PROTEINS: Structure, Function, and Genetics 46:330–339 (2002)
DOI 10.1002/prot.10043

A Study on Protein Sequence Alignment Quality

Arne Elofsson*
Stockholm Bioinformatics Center, Stockholm University, SE-10691, Stockholm, Sweden

ABSTRACT **One of the most central methods in bioinformatics is the alignment of two protein or DNA sequences. However, so far large-scale benchmarks examining the quality of these alignments are scarce. On the other hand, recently several large-scale studies of the capacity of different methods to identify related sequences has led to new insights about the performance of fold recognition methods. To increase our understanding about fold recognition methods, we present a large-scale benchmark of alignment quality. We compare alignments from several different alignment methods, including sequence alignments, hidden Markov models, PSI-BLAST, CLUSTALW, and threading methods. For most methods, the alignment quality increases significantly at about 20% sequence identity. The difference in alignment quality between different methods is quite small, and the main difference can be seen at the exact positioning of the sharp rise in alignment quality, that is, around 15–20% sequence identity. The alignments are improved by using structural information. In general, the best alignments are obtained by methods that use predicted secondary structure information and sequence profiles obtained from PSI-BLAST. One interesting observation is that for different pairs many different methods create the best alignments. This finding implies that if a method that could select the best alignment method for each pair existed, a significant improvement of the alignment quality could be gained. Proteins 2002;46:330–339.**
© 2002 Wiley-Liss, Inc.

Key words: sequence alignment; hidden Markov models; dynamic programming; homology modeling; fold recognition

INTRODUCTION

As the genome projects proceed, we are presented with an exponentially increasing number of protein sequences but with only a very limited knowledge of their structure or function. Because structure and function determination is a nontrivial task even for a single protein, the best way to gain understanding of these genes is if we can relate them to proteins with known properties by searching databases. Improving such algorithms is one of the fundamental challenges in bioinformatics today. By determining how sequences are related to known proteins, we can make predictions of their structural, functional, and evolutionary features. Relationships between proteins span a broad range from the case of almost identical sequences to apparently unrelated sequences that share only rough three-dimensional structure. This poses different challenges to the detection algorithms used, a method excellent at finding sequence similarity might not perform very well in the case of only structural relationship or vice versa.

Within the emerging field of structural genomics, the goal is to determine the structure of all proteins. Because it is not feasible to determine the structure of all proteins experimentally, a lot of effort is put into building homology models. To create a homology model, first, one or several homologs of known structure are identified; then, alignments are made, and finally, a three-dimensional model is created. If a close homolog is found, a high-quality model can often be created by using automatic methods. However, it has been shown than even if a fairly close homolog is identified, it is often not trivial to automatically create a good alignment.[1] Often a gray zone is identified where homologs with 20–40% identity can readily be identified but where the alignment might not be good enough for the creation of a good model.

There are many methods to create an alignment: single sequence based, such as Smith-Waterman[2] and Needleman-Wunch,[3] multiple sequence alignments,[4] profile based,[5,6] prediction based,[7–9] and structure based.[10] Furthermore, several groups have used other structural information from the template, for instance using special gap penalties in loop regions,[11] whereas other groups have been using special alignment techniques.[12] Although all these methods use different information and different methodologies, they all produce a score and an alignment. In several recent studies, the ability of different methods to detect proteins that share the same fold has been studied.[13–18] However, recognizing the correct fold is often not enough to deduce the function of a protein or build a good three-dimensional model of the protein. Obviously, a better model could be made if in addition to a correctly identified fold, a correct alignment was found. Such a

Abbreviations: Scop, the Structural Classification of Proteins database; family, protein domains that are closely related because of a common origin according to Scop; superfamily, protein domains of probable common origin according to Scop; fold, protein domains that have major structural similarities according to Scop.

Grant sponsor: Swedish Natural Science's Research Council; Grant sponsor: Swedish Research Council for Engineering Sciences.

*Correspondence to: Arne Elofsson, Stockholm Bioinformatics Center, Stockholm University, SE-10691 Stockholm, Sweden. E-mail: arne@sbc.su.se.

Received 19 March 2001; Accepted 28 September 2001

© 2002 WILEY-LISS, INC.

model might be easier to use for understanding the function of the protein. Practically, all methods developed for fold recognition are based on alignments; therefore, obtaining the alignment is mostly an integrated part of the method. However, most large-scale comparisons have completely ignored this aspect of these methods. It is not evident that a better fold recognition method also provides a better alignment. For instance, there are many examples where the correct fold is identified even when the alignment is wrong.[19] In other cases, only a small fraction of the protein is aligned correctly.

Several studies of alignment quality have been performed. However, most of them have used a small set of pairs and a small set of methods, whereas we here use a large benchmark and several different methods. This process enables us to get a more complete picture of the performance of different alignment methods on different types of targets. It has been shown that a threading method aligned distantly related proteins better than a sequence alignment method.[20] In another study, it was shown that a method that combined threading energies with sequence information improved the alignment quality.[21] In a third study, it was also shown that evolutionary information improved the alignment quality.[22] These studies were based on 127–190 structurally related pairs of proteins, respectively, whereas in this study we used more than 8000 pairs. Our set includes proteins from widely different folds and varying degree of similarity, whereas in these earlier studies, only distantly related proteins were included. Furthermore, we used a large set of alignment methods, whereas these studies only compared a few methods. In a third recent study,[23] a similar large-scale benchmark was used to compare the alignment quality between BLAST, intermediate sequence searches (ISS), PSI-BLAST, and CLUSTALW. However, in this study no methods using structural information were included.

One important aspect when studying alignment quality is how the alignment quality is measured. For instance in an earlier study,[23] BLAST can be seen to perform better or worse than CLUSTALW, depending on the choice of measure (f_M or f_D); see below. In the next section we briefly review some different measures and explain why we mainly used the *LGscore*[24] measure.

