Protein Science (2000), *9*:232–241. Cambridge University Press. Printed in the USA.

Comparison of sequence profiles. Strategies for structural predictions using sequence information

LESZEK RYCHLEWSKI,[1] LUKASZ JAROSZEWSKI,[2] WEIZHONG LI,[1] AND ADAM GODZIK[2]

[1]San Diego Supercomputer Center, La Jolla, California 92093
[2]The Burnham Institute, La Jolla, California 92037

(RECEIVED August 10, 1999; FINAL REVISION October 20, 1999; ACCEPTED October 22, 1999)

Abstract

Distant homologies between proteins are often discovered only after three-dimensional structures of both proteins are solved. The sequence divergence for such proteins can be so large that simple comparison of their sequences fails to identify any similarity. New generation of sensitive alignment tools use averaged sequences of entire homologous families (profiles) to detect such homologies. Several algorithms, including the newest generation of BLAST algorithms and BASIC, an algorithm used in our group to assign fold predictions for proteins from several genomes, are compared to each other on the large set of structurally similar proteins with little sequence similarity. Proteins in the benchmark are classified according to the level of their similarity, which allows us to demonstrate that most of the improvement of the new algorithms is achieved for proteins with strong functional similarities, with almost no progress in recognizing distant fold similarities.

It is also shown that details of profile calculation strongly influence its sensitivity in recognizing distant homologies. The most important choice is how to include information from diverging members of the family, avoiding generating false predictions, while accounting for entire sequence divergence within a family. PSI-BLAST takes a conservative approach, deriving a profile from core members of the family, providing a solid improvement without almost any false predictions. BASIC strives for better sensitivity by increasing the weight of divergent family members and paying the price in lower reliability. A new FFAS algorithm introduced here uses a new procedure for profile generation that takes into account all the relations within the family and matches BASIC sensitivity with PSI-BLAST like reliability.

Keywords: fold recognition; PSI-BLAST; sequence profile

A simple observation that homologous proteins have similar folds and strong similarities in their functions forms a cornerstone of most methods of predicting protein structure and function from sequence. Structure and/or function prediction is usually based on establishing homology between a newly sequenced protein and an already known and characterized protein group. Once the homology is established, it is possible to make various inferences about the structure, activity, and function of the new protein.

Unfortunately, deciding whether or not two proteins are homologous, i.e., related by evolution is not always easy. The usual approach is to look for similarity between their amino acid sequences. Dynamic programming (Needleman & Wunsch, 1970) provides a very powerful and fast method to compare two sequences. Extensive experience with this approach established quite precise thresholds when the similarity is strong enough to infer that the two proteins are related. The rule of thumb is that for proteins of the approximate length of 100 amino acids, two proteins with sequence similarity around the level of 25% of identities have about 50% chance of being related. The range of sequence similarity around and beyond this threshold is often referred to as a twilight zone. The specific value of the threshold depends on the length of the alignment (Sander & Schneider, 1993). Also, other measures of sequence similarity, such as statistical significance of the alignment, allow for better distinction between homologies and spurious sequence similarities (Brenner et al., 1998).

Quickly growing databases of well-characterized proteins provide many examples of proteins with sequence similarity below any reasonable threshold, and yet many similarities in structure and function. For some protein groups, the debate about the exact relations between their members is still ongoing, but in the majority of cases the consensus is that such proteins represent very distant evolutionary relations. The existence of many pairs and groups of such proteins has inspired the efforts to increase the sensitivity of sequence comparison and to detect homology between proteins based on weak sequence similarities or other additional information.

Reprint requests to: Adam Godzik, Program in Bioinformatics and Biological Complexity, The Burnham Institute, 10901 N. Torrey Pines Road, La Jolla, California 92037; e-mail: adam@burnham-inst.org.

One of the possible approaches to increase the sensitivity of sequence comparison is to change the definition of sequence similarity. For instance, instead of comparing two proteins by building an alignment of their sequences, it is possible to compare two protein families, using information contained in the sequences of all proteins in each family. There are many ways this information can be used to enhance sequence comparison, for instance profile methods (see later for a detailed discussion), sequence signature libraries (Bairoch et al., 1996), hidden Markov models (Krogh et al., 1994), and the intermediate sequence search procedure (Park et al., 1997). The most common way of processing and representing the sequences of a family of proteins is to build a multiple alignment of all sequences and describe it by a profile—a generalized sequence, containing information about mutations allowed at each position in the protein family (Gribskov et al., 1987; Bork & Gibson, 1996). On the algorithmic level, a profile can be used in a dynamic programming subroutine (or any other alignment method), with the only change being that the scoring becomes position dependent. It has been shown that replacing a single sequence by a profile increases the accuracy of comparison methods as compared to the simple sequence alignment (Altschul et al., 1997; Park et al., 1998; Aravind & Koonin, 1999). Various claims were made as to relative advantages of one method over the other, but evaluation and comparison of different algorithms are rather difficult, because of the lack of universally accepted measures of success. Recently, several algorithms including PSI-BLAST (Altschul et al., 1997), hidden Markov models (Krogh et al., 1994), and the intermediate sequence search procedure (Park et al., 1997) were compared on a large benchmark (Park et al., 1998) based on the SCOP classification of protein structures (Murzin et al., 1995).

In this spirit, we evaluate several possible strategies of building and comparing sequence profiles on a benchmark developed from the exhaustive clustering of all known protein structures. In this publication, we focus on closely related algorithms that differ in well-defined and well-understood ways and yet (as shown in the paper) differ significantly in their homology recognition sensitivity.

Evaluation of sequence alignment algorithms on structure similarity benchmarks is limited by necessity to proteins with known structure, but at the same time, allows the use of independent verification criteria. On a practical side, by limiting the database of proteins used for comparison to proteins with known structures, these sequence-only methods can be applied to the problem of structure prediction and directly compete with fold recognition or threading algorithms. Surprisingly, despite their different points of origin, both profile and threading methods seem to give similar results and reliability estimates, at least in limited tests (Rychlewski et al., 1998). Therefore, in several previous papers (Rychlewski et al., 1998, 1999; Pawlowski et al., 1999) we have used a sequence based method for fold assignments. Detailed comparison of sequence-only and threading algorithms is a subject of a separate publication.

