Improving the quality of twilight-zone alignments

LUKASZ JAROSZEWSKI,¹ LESZEK RYCHLEWSKI,² AND ADAM GODZIK¹

¹The Burnham Institute, 10901 North Torrey Pines Road, La Jolla, California 92037

²International Institute of Molecular and Cell Biology, Trojdena 4, Warszawa 02-109, Poland

(RECEIVED January 31, 2000; FINAL REVISION May 25, 2000; ACCEPTED June 8, 2000)

Abstract

Several recent publications illustrated advantages of using sequence profiles in recognizing distant homologies between proteins. At the same time, the practical usefulness of distant homology recognition depends not only on the sensitivity of the algorithm, but also on the quality of the alignment between a prediction target and the template from the database of known proteins. Here, we study this question for several supersensitive protein algorithms that were previously compared in their recognition sensitivity (Rychlewski et al., 2000). A database of protein pairs with similar structures, but low sequence similarity is used to rate the alignments obtained with several different methods, which included sequence–sequence, sequence–profile, and profile–profile alignment methods. We show that incorporation of evolutionary information encoded in sequence profiles into alignment calculation methods significantly increases the alignment accuracy, bringing them closer to the alignments obtained from structure comparison.

In general, alignment quality is correlated with recognition and alignment score significance. For every alignment method, alignments with statistically significant scores correlate with both correct structural templates and good quality alignments. At the same time, average alignment lengths differ in various methods, making the comparison between them difficult. For instance, the alignments obtained by FFAS, the profile–profile alignment algorithm developed in our group are always longer than the alignments obtained with the PSI-BLAST algorithms. To address this problem, we develop methods to truncate or extend alignments to cover a specified percentage of protein lengths. In most cases, the elongation of the alignment by profile–profile methods is reasonable, adding fragments of similar structure. The examples of erroneous alignment are examined and it is shown that they can be identified based on the model quality.

Keywords: alignment accuracy; FFAS; homology modeling; PSI-Blast; sequence profiles

The analysis of known protein structures enables us to formulate many empirical rules of protein folding that can be used in protein structure prediction. The first and most important rule is that proteins with similar sequences fold to similar structures. The whole field of homology modeling is based on this observation. In a typical modeling procedure, a known structure is used as a template and the new structure is built around it, introducing minimal changes so that the new sequence can be accommodated without violating excluded volume and other constraints.

Many successful applications of homology modeling have been reported in the literature (Svensson et al., 1990; Sanchez & Sali,

1998) and several recent reviews are available (Rost & Sander, 1996; Burley et al., 1999; Moult, 1999). To start the modeling process, we have to identify the template and define an alignment, i.e., a set of residue-by-residue equivalencies between the template and target sequences. This is the single most crucial step in a modeling process. Any errors at this stage are usually impossible to correct later and lead to significant errors in the models (Sali et al., 1995).

If the target and the template proteins are closely related, there is no problem both with the template identification and creation of the alignment. Well-developed tools of protein sequence analysis can be used for these purposes. However, the problem gets increasingly difficult, as the sequence homology between the target and possible template protein becomes more distant and sequence similarity gets weaker. Sequence alignments become unstable with sequence similarity lower than 40% of identical residues and their quality suddenly drops as the sequence similarity enters the “twilight zone” of below 25% of identical residues (Vogt et al., 1995). This is the most important reason why the quality of models gets worse with decreasing sequence similarity. In general, we can expect that the errors in the alignments may include both shifts of the two sequences in respect to each other, as well as extending the alignment beyond (or truncating it short of) the region of actual

Reprint requests to: Adam Godzik, The Burnham Institute, 10901 North Torrey Pines Road, La Jolla, California 92037; e-mail: adam@burnham-inst.org.

Abbreviations: COV(%), “coverage,” e.g., percent of the target sequence that is “covered” with the alignment; GAPS, number of gaps in the alignment; SF0 (%), percent of the structural alignment correctly reproduced by a given alignment; RMSD, root-mean-square deviation of the target and the template calculated according to the alignment; CMO, contact map overlap of the target and the template calculated according to the alignment; HSP, high scoring segment pair, a continuous fragment of the alignment identified by the BLAST algorithm.

similarity between the two proteins. There are other problems facing comparative modeling, such as structural differences between homologous proteins that accumulate with sequence divergence or modeling parts of the structure that do not have corresponding elements in the template structure. However, these latter problems are outside the scope of this paper.

With the flood of sequences available from genomic projects, the importance of the sensitive methods of recognizing distant homologies increased. Such methods remain the main source of annotations for newly sequenced genes. Several new methods, such as Hidden Markov Models (Eddy, 1998; Hargbo & Elofsson, 1999; Karplus et al., 1999; Bateman et al., 2000) or various flavors of threading (Domingues et al., 1999; Jones, 1999; Panchenko et al., 1999) were introduced in recent years, greatly increasing the range of applications of distant homology recognition. One of the most popular of the recent programs is PSI-BLAST, the position specific iterative BLAST (Altschul et al., 1997). Its application led to the discovery of many hitherto unsuspected relations between proteins, adding to our understanding of many important biological processes (Aravind & Koonin, 1999). Our group has been developing both threading (Jaroszewski et al., 1998) and sequence based fold recognition algorithms (Rychlewski et al., 2000), applying both to large scale, automated genome fold and function predictions (Rychlewski et al., 1999).

