PROTEINS: Structure, Function, and Bioinformatics 54:351–360 (2004)

A Shannon Entropy-Based Filter Detects High-Quality Profile–Profile Alignments in Searches for Remote Homologues

Emidio Capriotti,[1,2,3] Piero Fariselli,[3] Ivan Rossi,[2,3] and Rita Casadio[3]*
[1]*Department of Physics, University of Bologna, Bologna, Italy*
[2]*BioDec srl, Bologna, Italy*
[3]*Department of Biology and CIRB, University of Bologna, Bologna, Italy*

ABSTRACT **Detection of homologous proteins with low-sequence identity to a given target (remote homologues) is routinely performed with alignment algorithms that take advantage of sequence profile. In this article, we investigate the efficacy of different alignment procedures for the task at hand on a set of 185 protein pairs with similar structures but low-sequence similarity. Criteria based on the SCOP label detection and MaxSub scores are adopted to score the results. We investigate the efficacy of alignments based on sequence–sequence, sequence–profile, and profile–profile information. We confirm that with profile–profile alignments the results are better than with other procedures. In addition, we report, and this is novel, that the selection of the results of the profile–profile alignments can be improved by using Shannon entropy, indicating that this parameter is important to recognize good profile–profile alignments among a plethora of meaningless pairs. By this, we enhance the global search accuracy without losing sensitivity and filter out most of the erroneous alignments. We also show that when the entropy filtering is adopted, the quality of the resulting alignments is comparable to that computed for the target and template structures with CE, a structural alignment program. Proteins 2004; 54:351–360.** © 2003 Wiley-Liss, Inc.

Key words: alignment; PSI-BLAST; sequence profile; fold recognition; remote homologues

INTRODUCTION

The ever-increasing size of databases of protein sequences has promoted the development of new approaches in the field of fold recognition. It is generally accepted that in proteins with high-sequence identity (>25%), structures are similar.[1–3] Many examples in the literature describe the success of this approach.[4–7] However, the problem gets increasingly difficult when homology between target and template sequences becomes low and sequence identity is <25%; under these conditions, alignments become unreliable.[8,9] Therefore, the search into the so-called "twilight zone" of sequence similarity (<25%) requires the development of methods suited to find new protein structures (remote homologues[1]).

In the last few years, new algorithms incorporated evolutionary information through multiple-sequence alignments. Methods, such as PSI-BLAST,[4] 3D-PSSM,[5] GenTHREADER,[6] and BASIC[7] have improved the search for remote homologous sequences.

Generally speaking, starting from the sequence, the search can be performed by comparing two sequences (sequence–sequence), a sequence to a profile (sequence–profile), and two profiles (profile–profile). It has been reported that methods based on profile–profile alignments are in general more sensitive than those based on sequence–profile, such as PSI-BLAST.[10]

In this work, we describe a method similar to a previously developed procedure (BASIC)[7] for the profile–profile comparison, and we apply it to the fold recognition problem. Then, we contrast our results with those obtained with sequence–sequence, sequence–profile, and structural alignments. We prove that a filter based on Shannon entropy[11] is capable of selecting good alignments among meaningless ones generated with profile–profile comparisons and to enhance the global accuracy of fold recognition. Moreover, when the Shannon entropy filter operates, the resulting alignments are comparable to those obtained with the CE[12] structural alignment program that operates starting from the target and template structures.

MATERIALS AND METHODS

Data Set

Protein structures are clustered into families based on hierarchy of functional and structural similarities. Databases, such as SCOP[13] and CATH[14,15] (http://www.biochem.ucl.ac.uk/bsm/cath_new), are built by human expertise and agreed-upon criteria to cluster proteins of similar structure and function. Because we are interested in fold recognition of proteins with low-sequence similarity, we select a representative set containing typical folds.

Grant sponsor: Ministero della Università e della Ricerca Scientifica e Tecnologica (MURST); Grant sponsor: SPINNER Consortium (Regiona Emilia-Romagna).

*Correspondence to: Rita Casadio, Department of Biology and CIRB, University of Bologna, Via Irnerio, 40126 Bologna, Italy. E-mail: casadio@alma.unibo.it

