J. Mol. Biol. (1997) **273**, 355–368

JMB

Do Aligned Sequences Share the Same Fold?

Ruben A. Abagyan* and Serge Batalov

The Skirball Institute of Biomolecular Medicine Biochemistry Department NYU Medical Center 540 First Avenue, New York NY 10016, USA

**Corresponding author*

Sequence comparison remains a powerful tool to assess the structural relatedness of two proteins. To develop a sensitive sequence-based procedure for fold recognition, we performed an exhaustive global alignment (with zero end gap penalties) between sequences of protein domains with known three-dimensional folds. The subset of 1.3 million alignments between sequences of structurally unrelated domains was used to derive a set of analytical functions that represent the probability of structural significance for any sequence alignment at a given sequence identity, sequence similarity and alignment score. Analysis of overlap between structurally significant and insignificant alignments shows that sequence identity and sequence similarity measures are poor indicators of structural relatedness in the ''twilight zone'', while the alignment score allows much better discrimination between alignments of structurally related and unrelated sequences for a wide variety of alignment settings. A fold recognition benchmark was used to compare eight different substitution matrices with eight sets of gap penalties. The best performing matrices were *Gonnet* and *Blosum50* with normalized gap penalties of 2.4/0.15 and 2.0/0.15, respectively, while the positive matrices were the worst performers. The derived functions and parameters can be used for fold recognition *via* a multilink chain of probability weighted pairwise sequence alignments.

Keywords: bioinformatics; fold recognition; sequence alignment; modeling by homology; protein structure prediction

Introduction

Evaluating the significance of an alignment is an important problem in fold recognition. The actual statistics, however, depend on the alignment algorithm, choice of the substitution matrix and gap penalties. These problems have been described in detail recently in several excellent reviews (Bryant & Altschul, 1995; Altschul & Gish, 1996; Pearson, 1996; Vingron, 1996; Henikoff, 1996). Analytical expressions for the statistical significance of sequence comparisons by the fast maximal segment pair algorithm used in a popular program BLAST (Altschul *et al*., 1990) were derived by Karlin & Altschul (1990). Later, they derived the statistics for multiple high-scoring segments (Karlin & Altschul, 1993). Waterman & Vingron (1994a,b) extended the approach to alignments with high gap penalties.

Sequence comparisons can be used to infer the similarity of three-dimensional topologies (Lesk & Chothia, 1980; Chothia & Lesk, 1986); in this case, information about three-dimensional (3D) structures is used directly to derive probability distributions. An important number resulting from this analysis is the threshold of sequence identity that guarantees 3D similarity. From the analysis of homologous structures, Chothia & Lesk (1986) established a sequence threshold of 30% and the ''twilight zone'' of sequence identities between 15 and 30%. Kabsch & Sander (1984) noticed that this threshold should depend on the length of sequences being compared: five and even six residue fragments with identical amino acid sequences may adopt totally unrelated 3D folds. Recently the six-residue record was broken by Minor & Kim (1996), who designed a protein in which two 11 residue fragments with identical sequences adopted different 3D folds. The idea of length-dependent sequence identity threshold led to the

Abbreviations used: 3D, three-dimensional; HSSP, homology-derived structure of proteins (Sander & Schneider, 1991); ICM, internal coordinate mechanics; SCOP, structural classification of proteins (Murzin, *et al*., 1995); ZEGA, zero end-gap global alignment.

0022–2836/97/410355–14 $25.00/0/mb971287

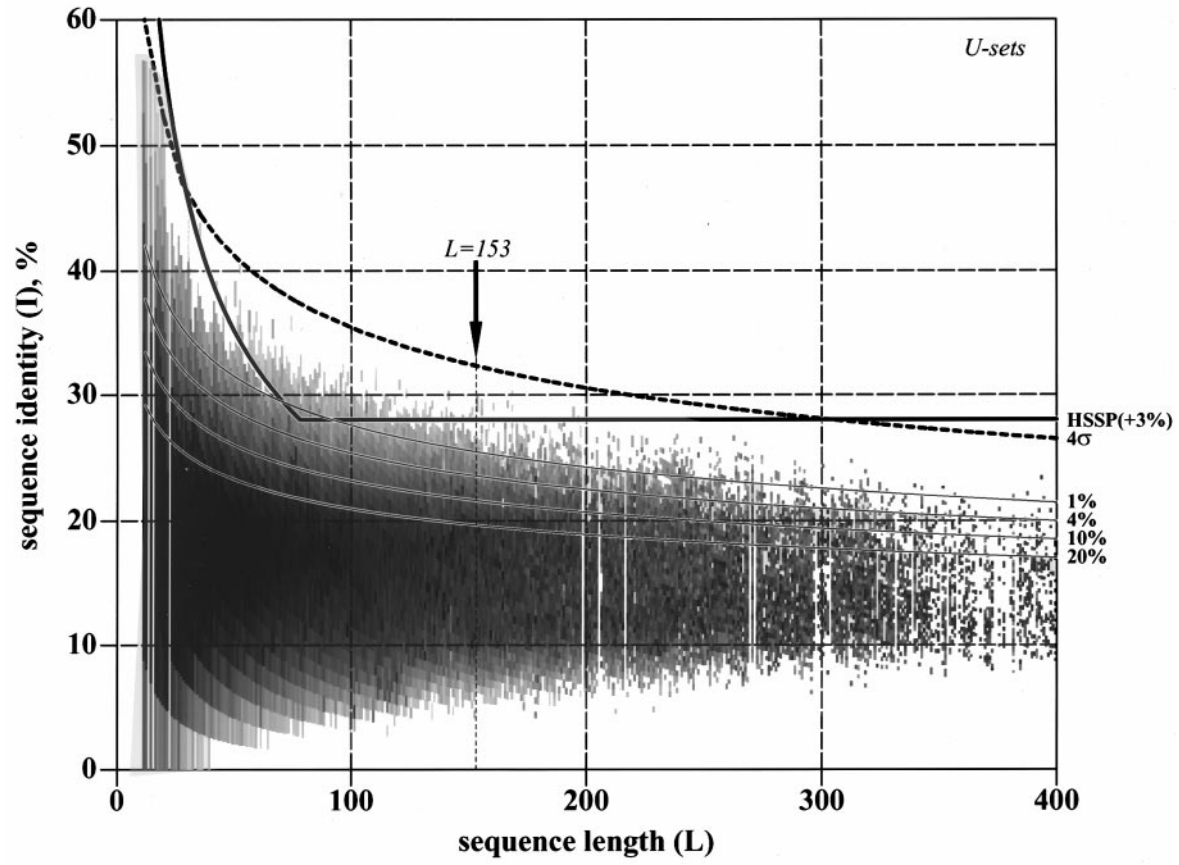

Figure 1. Distribution of sequence identities in 1,330,931 pairwise Needleman & Wunsch (1970) alignments with zero end gap penalties between sequences of structurally unrelated protein domains (*U*-set) as a function of length of the shorter sequence (L). Only alignments with sequence overlap greater than 50% were retained. Shading represents the probability density (i.e. the fraction of alignments with the given number of identical residues and minimal sequence length). White corresponds to zero density (no alignments). At low sequence identities, two sequences cannot be aligned with the global alignment algorithm utilizing a non-positive comparison matrix. This explains the white zone at very low identities. The upper continuous curve combined with the straight line shows the Sander & Schneider (1991) dependence of sequence identity threshold (equation (1)) with the safety margin $m = 3\%$, as used in the HSSP database. The broken line is the 4σ-level threshold for this comparison setup (equation (2)). The four continuous curves represent the derived thresholds at the following levels: 1, 4, 10 and 20%, from top to bottom.

HSSP database (Sander & Schneider, 1991) and the following dependence:

$$t(L) = 290.15L^{-0.562} + m \text{ for } 10 < L < 80 \text{ residues}$$
$$t(L) = 25 + m \text{ for } L > 80 \text{ residues} \quad (1)$$

where t is the percentage identity, L is the alignment length and m is the safety margin parameter (3% in HSSP).

Statistics of sequence comparisons depend on the random model and on the alignment model. Recent comparisons of a number of different amino acid exchange matrices (Vogt *et al.*, 1995) showed good and similar performance of several matrices, including those of Gonnet *et al.* (1992), Henikoff & Henikoff (1992), Johnson *et al.* (1993), and others. The best performers were positive matrices in which a constant is added to make all the elements greater than zero. Vogt *et al.* (1995) analyzed the accuracy of sequence alignments as a criterion of quality of the sequence comparison procedure.

Local alignment algorithms and a number of comparison matrices were carefully tested and ranked according to their recognition efficiency in searching 134 query sequences through 67 protein superfamilies (Pearson, 1995).

Here we used the global Needleman & Wunsch (1970) sequence alignment algorithm with zero end-gap penalties (further referred to as ZEGA) in an exhaustive cross-comparison of 2819 sequences of protein domains of known 3D structures as defined by the SCOP database (Murzin *et al.*, 1995) to answer the following questions. (a) What is the probability that a sequence alignment is structurally significant at a given length, sequence identity, sequence similarity and alignment score? (b) Which of the three mentioned criteria provides the most sensitive criterion of structural significance? (c) What residue substitution matrix and gap penalties are optimal for the global alignment based fold recognition? To address these problems we derived a family of analytical probability functions for eight

different comparison matrices and a variety of gap penalties. A subfamily of sequence identity distribution functions resulted in an updated length-dependence of the HSSP identity threshold. A rigorous fold recognition performance test allowed ranking matrices and gap penalties.