Measuring alignment quality can be done in two fundamentally different ways. One method is to obtain a "true alignment" and then compare the obtained alignment with this. The true alignment is often obtained from a structural alignment.[23,25] The alternative to using a true alignment is to build a model from the alignment and evaluate the similarity between the model and the structure. This measure can then be used to measure the quality of the alignment. The first approach was used in the studies by Domingues et al.[20] and Sauder et al.,[23] whereas the quality of the model is compared in this study and in the study by Panchenko et al.[21]

The true alignment-based measures are sometimes referred to as template-based measures, because they compare the difference between two alignments to a template. The measure reported from these methods is related to the number of aligned residues that are identical between the structural alignment and the sequence alignment. One problem with this approach is that alignments obtained by different structural alignment programs might differ substantially, especially when the sequence similarity is low,[26,27] that is, the true alignment might not be unique. In the studies by Domingues et al.,[20] this problem was solved by using several alternative structural alignments and selecting the one that has the highest degree of agreement with the sequence alignment.

One commonly used type of model-based measures is often referred to as "alignment-dependent measures."[24] These measures are based on the detection of a common segment between a model and a structure. This segment should be as long as possible but also as similar as possible. The measure *LGscore* used in this study and the contact specificity used in Panchenko et al.[21] are examples of such measures. The difference between different alignment-dependent measures is mainly due to how the similarity and length are combined into a score. This type of measure has been used to detect correct structures in CAFASP2[28] and CASP4 and also in the LiveBench experiments.[29] The measure used here, *LGscore,* was one of three measures used in CAFASP2 and LiveBench, whereas in CASP4 another alignment-dependent measure was used in the threading evaluation. In a recent review we compared a large set of these measures, and we found that given enough targets, several measures would have reproduced the manual ranking.[24]

We do not believe that the fundamental differences between template-based measures and alignment-dependent measures are of great importance. However, what might be crucial are the details in what is measured. An example can be taken from the study by Sauder et al.,[23] in which they used two measures, f_M and f_D. Both of these measure the fraction of correctly aligned residues in comparison with a structural alignment. The difference is that f_M measures the number of correctly aligned residues divided by the number of aligned residues, in the sequence alignment, whereas f_D compares the correct residues with the number of aligned residues in the structural alignments. Obviously, a local alignment that is shorter will benefit from using f_M than f_D. In Figure 7 of Sauder et al.,[23] the difference can be seen; when using f_M CLUSTALW performs significantly worse than BLAST; however, when using f_D, it performs better. In this study, we chose to use a measure that focuses on the correct sections of an alignment, that is, a measure that should correlate more with f_D than f_M.

RESULTS

In this study we compared a number of alignment methods by using a benchmark containing 17,118 pairs of proteins. For each pair and method, a model is generated and the quality of this model is evaluated. There are several possibilities how these data then can be analyzed. Below, we describe some of these methods. One important conclusion from large-scale benchmarks of fold recognition methods is that the same methods do not necessarily

TABLE I. The Best Sequence Alignment Parameters as Measured by the Average Log (*LGscore*)

Algorithm	Matrix	GO	GE	Family	Superfamily	Fold
global	BLOSUM-62	−15	−1	15.1	2.9	1.4
global	BLOSUM-50	−20	−1	15.1	2.9	1.4
global	BLOSUM-45	−20	−1	15.0	2.9	1.4
global	PAM-250	−15	−1	14.8	2.8	1.2
local	BLOSUM-62	−5	−1	13.3	1.8	0.5
local	BLOSUM-50	−10	−1	13.9	1.9	0.6
local	BLOSUM-45	−10	−1	14.1	2.0	0.7
local	PAM-250	−10	−1	14.0	1.8	0.6

TABLE II. Description of Alignment Methods

Name	Version	Parameters	Description
local	—	BLOSUM-45, GO = −10, GE = −1	Local sequence alignment
global	—	BLOSUM-62, GO = −15, GE = −1	Global sequence alignment
PSI		GO = −11, GE = −1	Global alignment against a profile obtained from PSI-BLAST
SAM	SAM-T98	default	Alignment against an HMM built from the multiple sequence alignments obtained by PSI-BLAST
HMMER	2.1	default	Alignment against a HMM built from the multiple sequence alignments obtained by PSI-BLAST
CLUSTALW	1.8	default	All sequences from two PSI-BLAST searches
SS	—	GO = −11, GE = −1, S = 5, S′ = −0.5, PAM-250	Global alignment including predicted secondary structure information
SSPSI	—	GO = −11, GE = −1, S = 5, S′ = −0.5	Global alignment against a PSI-BLAST profile using predicted secondary structure information
LG score2 structural		default	Structural alignments

perform best on targets of different difficulty.[18] Therefore, we sort all pairs by difficulty in two different ways. First, we separate models into family, superfamily, or fold-related pairs; second, we separate the models by the sequence identity after a structural superposition.

Average *LGscore*

The first approach we applied to evaluate the performance of different methods is to simply calculate the average log (*LGscore*) for each pair. This is identical to how the *LGscore* has been used in CAFASP[28] and LiveBench.[29] In Table I, the average *LGscore* for the best choice of penalties for each substitution matrix is shown. There is a strong difference between the average *LGscore* for family-related pairs (about 15) and more distantly related pairs (an average *LGscore* of <4). The best performance is obtained by using the global alignment algorithm and BLOSUM-62 or BLOSUM-50. Many combinations of different gap penalties and global alignments produce almost identical results. For instance, using BLOSUM-62 and a gap-opening penalty of −10, −15, or −20 with the gap extension penalty of −1 produce average *LGscore* of 14.9, 15.1, and 14.9 for the family-related pairs. It is of interest that using local alignments BLOSUM-45, BLOSUM-50, or PAM-250 produce better alignments than BLOSUM-62. The worst results are in general obtained by using a gap extension penalty of 0, but also a combination of high opening and extension penalties perform badly; data not shown.