Received 19 February 2003; Accepted 9 June 2003

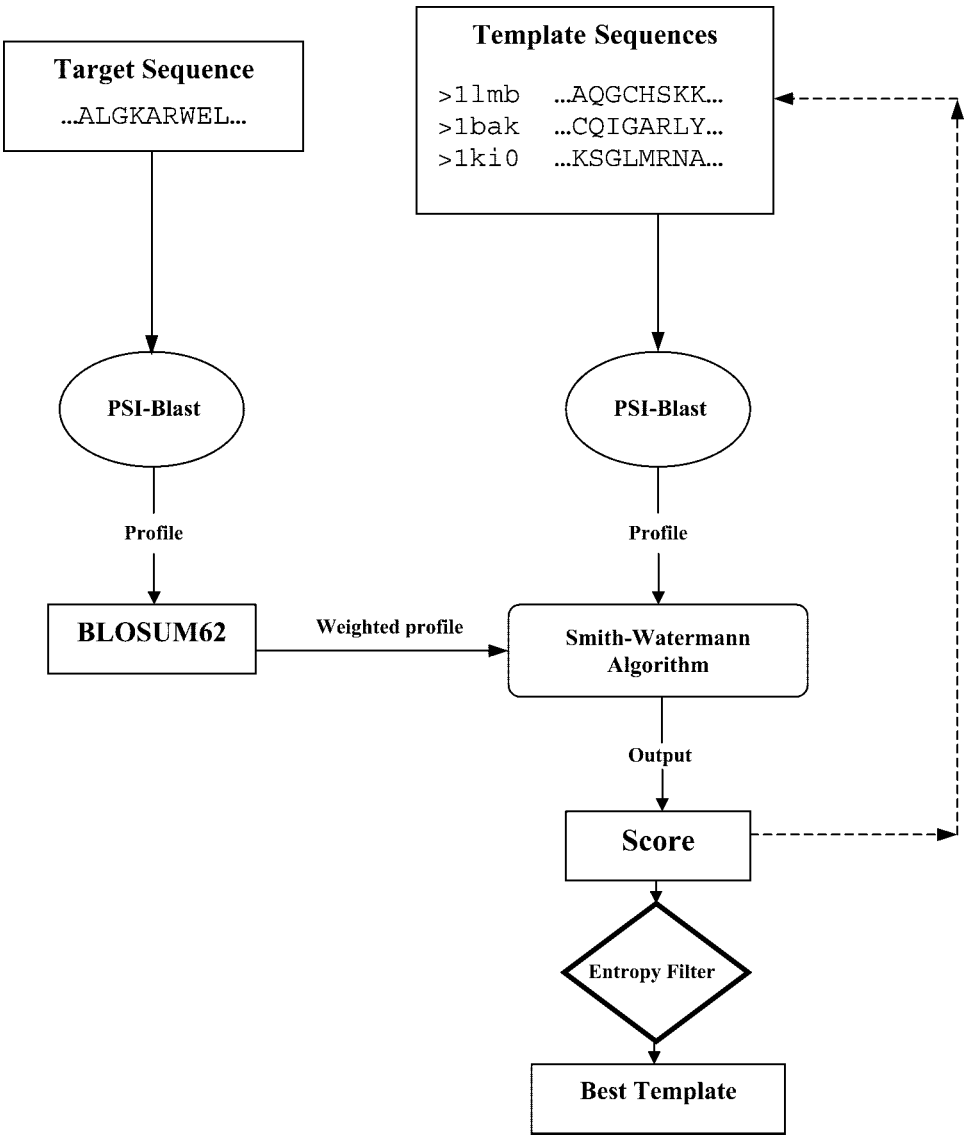

Fig. 1. Flow chart of the method. Here we describe all steps by using the SWA-PP method. In the first step, we build target and template profiles by PSI-Blast. In the second phase, we align the profiles by using BLOSUM62 mutation matrix and Smith and Watermann algorithm. We repeat this operation for all templates in our database. In the final step, we assess our results by the Shannon entropy criterion.

The procedure is as follows. First, we use the representative list of protein families downloaded from http://www.ebi.ac.uk/dali/fssp/TABLE1.html. This list contains chains with sequence identity $< 25\%$ and known structures in the PDB. Second, we discharge the chains whose entries in the corresponding PDB files (http://www.rcsb.org/pdb) are not completed (we require continuity in the structures). Third, we keep only those chains that had a corresponding SCOP code. Finally, we chose a protein pair in the previous list for each different SCOP label (4 digits) (http://scop.mrc-lmb.cam.ac.uk/scop/). Our final list consists of 185 protein pairs with sequence identity $< 25\%$ and known structure, with SCOP labels, and is available at (http://www.biocomp.unibo.it/emidio/thlist.html). This set is then used to score our approach.

Evaluation of Fold Recognition

For proteins of known structures, we can evaluate the sequence alignments obtained by a threading method. Indeed, we can measure how well the putative protein structure fits into its high-resolution three-dimensional structure. This task can be addressed by using several methods, such as comparing the identical pairs in sequence and structure alignments,[8] measuring the contact map overlap after optimal superposition,[16,17] or calculating the structural similarity between targets and templates.[18–20]

In this article, we measure the overlapping of the putative structure evaluated after finding of the template with the known sequence structure in the SCOP database.

TABLE I. Assessment of the Different Alignment Procedures on the 185 Protein Pairs of the Data Set

Methods[a]	% CF[c]	% CSF[b]	<MaxSub>[b]
SWA-SS	52.4	54.1	1.7 ± 0.2
SWA-SP	62.7	63.8	2.7 ± 0.2
SWA-PS	64.3	68.1	2.6 ± 0.2
Max (PS,SP)	64.9	67.0	2.8 ± 0.2
SWA-PP	67.6	70.8	3.1 ± 0.2
SWA-PP+SF[b]	97.5	99.2	4.6 ± 0.2

[a]SW, Smith–Waterman algorithm; SWA-SS, SW with sequence–sequence score; SWA-SP, SW with sequence–profile score; SWA-PS, SW with profile–sequence score; Max(PS,SP), maximum between SWA-PS and SWA-SP; SWA-PP, SW with profile–profile score.
[b]Indexes: CF is the number of correct fold predictions at family level; CSF is the number of correct superfamily predictions. MaxSub score is the average value over 185 pair proteins data set.
[c]Shannon entropy filter reduces the number of protein pairs to 119, because the alignments in which at least one profile is endowed with an entropy value of <0.5 are rejected.

This is performed simply by computing the match of the SCOP indexes both at the family (4 digits) and superfamily (3 digits) level and introducing two indexes that evaluate the percentage match at the family (FC) and superfamily (SFC) level, respectively.

A second measure of the efficiency of our procedure is computed with the MaxSub algorithm.[21] The difficulty in assessing models is that simple measures, such as root-mean-square deviation (RMSD),[20] computed over all the atoms is a very poor indicator of the quality of a model, especially when only part of the model is well assigned. In this case, the wrongly predicted regions spoil the RMSD computation. Thus, a way to identify and score only the well-predicted regions is needed. We evaluate our results by using the MaxSub algorithm, which has also been used in CAFASP3 experiments for the same purpose. MaxSub aims to identify the largest subset of C_α atoms of a model that superimpose well over the experimental structure and ranges from 0 to 10.

We also computed an upper bound to the alignment accuracy using a structural alignment algorithm. There are many methods to evaluate similarity of protein structures, such as CE,[12], DALI[22,23] (http://www.ebi.ac.uk/dali/), and VAST[24] (http://www.ncbi.nlm.nih.gov/Structure/VAST/vast.shtml).

We used the CE method as available on the web (http://cl.sdsc.edu/ce.html) with the default parameters.