This paper is built as follows: a fold recognition benchmark is constructed from several existing classifications of protein structures. In a further step, different profile building and comparing strategies are tested on this benchmark. Among others, we present details of the methods used in our group to assign folds to proteins from several genomes (Rychlewski et al., 1998, 1999; Pawlowski et al., 1999). Here, we compare it to the newest version of the PSI-BLAST algorithm (Altschul et al., 1997) and to the next generation profile-profile alignment method developed in our group.

Results

The benchmark

There are several classifications dividing known protein structures into families based on their structural similarities. Some, such as CE (Shindyalov & Bourne, 1998), VAST (Gibrat et al., 1996), or DALI/FSSP (Holm & Sander, 1998), are based on all-by-all comparisons of all protein structures. Some others, such as SCOP (Murzin et al., 1995) or CATH (Orengo et al., 1997), are based on a hierarchical tree of similarities between proteins, with functional and structural similarities both defining a position of a given protein on the tree. Similar classifications can be made using only sequence similarity, without any reference to structure. Comparing the two classifications, one based on structure, one based on sequence, we can identify pairs of proteins that belong to the same structural family but to two different sequence families. Such pairs can be used in a hypothetical experiment: could we have predicted the structural similarity of the two proteins, knowing only their sequences, or perhaps the structure of one and the sequence of the other. Statistics on such experiments can be used to compare various fold prediction and recognition strategies, especially when the number of such pairs is large enough to exclude simple memorization effects.

In this paper, we are using a list of over 900 proteins pairs identified from DALI (Holm & Sander, 1998) and SCOP (Murzin et al., 1995) database. The DALI database was used for selection of protein pairs of significant structural similarity but low sequence similarity (see Materials and methods). SCOP was used to verify the structural similarity of the pair and to assess the level of similarity (fold, superfamily, family). The full benchmark list (as well as a full list of benchmark results for all methods discussed here) is available from our WEB server at bioinformatics.burnham-inst.org/benchmarks.

To assess the quality of a given prediction method, each of the test proteins was compared to all of the proteins in the Protein Data Bank (PDB) database of proteins with known structure, or the nonredundant database of protein sequences in the case of PSI-BLAST. Proteins were not divided into domains, neither on the query side nor on the database side. This way the fold assignment is closer to the real-life situation where domain boundaries are often not known.

Benchmark statistics

In this paper, we compare four different sensitive sequence comparison algorithms. Full details of various algorithms and their implementation are given in Materials and methods. The first algorithm is PSI-BLAST (Altschul et al., 1997), the latest version of the BLAST algorithm. In this version, the profile built from already recognized members of the homologous family is compared to a large sequence database. In consecutive iterations, new proteins are recognized and added to the family, the profile is recalculated, and the procedure is run until convergence or for a predetermined number of steps. In this paper, PSI-BLAST is used for structure prediction by scanning its output for proteins with known structures, identified by the "PDB" keyword in the sequence database.

The second method, which we call here PDB-BLAST, is a specific implementation of the PSI-BLAST algorithm, where the actual profile from the last step of the previous procedure is saved

Table 1. *The comparison of methods at different levels of significance*[a]

E-value	PSI-BLAST	PDB-BLAST	Z-score	BASIC	FFAS
10^{-10}	250/0	281/0	50	154/0	145/0
10^{-5}	270/0	**330/1**	20	252/1	266/0
0.01	274/0	357/4	15	**293/2**	302/0
0.1	281/0	371/7	10	337/16	349/0
0.5	287/0	383/16	8	364/20	**373/0**
1	**292/1**	396/56	7	391/29	390/26
5	296/5	402/97	6	424/38	427/59
10	302/6	409/418	0	**510/419**	**526/403**

[a]The number of correct predictions/number of false positives is shown for each method and significance level.

and used to scan a database of proteins with known structures, which is a subset of the entire sequence database.

The third algorithm is a profile–profile alignment algorithm BASIC (Rychlewski et al., 1998) developed in our group and used previously to assign folds to proteins from several genomes (Rychlewski et al., 1998, 1999; Pawlowski et al., 1999), in a fold prediction competition CASP3 (Murzin, 1999) and experimental fold prediction competition between automated fold prediction servers CAFASP (Kelley et al., 1999). This algorithm compares a sequence profile of a query protein to a library of profiles representing known protein structures. The two main differences between this algorithm and PDB-BLAST algorithm is a different, simplified procedure for profile calculation and using profiles on both sides of the alignment.

Finally, a new profile–profile alignment algorithm FFAS is introduced and compared to others. This algorithm is similar to BASIC in using profile information of both sides of the alignment, but it is based on a novel procedure for profile preparation from the multiple alignment of sequences in the family of homologous proteins.

All four methods are applied to the benchmark, with results presented in Tables 1–3. In the first table, the number of correctly and incorrectly predicted structures is shown as a function of prediction significance. The methods were compared in terms of the numbers of correct predictions, with a distinction made between reliable and tentative predictions. The prediction is counted as reliable when the score is better than the reliability threshold, chosen in such a way that there is less that 1% of false predictions above it. Tentative predictions are based on fold assignments with the highest score. Because the way the score significance is calculated differs significantly between different methods, the positions that should be compared directly, i.e., the number of correct reliable predictions and number of correct tentative predictions are highlighted by underscore or bold characters, respectively.

As seen the Table 1, the PSI-BLAST alone can correctly recognize 31% of proteins in the test set. This value is lower than the one reported previously (44%) on a similar benchmark (Park et al., 1998) because of the lower similarity thresholds for protein pairs included in the benchmark (30% in the current manuscript vs. 40 (Park et al., 1998)). It is remarkable that a majority of proteins can be recognized with e-value significance smaller than 1, with no false positive found for this significance level. On the other hand, only a very small number of proteins can be found within the significance levels between 1 and 10, with about a 50/50 chance of the prediction in this significance level being wrong. It is interesting to note that the number of correctly recognized structures increases to 35% in PDB-BLAST, where the information used in prediction does not change, only the choices given to the program are more limited. The most remarkable improvement happens at the significance levels, which already include some false positives. For predictions at all significance levels, the difference between PSI-BLAST and PDB-BLAST exceeds 30%, i.e., there are 107 new, correct predictions. This result is truly puzzling because both methods use exactly the same profile and alignment procedure. The difference between these two methods will be further discussed in one of the following paragraphs.