In this study we do not only use pairwise sequence alignment methods but also multiple sequence-based methods and threading methods; See table II. The average *LGscore* for these are reported in Table III. Independently of the relationship between the two proteins the best average performance is obtained by *SSPSI,* which uses both predicted secondary structure information and multiple sequence alignment information. The improvement only seems significant for distantly related pairs. The only method that completely ignores sequence information, THREADER,[10] performs significantly worse than the other methods.

TABLE III. Average Log (*LGscore*)

Method	Family	Superfamily	Fold
local	14.1	2.0	0.7
global	15.1	2.9	1.4
PSI	15.8	3.3	1.4
HMMER	14.7	2.7	1.3
SAM	14.6	2.3	1.0
CLUSTALW	15.1	3.4	1.7
SS	15.2	3.8	2.4
SSPSI	16.0	4.1	2.6
THREADER	5.2	1.2	1.0
structural	19.4	9.1	8.0

We have also compared the average *LGscore* versus the sequence identity for the pairs; see Figure 1. Figure 1 shows that the curves are quite similar for all methods, except THREADER. All methods indicate a sharp increase in the model quality between 20 and 30% identity. The best average quality is obtained by *SSPSI* for pairs with

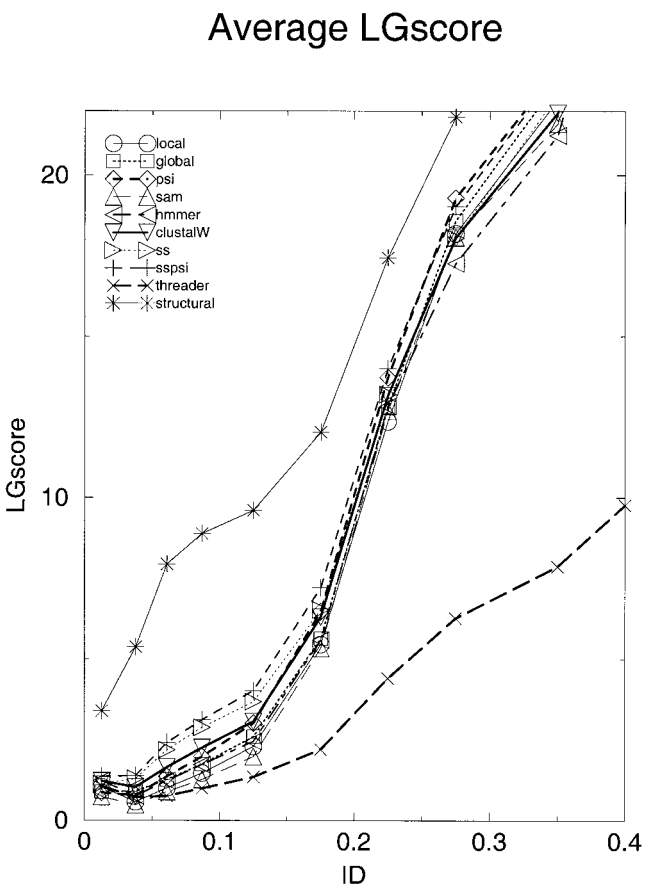


Fig. 1. The average *LGscore* for all methods versus sequence identity.

<30% sequence identity, whereas for more similar pairs *PSI* performs as well. It can also be noted that the structural alignments are only slightly better then sequence alignment methods in the higher sequence identity range. The main performance differences between different methods can be found at <20% sequence identity.

Fraction Correct Models

An alternative method to analyze the data is to examine how many models that are correct. Obviously, this method depends on how you define if a model is correct. We used two different cutoffs of *LGscore* to define correctness as well as two different root-mean-square deviation (RMSD) cutoffs (3 Å and 5 Å). The looser *LGscore* cutoff (10^{-3}) is the same as that used in LiveBench to define a correct model, whereas the more stringent cutoff (10^{-5}) is used to identify models of higher quality. For a 100-residue-long fragment, without any gaps, the two cutoffs correspond to RMS differences of 6.2 and 4.6 Å. The two RMSD cutoffs used (5 Å and 3 Å) would in these cases be slightly tougher, but for a shorter protein they could be more relaxed.

Tables IV and V show the fraction of correct models for the different methods. The highest fraction of correct models, in all categories, is obtained by *SSPSI,* followed by *SS*. By using the *LGscore* measure, (see Table IV), *PSI* performs on par with *SS,* whereas by using the RMSD measure, none of the multiple sequence-based methods seem to perform better than standard global alignments.

In Figures 2 and 3, the fraction of "correct" models (using the looser cutoffs) versus sequence identity is plotted. It can be seen that the number of correct pairs increases sharply at 20% identity. At low sequence identity, *SSPSI* and *SS* create the most correct models, whereas at higher sequence identity, all methods, except THREADER,[10] create mainly "correct" models. However, three methods, CLUSTALW,[4] SAM,[30] and HMMER,[31] seem to make slightly more mistakes at high identities than the other methods. By using the RMSD measures, even *PSI* seems to misalign some of the easier targets. All these methods are based on multiple sequence information. The increase of fraction good models at very low sequence identities is due to some few distantly related short proteins. This finding indicated one of the problems with use of the RMSD measure.

Best Models

It is also possible to study what method makes the best model for a specific pair. Below we refer to this as the best models. In this analysis, we ignored all pairs where no method made a correct model (using the 10^{-3} and 5 Å as cutoffs); these models are marked as none. In Tables VI and VII, it is seen that for family-related models, several different methods create a significant part of the best models, whereas for fold-related pairs, most come from *SSPSI* and *SS*. It can also be noted that the best single method (*SSPSI*) only makes correct models for 6% (14% using the RMSD measure) of the fold-related pairs, whereas all methods together make correct models for almost twice as many (11%/23%) of the pairs.