The Algorithm

Our algorithm is based on a local dynamic programming,[25] with BLOSUM62 as scoring matrix.[26] In the dynamic programming calculation, we optimize the gap penalties values, implemented as the usual linear gap ($g(x) = kx + q$). We implemented several alignment procedures, all based on the local Smith–Waterman algorithm (SWA).[25] Different scoring functions are computed, considering, respectively, information derived by sequence–sequence (SS), sequence–profile (SP), profile–sequence (PS), or profile–profile (PP) alignments. Results are also compared with those of the PDB-BLAST[10] protocol.

Fig. 2. Example of a good SWA-PP alignment. Here we show the alignment between 1GAKA model (gray) and 1 GAKA protein (black). The model is obtained by using 1LIS_ template. The two proteins have a.19.1.1 SCOP code. The comparison between the two structures with MaxSub program give 4.94 points. The identity in the sequence alignment is 19%. The 1GAKA and 1LIS_ protein pair is detected by using the SWA-PP method, which gives a score of 39.7. Other methods, such as SWA-SS, SWA-PS, and SWA-SP, are not able to recognize the correct template.

To obtain sequence profiles, we use three iterations of PSI-BLAST with an E-value inclusion threshold of 10^{-3}. Our PSI-BLAST runs were performed against a nonredundant database[27] consisting of 705,002 protein sequences.

The alignment score from the position i of target sequence profile $P_A(i)$ and the position j of template sequence profile $P_B(j)$ is indicated as $D_{AB}(i,j)$ and is calculated as

$$D_{A,B}(i,j) = P_A(i)^T M P_B(j) \quad (1)$$

where M is the 20×20 BLOSUM62 substitution matrix.

The formalism introduced with Eq. 1 applies smoothly to both profiles and sequence scores, because the sequences can be converted into profiles that have only 0 or 1 element.

When necessary, the values for the gapping parameters used in SWA-PP are $k = 1$ and $q = 3$, obtained after a search in the parameter space, to optimize the results. The value of gapping parameters reported in the literature for sequence–profile alignments, as in PSI-Blast, are $k = 1$ and $q = 10$ and for the sequence–sequence ones, as in LALIGN,[28] are $k = 4$ and $q = 10$.

The flowchart of our profile–profile-based method is shown in Figure 1.

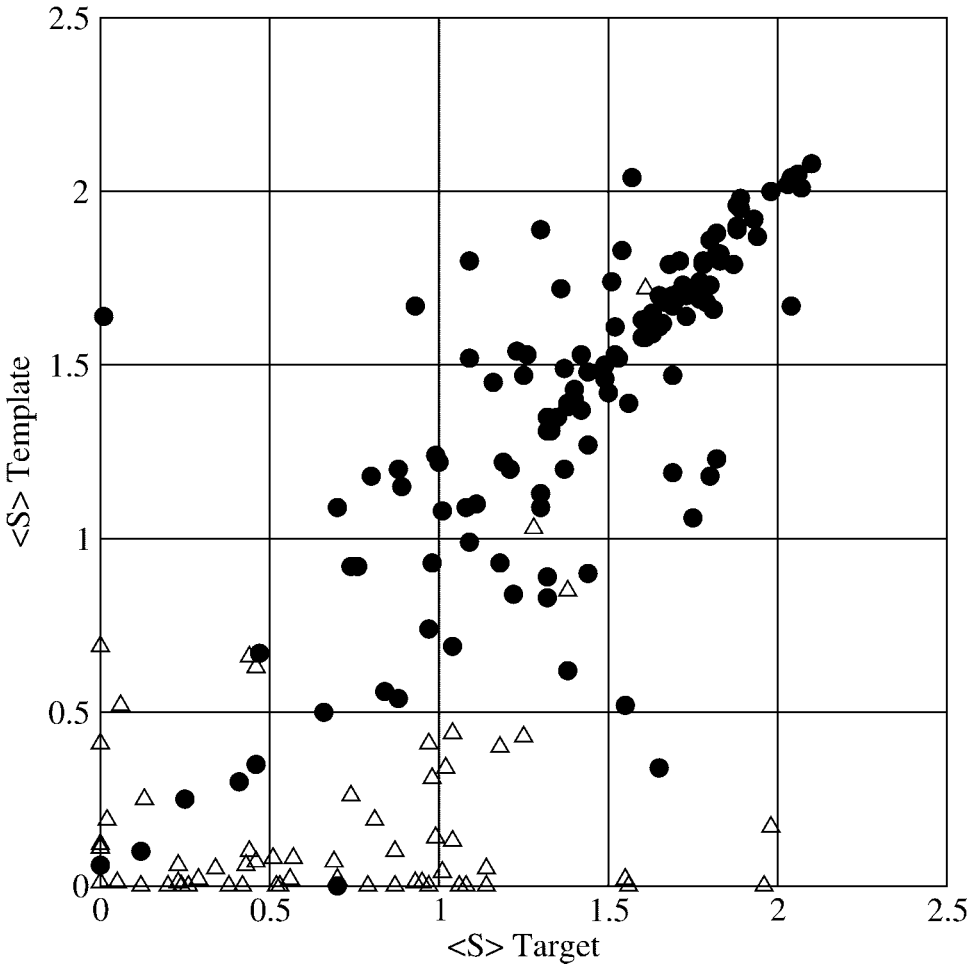

Fig. 3. Shannon entropy information and fold prediction. Template versus target average Shannon entropy. Filled circles represent the correct SCOP fold labels, whereas white triangles are erroneous assignments. However, two of the three wrong assingments whose entropy is higher than 0.5 for both target and template are correct at the SCOP superfamily level.

Statistical Indices

Among the standard statistical indices to evaluate an alignment accuracy, we select the following two. The first one is the z score, which is computed as

$$Z \text{ score} = \frac{D - \langle D \rangle}{\sigma_D} \tag{2}$$

where D is the score obtained from the alignment of two proteins, $<D>$ is the mean scoring point over all possible alignment in our data set, and σ_D is the standard deviation.