There is no doubt that the new algorithms greatly increased the range of application of distant homology recognition into what is often called “the twilight zone” of previously undetectable sequence similarities (Murzin, 1999). However, the question of the quality of alignment obtained with the new methods was rarely addressed. In fact, a lot of anecdotal evidence suggested that the new, supersensitive algorithms often give distorted view of the similarity between the prediction target and its better characterized homologue. At the same time, successful application of protein homology modeling critically depends on the quality of alignments between protein structures used as templates and proteins for which structure is being modeled. Therefore, in this contribution, we study the question of alignment accuracy for two of the distant homology recognition programs, the popular PSI-BLAST algorithm, as implemented at the NCBI-NIH WEB server and the FFAS algorithm, developed in our group.

In a recent paper, we have demonstrated that the FFAS profile-profile matching algorithm, if used as a fold recognition tool, is more sensitive than the popular sequence-profile matching PSI-Blast algorithm (Rychlewski et al., 2000). In this study, we compared the accuracy of FFAS alignments and the alignments calculated with PSI-BLAST. For comparison, several other mutation matrix based alignment algorithms in several versions (sequence-sequence, profile-sequence or profile-profile) were also included in the analysis.

To objectively measure the alignment accuracy, one must exclude the effect of the alignment length. The most commonly used measure of structure similarity, the RMSD of equivalent atoms upon optimal alignment, strongly depends on the alignment length, with shorter alignments having smaller RMSDs (Hubbard, 1999; Jones & Kleywegt, 1999). Other measures may have opposite behavior, for instance the contact map overlap (CMO) values are usually higher (better) for longer alignments. Since different methods systematically give alignments of various lengths to allow the unbiased evaluation of the differences of the alignment accuracy between various methods, we modified various alignment algorithms to yield exactly the same average alignment length. Ideally,

for each pair of proteins alignments produced by different methods should be of the same length. However, even small changes of parameters (gap penalties and mutation matrix zero level) cause rapid, discontinuous changes of the alignment length. Therefore, it is impossible to “tune up” the length of the individual alignment. However, it is possible to “tune up” the average length for the set of the alignments. For instance, the previously developed FFAS algorithm was modified to give the alignments of the same average length as PSI-BLAST (this version was called FFAS-R).

We also attempted to separate the effect of using various alignment algorithms (dynamic programming vs. HSP extension) from the effect of using various types of sequence information (sequence vs. profile). Thus, in addition to previously described methods (PSI-BLAST and FFAS), we used the dynamic programming alignment algorithm in three versions where the profiles or sequences are used in various combinations.

The average values of alignment accuracy calculated over large benchmarks show general trends and are very useful for comparing different alignment protocols. However, as can be seen in Figures 1 and 2, the accuracy of specific alignments produced by all methods is widely distributed ranging from satisfactory to completely wrong. Therefore, recognizing correct and incorrect alignments from various methods is necessary to develop alignments (and subsequently models) of uniformly good quality. This problem is also addressed here.

In the following section, we present the results of the evaluation of the alignments calculated with FFAS, PSI-BLAST, and simple mutation matrix based method and address the question of how to recognize accurate alignments or parts thereof. Next, we present examples of the models based on FFAS alignments and compare them with the models obtained using PSI-BLAST alignments. In Materials and methods, we describe the alignment evaluation benchmark, different measures of alignment accuracy, and alignment protocols evaluated in this contribution. In the Discussion, we describe the most important factors influencing the alignment accuracy and discuss possible improvements of alignment calculation protocols.

Fig. 1. The comparison of distributions of RMSD values calculated for FFAS-R, PSI-BLAST, and FFAS alignments.

Fig. 2. The comparison of distributions of contact map overlap values calculated for FFAS-R, PSI-BLAST, and FFAS.

Results

The overview of FFAS, PSI-BLAST, and mutation matrix alignments

The average values of accuracy measures of the alignments calculated with different methods are presented in Table 1. The parameters of structural alignments calculated with CE algorithm (Shindyalov & Bourne, 1998) are also shown in Table 1 as the reference. As can be seen, there are significant differences in the average length of the alignments calculated with different methods. These differences make the exact and objective comparison very difficult, because, as we mentioned earlier, the measures of alignment accuracy strongly depend on the alignment length. However, some general observations can be made.

In the low sequence similarity region that we explored, sequence and profile alignments are significantly less accurate than structural alignments. Structural alignments obtained with the CE algorithm cover on the average 66% of the target sequence. Despite that their average RMSD is lower than all types of sequence based alignments, including alignments that are shorter, such as PSI-BLAST alignments. At the same time, CE structural alignments surpass sequence alignments (even longer ones) also in terms of CMO, which is usually not true for alignments that are entirely

based on RMSD optimization (Godzik, 1996). This suggests that the CE algorithm is a good compromise between global and local similarity, by correctly reproducing interresidue contacts and at the same time keeping reasonable RMSD value.

Two problematic features of CE structural alignments are the large average number of gaps and the relative small part of proteins that is covered with the alignment. The first feature suggests that the alignments are built from rather short fragments. In sequence based alignments, it is possible to lower gap penalties to allow more gaps, but this causes significant drop of accuracy as indicated with average RMSD and CMO values (data not shown). Apparently, the sequence or profile similarity scoring require longer fragments to correctly recognize their similarity. If we treat structural alignments as a “standard of truth,” it may suggest that with increasing divergence the alignment becomes very fragmented and this causes problems for dynamic programming based sequence alignments with linear gap penalties. The second problem is that on average, only 66% of the shorter protein in the pair being compared is covered by the alignment. The quality of the alignment is a function of its length and the decision on where to stop the alignment is quite arbitrary. However, the visual inspection of most of the aligned pairs suggests that the alignment could be extended without significant loss in accuracy.