DISCUSSION

In several earlier studies, a predefined classification of proteins into folds, or families, was used to benchmark different fold recognition methods.[13–18] These benchmarks measured the ability to recognize the fold but ignored the quality of the alignment. In this study, we addressed the alignment problem by using a similar type of large benchmark. This benchmark was quite similar to the benchmark used in Sauder et al.[23] However, we tested another method to evaluate the accuracy of the alignments and a larger set of alignment methods, including threading methods.

Independent of alignment methods and evaluation method it is quite obvious that at approximately 20% identity there is a sharp increase in quality; see Figures 1, 2, and 3. The main difference between the different alignment methods lies in the exact position of this increase. For instance, it can be seen that *SSPSI* creates 50% correct models at 15% identity (using the *LGscore* measure), whereas the local alignment algorithm does not reach 50% until 19% identity (Fig. 2). The strong correlation between alignment quality and identity was a surprise to us, especially because it had been shown earlier that alignment quality is not a very good measure for the ability to detect related proteins.[13] In the study by Sauder

TABLE IV. Fraction of Models That Are Correct in %

	LGscore $< 10^{-3}$			*LGscore* $< 10^{-5}$		
Method	Family	Superfamily	Fold	Family	Superfamily	Fold
local	66	10	1	46	2	0
global	70	12	1	49	3	0
PSI	72	18	2	51	4	0
HMMER	68	13	2	49	3	0
SAM	66	12	1	47	3	0
CLUSTALW	69	18	2	49	4	0
SS	72	18	4	50	4	0
SSPSI	73	21	6	53	5	0
THREADER	37	4	1	12	0	0
structural	86	60	51	66	21	21

TABLE V. Fraction of Models That Are Correct in %

	RMSD < 5 Å			RMSD < 3 Å		
Method	Family	Superfamily	Fold	Family	Superfamily	Fold
local	60	8	1	45	3	1
global	64	15	6	46	5	2
PSI	60	13	6	43	5	2
HMMER	63	13	6	47	5	1
SAM	62	10	4	46	4	1
SS	70	22	13	48	6	2
SSPSI	71	24	14	51	7	2
THREADER	10	1	1	4	0	0
structural	87	76	77	83	59	50

et al.[23] they did not observe such a sharp increase. However, this is most likely due to what was measured. They measured the number of residues that were correctly aligned, whereas we measured either the average quality of the models or the fraction of models that were correct. Even if the number of correctly aligned residues increases quite linearly, it is possible that the number of correct models increase nonlinearly.

It should also be noted that the differences between the different methods is quite small. Taking this into account and that no single measure of alignment quality is perfect, it is important not to overstate the significance of the differences between different methods.

First, we analyzed standard sequence alignments. For each substitution matrix and alignment method we examined a total of 16 gap penalties. The results for the best choice of parameters are shown in Table I. It can be noted that for all substitution matrices the best gap extension penalty is -1, whereas the gap-opening penalty is -15 or -20 for global alignments and -5 or -10 for local alignments. One reason why a lower gap-opening penalty is preferred for local alignments is that the alignments would be too short using a higher gap penalty. For global alignment all the BLOSUM matrices performed slightly better than PAM-250, whereas for local alignments BLOSUM-62 performed worse than the other matrices. Furthermore, slightly lower gap-opening penalties are preferred for matrices with lower PAM values. When local alignments are used, the alignments only contain the part of the two proteins that are similar enough. If a matrix with a low PAM value, such as BLOSUM-62, is used, the alignment might only contain the parts that are highly similar. This will result in a shorter alignment than if a matrix of a higher gap penalty was used. To compensate for this, it is necessary to use lower gap penalties. For local alignments, this combination does not seem to be ideal, whereas it works well for global alignments.

In general, global alignments performed better than local alignments. This is in contrast with earlier observation.[23,25] However, as described above, this difference is due to differences in evaluation methods. Global alignment methods actually align more residues correctly; however, fewer of the aligned residues are aligned correctly.

Multiple Sequence Information Makes, at Most, Small Improvements to the Alignment Quality

It is well established that multiple sequence information helps to recognize distantly related proteins.[13–18] It is also expected that the alignments should be better by using multiple sequence information. In a recent study it was shown that PSI-BLAST and ISS align significantly more residues correctly than BLAST does, whereas CLUSTALW did not perform as well as PSI-BLAST.[23] In these studies, a smaller database was used for the CLUSTALW alignments than for the PSI-BLAST searches, whereas in this study we used the same set of sequences for a fair comparison between the two algorithms. It should be noted that the computational time for the CLUSTALW alignments is significantly longer than for all other alignment methods. This limits the use of CLUSTALW for large-scale alignment projects. In the *PSI* procedure, a