The second is the Shannon entropy information,[11] which for each position i in the sequence profile it is defined as

$$S_i = -\sum_{a=1}^{20} p_i(a)\log(p_i(a)) \tag{3}$$

where the sum runs over the 20 amino acids and $p_i(a)$ is the frequency that the residue a in the i-th position.

Entropy-Based Filter

The filter procedure is applied to score profile versus profile. The method is essentially based on the computation of the average entropy value for each alignment as described by Eq. 3. When this value is below a given threshold (routinely 0.5), the alignment is rejected. Conversely, alignments of both the target and template with the highest entropy values are retained (see Fig. 3 and text below, for explanations). The rationale for this procedure is based on the observation that when the entropy of either one or both the profiles is low (less than the 0.5 threshold), high scoring alignments, with an unrealistically large number of gaps are routinely generated. This is possibly due to the fact that gapping parameters used in SWA-PP become inadequate; particularly, the gap opening cost q becomes too low.

RESULTS AND DISCUSSION

Assessment of the Different Alignment Procedures

Starting from the protein sequence, we can search the database of sequences corresponding to known structures

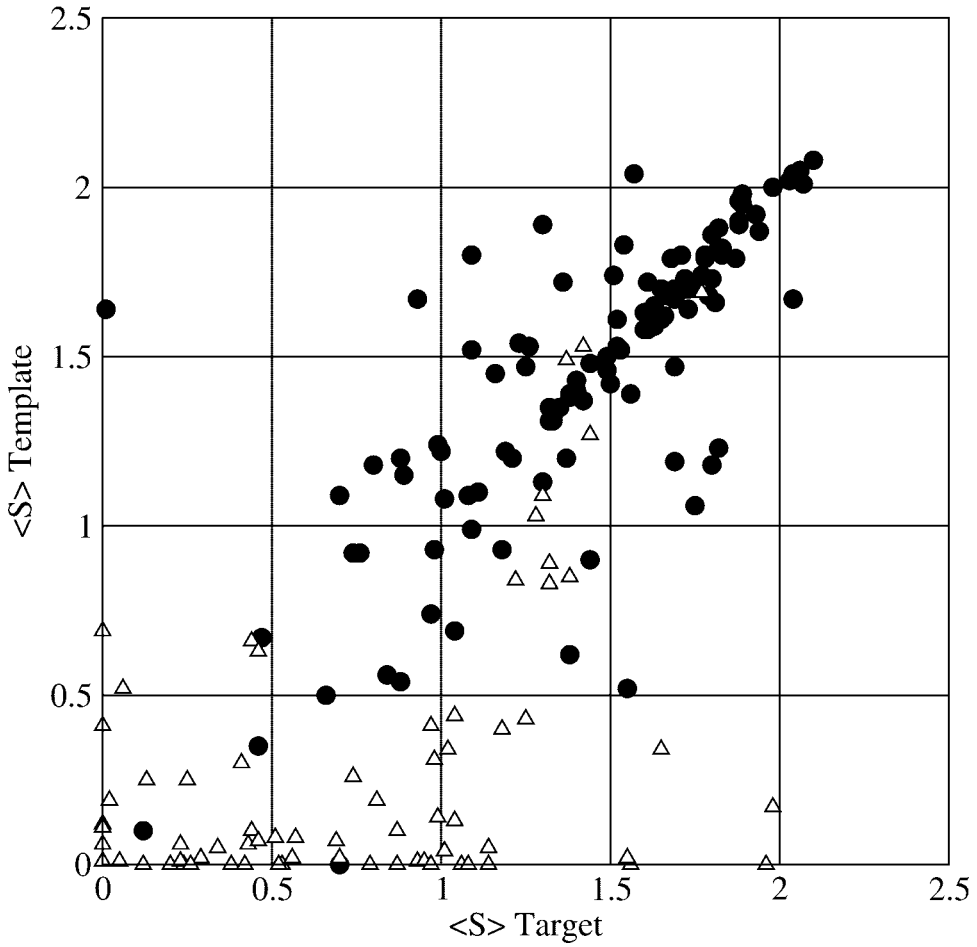

Fig. 4. Alignment entropy in terms of Shannon entropy information. Template versus target average Shannon entropy. Filled circles and white triangles represent alignments on which MaxSub score is positive or 0, respectively.

in different ways. Because we know the solution of our problem (the structure of the target), after searching, we compare the real protein structure with the computed one and evaluate the efficacy of different alignment procedures.

In this section, we compare the results obtained with different types of alignments. Basically, we can adopt at least five different procedures and score them as described by Eq. 1. The scoring functions are evaluated by using the BLOSUM62 matrix and five different alignment procedures: sequence versus sequence (SS), target sequence versus template profile (SP), target profile versus template sequence (PS), the maximum scoring alignment between each corresponding PS and SP, and profile versus profile (PP) (Eq. 1). By this, we can then compare the efficacy of the five different approaches to trace the most reliable template. This is done on the basis of the alignment scores and the number of correct fold predictions evaluated as described above.

It is worth mentioning that we consider only the first best scoring template of each run (no second best is used). This is a somewhat stringent condition, but we believe that it is needed to better highlight the differences in the performance of the different procedures.

Table I reports the results obtained by using our data set of 185 protein pairs. It is evident that when we score fold detection, the profile–profile-based method (SWA-PP of Table I) is the best performing, both at the family (%CF) and superfamily levels (%CSF). Indeed, SWA-PP score some 18% higher than the sequence–sequence-based method (SWA-SS).

A more stringent test, based on the alignment of the target with the template, once the procedure has focused it, is given by the MaxSub value. In this case, the results also clearly indicate that increasing the complexity of the alignment methods promotes a better template identification. Evidently, the results confirm the notion that evolutionary information improves fold assignment and are in agreement with a previous analysis reported by Rychlewski et al.[31] in which a similar algorithm (FFAS-R) was described.