Standard local sequence–sequence alignments calculated using *gonnet_p* mutation matrix with optimal parameters suggested in Vogt et al. (1995) (see Materials and methods) result in the longest alignments from all methods evaluated in this analysis. On average, they cover 90% of the target. Their average SF0 and CMO values are lower than average SF0 and CMO of FFAS alignments, despite smaller average length of the latter.

FFAS alignments are longer than CE (structural) and PSI-BLAST alignments, covering on the average 84% of the target. They have higher average SF0 and CMO values, but their average RMSD value is worse than PSI-BLAST alignments. Since such differences could be partly explained by their length differences, we counted the target–template pairs where alignments prepared with one method surpass alignments prepared with another method *both* in terms of RMSD and CMO. The FFAS alignments are univocally better in 56 of 190 benchmark pairs, while the reverse is true for 17 pairs. For the remaining targets PSI-BLAST alignment has better RMSD and FFAS alignment has better CMO, which might be connected to the significant length difference.

To allow the unbiased comparison of both methods, a variant of the FFAS method (FFAS-R) was developed to yield the same average alignment length as PSI-BLAST. As discussed in Materials and methods, this can be achieved by changing the average

Table 1. The comparison of the average parameters of the alignments evaluated in this paper

Alignment method	⟨COV(%)⟩	⟨GAPS⟩	⟨SF0(%)⟩	⟨RMSD⟩	⟨CMO(%)⟩
CE (structural)	66	8.5	100	3.1	42
Sequence (<i>gonnet_p</i>)	90	13.0	42	12.5	38
PSI-BLAST	54	4.7	45	5.9	32
FFAS-R	54	4.1	48	4.4	34
FFAS	84	7.5	56	9.3	43
PSI-BLAST $\cap$ FFAS	58	4.4	59	3.8	38

value of the mutation matrix. As shown in Table 1, this algorithm surpasses PSI-BLAST in all measures of alignment quality (SF0, RMSD, and CMO). There are several possible reasons of such a difference and this is further examined in the next paragraph and analyzed in Discussion.

The importance of the evolutionary data for alignment calculation methods

In the above subsection, we compared some standard and easily available alignment methods and the profile–profile alignment method FFAS. The latter was shown to be the most accurate alignment method from the group compared here. However, the evaluated methods differ in many aspects, including the details of scoring and alignment building. For example, FFAS and the sequence–sequence method directly use dynamic programming for alignment calculation and PSI-BLAST alignment is based on the search for high scoring segment pairs that are later connected (Altschul et al., 1997). It is not clear what features are actually responsible for the improvement in alignment quality. To directly assess the importance of evolutionary information for alignment accuracy, we introduce an “intermediate” algorithm that uses the same alignment algorithm (local–local dynamic programming) as the FFAS method, but uses sequence–profile scoring, such as PSI-BLAST. The sequence–sequence alignment method was modified to keep the average alignment length to be similar to that of PSI-BLAST and FFAS-R. As seen in Table 2, the incorporation of the evolutionary information in the alignment calculation leads to the gradual increase of the alignment accuracy. Profile–profile alignments are more accurate than profile–sequence alignments, which are better than sequence–sequence alignments. As expected, the profile–sequence alignment method yields slightly more accurate alignments than PSI-BLAST (compare Tables 1 and 3), which indicates that direct use of dynamic programming is more accurate than the PSI-BLAST heuristic based on HSPs (see Discussion for more comments).

How to distinguish the correct and incorrect alignments

The alignments having statistically significant scores were evaluated separately (see Table 3) from alignments with scores falling into the distribution of random scores. It is clear that alignment score significance allows quite reliable assessment of alignment accuracy without knowing the target structure. The average values of alignment accuracy for the alignments having significant scores

Table 2. *The comparison of sequence–sequence, profile–sequence and profile–profile alignments calculated with S/W algorithm as implemented in FFAS^a*

Alignment method	⟨SF0(%)⟩	⟨RMSD⟩	⟨CMO(%)⟩
Sequence–sequence	39	7.8	28
Profile–sequence	47	5.3	32
Profile–profile (FFAS-R)	48	4.4	34

^aThe parameters of sequence–sequence and profile–sequence matching algorithm were tuned up to yield the same average alignment length as PSI-BLAST and FFAS-R methods.

Table 3. *The correlation between alignment accuracy and FFAS alignment score significance^a*

Alignments	⟨SF0(%)⟩	⟨RMSD⟩	⟨CMO(%)⟩
All (190)	56	9.3	43
Significant score (138)	70	7.2	51
Insignificant score (52)	17	14.7	21

^aThe average values of alignment accuracy measures are given for all the alignments, the alignments with statistically significant score and the alignments with insignificant score.

are much better than the average values (see Table 1). However, does that help when one asks the question, “How confident we may be about the quality of a given alignment?”

To address this question, we divide the alignments into two broad classes: (1) The alignments with CMO > 20% (labeled as accurate); (2) the alignments with CMO < 20% (labeled as inaccurate).

The pairs for which the structural alignment has a CMO value of <20% were excluded from this classification as their structural similarity may be undetectable by the CMO value. We decided to use the CMO value as the criterion of the alignment correctness because as we mentioned earlier, large RMSD value is not always an indicator of erroneous alignment. The threshold of 20% of correctly reproduced contacts is two times the values obtained when aligning randomly selected protein structures with no structural similarity (data not shown). The alignment that reproduces more than 20% contacts always indicates significant structural similarity.