Results

How high can sequence identity be between unrelated sequences?

All pairs of structurally unrelated domain sequences were aligned using the HSSP matrix and gap penalties (Sander & Schneider, 1991). Alignments with less than 50% sequence overlap were discarded. Figure 1 contains the actual distribution of sequence identities of over a million ZEGA alignments of sequences with different SCOP folds. The sequence identity percentile (***I***) is the most traditional measure of sequence similarity, because it is easily determined and does not depend on residue exchange matrix or gap parameters once the alignment has been formed. The upper part of the distribution is the border of the twilight zone. The absence of alignments with identities below 5 to 10% is due to the fact that a global alignment procedure with non-positive comparison matrices cannot align sequences that are so different that the alignment becomes negative. The procedure prefers to make the two sequences overlap over a small fragment (less than 50% of the shortest length) and push the rest into hanging ends.

The distribution shown in Figure 1 was statistically analyzed as described in Methods and parameters of the analytical functions representing the twilight zone of sequence identities were derived. The distribution in the twilight zone at each length L is described by a normal distribution with parameters:

$$m_I(L) = 31L^{-0.124},\ \sigma_I(L) = 18.2L^{-0.305}$$

The analogue of the HSSP threshold dependence t at the 4σ level is:

$$t(L) = 31L^{-0.124} + 4 \cdot 18.2 \cdot L^{-0.305} \qquad (2)$$

leading to 41.2% at 50 residues, 35.4% at 100 residues, 30.5% at 200 residues and 28% at 300 residues. This dependence represents the identity threshold better than the HSSP dependence and specifies the level of significance. However, the threshold does depend on the comparison matrix and gap penalties, e.g. lower gap penalties may result in higher identity for the same pair of sequences.

Analytical functions for identity, similarity and score in 64 comparison settings

Similar distributions were built and analyzed for sequence identity (***I***) and for the complete alignment score (***A***) and the sequence similarity measure (***S***) defined as the normalized alignments score without gap penalties (see Methods for an exact definition). For each criterion, eight residue comparison matrices and eight sets of gap penalties were tested. For each of the 64 settings an exhaustive cross-comparison of the SCOP sequences was performed. Analytical formulae for $m(L)$ and $\sigma(L)$ were carefully derived from the distributions using the weighted twilight zone fitting within each L-set (to derive m_L and σ_L) as well as final weighted fitting to m_L and σ_L (see Methods). Twilight zone fitting means that instead of just calculating the mean and the standard deviation for each L-set, which would be appropriate for a normal distribution, we fitted the tail of the integrated observed distribution to the integrated theoretical distribution. The theoretical distribution was Gaussian (normal) for the ***I*** and ***S***-distributions (integrated distribution being the complementary error

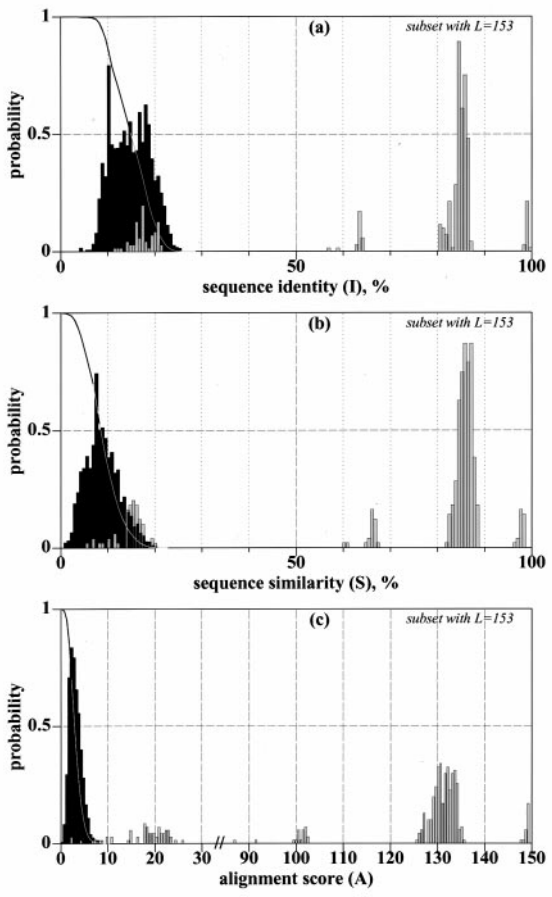

Figure 2. A sample distribution of (a) sequence identity ***I***, (b) sequence similarity ***S***, and (c) alignment score ***A***, in the subset $L = 153$ residues. This distributions illustrates the calculation of probability distributions and the overlap W between the *U*-set (black bars) and *R*-set (gray bars), i.e. between a set of alignments of unrelated and related sequences, respectively. Continuous lines show the cumulative probability distribution for the *U*-set. Comparison between (a), (b) and (c) illustrates the superiority of the alignment score over other similarity measures for this particular *L*-set.

Table 1. Distribution parameters (the mean and the standard deviation) for sequence identity, sequence similarity ($m(L) = AL^{\alpha}$, $\sigma(L) = BL^{\beta}$) and alignment score ($m_A(L) = D + L \cdot \sigma_A(L) = E + L \cdot \varepsilon$) derived from structurally insignificant alignments

