Quality of alignment comparison by COMPASS improves with inclusion of diverse confident homologs

Ruslan I. Sadreyev and Nick V. Grishin*

Howard Hughes Medical Institute, and Department of Biochemistry, University of Texas Southwestern Medical Center, 5323, Harry Hines Blvd, Dallas, TX 75390-9050, USA

Received on March 19, 2003; revised on October 16, 2003; accepted on October 17, 2003
Advance Access publication January 29, 2004

ABSTRACT
Motivation: Adding more distant homologs to a multiple alignment and thus increasing its diversity may eventually deteriorate the numerical profile constructed from this alignment. Here, we addressed the question whether such a diversity limit can be reached in the alignments of confident homologs found by PSI-BLAST, and we analyzed the dependence of the quality of the profile–profile comparison made by COMPASS on the sequence diversity within these alignments.

Results: Protein families that have a greater number of diverse confident homologs in the current sequence databases provide an increased quality of similarity detection in profile databases, but produce on average less accurate profile–profile alignments with their remote relatives. This lower alignment accuracy cannot be improved when the most distant members of these families are excluded from their profiles. On the contrary, the presence of more diverse members results in more accurate alignments. For families with a high diversity of confident homologs, the lower quality of profile alignments with their remote relatives seems to be an attribute of these families or their alignments, rather than to be caused by the large number of diverse sequences itself. Our results suggest that at any level of profile diversity, one should include in the multiple alignment as many confident sequence homologs as possible in order to produce the most accurate results.

Contact: grishin@chop.swmed.edu

INTRODUCTION
Development of more sensitive and accurate automatic methods for sequence comparison is essential for the prediction of protein structure and function, the identification of functionally important regions in protein sequences, and the analysis of remote evolutionary relationships within protein families. Whereas pairwise sequence alignment methods are effective for the comparison of proteins with relatively high sequence identity, more rigorous protocols based on the profile comparison (Altschul et al., 1997; Eddy, 1998; Karplus et al., 1998; Schaffer et al., 1999, 2001; Pietrokovski, 1996; Kunin et al., 2001; Rychlewski et al., 2000; Yona and Levitt, 2002; Sadreyev and Grishin, 2003; Sadreyev et al., 2003) were shown to improve both the sensitivity of homology detection and the accuracy of produced alignments. Sequence–profile and profile–profile comparisons utilize information about general features in protein families, even if the query sequence lacks these features. The most popular current protocols involve comparing multiple alignments to individual sequences. Meanwhile, recent reports (Pietrokovski, 1996; Kunin et al., 2001; Rychlewski et al., 2000; Yona and Levitt, 2002; Sadreyev and Grishin, 2003; Panchenko, 2003) suggest that the methods for profile–profile comparison are more sensitive and accurate. The methods proposed up to date vary in the employed procedures for the profile generation from multiple alignments, scoring systems, alignment extension algorithms and protocols for statistical evaluation of produced profile–profile alignments (Pietrokovski, 1996; Kunin et al., 2001; Rychlewski et al., 2000; Yona and Levitt, 2002; Sadreyev and Grishin, 2003; Panchenko, 2003).

The accurate profile representation of specific features of a protein family strongly depends on the input multiple alignment. Analysis of this dependence is important for more accurate comparison of multiple alignments. In this work, we focus on one property of the multiple alignments: the profile diversity, i.e. the diversity of the set of sequences comprising the multiple alignment, measured by the average number of different symbols at the alignment positions.

We analyzed the effect of profile diversity on the quality of the profile–profile alignments and similarity detection between remote protein families. One might hypothesize that adding more remote homologs to the multiple alignment of a protein family may eventually reduce the profile quality, due to less conserved patterns of residue representation at the profile positions. Indeed, a recent report (Panchenko, 2003) suggested that there is an optimal profile diversity that provides the most accurate detection of remote sequence similarities by profile–profile comparisons. Using sequence families of various diversities, the author observed that profiles of highly
diverse families showed a reduction in the detection quality compared to the families with less diverse profiles.

The existence of a diversity optimum would have immediate practical implications for the procedures of the profile comparison. This upper limit of profile diversity would restrict application of any profile–profile comparison method, since the use of extremely diverse profiles would result in a poorer performance. If this is the case, the procedures of the profile generation should be modified in order to avoid the excessive diversity (e.g. by purging the most remote homologs, or splitting the family into less diverse subfamilies).