The distribution of statistically significant and random alignments into these two groups is shown in Table 4. As one can see, the statistical significance of the alignment score is a good indicator of alignment accuracy. The FFAS alignment with a statistically significant score has about a 2% chance to be incorrect. Despite quite coarse classification of the alignment accuracy (e.g., two distinct classes) and simple significance criterion, the results show that we are able to differentiate between correct and incorrect alignments, at least in the statistical sense.

Another possible alignment reliability criterion is its stability between various methods. We attempted to apply this criterion by identifying identical part in FFAS and PSI-BLAST alignments as described in Materials and methods. The results of the evaluation

Table 4. *The distribution of the FFAS alignments into two coarse classes with CMO > 20% and CMO < 20% (labeled as accurate and inaccurate, respectively)^a*

Alignments	Accurate	Inaccurate
All (144)	89	11
Significant score (108)	98	2
Insignificant score (36)	58	42

^aThe percent of accurate and inaccurate alignments is given for all the alignments, the alignments with statistically significant score and the alignments with insignificant score.

of these “common parts” are shown in Table 1 labeled “PSI-BLAST $\cap$ FFAS.” The average values of the accuracy measures indicate that the parts of the alignments identical in PSI-BLAST and FFAS alignments are much more accurate than either PSI-BLAST or FFAS alignments. We note that for 51 target–template pairs the common part of the alignments is empty—these tend to be the worst quality alignments. These pairs are naturally not included in the calculation of average accuracy measures and for these two reasons the average coverage of the PSI-BLAST $\cap$ FFAS is better than the coverage of the PSI-BLAST alignments themselves.

The examples of the models based on the alignments

The most important application of the alignments is in homology modeling. The alignment calculation is a crucial step in the modeling process, any errors at this stage are usually impossible to correct and lead to errors in the models (Sali et al., 1995). On the other hand, it was shown that the errors in the alignment can be detected by the evaluation of the resulting model (Pawlowski et al., 1997). The typical examples of the models based on the PSI-BLAST and FFAS alignments are shown in Figures 3A and 3B. Target structures are also shown for comparison. The models 3A and 3B represent typical cases, when PSI-BLAST alignment is obviously too short and FFAS alignments reflect the real target–template similarity and allow construction of the complete model. The model 3C is more interesting. It was deliberately selected in such a way that both FFAS and PSI-BLAST alignments are of identical lengths and PSI-BLAST alignment is more accurate than FFAS alignment. More precisely, there are some shifts in the central part of FFAS alignment. PSI-BLAST and FFAS alignment have RMSDs of 2.0 and 3.3 Å, respectively, and CMOs of 78 and 66%, respectively. The resulting models are similar but, as expected, the central part of the model based on the FFAS alignment is partly distorted. Both models were evaluated with the structure evaluation option of MatchMaker threading program (Godzik et al., 1992) included in Sybyl package (1998). The average energy per residue of the model based on PSI-BLAST alignment is equal to -0.24 kT, which is two times lower than -0.12 kT obtained for the model based on the FFAS alignment. MatchMaker residue-by-residue energy profile of the model based on FFAS alignment contains the region of positive values matching the erroneous portion of the alignment. The same models were also evaluated with Verify3D Structure Evaluation Server (Eisenberg et al., 1997) available at UCLA-DOE Laboratory of Structural Biology and Molecular Medicine. These results also clearly indicated better quality of the model based on PSI-BLAST alignment. This example shows that the better model can be identified by the energy criterion. Extensive analysis of various energy based criteria to recognize correct alignments will be a subject of a separate publication.

Discussion

It is known that the incorporation of evolutionary information encoded in sequence profiles improves the sensitivity of detection of distant homologies (Park et al., 1998; Karplus et al., 1999; Rychlewski et al., 2000). We have shown that the same is true for alignment accuracy. From a practical viewpoint, this means that the FFAS profile–profile matching algorithm developed in our group surpasses the popular PSI-BLAST algorithm in unbiased

comparison of alignment accuracy. This is what we expected since FFAS is a natural extension of PSI-BLAST method and it should be more sensitive and accurate. The PSI-BLAST algorithm was not optimized for alignment accuracy. The authors of this most popular and most effective sequence similarity search method had to sacrifice the alignment accuracy for speed to some extent, which is the fundamental parameter when performing searches in large sequence databases. The PSI-BLAST alignments are calculated with a heuristic algorithm that first looks for high scoring segment pairs, which are then extended and connected (Altschul et al., 1990, 1997). This heuristics is very efficient but does not guarantee that the alignment is optimal and may not detect regions with weak similarities that are below thresholds for HSPs recognition (Altschul et al., 1997). On the contrary, FFAS uses Smith–Waterman dynamic programming routine, which gives the optimal alignment for given scoring matrix and gap penalties.

We have shown that the greater length of FFAS alignments as compared to the PSI-BLAST alignments is in most cases reasonable. FFAS-R is the only method that surpasses PSI-BLAST in *all* alignment accuracy criteria but it is hard to argue that it is better than FFAS. The alignments produced with FFAS method surpass PSI-BLAST and FFAS-R alignments in CMO and have better agreement with structural alignment. We have shown that higher RMSD value does not necessarily mean that the alignment does not reflect structural similarity (see Fig. 4). Higher RMSD values of FFAS alignments are the mere effect of their greater length.

