STATISTICAL SIGNIFICANCE OF UNGAPPED SEQUENCE ALIGNMENTS

N.N. ALEXANDROV, V.V. SOLOVYEV
Amgen Inc, Thousand Oaks CA, USA

Statistical significance of a local sequence alignment depends not only on the similarity score and on the sequence lengths, but also on a length of the alignment. Dependence of the alignment significance on the length of the sequences has been analyzed earlier, and is based on the idea that the longer sequences have more chances to share a local similarity with a bigger score. To the best of our knowledge, a dependence of the statistical significance on the length of an alignment has not been used in selecting the best aligmments. We have applied to real proteins formulas for assessing the statistical significance of ungapped local alignments. Let L be a length of the alignment, then the expected value of a similarity score is S_{exp} = <m> * L, where <m> is the expected similarity between two randomly chosen residues. Value of <m> can be calculated from a similarity (substitution) matrix M and amino acid frequencies P: $<m> = \sum_{ij} p_i * p_j * m_{ij}$. The probability of observing a score S greater than or equal to x for an alignment of length L is given by the normal distribution: $Prob(S \geq x) = 1 - \int N((S-S_{exp})/\sigma) = 1 - \int N((S-<m>*L)/\sigma_m\sqrt{L})$, where σ_m is a standard deviation of m. From these formula, we conclude, that we should select the best alignment using a normalized value of the similarity score as follows: $S' = \max\{(S-<m>*L)/\sigma_m\sqrt{L}\}$. The proposed normalization of the similarity score has been tested on the representative benchmark. To evaluate a performance of the normalization, we have calculated several measures of the recognition quality. Our normalization has improved all these measures. This procedure is important for choosing the correct alignment for homology modelling as well as for selecting distantly related sequences in databases.

1. Introduction

Sequence comparison based on sequence alignment is the most powerful tool for inferring the biological function of a gene or the protein that it encodes (Pearson, 1996). A central question in sequence comparison is the statistical significance of the observed similarity. For local alignment without gaps this problem has been approached by several groups (Arratia et al.,1988; Karlin & Altschul, 1990). Local similarity scores are described by the extreme value distribution. The probability of finding a score $S' > x$ can be calculated as: $P(S' \geq x) = 1-\exp(-e^{-x})$, where S' is the normalized similarity score $S'=\lambda*S-\ln(K*m*n)$; m and n are the lengths of sequences; λ and K are derived from the scoring matrix and the amino acid composition of sequences (Karlin & Altschul,1990; Altschul et al.,1994). Thus, in this theory the statistical significance of an alignment depends only on sequence lengths, letter distribution and scoring weights. In our work we show that it is necessary to include the length of alignment in such estimations. It is especially important when we try to recognize a short fragment (like a domain or EST encoded protein sequence interrupted by frame shift error) surrounded by unrelated sequences. Intuitively we can think that the match within such domain should be more

significant than a longer region when both of them have the same score, because a longer alignment can have bigger fluctuation of score values by chance. Using an appropriate random model, we present numerical formulas for normalizing a similarity score taking into account the length of alignment. Analysis of data base search results on a representative set of protein sequences has proved that the suggested normalization improve, in average, the quality of recognition.

2. Method

We will consider the random sequence model where the elements of a sequence are chosen independently from an alphabet of a letters with respective probabilities p_i (i=1,..,a). The pair of letters i of the first sequence and j of the second sequence occurs with probability p_i*p_j. Let the score for such a paring be m_{ij}. Usually the score matrix for aligning pairs of amino acids provides negative expected pair score $<m>=\sum_{ij}p_ip_jm_{ij}$, that permits to apply extreme value distribution statistics (Karlin & Altschul, 1990). For example, Dayhoff's score matrix PAM-250 has $<m>=-0.79$ and standard deviation of this score $\sigma_m= \sqrt{v} = \sqrt{\sum_{ij}(m_{ij}-<m>)^2p_ip_j}$ is equal to 2.81. For an identity matrix, when all the elements are zeros, except the diagonal ones, which equal to 1: $<m> = 0.058$, $\sigma_m = 0.23$.

For a given ungapped alignment of length L the expected aggregate score S of the alignment is $S = <m>*L$. The score S, as the sum of many independent random variables (for large enough L) yields a normal distribution N with the expectation S_{exp} and the standard deviation σ. The variance V of N is the sum of variances v: $V = \sum_L v = v*L$ and thus $s = \sqrt{V} = \sigma_m*\sqrt{L}$.

