doi:10.1016/S0022-2836(02)01371-2 J. Mol. Biol. (2003) 326, 317–336

JMB

Available online at www.sciencedirect.com

SCIENCE DIRECT®

COMPASS: A Tool for Comparison of Multiple Protein Alignments with Assessment of Statistical Significance

Ruslan Sadreyev and Nick Grishin*

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

*Corresponding author

We present a novel method for the comparison of multiple protein alignments with assessment of statistical significance (COMPASS). The method derives numerical profiles from alignments, constructs optimal local profile–profile alignments and analytically estimates *E*-values for the detected similarities. The scoring system and *E*-value calculation are based on a generalization of the PSI-BLAST approach to profile–sequence comparison, which is adapted for the profile–profile case. Tested along with existing methods for profile–sequence (PSI-BLAST) and profile–profile (prof_sim) comparison, COMPASS shows increased abilities for sensitive and selective detection of remote sequence similarities, as well as improved quality of local alignments. The method allows prediction of relationships between protein families in the PFAM database beyond the range of conventional methods. Two predicted relations with high significance are similarities between various Rossmann-type folds and between various helix-turn-helix-containing families. The potential value of COMPASS for structure/function predictions is illustrated by the detection of an intricate homology between the DNA-binding domain of the CTF/NFI family and the MH1 domain of the Smad family.

Keywords: sequence similarity searches; profile–profile comparison; sequence profiles; protein structure prediction; CTF/NFI

Introduction

With the rapid growth of the number of known protein sequences, the development of improved automated methods to determine remote sequence similarities becomes increasingly important. The detection of such similarities provides valuable information about the structural and functional relationships between proteins. In the case of a novel protein with unknown structure and function, this information can lead to further characterization of the protein. In the case of a protein family analysis, comparing multiple sequences might provide clues about the structure, function and evolution of the family as a whole.

Current methods for pairwise sequence comparison provide confident detection of similarity between sequences with more than ~30% identity.[1,2] The region of residue identity somewhere between 20% and 35% does not generally allow statistically trustable results of pairwise comparison and is traditionally called the twilight zone.[1,3] A series of successful efforts have been made to improve the inference of remote homologs from protein sequences in the twilight zone. Perhaps the most powerful methods involve the comparison of multiple protein alignments to single sequences or to other multiple alignments. The rationale for the use of multiple alignments is that the information extracted from aligned related sequences may represent general features of the family and allow the prediction of similarity to a remote sequence (or family), even if its similarity to each of the individual aligned sequences is insignificant. The residue composition of the multiple alignment is statistically represented in the form of a numerical profile,[4,5] which is used in further comparison procedures. The methods involving profile–sequence comparisons include several widely accepted searching protocols. PSI-BLAST[6] and IMPALA[7] share the same profile representation and scoring system. PSI-BLAST is an iterative method for sequence database searches with a profile constructed from the hits obtained after the previous iteration step. IMPALA is designed to search a database of profiles with a given sequence. The SAM-T99[8] and HMMER[9]

Abbreviations used: EVD, extreme value distribution; PDB, Protein Data Bank.

E-mail address of the corresponding author: grishin@chop.swmed.edu

packages represent another successful approach to profile–sequence comparison, using the formalism of hidden Markov models (HMM).[10,11]

As a further step in the use of the alignment information, several methods have been developed for the comparison of multiple alignments to multiple alignments. Gotoh[12] introduced iterative methods based on a straightforward but computationally costly sum-of-pairs scoring system, which is used for multiple alignment construction. The protocol of the profile–profile comparison (LAMA) with no gaps permitted was developed by Pietrokovski[13] for the comparisons of pairs of blocks from the BLOCKS database.[14,15] This protocol was further used in the CYRCA method[16] to identify multiple consistently aligned blocks within two compared alignments.

Here, we introduce the COMPASS (comparison of multiple protein alignments with assessment of statistical significance) method, which involves the construction of local profile–profile alignments allowing gaps by means of a dynamic programming algorithm. To our knowledge, two similar methods for the construction of local profile–profile alignments have been reported: FFAS[17] by Rychlewski *et al.*, and prof_sim[18] by Yona & Levitt. The main differences between these two methods include the protocols used to produce profiles from the multiple alignments and the scoring systems used for the alignment construction. To assess the similarity between profile columns, FFAS uses the "dot-product" scores that are related to the correlation coefficients between the amino acid frequencies within the two columns. Prof_sim employs a more sophisticated scoring approach, applying Jensen–Shannon measure for the divergence between two probability distributions and computing two terms that are interpreted as the divergence score and the significance score. These terms are combined to produce a single similarity score.[18] In both methods, the calculated substitution scores for the profile columns are further adjusted by means of simple linear transformations, and a dynamic programming algorithm is applied with optimized gap penalties. To characterize the reliability of the detected similarity, the statistical significance for the produced local alignment is estimated by constructing the empirical score distribution obtained from a number of comparisons between unrelated families,[18] or between the given protein family and other families from the database.[17,18]

Among the methods for detection of sequence similarity, PSI-BLAST is considered one of the most powerful and successful. An important advantage of BLAST and its successors (e.g. PSI-BLAST) is that the statistical significance of the local alignment (*E*-value) allows fast and simple analytical estimation.[6] To our knowledge, no similar analytical estimation of *E*-value has been proposed for profile–profile comparisons. Therefore, we intended to develop a method for constructing local profile–profile alignments, which would be based on a simple generalization of the PSI-BLAST approach to the scoring system and to assessing *E*-value. Our main expectations were (i) to increase the sensitivity and selectivity of the detection of remote similarities between protein groups; (ii) to improve the quality of the produced local alignments; (iii) to search for the previously unknown relationships between the protein families.

When testing COMPASS for the quality of produced alignments and for the detection of remote similarities between protein families, we used alignments of known protein structures as the reference. Thus, our goal was to improve the prediction of similarities between the proteins in the sense of their structural relationship. We compared the performance of COMPASS to that of PSI-BLAST as a method for profile–sequence comparison (the blastpgp program was downloaded from the NCBI site)†, and to that of prof_sim as a method for profile–profile comparison (the prof_sim program was generously provided by Dr G. Yona).

The COMPASS program can be downloaded from our web site‡.

Theory

Several major steps are required to produce a local alignment of two multiple alignments: (i) construction of numerical profiles from the two input alignments; (ii) calculation of scores for matches of positions in the two constructed profiles; (iii) applying an algorithm for aligning the profiles using scores for position matches; (iv) statistical evaluation of the resulting alignment.

Construction of profiles from the alignments

Compensating for sequence redundancy (effective counts)

A profile is a position-specific numerical representation of the residue content of a multiple alignment. For an alignment of length n, the profile is the matrix $n \times 21$, where each column corresponds to a position in the alignment and includes 20 numbers for each type of amino acid residue, plus one number for gap symbols. It is important to avoid situations where a large number of closely related sequences make a greater contribution to the profile than a small number of divergent sequences, which results in losing valuable information about the alignment. A common way to compensate for the redundancy of similar sequences is to down weigh the contributions of the residues from redundant sequences.

To perform such a weighting, we use a method based on the scheme of position-specific independent

† ftp://ftp.ncbi.nih.gov/blast/
‡ ftp://iole.swmed.edu/pub/compass/

counts (PSIC).[19] Residue content at each position is derived not from the overall weights for the sequences of the alignment but from the similarity of the sequence subset, which contains the given residue at the given position. This scheme is implemented as described, with one further modification. We calculate 21 counts $n_{\mathrm{eff}}^{\mathrm{PSIC}}$ for each symbol in the alignment column (including gaps, which are considered the 21st symbol), and then apply the following transformation:[20]

$$n_{\mathrm{eff}} = -\ln\frac{20 - n_{\mathrm{eff}}^{\mathrm{PSIC}}}{20} \quad (1)$$

Here, n_{eff} corresponds to the number of random sequences in the random alignment that has the average number of different residues per position equal to $n_{\mathrm{eff}}^{\mathrm{PSIC}}$ (for more details, see Ref. 20). If all sequences that contain the given symbol at the given position are independent then n_{eff} is equal to the number of these sequences; if they are identical then $n_{\mathrm{eff}} = 1$.

Purging columns with high effective gap content

Aligning regions that include positions with high gap content can present a major problem to the construction of extended local alignments. These gapped regions of the input alignments correspond to insertions in a small portion of sequences and do not reflect general features of the protein family. Including such regions in local alignments is often problematic for the Smith-Waterman algorithm, which tends to stop extending the local alignment rather than to introduce a number of gaps. This tendency would result in a short optimal alignment that would not include many possible regions of high similarity because these regions could not be "linked" in the process of the alignment extension. As a simple way to overcome this problem, only the positions with low effective gap content are considered in the process of the alignment construction. Specifically, if the effective count for gaps is not greater than the sum of the effective counts for all residue types, the position is used for the alignment construction described below; otherwise it is disregarded (in the final output alignment, such positions are marked as "non-aligned").

Estimation of target frequencies

To calculate the scores for position matches, we use the main elements of the scoring system used in PSI-BLAST[6] for profile–sequence comparison and generalize them for the case of profile–profile comparison. To generate scores in a log-odds form, it is necessary to estimate probabilities $\{Q_i\}_1^{20}$ for the residues to be found at a given position (predicted frequencies). The observed residue frequencies $\{f_i\}_1^{20}$ may be biased compared to the expected probabilities $\{Q_i\}_1^{20}$ due to the small effective size of the alignment, and employing information about the correlations between the occurrence of the different amino acid types at aligned positions[6,7,17,18,21–24] has proven useful. For the estimation of the predicted frequencies, we used the simple pseudocount method proposed by Tatusov *et al.*[23], which is implemented in PSI-BLAST[6] and IMPALA.[7] Given the effective frequency f_i of residue type i in a column, we estimate Q_i as the mixture of f_i and pseudocount frequency g_i:

$$Q_i = \frac{\alpha f_i + \beta g_i}{\alpha + \beta} \quad (2)$$

where

$$g_i = \sum_j f_i \frac{q_{ij}}{p_i} \quad (3)$$

(q_{ij} is the matrix of the probabilities of occurrence of residue pairs (i–j) corresponding to the substitution matrix s_{ij}, whereas p_i are the background frequencies of the residues). Parameters α and β determine the proportion of pseudocounts in the mixture. A reasonable setting is $\alpha = N_C - 1$, where N_C is the mean number of different symbols (including gaps) in the columns of the alignment.[6] β remains a free parameter. After testing several values, we found that a good alignment quality is produced by our method with $\beta = 10$, the same value as was initially introduced in PSI-BLAST.[6]

Scoring system

To generate the scores for matching positions of the two constructed profiles, we use the scheme of log-odds ratios, which is reasonable from the theoretical point of view[25] and has been extensively tested in practical applications[6,7,22,23,26–32] The general formula of the score for the match of two profile columns 1 and 2 is as follows:

$$S = c_1 S_{12} + c_2 S_{21} = c_1 \ln\frac{P(1|2)}{P(1|0)} + c_2 \ln\frac{P(2|1)}{P(2|0)} \quad (4)$$

where S_{12} and S_{21} are symmetrical log-odds ratios corresponding to the probabilities of occurrence for columns 1 and 2, respectively; parameters c_1 and c_2 determine the relative weights of both terms. $P(1|2)$ is the probability to observe the set of effective residue counts of column 1, $\{n_i^1\}_1^{20}$, given the set of the target frequencies of column 2, $\{Q_i^2\}_1^{20}$. $P(2|1)$ is the probability to observe the set of effective residue counts of column 2, $\{n_i^2\}_1^{20}$, given the set of the target frequencies of column 1, $\{Q_i^1\}_1^{20}$ $P(1|0)$ and $P(2|0)$ are the probabilities to observe the effective counts $\{n_i^1\}_1^{20}$, and $\{n_i^2\}_1^{20}$, respectively, given the background frequencies of the residues, $\{p_i\}_1^{20}$.