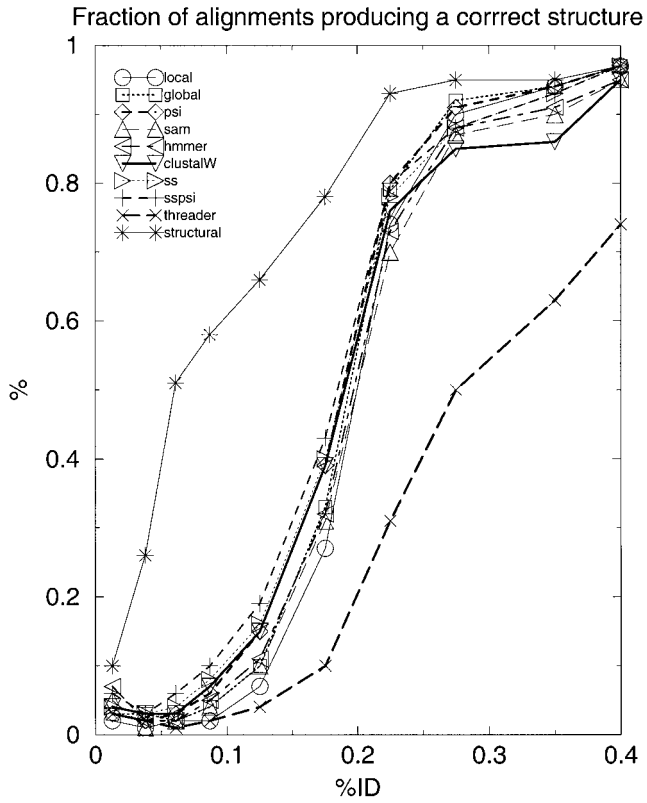


Fig. 2. The fraction of correct models versus identity, using a *LGscore* $< 10^{-3}$ as a definition of a correct model.

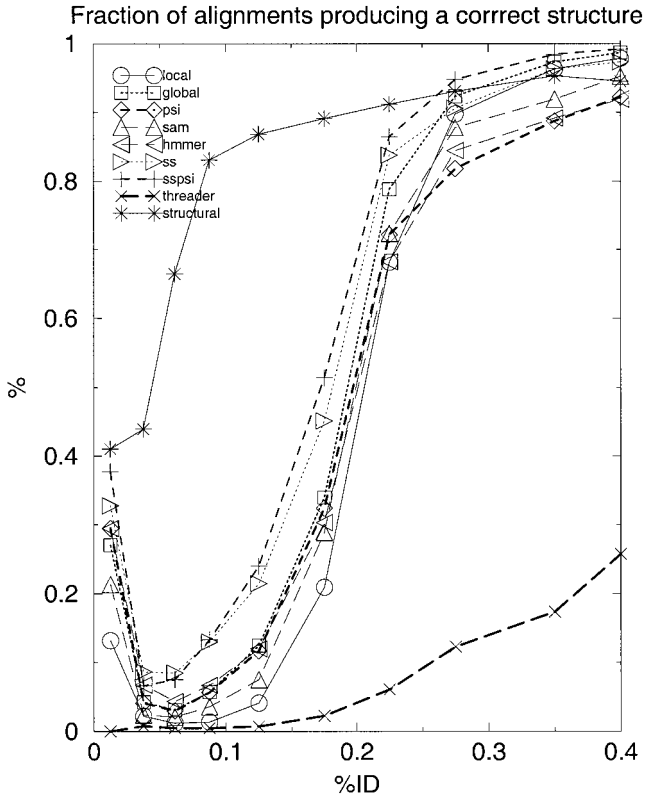


Fig. 3. The fraction of correct models versus identity, using RMSD < 3 Å as a definition of a correct model.

TABLE VI. Frequency When a Specific Method Perform Best in % Using the *LGscore* Measure

Method	Family	Superfamily	Fold
local	4	1	0
global	6	2	1
PSI	10	4	1
HMMER	17	3	1
SAM	13	3	1
CLUSTALW	13	7	1
SS	7	5	2
SSPSI	10	7	3
THREADER	1	1	1
NONE	19	67	89

TABLE VII. Frequency When a Specific Method Perform Best in % Using RMSD as a Measure

Method	Family	Superfamily	Fold
local	18	2	1
global	8	4	2
PSI	8	4	2
HMMER	14	6	3
SAM	12	3	1
SS	9	9	7
SSPSI	8	9	6
THREADER	0	0	0
NONE	21	63	77

global alignment method against a profile obtained from PSI-BLAST was used. This is similar to how PSI-BLAST was used for fold recognition in earlier studies.[32,33]

In addition to the comparison between PSI-BLAST profiles and CLUSTALW, we also used two hidden Markov models, HMMER and SAM, using the multiple sequence alignments from PSI-BLAST. For CLUSTALW, SAM, and HMMER we used default parameters, whereas for *PSI* we optimized the parameters by using an independent training set; see Materials and Methods.

In general, it can be seen that multiple sequence alignment methods perform better than single sequence methods using the *LGscore* measure. However, with use of the RMSD measure, no improvement can be seen. Taking the difficulties to detect small improvements, our data do not support the earlier observation that multiple sequence information improves the alignment quality significantly.

Also noticeable is that SAM, HMMER, and CLUSTALW (and possibly also *PSI*) all make some mistakes on pairs with high sequence identity (Fig. 2). One reason for these models is that the query and template proteins belong to the same subfamily of a larger family, which makes it is easier to align the two sequences to each other than to align both of them to the complete family. Another source of errors is the inclusion of unrelated files by the PSI-BLAST search.

In the earlier study by Sauder et al.[23] it was shown that ISS and PSI-BLAST helped the alignment quality. A more careful study shows that f_D, which is the measure that should be most similar to our measure, is about 10% better for PSI-BLAST than for BLAST. This difference seems

quite similar over the whole studied range of similarities (0–30%). These results are quite comparable to the improvements that we see by using the *LGscore* measure. However, when the f_M measure is used, they only see an improvement for PSI-BLAST over BLAST at <15% sequence identity. This finding also indicates that the improvements in alignment quality obtained from multiple sequence information is quite small and that the results might also depend on the choice of evaluation method.

Predicted Secondary Structures Improves the Alignments

Many methods that use structural information to detect distantly related proteins have been developed. In recent large-scale benchmarks, it was indicated that these methods perform better than purely sequence-based methods proteins.[29]

Here, the best overall alignments are obtained by using a combination of predicted secondary structures and evolutionary information in the *SSPSI* method; see Tables III, IV, and V. The improvement from predicted secondary structure information is largest at lower sequence identity; see Figures 1, 2, and 3.