We addressed the question whether, among the alignments of confident homologs, there is a limit for profile diversity above which the profile quality deteriorates. To approach this question, we used reliable sets of undoubted yet diverse homologs for a number of protein families and produced multiple sequence alignments of various diversities that correspond to the same families. As a testing tool for the profile–profile comparison, we used recently proposed method COMPASS (Sadreyev and Grishin, 2003) in order to analyze the effect of profile diversity on (i) the quality of the profile–profile alignments of remotely related families, and (ii) the quality of the profile–profile similarity detection in profile databases.

METHODS

Measure of alignment diversity

As a simple measure of alignment diversity, we used the average number of different symbols in the alignment columns that have low enough gap content. The effective gap content was calculated based on position-specific independent counts (Sunyaev et al., 1999) as described in Sadreyev and Grishin (2003). The columns that had the effective gap content 0.5 or higher were excluded from the consideration, since these columns are the result of insertions in a small subset of sequences and thus do not reflect general features of the alignment. In each of the remaining columns, the number of different symbols was counted, including 20 residue types and the gap symbol. As an estimate of the alignment diversity, we used the average number of different symbols in all considered columns ($N_{\text{eff}}$), which ranges between 1.0 and 21.0. A similar approach to the estimation of profile diversity has been previously used by others (e.g. Altschul et al., 1997; Schaffer et al., 1999, 2001; Panchenko, 2003). We chose this approach to make our results consistent with previous works, in particular with the recent results (Panchenko, 2003) relevant for this study.

Purging remote homologs to produce alignments of lower diversity

In order to compare the profiles of various diversities that correspond to the same family, we used the following procedure. From each alignment of the high diversity ($N_{\text{eff}} > 15$), we produced three less diverse alignments in three lower ranges of $N_{\text{eff}}$ (1–5, 5–10 and 10–15) by purging the most remote homologs located in the bottom part of the multiple alignment produced by PSI-BLAST. To achieve the diversity of the ‘trimmed’ alignment that belongs to a given range (e.g. $N_{\text{eff}} = 5–10$), we used the following iterative procedure. The set of sequences in the initial alignment was split into upper and lower halves, and the diversity ($N_{\text{eff}}$) was calculated for the upper half of the alignment. If the diversity of this sub-alignment was within the specified range (e.g. $5 \leq N_{\text{eff}} < 10$) then the procedure was stopped and the resulting sub-alignment was further used as the ‘trimmed’ alignment. If the diversity of the upper half was higher than necessary (e.g. $N_{\text{eff}} \geq 10$) then this sub-alignment was further split in halves, and the new iteration started. Otherwise, if the diversity of the top half was below the specified range (e.g. $N_{\text{eff}} < 5$), the bottom half was split in two equal parts and the upper of these parts was added to the top half. The resulting sub-alignment entered the next iteration. Multiple iterations of this dichotomy search provided the sub-alignment whose diversity belonged to the specified range.

Assessment of alignment quality

As a benchmark for the evaluation of alignment quality, we generated and further processed PSI-BLAST alignments for pairs of protein domains that are structurally related according to the FSSP database (Holm and Sander, 1996; Dietmann et al., 2001). We used these profile pairs as input for COMPASS and PSI-BLAST, and compared the predicted local alignments with the structure-based FSSP alignments.

We were most interested in the analysis of the method performance in twilight zone (sequence identities below ~25%), where the performance of methods for sequence–sequence comparison deteriorates. Since in this zone the alignment quality strongly depends on the similarity between input profiles (Sadreyev and Grishin, 2003), we restricted the sets of benchmark sequence pairs to the narrow ranges of identity. The sequence identities were taken from FSSP, where they are rounded and represented as integers. We focused on two FSSP identity ranges: 25 ± 1% at the upper bound of twilight zone and 15 ± 1%, where the accurate performance is much more problematic yet possible for many sequence pairs.

From each FSSP family, we extracted the parent sequence and all sequences of a significant structural similarity to the parent (Z-score greater than 5.0), with sequence identity to the parent within a given range. We found totally 494 and 1406 sequence pairs with identities 25 ± 1% and 15 ± 1%, respectively. These numbers were reduced by purging symmetric pairs and manual inspection of the remaining domains for the presence of repeats and low-complexity regions, which render the global structure based alignments inappropriate standard for sequence-based methods. For further analysis, we used 251 sequence pairs with identity 25 ± 1% and 340 pairs with identity 15 ± 1%, each pair representing a single
unique FSSP family. For each sequence, we ran five iterations of PSI-BLAST 2.2.1 against the NCBI nr database (E-value threshold for inclusion in the next iteration 0.005, BLOSUM62 matrix) and obtained multiple alignments of found homologs. In order to confirm the absence of false positives, we specifically tested these sets of homologs for the presence of sequences whose homology is unclear. In particular, we considered the sets with the largest numbers of diverse sequences, where false positives are most likely to occur, and ran PSI-BLAST starting from every sequence as a query. These back searches confidently detected the similarity to the initial parent sequence, which confirms the absence, or at least very low content of, spurious hits in our sequence sets.