The probability of observing a score S greater than or equal to x for an alignment of length L purely by chance is given by the formula:

$$Prob(S \geq x) = 1 - \int N((S-S_{exp})/\sigma) = 1 - \int N((S-<m>*L)/\sigma_m\sqrt{L}).$$

In other words, the significance of an alignment depends on its length and to rank the alignments properly, we should select the best alignment using normalized similarity score on a length of alignment:

$S' = \max\{(S-<m>*L)/\sigma_m\sqrt{L}\}$.

Notice that the alignment significance also depends on $<m>$ and σ_m, i.e. on the scoring matrix and amino acid frequencies.

Let us consider some interesting properties of such normalization. Fig. 1 shows behavior of the normalized score S' for several typical values of the raw score S. S' has a minimum S'_{min} at the length L_{min}, which is different for different S. We can see that for $L< L_{min}$, the shorter the alignment, the greater its statistical significance, and the opposite is true for $L>L_{min}$.

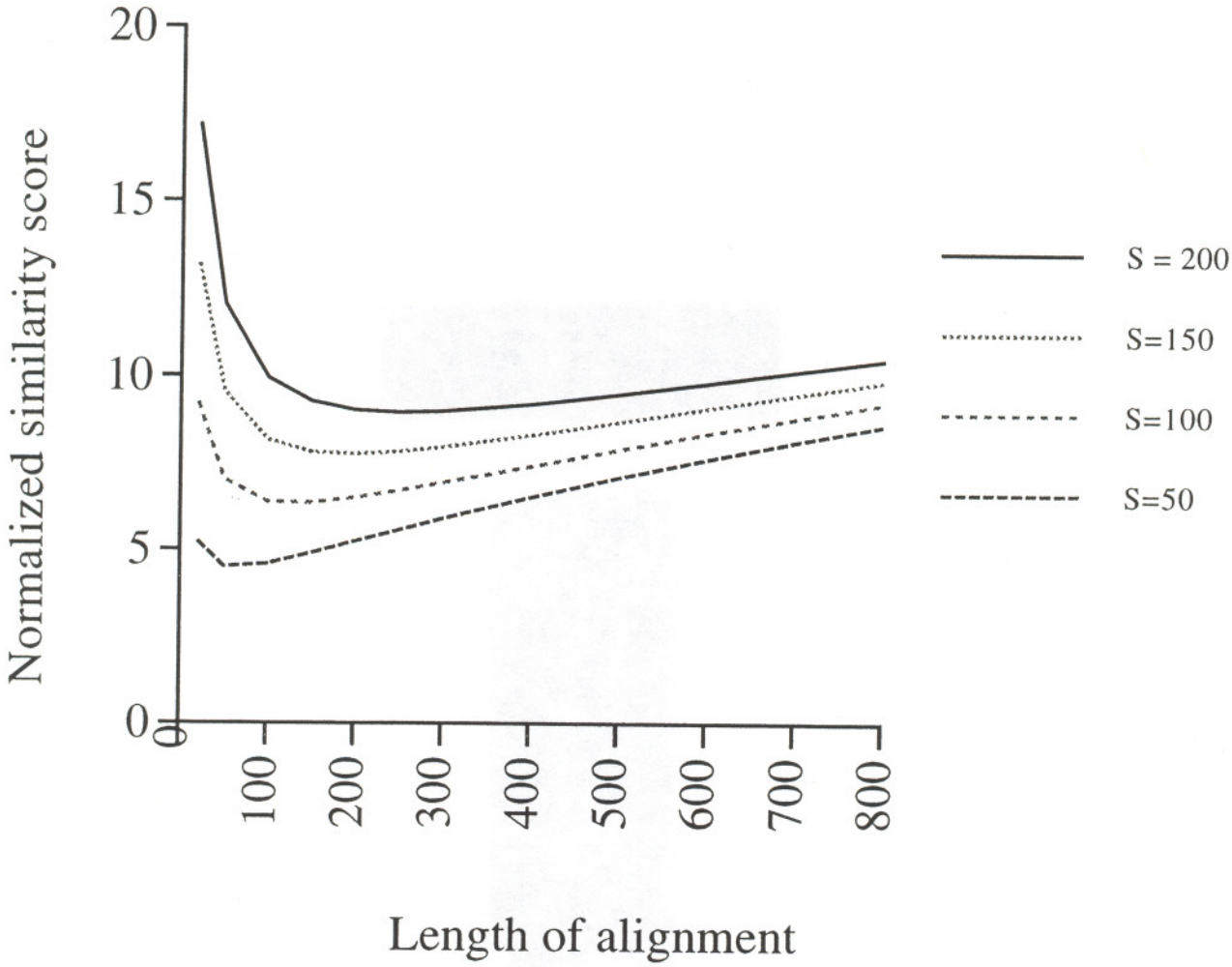