However, when we apply the entropy-based filter, the SWA-PP score further increases, clearly showing that in this case we can focus more effectively on the significant alignments and templates. The efficacy is about 99% and

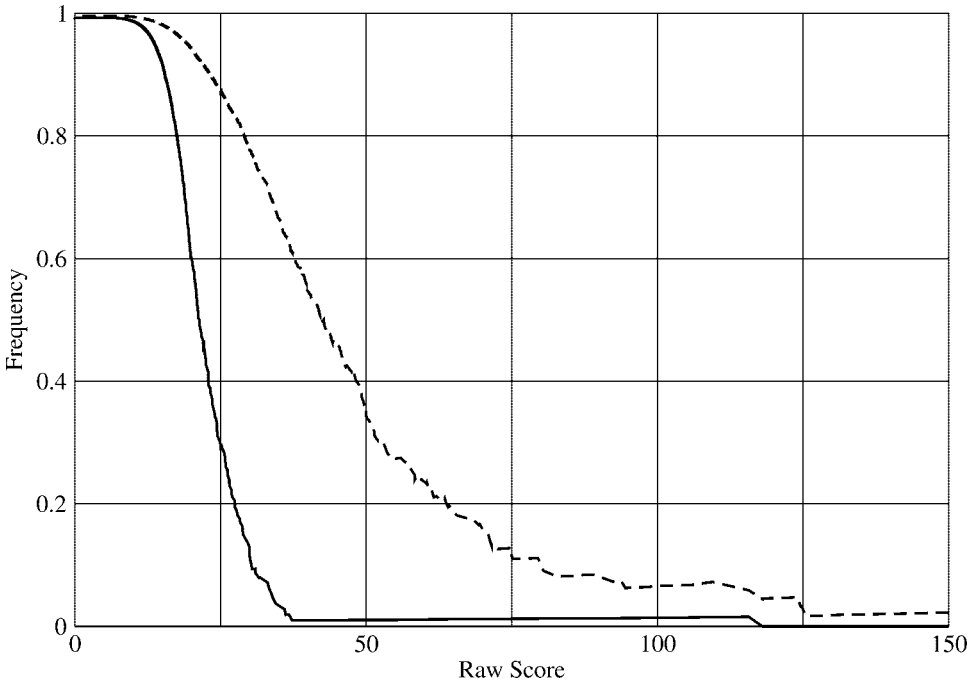

Fig. 5. Effect of entropy filtering on SWA-PP accuracy. The plot shows the probability of getting a positive MaxSub score as a function of the alignment raw score. The dotted line is obtained over the unfiltered data, the continuous line on the filtered ensemble. For filtered data, the frequency of a 0 MaxSub score is close to 1%, when the score is around 36 (z score $\approx$ 2).

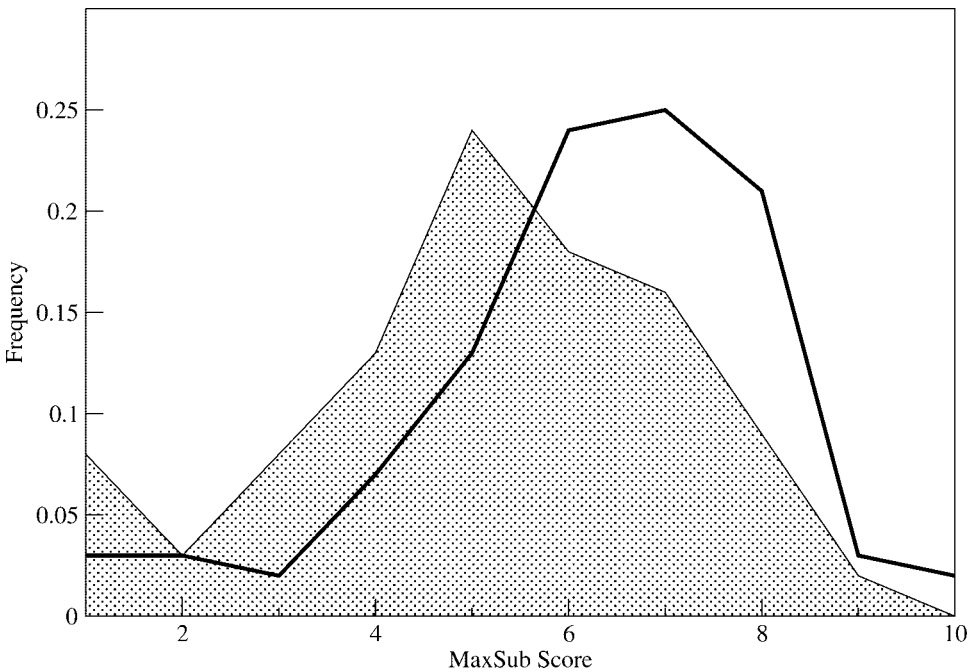

Fig. 6. Comparison between SWA-PP method and CE results. Here we compare the quality of the entropy-filled SWA-PP alignments (pointed area) with those obtained by the structural alignment program CE (bold line). We plotted in point j the number of MaxSub point between j-l and j divided by the total sample of data. In this way, the mean value of MaxSub points obtained by our method is 4.59 and by CE if 5.75.

98% at the superfamily and family level, respectively (some 30% more than without filter). Concomitantly, also the MaxSub value increases by about 1.5 folds.

For sake of clarity, we show in Figure 2 the structural alignment of two proteins detected by the SWA-PP approach. It is worth noting is that this fold would have not

TABLE II. Comparison of the Results Obtained With SWA-PP, PDB-BLAST, and CE on 119 Protein Pairs Data Set

Methods	% NP[a]	<MaxSub>
PDB-BLAST	58.9	2.8 ± 0.2
SWA-PP+SF	91.7	4.6 ± 0.2
CE	92.4	5.0 ± 0.2

[a]NP is the number of predictions that have average MaxSub score > 0.

been detected by using the SWA-SS or SWA-PS methods (protein PDB codes: 1GAKA, 1LIS_). The model obtained with SWA-PP is superimposed to the real protein structure by the MaxSub program, and the final superposition is shown in Figure 2 using Rasmol (http://www.umass.edu/microbio/rasmol/).