When comparing an alignment to a sequence, PSI-BLAST performs pre-processing of the input alignment before the profile is produced (Altschul et al., 1997). To perform the profile–profile comparison with COMPASS, we used a similar procedure of the alignment pre-processing. In particular, only one copy was retained of any rows that were >97% identical to one another and the columns with gaps inserted into the first (query) sequence were purged. The resulting pairs of alignments were used for the construction of their local alignment by COMPASS and its evaluation. As a method for sequence–profile comparison, we used PSI-BLAST 2.2.1 to align the parent FSSP sequence to the profile for the alignment based on the child FSSP sequence. We considered only the PSI-BLAST alignments that were actually produced; the cases where PSI-BLAST failed to find any word matches comprised less than 5% of the set and were excluded from the analysis.

As a reference to assess the quality of the produced alignments, we used structure-based FSSP alignments, i.e. the residue equivalences between confidently aligned protein positions marked as capital letters in FSSP. To measure the quality of profile–profile (or sequence–profile) alignment, we used the parameters proposed earlier (Sauder et al., 2000; Yona and Levitt, 2002; Sadreyev and Grishin, 2003). The quality from the modeler’s point of view ($Q_{mod}$) is the ratio of the number of correctly aligned positions to the total number of positions in the evaluated alignment. Thus, $Q_{mod}$ reflects the proportion of the position matches in the evaluated alignment that correctly reproduce the reference alignment. The quality from the developer’s point of view ($Q_{dev}$) is the ratio of the number of correctly aligned positions to the number of positions in the structural alignment. In the setting when two versions of a profile–profile alignment are compared using the same structure based alignment as a reference, $Q_{dev}$ is proportional to the total number of correctly aligned positions in the evaluated alignments. A higher $Q_{mod}$ means that a greater portion of the evaluated alignment is correct, whereas a higher $Q_{dev}$ means a larger absolute number of correctly reproduced position matches from the reference. If one profile–profile alignment provides higher $Q_{mod}$ without compromising $Q_{dev}$, or higher $Q_{dev}$ without compromising $Q_{mod}$, we consider such alignment more accurate than the other.

For local alignments, it is reasonable to assess the local prediction for only those regions of the structural alignment that are included in the evaluated alignment. Thus in addition to $Q_{dev}$, we used a measure of ‘local accuracy’ (Sadreyev and Grishin, 2003): $Q_{local} = N_{acc}/L$, where $N_{acc}$ is the number of correct matches, $L$ is the length of the region of the structural alignment that includes the pairs of the profile positions from the evaluated alignment. $Q_{local}$ is close to 1.0 for alignments with the correct prediction of structural matches, even if they are very short. To assess directly the length of the region covered by the alignment, we introduced the additional measure of coverage independent of accuracy. To calculate the coverage, we determined the length of the region in the structural alignment that includes all the profile positions from the evaluated alignment and divided it by the overall length of the structural alignment. If one profile–profile alignment provides higher $Q_{local}$ without compromising coverage, or higher coverage without compromising $Q_{local}$, we consider such alignment more accurate than the other.

Assessment of the ability to detect remote sequence similarities

The quality of remote similarity detection was assessed on two sets of multiple alignments: the alignments of the homologs produced by PSI-BLAST for the sequences from the FSSP database (Holm and Sander, 1996; Dietmann et al., 2001), and the alignments from the PFAM database (Bateman et al., 2002) that contained at least one sequence from FSSP.

As the first testing set of multiple alignments, we used the PSI-BLAST alignments with the FSSP entries as queries that were used for the assessment of the alignment quality (previous section). After manual inspection for the presence of repeats and low-complexity regions, we used 681 PSI-BLAST alignments that corresponded to the queries of known tertiary structure. Within the resulting database, we performed exhaustive COMPASS searches with each alignment as a query. In order to perform PSI-BLAST searches, we prepared the database of all 2735061 sequences extracted from these alignments, and ran PSI-BLAST searches using each of the alignments as a query (one round of the PSI-BLAST 2.2.1 search with PSI-BLAST numerical profile derived from the alignment; the template sequence was set to the first sequence of the query alignment; the maximal number of displayed hits and the maximal E-value were both set to 10000). Then a list of sequence hits for each query alignment was transformed into the list of similarities between the query and other alignments. Specifically, for each alignment from the dataset, the best E-value among all its sequences was chosen as a measure of similarity between this alignment and the query.