Each of the probabilities $P(a|b)$ can be expressed in the form of multinomial distribution generalized for the non-integer area:

$$P(1|2) = \prod_i C(n_i^1, \{n_j^1\}_1^{20}) \cdot (Q_i^2)^{n_i^1} \quad (5)$$

$$P(1|0) = \prod_i C(n_i^1, \{n_j^1\}_1^{20}) \cdot (p_i)^{n_i^1} \quad (6)$$

$$P(2|1) = \prod_i C(n_i^2, \{n_j^2\}_1^{20}) \cdot (Q_i^1)^{n_i^2} \quad (7)$$

$$P(2|0) = \prod_i C(n_i^2, \{n_j^2\}_1^{20}) \cdot (p_i)^{n_i^2} \quad (8)$$

where $C(n_i^1, \{n_j^1\}_1^{20})$ and $C(n_i^2, \{n_j^2\}_1^{20})$ are generalized multinomial coefficients. Their exact form is not important: since these coefficients are the same within the pair (5) and (6) and pair (7) and (8), they are cancelled after the substitutions made in equation (4). Then equation (4) transforms into:

$$S = c_1 \sum_i n_i^1 \ln \frac{Q_i^2}{p_i} + c_2 \sum_i n_i^2 \ln \frac{Q_i^1}{p_i} \quad (9)$$

The terms of the sum (9) depend on the effective residue counts in columns 1 and 2 (n_i^1 and n_i^2). If $c_1 = c_2$ the scales of these two terms can be very different when the effective counts in columns 1 and 2 are different (e.g. when the two alignments have different "thickness"). In this case the score is mostly determined by the set of target frequencies in one column, with almost no contribution from the other column. In other words, the scores for the matches of columns from the two profiles may turn into the scores for the columns of one of the profiles, which reduces the quality of the alignment. To balance the terms in the sum (9), we tested a number of expressions for the coefficients c_1 and c_2. In addition to the ability to compensate for the possible different scales of $\{n_i^1\}_1^{20}$ and $\{n_i^2\}_1^{20}$, we demanded that the setting of c_1 and c_2 should transform equation (9) into the PSI BLAST score $S = \ln(Q_i/p_i)$ in the special case of sequence-alignment comparison. The best alignment quality was provided by the following setting:

$$c_1 = \frac{\sum_i n_i^2 - 1}{\sum_i n_i^1 + \sum_i n_i^2 - 2}, \qquad c_2 = \frac{\sum_i n_i^1 - 1}{\sum_i n_i^1 + \sum_i n_i^2 - 2} \quad (10)$$

If one of the compared alignments contains only a single sequence then equation (9) with coefficients (10) reduces to the PSI-BLAST score. In the process of alignment construction, the score values rounded to the nearest integer are used. To increase the computational precision, scores (9) and (10) are multiplied by a factor of 32. The scores are then rescaled as described below, rounded to integer values and used to construct the optimal local alignment with the Smith & Waterman algorithm.[33]

Construction of optimal alignment and estimation of its statistical significance

Calculation of E*-value*

After generating the substitution scores for position matches, we rescale them to ensure correspondence to some standard setting universal for all profile pairs. This procedure simplifies the estimation of statistical significance of optimal alignments (i.e. their *E*-values). For the calculation of *E*-value, we (a) rescale the optimal scores produced by our method to obey the extreme value distribution (EVD),[34,35] and (b) use the simple formula proposed by Karlin & Altshul[36] for the case of EVD:

$$E = Kmne^{-\lambda S} \quad (11)$$

where m and n are lengths of the two profiles (in the case of a database search, the lengths of the query and the database), S is the score of the optimal alignment, and λ and K are statistical parameters of EVD, which depend on the scoring system and on the profiles that are compared. Since λ enters the expression exponentially, it is the key parameter that has the most substantial influence on the *E*-value.

Expression (11) was initially proposed for the alignments of single sequences without gaps.[36] It implies very wide assumptions about the nature of the sequences and the substitution scores. The theory[37,38] can be applied to the alignments of two random strings 1 and 2 composed of symbols a_i that are independently sampled from a finite alphabet $A = \{a_i\}_1^N$ with probabilities $\{p_i^1\}_1^N$ and $\{p_i^2\}_1^N$, respectively. Each match $(a_i - a_j)$ has the score s_{ij}. These assumptions are valid in the case of the comparison of two profiles, with a_i being the profile columns. In practice, the comparisons are made within a large but finite database of profiles that contain a large but finite alphabet of columns. Moreover, the distribution of profile columns in multiple protein alignments has a distinct structure,[22] with the tendency to accumulate in a low-dimensional region of the space of frequency sets. Thus, the effective size of the alphabet of columns may be much lower than the total number of columns in the database.

The application of formula (11) requires two conditions: at least one score should be positive, and the expected score per column pair should be negative.[36] Both of these conditions are fulfilled in the vast majority of profile–profile comparisons by our method. (Theoretically possible exceptions might occur, for example, when scores for all pairs of columns from the two profiles have the same sign. According to our observations, such exceptions are extremely rare.) Thus, if two profiles are constructed from columns randomly sampled from a finite set, the *E*-value for their local ungapped alignment allows the analytical expression (11).

The maximal segment theory[36–38] predicts the form of the optimal score distribution only for an ungapped scoring system. As has been empirically shown for the gapped scoring systems, the distribution of optimal scores for gapped alignments can also be approximated by EVD, however with different values λ and K. This was demonstrated for both sequence–sequence[6] and sequence–profile[6,7] comparisons. A similar observation was made by Yona & Levitt [18] when they applied their prof_sim method to the comparison of unrelated real profiles. Further in this section, we will show that our method also allows the use of EVD to describe the distribution of optimal scores for "random profiles".

Although the EVD parameters λ_u and K_u for ungapped scoring systems can be obtained from analytical equations,[36] the problem of estimation of these parameters for a given gapped scoring system (λ_g and K_g) is not analytically solved. λ_g and K_g may be precomputed by extensive simulations using the average residue composition of the database.[6,39,40] However, in the case of sequence–profile comparisons, unusual amino acid compositions of the sequence or the profile may imply different values of λ_g.[6,39] The profile–profile substitution scores (9) and (10) additionally depend on the effective residue counts in the columns. Therefore λ_g for the scores (9) and (10) may vary with both residue composition and effective number of sequences in the alignments.

To avoid random simulations and estimation of λ_g for each profile–profile pair, we adopt the score rescaling strategy similar to that used in PSI-BLAST and IMPALA. The aim of this technique is to "force" the substitution scores to a fixed scale. The easily calculated parameter λ_u is used as a measure of the score scale. For each profile–profile pair, the substitution scores are rescaled to achieve $\lambda_u = \lambda_u^0$, where λ_u^0 is the precomputed value for the reference sequence–sequence scoring system. Introducing gaps into such a rescaled scoring system will change λ_u^0 to the precomputed gapped parameter λ_g^0, as it would happen for the reference scoring system. Random simulations of sequence–profile comparisons support this assumption.[6,7] This rescaling technique remarkably increased the sensitivity of PSI-BLAST.[39]

For the sequence–sequence comparison, λ_u can be found as the unique positive solution of the equation:[36]

$$\sum_{i,j} p_i^1 p_j^2 \, e^{\lambda s_{ij}} = 1 \qquad (12)$$

where p_i^1 and p_j^2 are the frequencies of the residue types in the two sequences, s_{ij} is the matrix of the substitution scores for the residue pairs. PSI-BLAST and IMPALA apply expression (12) directly to the case of the sequence–profile comparison, considering p_i^1 and p_j^2 as the residue frequencies in the sequence and the profile, and s_{ij} as the substitution score for the pair of residues i and j (e.g. BLOSUM62). λ_u derived from equation (12) depends only on the residue composition of the profiles or sequences. The additional dependence of scores (9) and (10) on the effective residue counts makes the direct use of expression (12) meaningless, since λ_u does not correlate with the scale of the scores anymore. To estimate the composition-dependent λ_u in the context of the substitution scores (9) and (10), we use a more general form of equation (12):

$$\langle e^{\lambda s} \rangle = 1 \qquad (13)$$

where s are the scores for all possible pairs of positions in the two sequences, and brackets denote averaging over all such pairs. Presenting two profiles as two sequences of columns, we can find λ_u as the unique positive equation of (13) in the form:

$$\frac{1}{l_1 l_2} \sum_{a,b} e^{\lambda s_{ab}} = 1 \qquad (14)$$

where l_1, l_2 are the lengths of the two profiles; a, b are positions in the profiles; s_{ab} are the scores (9) and (10) for each pair of the positions. Thus summation over the types of symbols in equation (12) is changed in equation (14) to the equivalent summation over the positions, which can be performed for a given profile pair.

After estimating λ_u, the scores s_{ab} are rescaled to the values $s'_{ab} = s_{ab}\lambda_u/\lambda_u^0$, where λ_u^0 is the known value for the reference scoring system (as the default we use $\lambda_u^0 = 0.3176$ for BLOSUM62). To produce the optimal local alignment of the given profiles, we use the scores s'_{ab} and apply the Smith–Waterman algorithm with affine gap penalties $(11 + k)$, the optimized gap penalties for the BLOSUM62[6] substitution matrix. As we found empirically, these gap costs produced reasonable quality of the local alignments by our method.

The whole approach to the E-value calculation described in this section is valid only if the distribution of the optimal alignment scores allows the EVD approximation. In the next section, we will show the experimental results that support this assumption and allow estimation of λ and K for a given pair of profiles.

Estimation of the parameters of the optimal score distribution

In order to characterize the form of the optimal score distribution for the comparison of random profiles, we constructed a large number of profiles from randomly sampled columns derived from real alignments. Their optimal local alignments were produced as described above, and the distributions of the alignment scores were analyzed. These distributions can be reasonably well approximated with EVD and allow estimation of E-value for the alignment of a given pair of profiles.

We expected the parameters of the optimal score distribution to depend on two main properties of

the compared alignments, the length (number of columns, l) and the thickness (a characteristic of the effective number of divergent sequences; we used the average sum of effective residue counts for all columns, N). To study these dependencies, we constructed 16 datasets of "random alignments" for 4×4 different combinations of the lengths (100, 200, 330, and 500 columns) and the number of sequences (50, 100, 300, and 500 sequences). To sample the alignment columns, we used the alignments from the PFAM 6.6 database and extracted the columns with the effective gap content lower than the 50% cutoff.

For each combination of the alignment lengths and thicknesses, we produced 10,000 optimal alignment scores. The distribution of the scores was fitted to EVD using the maximum likelihood method.[41] The observed distributions were reasonably well approximated by EVD. Figure 1 shows a typical example of a fit for random alignments; a chi-square goodness-of-fit test in this case produced a value of 53.9 for 42 degrees of freedom. This value indicates that the given fit would be better than 10% of fits for the similar random trials, even if the theory for the ungapped sequence–sequence alignments were fully applicable to the gapped profile–profile alignments described above.

Having the estimates of the EVD parameters (λ and K), we studied their dependence on the properties of the compared alignments (N_1, l_1, N_2, l_2). Since the E-value exponentially depends on λ (equation (11)), a good precision in the estimation of this parameter is far more important than that of K; yet for completeness, we present the results for both parameters.

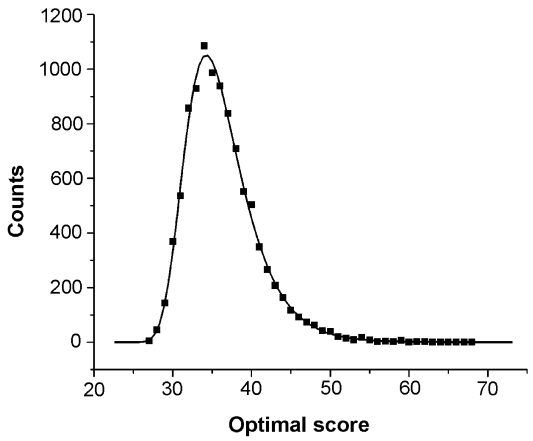

Fig. 1. Distribution of optimal local alignment scores produced for 10,000 pairs of alignments composed from randomly sampled columns of PFAM alignments (see the text for details). The score distribution produced by COMPASS for length 500 and number of sequences 300 is shown. The extreme value distribution that best fits the data is plotted. The chi-square goodness-of-fit test in this case produced a value of 53.9 for 42 degrees of freedom, which corresponds to a P-value of 0.10.

For sequence–sequence[42] comparisons, it has been shown that λ depends on the length of the sequences. This dependence allows a close approximation:[40,42]

$$\lambda = \lambda_0 + \alpha\left(\frac{1}{m} + \frac{1}{n}\right) \tag{15}$$

We assumed that in the case of profile–profile comparison the general form of expression (15) might also be applied to the length-dependence of λ. We directly introduced reasonable approximations of the observed values λ and K into the E-value formula (11). Thus for the alignments of two fixed thicknesses, we sought estimation of λ in the form:

$$\lambda = \lambda_0(N_1, N_2) + \alpha(N_1, N_2)\left(\frac{1}{l_1} + \frac{1}{l_2}\right) \tag{16}$$

where N_1 and N_2 are the average sums of effective residue counts in the columns, l_1 and $l_2$are the lengths of the profiles. To estimate $\lambda_0(N_1, N_2)$ and $\alpha(N_1, N_2)$, one can analyze the score distributions for the alignments of equal length and consider the dependence $\lambda = \lambda_0(N_1, N_2) + 2\alpha(N_1, N_2)/l$.

Dependence of λ and K on the effective number of sequences in the alignment

For fixed profile lengths, we studied the dependencies of λ and K on the thickness of the compared alignments, N_1 and N_2. Surprisingly, using the settings of our method described above, no considerable dependence was observed for any tested alignment length. The difference between the values was within 3% among all dataset combinations of various thickness for a given length l (data not shown).

