein data bank (PDB) [BWF$^+$00] of known structures contains only about 25000 structures. One example of the effort spent on resolving more protein structures experimentally is the Structural Genomics Inititiative of the National Institute of Health (NIH), a "worldwide initiative aimed at determining a large number of protein structures in a high throughput mode"[1]. Progress in structure prediction is therefore still a vital part of current bioinformatics research.

The structure prediction task is often divided into different categories depending on the level of homology of the target sequence (the amino acid sequence for which the structure shall be found) and the sequences of the known structures. Under the term *fold recognition* one usually subsumes all prediction problems where the target sequence has no close homologues among the sequences with resolved structure but its structure still falls into

[1]quoted from http://www.rcsb.org/pdb/strucgen.html

one of the known fold classes as given e.g. by databases like SCOP [MBHC95] and CATH [OMJ+97]. Here, the major task is assigning a protein of unknown structure but known sequence (the so-called *target*) to one of the known fold classes, e.g. by comparing it to a set of representatives for these classes (so-called *templates*).

Recent approaches for tackling the fold recognition problem follow two major directions, namely machine learning methods and alignment methods. Representatives of the first direction are e.g. the methods by Ding and Dubchak [DD01] (neural networks and support vector machines (SVMs)) and Chinnasamy et al. [CSM04] (tree-augmented naïve Bayesian classifiers). Well-known alignment oriented methods are e.g. GenTHREADER [Jo99a] (sequence-profile alignment, evaluation by energy potentials) and MANIFOLD [BCH+03] (sequence and secondary structure alignments combined with enzyme codes).

The work presented here belongs to the class of alignment oriented methods. We combine two alignment approaches that have been proven to work well for detecting remote homologues, log average profile-profile alignment [vÖZ01, vÖSZ03], and secondary structure element alignment [PAR99, MBJ01, MJ02]. Profile-profile alignment is an essential part of the Arby structure prediction server [vÖSZL04] that was ranked among the best independent fold recognition servers in the fold recognition category of the CAFASP 3 experiment [FRD+03]. Recent work by Bindewald et al. [BCH+03] shows that, for fold recognition benchmark sets with low sequence homology, secondary structure element alignment significantly outperforms sequence-based alignment methods.

For the evaluation we make use of two well-known benchmark sets published by McGuffin and Jones [MJ02] and Ding and Dubchak [DD01]. A third set we compiled from the 25% sequence identity subset of the ASTRAL compendium [CHW+04] based on SCOP [MBHC95] version 1.65.

2 Methods

2.1 Log Average Profile-Profile Alignment

The profile-profile alignment approach has recently become popular as it has proven to provide superior alignment quality as well as high fold recognition performance [YL02, RJLG00]. Several scoring functions for profile-profile alignment have been proposed by various authors. The log average scoring function was developed as an extension of the amino acid similarity score for sequences [vÖZ01]. Let α and β represent vectors of amino acid frequencies at two aligned positions, then the log average score of these two profile positions is calculated by the formula

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

where and $p_{\text{rel}}(i,j)$ and p_i originate from the used amino acid similarity model. The total alignment score is then calculated by summation over all aligned positions and applica-

tion of an affine gap cost model and local alignment mode. The similarity model used in the setup is the BLOSUM 62 model. The frequency profiles were generated using five iterations of PSI-BLAST [AMS97] against the NR database with a threshold of 0.001. Profile-profile alignment using this scoring function has been shown to compare favorably against other profile-profile approaches and against several alignment based fold recognition methods [vÖZ01, vÖSZ03].

2.2 Secondary Structure Element Alignment

Secondary structure element alignment (SSEA) was introduced by Przytycka et al. in 1999 [PAR99]. McGuffin et al. [MBJ01] then adopted the idea and compared a slightly modified SSEA to other alignment methods. The result was that SSEA performed best of all tested sequence and secondary structure alignment methods. McGuffin and Jones' alignment procedure reproduced here is astonishingly simple: Given two secondary structure sequences in three-letter code (C: coil, H: helix, E: strand),

1. Represent each sequence as the sequence of secondary structure elements and annotate the length of the elements. Discard leading and trailing coils. The following example illustrates this step: CCCHHHHHCCCCCEEEECC becomes (HCE,554).