The query alignments were divided into categories by their diversity ($N_{eff} = 1–5, 5–10, 10–15$ and $>15$). For each category, all found profile similarities were pooled together and
ranked by their $E$-value. To evaluate a hit as a true positive, we required that it should be consistent with the structural relationship in FSSP, i.e. that the FSSP entry from the hit should belong to the same FSSP family as the entry from the query, with the structural similarity $Z$-score $> 2.0$. Sensitivity curves (plots of the number of true positives versus the number of false positives) were constructed for each diversity range, and receiver operating characteristics for the top of false positives were calculated as ROC$\_n = (1/nT) \sum_{i=1}^{n} t_i$, where $t_i$ is the number of true positives that were ranked ahead the $i$th false positive in the list, and $T$ is the total number of true positives in the dataset. The values of $T$ for the dataset of the PSI-BLAST alignments were 19,024, 36,181, 61,188 and 225,886, respectively for the queries with the diversity within the four ranges. The standard deviation of ROC was calculated as previously proposed (Schaffer et al., 2001).

To generate the second testing set of multiple alignments, we collected all 1354 alignments from the PFAM database (Bateman et al., 2002) (version 6.6) that contained at least one sequence from the FSSP database (Holm and Sander, 1996; Dietmann et al., 2001). The exhaustive COMPASS and PSI-BLAST searches with each alignment as a query were performed in this database in the same manner as described above. None of the considered PFAM alignments had $N_{\text{eff}} > 15$; therefore, the ranges of diversity in for the PFAM alignments were chosen as $N_{\text{eff}} = 1–5$, $N_{\text{eff}} = 5–10$ and $N_{\text{eff}} > 10$. The total numbers of true positives in the dataset ($T$) for the queries with profile diversity within these ranges were 186,838, 417,204 and 446,246, respectively. The total number of the sequences extracted from these alignments and used as a database for the PSI-BLAST searches was 311,753.

RESULTS

Quality of produced alignments

For the pairs of sequences from FSSP with identity within the ranges of $15 \pm 1\%$ and $25 \pm 1\%$, we produced multiple alignments of their homologs found by PSI-BLAST. We used these alignment pairs as input for the COMPASS comparisons of profile1 versus profile2 and for the PSI-BLAST comparisons of sequence1 versus profile2. The quality of the resulting local alignments was assessed by their consistency with the structure-based alignments in FSSP (see Methods section). The average quality values were calculated for the alignments that were assigned $E$-value $<10\^{-3}$ by COMPASS, in four bins of diversity of profile2 ($N_{\text{eff}} \leq 5$, $N_{\text{eff}} = 5–10$, $N_{\text{eff}} = 10–15$ and $N_{\text{eff}} > 15$). The local ($Q_{\text{local}}$ and coverage) measures of quality are shown in Figure 1. The global ($Q_{\text{mod}}$ and $Q_{\text{dev}}$) measures of quality are shown in Figure 2. A higher $Q_{\text{mod}}$ means that a greater portion of the evaluated alignment (compared to its length) is correct, whereas a higher $Q_{\text{dev}}$ means a larger absolute number of correctly reproduced position matches from the reference. If one profile–profile alignment provides higher $Q_{\text{local}}$ and $Q_{\text{mod}}$ than another (without compromising coverage and $Q_{\text{dev}}$), or higher coverage and $Q_{\text{dev}}$ (without compromising $Q_{\text{local}}$ and $Q_{\text{mod}}$), we consider such alignment more accurate than the other.

Figures 1 and 2 show the alignment quality for three types of experiments: (i) sequence–profile comparisons by PSI-BLAST (dashed lines, empty circles) for different families binned by the diversity of profile2 (profile1 being a single sequence); (ii) profile–profile comparisons by COMPASS (solid lines, filled circles) for the pairs of families used in (i); and (iii) profile–profile comparisons by COMPASS (gray triangles) for the ‘trimmed’ alignments produced from the set of highly diverse families (see below).