Dependence on the length of alignments

Given no or little dependence of λ and K on N_1 and N_2, our task to approximate them in the form (16) was reduced to the form (15). The corresponding coefficients derived from the observed dependencies (Figure 2(a) and (b)) were $\lambda_0 = 0.277(\pm 0.002)$, $\alpha_\lambda = 2.25(\pm 0.15)$ and $K_0 = 0.044(\pm 0.005)$, $\alpha_K = 7.40(\pm 0.45)$, which allows the following simple approximations:

$$\lambda = 0.277 + 2.25\left(\frac{1}{m} + \frac{1}{n}\right) \tag{17}$$

$$K = 0.044 + 7.40\left(\frac{1}{m} + \frac{1}{n}\right) \tag{18}$$

For the case of sequence–sequence comparison, Altschul *et al.* [40]estimated λ as $\lambda = 0.2670 + 1.90(1/m + 1/n)$. According to the E-value formula (10), the 3.6% difference in the estimates of λ_0 for the sequence–sequence and profile–profile comparisons would lead to a fairly modest ~2.5-fold

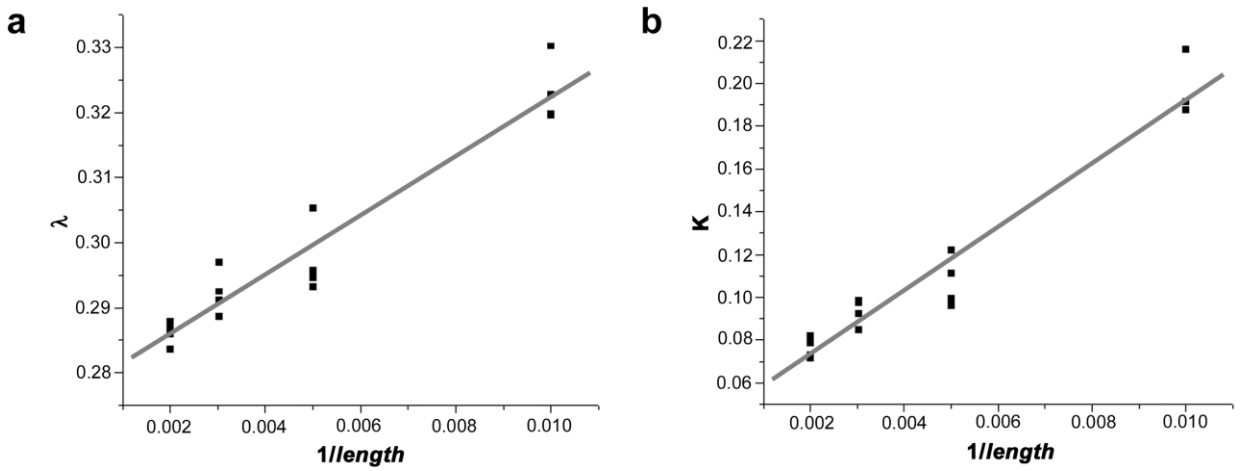

Fig. 2. Estimates of λ and K derived from comparisons of the random alignments (composed from randomly sampled PFAM columns) plotted as functions of 1/length, where length is the length of alignments. For a given length, four sets of random alignment pairs were prepared from randomly picked columns of PFAM alignments. Each set contained 10,000 pairs of alignments with a given number of rows (50, 100, 300 or 500). The distribution of optimal local alignment scores produced for each set was used to derive the estimates of λ and K. The plotted lines represent the linear approximations of the dependencies. The deviation of λ values from the line did not exceed ~2%.

difference in the E-values for typical marginally significant hits with $\lambda S \sim 25$.

Results and Discussion

Evaluation of alignment quality

To evaluate the performance of our method, we tested two aspects of its performance, the ability to produce accurate local alignments and the ability to detect profiles in a database that are related to the query. In both cases, we based our evaluation on the comparison of the produced alignments to the structural alignments from the FSSP database[43,44] generated by the DALI method.[45] The results of the evaluation were compared to the results for the corresponding sequence–profile comparisons by PSI-BLAST, and to the results for profile–profile comparisons by prof_sim[18]

Benchmark for the evaluation of the alignment quality

To create 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 FSSP. We used these pairs of multiple alignment as input for COMPASS, prof_sim and PSI-BLAST, and compared the predicted local alignments with the structural alignments in FSSP.

To assess the performance of each method for close, intermediate and far relationships, we generated datasets of sequence pairs for three ranges of sequence identity (according to the FSSP alignments), 0–15%, 15–30% and 30–98%. The values of 15% and 30% were chosen as boundaries based on our preliminary observations of COMPASS performance over the whole range of sequence identity (see below). From randomly chosen 500 FSSP families, we extracted the parent sequence and three random family sequences of a significant structural similarity to the parent (Z-score greater than 5.0), with sequence identity to the parent within the three ranges. 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 thus obtained 500 pairs of PSI-BLAST alignments for each identity range. In order to ensure that the sequences in each PSI-BLAST alignment were closer to the initial query than to its "partner" in the pair, we purged all the sequences whose similarity to the query (PSI-BLAST score) was lower than that of the query's partner. The sequences common to both alignments were purged in the alignment where the sequence had the lower PSI-BLAST score. Finally, 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 pairs of alignments were used for the construction of their local alignment and its evaluation.

Alignment quality evaluation parameters

To measure the quality of the prediction of a structural alignment by the corresponding profile–profile alignment, we used the parameters (Q_{modeler}, $Q_{\text{developer}}$ and Q_{combined}) proposed earlier.[18,46] The quality from the modeler's point of view (Q_{modeler} in the notation of Yona & Levitt[18]) is the ratio of the number of correctly aligned positions to the total number of positions in the evaluated alignment. The quality from the developer's point of view ($Q_{\text{developer}}$[18]) is the ratio of the number of

correctly aligned positions to the number of positions in the structural alignment. To calculate the "combined" quality ($Q_{combined}$[18]), the number of correct matches is divided by the total number of positions that are aligned in either the structural alignment or evaluated alignment.

For local alignments, it is reasonable to assess the local prediction for only the regions of the structural alignment that are included in the evaluated alignment as opposed to the prediction of the whole structural alignment. Thus in addition to $Q_{developer}$, we introduced two measures of "local accuracy". Q^1_{local} is defined as $Q^1_{local} = N_{acc}/L$, the ratio of the number of correctly aligned positions N_{acc} to the length L of the region in the structural alignment that includes the pairs of profile positions from the alignment under evaluation. Q^2_{local} is a modification of Q^1_{local}, which takes into account slight shifts between the positions aligned in the evaluated and the reference alignment. To calculate Q^2_{local}, we consider all the matches in the part of the structural alignment that correspond to the alignment under evaluation. For each pair of positions aligned by DALI, we find the shift Δ_i (the number of positions dividing the pair) that is introduced by the evaluated alignment. The definition of Q^2_{local} is $Q^2_{local} = \sum_i (0.5)^{\Delta i}/L$, where summation is made over all the position pairs from the structural alignment that are included into the evaluated alignment. Slight shifts of one to two positions would make some additional contribution to Q^2_{local}, whereas the contribution of the positions shifted by $\Delta > \sim 5$ would be $\geq \sim 10^2$ times lower than that of the correct matches.

Both measures Q_{local} are 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", which is independent of the accuracy of the prediction. To calculate the coverage (Q_{cov}), we determined the length of the region in the structural alignment that includes all the positions from the evaluated alignment and divided it by the overall length of the structural alignment.

Quality of the alignments by COMPASS compared to those by PSI-BLAST and prof_sim

To estimate the boundaries of zones with low, intermediate and high sequence similarity, we split the whole identity range into small bins and observed plots for Q_{local} and Q_{cov} (data not shown). The growth of these measures between 0% and 30% is roughly divided into two phases: a rapid growth phase (between 0% and 15% identity), and a slower phase of reaching the plateau (between 15% and 30% identity). Based on this observation, the identities of 15% and 30% were chosen as the delimiters between the zones of low, intermediate and high similarity.

Figure 3 shows the average quality measures for the alignments obtained with different methods for two zones of identity, 0–15% and 15–30%. The profile–profile local alignments were obtained by submitting the pairs of benchmark alignments to prof_sim (program was generously provided by Dr G. Yona) and COMPASS. The profile–sequence PSI-BLAST alignments were obtained by submitting the full alignment 1 and the first (query) sequence of alignment 2 from each pair in the datasets. As initial reference of the alignment quality, we used the Smith–Waterman alignments of single query sequences from each benchmark pair.

The quality of the alignments produced by COMPASS was higher than that of the alignments produced by prof_sim and PSI-BLAST. Compared to prof_sim, COMPASS produced slightly lower coverage (Q_{cov}), which may indicate a tendency for slightly shorter alignments. Combined with the increase of the local accuracy (Q_{local}) and of the integral quality parameters ($Q_{modeler}$, $Q_{developer}$, $Q_{combined}$), this reduction of the coverage suggests that the local alignments produced by COMPASS tend to include less spurious matches. The increase of Q_{local}, $Q_{modeler}$, $Q_{developer}$ and $Q_{combined}$ for COMPASS compared to other methods was more pronounced in the low range of identities (0–15%). In this range, the increase of the quality parameters for COMPASS compared to PSI-BLAST was approximately the same as the increase for PSI-BLAST compared to the sequence–sequence alignment (Figure 3).

Evaluation of the ability to detect remote similarity between profiles

Evaluation protocol

As a criterion for assessing COMPASS as a potential profile-based searching method, we chose the ability to predict structural relationships between protein domains. For the evaluation of the profile–profile similarity detection, we used the largest available source of accurate semi-automatic multiple sequence alignments, the PFAM database[47] (version 6.6), and the largest available source of automatic structural alignments, the FSSP database[43,48] (update of December 2001). We collected all PFAM alignments containing at least one sequence that belonged to an FSSP family. Within the resulting database of 1354 PFAM alignments, we performed exhaustive prof_sim and COMPASS searches, with each alignment as a query.

In order to perform PSI-BLAST searches, we prepared the database of all 311,753 sequences extracted from these alignments. In this database, we ran 1354 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

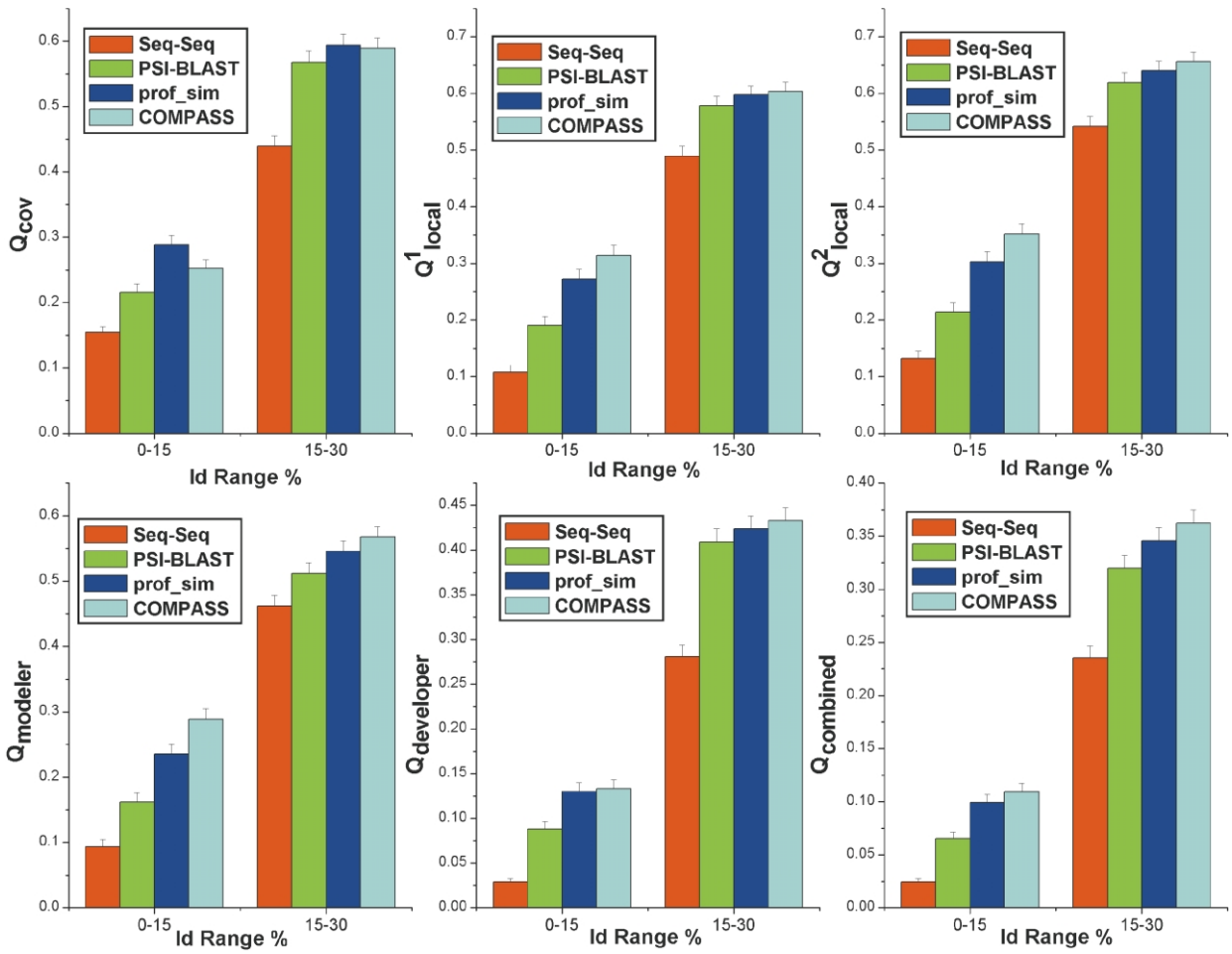

Fig. 3. Evaluation of alignment quality. In two ranges of sequence identity, the quality of the produced local alignments was assessed by six parameters for four different alignment methods (pairwise sequence alignment using BLOSUM62 matrix and Smith-Waterman algorithm, profile–sequence alignment using PSI-BLAST, profile–profile alignments using prof_sim and COMPASS). As a reference, DALI structural alignments from the FSSP database were used. Q_{cov} corresponds to the portion of the length of the structural alignment that was covered by the sequence alignment, regardless of the actual accuracy. Q^1_{local} and Q^2_{local} correspond to the accuracy of the local prediction for only the regions that are included in the evaluated alignment. $Q_{modeler}$, $Q_{developer}$ and $Q_{combined}$ are previously suggested measures of integral accuracy of the alignment from the modeler's, developer's and combined points of view (see the text for details). Means + standard errors are shown.

and the maximal *E*-value were both set to 10,000). As a result, a list of hits from the sequence database was produced for each query alignment. Then we transformed each list of the sequence hits into the list of similarities between alignments. For a given alignment, we compared the PSI-BLAST *E*-values for all sequences from this alignment, and assigned the best *E*-value to the similarity between this alignment and the query. Using the corresponding best sequence–profile alignment as template, the profile–profile alignment was constructed and further assessed.