Figure 1. Dependence of the normalized similarity score on the length of alignment. The dependence is particular important for short alignments. From this plot one can see, that short alignment with a raw score of 100 can be more statistically significance than a longer alignment with a raw score of 150.

The curves in figure 2 are steeper for the $L < L_{min}$ than they are past L_{min}. In other words, the effect of normalization is greater for small fragments of sequences. L_{min} could be calculated analytically that gives $L_{min} = -S/\langle m \rangle$ (fig. 2).

For reasonable sizes of the aligned fragments with $L < L_{min}$, not unusual for natural

proteins (of about 100 amino acids), a raw similarity score underestimates the relative significance of these alignments, and can lead to errors in selecting the correct alignments, and, consequently, in detecting related sequences in databases.

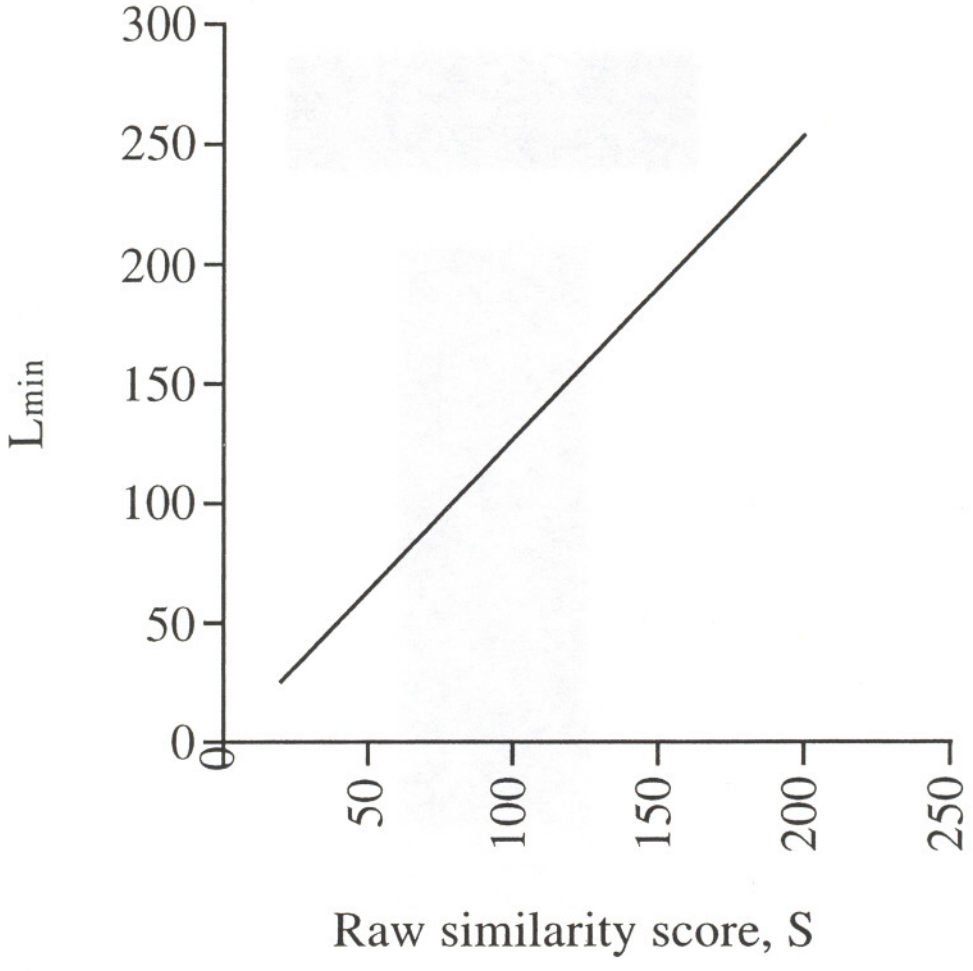

Figure 2. An expected length of the alignment (L_{min}) is proportional to the raw similarity score.