Matrix (av.diag.[a])	Normalized gapOpen/Ext	Identity A_I	α_I	B_I	β_I	Similarity A_s	α_s	B_s	β_s	Alignment score D	δ	E	ε
Blosums45	2.8/0.15	29.7	−0.135	15.0	−0.274	25.3	−0.211	41.8	−0.502	2.24	0.006	1.15	0.0014
(6.25)	2.8/0.1	28.4	−0.116	16.6	−0.298	23.6	−0.169	45.1	−0.513	2.26	0.008	1.14	0.0020
	2.4/0.15	29.7	−0.116	16.4	−0.291	27.7	−0.187	46.7	−0.517	2.30	0.009	1.16	0.0023
	2.4/0.1	28.7	−0.099	18.5	−0.320	26.1	−0.149	50.4	−0.529	2.26	0.014	1.15	0.0032
	2.0/0.15	29.3	−0.091	18.2	−0.312	28.4	−0.147	50.4	−0.523	2.28	0.018	1.17	0.0039
	2.0/0.1	29.1	−0.079	20.8	−0.341	28.7	−0.124	54.9	−0.537	2.09	0.027	1.18	0.0047
	1.6/0.15	30.3	−0.071	21.0	−0.341	31.9	−0.120	54.4	−0.529	2.00	0.037	1.25	0.0055
	1.6/0.1	30.5	−0.062	23.9	−0.365	32.8	−0.103	60.2	−0.544	1.75	0.049	1.29	0.0061
Blosum50	2.8/0.15	32.1	−0.151	15.1	−0.268	28.3	−0.235	43.8	−0.503	2.10	0.003	1.20	0.0002
(6.68)	2.8/0.1	30.3	−0.128	17.1	−0.301	25.6	−0.185	49.0	−0.526	2.15	0.004	1.19	0.007
	2.4/0.15	30.7	−0.122	16.5	−0.285	28.1	−0.193	49.9	−0.525	2.25	0.005	1.21	0.0010
	2.4/0.1	29.9	−0.106	18.8	−0.319	27.9	−0.161	54.5	−0.542	2.27	0.007	1.19	0.0019
	2.0/0.15	30.9	−0.101	17.9	−0.302	31.2	−0.165	52.4	−0.526	**2.36**	**0.010**	**1.21**	**0.0026**
	2.0/0.1	30.5	−0.088	20.8	−0.338	31.0	−0.139	57.8	−0.544	2.27	0.017	1.21	0.0038
	1.6/0.15	31.5	−0.078	20.5	−0.329	34.1	−0.131	55.8	−0.528	2.24	0.027	1.26	0.0048
	1.6/0.1	31.8	−0.071	23.7	−0.361	35.1	−0.117	62.7	−0.551	1.99	0.039	1.30	0.0056
Blosum62	2.8/0.15	33.5	−0.157	15.4	−0.266	23.3	−0.200	35.3	−0.455	1.78	0.003	1.08	0.007
(5.23)	2.8/0.1	31.6	−0.134	17.5	−0.301	21.7	−0.154	39.8	−0.483	1.82	0.004	1.08	0.009
	2.4/0.15	32.4	−0.129	16.4	−0.278	25.2	−0.170	40.1	−0.477	1.89	0.004	1.09	0.0011
	2.4/0.1	31.0	−0.111	18.8	−0.315	24.5	−0.137	43.5	−0.498	1.95	0.005	1.09	0.0017
	2.0/0.15	31.8	−0.102	17.4	−0.290	27.5	−0.140	41.3	−0.476	2.08	0.007	1.11	0.0024
	2.0/0.1	31.2	−0.089	20.3	−0.329	27.4	−0.115	44.4	−0.491	2.11	0.012	1.09	0.0035
	1.6/0.15	32.2	−0.079	19.2	−0.311	30.6	−0.112	42.2	−0.471	2.16	0.020	1.13	0.0047
	1.6/0.1	32.3	−0.069	22.4	−0.343	31.5	−0.093	46.1	−0.486	1.93	0.031	1.15	0.0058
Gonnet	2.8/0.15	23.3	−0.093	24.0	−0.396	36.5	−0.165	72.4	−0.568	3.08	0.031	1.51	0.0046
(4.51)	2.8/0.1	23.5	−0.087	26.2	−0.414	36.9	−0.148	81.1	−0.588	2.85	0.039	1.53	0.0053
	2.4/0.15	23.9	−0.081	26.6	−0.416	38.9	−0.145	81.5	−0.587	**2.82**	0.046	1.57	0.0056
	2.4/0.1	24.5	−0.077	28.5	−0.428	40.3	−0.135	89.8	−0.602	2.58	0.057	1.61	0.0062
	2.0/0.15	24.9	−0.070	29.4	−0.433	42.2	−0.127	91.0	−0.604	2.52	0.067	1.67	0.0065
	2.0/0.1	26.0	−0.069	31.4	−0.442	44.9	−0.123	98.7	−0.613	2.31	0.080	1.72	0.0072
	1.6/0.15	26.8	−0.061	32.9	−0.451	47.2	−0.113	102.6	−0.620	2.24	0.096	1.81	0.0076
	1.6/0.1	28.8	−0.065	34.4	−0.451	51.8	−0.116	108.1	−0.619	2.09	0.110	1.87	0.0085
HSSP	2.8/0.15	32.7	−0.147	16.7	−0.280	15.7	−0.109	24.5	−0.378	1.52	0.006	0.97	0.0019
(7.45)	2.8/0.1	31.0	−0.124	18.2	−0.305	14.6	−0.064	27.3	−0.402	1.57	0.006	0.98	0.0021
	2.4/0.15	32.4	−0.123	17.4	−0.286	17.3	−0.085	28.1	−0.402	1.64	0.007	1.00	0.0023
	2.4/0.1	31.2	−0.106	19.2	−0.317	17.3	−0.062	30.2	−0.423	1.70	0.008	1.01	0.0028
	2.0/0.15	32.3	−0.102	17.8	−0.292	20.0	−0.073	28.8	−0.404	1.80	0.010	1.03	0.0034
	2.0/0.1	31.4	−0.086	19.8	−0.321	20.3	−0.054	29.9	−0.414	1.84	0.014	1.03	0.0044
	1.6/0.15	32.6	−0.078	18.7	−0.302	22.9	−0.053	28.6	−0.396	1.85	0.021	1.07	0.0057
	1.6/0.1	31.9	−0.063	21.0	−0.329	23.3	−0.035	29.9	−0.404	1.56	0.033	1.09	0.0068
McLachlan	2.8/0.15	25.2	−0.161	16.8	−0.385	44.7	−0.013	10.5	−0.325	1.42	0.394	1.04	0.0063
(8.07)	2.8/0.1	23.8	−0.146	16.0	−0.367	44.0	−0.009	9.9	−0.303	1.36	0.395	1.04	0.0064
	2.4/0.15	22.9	−0.130	17.1	−0.378	43.8	−0.005	11.3	−0.329	1.27	0.397	1.04	0.0063
	2.4/0.1	21.8	−0.113	16.5	−0.363	43.2	−0.001	10.8	−0.310	1.20	0.399	1.03	0.0065
	2.0/0.15	21.3	−0.099	18.1	−0.379	43.2	0.002	12.6	−0.340	1.10	0.402	1.03	0.0065
	2.0/0.1	20.6	−0.085	18.1	−0.370	42.9	0.006	12.7	−0.332	1.01	0.404	1.02	0.0068
	1.6/0.15	21.1	−0.077	20.4	−0.390	43.6	0.006	14.8	−0.363	0.90	0.409	1.02	0.0068
	1.6/0.1	20.9	−0.067	20.3	−0.381	43.7	0.009	14.8	−0.352	0.82	0.412	1.01	0.0073
*Blosum50*_p	2.8/0.15	25.4	−0.162	16.7	−0.384	50.9	−0.048	8.6	−0.295	2.15	0.372	0.84	0.0064
(11.65)	2.8/0.1	24.2	−0.148	16.1	−0.370	50.2	−0.044	8.5	−0.281	2.09	0.373	0.83	0.0065
	2.4/0.15	23.5	−0.134	17.5	−0.385	50.0	−0.041	10.4	−0.322	2.01	0.376	0.84	0.0067
	2.0/0.15	22.1	−0.107	18.9	−0.389	49.5	−0.034	12.9	−0.354	1.84	0.381	0.83	0.0066
	2.0/0.1	21.5	−0.093	19.1	−0.382	49.2	−0.030	13.5	−0.352	1.76	0.383	0.82	0.0070
	1.6/0.15	22.1	−0.086	21.3	−0.401	50.1	−0.030	16.7	−0.392	1.66	0.388	0.83	0.0069
	1.6/0.1	22.0	−0.076	21.5	−0.392	50.3	−0.027	17.1	−0.384	1.58	0.391	0.81	0.0075
*Gonnet*_p	2.8/0.15	21.5	−0.155	22.4	−0.449	64.6	−0.038	8.7	−0.308	2.21	0.506	0.78	0.0067
(9.68)	2.8/0.1	20.6	−0.143	21.7	−0.438	63.9	−0.035	8.4	−0.290	2.15	0.507	0.77	0.0068
	2.4/0.15	20.0	−0.131	22.4	−0.441	63.6	−0.033	10.0	−0.324	2.07	0.509	0.78	0.0067
	2.4/0.1	19.3	−0.118	21.9	−0.432	63.1	−0.030	10.0	−0.314	2.00	0.511	0.77	0.0069
	2.0/0.15	19.1	−0.108	23.4	−0.441	63.2	−0.028	12.6	−0.361	1.91	0.514	0.77	0.0068
	2.0/0.1	18.6	−0.096	23.2	−0.434	62.8	−0.025	12.9	−0.354	1.83	0.516	0.76	0.0072
	1.6/0.15	19.0	−0.090	25.2	−0.446	63.6	−0.025	16.3	−0.402	1.72	0.520	0.76	0.0071
	1.6/0.1	18.9	−0.081	25.2	−0.438	63.6	−0.023	16.8	−0.394	1.64	0.523	0.75	0.0076

Values of m and σ can be transformed into threshold and probability distributions (see Results).

[a] Normalized gap penalties are calculated with respect to the amino acid frequency weighted average values of the diagonal elements of the comparison matrix.

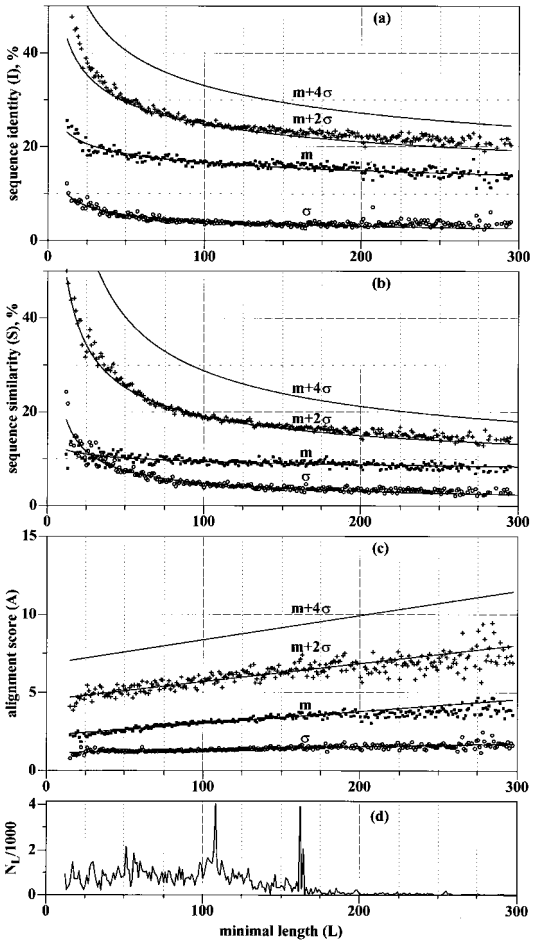

Figure 3. Distribution of the derived median (m_L, squares) and standard deviation (σ_L, open circles) parameters for all L-subsets for (a) sequence identity ***I***, (b) sequence similarity ***S***, and (c) alignment score ***A***, for *Blosum45* matrix and gap penalties 17.5 and 0.9 The marked continuous lines show the interpolations of local distributions for all sequence lengths the contributions being weighted according to N_L. Crosses represent the value above which a 2.28% fraction of the *U*-set is found (corresponding to the $m_L + 2\sigma_L$ value for the normal distribution). (d) The numbers N_L of alignments in L-subsets.

function erfc) and the extreme value distribution (Gumbel, 1962) for the ***A***-distribution. This derivation gives a more accurate description of the twilight zone even for large deviations from the theoretical distribution. A cross-section of the ***I***-distribution at $L = 153$ residues (Figures 1 and 2) shows a sample distribution from which intermediate parameters such as mean m_I, standard deviation s_I and overlap W_I were derived for $L = 153$. Similar cross-sections were calculated for all L and then smoothed with sample-size dependent weights and combined into an analytical function of L. ***I***-distribution, ***S***-distribution and ***A***-distributions were processed similarly.