For each method, all found similarities between alignment pairs were pooled together and ranked by their *E*-value. The prof_sim program generated only *P*-values for each pair of profiles; in order to obtain an estimate of *E*-value, *P*-values had to be multiplied by the number of profiles in the database (G. Yona, personal communication). To determine whether a hit was a true or a false positive, we used one of the several criteria described below. Having the ranked lists of true and false positive hits for each criterion, we generated and compared sensitivity plots for different methods.

Evaluation criteria

To evaluate a hit as a true positive, we required that it should be consistent with the structural relationship in FSSP. In order to test for this consistency, from both the "query" and the "hit" alignments we extracted sequences that belonged to FSSP. (If a PFAM alignment contained more than one sequence from FSSP, we chose a single representative sequence closest to the top of the alignment.) The "relaxed" evaluation criterion for a true positive hit demands that the FSSP entry

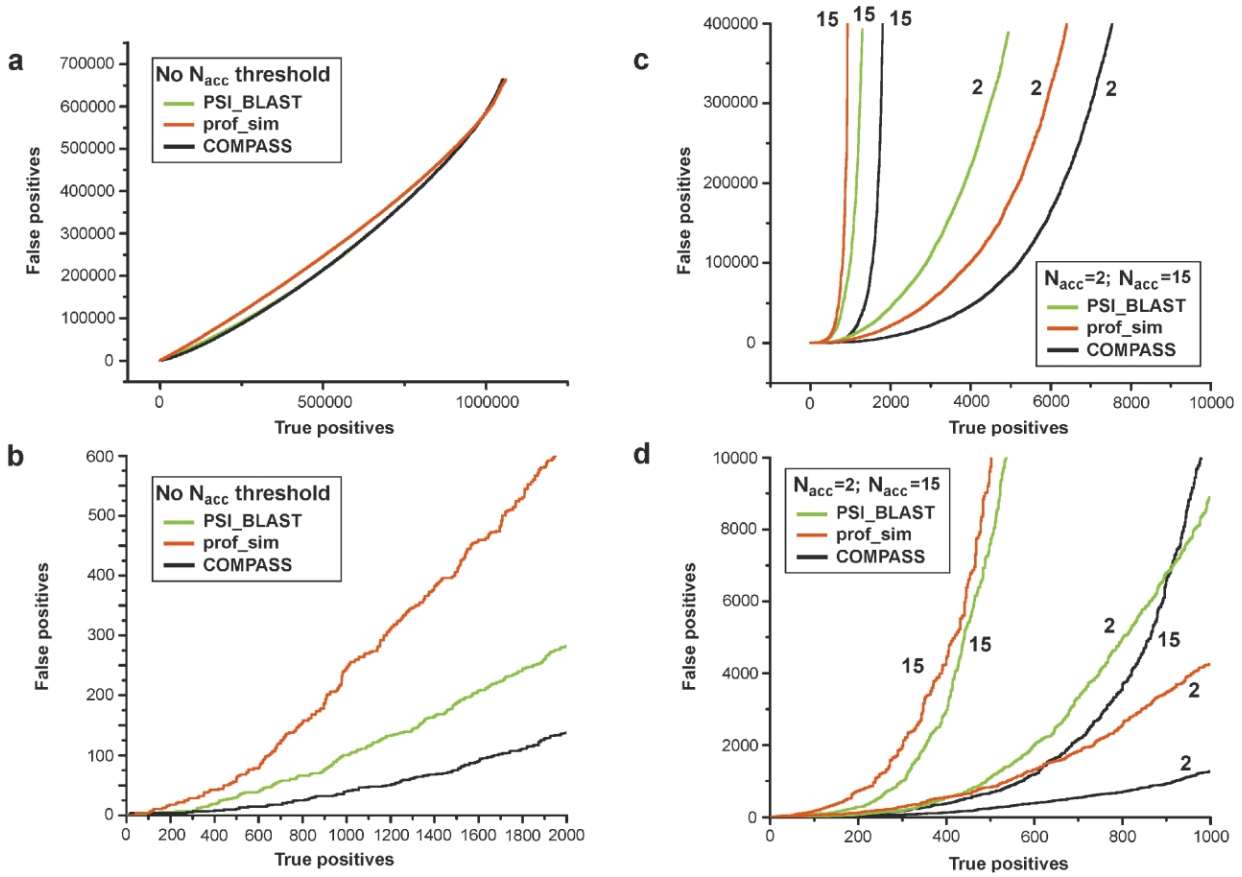

Fig. 4. Sensitivity curves of PSI-BLAST, prof_sim and COMPASS for different criteria of true positive hit assignment. (a) and (b) The curves for the relaxed criterion requiring that the sequences from the two alignments share the same FSSP family. (a) The full-scale graph for the whole experiment (at this scale, the PSI-BLAST and COMPASS curves are close). (b) The same curves for the first 2000 true positives. (c) and (d) The curves for the stringent criteria requiring accurate prediction of $N_{acc} = 2$ and $N_{acc} = 15$ matches. (c) The full-scale graph. (d) The same curves for the first 1000 true positives.

from the hit should belong to the same FSSP family as the entry from the query, with the structural similarity Z-score > 2.0. Among 1354 PFAM alignments that contained FSSP entries, only 56 could not be linked to any other alignments in this way; the remaining 1298 alignments included FSSP sequences that shared a family with at least one sequence from other alignments of the dataset.

A more stringent set of criteria requires a certain level of consistency between the predicted alignment and the alignment derived from FSSP. To implement these criteria, we produced the local alignment of the FSSP entries extracted from the query and the hit, based on the profile–profile alignment for the query and the hit. The resulting local alignment of the two sequences was compared to the alignment of these sequences in FSSP.

To consider the hit a true positive, we demanded that the alignment produced by the searcher should correctly predict al least N_{acc} matches in the FSSP alignment that has Z-score > 2.0. Variation of the required N_{acc} determines the stringency of the criterion. We will present the results for $N_{acc} = 2$ (more than one match consistent with the structural alignment) and $N_{acc} = 15$ (at least 15 consistent matches, which may correspond roughly to a correctly predicted element of secondary structure).

This type of criteria for the true/false positive hit evaluation demands a modest level of alignment accuracy. In many cases, the sequence fragments aligned by a search program were not considered as reliably aligned by DALI (they were represented in lowercase letters in the FSSP alignment). We treated these regions as problematic for DALI and excluded such "indecisive" cases from the sensitivity plots. To compare the alignment produced by a search program to the corresponding FSSP alignment, we used the position numbering provided for the two sequences by PFAM and FSSP. In a number of cases, the starting positions of the PDB sequences indicated in the PFAM alignment were not consistent with those in the FSSP alignment; such hits were disregarded. Since we did not find any specific pattern of such inconsistencies, we assumed that disregarding these cases did not favor any particular searching method.

Table 1. Similarities that were accurately predicted with low *E*-value by PSI-BLAST and received high *E*-value by COMPASS

PFAM name1	PDB ID1	PFAM name2	PDB ID2	CMPSS *E*-value	N_{acc}	Minimal PSI-BLAST *E*-value: 1 *versus* 2	N_{acc}	Minimal PSI-BLAST *E*-value: 2 *versus* 1	N_{acc}
DIM1	1qgvA	thiored	1mek	2.53×10	39	9.00×10^{-4}	33	1.90×10^{-1}	44
B12-binding	2reqA	TMP-TENI	2tpsA	1.77	18	9.00×10^{-3}	17	1.76×10^{2}	0

Similarities between PFAM families that were accurately predicted ($N_{acc} \geq 15$) by PSI-BLAST with minimal *E*-value < 0.01, and were assigned *E*-value > 1.0 by COMPASS. PFAM names for the alignments and PDB Ids for their representatives with solved structure are indicated, along with *E*-value estimates by COMPASS (CMPSS *E*-value), and minimal *E*-values produced by PSI-BLAST for the first (1 *versus* 2) and the second (2 *versus* 1) alignments used as query. Numbers of accurate matches in the local alignments (N_acc) are shown after corresponding *E*-values.

Performance of COMPASS as a search program compared to prof_sim and PSI-BLAST

Sensitivity curves (plots of the number of true positives *versus* the number of false positives) were constructed for the searches by COMPASS, prof_sim and PSI-BLAST using the hit evaluation criteria described above. The results for the relaxed criterion (requiring the correct prediction for sharing the same FSSP family) are shown in Figure 4(a) and (b). The number of false positives produced by COMPASS was considerably lower than those produced by other methods. For the top 1000 true positive hits, COMPASS generated

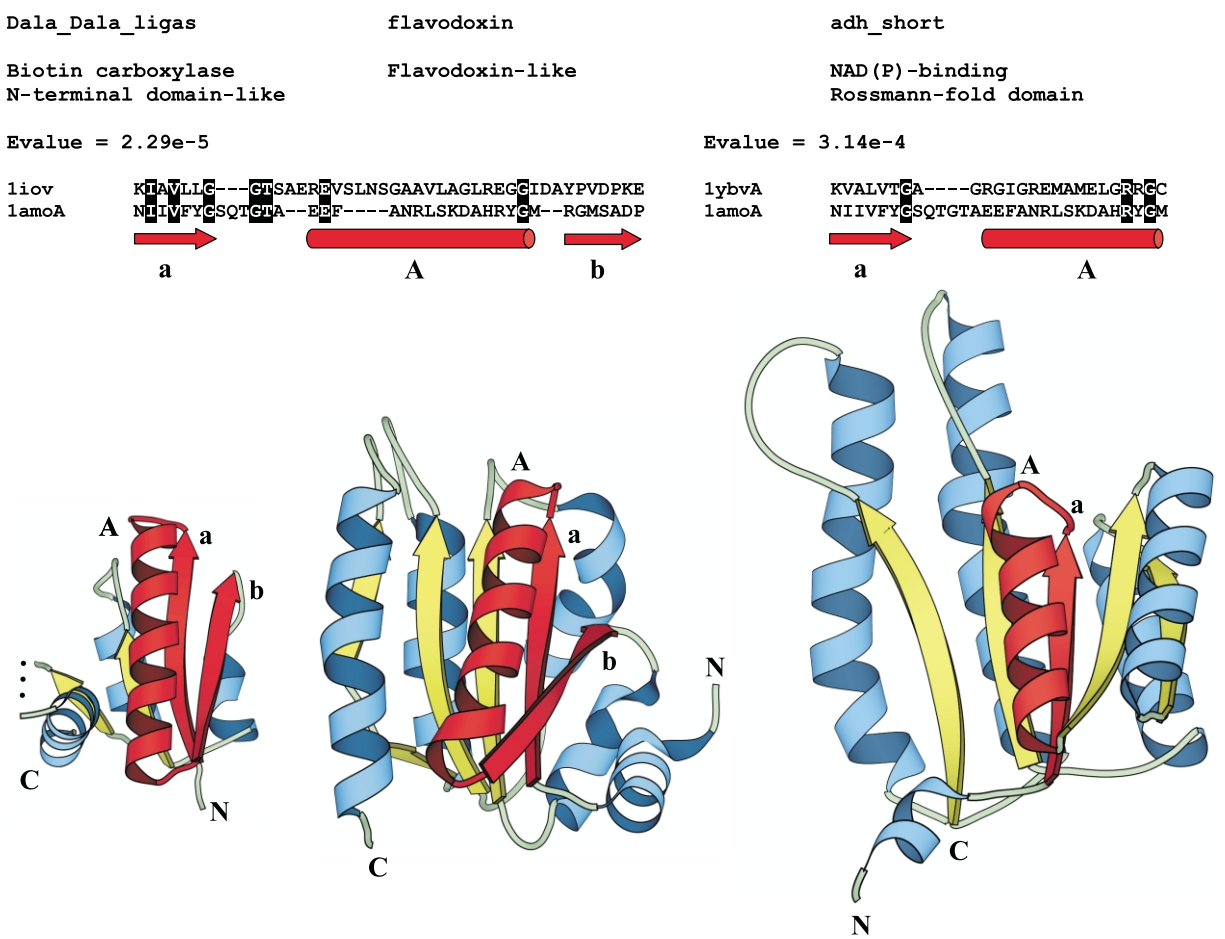

Fig. 5. Examples of predicted similarities between Rossmann-type folds. The structures of the representatives of three PFAM families are shown, along with the sequence alignments produced from two COMPASS hits (invariant residues are boxed with black). Names of the PFAM alignments, corresponding protein folds and PDB Ids of the representative structures are indicated. Below the sequence alignments, conserved secondary structure elements are shown. The α-helices and β-strands are displayed as arrows and cylinders, respectively. The ribbon diagrams of the three domains (PDB Id 1iov, residues 1–87, 292–306; PDB Id 1amoA, residues A66–232; PDB Id 1ybvA, residues A20–199) were drawn by MOLSCRIPT.[63] The regions of predicted similarity in the protein structures are highlighted in red, the labels of secondary structure elements corresponding to those shown below the alignments. Other secondary structure elements are colored in blue (α-helices) and yellow (β-strands).