Shannon Entropy Information and Alignment Accuracy

The major contribution of this article is the introduction of an index related to the accuracy of the profile–profile alignments. By analyzing the wrongly assigned high scoring pairs, we noticed that routinely either sequences of very high homology or very few sequences were included in the alignment used to build the corresponding profiles. In this case, there is very little difference between the profile and the sequence alone, so there is not much evolutionary information included in the profile.

In this article, we propose that the Shannon entropy is a parameter suited to separate accurate profile–profile alignments from bad ones. To support our claim, in Figure 3 we plot each pair assigned by SWA-PP as a function of the target and template average entropies calculated on the aligned profile segments (filled circles represent the correct SCOP fold labels and white triangles are erroneous assignments). It is evident that for $<S> > 0.5$ for both target and template, >97% (116 of 119) of SWA-PP predictions are in agreement with the SCOP folds. Noticeably, two of the three wrong assignments, whose entropy is higher than 0.5 for both target and template, are at least correct at the SCOP superfamily level, reaching in this case a score of 99% (118 of 119).

To further support this finding, a more stringent test is conducted to evaluate the MaxSub score (Fig. 4). In this case, only the alignment pairs to whom MaxSub assigns points are considered "corrects." In Figure 4, the correct aligned pairs are represented, as before, with filled circles, and the wrong ones are depicted with white triangles. The scores are depicted as a function of the average entropy. It can be noticed that again for $<S> > 0.5$, about 92% (109 of 119) of the alignments has a MaxSub score greater than 0.

The relevance of the entropy criterion is also highlighted in Figure 5. In this case, for each score, the probability that a given alignment receives 0 MaxSub points is plotted as a function of the alignment score. The alignment score is computed from the distribution of the all possible alignment scores using the whole data set (34,225 alignments) with a mean and a standard deviation of 12.5 and 12.1, respectively. Both the whole data set pairs (dotted line) and the filtered (solid line) ones are shown in the plot. In this second case, only those pairs whose $<S> > 0.5$ for the template and target profiles are taken into consideration (16,795 alignments).

By considering the unfiltered data, it is evident that there is still a significant chance of having an incorrect alignment even when the score of a given pair is >100. Conversely, the chance of a wrong pair alignment with a score > 40 is negligible for the set after filtering. In other words, we can say that if we are using the entropy filter, a pair of target-template profiles with a z score > 2 (alignment score > 36.7 and Eq. 2) corresponds very likely (about 99%) to a good candidate for the target query.

Therefore, we can conclude that entropy filtering eliminates bad alignments with a high score. This kind of alignment is obtained when a protein target or a template has no sequence similarity to any protein in the nonredundant database. In this case, it is impossible to exploit evolutionary information.

Fig. 7. Sometimes SWA-PP is better than CE. A rare example where the SWA-PP alignment is better than the CE one. The two protens are 1A0AA and 1AM9A, and their SCOP code is a.38.1.1. The SWA-PP alignment has 28% sequence identity. On the left, we show the SWA-PP alingment (6.05 MaxSub), whereas on the right, the CE structural alignment (4.05 MaxSub) is shown. SWA-PP better aligns the loop zone.

Two fold recognition methods that make use of information theory-based scoring functions have recently been published.[29,30] In these articles, different information theory descriptors are involved in the definition of the scoring funtions. Alternatively, with our approach, Shannon entropy is used to *filter* the result of a traditional scoring (Eq. 1).

Comparison With CE and PDB-BLAST

To show the increased capability of our procedure, the results obtained with the filtered SWA-PP method are compared with those obtained by applying a structural alignment program (CE). In this case, the target and the template are aligned according to CE, and the alignment is afterwards scored with MaxSub. Results are shown in Figure 6. It is interesting that a shift toward higher MaxSub points is noticed in CE. This is expected because CE performs the structural alignment of a known three-dimensional (3D) target to a known 3D template, namely the correct SCOP pairs. This is based on the assumption that the correct pairs are known and that the remote homology search (the location of the correct templates) is perfectly accomplished (100% accurate).

These results can be regarded as an upper limit with respect to MaxSub values obtained after applying filtered SWA-PP. However, the two distributions have a large overlap, with average values of the MaxSub values of the filtered SWA-PP and CE distributions of 4.6 and 5.7, respectively. This is noteworthy, if we consider that different from CE, our procedure builds a model starting from the target sequence.

Our procedure is also contrasted with the PDB-BLAST protocol (see Materials and Methods). PDB-BLAST consists of two stages: 1) the protein target profiles are generated by three rounds of PSI-BLAST algorithm against the nonredundant database, and 2) a second run is performed by PSI-BLAST against the template sequence database.

The results (Table II) clearly indicate that the upper bound accuracy (CE) is superior to the results obtained by using profile–sequence information (PDB-BLAST), as expected considering also our results shown above (Table I). The entropy profile-based method scores close to CE.

However, it is important to remark that CE fails in some cases. More precisely, about 8% of the CE alignments receives zero MaxSub points (Table II).

It is surprising that there are few cases in which the SWA-PP + SF outperforms the CE structural alignment. An example of this finding is reported in Figure 7.

CONCLUSIONS

In this article, we support the idea that the introduction of evolutionary information improves the quality of the fold predictions (Tables I and II). In general, methods, such as BASIC,[7] FFAS,[31] or our SWA-PP, which are different from other faster approaches (e.g., PSI-Blast and GeneThreader), take advantage of evolution information in the form of sequence profiles both for the target and template sequences. Therefore, with SWA-PP, it is possible to harvest for a given query a homologous sequence (template) that has diverged beyond the point where its homology can be recognized by a simple direct comparison. In this case, a third sequence, intermediate between the two, can relate them through their sequence profiles.[32] For instance, there are cases in which protein A can be reliably identified as being homologous to B and B is reliably homologous to C: this enables A and C to be classified as homologous, despite the fact that A and C cannot be directly recognized as related. One example is the one we show in Figure 7, where the two proteins 1gakA and 1lis_ are detected as remote homologues only when SWA-PP and the entropy filter criterion are used.