What is the threshold between significant alignments and random matches in terms of the percentage of the identical residues in the alignment? To compute an expected number of the identical residues we apply formula $S_{exp} = <m>*L$. A level of the significance is expressed in the units of standart deviation Z. Thus alignment is significant if a number of the identical residues in it $S \geq S_{exp} + Z*\sigma = <m>*L + Z*\sqrt{L}*\sigma_m$, where $<m> = 0.058$, $\sigma_m = 0.23$. The same value, expressed as a percentage of identical residues is: $I = 100*S/L = <m> + Z*\sigma_m/\sqrt{L}$. In Figure 3 we plot this threshold for several significance levels. A shape of the curves is very similar to the emperical plot derived by Sander&Schneider, 1991.

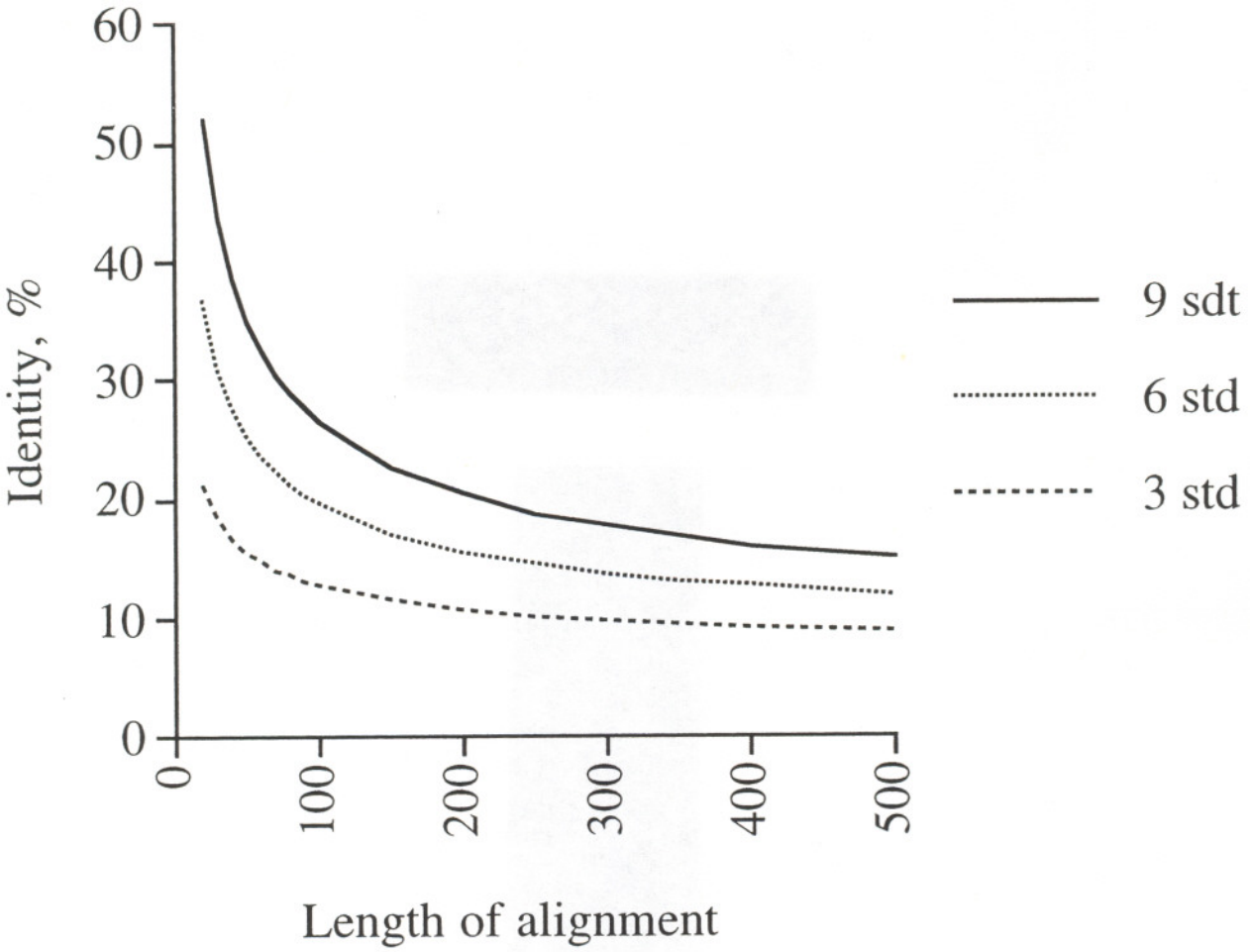

Figure 3 A threshold between significant alignments and random matches as a function of the alignmennt length. Three curves present three levels of significance: three, six, and nine standard deviations. Even at the very high confidance level of nine standard deviation long ungapped alignments (>200 a.a.) with 20% identity are significant.