Consistent with previous results (Petrokovskiy, 1996; Rychlewski et al., 2000; Yona and Levitt, 2002; Sadreyev and Grishin, 2003; Panchenko, 2003), distantly related sequences (15% identity) showed a more significant improvement in the quality of profile–profile comparisons with respect to sequence–profile comparisons (up to $3\times$-fold increase in $Q_{\text{local}}$, $Q_{\text{mod}}$ and $Q_{\text{dev}}$, and a higher coverage, as compared to PSI-BLAST, Figs. 1a, b and 2a, b), whereas at the upper bound of the twilight zone the performance of the two protocols becomes more similar (Figs. 1c, d and 2c, d).

In both identity ranges, most of the measures show a reduction of the alignment quality at the high level of profile diversity ($N_{\text{eff}} > 15$). At this level, the portion of the correctly aligned positions in the alignments drops, as reflected by $Q_{\text{local}}$ (Fig. 1a, c) and $Q_{\text{mod}}$ (Fig. 2a, c). The average values of coverage (Fig. 1b, d) and $Q_{\text{dev}}$ do not show statistically significant changes, which correspond to the profile alignments that cover similar lengths of the reference alignments, but include a lower portion of correctly predicted position matches. Higher level of diversity of profile2 strongly correlates with a higher level of diversity of profile1 (data not shown). Therefore, the observed effect is not due to a lower diversity of the counterpart profile1, which might affect the alignment quality. The lower performance of a protocol for profile–profile comparison on the set of highly diverse profiles is consistent with the previous observation by Panchenko (2003). Under the assumption that the profiles of different diversity otherwise have the same properties, this observation might suggest that the performance reduces at the high diversity levels due to the diversity increase. This reduction would mean the existence of the optimal level of profile diversity below the maximum. However, the described effect does not yet allow inferring a direct dependence of the performance on profile diversity.

To directly analyze the connection between profile diversity and the quality of profile–profile alignments, we used the sets of highly diverse multiple alignments ($N_{\text{eff}} > 15$) in order to produce ‘trimmed’ alignments of lower diversity. Then we
compared the COMPASS performance on the ‘trimmed’ and the initial non-trimmed alignments of the same families. The bottom sequences of multiple alignments (representing the most remote homologs found by PSI-BLAST) were purged so that the resulting thinner alignment had $N_{\text{eff}}$ in a given range (1–5, 5–10 or 10–15; see Methods). Having these profiles of various diversities for each family, we used COMPASS to construct the profile–profile alignments of unchanged profile1 and trimmed profile2, for each level of diversity of profile2. The average quality of the resulting alignments is shown as gray triangles in Figure 1 ($Q_{\text{local}}$ and coverage) and Figure 2 ($Q_{\text{mod}}$ and $Q_{\text{dev}}$). In addition, Table 1 illustrates the statistics of individual paired comparisons between the results for the initial profile2 and the trimmed profiles 2 ($N_{\text{eff}} \leq 5$ for profile2) for each family.

The dependence of the alignment quality on profile diversity for the trimmed alignments was significantly different from the behavior observed for the initial set of alignments divided into diversity bins. For both tested levels of sequence similarity, this dependence was non-decreasing within the statistical errors and did not show a reduction of accuracy at higher profile diversity. The dependence was more pronounced for more similar profile pairs (sequence identity 25 ± 1%). For this identity, increasing the diversity of profile2 results in the improvement of local alignment accuracy (Fig. 1c) and no significant change of coverage (Fig. 1d), which leads to the growth of integral alignment quality ($Q_{\text{mod}}$, Fig. 2c and $Q_{\text{dev}}$, Fig. 2d). These results are supported by paired $t$-tests (Table 1). These changes suggest that at higher level of the sequence similarity, profiles of greater diversity produce the
Sequence diversity improves alignment comparison

Fig. 2. Quality of COMPASS profile alignments of remote families: integral quality measures (quality from modeler’s and developer’s points of view, $Q_{\text{mod}}$ and $Q_{\text{dev}}$) for various ranges of the input profile diversity ($N_{\text{eff}}$, the centers of bins are shown). Trimming alignments of the same set produces a different trend (gray triangles) than separating different families by profile diversity (filled circles) (see details in the text). Empty circles, dashed lines: quality of PSI-BLAST sequence–profile alignments for the families binned by profile diversity (the same set as for the curve shown with filled circles). (a, b) Alignment quality for lower sequence identity (15%): (a) $Q_{\text{mod}}$, (b) $Q_{\text{dev}}$. (c, d) Alignment quality for higher sequence identity (25%): (c) $Q_{\text{mod}}$, (d) $Q_{\text{dev}}$.