Unfortunately, it is not rare that an incorrect pair alignment obtains the highest score among the various possible templates. This is often due to the fact that one sequence profile (or both target and template profiles) consists of too few highly similar sequences. In this case, Eq. 1 can give a high score value, despite the fact that the corresponding alignment may have a large number of discontinuous gaps.

To overcome this inconvenience, we propose that the alignments characterized by an entropy value lower than a given threshold (<0.5) be discarded. This is equivalent to discharge all those sequences that are improperly aligned, possibly when the alignment parameters become unsuited to select specific templates in the low homology set. Therefore, our results are indicating that with profiles containing information about remote homologues, the SWA-PP program endowed with an entropy filter has a fold prediction accuracy of 98% and a probability of about 92% to have a significant MaxSub score. The efficacy of our method is also supported by comparing with what the structural alignment program CE obtains on the same set.

Our findings support the idea that information theory-based descriptors play a relevant role in improving the sensitivity of alignment methods (see also Refs. 29 and 30) highlighting a pattern for fold recognition based on the entropy evaluation of profile–profile alignments.

ACKNOWLEDGMENTS

This work was partially supported by grants of the Ministero della Università e della Ricerca Scientifica e Tecnologica (MURST) for the project "Hydrolases from Thermophiles: Structure, Function and Homologous and Heterologous Expression" and for the project "Development and Implementation of Algorithms for Protein Structure Prediction," a PNR 2001–2003 (FIRB art. 8) project on Bioinformatics, delivered to RC. EC and IR acknowledge the financial support from the SPINNER Consortium (Regione Emilia-Romagna) through a grant to the BioDec project.

REFERENCES

1. Flockner H, Braxenthaler M, Lackner P, Jaritz M, Ortner M, Sippl MJ. Progress in fold recognition. Proteins 1995;23:376–386.
2. Rost B, Sander C. Bringing the protein sequence-structure gap by structure predictions. Annu Rev Biophys Biomol Struct 1996;25: 113–36.
3. Moult J. Predicting protein three-dimensional structure. Curr Opin Biotechnol 1999;10:583–588.

4. Altschul SF Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DI. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25: 3389–3402.
5. Kelley LA, MacCallum RM, Sternberg MJE. Enhanced genome annotation using structural profiles in the program 3D-PSSM. J Mol Biol 2000;299:499–520.
6. Jones DT. GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. J Mol Biol 1999;287: 797–815.
7. Rychlewski J, Zhang B, Godzik A. Fold and function predictions for Mycoplasma genitalium proteins. Fold Des 1998;3:229–238.
8. Vogt G, Etzold T, Argos P. An assessment of amino acid exchange matrices in aligning protein sequences. The twilight-zone revisited. J Mol Biol 1995;249:816–831.
9. Burkhard R. Twilight zone of protein sequence alignments. Protein Eng 1999;12:85–94.
10. Rychlewski L, Jaroszewski L, Weizhong L, Godzik A. Comparison of sequence profiles. Strategies for structural predictions using sequence information. Protein Sci 2000;9:232–241.
11. Shannon CE, Weaver W. The mathematical theory of communication. Urbana, IL: University of Illinois Press; 1949.
12. Shindyalov IN, Bourne PE. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. Protein Eng 1998;11:739–747.
13. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of protein database for the investigation of sequences and structures. J Mol Biol 1995;247:536–540.
14. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM. CATH—a hierarchy classification of protein domain structures. Structure 1997;5:1093–1108.
15. Pearl FMG, Lee D, Bray JE, Sillitoe I, Todd AE, Harrison AP, Thornton JM, Orengo CA. Assigning genomic sequences to CATH. Nucleic Acids Res 2000;28:277–282.
16. Godzik A, Skolnick J, Kolinski A. Regularities in interaction patterns of globular proteins. Protein Eng 1993;6:801–810.
17. Godzik A. The structural alignment between two proteins: is there a unique answer? Protein Sci 1996;5:1325–1338.
18. Hubbard TJ. RMS/coverage-graph: a quality method for comparing three-dimensional structure predictions. Proteins 1999;S3:15–21.
19. Jones TA, Kleywegt GJ. CASP3 comparative modeling evaluation. Proteins 1999;S3:83–86.
20. Kabsch W. A discussion of the solution for the best rotation to relate two sets of vectors. Acta Crystallogr 1978;A34:827–828.
21. Siew N, Elofsson A, Rychlewski L, Fischer D. MaxSub: an automated measure for the assessment of protein structure prediction quality. Bioinformatics 2000;16:776–785.
22. Holm L, Sander C. Mapping the protein universe. Science 1996;273: 595–602.
23. Holm L, Sander C. Dictionary of recurrent domains in protein structures. Proteins 1998;33:88–96.
24. Gibrat JF, Madej T, Bryant SH. Surprising similarities in structure comparison. Curr Opin Struct Biol 1996;6:377–385.
25. Smith TS, Waterman MS. Identification of common molecular subsequences. J Mol Biol 1981;147:145–147.
26. Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci USA 1992;89:10915–10919.
27. Hermjakob H, Lang F, Apweiler R. SPTR—a comprehensive, non-redundant and up-to-date protein sequence database. Bioinformer 1998;4:http://bioinformer.ebi.ac.uk/newsletter/archives/4/sptr.html.
28. Huang X, Miller M. A time-efficient, linear-space local similarity algorithm. Adv Appl Math 1991;12:337–367.
29. Yona G, Levitt M. Within the twilight zone: a sensitive profile–profile comparison tool based on information theory. J Mol Biol 2002;315:1257–1275.
30. Panchenko AR. Finding weak similarities between proteins by sequence profile comparison. Nucleic Acids Res 2003;31:683–689.
31. Rychlewski L, Jaroszewski L, Godzik A. Improving the quality of twilight—zone alignments. Protein Sci 2000;9:1487–1496.
32. Park J, Teichmann SA, Hubbard T, Chothia C. Intermediate sequences find distant sequence homologues. J Mol Biol 1997;273: 349–354.