The final dependencies for the thresholds at a given probability and length, as well as the probability at a given criterion value and length for ***I***, ***S*** and ***A*** can be obtained from analytical dependencies of parameters of theoretical distribution, $m(L)$ and $\sigma(L)$, as functions of length. Each distribution was described by four parameters A, α, B, β for ***I*** and ***S***-distributions or D, δ, E, ε for the ***A***-distribution, respectively. On the basis of visual analysis and χ^2-goodness-of-fit test, dependencies $m_A(L)$ and $\sigma_A(L)$ were represented by linear functions rather than the power functions chosen for ***I*** and ***S*** dependencies. The coefficients for the formulae are given in Table 1. The formulae read:

$$m_I(L) = A_I L^{\alpha_I},\ \sigma_I(L) = B_I L^{\beta_I} \tag{3}$$

$$m_S(L) = A_S L^{\alpha_S},\ \sigma_S(L) = B_S L^{\beta_S} \tag{4}$$

$$m_A(L) = D + L \cdot \delta,\ \sigma_A(L) = E + L \cdot \varepsilon \tag{5}$$

The mean values and standard deviations for the *Blosum45* matrix and normalized gap penalties of 2.8 and 0.15 are shown in Figure 3 along with the number of alignments in each L-subset. Other dependencies can be calculated using coefficients from Table 1.

For a criterion Y, where Y is sequence identity, similarity or alignment score, the threshold dependence on length at a given sigma-level y can be obtained as:

$$t(L) = m_Y(L) + y \cdot \sigma_Y(L) \tag{6}$$

The probability that criterion ***I***, ***S*** or ***A*** is higher than t can be expressed by the normal (for ***I*** or ***S***) and the extreme value distribution (for ***A***: see Methods and Appendix):

$$P_Y(> t) = \frac{1}{2}\ \mathrm{erfc}\left(\frac{y}{\sqrt{2}}\right),$$

where $$y = \frac{t - m_Y(L)}{\sigma_Y(L)} \text{ for } Y = \boldsymbol{I} \text{ or } \boldsymbol{S} \tag{7}$$

$$P_A(> t) = 1 - e^{-e^{-1.618y}},$$

where $$y = \frac{t - m_A(L)}{\sigma_A(L)} \text{ for } \boldsymbol{A} \tag{8}$$

The above family of functions can be used to specify cutoff values in database searches as well as to assign a probability of structural insignificance to an alignment on the bases of selected criterion ***I***, ***S*** or ***A***. The alignment accuracy correlates with probability (Figure 4).

Alignment score indicates structural resemblance much better than sequence identity or similarity

To find the most indicative measure of structural similarity we compared the integral overlap of Y-values ($Y = \boldsymbol{I},\boldsymbol{S},\boldsymbol{A}$) for alignments between sequences with the same folds (R-set) and

sequences with different folds (*U*-set). Figure 2 gives an example of the separation of the sets at $L = 153$. In this case the overlap and hence the recognition efficiency is worst for sequence identity and best for alignment score. An integrated measure, the total overlap value *W*, calculated as described in Methods for all *L*-sets shows that with *Blosum45* matrix and normalized gap penalties of 2.8/0.15, ranking is the same, with the overlap being 0.230, 0.194, 0.076 for sequence identity, sequence similarity and alignment score, respectively. Therefore for the above comparison setup the score provided the best separation between the *R* and *U*-sets.

Numerical optimization of function $W(Y)$ was performed to find a linear combination of three scores $Y = f_I \boldsymbol{I} + f_S \boldsymbol{S} + (1 - f_I - f_S)\boldsymbol{A}$, minimizing the overlap between the two sets in the space of parameters f_S and f_I. If parameters f_S and f_I are restricted to ranges [0,1], the best combination is $f_S = 0$ and $f_I = 0$. The overlap function in this vicinity has a shallow minimum and can be slightly improved at f_S and f_I between 0 and -0.3. However, the gain is not strong enough to justify the usage of this combination.

The overlap value *W* could not be used to compare ***I***, ***S*** and ***A*** for all eight matrices and eight sets of gap penalties because the number of formed alignments satisfying the 50% criterion varied strongly between settings. For example, the number of alignments in the *U*-set (unrelated sequences) varied from 1,257,300 (100%) for all-positive matrices to 121,958 (9.7%) for the HSSP matrix with normalized penalties 2.8/0.15 (Table 2), and the number of alignments in the *R*-set varied from 73,631 (100%) to 58.7%, respectively, resulting in different normalization of the overlap values. Therefore, we used another performance criterion, the total number of sequence pairs with the alignment criterion (***I***, ***S*** or ***A***) higher than any false positive at a given length (see Methods for the exact definition). This performance test showed that alignment score recognizes structural significant alignments better than sequence identity or similarity consistently in all 40 setups with non-positive matrices. All-positive matrices had low discriminating ability using all three criteria. The performance of similarity was a little better than that of identity (Table 2).

To further compare ***I***, ***S*** and ***A***, by identifying difficult fold recognition cases and evaluating performance in different alignment settings, we divided all the meaningful alignments (i.e. alignments between sequences from the same SCOP superfamily) into three categories (Figure 5). Each alignment was considered to be recognized as significant if its criterion was above the highest false positive in its length category. Three criteria, sequence identity, similarity or the alignment score (***I***, ***S*** and ***A***) were analyzed separately. For each criterion, eight different gap penalty settings with five non-positive comparison matrices, 40 settings in total, were tested. The alignments recognized in all the 40 settings were assigned to the obvious zone, those not recognized with any setting were assigned to the unresolved zone, and the rest constituted the ambiguous zone. The success of recognition in the ambiguous zone depended on the choice of matrix and penalties. The size of the ambiguous zone shows the number of alignments recognized differently under different comparison settings.

The location of the ambiguous zone for the three criteria (***I***, ***S*** and ***A***) again illustrates the difference in performance (Figure 5). While using the sequence identity with the best comparison setup, 54% of the alignments could not be recognized, only 50.3% of the alignments were not recognized using alignment score with even the worst matrix and gap penalties. The similarity measure performed somewhat better than sequence identity but definitely worse than the alignment score.

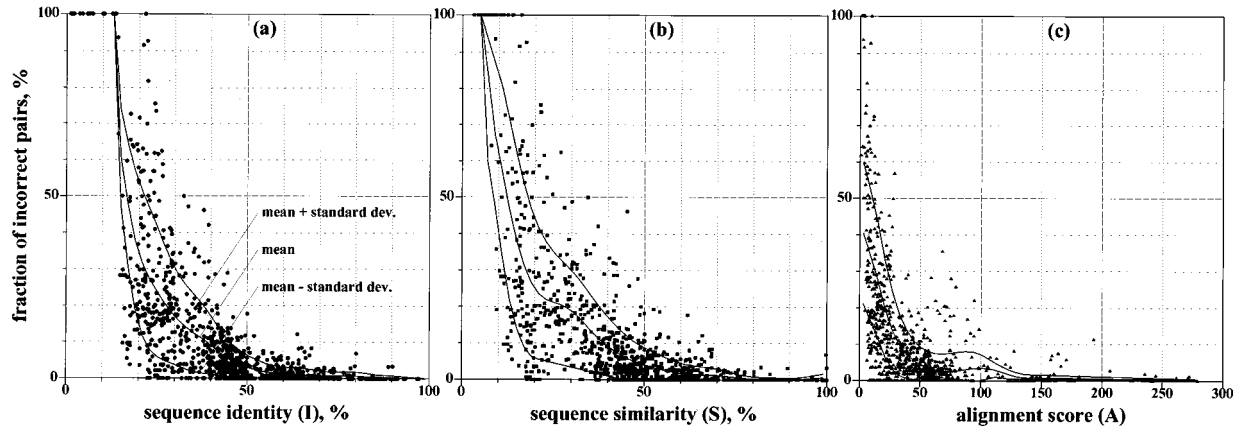

Figure 4. Accuracy of alignments at different levels of similarity for (a) sequence identity ***I***, (b) sequence similarity ***S***, and (c) alignment score. The *y*-axis shows the fraction of incorrect residue pairs. The comparison was made between the structural alignment from a database of structurally aligned protein families (Johnson *et al.*, 1993) and the Needleman & Wunsch (1970) sequence alignment.