Table 1. Paired comparison of alignment quality for COMPASS alignments produced with input multiple alignments of high diversity (‘Thick’, $N_{\text{eff}} > 15$) versus the same multiple alignments trimmed to lower diversity (‘Thin’, $N_{\text{eff}} \leq 5$)

<table>
<thead>
<tr>
<th>Quality measure</th>
<th>Coverage</th>
<th>Local accuracy</th>
<th>$Q_{\text{mod}}$</th>
<th>$Q_{\text{dev}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sequence identity</td>
<td>15%</td>
<td>25%</td>
<td>15%</td>
<td>25%</td>
</tr>
<tr>
<td>Thick &gt; thin</td>
<td>85</td>
<td>65</td>
<td>68</td>
<td>96</td>
</tr>
<tr>
<td>Thin = thick</td>
<td>26</td>
<td>45</td>
<td>21</td>
<td>14</td>
</tr>
<tr>
<td>Thick &lt; thin</td>
<td>32</td>
<td>55</td>
<td>54</td>
<td>55</td>
</tr>
<tr>
<td>Paired $t$-test $P$-value</td>
<td>$2.7 \times 10^{-9}$</td>
<td>0.99</td>
<td>$0.026$</td>
<td>$4.7 \times 10^{-9}$</td>
</tr>
</tbody>
</table>

Two levels of sequence similarity are considered (sequence identity 15% and 25%). For each quality measure, numbers of cases with increased, reduced, and unchanged quality are indicated, along with $P$-value produced by paired $t$-test ($P < 0.05$ are highlighted in bold).
alignments that cover approximately the same portions of the reference alignments, but the portion of correctly predicted position matches within these profile alignments is significantly higher.

For the profiles of lower similarity (sequence identity $15 \pm 1\%$), the dependence was different: increasing profile diversity leads to an increase of coverage, whereas the improvement of local accuracy is less pronounced (Table 1). The changes of the average values of local accuracy and coverage are within the sum of the standard errors (Fig. 1a, b). However, the more powerful paired comparison shows a significant difference between thick and thin alignments, which is confirmed by low $P$-values produced by paired $t$-test (Table 1). The portion of correctly reproduced matches of the structure-based alignments increases ($Q_{dev}$, Fig. 2b and Table 1); whereas the proportion of correct and incorrect matches within the profile–profile alignment itself (reflected by $Q_{mod}$, Fig. 2a and Table 1) does not change significantly. These changes suggest that at lower level of the sequence similarity, profiles of greater diversity produce the alignments that cover significantly longer parts of the reference alignments, without compromising the portion of correctly predicted position matches compared to the alignment length.

In summary, when different protein families are compared, the alignment quality reduces at high levels of profile diversity. However, this reduction is not observed when alignments of different diversity for the same family are considered. These results suggest that the families having a greater number of diverse (confident) homologs in the current sequence databases produce on average less accurate profile–profile alignments with their remote relatives, which cannot be improved when the most diverse members of these families are not included in their profiles. On the contrary, the presence of more diverse members results in more accurate alignments.

**Distributions of positional shifts among inaccurate matches for different levels of sequence similarity**

To analyze in more detail the difference between the described behaviors of alignment quality at 15% and 25% sequence identity, we considered only the profile positions that were misaligned by COMPASS. In particular, we analyzed the positions of two profiles that were predicted in the COMPASS alignments as not equivalent but were paired in the structure-based FSSP alignments. Within the sets of such positions, we compared the distributions of positional shifts introduced by COMPASS with respect to the reference alignments. These distributions were normalized by the total numbers of misaligned positions in a set. They show the relative frequencies of positional shifts among incorrect matches. For each identity range, we constructed such distributions for two levels of diversity of profile2: for the case of input alignments with $N_{eff} > 15$, and for the case of trimmed alignment 2 with $N_{eff} \leq 5$.

For the two levels of sequence similarity, we observed two distinct patterns of the distributions and of their changes with profile diversity (Fig. 3a, b). At 15% sequence identity, large shifts ($>\sim 10$ positions) are much more frequent than at 25% identity. These distant shifts remain considerably frequent when the diversity of input profiles grows (Fig. 3a). At 25% sequence identity, large positional shifts are infrequent, and their portion is dramatically reduced when profile diversity increases (Fig. 3b). As a result, the portion of only slightly misaligned positions increases [for the shifts less than 5, from...
73% to 92% (Fig. 3a), as opposed to a modest increase from 70% to 74% at 15% identity (Fig. 3b). Thus, in addition to the growing number of the correctly aligned positions, there is a reduction in the introduced shifts for the misaligned positions, which suggests that the profile alignment in general becomes more consistent with the reference alignment. This effect is more pronounced for the closer profile pairs. This is consistent with a major role of the local refinement (corrected placement of the profile positions that were previously misaligned) in the improvement of the alignment quality at 25% identity, as opposed to the extension of the alignment length, which is a major factor at 15% identity.