To answer the general question about the importance of evolutionary information for alignment accuracy, we constructed three alignment methods yielding the alignments of the same average length and differing only in the level of evolutionary information used for the calculation of the alignment. The results of this test show that the profile–profile alignments are more accurate than profile–sequence alignments, which in turn surpass sequence–sequence alignments. The alignment accuracies of protein pairs with low sequence similarity are widely distributed (see Figs. 1, 2), which indicates the importance of methods allowing differentiation between accurate and erroneous alignment without referring to the target structure. Two such methods were described here.

The alignment score is the most obvious measure of the similarity of two sequences. Its significance can be assessed when compared with the distribution of the alignment scores in the database. This corresponds to the fold recognition task. It is natural to expect that the target sequence that can be recognized as being similar to the template is more likely to be accurately aligned than the sequence that is not recognized in the fold recognition test. Our tests showed that this is exactly the case—the alignments with statistically significant scores are much more accurate than the average. However, this could not be used as the only criterion for alignment accuracy. Occasionally, the high significance alignments can be wrong (or other methods may result in better alignments). Also, for some reason we may want to align protein pairs that do not result in significant score, for instance where there is a strong biochemical evidence that the two proteins are homologous. In such cases, we may need an independent evaluation of alignment accuracy.

The second method connects the alignment accuracy with its stability. The connection between alignment accuracy and stability was widely explored. In the most elegant approach, the alignment stability is assessed by calculation of suboptimal alignments that enables calculation of the reliability index for each pair of aligned

Fig. 3. The models based on the PSI-BLAST alignments (on the left) vs. models based on FFAS (in the middle) alignments (target–template pairs; (A) 3sdhA – 1flp_, (B) 2fx2_ – 1qrdA, (C) 1sriA – 1aviA). The correct structures (the targets) are shown on the right. 3sdh is hemoglobin I from *Scapharca inaequivalvis*, 1flp is hemoglobin I form *Lucina pectinata*, 2fx2 is flavodoxin from *Desulfovibrio vulgaris*, 1qrdA is quinone-reductase chain A from rat, 1sriA is streptavidin chain A from *Streptomyces avidinii*, 2aviA is avidin from hen.

residues. This method was applied for dynamic programming alignment protocol (Zuker, 1991; Mevissen & Vingron, 1996) and for hidden Markov model algorithm (Yu & Smith, 1999). We tested a simplified version of such an approach. We assessed the alignment simply by comparing the alignments calculated with two methods. The alignments calculated with FFAS were compared position-by-position with the alignments obtained from PSI-BLAST. The parts where both alignments are identical were evaluated. We note that there are as many as 51 pairs where the common part of the

PSI-BLAST and FFAS alignments was empty. The “common parts” of the alignments are much more accurate than PSI-BLAST or FFAS alignments. The simple test when one calculates the alignment with two methods and compares them allows the detection of most accurate alignment segments. This simple test proves that the connection between alignment stability and accuracy is valid for profile alignments. We also plan the application of more precise alignment local reliability calculation methods for profile–profile alignments.

Fig. 4. The example of target–template pair, where alignment RMSD is very poor and CMO is acceptable. Glutathione biosynthetic ligase from *E. coli* (2glt) and succinyl-CoA synthetase chain B from *E. coli* (1scuB) both consist of three α/β domains and show significant topological similarity, but the different domain arrangement makes global alignment with low RMSD impossible.

Finally, it is possible to perform an extensive post-alignment analysis of models built on different alignments. In principle, this approach may provide more accurate models than relying on any single alignment method. Indeed, this approach was shown to be successful in the case of simple sequence–sequence alignments using different mutation matrices (Pawlowski et al., 1997), and in this work it was illustrated on one example for various profile–profile alignment methods. A large comprehensive test of such an approach for profile–profile alignments will be a subject of a separate paper.

Materials and methods

The alignment evaluation benchmark

The alignment evaluation benchmark consists of pairs of proteins with similar structures and relatively low sequence similarity. Such pairs are used in the experiment: how accurately could we predict the structure of the first protein using the second protein as the template. Such benchmarks can be used to evaluate various alignment building protocols and homology modeling algorithms.

There are many methods of detecting similarities of protein structures. The methods like CE (Shindyalov & Bourne, 1998), VAST (Gibrat et al., 1996), or DALI/FSSP (Holm & Sander, 1998) are automated. The protein universe can also be organized in the “family tree” based on the hierarchy of functional and structural similarities. Such hierarchical databases, such as SCOP (Murzin et al., 1995) or CATH (Orengo et al., 1997), are usually constructed by human inspection of protein structures and features.

When the sequence similarity between two proteins is higher than 40%, the alignment calculated with sequence–sequence matching protocol is usually stable and reliable. We are particularly interested in “problematic” cases, hence the benchmark of protein pairs was constructed of protein pairs with less than 40% sequence identity.

In the alignment evaluation benchmark, we used a list of protein pairs selected from the SCOP (Murzin et al., 1995) database clustered at 40% of sequence identity PDB40D-J (Park et al., 1998). The benchmark pairs represent different levels of similarity (fold, superfamily, family). In the SCOP classification, the proteins are divided into domains. We selected target–template pairs sharing a similar domain as identified by SCOP but we used the entire sequences as targets and templates. This way, our benchmark is closer to the real life situation where domain boundaries are often not known. A sensitive alignment algorithm is expected to recognize and align only the similar domains. To avoid biasing our results by more popular folds, we selected only one pair to represent each fold type at a given level of similarity. The list of 190 benchmark pairs is available from our WWW server bioinformatics.burnham-inst.org/benchmarks.