Table 2. Fraction of structurally significant alignments correctly recognized in the ambiguous zone (see Figure 5) for different matrices and gap parameters

Residue comparison matrix (average diagonal[a])	Normalized to average diagonal[a] gapOpen/ Extension	Unnormalized[a] gapOpen/ Extension	Fraction of alignments formed with >50% overlap		Total significant alignments over the 1st false positive			Fraction of recognized significant alignments by alignment score in the ambiguous zone[d] p_A
			Unrelated[b] n_U	Related[c] n_R	Identity	Similarity	Alignment score	
Blosum45	2.8/0.15	17.5/0.94	37.3	73.1	31519	34528	39011	57.0
(6.25)	2.8/0.1	17.5/0.63	45.2	76.8	31168	33495	39130	59.8
	2.4/0.15	15./0.94	52.4	82.5	31056	33077	39427	66.7**
	2.4/0.1	15./0.63	61.9	84.5	30473	32108	39416	66.4**
	2.0/0.15	12.5/0.94	71.0	88.3	30154	31614	39543	69.4**
	2.0/0.1	12.5/0.63	78.6	90.3	29842	30629	39476	67.8**
	1.6/0.15	10./0.94	86.4	95.0	29718	30371	39460	67.4**
	1.6/0.1	10./0.63	90.3	97.0	29138	29708	39162	60.5*
Blosum50	2.8/0.15	18.7/1	22.7	65.0	31877	34650	39338	64.6*
(6.68)	2.8/0.1	18.7/0.67	28.7	68.3	31326	33307	39398	66.0**
	2.4/0.15	16./1.	33.9	70.7	31177	33240	39590	70.5***
	2.4/0.1	16./0.67	45.1	78.0	30620	32381	39582	70.3***
	2.0/0.15	**13.4/1.**	56.2	82.8	30171	31645	**39655**	**72.0*****
	2.0/0.1	13.4/0.67	67.9	87.9	29815	30743	39477	67.8**
	1.6/0.15	10.7/1	79.8	91.6	29554	30352	39493	68.2**
	1.6/0.1	10.7/0.67	86.1	96.0	29137	29666	39179	60.9*
Blosum62	2.8/0.15	14.6/0.78	13.6	60.1	32510	34939	38959	55.8
(5.23)	2.8/0.1	14.6/0.52	17.5	62.5	31955	33930	39095	59.0
	2.4/0.15	12.6/0.78	21.5	64.4	31444	32865	39386	65.7**
	2.4/0.1	12.6/0.52	29.9	70.3	30688	31980	39409	66.3**
	2.0/0.15	10.5/0.78	39.9	74.3	30140	31513	39511	68.6**
	2.0/0.1	10.5/0.52	54.1	83.6	29818	30677	39397	66.0**
	1.6/0.15	8.4/0.78	69.8	90.9	29581	30154	39313	64.0*
	1.6/0.1	8.4/0.52	79.7	94.9	29068	29482	39071	58.4
Gonnet	2.8/0.15	12.6/0.68	84.3	90.1	31770	32778	39738	73.9**
(4.51)	2.8/0.1	12.6/0.45	87.9	91.3	31405	31420	39589	70.4***
	2.4/0.15	**10.8/0.68**	90.9	92.1	31371	31377	**39799**	**75.3*****
	2.4/0.1	10.8/0.45	93.1	94.0	30826	30561	39522	68.9**
	2.0/0.15	9./0.68	95.3	94.4	30726	30516	39678	72.5***
	2.0/0.1	9./0.45	96.5	98.8	30065	29713	39291	63.5*
	1.6/0.15	7.2/0.68	97.9	99.7	29727	29419	39279	63.2*
	1.6/0.1	7.2/0.45	98.4	99.8	29433	28551	38726	50.4
HSSP	2.8/0.15	20.9/1.12	9.7	58.7	31853	34066	37842	29.9
(7.45)	2.8/0.1	20.9/0.75	12.9	60.7	31319	32926	37941	32.2
	2.4/0.15	17.9/1.12	16.2	62.0	31102	32468	38417	43.3
	2.4/0.1	17.9/0.75	23.4	66.1	30508	31672	38400	42.9
	2.0/0.15	14.9/1.12	32.1	72.2	30155	31040	38709	50.0
	2.0/0.1	14.9/0.75	45.3	84.1	29779	30289	38571	46.8
	1.6/0.15	11.9/1.12	61.1	90.2	29436	29885	38621	48.0
	1.6/0.1	11.9/0.75	72.3	94.0	28990	29349	38265	39.7
McLachlan	2.8/0.15	22.6/1.21	100	100	33897	35833	34996	n/a[e]
(8.07)	2.8/0.1	22.6/0.81	100	100	33508	35569	35241	n/a
	2.4/0.15	19.4/1.21	100	100	33864	36241	35868	n/a
	2.4/0.1	19.4/0.81	100	100	33576	35844	36145	n/a
	2.0/0.15	16.1/1.21	100	100	33727	36209	36933	n/a
	2.0/0.1	16.1/0.81	100	100	32892	35619	37162	n/a
	1.6/0.15	12.9/1.21	100	100	32738	35796	38037	n/a
	1.6/0.1	12.9/0.81	100	100	31864	34325	38161	n/a
*Blosum50*_p	2.8/0.15	32.6/1.75	100	100	33633	36532	35179	n/a
(11.65)	2.8/0.1	32.6/1.17	100	100	33507	36075	35410	n/a
	2.4/0.15	28/1.75	100	100	33625	36961	36068	n/a
	2.4/0.1	28/1.17	100	100	33223	35743	36335	n/a
	2.0/0.15	23.3/1.75	100	100	33481	36342	37215	n/a
	2.0/0.1	23.3/1.17	100	100	32536	35285	37539	n/a
	1.6/0.15	18.6/1.75	100	100	32124	34772	38317	n/a
	1.6/0.1	18.6/1.17	100	100	31393	33345	38522	n/a
*Gonnet*_p	2.8/0.15	27.1/1.45	100	100	34235	35829	34296	n/a
(9.68)	2.8/0.1	27.1/0.97	100	100	33983	35543	34420	n/a
	2.4/0.15	23.2/1.45	100	100	34481	36531	34864	n/a
	2.4/0.1	23.2/0.97	100	100	34322	35853	35068	n/a
	2.0/0.15	19.4/1.45	100	100	34194	35896	35694	n/a
	2.0/0.1	19.4/0.97	100	100	33898	35378	35917	n/a
	1.6/0.15	15.5/1.45	100	100	33750	35351	36863	n/a
	1.6/0.1	15.5/0.97	100	100	33170	34396	37136	n/a

[a] Normalized gap penalties are calculated with respect to the amino acid frequency weighted average values of the diagonal elements of the comparison matrix indicated in parentheses.

[b] The maximal number of possible alignments between unrelated sequences was 1,257,300 (100%)

[c] The maximal number of possible alignments between structurally related sequences was 73,631 (100%)

[d] The number of asterisks denotes the surplus over 60% value in 5% steps.

[e] All-positive matrices McLachlan, *Blosum50_p* and *Gonnet_p* produced substantially smaller number of significant alignments and were not taken into account in determination of the ambiguous zone. They would have had negative p_A *values in this scale.*

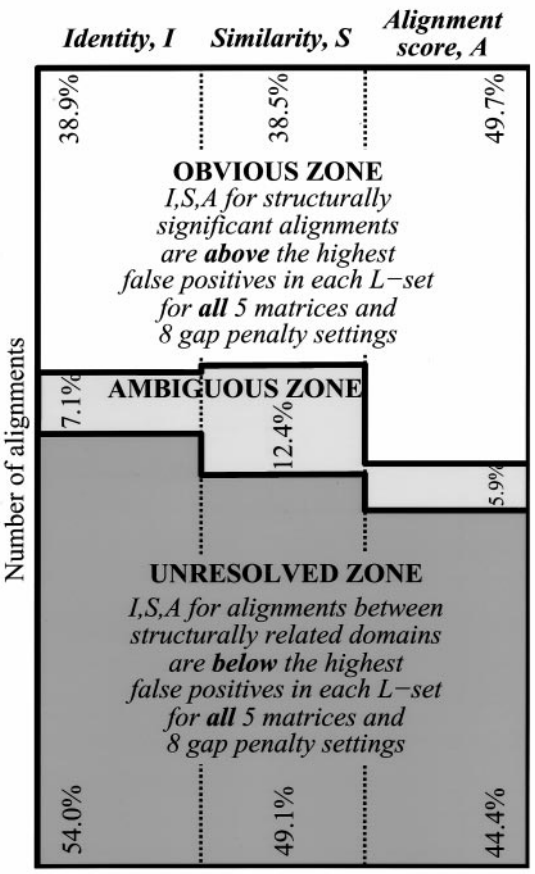

Figure 5. Separation of 73,631 alignments between structurally related sequences into three zones: the obvious (above false positives in all 40 non-positive alignment setups), the ambiguous (recognition depends on the setup) and the unresolved (below false positives in all 40 non-positive alignment setups).

Ranking of comparison matrices and gap penalties in a fold recognition test

The above analysis shows that alignment score is a good indicator of structural similarity. However, identification of structurally significant alignments is affected by the choice of comparison matrix and gap penalties for the 4365 alignments (5.9% of all alignments in the *R*-set) belonging to the ''ambiguous'' zone. To rank matrices and gap penalties by their performance in fold recognition, we define the performance criterion p_A of a given alignment setup as the relative number of the ambiguous alignments that were recognized.

Table 2 contains p_A values for 64 different setups. Positive matrices were clearly the worst performers. The HSSP matrix reached performance of 50.2% with normalized penalties of 2.0 and 0.15. Four best matrices, *Gonnet*, *Blosum50*, *Blosum45* and *Blosum62*, had close performance of 75.3, 71.7, 69.4 and 68.3%, respectively. The performance criterion shows that the optimal value of gap opening penalty is within the tested interval [1.6:2.8] for non-positive matrices. The three *Blosum* matrices showed best performance with normalized gap penalties of 2.0/0.15, and the *Gonnet* matrix with 2.4/0.15.