The BASIC algorithm closely reproduces the results of the PSI-BLAST algorithm for high reliability predictions, with 293 correct predictions for the Z-score of 15, as compared to 292 for the e.value of 1. Beyond that point, the number of false positive predictions is growing rapidly, but at the same time the number of correct predictions is growing too. This is in direct contrast to the PSI-BLAST, which is either correct or does not give any answer. At the lowest significance level, the difference between BASIC or FFAS and PSI-BLAST reaches almost 70%. For the former two methods, the number of correct predictions on the low significance levels approaches 55%. Clearly the strategy of extending the homologous family (see Material and methods) results in increasing both the number of correct predictions and the number of false positives.

From the results of the large benchmark presented here, it is clear that the significance levels used in earlier genome fold assignments were overly optimistic, being based on benchmarks with a small number of examples. The BASIC algorithm was developed on the benchmark of 68 protein pairs and tested on a set of three benchmarks with a total number of about 50 protein pairs (Rychlewski et al., 1998). For instance, the threshold of Z-score equal to 7 was used in the fold assignments for the *Escherichia coli* and

Table 2. *The comparison of the results of different methods and SCOP similarity levels*[a]

	BLAST	BLAST2	PSI-BLAST	PDB-BLAST	BASIC	FFAS
SCOP-level 3 (279)	0/3	0/2	5/11	5/12	4/47	7/41
SCOP-level 4 (262)	1/13	16/26	55/71	55/90	51/129	80/135
SCOP-level 5 (388)	17/44	130/151	232/243	270/308	238/334	286/350

[a]The number of *reliable* and *tentative* predictions is shown for each method and each SCOP similarity level.

Table 3. *The results of the benchmark for PSI-BLAST and PDB-BLAST methods for low significance range*[a]

e-Value threshold	PSI-BLAST	PDB-BLAST
10	307/11	409/418
200	321/151	410/509
2000	325/404	

[a]The number of hits/false positives is shown for each method and each e-value threshold.

Helicobacter pylori genomes. Based on the results from this study, we can expect about 7% of false positives at this level. On the other hand, the number of false positives is still only 14% at the level of Z-score equal to 5. Thus, many previously disregarded fold predictions have a good chance of being correct.

Finally, the FFAS algorithm gives the highest number of correct predictions with the relatively low number of false positives. For a threshold of Z-score equal to 8, it achieves 373 correct predictions with no false positive, which is about 12% better than the PDB-BLAST and 26% better than standard PSI-BLAST. The number of correct predictions at the low significance level is also higher than that for other methods, including the BASIC algorithm. The more elaborate profile weighting scheme made it possible to find a balance between sensitivity (how many correct predictions) and reliability (how many false positives).

Homology recognition vs. fold assignment

The 929 protein pairs in the benchmark could be divided into three groups depending on the probable relationship between them. SCOP classification allowed us to divide the entire list into subgroups with different levels of structural similarity. Using the SCOP database, all protein pairs in the list could be divided into three levels of similarity:

- SCOP level 5, i.e., family level (388 pairs). On this level, proteins are believed to be homologous despite limited sequence similarity, based on significant similarities in their function.
- SCOP level 4, i.e., the superfamily level (262 pairs). On this level, there are some analogies between function of different proteins, but the homology is not obvious.
- SCOP level 3, i.e., fold family (279 pairs). At this level, there is no similarity between function and proteins are believed not to be homologous.

Table 2 illustrates the prediction rate of various algorithms on different groups of proteins. All the methods recognize predominantly homologous proteins from the SCOP level 5. In this group the success rate approaches 75% for the FFAS method, and if low significance predictions are included, it reaches 85%. Even the best method reaches only 2% success (16%, if low reliability predictions from FFAS are included) rate for the SCOP level 3, where proteins are thought not to be homologous. Also the differences between methods are minimal—the same protein pairs are recognized by all methods. The SCOP level 4, where some level of functional similarity between proteins is present and their homology is possible, the success rate is slightly higher, reaching 28% at the reliable level and getting close to 50% for low significance predictions. At this point it is not clear if the cases of successful predictions in the SCOP categories 3 and 4 represent cases where profile methods were able to generalize the sequence to arrive at fold definition, thus becoming de facto threading methods, or perhaps these pairs are actually homologous but were classified in the wrong category by SCOP.

However, at this point it is clear that true to their roots, all the methods studied in this manuscript are predominantly based on recognition of distant homologies.

The comparison of the methods

The four methods compared in the present study are based on very similar ideas (see Table 5). Each one is trying to extract information contained in the multiple alignments of protein families to arrive at position specific mutation matrix. And yet, as shown in the previous paragraph, the success rate in fold recognition can vary by as much as 50% between different methods with some apparently trivial changes resulting in surprisingly large differences. In this paragraph we attempt to "dissect" various methods to arrive at better understanding of factors contributing to successes and failures of different methods.

One of the most intriguing differences is the difference between PSI-BLAST and PDB-BLAST results. Both methods use the same algorithm both for profile calculations and for alignment and the only difference is the smaller number of choices given to the PDB-BLAST algorithm. Results presented in Table 1 were obtained with standard PSI-BLAST parameters, which limit the number of proteins included in the output. As a result, there were numerous cases where no PDB structure was included in the PSI-BLAST output. This is the reason why the number of correct answers and number of false positives did not add up for PSI-BLAST to the total number of cases, as they did for BASIC, FFAS, and PDB-BLAST.

To test whether this effect is responsible for the observed results, the parameters were changed to increase the number of protein included in the output. The results are summarized in Table 3.

Even when the e-value threshold was increased to 2,000, there were still over 200 families where no structural prediction, right or wrong, can be made. In some of these cases, limiting the search only to proteins with known structure, as done in PDB-BLAST, makes it possible to arrive at a correct answer. This effect contributed to the more than 30% difference between PSI-BLAST and PDB-BLAST predictions. Thus, the PDB-BLAST strategy is a purely technical trick allowing one to bypass one of the trivial, but difficult to solve problems in PSI-BLAST applications, namely the output size.