The good performance of *SSPSI* inspired us to examine if it is possible to increase the alignment quality even if the structure is not know. This can be done by using the predicted secondary structure of the template sequence instead of the correct secondary structure. By using this approach, the alignment quality is at a level between the *PSI* and *SSPSI* methods. The average *LGscore* values for this method are 15.8, 3.8, and 2.4 for the family, superfamily, and fold related proteins, respectively.

It can also be noted that THREADER performs significantly worse than all other methods. As had been observed earlier, THREADER is not very specific, but it detects more distantly related proteins than sequence-based methods.[18] However, here we show that, on average, the alignment quality is not better even for distantly related proteins.

What Method Makes the Best Models?

For each pair, obviously some method will produce the best model. In Tables VI and VII, the frequencies that a method produces this model are shown. For proteins that are related on the family level, most best models are created by the purely sequence-based alignment methods, whereas for the fold-related models, most best models are created by *SSPSI*. This finding emphasizes the conclusion that predicted secondary structures are mostly useful for distantly related proteins, whereas for closely related proteins the improvement of alignment quality is marginal.

All methods sometimes produce the best model, and all the methods taken together perform much better than the best individual method. For instance, the best method, *SSPSI* is able to produce a correct model for 6% (14% using the RMSD measure) of the fold-related pairs, but if the best method was used for each pair, 11% (23%) of the models would be correct. The observation that not a single method always produces the best alignment is in agreement with experience from studies of sequence alignments. In these, it is often noted that it is possible to obtain a good alignment even for distantly related proteins given the right choice of parameters. However, the right choice of parameters is not identical in all cases, that is, no single choice of gap penalties and methods always produces the best alignment. Additional evidence for this conclusion can be found by the benchmarking of web-based fold recognition servers. In a recent study[29] we concluded that each server had a number of cases with a correct assignment, where the assignments of all the other servers were wrong. However, in an earlier study we were not able to combine the score from several alignment methods without compromising the specificity.[18] However, it is possible that by using another method it is possible to detect the best alignment. We recently introduced a consensus method, Pcons, that does this.[34] It was shown that Pcons performed significantly better than any individual method.

CONCLUSIONS

In this study we performed a large-scale analysis of alignment quality using both sequence methods and threading methods. In general, it is seen that if two pairs share >25% sequence identity, most alignment methods can be used to make a good model; however, if they share <15% identity, only rarely can a good model be created. In general, the difference in overall model quality between different methods is quite small. Because we chose to measure the quality of the model, we obtained better results using global alignments than local alignments. We show that predicted secondary structure information improves the alignment quality. The improvement is largest at lower sequence identities (<20%). It is of interest that even when the structure is not known, the model quality, for distantly related proteins, can be increased by the use of two predicted secondary structures for the proteins.

A final observation is that for each pair of proteins, the best alignment is obtained by different methods. The highest number of best alignments is obtained by sequence alignment methods at higher identities and by *SSPSI* at lower identities. If it were possible to choose a particular method for a particular pair, a significantly better overall result would be obtained, because for fold related the best method (*SSPSI*) only creates correct models for 6% (14%) of the pairs, whereas all methods combined create correct models for 11% (23%) of them.

MATERIALS AND METHODS

Creation of Models

From an alignment of a query protein to a target protein, a pdb file of the query protein is created, containing only the Cα atoms. The coordinates of the target Cα atoms are assigned to the Cα atoms of the query protein.

The *LGscore* Measure

Recently, a measure to calculate the significance of the similarity between two structures after a structural super

position was introduced by Levitt and Gerstein.[35] This measure is based on the following score:

$$S_{str} = M\left(\sum \frac{1}{1 + (d_{ij}/d_0)^2} - \frac{N_{gap}}{2}\right)$$

where M is equal to 20, d_{ij} is the distance between residues i and j, d_0 is equal to 5 Å and N_{gap} is the number of gaps in the alignment.

To calculate the statistical significance of this score, a set of structural alignments of unrelated proteins was used to calculate a distribution of S_{str} dependent on the alignment length, l was used.[35] From this distribution, a p value dependent on S_{str} and l was calculated.

Algorithms for Calculation of *LGscore*

It is common that only a fraction of a model is similar to the correct structure; therefore, it is necessary to detect the most significant subset of the model. It is our assumption that the most similar subset is the one with the best p value. To find the most significant fragment, we used two different algorithms, referred to as the top-down and bottom-up algorithms. The top-down algorithms works as follows:

```
while Number of aligned residues > 25
      Super position all residues in the model
        and in the correct structure.
      Calculate and store the p value for this
        super position
      Delete the pair of residues that is fur-
        thest apart from each other in the model
        and the correct structure.
      return the best p value.
```

and the bottom-up algorithm:

```
for i = 0 to length − 4
  j = 4
  add residues i to i + j to the selected set S
  while (j < length)
  Super position the selected set S of the model
    and the correct structure.
  Calculate the p value for the super posi-
    tioned residues
  if (j > 25) store the p value
  Add the pair of non aligned residues that is
    closest to each other to S.
  j++
  return the best p value
```

Most frequently, the bottom-up algorithm found the best subset, but sometimes the best subset is found by using the top-down algorithm; therefore, we used both. The bottom-up algorithm is very similar to algorithms developed earlier.[36–38] As can be seen in the algorithms above, we do not consider fragments shorter than 25 residues, because shorter fragments often were given unrealistically good p values. For comparison purposes, we used the negative logarithm of the p value. We refer to models with a *LGscore* better than 10^{-3} as correct models, as was done in LiveBench.[29] Program and a more detailed description of the *LGscore* is available at http://www.sbc.su.se/~arne/lgscore/.

The RMSD Measure

In addition to the *LGscore* measure, we chose to use a simple but easily understandable measure of model quality. RMSD is the most used measure of similarities between proteins. However, one problem with RMSD as a measure is that it depends on the size of the model. Therefore, if one alignment method produces a shorter alignment, this model will have a lower RMSD than a longer model. To avoid this problem, we defined the measure as follows. The RMSD measure is the lowest RMSD for any fragment that contains at least 50% of the residues in the protein. This measure was calculated by using the same algorithms as for the *LGscore*.