Discussion

Here, we derived analytical functions (equations (3) to (8)) for the structural significance of three measures of global sequence alignment (ZEGA) by exhaustive comparison between sequences of protein domains with known 3D structure (Murzin *et al*., 1995). This approach is complementary to a well-established method of evaluating the statistical significance of an alignment on the basis of randomly generated sequences (Needleman & Wunsch, 1970; Fish, 1983; Altschul & Gish, 1996; Bryant & Altschul, 1995), since definition of the random model exactly according to the goal, which is discrimination between protein folds, is preferable. Automatic compilation of domain assignments for all the PDB structures is not a trivial task (e.g. see Orengo *et al*., 1993; Siddiqui & Barton, 1995; Sowdhamini & Blundell, 1995), since both domain definition and clustering procedures have adjustable parameters and it is difficult to tune them so that decisions for all different structures are equally satisfactory. The interactively assigned fold and superfamily categories of the SCOP database (Murzin *et al*., 1995) determine a level of 3D similarity appropriate for our analysis. It is essential that we test a global alignment procedure with zero end gaps. In contrast to local alignment procedures such as Smith & Waterman (1981), FASTA, and BLAST, this procedure will not be efficient if two multidomain sequences are compared or if a domain sequence is interrupted by a long insertion, however the ZEGA procedure is quite efficient in comparison of continuous domains.

In our derivation of smooth functions from a set of about a million data points we had to deal with the problem of deviation from both normal and extreme value distribution and sparse statistics in *L*-sets. Bryant & Altschul (1995) note that ''there is no reason to believe that the random score distribution is normal''. This is equally true for the distribution of identities, similarities and alignment scores in our *U*-set, i.e. gapped global alignments with zero end penalties between sequences with different folds. The extreme value distribution was previously used to describe distribution of scores of ungapped local alignments (Karlin & Altschul, 1990, 1993; Altschul & Gish, 1996) and local alignments with high gap penalties (Waterman & Vingron, 1994a). We found the alignment scores to be distributed close to the extreme value distribution, as one might expect for a value that is maximized by the alignment procedure, while the sequence similarity ***S*** and especially sequence identity ***I*** were distributed rather normally, since they are not directly maximized by the procedure. Furthermore, we attempted to find an accurate

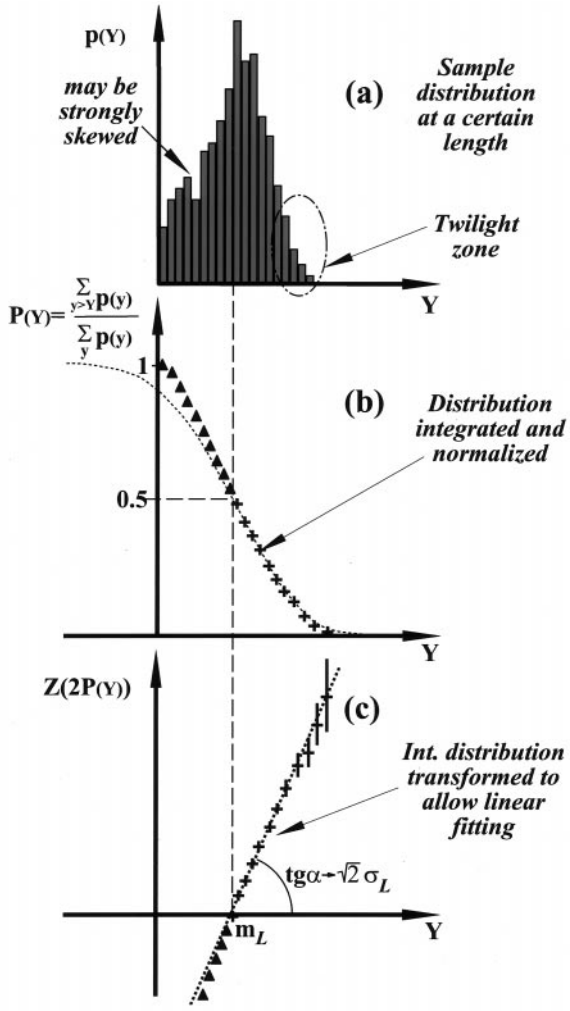

Figure 6. A diagram illustrating how the median (m_L) and standard deviation (σ_L) were derived to only fit the tail of the twilight zone distribution and reduce the impact of the deviation from the normality (''skewness'') at lower values of the alignment criterion Y distribution. (a) A histogram of the Y-distribution at a certain length. (b) The same distribution integrated and normalized to 1. (c) The integral distribution transformed to become linear. The Z-function transforming the normalized integrated distribution (b) into a linear function is the inverse of the theoretical distribution function. Points with negative Z values (filled triangles) are discarded in the fitting and contributions are weighted according to the transformed expected errors. The intersection with the y-axis and the slope of the fitted line give m_L and σ_L, respectively. The broken curve in (b) and the line in (c) represent the same cumulative theoretical distribution.

representation of the twilight zone (upper gray area in Figure 1). To achieve this, we fitted the tail of the integrated distribution (Figures 2 and 6) rather than all the points. In this way, deviations from the theoretical distribution in the uninteresting range of low similarities did not affect the probability functions.

The probability of structural significance for an alignment as a function of sequence identity was meant to replace the previously derived HSSP dependence (Sander & Schneider, 1991) shown in Figure 1. This type of function is of general importance, since one can simply count sequence identities and easily evaluate the significance of an alignment using a plot or a pocket calculator. The downside of this convenience is a low level of accuracy, since other indicators, especially the alignment score, are more powerful, and the identity statistics depend on the substitution matrix and gap penalties. The new curve (Figure 1) $m_I + 4\sigma_I$ derived with the HSSP comparison parameters is drawn at the P level of 1/31,000 to guarantee the significance of alignments with identities above the curve in a Swissprot search. One can see that the HSSP curve seriously underestimates the identity requirements for lengths between 80 and 170 residues even at the safety margin of 3%. Therefore a larger ''safety'' margin was frequently used. The HSSP curve was a function of the length of the sequence of the parent PDB entry, referred to as an alignment length. Insertions in found homologous sequences were ignored. In this work we used a close but more symmetric definition of the alignment length, as the length of the shortest sequence (L). It is quite appropriate for a global alignment with zero end gap penalties. The two differences are quite close and the same formulae can still be used, especially given the weak dependence on length for average-size proteins. Two other possible definitions of the alignment length, as the number of alignment residue pairs ($\leqslant L$) and the alignment length with gaps ($\geqslant L$), can also be used with some length correction factors depending on the gap penalties (about 20% on the average).

The HSSP statistics were extended in this work in the following directions: (i) the dependence was more rigorously derived from a larger data set; (ii) instead of one threshold dependence, general formulae were found to determine a threshold at any probability level; (iii) the inverse functions were also obtained; (iv) the functions were calculated for eight different matrices and eight sets of gap penalties.

Functions for sequence similarity and alignment score (equations (3) to (8) and Table 1) are actually more important for the evaluation of structural significance, but they directly depend on the residue exchange matrix and gap penalties. The exchange matrices were normalized (see Methods) to make the dependences more universal. The formulae for the most sensitive significance criterion, the alignment score, are the most useful, since they allow evaluation of structural significance for the alignment in terms of probability.

Comparison between the discrimination ability of sequence identity, sequence similarity and alignment score demonstrated clearly that (i) all-positive matrices performed poorly and (ii) the alignment score is the most powerful discriminator between true and false positives for all 40 non-positive alignment setups analyzed (Figure 5 and Table 2).

Any linear combination of I, A and S is also less efficient than the score alone. It is possible, though, that linear combination log$P$ values rather than the actual scores, further combination of the score with threading scores (Jones & Thornton, 1996; Abagyan *et al*., 1994) and, possibly, introduction corrections, e.g. the residue composition correction (for reviews, see Bryant & Altschul, 1995; Pearson, 1996) or **ln**()-correction (Pearson, 1996) will result in a better separation. Here, we did not introduce the length correction because we used a global alignment algorithm and with the intention of using it in comparison of a domain sequence with a much longer sequence of a multidomain protein. This would be the case if a domain sequence were searched against a sequence database, or if a full unknown sequence were searched against a database of domain sequences. Local comparison methods (e.g. Smith & Waterman (1981); FASTA, Pearson & Lipman, (1988); and BLAST, Altschul *et al*. (1990)) do not have this problem and can efficiently compare two multidomain protein sequences. However, even for a global alignment with zero end gap penalties, the length of the aligned fragment of the longer sequence can still be used in a log-correction term. The derived functions are essential for evaluating the similarity between a query sequence and a known 3D fold through a multilink chain of pairwise comparisons (Abagyan *et al*., 1997). In this approach the folds are ranked according to a product of probabilities $...(1 - p_A^{(i)}) \cdot (1 - p_A^{(i+1)})...$ in the best chain of sequential pairwise alignments bridging the query sequence with the fold sequence.