2. Align the two element sequences using dynamic programming with zero gap costs. The scoring function is defined as follows: H-H, E-E and C-C are scored with the minimum length of the two aligned elements, H-C and E-C are scored with half the minimum length, and H-E scores 0 (H denotes helix, C denotes coil and E denotes strand). The total score is the sum of all aligned element pair scores.

3. Normalize the score by normscore $= 2\ \frac{\text{rawscore}}{l1+l2}$, where $l1$ and $l2$ denote the length of the first and the second input sequence, respectively.

2.3 Preselection and Refinement

In order to combine the strengths of both methods, we follow a two-stage approach to fold recognition. In the first stage, we align a target sequence to all template sequences using SSEA. Based on the top scores for each template fold class, we select the highest-scoring five percent of all fold classes. For small numbers of classes we use a minimum number of 5 such template classes. We then compute the local profile-profile alignments for the target sequence against all members of these fold classes within the template set.

As target fold class we predict the fold class of the highest scoring template of all template fold classes with respect to profile-profile alignment raw score. As gap costs for local profile-profile alignments we use a gap opening penalty of 9.3540 and a gap extension penalty of 1.369. These values were optimized for individual performance on an independent dataset by NvO.

The reason for this approach is that SSEA often finds the correct fold class within the top scoring classes, with only marginal score differences between them. Basically, SSEA compares secondary structure topologies and therefore does not provide as much specificity as sequence based methods. Profile-profile alignments against all templates within these candidate classes as the next step yield a more precise measure of similarity and therefore allow to detect potential templates with higher specificity.

3 Results

3.1 Datasets and Evaluation Procedure

For the evaluation we use three different benchmark sets, one well-known "difficult" set, one newly compiled "intermediate" set, and one well-known, "easy" set.

The first set was introduced by McGuffin and Jones [MJ02]. It contains 542 nonredundant domains based on CATH [OMJ+97] version 1.7 and is divided into a subset of 252 "known" folds which have at least one other match in the set, and 290 "unique" folds, i. e. domains which have folds unique with respect to this set. In order to compare our method to the results of Bindewald et al. [BCH+03], we use their approach by selecting the set of known folds as targets and the complete set as templates, excluding identical hits.

The second set was compiled from the ASTRAL subset with less than 25% sequence identity based on SCOP version 1.65[2]. We performed leave-one-out tests on all fold classes containing at least 2 members (3999 domains in 441 fold classes).

The third set is the test set provided by Ding and Dubchak [DD01]. It contains 386 SCOP domains in 27 SCOP folds. This set is known to contain distant homologues [BCH+03], a fact that leads to higher recognition rate for such target-template pairs. Ding and Dubchak also provide a training set of SCOP domains which is used e.g. by the MANIFOLD method to train their neural net. Our method is parameter free and therefore does not need this set. We again follow the MANIFOLD procedure by performing leave-one-out tests on the test set only (Silvio C. E. Tosatto, personal communication).

3.2 Quoted Methods

For the sets obtained from the literature, we are able to compare our results directly to the accuracy values reported for other methods:

MANIFOLD (MF). The MANIFOLD method introduced by Bindewald et al. in 2003 [BCH+03] is the most interesting comparison, since it also makes use of secondary structure element alignment. The results are combined with PDB-BLAST and enzyme code similarity by training a two-layer neural net for weighing the three contributions.

[2] provided by http://astral.berkeley.edu

PDB-BLAST (PB). From Bindewald et al. [BCH+03] we quote their results for the PDB-BLAST method as introduced by Rychlewski et al. [RJLG00] which generates PSI-BLAST [AMS97] profiles for each target and aligns them to all template sequences using BLAST [AGM+90].