Benchmark Database

To create a useful benchmark, it is important to use a large and broad set of related and unrelated protein domains of high quality. To achieve this, we decided to use a set of protein that, according to Scop, shared the same fold.[39] We started from the 12,805 domains in Scop version 1.39. From these domains a subset of proteins representing all families and that all have <50% sequence identity to any other member of the set was selected by using the Hobohm algorithm for homology reduction,[40] leaving us with 1245 domains; 187 of these were used for manual optimizing gap penalties and other parameters, and the remaining 1045 were used for benchmarking. The proteins in the first (training) set was selected to contain all proteins from a number of folds, to ensure that is was completely independent of the second (test) set. The data sets are available from http://www.sbc.su.se/~arne/alignments/

From the list of domains in the benchmark database, we extracted all pairs of proteins that belong to the same fold, because there was no reason to try to align proteins that did not share the same fold. In the test set there were 9983 pairs of protein domains that shared the same fold. Of these, 4979 only shared the same fold, 3101 belonged to the same superfamily but different families, and 1903 belonged to the same family. Because alignments are nonsymmetrical, when a model was built, we actually had twice as large a test set containing 19966 pairs.

Alignments Methods

Most alignments in this study were performed by using standard sequence[2,3] or profile alignment techniques.[5] The computer program, palign, used for this is freely available at http://www.sbc.su.se/~arne/pscan/palign.tar.gz.

The first section of this article analyzes standard local and global alignments by using four different substitution tables: PAM-250, BLOSUM-45, BLOSUM-50, and BLOSUM-62. For each substitution table we used four different gap-opening penalties (5, 10, 15, and 20) and four different gap extension penalties (0, 1, 3, and 5), that is, altogether 16 different penalties were tried.

Second, we evaluated several other alignment methods, briefly described in Table II and below. Because of computational limitations it was not possible to exhaustively optimize all parameters. Therefore, for these methods, gap and other parameters were optimized by using the smaller training set.

We tried five methods that use multiple sequence alignment information: profile alignments, CLUSTALW, two hidden Markov models (HMMER[41] and SAM[30]), and a method that uses predicted secondary structures *SSPSI* described below.

All these methods were based on multiple sequence information obtained from a PSI-BLAST[6] search, of a sequence against KIND[42] allowing a maximum number of three iterations using default parameters. This means that the same sequences were used in all these studies, that is, we did not test the performance of the different methods to detect related sequences. The main reason for this was that it was computationally too demanding to do this by using any other methods than PSI-BLAST, if such a large database as KIND is used. Furthermore, exactly the same multiple sequence alignments were used in the profile searches and in the hidden Markov models.

In the profile method *PSI,* a sequence was aligned against the final PSI-BLAST profile by using global alignments.[5] This is identical to how PSI-BLAST was used in the PDBBLAST methods.[29] The multiple sequence alignments from the last iteration of PSI-BLAST were used when building the HMMs. Default parameters were used. To align two proteins with CLUSTALW, all the proteins that were found with PSI-BLAST for any of the two proteins were used. CLUSTALW was then run by using default parameters, and the alignment of the two test proteins was reported.

We also used two methods that use predicted secondary structures. This alignment algorithm is identical to the one used in earlier studies.[7,8] In these methods, a score (S) is added to the sequence alignment score if the predicted secondary structure matches the secondary structure of an aligned residue. If these residues do not match a score (S') is subtracted:

$$SCORE = \sum(S[i, j] + f(ss_i, ss_j))$$

where $S[i, j]$ is the standard alignment score for aligning residues i and j, and $f(ss_i, ss_j)$ is a score dependent on the (predicted or real) secondary structures of residues i and j. In this, and earlier, work $f(ss_i, ss_j)$ is:

$$f(ss_i, ss_j) = S \text{ if } ss_i = ss_j$$

$$f(ss_i, ss_j) = S' \text{ if } ss_i \neq ss_j$$

In *SS* a single sequence was used, whereas in *SSPSI* the prediction score was added to the same PSI-BLAST profile used in *PSI.* The profile-based algorithm is very similar to the method used in 3D-PSSM.[43] After the parameters on the training set were optimized, it was found that the best performance was obtained by using $S = 5$ and $S' = -0.5$. We also made alignment by using THREADER[10] with default parameters.

Finally, we compared the alignment methods with alignments obtained from a structural alignment program *LGscore2.* The program *LGscore2* uses an algorithm similar to the structural alignment algorithm used by Levitt and Gerstein,[35] and it is freely available from http://www.sbc.su.se/~arne/lgscore/. It should be noted that this structural alignment program probably does not perform as well as the state-of-the-art programs.

ACKNOWLEDGMENT

We thank Susana Cristobal, Daniel Fisher, Leszek Rychlewski, Bob MacCallum, and David Liberles for valuable discussions and help. We also thank one reviewer for the suggestion to additionally use the RMSD measure.