The two methods presented here, BASIC and FFAS, differ in two main points: what proteins are included in the multiple alignment used for profile generation and in a weighting scheme used to calculate the profile.

Two variants of the FFAS method were tested—one, denoted FFAS/M in Table 4, where multiple alignment was built in the same way as in the BASIC algorithm, i.e., with only one PSI-BLAST iteration and with e-value threshold extended to 0.1 instead of standard parameters of 5 iterations and e-value threshold of 1e-3 (see Materials and methods). As seen in Table 4, this variant results in recognition rates that are slightly worse than that of BASIC, i.e., in this case the new weighting scheme did not improve the results, but instead made them slightly worse. Clearly

Table 4. *The comparison of the results of different profile-to-profile methods for different SCOP similarity levels*

	BASIC	FFAS/W	FFAS/M	FFAS
SCOP-level 3 (279)	4/47	6/26	3/42	7/41
SCOP-level 4 (262)	51/129	75/116	44/134	80/135
SCOP-level 5 (388)	238/334	290/336	226/336	286/350

the balance between weight from distant homologues can be regulated in two ways: one by increasing the number of proteins included in the multiple alignment (the BASIC way) or by more careful weighting scheme (the FFAS way). When both methods are used, the effect it results in overcompensation. The second FFAS variant tested here, denoted as FFAS/W in Table 4, uses a variant of the weighting scheme that gives a smaller weights to more distant homologues. As seen in Table 4, this difference does not influence the number of reliable predictions, but decreases the number of low reliability predictions.

The diversity of the profile and fold recognition

Interestingly, all methods fail to directly recognize 100% of benchmark pairs even on the SCOP family level, where an evolutionary relationship exists. It is interesting to investigate what decides about the success or failure of the relationship detection. The diversity of the sequences belonging to the profile is known to be one of the most important factors influencing the detection (Aravind & Koonin, 1999). We have investigated the correlation between the diversity of profile sequences and the recognition result. The sequence diversity can be roughly measured by the geometrical average of e-values of all the sequences contained in the PSI-BLAST profile in the first iteration. (Increasing the average e-value for this profile means a greater number of distant homologues in that profile.) Indeed, the benchmark statistics show that detection result for a given pair is strongly connected with the abundance of distant homologues in query and template sequence profiles (see Fig. 1). The percentage of successful predictions reaches its maximum when the average PSI-BLAST profile e-value is between 1e−10 and 1e−3 (see Fig. 1). In this e-value range, the sequences are unquestionably related to the sequence for which the profile was created and, on the other hand, different enough to yield important information to the profile.

Table 5. *The differences between the methods compared in this study*

PSI BLAST
- Multiple alignment: five iterations with 10^{-3} e-value threshold
- Profile: preclustering with 98% identity cutoff; pseudo count based variability estimation—background amino acid frequencies
- Database: database of nonredundant sequences (NR)

PDB-BLAST
- Multiple alignment: same as PSI-BLAST
- Profile: same as PSI-BLAST
- Database: sequences of all proteins from PDB database

BASIC
- Multiple alignment: two PSI-BLAST iterations with 0.1 e-value threshold
- Profile: preclustering with 97% identity cutoff; amino-acid composition filter; distant homologues have smaller weights
- Database: profiles of proteins from PDB

FFAS
- Multiple alignment: same as PSI-BLAST
- Profile: preclustering with 97% identity cutoff; amino-acid composition filter; sequence diversity based weight
- Database: profiles of proteins from PDB

Interdependence of various methods

The results presented in Table 1 rank-order various methods based on the total number of correct predictions. With different strategies of profile calculations and scoring systems, it is interesting to check if the same targets are being recognized by all methods. The interdependence between predictions of BASIC, FFAS, and PDB-BLAST is illustrated in Figure 2 (PSI-BLAST results form a complete subset of PDB-BLAST results and were omitted from the figure). The figure shows that the majority of predictions is made simultaneously by all methods, but each method is able to recognize some pairs missed by all other methods. For instance, there are 419 proteins that are predicted by at least one method with high significance, which is 12% better that any single method.

Why profile-to-profile methods work

Apart from different weighting schemes used by PSI-BLAST and BASIC or FFAS, the more fundamental difference is that in the two latter methods the profiles are used on both sides of the alignment. However, other approaches, such as Intermediate Sequence Search (Park et al., 1997) or starting PSI-BLAST search from different proteins from the homologous family (Aravind & Koonin, 1999), can de facto explore the sequence divergence on both sides of the comparison. In the ISS method, one searches for the sequence or series of sequences, which are simultaneously similar to the query sequence and to one of the sequences from database. Anecdotal evidence suggests that there are cases where protein A can be reliably identified as being homologous to B, and B is reliably homologous to C, which allows to classify A and C as

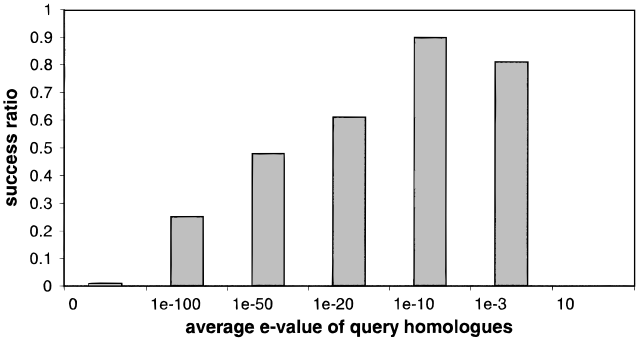

Fig. 1. The distribution of homologues of the query sequence in NR database and the probability of correct fold assignment for this sequence.