In this work we have applied the statistics for ungapped alignment, but we plan to extend the estimates for a general case as it was done for extreme value distribution statistics (Waterman & Vingron, 1994; Altschul & Gish, 1996; Pearson, 1996), which was initially devised for ungapped alignments. Ungapped alignments have clear practical value, being implemented in BLAST -- a widely used program for rapid sequence comparison (Altschul et al., 1990).

3. Results

We have tested the performance of our normalization scheme on a benchmark, used by Pearson to compare different methods for sequence comparison. The benchmark

consists of 67 family representatives and a database of about 12,000 amino acid sequences. Each family representative is compared with all the database sequences. A recognition quality is evaluated based on the discrimination of the related sequences from the others. An average recognition quality for all 67 sequences assesses a quality of the method.

Because we are interested in detecting local similarities with variable alignment length, we have chosen this benchmark, in which query sequences contain only a fragment of the real protein sequence, whereas the rest of the sequence is random. Random parts of the sequences were generated according to the amino acid distribution in the whole database.

3.1 Measuring the quality of recognition

For each query sequences three measures of the recognition quality were used: (i) an equivalence number, (ii) a separation score (S-score), and (iii) an error score.

Equivalence number has been used by Pearson in his benchmark and is the number of false positives (or the number of false negatives) when the threshold value of the similarity score is chosen to make these numbers equal. In the case of perfect recognition, the equivalence number is zero, in the worst case, it equals to the size of a family.

Separation score shows an overall separation between a family and the rest of the database sequences. it is computed as:

S-score = $(\langle S_f \rangle - \langle S_o \rangle)/0.5(\sigma_f + \sigma_o)$,

where $\langle S_f \rangle$ and $\langle S_o \rangle$ are average similarity scores for the family and the other sequences; σ_f and σ_o are average standard deviations of the family and of the other sequences, respectively. The better is the separation between a family and the other proteins, the bigger is the S-score.

While S-score characterizes the overall separation, an error score concentrates on a twilight zone, where related and unrelated sequences are mixed together:

E-score = $(\langle Z_{fp} \rangle - \langle Z_{fn} \rangle)/Z_{eq}$,

where Z_{eq} is a threshold of similarity score corresponding to the equivalence number; Z_{fp} and Z_{fn} are sums of similarity scores of all false positives and of all false negatives. Similarity scores for this formula are expressed in the units of standard deviation (Z-scores). When all the family proteins are separated from the rest of the database, the E-score =0, a bad separation results in the large E-score.

In Figure 4 we show two distributions for query sequence TISYO: one for raw and another for normalized similarity scores.