Quality of similarity detection

To measure the quality of similarity detection, we chose the ability to predict structural relationships between protein domains in the largest available source of automatic structural alignments, the FSSP database (Holm and Sander, 1996; Dietmann et al., 2001) (update of December 2001). We used two datasets of multiple alignments: the set of the PSI-BLAST alignments produced from the FSSP entries as queries (the same set as was used above for assessing the alignment quality), and the set the PFAM alignments that contain at least one sequence from FSSP. Each alignment was used as a query for COMPASS search in the dataset of individual sequences extracted from all alignments (see Methods section). The hits were evaluated as true or false positives by their consistency with structural similarity predicted in FSSP database.

To observe the dependence on the diversity of the query profile, we divided the alignments by their diversity $N_{\text{eff}}$ (see Methods section) and constructed separate sensitivity curves from the results of the searches with queries of different diversities. These curves for PSI-BLAST and COMPASS are compared in Figure 4a (data for the set of PSI-BLAST alignments) and Figure 4b (data for the set of PFAM alignments). Figure 4c and d show the receiver operating characteristics (ROC100) derived from these curves. In our case, the tests were performed using the same database but different sets of queries, which had different total numbers of true positives ($T$) that could be found in this database. The difference in $T$ for different benchmarks is accounted for by the normalization by $T$ in the ROC$_n$ formula (see Methods section). Thus, we believe ROC can be used as a comparative estimate of detection quality in our case. For both types of profiles used (PSI-BLAST and PFAM), the receiver operating characteristics (ROC100) of PSI-BLAST and COMPASS increase with the diversity of the query alignments.

COMPASS shows greater detection quality, especially at higher levels of alignment diversity, where the ratio of the ROCs produced by COMPASS and PSI-BLAST is $2.9 \pm 0.2$ for the case of the PSI-BLAST profiles (Fig. 4c) and $1.9 \pm 0.3$ for the PFAM profiles (Fig. 4d).

The rates of the true positive detection and the ROCs produced in our tests were lower than for previously reported benchmarks of manually collected homologs (e.g. Schaffer et al., 2001), due to two main factors. First, the used benchmarks presented a high level of difficulty for all methods, because they included very remote protein families whose relationship was based on the structure similarity. Second, the size of the database, and the total numbers of possible true positives were much higher than in the previously used benchmarks. Nonetheless, the absolute numbers of the true positives detected before 100 false positives were large enough to make statistically confident ROC100 estimates. This is confirmed by reasonable ROC errors (Fig. 4), which allowed distinguishing between the ROC values in most cases.

DISCUSSION

Consistent with previous results (Pietrokovski, 1996; Kunin et al., 2001; Rychlewski et al., 2000; Yona and Levitt, 2002; Sadreyev and Grishin, 2003; Panchenko, 2003), profile–profile comparison is more powerful than sequence–profile comparison, both in the sense of sequence similarity detection and of alignment quality. The advantage is especially significant for sequences of remote similarity.

Analyzing the quality of the profile–profile alignments at various levels of the input profile diversity, we found a discrepancy between the results obtained for the alignments of different families and for the alignments of different diversity produced for the same family. For the subset of sequences whose homologs found by PSI-BLAST comprised highly diverse profiles, the alignments were on average less accurate than for less diverse families. To directly observe the dependence of the alignment quality on profile diversity, we compared the performance of COMPASS for the profiles that represent the same family but have different diversity. This comparison can be viewed both as trimming the diverse alignments and as adding more remote sequences to the trimmed alignments. The comparison showed that starting from extremely diverse alignments of confident homologs and consequently reducing profile diversity by purging the most remote sequences, one cannot improve the quality of profile–profile alignments between these families and their remote relatives. On the contrary, the alignments were less accurate for the ‘trimmed’ alignments. This comparison also suggests that inclusion of more remote confident homologs into less diverse alignments does not reduce the average quality of the profile–profile alignments, even at very high levels of profile diversity.