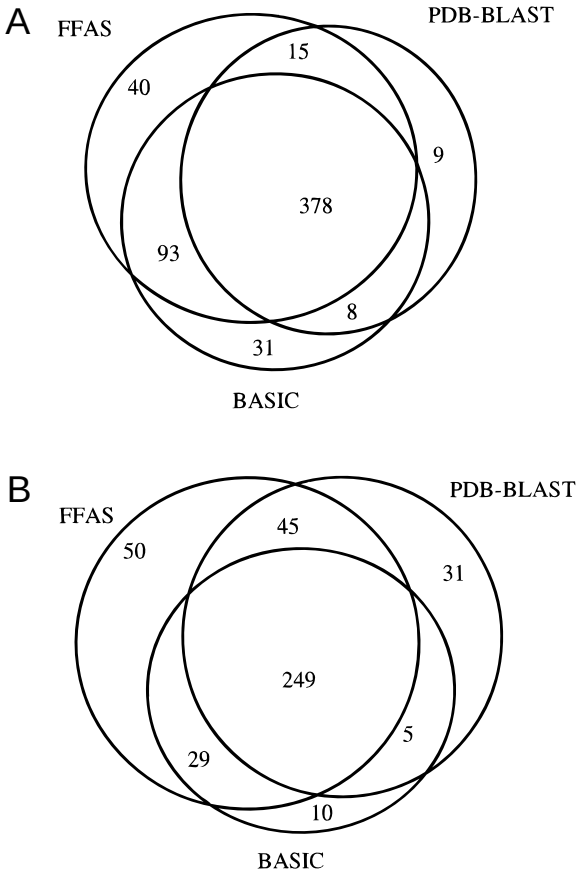

Fig. 2. Venn diagrams of correct predictions made with different methods. The number of correct predictions is shown for *reliable* (**A**) and *tentative* (**B**) significance level.

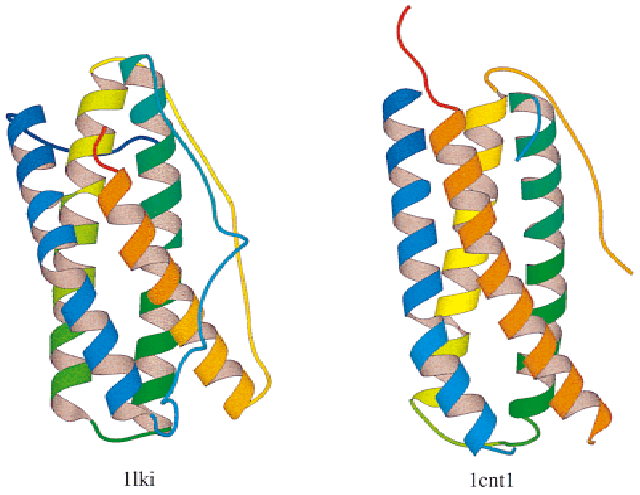

Fig. 3. The structures of helical cytokines used as an example of structural and functional relationship impossible to detect with PSI-BLAST and detectable with FFAS method.

homologous, despite the fact that A and C are not being directly recognized as homologous. Protein B is than called an "intermediate" or "bridging" protein between A and C.

Some results for the ISS strategy suggest that on a comparable benchmark it performs worse than PSI-BLAST (Park et al., 1998), but it is not clear that some other algorithmic implementation of this idea would not perform better. We are currently working on implementing a similar search strategy using the FFAS algorithm. However, it is not difficult to find an example of a protein pair, which is seen as similar by FFAS, but not by PSI-BLAST or ISS-like search strategies performed by hand.

Two proteins, Leukemia Inhibitory Factor (gi:999942, PDB code 1lki) and Ciliary Neurotrophic Factor (gi:116585, PDB code 1cnt), belong to a family of helical cytokines with a characteristic four helical up-up-down-down topology. Both proteins have very similar structures, with the root-mean-square deviation of 1.9 Å for 158 C_α positions with sequence similarity of 16% of identical amino acid (see Fig. 3). Similarity between both proteins can be seen by the FFAS algorithm with a Z-score of 8.

The sequence families of the Leukemia Inhibitory Factor and Ciliary Neurotrophic Factor represented in the database were found to consist of 11 and 7 proteins, respectively (PSI-BLAST search in NR database). All proteins from both families were used to find possible homologues. For all proteins from both proteins, PSI-Blast searches were conducted and in all cases the method converged at the previously found family boundaries. All nonsignificant hits with E-value cutoff of 10.0, where additionally verified by performing PSI-Blast searches and screening for hits to proteins belonging to the families under investigation. As a result, no additional sequence was found that would have a PSI-Blast hit to any of the members of the family. Thus, this pair could not be identified with any variant of the ISS-like strategy and at least in this case, it is only a generalization of features of both sequences via profile building that allowed their similarity to be recognized on the sequence level.

Discussion

A large fold recognition benchmark was prepared on the basis of existing classifications of protein structures into families, superfamilies, and fold families. This allowed the comparison of several different strategies of extracting information from multiple alignment of homologous families for the purpose of recognition of very distant homologies. Two novel algorithms were compared to the newest generation of the BLAST family of programs, the position specific iterative BLAST (Altschul et al., 1997).

The results clearly illustrate that it is possible to improve the recognition rate by developing more elaborate ways of profile calculation and by increasing the information about the proteins that are used for comparison. A multiple alignment contains a lot of information and there are different ways to translate it into a profile. A most important choice that must be made is how to balance two effects: introducing of new information, which leads to increased sensitivity in recognizing distant homologues with avoidance of errors, leading to incorrect homology assignments. The authors of the PSI-BLAST algorithm did a perfect job in avoiding the latter problem, while making immense progress on the former. However, other choices are possible, leading to improved recognition sensitivity. Recognition of distant homologues with low significance requires additional tools to verify the homology. Other tools, such as modeling, function analysis, etc., might be needed to improve the reliability of such predictions.

It should be noted that profile-profile matching methods (like BASIC or FFAS) are much more time and memory demanding than profile-sequence matching methods (like PSI-BLAST). First of all, FFAS uses the dynamic programming alignment, which is significantly slower than the BLAST algorithm. On the same CPU, the average search of a 100 residue protein against a database of representative proteins from PDB takes about the same time as five iterations of PSI-BLAST search against the nonredundant sequence database. The second reason is that for FFAS the sequence profile must be precalculated for all proteins from database the new sequence is compared to. On the other hand, this calculation must be done only once, so it is not so important in repeated calculations.