Table 2. Similarities that were accurately predicted with low *E*-value by COMPASS and received high *E*-value by PSI-BLAST

PFAM name1	PDB ID1	PFAM name2	PDB ID2	CMPSS *E*-value	N_acc/ L_algn	Minimal PSI-BLAST *E*-value: 1 *versus* 2	N_acc	Minimal PSI-BLAST *E*-value: 2 *versus* 1	N_acc
Helix-turn-helix motifs									
HTH_5	1smtB	HTH_7	1hcrA	6.59×10^{-8}	22/31	1.30×10	2	2.65×10^{2}	14
HTH_3	1adr	HTH_7	1hcrA	2.38×10^{-6}	23/33	6.30	26	5.20×10	15
HTH_7	1hcrA	lacI	1pru	3.75×10^{-6}	21/22	1.17×10^{3}	0	5.90	19
HTH_7	1hcrA	HTH_8	1ntcA	4.56×10^{-6}	20/21	3.58×10^{2}	14	1.30	20
RB_B	1guxB	transcript_fac2	1volA	9.95×10^{-5}	54/62	1.30	6	1.10	25
HTH_8	1ntcA	ModE	1b9nB	2.51×10^{-4}	31/33	4.20	9	3.30×10	14
Arg_repressor	1b4aA	HTH_5	1smtB	2.80×10^{-4}	25/62	1.70	0	4.19×10^{2}	11
HTH_8	1ntcA	HTH_AraC	1d5yA	3.42×10^{-4}	31/42	1.40×10	0	6.40	15
HTH_7	1hcrA	Trp_repressor	1troG	1.16×10^{-3}	20/32	1.87×10^{2}	12	3.20×10	12
E2F_TDP	1cf7B	HTH_5	1smtB	1.70×10^{-3}	17/25	1.68×10^{3}	0	3.35×10^{2}	0
HTH_7	1hcrA	gntR	1e2xA	8.21×10^{-3}	26/35	5.34×10^{2}	10	1.70×10	20
Crp	1berB	gntR	1e2xA	4.66×10^{-5}	29/29	1.30	28	2.40	28
Fe_dep_repress	1ddnA	GerE	1a04A	3.06×10^{-4}	16/27	4.30	22	1.30×10	26
GerE	1a04A	lacI	1pru	2.33×10^{-3}	22/22	2.90×10	21	2.00×10	13
Arg_repressor	1b4aA	Fe_dep_repress	1ddnA	2.75×10^{-3}	20/41	2.20×10	8	1.01×10^{3}	2
Rossmann-like folds (*beta-alpha-beta units*)									
AAA	1g3iV	CbiA	1a82	$5.91 \times 10 - 13$	17/27	3.20	0	2.30	16
Amino_oxidase	1b37A	GMC_oxred	1gpeA	7.11×10^{-8}	16/29	3.90×10	23	4.60×10	22
ArsA_ATPase	1f48A	SKI	1shkB	7.43×10^{-8}	19/36	6.90×10	21	5.20×10	23
AAA	1g3iV	fer4_NifH	1n2cE	9.70×10^{-8}	15/22	1.30×10	16	2.97×10^{2}	0
GMC_oxred	1gpeA	adh_short	1ybvA	1.16×10^{-7}	22/29	4.90	14	7.90	6
GMC_oxred	1gpeA	transketolase_C	1qs0B	8.01×10^{-7}	15/34	4.76×10^{2}	0	5.49×10^{2}	0
GMC_oxred	1gpeA	UDPG_MGDP_dh	1dliA	9.76×10^{-7}	15/29	4.10×10^{2}	8	2.46×10^{2}	0
DAO	1daoA	Octopine_DH_N	1bg6	1.16×10^{-6}	23/24	3.60×10	21	2.40	23
FAD_binding_3	1d7lA	Octopine_DH_N	1bg6	1.66×10^{-6}	26/26	6.90	19	1.10	32
DAO	1daoA	Epimerase	1bxkA	3.41×10^{-6}	18/21	5.90	16	1.50	14
Beta_elim_lyase	2tplA	Cys_Met_Meta_PP	1cs1A	6.64×10^{-6}	106/226	3.90	34	2.80	12
AdoHcyase	1a7aB	UDPG_MGDP_dh	1dliA	7.03×10^{-6}	37/85	7.50×10	7	2.54×10^{2}	1
FGGY	1glaG	HSP70	1hjoA	8.55×10^{-6}	16/16	1.50	20	1.20×10	19
CheR	1bc5A	adh_zinc	1dehA	1.12×10^{-5}	36/148	3.30×10	15	9.10×10	0
Ras	1ek0A	recA	2reb	1.31×10^{-5}	15/51	1.15×10^{2}	18	9.20	6
SIS	1moq	adh_short	1ybvA	1.52×10^{-5}	16/22	3.10×10	0	4.30	4
Asparaginase	1djoA	ECH	1dciA	1.99×10^{-5}	35/97	8.50×10	21	5.56×10^{2}	27
AlaDh_PNT	1pjbA	PALP	2wsyB	2.29×10^{-5}	34/152	6.20×10	0	2.65×10^{3}	0
Dala_Dala_ligas	1iov	flavodoxin	1b1cA	2.29×10^{-5}	20/44	7.13×10^{2}	0	8.65×10^{2}	0
AlaDh_PNT	1pjbA	Semialdhyde_dh	1brmA	2.66×10^{-5}	28/131	1.90×10	8	9.80	0
adh_short	1ybvA	tRNA-synt_1d	1bs2A	4.40×10^{-5}	24/55	1.06×10^{2}	17	1.10	0
GLFV_dehydrog	1aup	Octopine_DH_N	1bg6	5.00×10^{-5}	16/16	7.80×10	27	6.00	34
AlaDh_PNT	1pjbA	GMC_oxred	1gpeA	5.86×10^{-5}	18/32	1.70×10^{2}	19	1.50×10	21
Lipase_3	1lgyA	abhydrolase_2	1auoA	6.15×10^{-5}	15/15	1.54×10^{3}	0	1.30×10	0
GLFV_dehydrog	1aup	UDPG_MGDP_dh	1dliA	6.34×10^{-5}	24/36	1.70	0	1.60×10	0
DLH	1din	Peptidase_S9	1e5tA	7.41×10^{-5}	24/30	2.90	11	2.50	21
Acyl_transf	1mla	abhydrolase_2	1auoA	8.83×10^{-5}	16/16	3.03×10^{3}	16	8.40×10	1
GMC_oxred	1gpeA	Octopine_DH_N	1bg6	8.91×10^{-5}	16/16	2.96×10^{2}	12	9.80×10	16
Flavodoxin	1b1cA	tRNA-synt_1d	1bs2A	1.31×10^{-4}	16/49	3.85×10^{2}	0	8.20×10	0
Epimerase	1bxkA	GMC_oxred	1gpeA	1.79×10^{-4}	27/34	3.87×10^{2}	0	1.42×10^{2}	0
GARS_N	1gsoA	adh_short	1ybvA	1.84×10^{-4}	19/33	2.30×10	10	5.60×10	4
AlaDh_PNT	1pjbA	Methyltransf_3	1vid	1.94×10^{-4}	18/37	2.27×10^{2}	0	2.30×10	0
Semialdhyde_dh	1brmA	adh_zinc	1dehA	2.05×10^{-4}	18/19	1.50	15	5.66×10^{2}	6
DHDPS	1f5zA	Peripla_BP_like	1efaC	2.27×10^{-4}	29/50	2.70	0	1.20	0
B12-binding	2reqA	tRNA-synt_1d	1bs2A	2.49×10^{-4}	19/51	6.20	0	3.29×10^{2}	0
ATP-synt_ab	1bmfA	PRK	1esmB	2.60×10^{-4}	18/18	1.30×10	20	4.00	20
adh_short	1ybvA	flavodoxin	1b1cA	3.14×10^{-4}	16/31	5.60	0	2.30	0
PFK	3pfk	adh_short	1ybvA	3.22×10^{-4}	17/36	1.44×10^{3}	0	4.37×10^{2}	0
Octopine_DH_N	1bg6	Semialdhyde_dh	1brmA	3.68×10^{-4}	16/17	8.80×10	6	2.23×10^{2}	0
CbiA	1a82	fer4_NifH	1n2cE	4.03×10^{-4}	31/43	1.20×10	0	4.80×10	0
FAD_binding_3	1d7lA	ldh	5ldh	4.34×10^{-4}	15/18	1.20	19	1.48×10^{2}	21
Asparaginase	1djoA	aminotran_3	1d7rA	4.50×10^{-4}	24/36	2.79×10^{2}	0	1.24×10^{3}	0
CheR	1bc5A	adh_short	1ybvA	4.61×10^{-4}	18/42	6.80×10	0	1.60×10	0
PRK	1esmB	recA	2reb	4.61×10^{-4}	19/34	1.10×10	18	6.00	9
AdoHcyase	1a7aB	FAD_binding_3	1d7lA	5.81×10^{-4}	28/28	1.50×10	33	2.00	0
B12-binding	2reqA	Peripla_BP_like	1efaC	6.04×10^{-4}	22/71	5.20×10	10	2.92×10^{2}	24
CheR	1bc5A	Methyltransf_3	1vid	6.14×10^{-4}	55/158	1.02×10^{2}	0	3.30×10	0
MethyltransfD12	2dpmA	adh_short	1ybvA	6.59×10^{-4}	15/33	1.05×10^{2}	0	2.76×10^{2}	0
AlaDh_PNT	1pjbA	transketolase_C	1qs0B	6.96×10^{-4}	28/35	2.26×10^{2}	0	2.32×10^{2}	0
GTP_EFTU	1tuiA	adenylatekinase	3adk	6.96×10^{-4}	17/17	2.64×10^{2}	0	1.27×10^{2}	15

(*continued*)

Table 2 Continued

PFAM name1	PDB ID1	PFAM name2	PDB ID2	CMPSS *E*-value	N_acc/ L_algn	Minimal PSI-BLAST *E*-value: 1 *versus* 2	N_acc	Minimal PSI-BLAST *E*-value: 2 *versus* 1	N_acc
GMC_oxred	1gpeA	adh_zinc	1dehA	8.36×10^{-4}	18/32	1.07×10^{2}	0	1.39×10^{2}	0
Peptidase_S9	1e5tA	abhydrolase_2	1auoA	8.51×10^{-4}	17/17	1.60	26	4.00×10	0
Fibrillarin	1fbnA	RrnaAD	1yub	9.31×10^{-4}	42/71	4.00×10	25	6.30	0
ArsA_ATPase	1f48A	Glycos_transf_3	1azyA	1.06×10^{-3}	16/27	1.76×10^{2}	5	2.92×10^{2}	0
GTP_EFTU	1tuiA	recA	2reb	1.06×10^{-3}	22/186	1.10×10	27	4.00	15
GMC_oxred	1gpeA	malic	1qr6A	1.11×10^{-3}	17/17	3.38×10^{2}	18	1.30×10	0
Malic	1qr6A	pyr_redox	1d7yA	1.44×10^{-3}	20/20	4.60×10	0	2.40×10	21
GTP_EFTU	1tuiA	fer4_NifH	1n2cE	2.09×10^{-3}	21/32	4.50×10	28	4.30×10	0
APS_kinase	1d6jA	GTP_EFTU	1tuiA	2.32×10^{-3}	20/20	1.80×10	30	6.30×10	22
CbiA	1a82	recA	2reb	2.25×10^{-3}	21/35	1.17×10^{2}	19	2.90	5
GDI	1gnd	UDPG_MGDP_dh	1dliA	2.63×10^{-3}	21/31	3.30×10	0	3.49×10^{3}	0
Asn_synthase	1ct9A	PAPS_reduct	1sur	2.64×10^{-3}	15/15	4.80×10	17	4.10	20
ECH	1dciA	THF_DHG_CYH	1a4iA	2.68×10^{-3}	20/20	1.97×10^{2}	20	4.40×10	22
GFO_IDH_MocA	1evjA	aminotran_1_2	1bs0A	2.84×10^{-3}	23/40	1.20×10	0	5.40	0
SIS	1moq	ldh	5ldh	2.89×10^{-3}	15/21	2.51×10^{2}	10	3.23×10^{2}	0
APS_kinase	1d6jA	recA	2reb	3.10×10^{-3}	30/39	1.30	30	4.20	30
AdoHcyase	1a7aB	RrnaAD	1yub	3.24×10^{-3}	18/52	9.10×10	0	2.23×10^{2}	0
Dala_Dala_ligas	1iov	Flavoprotein	1e20A	3.86×10^{-3}	25/42	1.70×10^{2}	0	1.63×10^{2}	9
Thymidylate_kin	4tmkA	fer4_NifH	1n2cE	3.97×10^{-3}	20/28	3.80	21	1.43×10^{3}	0
DapB	1arzB	adh_zinc	1dehA	4.11×10^{-3}	19/19	3.50	21	3.50×10	11
DHDPS	1f5zA	aminotran_1_2	1bs0A	5.21×10^{-3}	22/72	4.50	0	4.70	0
FAD_binding_3	1d7lA	THF_DHG_CYH_C	1b0aA	6.06×10^{-3}	29/33	6.20×10	7	5.13×10^{2}	8
FAD_binding_3	1d7lA	SIS	1moq	6.06×10^{-3}	15/20	3.20×10^{2}	0	1.33×10^{3}	5
Octopine_DH_N	1bg6	adh_short	1ybvA	6.82×10^{-3}	34/37	2.20×10	8	4.40×10	8
Asp_Glu_race	1b73A	flavodoxin	1b1cA	7.69×10^{-3}	16/33	9.20×10^{2}	6	2.34×10^{3}	0
adh_zinc	1dehA	malic	1qr6A	8.00×10^{-3}	15/42	2.30	13	1.60×10	0
6PGD	1pgn	UDPG_MGDP_dh	1dliA	8.14×10^{-3}	48/116	2.60×10	24	1.23×10^{3}	0
DLH	1din	Lipase_3	1lgyA	8.18×10^{-3}	18/18	7.40×10	1	4.16×10^{2}	0
CobU	1c9kA	adenylatekinase	3adk	8.70×10^{-3}	20/30	1.50	18	3.00	20
ATP-synt_ab	1bmfA	adenylatekinase	3adk	9.37×10^{-3}	17/17	9.80×10	0	2.20×10	0
AlaDh_PNT	1pjbA	GDI	1gnd	9.52×10^{-3}	19/36	6.30×10^{2}	0	3.38×10^{2}	0

Similarities between PFAM families that were accurately predicted ($N_{acc} \geq 15$) by COMPASS with *E*-value < 0.01, and were assigned minimal *E*-value > 1.0 by PSI-BLAST. PFAM names for the alignments and PDB IDs for their representatives with solved structure are indicated, along with *E*-value estimates by COMPASS (CMPSS *E*-value), and minimal *E*-values produced by PSI-BLAST for the first (1 *versus* 2) and the second (2 *versus* 1) alignments used as query. Lengths of COMPASS local alignments (L_{algn}) and numbers of accurate matches in the local alignments (N_{acc}) are shown after corresponding *E*-values.