REFERENCES

1. CASP. The casp www-site. http://predictioncenter.llnl.gov/casp3/Casp3.html, 1999.
2. Smith T, Waterman M. Identification of common molecular subsequences. J Mol Biol 1981;147:195–197.
3. Needleman S, Wunsch C. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 1970;48:443–453.
4. Thompson JD, Higgins D, Gibson T. Clustal w: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. Nucleic Acids Res 1994;22:4673–4680.
5. Gribskov M, McLachlan AD, Eisenberg D. Profile analysis: detection of distantly related proteins. Proc Natl Acad Sci USA 1987;84:4355–4358.
6. Altschul S, Madden T, Schaffer A, Zhang J, Zhang Z, Miller W, Lipman D. Gapped blast and psi-blast: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389–3402.
7. Fischer D, Eisenberg D. Protein fold recognition using sequence-derived predictions. Protein Sci 1996;5:947–955.
8. Rost B, Schneider R, Sander C. Protein fold recognition by prediction-based threading. J Mol Biol 1997;270:471–480.
9. Rice D, Eisenberg D. A 3D-1D substitution matrix for protein fold recognition that includes predicted secondary structure of the sequence. J Mol Biol 1997;267:1026–1038.
10. Jones D, Taylor W, Thornton J. A new approach to protein fold recognition. Nature 1992;358:86–89.
11. Sanchez R, Sali A. Large-scale protein structure modeling of the saccharomyces cerevisiae genome. Proc Natl Acad Sci USA 1998;95:13597–13602.
12. Alexandrov N, Luethy R. Alignment algorithm for homolgy modeling and threading. Protein Sci 1998;7:254–258.
13. Abagyan RA, Batalov S. Do aligned sequences share the same fold? J Mol Biol 1997;273:355–368.
14. Brenner SE, Chothia C, Hubbard T. Assessing sequence comparison methods with reliable structurally identified evolutionary relationships. Proc Natl Acad Sci USA 1998;95:6073–6078.
15. Park J, Teichmann SA, Hubbard T, Chothia C. Intermediate sequences increase the detection of homology between sequences. J Mol Biol 1997;273:249–254.
16. Park J, Karplus K, Barrett C, Hughey R, Haussler D, Hubbard T, Chothia C. Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. J Mol Biol 1998;284:1201–1210.
17. Di Francesco V, Geetha V, Garnier J, Munson PJ. Fold recognition using predicted secondary structure sequences and hidden Markov models of proteins folds. Proteins 1997;Suppl 1:123–128.
18. Lindahl E, Elofsson A. Identification of related proteins on family, superfamily and fold level. J Mol Biol 2000;295:613–625.
19. Rost B. Better 1d predictions by 'experts' with machines. Proteins 1997;Suppl 2:192–197.
20. Domingues F, Lackner P, Andreeva A, Sippl M. Structure based evaluation of sequence comparison and fold recognition alignment accuracy. J Mol Biol 2000;297:1003–1013.
21. Panchenko A, Marchler-Bauer A, Bryant SH. Combination of threading potentials and sequence profiles improves fold recognition. J Mol Biol 2000;296:1319–1331.

22. Jaroszewski L, Rychlewski L, Godzik A. Improving the quality of twilight-zone alignments. Protein Sci 2000;9:1487–1496.
23. Sauder J, Arthur J, Dunbrack RL, Dunbrock RL Jr. Large-scale comparison of protein sequence alignment algorithms with structure alignments. Proteins 2000;40:6–22.
24. Cristobal S, Zemla A, Fischer D, Rychlewski L, Elofsson A. A study if quality measured for protein threading models. BMC Bioinformatics 2001;2:5.
25. Thompson J, Plewniak F, Poch O. A comprehensive comparison of multiple sequence alignment programs. Nucleic Acids Res 1999;27: 2682–2690.
26. Feng Z-K, Sippl M. Optimum superimposition of protein structures, ambiguities and implications. Fold Des 1996;1:123–132.
27. Godzik A. The structural alignment between two proteins. Is there a unique answer? Protein Sci 1996;5:1325–1338.
28. Fischer D, Elofsson A, Rychlewski L, Pazos F, Valencia A, Rost B, Ortiz A, Dunbrack R. Cafasp2: the critical assessment of fully automated structure prediction methods. Submitted for publication.
29. Bujnicki J, Elofsson A, Fischer D, Rychlewski L. Livebench: continuous benchmarking of protein structure prediction servers. Protein Sci 2001;10:352–361.
30. Karplus K, Barrett C, Hughey R. Hidden markov models for detecting remote protein homologies. Bioinformatics 1998;14:846–856.
31. Eddy S. Hmmer-hidden Markov model software url: http://genome.wustl.edu/eddy/hmmer.html, 1997.
32. Rychlewski L, Jaroszewski L, Li W, Godzik A. Comparison of sequence profiles. Strategies for structural predictions using sequence information. Protein Sci 2000;9:232–241.
33. Müller A, MacCallum R, Sternberg MJE. Benchmarking psi-blast in genome annotation. J Mol Biol 1999;293:1257–1271.
34. Lundström J, Rychlewski L, Bujnicki J, Elofsson A. Pcons: a neural network based consensus predictor that improves fold recognition. Protein Sci Forthcoming.
35. Levitt M, Gerstein M. A unified statistical framework for sequence comparison and structure comparison. Proc Natl Acad Sci USA 1998;95:5913–5920.
36. Zemla A, Veclovas C, Moult J, Fidelis K. Processing and analysis of casp3 protein structure predictions. Proteins 1999;3:22–29.
37. Hubbard T. Rmsd/coverage graphs: a qualitative method for comparing three-dimensional protein structure predictions. Proteins 1999;Suppl 3:15–21.
38. Siew N, Elofsson A, Rychlewski L, Fischer D. Maxsub: an automated measure to assess the quality of protein structure predictions. Bioinformatics 2000;16:776–785.
39. Murzin A, Brenner S, Hubbard T, Chothia C. Scop: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 1995;247:536–540.
40. Hobohm U, Scharf M, Schneider R, Sander C. Selection of representative protein data sets. Protein Sci 1992;1:409–417.
41. Eddy S. Profile hidden markov models. Bioinformatics 1998;14: 755–763.
42. Kallberg Y, Persson B. Kind-a non-redundant protein database. Bioinformatics 1999;15:260–261.
43. Kelley L, MacCallum R, Sternberg M. Enhanced genome annotation using structural profiles in the program 3d-pssm. J Mol Biol 2000;299:523–544.