For all algorithms, best results were obtained for recognition of homologous proteins, sharing similarities in function. For the benchmark used in this work, almost all proteins from this group were recognized by at least one method. None of the methods studied here achieved much success in recognizing nonhomologous proteins sharing similar folds. Clearly, additional sources of information, such as those used by threading algorithms, would be necessary to achieve success for proteins from this group.

Materials and methods

Three paragraphs in this section cover three basic areas of development of new methods: how they are tested (benchmark), how the profile is created, and how the profile is compared to another profile (or sequence).

Benchmark calculation

Preparation of the benchmark

The benchmark set used in the method tests consists of 929 pairs of structurally similar proteins with low sequence similarity. The selection of pairs was based on the database of recurrent domains in protein structures (DALI) (Holm & Sander, 1998). The domain pairs of high structural similarity (*Z-score* >10) and low sequence similarity (sequence identity $<30\%$) have been selected. We have found 1,730 pairs that fulfill this criterion. Then, the redundant sequence pairs consisting of similar query and target sequences (sequence similarity of *both* query and target sequences higher than 75%) have been eliminated. Finally, the similarity of all sequence pairs have been additionally verified using SCOP database. The pairs, which were not found to be similar in SCOP, have been removed from the benchmark. At this stage, the set of benchmark pairs has been divided into three subsets:

1. The pairs that are classified as similar at the SCOP family level (388 pairs). There is extensive function similarity and significant sequence similarity at this level. The proteins within family are expected to be homologous.
2. The pairs that are classified as similar at the superfamily level of SCOP (262 pairs). There are some analogies in function at this level, but the evolutionary relation between the proteins is not certain.
3. The pairs of proteins having the same fold (279 pairs). There are no similarities in function, the proteins are thought to be unrelated.

Both SCOP and DALI databases contain protein domains. The benchmark, however, was intended to simulate the real sequence database search, when the protein structures and consequently the division of the proteins into domains remains unknown. Hence, instead of protein domain sequences, the whole protein sequences have been used. The similarity of most closely related domain pair of two sequences is taken as the similarity of a given sequence pair.

The benchmark results calculations

For each query sequence, the database search was performed. The sequences that are structurally similar to the query according to the DALI database, but different than benchmark target for this query, have been removed from the results set for each query sequence. All the sequences having greater e-values (in the case of PSI-BLAST) or lower Z-scores (in the case of FFAS) than a given threshold have been also removed form the results set. If the highest scoring sequence from the remaining results set is the target for a given query, then the result for this benchmark pair has been counted as the correct prediction. If this is not true, then the result is counted as a false positive (as all proteins structurally similar to the query and different than the target have been previously excluded from the results set).

Each subset of the benchmark has been recalculated for several values of *Z-score* or *e-value* threshold starting from the results sets of high significance and ending with the results set where no constraint is applied for *Z-score* or *e-value* and simply the best scoring result was taken. The largest results set with the percent of false positives lower than 1% was denoted as *reliable* level for each method. The results set obtained without any constraints on the scores was denoted as a *tentative* prediction level for each method (see Table 2).

Profile building and comparison

There are three main steps in designing the profile alignment method:

1. Preparation of the multiple alignment of the entire homologous family of the prediction target.
2. Transformation of the multiple alignment into a position specific scoring matrix or profile.
3. Alignment and scoring procedure used to compare the profile prepared in point 2 with the database of sequences or profiles.

Multiple alignment

In all four cases the multiple alignment was prepared based on the PSI-BLAST output. However, the BASIC method uses a different significance threshold for acceptance in the homologous family (E-value of 0.1 instead of 10^{-3} for all other methods). This results in a broader and faster approximation of the homologous family, which often includes spurious similarities, contributing to the high rate of wrong fold assignments by the BASIC method.

Profile

The second step is the generation of the profile. This is the most important difference between all profile-based methods. Different algorithms differ by the precise way the weights are assigned and the way sequences and positions with a low level of sequence variation are treated. The simplest solution is to average all sequences from the multiple alignment with equal weights. This approach is for instance used in the Profile algorithm from the GCG software package (Group, 1991), originally developed by Gribskov and Eisenberg (Gribskov et al., 1987). This approach is

very sensitive to redundancies in the database, where closely related or even multiple entries of identical proteins are often found. All methods used here perform a filtering procedure removing highly identical sequences, leaving only a set where all sequences have an identity lower than 97 or 98% to each other.

The profile generation routine used in PSI-Blast is described in (Altschul et al., 1997). In short, the multiple alignment is purged leaving sequences with mutual identity lower than 98%. The sequences are weighted based on a procedure described in Henikoff and Henikoff (1994). A pseudo count method is used to estimate the number of independent observations. The row profile weighted by the pseudo counts is combined with a balanced profile generated using data-dependent target frequencies (Tatusov et al., 1994).

BASIC uses sequences with E-value up to 10, taken from PSI-Blast output. It purges them with a 97% cutoff. A composition filter is used to remove sequences with a different amino acid composition than the query sequence. This procedure deletes mainly low complexity sequences. The sequences are weighted lower if they show very weak similarity to the query sequence. The weighting of all sequences with E-value <0.1 is almost equal 1. Only sequences with E-value between 10.0 and 0.1 are seriously affected by the weighting scheme. The sum of weights is used as pseudo count, and the procedure of balancing similar to the routine in Blast is followed.

FFAS uses sequences with E-value below 0.1 and has a more sophisticated weighting scheme. Weights are assigned based on the dissimilarity of the sequence in respect to the family. The higher the divergence of the sequence the higher the weight. The evaluation of dissimilarity is based on comparing all sequences of the family with each other. The sum of weights is also used as pseudo count for the profile balancing routine.

Alignment

The sequence profile representing a family of homologous proteins can be compared to a database of protein sequences or a database of profiles, representing a specific set of protein families. PSI-BLAST and PDB-BLAST take the first approach, but with different databases. BASIC and FFAS use the second approach, with the database of profiles prepared in an identical way to that of the profile for the prediction target.

PSI-BLAST and PDB-BLAST use a specific alignment method based on recognition and expansion of high scoring fragments (Altschul et al., 1990). BASIC and FFAS use a standard local-local dynamic programming alignment (Needleman & Wunsch, 1970).