40 false positives, whereas PSI-BLAST and prof_sim generated 100 and 244 false positives, respectively.

In order to assess whether the *E*-values produced by COMPASS were within a reasonable range, we compared the *E*-values of the 100th false positive hit with the estimate of *E*-value for the first 100 hits in the random database of the same size. Since we pooled together all searches for each alignment as query, the estimated *E*-value should be much lower than the *E*-value for the first 100 random hits in a single search. For the setting with no additional conditions of alignment accuracy, we considered totally $\sim 1.8 \times 10^{6}$ hits; thus the *E*-value for the 100th hit in the random database should be of the order $\sim 100/1.8 \times 10^{6} = 5.6 \times 10^{-5}$. In our experiment, the *E*-value of the 100th false positive hit for COMPASS was 4.02×10^{-4}, which is $\sim$seven times higher than the value expected for the random database. The prof_sim *P*-value assigned to the 100th false positive was 8.07×10^{-4}; the estimate of *E*-value obtained from the *P*-value after the recommended multiplication by the database size (G. Yona, personal communication) was 1.09, which is $\sim$four orders of magnitude higher than the theoretical *E*-value for the 100th hit in the random database. The protocol that we used to process the results of PSI-BLAST searches included choosing the single best sequence hit from the whole set of the sequences representing a given alignment. Therefore the *E*-values for such hits were biased; they were used only for ranking the hits. Thus the use of PSI-BLAST *E*-values in their full sense was irrelevant for our setting. In fact, *E*-values for the 100th false positive hit produced in such a way were $\sim$seven orders higher than the theoretical expectations for the results of 1354 searches performed in the database of $\sim 3.1 \times 10^{5}$ sequences.

When additional restrictions of the alignment accuracy were imposed, the sensitivity plots for all methods changed dramatically (Figure 4(c) and (d)). Each method produced a much higher number of false positives, even if we demanded only $N_{acc} = 2$ positions to be aligned exactly as in the FSSP alignments. However, the performance of COMPASS was better compared to other methods

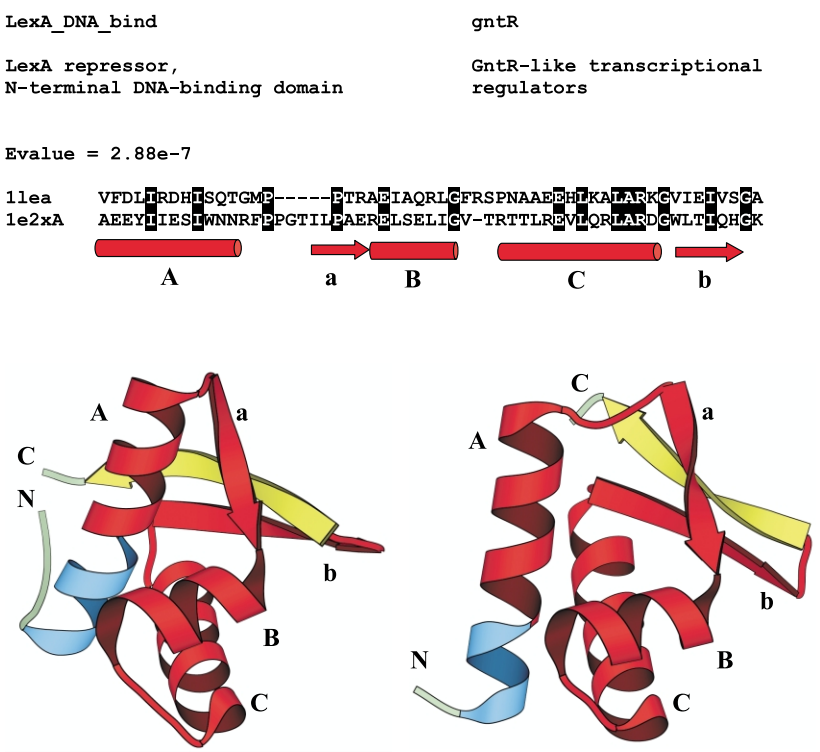

Fig. 6. Detection of common helix-turn-helix motifs. The regions of structural similarity predicted by a COMPASS hit are shown for the representatives of two PFAM families, which are classified within different SCOP families. Names of the PFAM alignments, corresponding SCOP families and PDB Ids of the representative structures are indicated. In the sequence alignment, invariant residues are boxed with black. Conserved secondary structure elements are shown below the alignment. The α-helices and β-strands are displayed as arrows and cylinders, respectively. The ribbon diagrams of the representative domains (PDB Id 1lea, residues 1–72, and PDB Id 1e2xA, residues 6–74) were drawn by MOLSCRIPT.[63] The regions of predicted similarity in the protein structures are highlighted in red, the labels of secondary structure elements corresponding to those shown below the alignment. Other secondary structure elements are colored in blue (α-helices) and yellow (β-strands).

for all tested values of N_{acc}. The total number of true positive hits detected by COMPASS was considerably larger than those detected by prof_sim or PSI-BLAST, and the rate of false positives was lower. For the top 1000 true positive hits, based on the threshold $N_{acc} = 2$, PSI-BLAST and prof_sim produced 8901 and 4246 false positives, respectively, whereas COMPASS produced 1266 false positives. For the threshold setting $N_{acc} = 15$, the sensitivity curves were much steeper, and the proportion of false positives was much higher. However, the curve for COMPASS showed considerably better performance than the curves for PSI-BLAST and prof_sim (Figure 4(c) and (d)).

Analysis of the top false positive hits revealed some drawbacks of using FSSP as a reference to assess alignments produced by search programs. Sometimes, the alignments assessed as false positives were obviously correct but different from the FSSP alignments. We noticed several main categories of such "correct false positives". One of these categories includes the cases of domain duplication. For example, the beta subunit of DNA polymerase III contains three similar domains, which are represented by different PFAM alignments. When the similarity between these alignments DNA_pol3_beta and DNA_pol3_beta_2 was detected by COMPASS (with E-value $= 1.26 \times 10^{-14}$), it was considered a false positive, since the reference did not include such interdomain alignment. Another category of correct false positives includes functionally relevant and well-conserved regions that may be misaligned in the FSSP structural alignments. For example, in the alignment of shikimate kinase chain A (PDB ID 1shkA) and D2 domain of *N*-ethylmaleimide-sensitive fusion protein chain A (PDB ID 1d2nA) derived from the multiple alignment for the 1shkA FSSP family, the obvious alignment of Walker A motifs in the two sequences is not made correctly. However, Walker A motifs were aligned by COMPASS, which assigned an E-value of 2.78×10^{-14} to the detected similarity between the corresponding AAA and SKI alignments from PFAM. Although these discrepancies exist, the portion of such inconsistencies should be approximately the same for each searching method we tested. Therefore, they should not affect the comparison of the sensitivity curves obtained by different methods.

Thus, our conclusion that COMPASS performed better searches than other methods is valid for all types of criteria that were used for hit evaluation. This suggests not only a better prediction of the structural relationships between protein domains but also a better quality of the produced alignments.

Relationships between profiles that were reliably detected by COMPASS and were not detected by PSI-BLAST

In order to compare the sets of relationships between PFAM profiles predicted by two different methods 1 and 2, we analyzed the hits that (i) were correctly detected and assigned conservative

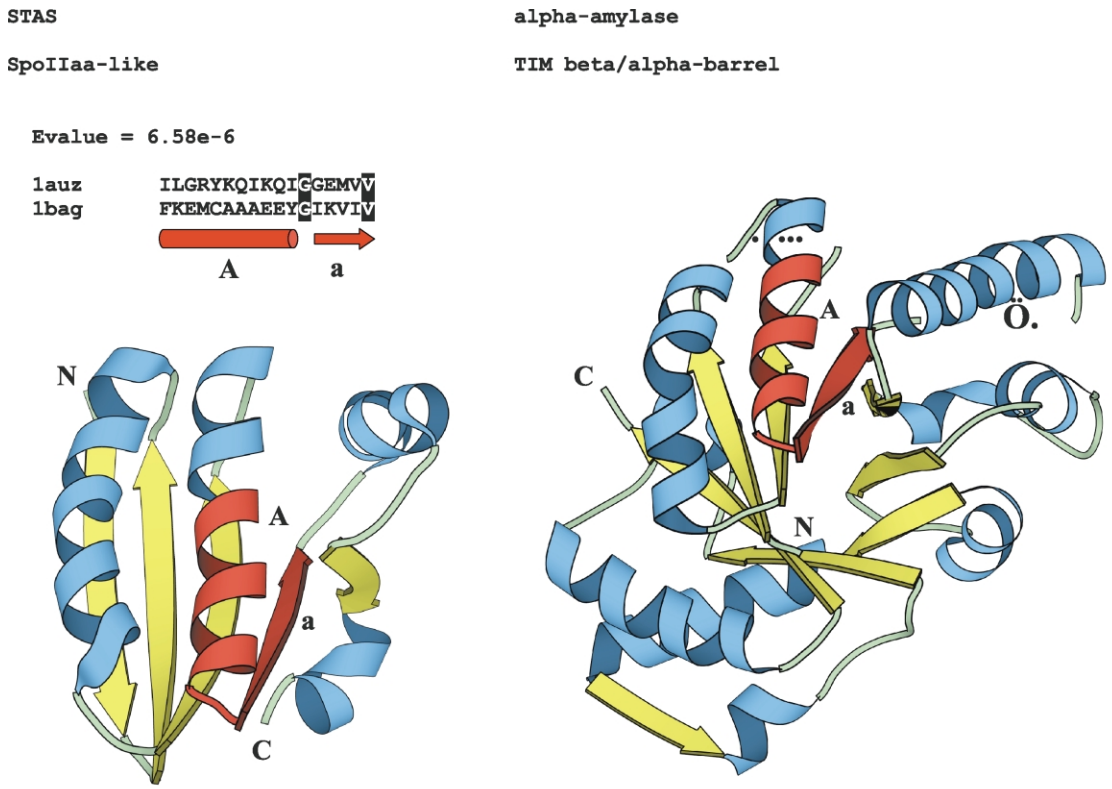

Fig. 7. Similarity between families possessing a Rossmann-type and TIM-barrel folds. The structures of the representatives of two PFAM families are shown, along with the sequence alignment produced from the COMPASS hits (invariant residues are boxed with black). Names of the PFAM alignments, corresponding protein folds and PDB Ids of the representative structures are indicated. Conserved secondary structure elements are shown below the sequence alignments. The α-helix and β-strand are displayed as arrow and cylinder, respectively. The ribbon diagrams of the domains (PDB Id 1auz, residues 1–111, and PDB Id 1bag, residues 9–43, 74–100, 150–307) were drawn by MOLSCRIPT.[63] The regions of predicted similarity in the protein structures are highlighted in red, the labels of secondary structure elements corresponding to those shown below the alignment. Other secondary structure elements are colored in blue (α-helices) and yellow (β-strands).

E-value <0.01 by method 1, and (ii) were assigned high *E*-value >1.0 by method 2, which corresponds to the expectation of at least one random hit with the same score. We requested relatively high accuracy of the predictions made by method 1: the cutoff for the number of matches consistent with the FSSP alignment was set to $N_{acc} = 15$. We considered two sets of hits that satisfied conditions (i) and (ii) for the pairs COMPASS *versus* PSI-BLAST and PSI-BLAST *versus* COMPASS. PSI-BLAST searches with PFAM alignments against the database of sequences were performed as described above, the minimal *E*-value for a profile–sequence hit being chosen among all the sequences that belonged to the "subject" alignment.