GenTHREADER (GT). From McGuffin and Jones 2002 [MJ02] we quote the results for GenTHREADER, a simple threading approach introduced by Jones [Jo99a] which uses a sequence profile-based algorithm and subsequently analyzes the alignments by using energy potentials.

BAYESPROT (BP). BAYESPROT utilizes tree-augmented naïve Bayesian classifiers. Here we quote the results from the original paper by Chinnasamy et al. of 2004 [CSM04].

Ding and Dubchak (DD). Ding and Dubchak studied support vector machines and neural nets for fold recognition. The results are quoted from the original paper of 2001 [DD01].

Since these results were not recomputed, it should be noted that there are a couple of differences to the quoted methods. We use PSIPRED [Jo99b] predictions while, for the Ding and Dubchak set, MANIFOLD makes use of consensus secondary structure predictions which were shown to be more accurate by Albrecht et al. [ATLV03]. Furthermore, we made use of an NR version of November 2003 to compute our profiles, so that these also will differ slightly from the profiles generated by Bindewald et al. for MANIFOLD. The final revision of their paper was in August 2003.

3.3 Fold Recognition Accuracy

The fold recognition accuracy on the two benchmark sets is shown in Figure 1. All values were rounded to full percentages. For the MANIFOLD results which were generated by averaging over several test runs only the mean value was used. The difficulty level of the benchmark sets decreases from left to right as indicated by the accuracy of the different methods for each set.

McGuffin and Jones. For the most difficult set we find that sequence based methods perform poorly (PDB-BLAST: 13%, GenThreader: 14%, profile-profile alignment: 16%). The main contribution comes from secondary structure element alignment with 30% accuracy. However, the combination of both sequence and secondary structure again increases accuracy to 34% for MANIFOLD, the best published result currently known to the authors, and 42% for our approach, which is an increase of about 40% over the single contributions of profile-profile alignment and secondary structure element alignment.

ASTRAL 25. For this set we have an orthogonal situation with respect to individual contributions when compared to the McGuffin and Jones dataset. With only 51% accuracy, secondary structure element alignment achieves significantly less hits than local profile-profile alignment with 70%. The combination of both yields 71%, this time only increasing accuracy by 1%.

Ding and Dubchak. On the easier benchmark set containing distant homologues we find that our approach again scores best in comparison to the MANIFOLD method with 80%