In all previously mentioned profile-to-profile alignment methods, the score of comparison of two profile positions is defined as the dot product between the profile vectors corresponding to those positions. In contrast, profile-to-sequence alignment methods like PSI-Blast do not require the computationally expensive dot product calculation. The evaluation of the score involves only a lookup of the value in the profile vector corresponding to the amino acid at the aligned position in the sequence.

In addition, FFAS performs a normalization of the matrix containing the comparison scores between all position of both aligned profiles before the best path is searched for. This procedure has proven to result in higher sensitivity and offers some advantages when changing the profile generation method. After normalization the values in the matrix have a more consistent distribution and it is easier to transfer previously optimized alignment parameters, like gap penalties. A detailed description of the FFAS method is given in the Appendix.

Acknowledgments

The research described in this manuscript was supported by the NIH grant GM-60049.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol 215(3)*:403–410.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res 25(17)*:3389–3402.

Aravind L, Koonin EV. 1999. Gleaning non-trivial structural, functional and evolutionary information about proteins by iterative database searches. *J Mol Biol 287(5)*:1023–1040.

Bairoch A, Bucher P, Hofmann K. 1996. The PROSITE database, its status in 1995. *Nucleic Acids Res 24(1)*:189–196.

Bork P, Gibson TJ. 1996. Applying motif and profile searches. *Methods Enzymol 266*:162–184.

Brenner SE, Chothia C, Hubbard TJ. 1998. Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. *Proc Natl Acad Sci USA 95(11)*:6073–6078.

Fischer D, Elofsson A, Rice D, Eisenberg D. 1996. Assessing the performance of fold recognition methods by means of a comprehensive benchmark. *Pac Symp Biocomput 97*:300–318.

Gibrat JF, Madej T, Bryant SH. 1996. Surprising similarities in structure comparison. *Curr Opin Struct Biol 6(3)*:377–385.

Gribskov M, McLachlan AD, Eisenberg D. 1987. Profile analysis: Detection of distantly related proteins. *Proc Natl Acad Sci USA 84(13)*:4355–4358.

Group GC. 1991. Program manual for the GCG package, version 7. (April 1991, 575 Science Drive, Madison, WI 53711).

Henikoff S, Henikoff JG. 1994. Position-based sequence weights. *J Mol Biol 243(4)*:574–578.

Holm L, Sander C. 1998. Dictionary of recurrent domains in protein structures. *Proteins 33(1)*:88–96.

Kelley LA, MacCallum RM, Sternberg M, Karplus K, Fischer D, Elofsson A, Godzik A, Rychlewski L, Pawlowski K, Jones D, et al. 1999. CAFASP-1: Critical assessment of fully automated structure prediction methods. *Proteins*. Forthcoming.

Krogh A, Mian IS, Haussler D. 1994. A hidden Markov model that finds genes in *E. coli* DNA. *Nucleic Acids Res 22(22)*:4768–4778.

Murzin AG. 1999. Structure classification-based assessment of CASP3 predictions for the fold recognition targets. *Proteins Suppl 3*:88–103.

Murzin AG, Brenner SE, Hubbard T, Chothia C. 1995. SCOP: A structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol 247(4)*:536–540.

Needleman SB, Wunsch CD. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol 48(3)*:443–453.

Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM. 1997. CATH—A hierarchic classification of protein domain structures. *Structure 5(8)*:1093–1108.

Park J, Karplus K, Barrett C, Hughey R, Haussler D, Hubbard T, Chothia C. 1998. Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. *J Mol Biol 284(4)*:1201–1210.

Park J, Teichmann SA, Hubbard T, Chothia C. 1997. Intermediate sequences increase the detection of homology between sequences. *J Mol Biol 273(1)*: 349–354.

Pawlowski K, Zhang B, Rychlewski L, Godzik A. 1999. The *Helicobacter pylori* genome: From sequence analysis to structural and functional predictions. *Proteins 36(1)*:20–30.

Pearson WR. 1998. Empirical statistical estimates for sequence similarity searches. *J Mol Biol 276(1)*:71–84.

Rychlewski L, Zhang B, Godzik A. 1998. Fold and function predictions for *Mycoplasma genitalium* proteins. *Fold Des 3(4)*:229–238.

Rychlewski L, Zhang B, Godzik A. 1999. Functional insights from structural predictions: Analysis of the *Escherichia coli* genome. *Protein Sci 8(3)*:614–624.

Sander C, Schneider R. 1993. The HSSP data base of protein structure-sequence alignments. *Nucleic Acids Res 21(13)*:3105–3109.

Shindyalov IN, Bourne PE. 1998. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng 11(9)*:739–747.

Tatusov RL, Altschul SF, Koonin EV. 1994. Detection of conserved segments in proteins: iterative scanning of sequence databases with alignment blocks. *Proc Natl Acad Sci USA 91(25)*:12091–12095.

Appendix

The description of the FFAS method

The calculation of the multiple sequence alignment

Every representative sequence from the clustered template database was used as an input for 5 PSI-Blast iterations with E-value cutoff equal 0.001 using nonredundant database (NR). All sequences that obtained E-value <0.001 in any iteration were included in the multiple sequence alignments. The resulting set of sequences was purged, leaving only sequences with less than 97% identity to other sequences. Sequences with the amino acid composition significantly different than the query sequence were removed from the profile (the composition difference threshold of 0.05 was used). The composition difference is defined as follows:

$$\Delta Comp = \frac{\sqrt{\sum_{a=1}^{20} (f_{i,a} - f_{j,a})^2}}{20}$$

$$\sum_{a=1}^{20} f_{i,a} = \sum_{a=1}^{20} f_{j,a} = 1 \qquad \text{(A1)}$$

where $\Delta Comp$ is the composition difference and $f_{i,a}$ is the fraction of amino acid of type a in the aligned part of sequence i.

The assignment of weights to the sequences in the profile

A profile consists of an array of 22-dimensional vectors, one for every position in the query sequence. Usually in the profile based methods, the sequences in the profile are unequally weighted. In the FFAS method, the weight assigned to a given sequence is based on the diversity of that sequence relative to other sequences in the family. When the sequence is less similar to other sequences in the profile, it has higher weight in the profile. The similarity score of two sequences is calculated using the same multiple sequence alignment, which is used in the family profile. The alignment is taken from PSI-Blast output, and the similarity scores are calculated directly from it using BLOSUM 62 mutation matrix. The alignment scores are than transformed into similarity scores, which have only values between 0 and 1.