The fold recognition performance criterion allowed us to compare different residue comparison matrices and gap penalties used for the global alignment procedure with zero end gaps. It is quite clear that ranking of matrices and gap penalties depends strongly on the performance criterion. The accuracy of the alignment criterion used by Vogt *et al*. (1995) leads to different values than the correct database ranking criterion (Pearson, 1995). By reducing gap opening penalty, for example, the noise level of a database search is increased, but the alignment accuracy may improve. Also, surprisingly, the positive matrices that were the best in the alignment test by Vogt *et al*. (1995) were the poorest in the fold recognition test (Table 2). The best matrices were *Gonnet* (Gonnet *et al*., 19922 $p_A = 75.3\%$) and *Blosum50* (Henikoff & Henikoff (1992) $p_A = 72.0\%$). *Blosum45* and *Blosum62* with optimized gap penalties were very close. Several interesting observations result. First, the *Gonnet* matrix aligns many more unrelated sequences (almost as many as the inefficient all-positive matrices) than the best Blosum matrix (90.9% *versus* 56.2%, respectively, Table 2) but the increased noise level does not result in bad performance. The second interesting fact is that the best performing *Gonnet* matrix predicts only 75.3% (rather than 100%) of the alignments in the ambiguous zone, i.e. 24.7% of the alignments were not recognized by the *Gonnet* matrix but were recognized by other less efficient matrices. This implies that the usage of several different matrices and gap penalties, rather than a single best performing matrix, might be preferable. This work makes a multi-matrix approach possible, since probability distributions were derived for 64 different comparison settings.

As the alignment parameters approach the twilight zone the positional accuracy of the alignment deteriorates (Vogt *et al*., 1995). Figure 4 shows that at 30% of sequence identity the mean fraction of misaligned residues is 20% with standard deviation of 10%. However, as we saw above matrices and penalties that are optimal for fold recognition are not necessarily optimal from the alignment accuracy point of view.

Methods

Database and alignment

The 7849 protein sequences of continuous protein domains as defined in the SCOP database (Murzin, *et al*., 1995), summer 1996 release, were extracted from the Protein Data Bank (Abola *et al*., 1987) and 2819 unique protein sequences with length greater than 12 residues were retained. Each of these 2819 sequences was assigned a protein fold tag according to the second category of the SCOP classification (class + fold). Pairs of sequences with the same fold were assigned to the set of related sequence (*R*-set), while sequences with different fold tags were assigned to the unrelated set (*U*-set).

Global Needleman & Wunsch (1970) sequence alignment was performed with the ICM program (Molsoft, 1996). The ZEGA alignment procedure employed zero end gap penalties. Only alignments in which the number of aligned residues was greater than one half of the length of the shorter sequence were retained for further analysis.

Residue substitution matrices

Eight residue substitution matrices were used. A family of three *Blosum* matrices (Henikoff & Henikoff, 1992), *Blosum45, Blosum50* and *Blosum62*, were derived from blocks of alignments extracted from several hundred protein families. These matrices contain negative elements. The *Gonnet* matrix (Gonnet *et al*., 1992) was built from exhaustive cross-comparison of sequences in a sequence database. The McLachlan (1971) matrix was derived from 89 pairwise alignments from 16 protein families. The HSSP matrix (Sander & Schneider, 1991) was built from the McLachlan matrix by scaling to the range −0.7:1.0. All positive matrices *Blosum50-p* and *Gonnet-p* are described in Vogt *et al*. (1995). The residue exchange matrices were normalized by multiplication of all the num-

bers by a factor such that the sum of occurrence-weighted diagonal elements (identities) of the matrix was 1.0. The following residue frequencies were used (A 7.85, C 2.55, D 5.17, E 6.95, F 4., G 6.52, H 2.12, I 5.45, K 5.66, L 8.86, M 2.51, N 4.59, P 4.67, Q 4.09, R 5.17, S 7.1, T 5.48, V 6.2, W 1.46, Y 3.05%).

Definitions of alignment significance criteria

Three significance criteria Y were analyzed. Sequence identity, ***I***, was defined as the number of identical residues divided by the length of the shortest sequence (L) and multiplied by 100%. Sequence similarity, ***S***, was calculated as the sum of similarity values from the normalized residue exchange matrix for all aligned residue pairs divided by L and multiplied by 100%. Since the residue exchange matrix is normalized so that the weighted average of a diagonal element is 1.0, the expected range of the similarity measure is comparable to the sequence identity measure. Alignment score, ***A***, was defined as the sum of similarity values for the aligned residue pairs minus the sum of $O + El_{\text{gap}}$, where O is gap opening penalty, E is gap extension penalty and l_{gap} is the gap length, for all but the terminal gaps.

Both ***S*** and ***A*** measures depend on normalization of the comparison matrix. To calculate probability and threshold dependencies (see equations (3) to (8) and Table 1) for an original unnormalized matrix, coefficients A_S, B_S, D, δ, E, ε from Table 1 must be multiplied by the average diagonal value from the same Table.

Derivation of probability distributions

To derive the probability functions and the recognition for a given criterion Y, the following steps were performed.

(1) The whole data set of alignments was divided into L-subsets, each of which had the same minimal length (L) of the sequences constituting alignments and contained N_L alignments. Each L-subset was divided into an R-set of alignments between structurally related sequences (having the same class/fold tag from SCOP classification) and a U-set of alignments between structurally unrelated sequences (different class/fold tags).

(2) For each subset L a histogram $p(Y)$ was calculated for the U-sets (Figure 6a).

(3) Each $p(Y)$ was integrated (summated) into the integral distribution $P(Y)$ representing the relative number of alignments with the criterion value exceeding Y (Figure 6b).

(4) For each subset L we performed the optimal fit of the tail of the integrated distribution $P(Y)$ to the tail of the integrated theoretical distribution (the so-called complementary error function erfc for ***I*** and ***S***, and extreme value distribution for ***A***) with parameters m_L and σ_L. To allow linear fitting $P(Y)$ was transformed (Figure 6c) with a certain function Z, which is a computable non-analytical function inverse to the integrated theoretical distribution function. In the case of normal distribution, the slope of the fitted line is $\sqrt{2} \cdot \sigma_L$ and the intersection with the abscissa gives m_L, which is the median of the distribution. The median coincides with the mean for a normal distribution but differs for skewed distributions. To calculate the fitting weights of points upon transformation, the uncertainties of the individual points were taken:

$$\varepsilon_Z = \varepsilon_P \frac{\mathrm{d}Z}{\mathrm{d}P} = \varepsilon_P \left(\frac{\mathrm{d}P}{\mathrm{d}Z}\right)^{-1} = C' \mathrm{e}^{Z^2}$$

since:

$$P(Z) = \mathrm{erfc}(Z) = C \int_Z^\infty \mathrm{e}^{-t^2}\, \mathrm{d}t$$

where the constants C, C' and ε_P are of no interest. In the case of the extreme value distribution the uncertainties can be calculated as:

$$\varepsilon_Z = \varepsilon_P \frac{\mathrm{d}Z}{\mathrm{d}P} = \frac{C''}{P \ln P}$$

since:

$$P(z) = 1 - \mathrm{e}^{-\mathrm{e}^{-\xi z}},\ Z(P) = \frac{1}{\xi} \ln(-\ln P)$$

The weights of points were then calculated as $1/\varepsilon_z^2$.

(5) All m_L and σ_L values were then used to derive two smooth, interpolated functions for $m(L)$ and σL) (Figure 3). (i) The alignment score distribution parameters were well approximated by simple linear functions (Figure 3c): $m_A(L) = D + \delta \cdot L$ and $\sigma_A(L) = E + \varepsilon \cdot L$. Parameters of the linear functions were derived by linear fitting to the m_L and σ_L points with the uncertainties of each data point taken as $1/\sqrt{N_L}$ and weight being N_L (see Numerical Recipes in C, 1992). (ii) Identity and similarity measures could not be approximated by the linear function and the power function AL^α was used instead (Figure 3a and b and see also Sander & Schneider, 1991). Parameters A and α of the AL^α function were found by best-linear fit of $\log m_L$ values *versus*. $\log L$, uncertainties and weights of each point being $1/m_L\sqrt{N_L}$ and $N_L m_L^2$ respectively, because:

$$\varepsilon_{\log m} = \varepsilon_m \frac{\mathrm{d} \log m}{\mathrm{d}m} = \frac{\varepsilon_m}{m} = \frac{1}{m\sqrt{N_L}}$$

Parameters B and β for σ_L were derived by the same procedure.

Comparison of the alignment criteria for fold recognition

To find the optimal criterion Y, we analyzed the linear combination of the above three measures:

$$Y = f_I \boldsymbol{I} + f_S \boldsymbol{S} + (1 - f_I - f_S)\boldsymbol{A} \qquad (9)$$

To derive the discrimination efficiency for a given criterion Y, we evaluated the total overlap between U-set and R-set through the following procedures. (1) The whole data set of alignments was divided into L-subsets, each of which had the same minimal length (L) of the sequences constituting alignments and contained N_L alignments. Each L-subset was divided into an R-set of alignments (N_R alignments) between structurally related sequences (having the same class/fold tag from SCOP classification) and a U-set of alignments (N_U alignments) between structurally unrelated sequences (different class/fold tags). (2) For each subset L, the overlap W_L between R-set ($i = 1,N_R$) and U-set ($j = 1,N_U$) distributions characterizing the discrimination ability of a given Y-criterion was calculated as:

$$W_L = 1 - \left| \frac{1}{N_U N_R} \sum_{i=1,N_R} \sum_{j=1,N_U} \delta(Y_{U_i}, Y_{R_j}) \right|, \qquad (10)$$

where $\delta(a,b)$ equals 1 if a is greater than b, 0 it they are equal, and -1 otherwise. The overlap defined in this way equals 0, if the distributions are divided apart, and reaches the maximum value of 1 if the distributions cover each other's range uniformly.