APPENDIX A

12ASA 1PYSA	1B35A 1B35C	1D1RA 1B6CB	1QGUA 1QGUB
19HCA 1NEW_	1B3AA 1DOKA	1D2NA 1QVAA	1SHCA 2NMBA
1A0AA 1AM9A	1B3TA 2BOPA	1D2VA 1MHLC	1TCA_ 1THG_
1A0CA 4XIS_	1B4CA 1PSRA	1D9CA 1FJCA	1UCYE 1UCYH
1A17_ 1E96B	1B4FA 1COKA	1D9NA 1QK9A	1WGJA 1ARB_
1A1Z_ 1NTCA	1B5EA 1BKPA	1DBTA 1DQWA	2BPA1 2BPA2
1A28A 1LBD_	1B5FB 1FKNA	1DBWA 1DZ3A	2IGD_ 2PTL_
1A34A 1AUYA	1B64_ 1GH8A	1DCEB 1FT1B	2QWC_ 3SIL
1A3AA 1A6JA	1B6E_ 1AYFA	1DCIA 1NZYA	
1A3K_ 1C1LA	1B6TA 1F9AA	1DHPA 1FBAA	
1A53_ 1NSJ_	1B87A 1BO4A	1DI0A 1CQKA	
1A5R_ 1UBI_	1B8OA 1ECPA	1DLXA 1QGIA	
1A6M_ 1ASH_	1B9HA 1BJ4A	1DMHA 3PCHA	
1A6O_ 2IF1_	1B9LA 1DHN_	1DT4A 1VIH_	
1A7TA 1SMLA	1BAQ_ 1EYVA	1DUN_ 1DUPA	
1A9V_ 1EHXA	1BBHA 1CPQ_	1DWNA 1MSC_	
1AAC_ 1BQK_	1BCFA 1DPSA	1E70M 1QOXN	
1ABA_ 1ERV_	1BCPD 1PRTF	1EAF_ 3CLA_	
1AC5_ 1IVYA	1BD3A 1DQNA	1EERA 1ETEA	
1ACP_ 2AF8_	1BD8_ 2MYO_	1EWIA 1OTCB	
1AD3A 1BPWA	1BDO_ 1FYC_	1EWXA 1QQ2A	
1ADEA 1BYI_	1BDYA 1RLW_	1EXG_ 1XBD_	
1AFRA 1MHYD	1BE3A 1BE3B	1F2DA 1OASA	
1AGDB 1RVV1	1BEFA 1JXPA	1FFKJ 1FFKL	

APPENDIX A (Continued)

1AGJA 2PRD_	1BG2_ 3KINB	1FFKK 1FJFK
1AH1_ 1CD8_	1BH9A 1BH9B	1FFKN 1FFKQ
1AH9_ 1D7QA	1BHE_ 1CZFA	1FIPA 1CY5A
1AIR_ 1EE6A	1BK7A 1BOLA	1FJ7A 2ILK_
1AIW_ 1ED7A	1BO9A 1DK5A	1FLTX 1TIT_
1AJ8A 1CSH_	1BPV_ 1C8PA	1FMB_ 1HVC_
1AJQA 1AJQB	1BQZ_ 1FAFA	1FQTA 1G8JB
1AKHA 1AKHB	1BS9_ 1CEX_	1FRB_ 1QRQA
1AKO_ 1BIX_	1BU7A 1EA1A	1FXD_ 2FDN_
1AL3_ 1ATG_	1BVB_ 1FGJA	1G24A 1LT3A
1ALY_ 1D4VB	1BVWA 1TML_	1G6GA 1QU5A
1AOEA 1D1GA	1BVYF 1RCF_	1GAKA 1LIS_
1AOHA 1NBCA	1BX4A 1RKD_	1GEN_ 1WBA_
1AOIA 1YTW_	1BXYA 1FFKT	1GKY_ 1NKSA
1AOJA 1AWJ_	1BY1A 1F5XA	1GPC_ 1GVP_
1AOXA 1ATZA	1BYKA 1DP4A	1HCRA 1TC3C
1AP0_ 1DZ1A	1BYUA 1CTQA	1HNR_ 1HUUA
1APYA 1APYB	1C0NA 1CL2A	1HYP_ 1RZL_
1AQ0A 1LTAC	1C1YB 1LFDA	1IRP_ 2I1B_
1AQB_ 1BBPA	1C3YA 1DQEA	1KVEA 1KVEB
1ARV_ 1BGP_	1C52_ 1CC5_	1KWAA 1HAVA
1AUIB 1CLL_	1C9FA 1D4BA	1LED_ 1NLS_
1AUWA 1FURA	1CAXB 1QI7A	1LMB3 1NEQ_
1AVAC 1HXN_	1CEM_ 1FCE_	1MFMA1YAIA
1AVOA 1AVOB	1CEWI 1EQKA	1MRJ_ 1DGWA
1AVPA 1EUVA	1CFE_ 1QNXA	1NEDA 1PMAA
1AW0_ 1CC8A	1CHKA 1EW4A	1NOX_ 1VFRA
1AWD_ 1BYFA	1CK9A 1FFKE	1ONC_ 7RSA_
1AWE_ 1BAK_	1CKTA 1HRYA	1OYC_ 2DORA
1AXJ_ 1CI0A	1CKV_ 1HQI_	1PAUA 1PAUB
1AYM1 1AYM2	1CMOA 1TUPA	1PBWA 1TX4A
1AZSA 1FX2A	1CNV_ 1EOKA	1POIA 1POIB
1B0UA 1F2TB	1CNZA 1ISO_	1PRU_ 1UXC_
1B16A 1BSVA	1CQQA 1PDR_	1PTY_ 1AOIB
1B20A 1RGEA	1CV8_ 1HUCA	1PYAA 1PYAB