The measures of alignment accuracy

For proteins with known structure, we can evaluate the sequence alignment by testing how well it describes the similarity between structures of both proteins. Because this comparison is free of any assumptions concerning the evolutionary process, structure comparisons are often used as the ultimate “standard of truth” in the evaluation of sequence alignments. For instance, the quality of sequence alignments in Vogt et al. (1995) was calculated by comparing the number of identical pairs in both sequence and structural alignments.

There is, however, a serious problem in using structural alignments as the “standard of truth.” There are many measures of protein structure similarity, which in general would produce different structural alignments (Godzik et al., 1993; Godzik, 1996). For weakly similar proteins, this difference becomes so big that various structural alignments can differ at all positions, and several different alignments can have a comparable similarity score (Godzik et al., 1993; Godzik, 1996). Therefore, to evaluate an alignment, in

this work, we calculated the structural similarity of the target and template as seen by the alignment to be evaluated. A similar approach was used to evaluate structure prediction results in the CASP meeting (Hubbard, 1999; Jones & Kleywegt, 1999). Other automated algorithms for alignment evaluation and comparison are being developed (Fischer, 1999), but at this point in time they are not widely available.

Two particular measures of protein structural similarity used here are CMO and RMSD after optimal superposition (Kabsch, 1978). CMO, introduced previously (Godzik et al., 1993; Godzik, 1996), has several advantages over other measures, and in particular, it is insensitive to extending the alignment to include less similar regions. The most popular structure similarity measure, RMSD, is particularly susceptible to this problem. A detailed analysis and comparison of various structural similarity measures and structural alignments obtained with them was the subject of a separate paper (Godzik, 1996). At this point, it is sufficient to say that the structural alignments based on contact map overlap are in most cases similar in terms of secondary structure elements assignment to the alignments optimizing either the C_α distance matrix or RMSD. However, they may differ strongly on the level of single residues.

There are many cases when the discrepancy between alignment accuracy measured with RMSD and CMO has structural reasons. For instance, there are protein pairs with similar topology but with significant rearrangements of smaller building blocks. Typical examples are multidomain proteins that share the same topology, but have different domain arrangements (see Fig. 4). In such cases, the alignment having low RMSD and covering the entire sequences is impossible to build and the structural alignments based on RMSD optimization are much shorter than sequence alignments. A well-known example of protein family with such structural features are calcium binding proteins (Pawlowski et al., 1996). Such pairs of proteins are easily detected as they have large alignment RMSD values but satisfactory CMO values. For this reason, in our tests we used CMO value as the structural criterion of alignment correctness.

The overlap of two contact maps, as defined by the number of contacts between equivalenced residues in two proteins that are simultaneously present in both structures (Godzik et al., 1993), is calculated by the following equation, where for the alignment between proteins A and B the summation is taken over all aligned pairs.

$$S_C(A, B) = \frac{1}{N_C} \sum_{i < j} C_{ij}^A C_{AB(i), AB(j)}^B \quad (1)$$

where the C_{ij}^A and C_{ij}^B are binary contact maps of a proteins A and B, respectively, $C_{AB(i), AB(j)}^B$ is a contact map of a protein B, but numbered according to the alignment AB between proteins A and B. The normalization factor N_C is equal to the number of contacts in the target protein.

As there is no ideal alignment accuracy measure, we decided to calculate the measures of accuracy for each method:

- SF0, the percent of the length of structural alignment correctly reproduced by a given alignment
- RMSD, root-mean-square deviation of target and template C_α atoms when assigned to each other according to the alignment
- CMO, contact map overlap of target and template structures when assigned to each other according to the alignment.

In the Tables 1 and 2, we presented the average values of these measures. It should be noted, however, that alignment accuracies are widely distributed for all the methods (see Figs. 1, 2), and the average values can give only some general insight.

Structural alignments

Structural alignments were calculated using combinatorial extension CE algorithm (Shindyalov & Bourne, 1998). The program source available from San Diego Supercomputer Center ftp site (<ftp://ftp.sdsc.edu/pub/sdsc/biology/CE/src/>) was downloaded and compiled on the Linux operating system. Default parameters of the CE algorithm were used. Protein structures were downloaded from Protein Data Bank (PDB) (<http://www.rcsb.org/pdb/>).

Sequence–sequence, profile–sequence, and profile–profile matching protocols

Simple mutation matrix based algorithm

The simple mutation matrix algorithm is based on the local dynamic programming algorithm (Needleman & Wunsch, 1970; Smith & Waterman, 1981). The accuracy of sequence alignment method depends strongly on the quality of the mutation matrices used. The overall assessment of various matrices was published by Vogt et al. (1995). These authors obtained most accurate alignments using gonnet_p mutation matrix. We used gonnet_p mutation matrix with optimal gap penalties identified in this paper.

PSI-Blast

Each target sequence from the benchmark pairs was used as the input for five PSI-BLAST iterations with E-value cutoff equal 0.001 on nonredundant database (NR). The profile from the last iteration is saved and used in one iteration of PSI-BLAST on the PDB database clustered at 40% of sequence identity that contains the template (PSI-BLAST-C and -R options were applied). The target–template alignment was taken from the PSI-BLAST output from that step.