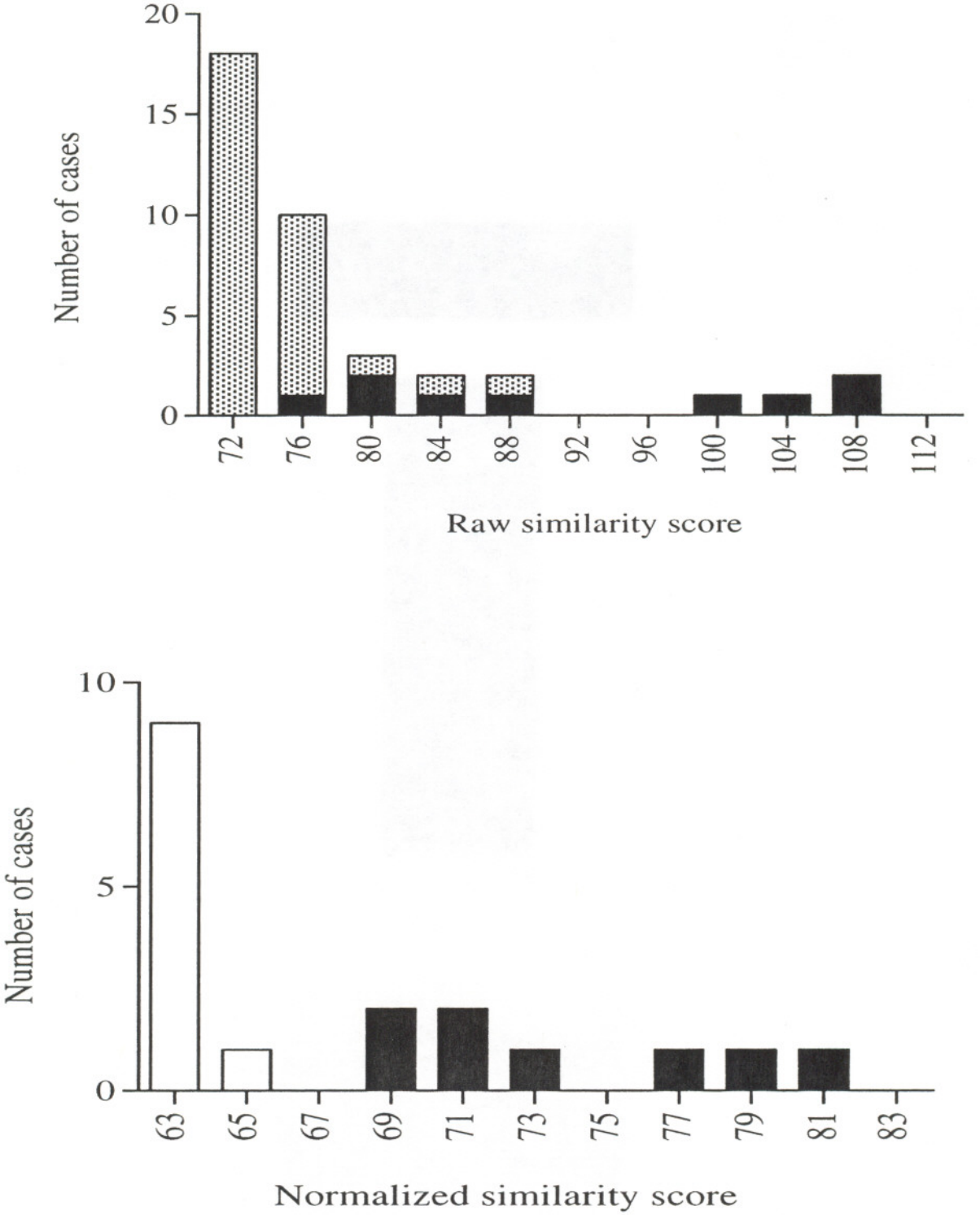

Figure 4. Distribution of the raw and normalized similarity scores for proteinase inhibitor family (query sequence TISYO). Only a region on a border between related proteins (black bars) and the others (white bars) is shown. Equivalence number for the raw score equals to three, whereas for the normalized score it is zero. S- and E-scores are also better for the normalized similarity score (table 1).

Table 1. Comparison of the discriminating power of raw and normalized similarity scores for 67 query proteins.