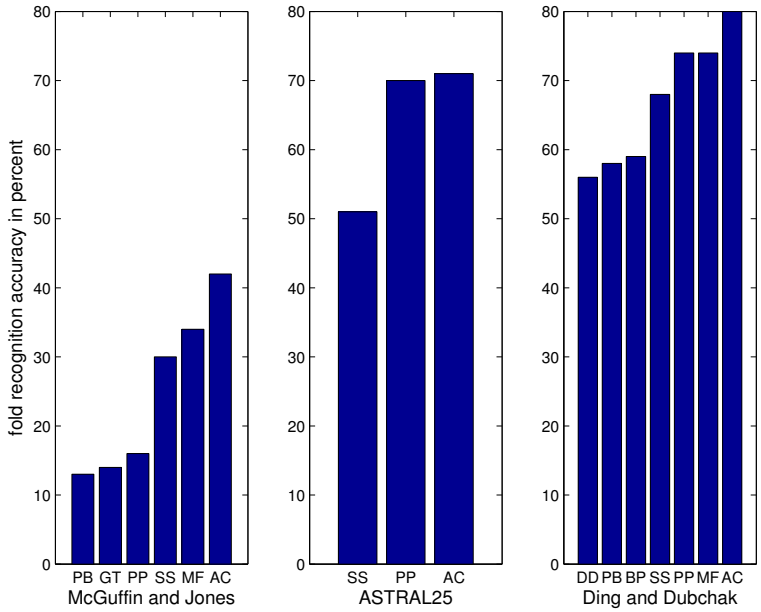

Figure 1: Fold recognition accuracy on three benchmark sets. Method labels are AC (alignment combination of PPA and SSEA), BP (BAYESPROT), DD (Ding and Dubchak), GT (GenThreader), MF (MANIFOLD), PB (PDB-BLAST), PP (profile-profile alignment), and SS (secondary structure element alignment). The values for PB and MF were obtained from Bindewald et al. [BCH+03], the value for BP from Chinnasamy et al. [CSM04], the value for DD from Ding and Dubchak [DD01], and the value of GT from McGuffin et al. [MJ02]. For MF only the mean values were shown.

accuracy compared to 74%, achieving 21% more fold recognition accuracy than the recently published BAYESPROT and even 24% more than the machine learning methods proposed by Ding and Dubchak. Local profile-profile alignment results in 74%, secondary element alignment in 68%. Thus, we observe an increase of 8% over the single contributions.

Overall, we find that, for each dataset, our approach is beneficial for fold recognition, increasing accuracy by up to 40% over the single alignment methods. With increasing difficulty level we observe that the performance of sequence based methods decreases when compared against secondary structure element alignment. We observe better fold recognition accuracy no matter whether the better individual performance comes from secondary structure element alignment (McGuffin and Jones) or from profile-profile alignment (Ding and Dubchak, ASTRAL25).

4 Discussion

We have introduced a simple way of combining two powerful alignment methods for fold recognition, local log average profile-profile alignment and secondary structure element alignment. We select potential fold classes according to their secondary structure topology and then rescore these classes using sequence profiles generated by PSI-BLAST. Direct comparison to recently proposed fold recognition methods shows that this approach is competitive with state-of-the-art approaches, although methods like MANIFOLD require parameter optimization procedures like e.g. Monte-Carlo optimization to train their machine learning based classifiers and weighting schemes. On both well-known benchmark sets obtained from the literature this simple procedure outperforms recently published results of other methods on these sets. Especially interesting is the improvement of accuracy for difficult targets, i.e. targets for which we do not find homologues in the template set, as it is the case for the McGuffin and Jones dataset.

We therefore propose to make use of a combination of both secondary structure element alignment and profile-profile-alignment for remote homology detection. More elaborate ways of combining the independent predictions like e.g. confidence measures [SZvÖ+02] or neural networks as in MANIFOLD may further improve the performance.

Acknowledgements. The authors would like to thank Alessandro Macri for helpful discussions. Silvio C. E. Tosatto kindly provided the data for the Ding and Dubchak set that was used in [BCH+03]. JG was funded by the German Research Council (DFG) under project grant PROSEQO II (Zi616/2). NvO was funded by the BMBF project "Development of Microbalance Array/ Mass Spectrometry as a Tool for Functional Proteomics" (0312708).

References

[AGM+90] Altschul, S., Gish, W., Miller, W., Myers, E., and Lipman, D.: Basic local alignment search tool. *J Mol Biol*. 215:403–410. 1990.

[AMS97] Altschul, S., Madden, T., and Schäffer, A. e. a.: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*. 25:3389–3402. 1997.

[ATLV03] Albrecht, M., Tosatto, S., Lengauer, T., and Valle, G.: Simple consensus procedures are effective and sufficient in secondary structure prediction. *Protein Engineering*. 16:459–462. 2003.

[BCH+03] Bindewald, E., Cestaro, A., Hesser, J., Heiler, M., and Tosatto, S.: MANIFOLD: Protein fold recognition based on secondary structure, sequence similarity and enzyme classification. *Protein Engineering*. 16(11):785–789. 2003.

[BWF+00] Berman, H., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T., Weissig, H., Shindyalov, I., and Bourne, P.: The protein data bank. *Nucleic Acids Research*. 28:235–242. 2000.