$S_{i,j} = \max[A_{i,j}/\min(A_{i,i}, A_{j,j}), 0]$ can adopt values between 0 and 1;

$S_{i,j}$ = similarity score of sequences i and j;

$A_{i,j}$ = alignment score of sequences i and j calculated with mutation matrix.

The diversity score of each sequence is calculated based on its similarity scores to all other sequences in the profile.

$$D_i = \frac{1}{1 + \sum_j s_{i,j}^2}$$

where

D_i = diversity score of sequence i;

$S_{i,j}$ = similarity score of sequences i and j.

The diversity score used as the weight of the sequence in the family profile. This weighting scheme has the effect that the homologues that are more distant from the family obtain higher weights.

The calculation of the profiles

Each position in the family profile is represented with a vector. The vector consists of 20 values representing the frequencies of occurrence of all amino acids, one value representing deletions (gaps), and one value equal to the sum of diversity scores of all sequences aligned at this position.

$$f_a = \frac{\sum_{i,\, a_i = a} D_i}{\sum_i D_i}$$

$$\sum_{a=1}^{21} f_a = 1$$

where

f_a = fraction of amino acid a (or gap) in the sequences aligned a given position;

D_i = diversity score for sequence i;

a_i = amino acid type in the sequence i at this position.

The transformation of the profiles

In the cases when the diversity of the sequences in the family profile is low, the profile does not contain sufficient information about mutation characteristics in the family. The expected distribution of amino acids is added to the profile to balance it out in cases when the family consists of only few sequences. The expected distribution is calculated using the average probabilities of mutations in the PDB database. For every amino acid type, the probability of mutation to another amino acid type was estimated using statistical mutation probabilities in the multiple sequence alignments calculated with PSI-Blast.

$$\sum_{b=1}^{21} p_{b,a} = 1$$

$p_{a,b}$ = probability of a mutation from amino acid b to a (or deletion).

This probability table is used to balance out the profiles:

$$f'_a = \sum_{b=1}^{20} f_b \cdot p_{b,a}$$

where

f'_a = new fraction score of amino acid a in the vector of the balanced profile;

f_b = fraction score of amino acid b in the original family profile;

$p_{b,a}$ = statistical probability of a mutation from amino acid b to a observed in sequence database.

The balanced family profile is multiplied by a constant weight of 5 and added to the original family profile. This value is based on the average diversity of sequence profiles in database. The weight of the original profile is determined based on the diversity of the sequences in the family

profile. The diversity of the family profile is defined as the sum of diversity scores of all sequences aligned at a given position. The diversity of the family profile is stored at the 22nd position of the family profile vector.

$$f_a'' = 5 \cdot f_a' + f_a \cdot \sum_i D_i$$

$$f_a''' = \frac{f_a''^2}{\sum_{a=1}^{21} f_a''^2}$$

where

f_a''' = final fraction score of amino acid a;

f_a' = fraction score of amino acid a in the vector of the balanced profile;

f_a = fraction score of amino acid a (or gap) in the original family profile;

D_i = diversity score of sequence i.

The calculation of the alignment of two profiles

Standard dynamic programming (Needleman & Wunsch, 1970) is used to align two profiles. This algorithm requires the score matrix containing the comparison score for each pair of positions in the sequences (or profiles) to be aligned. The comparison score for a given pair of positions in two profiles is equal to the dot product of vectors found at these positions.

$$C_{m,n} = \sum_{a=1}^{20} f'''_{m,a} \cdot f'''_{n,a}$$

where

$C_{m,n}$ = comparison score for positions m and n;

$f'''_{m,n}$ = final fraction score for amino acid a at position m of the first profile.

The matrix is then normalized so that its average score is equal a constant value (equal to the matrix zero shift) and its standard deviation is equal to 1. This allows using the same optimal gap initiation and extension penalties for different profile comparison protocols. The choice of gap penalties and matrix zero shift parameter is crucial for the effectiveness of the method, as shown in benchmark tests. The optimization of these three parameters using comprehensive benchmark of sequence pairs requires large computational resources. It takes about 24 h on a single Pentium II 400 MHz processor. To optimize gap penalty parameters at the 40 processors of a CRAY T3E with each 120 Mbytes of RAM memory were used. The number of correctly recognized pairs in the UCLA-DOE fold recognition benchmark (Fischer et al., 1996) was used as the objective function for the optimization. The benchmark consists of 68 query sequences and the database of 300 templates. The optimized gap initiation and gap extension penalties are equal to 4.71 and 0.37, respectively, and zero shift parameter optimal value is equal to -0.12.

The transformation of the alignment scores into Z-scores and E-values

The alignment scores for each pair of profiles are transformed into E-values and Z-scores to perform significance evaluation. The transformation is based on the hypothesis that the alignment scores obtained for the set of proteins follow the extreme value distribution (Pearson, 1998). For every profile the set of alignment scores is obtained by comparing it to all other profiles in the database. The set of scores is transformed to remove the logarithmic sequence length effect of the template protein, so that the new score has approximately the same expected value independent on the size of template protein. The extreme value distribution formula is then fitted to the resulting data. The extreme value distribution is calculated separately for every profile and is used to translate every score into a p-value. The p-value is defined as the expected probability to obtain a certain score by chance. This p-value is transformed into an equivalent gauss distribution Z-score (a Z-score that has the same p-value in a gauss distribution as the obtained p-value in the extreme value distributed scores). This Z-score (or p-value) can be also easily presented as a database size dependent E-value as used in Blast. The E-value describes the number of random hits expected to have a score better than a given score. The previously obtained gauss distribution Z-scores undergo an additional transformation. A pair of proteins can have two different Z-scores depending on which of the two proteins was used for extreme value distribution calculation. To make Z-score assignment method symmetrical, the minimum of two alternative Z-scores is used and is called M-score. The distribution of M-scores is used to express and report them as Z-scores again.