(3) to obtain the final W_Y, all W_L contributions were averaged with the weight of N_L, similarly to the above section:

$$W_Y = \frac{\sum_L W_L N_L}{\sum_L N_L}$$

Performance test

Each sequence alignment between structurally related protein domains was marked as correctly recognized if its alignment criterion (***I***, ***S*** or ***A***) was greater than the smoothed threshold T_L, or otherwise marked as unrecognized. This division was used in Figure 5 and Table 2. The smoothed threshold with the following procedures: (i) the highest false positives (the greatest criterion values for alignments from the U-set) were collected for each L-subset; (ii) highly fluctuating values of highest false positives were averaged within a sliding window $[L\text{-}3 : L + 3]$. The performance criterion p_A for given alignment settings is defined as the number of correctly recognized alignments from the ambiguous zone divided by the total number of alignments in the ambiguous zone.

Acknowledgments

We thank Tim Hubbard and Alexei Murzin for kindly providing us with a table of SCOP definitions of protein domains and Martin Vingron for interesting discussions. We thank Department of Energy (DoE grant DE-FG02-96ER62268) for financial support. DoE's support does not constitute an endorsement by DoE of the views expressed in the article.

References

Abagyan, R. A., Frishman, D. & Argos, P. (1994). Recognition of distantly related proteins through energy calculations. *Proteins: Struct. Funct. Genet.* **19**, 132–140.

Abagyan, R. A., Batalov, S., Cardozo, T., Maiorov, V. & Totrov, M. (1997). Computational approaches to understanding proteins. *Protein Sci.* **6**(suppl.2), 58.

Abola, E. E., Bernstein, F. C., Bryant, S. H., Koetzle, T. F. & Weng, J. (1987). Protein data bank in crystallographic databases-information content, software systems, scientific applications. In *Data Commission of the International Union of Crystallography* (Allen, F. H., Bergerhoff, & Sievers, R., eds), pp. 107–132.

Altschul, S. F. & Gish, W. (1996). Local alignment statistics. *Methods Enzymol.* **266**, 460–480.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410.

Bryant, S. H. & Altschul, S. F. (1995). Statistics of sequence-structure threading. *Curr. Opin. Struct. Biol.* **5**, 236–244.

Chothia, C. & Lesk, A. M. (1986). The relation between the divergence of sequence and structure in proteins. *EMBO J.* **5**, 823–826.

Fish, W. M. (1983). Random sequences. *J. Mol. Biol.* **163**, 171–176.

Gonnet, G. H., Cohen, M. A. & Benner, S. A. (1992). Exhaustive matching of the entire protein sequence database. *Science,* **256**, 1433–1445.

Gumbel, E. J. (1962). Statistical theory of extreme values (main results). In *Contributions to Order Statistics,* pp. 56–93, Wiley, New York.

Henikoff, S. (1996). Scores for sequence searches and alignments. *Curr. Opin. Struct. Biol.* **6**, 353–360.

Henikoff, S. & Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proc. Natl Acad. Sci. USA,* **89**, 10915–10919.

Johnson, M. S., Overington, J. P. & Blundell, T. L. (1993). Alignment and searching for common protein folds using a data bank of structural templates. *J. Mol. Biol.* **211**, 735–752.

Jones, D. T. & Thornton, J. M. (1996). Potential energy functions for threading. *Curr. Opin. Struct. Biol.* **6**, 210–216.

Kabsch, W. & Sander, C. (1984). On the use of sequence homologies to predict protein structure. Identical pentapeptides can have completely different conformations. *Proc. Natl Acad. Sci. USA,* **81**, 1075–1078.

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.

Karlin, S. & Altschul, S. F. (1993). Applications and statistics for multiple high-scoring segments in molecular sequences. *Proc. Natl Acad. Sci. USA,* **90**, 5873–5877.

Lesk, A. & Chothia, C. (1980). How different amino acid sequences determine similar protein structures: the structure and evolutionary dynamics of the globins. *J. Mol. Biol.* **136**, 225–270.

McLachlan, A. D. (1971). Tests for comparing related amino acid sequences. Cytochrome *c* and cytochrome *c*551. *J. Mol. Biol.* **61**, 409–424.
Minor, D. L., Jr & Kim, P. S. (1996). Context-dependent secondary structure formation of a designed protein sequence. *Nature,* **380**, 730–734.
Molsoft, L. L. C. (1996). *ICM Software Manual. Version 2.5*, http://molsoft.com.
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.
Needleman, S. B. & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48**, 443–453.
Numerical Recipes C., (1992). *The Art of Scientific Computing* (Press, W. H., Teukolsky, S. A., Vetterling, W. T. & Flannery, B. P., eds), 2nd edit., pp. 220–221 and 661666, Cambridge University Press, Cambridge.
Orengo, C. A., Flores, T. P., Taylor, W. R. & Thornton, J. M. (1993). Identification and classification of protein fold families. *Protein Eng.* **6**, 485–500.
Pearson, W. R. (1995). Comparison of methods for searching protein sequence databases. *Protein Sci.* **4**, 1145–1160.
Pearson, W. R. (1996). Effective protein sequence comparison. *Methods Enzymol.* **266**, 227–258.
Pearson, W. R. & Lipman, D. J. (1988). Improved tools for biological sequence comparison. *Proc. Natl Acad. Sci. USA,* **85**, 2444–2448.
Sander, C. & Schneider, R. (1991). Database of homology derived protein structures and the structural meaning of sequence alignment. *Proteins: Struct. Funct. Genet.* **9**, 56–68.
Siddiqui, A. S. & Barton, G. J. (1995). Continuous and discontinuous domains: an algorithm for the automatic generation of reliable protein domain definitions. *Protein Sci.* **4**, 872–884.
Smith, T. F. & Waterman, M. S. (1981). Identification of common molecular subsequences. *J. Mol. Biol.* **147**, 195–197.
Sowdhamini, R. & Blundell, T. L. (1995). An automatic method involving cluster analysis of secondary structures for the identification of domains in proteins. *Protein Sci.* **4**, 506–520.
Vingron, M. (1996). Near-optimal sequence alignment. *Curr. Opin. Struct. Biol.* **6**, 346–352.
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.
Waterman, M. S. & Vingron, M. (1994a). Rapid and accurate estimates of statistical significance for database searches. *Proc. Natl Acad. Sci. USA,* **91**, 4625–4628.
Waterman, M. S. & Vingron, M. (1994). Sequence comparison significance and poisson approximation. *Stat. Sci.* **9**, 401–418.

Appendix

Probability of a given value of sequence identity or similarity: approximation of the complementary error function erfc

The complementary error function $\mathrm{erfc}(x)$ is a cumulative probability of the normal (Gaussian) distribution with the mean value equal to 0 and standard deviation equal to $\sqrt{2}$. It represents doubled probability for the normally distributed random to have a value higher than $x\sqrt{2}$ standard deviations over the mean value and has the following integral representation:

$$\mathrm{erfc}(x) = \frac{2}{\sqrt{\pi}} \int_x^\infty \mathrm{e}^{-t^2}\, \mathrm{d}t$$

It is illustrated by Figure 6b (broken line) of the main text.

The probability of structural significance at a given value of alignment criterion Y can be calculated in two steps. First, calculate a normalized value y:

$$y = \frac{Y - m_Y(L)}{\sigma_Y(L)}$$

where $m(L)$ and $\sigma(L)$ are given by equations (3) to (5) with coefficients from Table 1 of the main text.

Second, calculate probability P_Y as:

$$P(y) = \frac{1}{2}\,\mathrm{erfc}\left(\frac{y}{\sqrt{2}}\right)$$

The erfc function can be approximated at positive y values as:

$$\mathrm{erfc}(y) \approx \frac{1}{1 + y/2} \cdot \mathrm{e}^{-y^2 - \frac{0.774y}{1+0.617y}}, y > 0$$

Therefore the probability can be calculated as:

$$P(y) \approx \frac{1}{2 + 0.7y} \cdot \mathrm{e}^{-\frac{y^2}{2} - \frac{0.547y}{1+0.436y}}, y > 0.$$

Note that the erfc function descends faster than exponentially and this approximation is quite accurate; in the most useful range $y > 3$ the relative inaccuracy is less than 1% and in the range $0 < y < 3$ it is less than 4%.

Probability of a given global alignment score: the extreme value distribution

The probability that criterion $\boldsymbol{A}$ is higher than t can be approximated by the extreme value distribution:

$$P_A(> t) = 1 - \mathrm{e}^{-\mathrm{e}^{-\xi y}}$$

where:

$$y = \frac{t - m_A(L)}{\sigma_A(L)} \text{ and } \xi = 1.618 \qquad \text{(A1)}$$

It is known that for the extreme value distribution the first and the second moments (the mean and the standard deviation) are expressed through the parameters u and λ of the distribution: $u = \langle x \rangle - 0.4500 \cdot \sigma$, $\lambda = 1.2825/\sigma$. We found that the $\boldsymbol{A}$-distributions could approximately be described by the extreme

value distribution (A1) with $\zeta = 1.618$ at $y > 0$. Parameter ξ was derived only for scores higher than the average ($y > 0$), the interesting part of the distribution, and hence is slightly different from the expected coefficient of 1.2825. For scores lower than the average ($y < 0$), the deviations from the extreme value distribution were substantial.

Edited by F. E. Cohen

(Received 10 December 1996; received in revised form 7 July 1997; accepted 7 July 1997)