Two pairs of alignments existed for which PSI-BLAST predictions had minimal *E*-value lower than 0.01 and $N_{acc} \geq 15$, whereas COMPASS *E*-values were higher than 1.0. These pairs were PFAM families DIM1 (contains thioredoxins and related proteins) *versus* thiored (contains thioredoxins), and B12-binding (contains B12-binding domains of several enzymes) *versus* TMP-TENI (contains thiamine monophosphate synthase and the regulatory protein TENI). The PFAM family names, representative PDB Ids, *E*-values and N_{acc} for the two methods are shown in Table 1.

On the other hand, COMPASS produced predictions with *E*-value < 0.01 and $N_{acc} \geq 15$ for 114 pairs of alignments that had PSI-BLAST *E*-values greater than 1.0. Analyzing the common structural features detected by these hits, we noticed that the majority of the predictions fell into two main groups (shown in Table 2). The most populated group ($\sim$80 hits) includes relationships detected between the protein domains of Rossmann-type folds (beta–alpha–beta units). Examples of such similarities are shown in Figure 5. In order to illustrate the sequence alignments and the corresponding structural fragments revealed by these hits, a single representative with known structure was chosen from each of three PFAM alignments: flavodoxin (contains various flavodoxins), Dala_Dala_ligas (contains D-alanine–D-alanine ligase), and adh_short (contains short-chain dehydrogenases). According to SCOP[49] classification, the three domains (PDB Ids 1amoA, 1iov and 1ybvA) belong

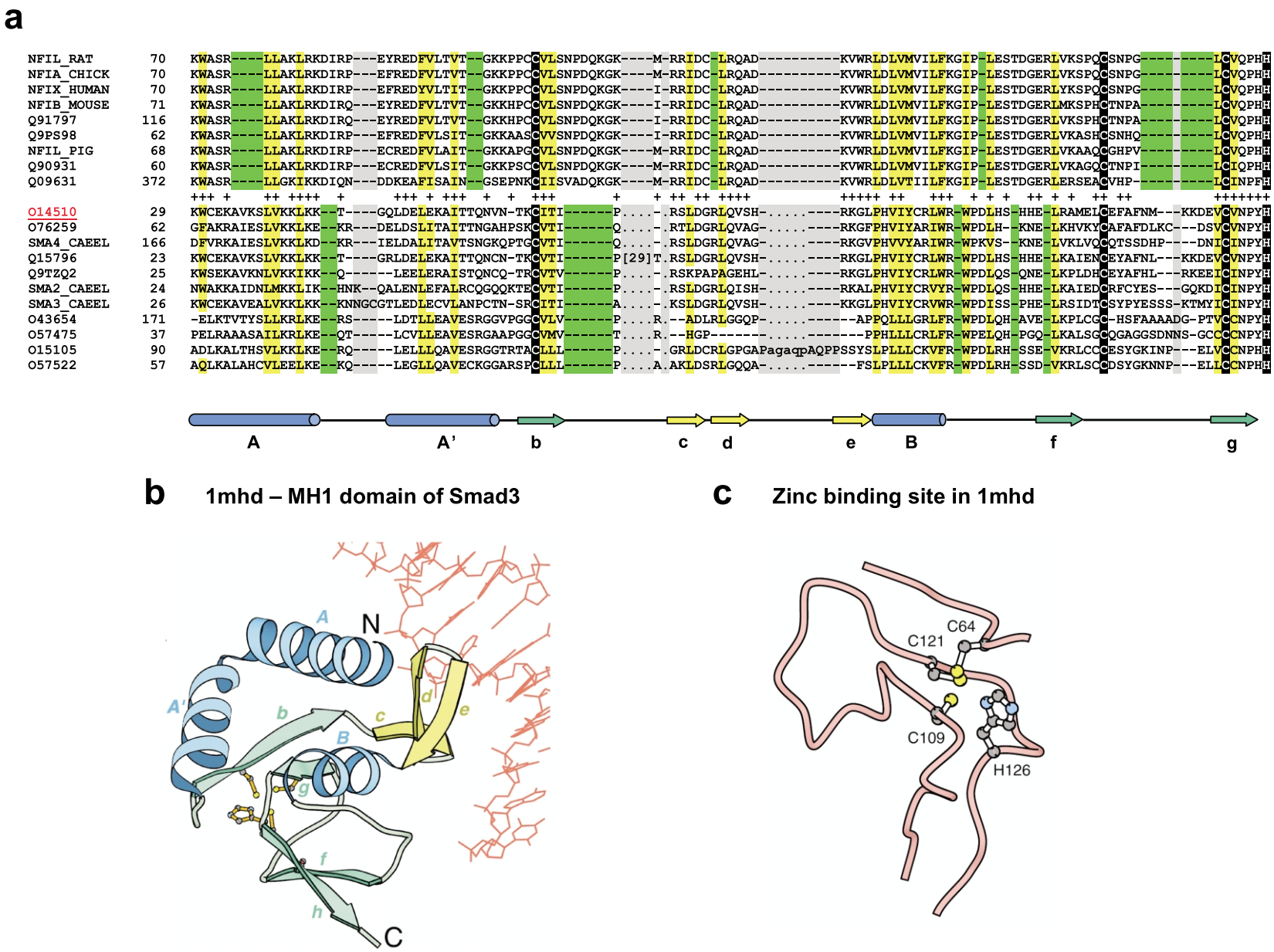

Figure 8 (*legend opposite*)

to different folds (flavodoxin-like, biotin carboxylase N-terminal domain-like and NAD(P)-binding Rossmann fold, respectively). COMPASS detected the similarity between the N-terminal fragments of the pair Dala_Dala_ligas—flavodoxin (β/α/β motif) and the pair adh_short—flavodoxin(β/α motif). In the Rossmann and Rossmann-type folds, these regions generally contain the conserved phosphate-binding motif: (N′) β-strand, loop, α-helix.

Similarities between the helix-turn-helix (HTH) motif-containing proteins comprise another group of structural relationships that are predicted with high accuracy and low *E*-value by COMPASS but are assigned high *E*-values by PSI-BLAST. This group includes 15 hits (see Table 2). An example of such a hit is shown in Figure 6. COMPASS assigned an *E*-value of 2.88×10^{-7} to the local alignment of two PFAM families, LexA_DNA_bind (contains LexA SOS regulon repressor) and gntR (contains a number of bacterial transcription regulation proteins). The alignment of two representative sequences (PDB Ids 1lea and 1e2xA) corresponds to the similarity between major parts of the domain structures. According to SCOP, these structures belong to different families of the same "winged helix" superfamily of the DNA/RNA-binding three-helical bundle fold.

These two major groups of relationships predicted by COMPASS and not predicted by PSI-BLAST are consistent with the overall composition of the fold similarities revealed by the block-based methods of profile–profile comparison, LAMA[13] and CYRCA.[16] Using ungapped alignments of the blocks, it was possible to predict a number of relationships between the folds, the largest sets of the related blocks representing HTH-motifs[13] (14 blocks) and phosphate-binding sites of Rossmann-type folds[16] (47 blocks, including several blocks from TIM-barrel folds). It is difficult to compare these results obtained on the BLOCKS database to our results obtained on the PFAM database. However, the more extended COMPASS alignments allowing gaps may not only detect the similarity between the short motifs but give more specific and diverse predictions of structural and functional similarities between the protein families (see an example below).

Along with the cases of clear homology detection, COMPASS produced accurate predictions of structural similarities that yet do not have an obvious evolutionary meaning. The largest group of such predictions contains nine hits revealing similar fragments in the domain structures of Rossmann-type and TIM-barrel folds. An example of such a hit is shown in Figure 7. The two folds are represented by anti-sigma factor antagonist SpoIIaa (PDB ID 1auz) and bacterial alpha-amylase (PDB ID 1bag), respectively. COMPASS alignment of the corresponding PFAM alignments (STAS and alpha-amylase) reveals the similarity between α/β pairs in these two structures. The evolution of a TIM-barrel from two $(\beta/\alpha)_4$ half-barrels has been previously hypothesized.[50] However, it is still unclear whether the observed hits reflect the evolutionary relationship between the families of different folds or the structural constraints that are similar for these types of α/β domains. Consistent with our results, the analysis of the block similarities by CYRCA[16] revealed short regions similar for Rossmann-type and several TIM-barrel families, which were assigned by the authors to the set of Rossmann-type fold ligand binding motifs.

Several other similarities between PFAM families that were accurately predicted by COMPASS and were assigned large *E*-values by PSI-BLAST included relationships between SH2 domains (in Cbl_N3 and SH2 families), between phosphotyrosine-binding domains (in ERM and PID families), and between the motifs typical for cysteine proteases (in Acetyltransf2 and Transglut_core families).

Example of COMPASS prediction for a PFAM family with unknown structure

To produce structural and functional predictions for PFAM families with unknown structure, we used these alignments as queries to run COMPASS

Fig. 8. Sequence similarity between the N-terminal domain of CTF/NFI family and MH1 domain of Smads infers structural and functional similarity. (a) Sequence alignment of the N-terminal region in representative sequences of CTF/NFI (PFAM family CTF_NFI, top) and Smads (PFAM family Dwarfin, bottom), as constructed by COMPASS. PFAM sequence Id and first residue number are shown for each sequence. The identifier of the sequence with known spatial structure (PDB Id 1mhd) is highlighted in red. Long insertions are not displayed: the numbers of omitted residues are specified in brackets. The regions of alignments with high content of gaps that were disregarded in the process of the local alignment construction, and corresponding gap symbols inserted in the other alignment are highlighted in gray. The gap symbols inserted in the course of the local alignment construction are highlighted in green. Potential zinc ligands are boxed in black, the uncharged residues (all amino acids except D, E, K, R) in mostly hydrophobic sites are highlighted in yellow. The secondary structure consensus is shown below the alignment, with secondary structure elements labeled and colored according to the scheme shown in (b). α-Helices and β-strands are displayed as arrows and cylinders, respectively. (b) The ribbon diagram of Smad3 MH1 domain (PDB Id 1mhd, residues A29–132) in complex with DNA that was drawn by MOLSCRIPT.[63] N and C termini are labeled. α-Helices are colored in blue, β-strands that donate zinc ligands are colored in green, other β-strands are colored in yellow. (c) The ribbon diagram of a zinc-binding site (Cys-His box) in Smad3 MH1 domain (PDB Id 1mhd, residues A63–68, A102–130). Potential zinc ligands are labeled and shown in ball-and-stick representation.

against the dataset of PFAM alignments containing at least one sequence with a solved structure. Although the complete discussion of the obtained results is beyond the scope of this paper, we would like to provide an interesting example of predictions based on one of COMPASS hits.

COMPASS assigned an E-value $= 1.73 \times 10^{-9}$ to the local alignment of N-terminal regions of the PFAM alignments CTF_NFI and Dwarfin (Figure 8). The CTF_NFI alignment contains proteins of the CTF/NFI family that have no members with known structure, whereas the Dwarfin alignment corresponds to the family of Smad proteins (i.e. dwarfins), which consist of two conserved domains separated by a linker. Crystal structures are available for both of the dwarfin domains. Both CTF/NFI and Smads are site-specific DNA-binding proteins. Smads are known as transcription factors,[51–53] whereas the CTF/NFI family is involved in the regulation of transcription and viral replication.[54] Although each of the families plays a significant role in signal transduction,[51–54] the regulatory functions of Smads, namely their place in the TGF-β cascade,[53,55,56] have been lately investigated to a greater extent. The N-terminal domain of CTF/NFI binds DNA[57,58], as well as does the N-terminal domain MH1 of most Smads.[52,59] The alignment produced by COMPASS suggested the similarity between these domains (Figure 8(a)). This similarity has not been previously reported in the literature, and was undetectable even with extensive PSI-BLAST searches.

The solved structure of MH1 domain in Smads consists of two subdomains[60] (Figure 8(b)). The functional segment of the first subdomain is the β-hairpin *de* that binds in the major groove of DNA.[60] The second subdomain is folded into a β−Ω−β unit. The significant structural similarity of MH1 to the zinc binding I-PpoI endonuclease, as well as the deviations from ideal chain geometry of MH1 strongly suggest the presence of the metal-binding site formed by three cysteine residues (C64, C109 and C121) and one histidine H126 (Figure 8(c)), which resembles the similar site in I-PpoI.[61]

The COMPASS alignment of Dwarfin and CTF/NFI includes the sequence regions of both MH1 subdomains. In addition to a significant similarity in the patterns of hydrophobicity and distribution of small residues, the alignment reveals remarkable conservation of the zinc binding Cys-His box motif in CTF/NFI family (Figure 8(a), zinc ligands boxed with black). The found sequence similarity suggests that the MH1 domain of Smads and the N-terminal domain of CTF/NFI shared a common ancestor and should be classified within the same superfamily. MH1 was hypothesized to be homologous to His-Me endonucleases.[61] Therefore, the N-terminal domain of CTF/NFI represents another family within this diverse superfamily, and is likely to be a modified endonuclease that lost its enzymatic activity but retained its ability to bind DNA. Our prediction is additionally supported by the recent inclusion of the CTF/NFI DNA-binding domains into the DWA family of the SMART database of manually curated alignments,[64] which now contains both Smad and CTF/NFI domains.