FFAS

The detailed description of the FFAS method was published recently (Rychlewski et al., 2000). Every target sequence from the benchmark was used as an input for five PSI-Blast iterations with E-value cutoff equal to 0.001 using a nonredundant database (NR). All sequences which obtained E-value <0.001 in any iteration were included in the family profile. The resulting profiles of the target sequence and template sequence were aligned with local dynamic programming algorithm (Needleman & Wunsch, 1970; Smith & Waterman, 1981). The gap initiation penalty of 4.71, gap extension penalty 0.37, and comparison matrix zero level of −0.12 were applied that were identified as optimal for fold recognition sensitivity (Rychlewski et al., 2000).

FFAS-R and other alignments with predefined length

It was shown in Vogt et al. (1995) that adding or subtracting a constant value from the amino acid mutation matrix causes the significant change of the alignment features. The average value or “zero level: of mutation matrix, together with two gap penalties (introduction and extension) determines the alignment features, even that the latter two parameters usually receive more attention. Lower zero level of the mutation matrix yields more accurate but

shorter alignments and the higher value is connected with longer but less reliable alignments. In other words, tuning up this parameter enables the balance of the alignment length and its accuracy. To allow direct comparison of profile–profile, sequence–sequence, and sequence–profile alignments with PSI-BLAST alignments, one needs the alignments of the same average length. By changing the zero of the mutation matrix, we obtained alignment methods that yield the alignments of the same average length that PSI-BLAST does. This allows for the comparison of the accuracy of alignments obtained with different algorithms without the bias of the length difference. The modified version of FFAS method producing “length restricted” alignments was called FFAS-R.

The statistical significance of the FFAS alignment score

All protein pairs included in our benchmark show significant structural similarity. Hence, ideal alignment algorithm should produce the accurate and statistically significant alignments for all protein pairs.

The wide distribution of FFAS alignment accuracies for the sequence pairs of low similarity (see Figs. 1, 2) indicates the importance of differentiation between accurate and incorrect alignments. The most evident parameter that should allow this is the alignment score. More precisely, we expect the alignment accuracy to be correlated with the statistical significance of the alignment score. It should be noted, however, that low statistical significance of the alignment score does not mean that the proteins are not similar, but in the case of our tests, it rather indicates, that the alignment algorithm did not detect this similarity.

The alignment score for a given target–template pair was compared with scores of the same target when aligned with 1,800 sequences from PDB40D-J database (Park et al., 1998). The sequences having significant similarity to the target were excluded, so the distribution of the alignment scores is a good approximation of the distribution of random alignment scores. If the alignment score for a given benchmark pair is better than 99% of random scores, it is classified as significant. The alignments having significant and insignificant scores were evaluated separately.

The high statistical significance of the alignment score is equivalent to successful recognition of fold similarity. The set of alignments having significant scores corresponds to the target–template pairs whose similarity is detectable by FFAS in fold recognition test.

The distribution of the alignment accuracy measured by CMO values is shown on Figure 5. The distribution was calculated for the set of 700 structurally similar protein pairs from PDB40D-J database (Park et al., 1998). (Larger benchmark was used to obtain better statistics.) Z-score values were calculated based on the distribution of alignment scores of each sequence aligned with 1,800 sequences from PDB40D-J database as described earlier in this paragraph. We used standard formula for the Z-score calculation: $Z\text{-score} = (\text{Score} - \text{Average_Score}) / \text{Standard_Deviation}$. The graph shows that there are very few alignments of high statistical significance and low accuracy, and there is significant correlation between alignment z-score and the CMO of the worst alignment with this this z-score (see Fig. 5).

The alignment stability test

One of the possible ways of recognizing the alignment accuracy is by studying its stability (Mevissen & Vingron, 1996). We at-

Fig. 5. The contact map overlap values shown as the function of the alignment z-score calculated for 700 pairs of structurally similar proteins from PDB40D-J database (Park et al., 1998). The average value of the alignment CMO and lower limit of the alignment CMO are approximated with linear functions (upper and lower line, respectively). We used negative z-score values to show that similarity score is lower (better) than average value.

tempted to test this by applying a simple stability test to the benchmark alignments. The alignments calculated with PSI-BLAST and FFAS differ significantly as indicated by accuracy measures. We can anticipate that the fragments that are identical in PSI-BLAST and FFAS alignments should be more accurate than the rest of PSI-BLAST or FFAS alignments.

PSI-BLAST and FFAS alignments were compared position-by-position and their “common parts” were evaluated in the same way as the alignments from other methods. When a “common part” of PSI-BLAST and FFAS alignments is empty, it means that there are no residues that are identically aligned in PSI-BLAST and FFAS alignment. This is true for 51 pairs, and for 40 of them, the FFAS alignment has insignificant score. This means that most of these alignments have low quality and removing them from the benchmark improves the average quality of the remaining alignments (see Table 3).

Acknowledgments

The authors are grateful to Dr. Pawlowski for many useful comments and discussion concerning this study. The research described in this paper 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* 215:403–410.
- 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* 25:3389–3402.
- Aravind L, Koonin EV. 1999. Gleaning non-trivial structural, functional and evolutionary information about proteins by iterative database searches. *J Mol Biol* 287:1023–1040.
- Bateman A, Birney E, Durbin R, Eddy SR, Howe KL, Sonnhammer EL. 2000. The Pfam protein families database. *Nucleic Acids Res* 28:263–266.
- Burley SK, Almo SC, Bonanno JB, Capel M, Chance MR, Gaasterland T, Lin D, Sali A, Studier FW, Swaminathan S. 1999. Structural genomics: Beyond the human genome project. *Nat Genet* 23:151–157.