[CHW+04] Chandonia, J., Hon, G., Walker, N., Lo Conte, L., Koehl, P., Levitt, M., and SE, B.: The ASTRAL compendium in 2004. *Nucleic Acids Research*. 32:D189–D192. 2004.

[CSM04] Chinnasamy, A., Sung, W., and Mittal, A.: Protein structure and fold prediction using tree-augmented naïve Bayesian classifiers. In: Altman, R., Keith, A., Hunter, L., Jung, T., and Klein, T. (Eds.), *Pacific Symposium on Biocomputing 2003*. volume 9. pp. 387–398. 2004.

[DD01] Ding, C. and Dubchak, I.: Multi-class protein fold recognition using support vector machines and neural networks. *Bioinformatics*. 17(4):349–358. 2001.

[FRD+03] Fischer, D., Rychlewski, L., Dunbrack, R. L., Ortiz, A. R., and Elofsson, A.: CAFASP3: The third critical assessment of fully automated structure prediction methods. *Proteins: Structure, Function, and Genetics*. 53:503–516. 2003.

[Jo99a] Jones, D.: GenTHREADER: An efficient and reliable protein fold recognition method for genomic sequences. *J Mol Biol*. 287:797–815. 1999.

[Jo99b] Jones, D.: Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol*. 292:195–202. 1999.

[MBHC95] Murzin, A., Brenner, S., Hubbard, T., and Chothia, C.: SCOP: A structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol*. 247:536–540. 1995.

[MBJ01] McGuffin, L., Bryson, K., and Jones, D.: What are the baselines for protein fold recognition? *Bioinformatics*. 17(1):63–72. 2001.

[MJ02] McGuffin, L. and Jones, D.: Targeting novel folds for structural genomics. *PROTEINS: Structure, Function, Genetics*. 48:44–52. 2002.

[OMJ+97] Orengo, C., Michie, A., Jones, S., Jones, D., Swindells, M., and Thornton, J.: CATH - a hierarchical classification of protein domain structures. *Structure*. 5:1093–1108. 1997.

[PAR99] Przytycka, T., Aurora, R., and Rose, G.: A protein taxonomy based on secondary structure. *Nature Structural Biology*. 6:672–682. 1999.

[RJLG00] Rychlewski, L., Jaroszewski, L., Li, W., and Godzik, A.: Comparison of sequence profiles. Strategies for structural predictions using sequence information. *Protein Science*. 9:232–241. 2000.

[SZvÖ+02] Sommer, I., Zien, A., von Öhsen, N., Zimmer, R., and Lengauer, T.: Confidence measures for protein fold recognition. *Bioinformatics*. 18(6):802–818. 2002.

[vÖSZ03] von Öhsen, N., Sommer, I., and Zimmer, R.: Profile-profile alignment: A powerful tool for protein structure prediction. In: Altman, R. B., Dunker, A. K., Hunter, L., Jung, T. A., and Klein, T. E. (Eds.), *Pacific Symposium on Biocomputing 2003*. pp. 252–263. World Scientific Publishing Co. Pte. Ltd., Singapore. 2003.

[vÖSZL04] von Öhsen, N., Sommer, I., Zimmer, R., and Lengauer, T.: Arby: Automatic protein structure prediction using profile-profile alignment and confidence measures. *Bioinformatics*. to appear. 2004.

[vÖZ01] von Öhsen, N. and Zimmer, R.: Improving profile–profile alignment via log average scoring. In: Gascuel, O. and Moret, B. M. E. (Eds.), *Algorithms in Bioinformatics, First International Workshop, WABI 2001, Aarhus, Denmark, August 2001, Proceedings*. volume 2149 of *Lecture Notes in Computer Science*. pp. 11–26. Springer-Verlag Berlin Heidelberg New York. 2001.

[YL02] Yona, G. and Levitt, M.: Within the twilight zone: A sensitive profile-profile comparison tool based on information theory. *J. Mol. Biol.* 315:1257–1275. 2002.