In addition to the overall homology and structural similarity of the MH1 domain and the N-terminal domain of CTF/NFI, two specific predictions can be made. The sequence region aligned to the DNA-binding segment of MH1 (β-hairpin *de*, Figure 8(a) and (b)) provides a potential DNA-binding segment in CTF/NFI, and we suggest that these two segments may employ a similar mode of binding DNA. The presence of the Cys-His box motif suggests that CTF/NFI bind metal ions similarly to His-Me endonucleases. Interestingly, mutations of the three Cys-His box cysteine residues abolished the DNA-binding activity of NFI, whereas mutation of another conserved cysteine (aligned with G73 in 1mhd) did not affect DNA binding.[62] These biochemical data further support our prediction and suggest that metal binding is required for the DNA-binding activity of the N-terminal domain of CTF/NFI.

Conclusion

Here, we present COMPASS, a new method for comparison of multiple protein alignments, which constructs local profile–profile alignments and analytically estimates E-values for the detected similarities. As compared to the existing methods of profile–sequence (PSI-BLAST) and profile–profile comparison (prof_sim), this method provides an increased ability to detect remote sequence similarities, as well as improved quality of local alignments. COMPASS was able to detect new relations between protein families that are consistent with similarities in their structures, although the families often possess different SCOP folds. The ability of COMPASS to extend the current limitations of remote sequence similarity detection may provide new insights into the structural and functional features of uncharacterized protein families. For example, COMPASS detected a novel relationship between the N-terminal domains of CTF/NFI and Smad. The corresponding COMPASS alignment (i) allows the overall structure prediction of the DNA-binding domain of CTF/NFI; (ii) suggests that this domain binds metal and predicts the metal binding site; and (iii) leads to a hypothesis about the mechanism of DNA binding by CTF/NFI, the family of transcription factors with unknown structure.

Acknowledgements

We thank Jimin Pei, Lisa Kinch and James Wrabl for discussion and critical reading of the manuscript.

References

1. Doolittle, R. F. (1981). Similar amino acid sequences: chance or common ancestry? *Science*, **214**, 149–159.
2. 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.
3. Doolittle, R. F. (1992). Stein and Moore Award address. Reconstructing history with amino acid sequences. *Protein Sci.* **1**, 191–200.
4. Gribskov, M., McLachlan, A. D. & Eisenberg, D. (1987). Profile analysis: detection of distantly related proteins. *Proc. Natl Acad. Sci. USA*, **84**, 4355–4358.
5. Luthy, R., Xenarios, I. & Bucher, P. (1994). Improving the sensitivity of the sequence profile method. *Protein Sci.* **3**, 139–146.
6. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res.* **25**, 3389–3402.
7. Schaffer, A. A., Wolf, Y. I., Ponting, C. P., Koonin, E. V., Aravind, L. & Altschul, S. F. (1999). IMPALA: matching a protein sequence against a collection of PSI-BLAST-constructed position-specific score matrices. *Bioinformatics*, **15**, 1000–1011.
8. Karplus, K., Barrett, C., Cline, M., Diekhans, M., Grate, L. & Hughey, R. (1999). Predicting protein structure using only sequence information. *Proteins: Struct. Funct. Genet.* **37**, 121–125.
9. Eddy, S. R. (1998). Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.
10. Krogh, A., Brown, M., Mian, I. S., Sjolander, K. & Haussler, D. (1994). Hidden Markov models in computational biology. Applications to protein modeling. *J. Mol. Biol.* **235**, 1501–1531.
11. Durbin, R. E., Krogh, A., Mitchison, G. & Eddy, S. (1999). *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*, Cambridge University Press, Cambridge, UK.
12. Gotoh, O. (1993). Optimal alignment between groups of sequences and its application to multiple sequence alignment. *Comput. Appl. Biosci.* **9**, 361–370.
13. Pietrokovski, S. (1996). Searching databases of conserved sequence regions by aligning protein multiple-alignments. *Nucl. Acids Res.* **24**, 3836–3845.
14. Henikoff, J. G., Greene, E. A., Pietrokovski, S. & Henikoff, S. (2000). Increased coverage of protein families with the blocks database servers. *Nucl. Acids Res.* **28**, 228–230.
15. Henikoff, S., Henikoff, J. G. & Pietrokovski, S. (1999). Blocksredundant database of protein alignment blocks derived from multiple compilations. *Bioinformatics*, **15**, 471–479.
16. Kunin, V., Chan, B., Sitbon, E., Lithwick, G. & Pietrokovski, S. (2001). Consistency analysis of similarity between multiple alignments: prediction of protein function and fold structure from analysis of local sequence motifs. *J. Mol. Biol.* **307**, 939–949.
17. Rychlewski, L., Jaroszewski, L., Li, W. & Godzik, A. (2000). Comparison of sequence profiles. Strategies for structural predictions using sequence information. *Protein Sci.* **9**, 232–241.
18. Yona, G. & Levitt, M. (2002). Within the twilight zone: a sensitive profile–profile comparison tool based on information theory. *J. Mol. Biol.* **315**, 1257–1275.
19. Sunyaev, S. R., Eisenhaber, F., Rodchenkov, I. V., Eisenhaber, B., Tumanyan, V. G. & Kuznetsov, E. N. (1999). PSIC: profile extraction from sequence alignments with position-specific counts of independent observations. *Protein Eng.* **12**, 387–394.
20. Pei, J. & Grishin, N. V. (2001). AL2CO: calculation of positional conservation in a protein sequence alignment. *Bioinformatics*, **17**, 700–712.
21. Dayhoff, M. O., Schwartz, R. M. & Orcutt, B. C. (1978). Atlas of Protein Sequence and Structure (Dayhoff, M. O., ed.), vol. 5, pp. 345–352, National Biomedical Research Foundation, Washington, DC.
22. Sjolander, K., Karplus, K., Brown, M., Hughey, R., Krogh, A., Mian, I. S. & Haussler, D. (1996). Dirichlet mixtures: a method for improved detection of weak but significant protein sequence homology. *Comput. Appl. Biosci.* **12**, 327–345.
23. Tatusov, R. L., Altschul, S. F. & Koonin, E. V. (1994). Detection of conserved segments in proteins: iterative scanning of sequence databases with alignment blocks. *Proc. Natl Acad. Sci. USA*, **91**, 12091–12095.
24. Eskin, E., Grundy, W. N. & Singer, Y. (2001). Using mixtures of common ancestors for estimating the probabilities of discrete events in biological sequences. *Bioinformatics*, **17**, S65–S73.
25. McCullagh, P. (1984). On the elimination of nuisance parameters in the proportional odds model. *J. R. Stat. Soc. B*, **46**, 250–256.
26. Staden, R. (1984). Measurements of the effects that coding for a protein has on a DNA sequence and their use for finding genes. *Nucl. Acids Res.* **12**, 551–567.
27. Schneider, T. D., Stormo, G. D., Gold, L. & Ehrenfeucht, A. (1986). Information content of binding sites on nucleotide sequences. *J. Mol. Biol.* **188**, 415–431.
28. Berg, O. G. & von Hippel, P. H. (1987). Selection of DNA binding sites by regulatory proteins. Statistical–mechanical theory and application to operators and promoters. *J. Mol. Biol.* **193**, 723–750.
29. Dodd, I. B. & Egan, J. B. (1987). Systematic method for the detection of potential lambda Cro-like DNA-binding regions in proteins. *J. Mol. Biol.* **194**, 557–564.
30. Stormo, G. D. & Hartzell, G. W., III (1989). Identifying protein-binding sites from unaligned DNA fragments. *Proc. Natl Acad. Sci. USA*, **86**, 1183–1187.
31. Lawrence, C. E., Altschul, S. F., Boguski, M. S., Liu, J. S., Neuwald, A. F. & Wootton, J. C. (1993). Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*, **262**, 208–214.
32. Henikoff, J. G. & Henikoff, S. (1996). Using substitution probabilities to improve position-specific scoring matrices. *Comput. Appl. Biosci.* **12**, 135–143.
33. Smith, T. F. & Waterman, M. S. (1981). Identification of common molecular subsequences. *J. Mol. Biol.* **147**, 195–197.
34. Gnedenko, B. (1943). Sur la distribution limite du terme maximum d'une serie aleatoire. *Ann. Mathemat.* **44**, 423–453.
35. Gumbel, E. J. (1958). *Statistics of Extremes*, Columbia University Press, New York.
36. Karlin, S. & Altschul, S. F. (1990). Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Natl Acad. Sci. USA*, **87**, 2264–2268.
37. Karlin, S., Dembo, A. & Kawabata, T. (1990). Statistical composition of high-scoring segments from molecular sequences. *Ann. Stat.* **18**, 571–581.

38. Dembo, A. & Karlin, S. (1991). Strong limit theorems of empirical functionals for large exceedances of partial sums of i.i.d. variables. *Ann. Probab.* **19**, 1737–1755.
39. Schaffer, A. A., Aravind, L., Madden, T. L., Shavirin, S., Spouge, J. L., Wolf, Y. I. *et al.* (2001). Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucl. Acids Res.* **29**, 2994–3005.
40. Altschul, S. F., Bundschuh, R., Olsen, R. & Hwa, T. (2001). The estimation of statistical parameters for local alignment score distributions. *Nucl. Acids Res.* **29**, 351–361.
41. Eddy, S. (1997). Maximum likelihood fitting of extreme value distributions. http://www.genetics.wustl.edu/eddy/publications/
42. Altschul, S. F. & Gish, W. (1996). Local alignment statistics. *Methods Enzymol.* **266**, 460–480.
43. Holm, L. & Sander, C. (1998). Touring protein fold space with Dali/FSSP. *Nucl. Acids Res.* **26**, 316–319.
44. Dietmann, S. & Holm, L. (2001). Identification of homology in protein structure classification. *Nature Struct. Biol.* **8**, 953–957.
45. Holm, L. & Sander, C. (1993). Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.* **233**, 123–138.
46. Sauder, J. M., Arthur, J. W. & Dunbrack, R. L., Jr (2000). Large-scale comparison of protein sequence alignment algorithms with structure alignments. *Proteins: Struct. Funct. Genet.* **40**, 6–22.
47. Bateman, A. *et al.* (2002). The Pfam protein families database. *Nucl. Acids Res.* **30**, 276–280.
48. Holm, L. & Sander, C. (1996). Mapping the protein universe. *Science*, **273**, 595–603.
49. Murzin, A. G., Brenner, S. E., 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.
50. Lang, D., Thoma, R., Henn-Sax, M., Sterner, R. & Wilmanns, M. (2000). Structural evidence for evolution of the beta/alpha barrel scaffold by gene duplication and fusion. *Science*, **289**, 1546–1550.
51. Heldin, C. H., Miyazono, K. & ten Dijke, P. (1997). TGF-beta signalling from cell membrane to nucleus through SMAD proteins. *Nature*, **390**, 465–471.
52. Massague, J. & Wotton, D. (2000). Transcriptional control by the TGF-beta/Smad signaling system. *EMBO J.* **19**, 1745–1754.
53. Moustakas, A., Souchelnytskyi, S. & Heldin, C. H. (2001). Smad regulation in TGF-beta signal transduction. *J. Cell. Sci.* **114**, 4359–4369.
54. Gronostajski, R. M. (2000). Roles of the NFI/CTF gene family in transcription and development. *Gene*, **249**, 31–45.
55. Heger, A. & Holm, L. (2001). Picasso: generating a covering set of protein family profiles. *Bioinformatics*, **17**, 272–279.
56. Massague, J. (1998). TGF-beta signal transduction. *Annu. Rev. Biochem.* **67**, 753–791.
57. Mermod, N., O'Neill, E. A., Kelly, T. J. & Tjian, R. (1989). The proline-rich transcriptional activator of CTF/NF-I is distinct from the replication and DNA binding domain. *Cell*, **58**, 741–753.
58. Gounari, F., De Francesco, R., Schmitt, J., van der Vliet, P., Cortese, R. & Stunnenberg, H. (1990). Amino-terminal domain of NF1 binds to DNA as a dimer and activates adenovirus DNA replication. *EMBO J.* **9**, 559–566.
59. Kim, J., Johnson, K., Chen, H. J., Carroll, S. & Laughon, A. (1997). Drosophila Mad binds to DNA and directly mediates activation of vestigial by Decapentaplegic. *Nature*, **388**, 304–308.
60. Shi, Y., Wang, Y. F., Jayaraman, L., Yang, H., Massague, J. & Pavletich, N. P. (1998). Crystal structure of a Smad MH1 domain bound to DNA: insights on DNA binding in TGF-beta signaling. *Cell*, **94**, 585–594.
61. Grishin, N. V. (2001). Mh1 domain of Smad is a degraded homing endonuclease. *J. Mol. Biol.* **307**, 31–37.
62. Bandyopadhyay, S. & Gronostajski, R. M. (1994). Identification of a conserved oxidation-sensitive cysteine residue in the NFI family of DNA-binding proteins. *J. Biol. Chem.* **269**, 29949–29955.
63. Kraulis, P. J. (1991). MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. *J. Appl. Crystallog.* **24**, 946–950.
64. Letunic, I., Goodstadt, L., Dickens, N. J., Doerks, T., Schultz, J., Mott, R., Ciccarelli, F., Copley, R. R., Ponting, C. P. & Bork, P. (2002). Recent improvements to the SMART domain-based sequence annotation resource. *Nucleic Acids Res.* **30**, 242–244.

Edited by M. Levitt

(Received 17 July 2002; received in revised form 19 November 2002; accepted 21 November 2002)