- Domingues FS, Koppensteiner WA, Jaritz M, Prlic A, Weichenberger C, Wiederstein M, Floeckner H, Lackner P, Sippl MJ. 1999. Sustained performance of knowledge-based potentials in fold recognition. *Proteins Suppl* 53:112–120.
- Eddy SR. 1998. Profile hidden Markov models. *Bioinformatics* 14:755–763.
- Eisenberg D, Luthy R, Bowie JU. 1997. VERIFY3D: Assessment of protein models with three-dimensional profiles. *Methods Enzymol* 277:396–404.
- Fischer D. 1999. Modeling three-dimensional protein structures for amino acid sequences of the CASP3 experiment using sequence-derived predictions. *Proteins* 37:61–65.
- Gibrat JF, Madej T, Bryant SH. 1996. Surprising similarities in structure comparison. *Curr Opin Struct Biol* 6:377–385.
- Godzik A. 1996. The structural alignment between two proteins: Is there a unique answer? *Protein Sci* 5:1325–1338.
- Godzik A, Kolinski A, Skolnick J. 1992. Topology fingerprint approach to the inverse protein folding problem. *J Mol Biol* 227:227–238.
- Godzik A, Skolnick J, Kolinski A. 1993. Regularities in interaction patterns of globular proteins. *Protein Eng* 6:801–810.
- Hargbo J, Elofsson A. 1999. Hidden Markov models that use predicted secondary structures for fold recognition. *Proteins* 36:68–76.
- Holm L, Sander C. 1998. Dictionary of recurrent domains in protein structures. *Proteins* 33:88–96.
- Hubbard TJ. 1999. RMS/coverage graphs: A qualitative method for comparing three-dimensional protein structure predictions. *Proteins* 53:15–21.
- Jaroszewski L, Rychlewski L, Zhang B, Godzik A. 1998. Fold prediction by a hierarchy of sequence, threading, and modeling methods. *Protein Sci* 7:1431–1440.
- Jones DT. 1999. GENTHREADER: An efficient and reliable protein fold recognition method for genomic sequences. *J Mol Biol* 287:797–815.
- Jones TA, Kleywegt GJ. 1999. CASP3 comparative modeling evaluation. *Proteins* 53:30–46.
- Kabsch W. 1978. A discussion of the solution for the best rotation to relate two sets of vectors. *Acta Crystallogr A* 34:827–828.
- Karplus K, Barrett C, Cline M, Diekhans M, Grate L, Hughey R. 1999. Predicting protein structure using only sequence information. *Proteins Suppl* 53:121–125.
- Mevissen HT, Vingron M. 1996. Quantifying the local reliability of a sequence alignment. *Protein Eng* 9:127–132.
- Moult J. 1999. Predicting protein three-dimensional structure. *Curr Opin Biotechnol* 10:583–588.
- Murzin AG. 1999. Structure classification-based assessment of CASP3 predictions for the fold recognition targets. *Proteins* 37:88–103.
- 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* 247:536–540.
- 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* 48:443–453.
- Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM. 1997. CATH—A hierarchic classification of protein domain structures. *Structure* 5:1093–1108.
- Panchenko A, Marchler-Bauer A, Bryant SH. 1999. Threading with explicit models for evolutionary conservation of structure and sequence. *Proteins Suppl* 53:133–140.
- 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* 284:1201–1210.
- Pawlowski K, Bierzynski A, Godzik A. 1996. Structural diversity in a family of homologous proteins. *J Mol Biol* 258:349–366.
- Pawlowski K, Jaroszewski L, Bierzynski A, Godzik A. 1997. Multiple model approach—dealing with alignment ambiguities in protein modeling. *Pac Symp Biocomput*:328–339.
- Rost B, Sander C. 1996. Bridging the protein sequence-structure gap by structure predictions. *Annu Rev Biophys Biomol Struct* 25:113–136.
- Rychlewski L, Jaroszewski L, Weizhong L, Godzik A. 2000. Comparison of sequence profiles. Structural predictions with no structure information. *Protein Sci* 8:232–241.
- Rychlewski L, Zhang B, Godzik A. 1999. Functional insights from structural predictions: Analysis of the *Escherichia coli* genome. *Protein Sci* 8:614–624.
- Sali A, Potterton L, Yuan F, van Vlijmen H, Karplus M. 1995. Evaluation of comparative protein modeling by MODELLER. *Proteins* 23:318–326.
- Sanchez R, Sali A. 1998. Large-scale protein structure modeling of the *Saccharomyces cerevisiae* genome. *Proc Natl Acad Sci USA* 95:13597–13602.
- Shindyalov IN, Bourne PE. 1998. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng* 11:739–747.
- Smith TF, Waterman MS. 1981. Identification of common molecular sub-sequences. *J Mol Biol* 147:195–197.
- Svensson B, Vass I, Cedergre, E, Styring S. 1990. Structure of donor side components in photosystem II predicted by computer modeling. *EMBO J* 9:2051–2059.
- Sybyl. 1998. Sybyl 6.5 Tripos Inc. www.tripos.com.
- Vogt G, Etzold T, Argos P. 1995. An assessment of amino acid exchange matrices in aligning protein sequences: The twilight zone revisited. *J Mol Biol* 249:816–831.
- Yu L, Smith TF. 1999. Positional statistical significance in sequence alignment. *J Comput Biol* 6:253–259.
- Zuker M. 1991. Suboptimal sequence alignment in molecular biology. Alignment with error analysis. *J Mol Biol* 221:403–420.