Thus, COMPASS alignment quality does not show a direct negative correlation with the diversity of input profiles, provided that they are comprised of confident homologs. This result suggests that a lower quality of profile alignments produced from highly diverse protein families may be caused not by their diversity per se but by other properties of these
Fig. 4. The quality of homology detection grows faster with the diversity of query profile in the case of profile–profile comparison (COMPASS) than in the case of profile–sequence comparison (PSI-BLAST). (a, b) Sensitivity curves for PSI-BLAST and COMPASS searches with query alignments of various diversity (diversity ranges are indicated above each curve pair). The data are shown for two types of alignment datasets: (a) PSI-BLAST alignments (the same set as used in Figs 1 and 2) and (b) PFAM alignments that contain sequences from the FSSP database. None of the alignments in the PFAM set had $N_{\text{eff}} > 15$. (c, d) Comparison of ROC$_{100}$ obtained for PSI-BLAST and COMPASS on the two profile datasets: (c) PSI-BLAST alignments, (d) PFAM alignments.

A similar observation was made recently (Panchenko, 2003) for the quality of the profile–profile similarity detection by another method. The author suggested the existence of an optimum of profile diversity that provides the best performance of the method for a given level of sequence similarity. The method employed a very similar procedure for profile construction but a different measure for scoring similarity between profile positions than that implemented in COMPASS. The scoring function used by Panchenko combines the log-odds ratios for the residue target frequencies in the families (or their PSI-BLAST alignments) that correlate with the presence of a large variety of homologous sequences in the database. A more detailed analysis of this interesting phenomenon would be important for further improvement of profile–profile comparison. The extreme diversity of confident homologs found by PSI-BLAST might serve as a marker of a lower ‘alignability’ of such families, the property that seems to be an attribute of the family and might reflect its inner structure, or the structure of its alignment produced by PSI-BLAST.
Pearson’s correlation coefficient (Panchenko, 2003), whereas the COMPASS includes the log-odds ratios in a relative entropy—type scoring function (Sadreyev and Grishin, 2003). According to our previous results (Mittelmann et al., 2003), in the upper part of twilight zone, the discriminative power of the measure based on Pearson’s correlation coefficient is comparable to the measures based on relative entropy. For very remote sequence similarities, however, Pearson’s correlation coefficient represents the position similarities less accurately than relative entropy does (Mittelmann et al., 2003). It would be interesting to analyze the performance of the method described in (Panchenko, 2003) not on the sets of protein families divided by profile diversity, but on the same set of families, with profile diversity adjusted to various levels.

In addition to the increase of the alignment quality in the sense of the numbers of correctly reproduced matches of the reference alignment, we found that, for the more diverse profiles of the same family, there is a change in the population of the misaligned positions. In particular, the portion of slightly misaligned positions (the profile positions that should be matched according to the reference and were placed only several residues apart) increases compared to other misaligned positions. This increase is more pronounced for closer profile pairs (25% sequence identity). This effect is consistent with a major role of the local refinement (a closer placement of profile positions that were previously misaligned) in the improvement of the alignment quality for more similar profile pairs, as opposed to the extension of the alignment length, which is a major factor for less similar pairs.

Our results are valid for a typical setting for the generation of multiple alignments of homologs from single sequences: several iterations of PSI-BLAST with rather conservative default E-value of 0.005. The absence of false positives was confirmed by the PSI-BLAST back searches using the found homologs as queries, which confidently detected the similarity to the initial parent sequence (see Methods section). Thus, even when the resulting homologs were highly diverse, their similarity was sufficient for confident detection by PSI-BLAST. Less conservative PSI-BLAST settings might introduce more distant homologs into the profiles yet additionally producing false hits. Adding spurious hits will finally reduce the quality of a profile, and finding the optimal point for the inclusion of the doubted homologs is an interesting future direction, which has a practical value. In particular, it is of interest to determine the range of E-value (or other measure of distance between homologs included in the multiple alignments) where our conclusions are valid.

**CONCLUSION**

We found that (i) protein families that have a greater number of diverse (confident) homologs in the current sequence databases produce on average less accurate profile—profile alignments with their remote relatives; (ii) this lower accuracy cannot be improved when the most diverse members of these families are not included in their profiles; on the contrary, the presence of more diverse members results in more accurate alignments; (iii) protein families with a greater number of diverse homologs provide increased sensitivity and selectivity of similarity detection in profile databases. Our results suggest the following implications. (a) High diversity of the confident homologs found by PSI-BLAST might be a marker of less ‘alignable’ families, which seems to be an attribute of their PSI-BLAST alignments, rather than to be caused by the high number of various sequences itself. (b) In order to produce the most accurate results, at any level of profile diversity one should include in the multiple alignment as many confident sequence homologs as possible. (c) When more diverse confident homologs are incorporated into a profile, the greatest improvement in the accuracy of profile alignments should be expected for closer families with identities, at the upper border of twilight zone and higher.

**REFERENCES**