PIR entry	Superfamily	Family size	Equivalence numbers		S-score		E-score	
			raw	norm	raw	norm	raw	norm
ACHUA1	acetylcholine receptor	23	0	0	3.56	3.79	0	0
AJHUQ	glutamate--ammonia ligase	39	12	12	2.24	2.40	12.18	11.09
AZBR	plastocyanin	38	27	27	1.05	.97	35.17	36.42
CCHU	cytochrome c	142	29	21	2.70	2.86	28.88	22.18
CFKKA	phycocyanin	25	1	3	1.67	1.94	0.16	0.72
CYBOA	alpha-crystallin	67	14	12	3.19	3.25	12.57	10.35
CYBOB	beta-crystallin	21	0	0	2.28	3.02	0	0
DEHUAA	alcohol dehydrogenase	27	11	11	1.99	1.97	10.24	9.64
DEHUGL	glyceraldehyde-3-phosphate dehydrogenase	46	1	1	5.80	5.50	0.54	0.63
DEPGLH	L-lactate dehydrogenase	26	4	4	2.06	2.10	2.68	2.99
DJHUAC	herpesvirus DNA-directed DNA polymerase	23	21	21	.60	.65	17.38	17.87
FEPE	ferredoxin 2[4Fe-4S]	96	59	55	.87	.97	76.2	78.03
FOVWH3	AIDS-related virus gag polyprotein	91	65	64	.96	1.04	65.14	57.72
GCHU	glucagon	44	13	11	1.87	2.35	8.38	6.66
HAHU	globin	505	50	55	3.20	2.98	71.06	69.63
HLHUB2	class I histocompatibility antigen	71	41	41	1.51	1.52	42.95	38.46
HMIVV	influenza virus hemagglutinin	85	18	18	3.05	3.03	24.31	23.22
HNNZS	paramyxovirus hemagglutinin-neuraminidase	49	7	7	2.46	2.72	6.29	5.18
HSHU1B	histone H1	22	5	6	2.43	2.64	3.52	2.74
IJHUCN	cadherin	28	2	2	2.23	2.46	0.97	0.88
IPHU	insulin	69	4	3	2.70	4.02	3.56	3.14
IVHUI6	interferon alpha	39	1	1	5.05	5.25	0.24	0.19
K1HUAG	immunoglobulin V region	280	107	91	1.96	2.06	94.07	80.2
K3HU	immunoglobulin C region	74	14	16	1.75	2.15	10.66	9.73
KIBET	herpesvirus thymidine kinase	27	4	3	1.99	2.22	1.47	1.26
KRHUE	cytoskeletal keratin	32	5	6	2.75	2.60	4.08	4.64
LCHU	prolactin	20	4	2	2.10	2.43	1.5	0.7
LNHU1	hepatic lectin	20	17	17	.72	.66	19.63	19.72
LUHU	annexin I	27	0	0	4.87	4.95	0	0
LWBOA	H+-transporting ATP synthase lipid-binding protein	28	11	13	2.41	2.24	3.25	6.05
LZHU	lysozyme c	28	2	2	4.31	4.46	1.95	1.82
MFNZS	parainfluenza virus matrix protein	24	10	10	2.02	2.05	6.58	7.18
MNIV2K	influenza virus nonstructural protein NS2	22	2	2	5.59	5.43	1.44	1.6
N2KF1U	snake toxin	109	19	10	1.91	2.43	9.4	6.21
NKVLAH	hepatitis B virus core antigen	25	2	1	3.84	4.02	0.23	0.29
NMIV	influenza virus exo-alpha-sialidase	27	14	14	1.65	1.74	12.05	11.95
NRBO	pancreatic ribonuclease	40	1	0	5.41	6.05	0.12	0
NTSRIA	scorpion neurotoxin	26	9	7	1.54	1.62	11.54	10.89
O4HUD1	cytochrome P450	35	12	11	2.22	2.13	10.84	10.22
OKHU2C	kinase-related transforming protein	183	61	60	1.35	1.51	34.91	34.46
OOHU	vertebrate rhodopsin	167	57	57	1.41	1.76	32.47	31.82
P2WL	papillomavirus L2 protein	27	1	1	2.00	2.58	0.04	0.12
PEHU	pepsin	20	4	4	2.55	2.64	3.4	3.52
PSHU	phospholipase A2	58	31	28	1.50	1.69	21.66	18.2
PWHU6	H+-transporting ATP synthase protein 6	23	3	3	2.09	2.15	1.63	1.47
PWHUA	H+-transporting ATP synthase alpha chain	46	3	4	2.76	2.84	0.82	1.38
QQBE1L	herpesvirus glycoprotein B	23	2	1	1.69	1.85	1.2	1.07
QRECBD	inner membrane protein malK	43	37	35	.72	.92	20.62	22.83
R6HUP2	rat acidic ribosomal protein P1	41	21	23	1.46	1.41	16.82	16.72

RKMDS	ribulose-bisphosphate carboxylase small chain	77	2	0	2.50	3.28	0.18	0
SMHU2	metallothionein	21	5	5	2.97	3.25	2.74	2.26
TISYO	Bowman-Birk proteinase inhibitor	23	3	0	2.94	3.45	0.28	0
TPHUCS	calmodulin	116	25	22	1.89	2.15	13.02	12.47
TRRT1	trypsin	72	31	17	1.47	2.31	20.2	9.21
TVHUM	myc transforming protein	91	67	67	.83	.97	75.01	74.24
TVHURA	ras transforming protein	45	5	9	2.55	2.81	2.29	3.76
TYTUY2	protamine Y2	24	17	5	2.72	3.13	4.63	0.84
UART	lipocalin	21	11	11	.96	1.01	10.4	10.08
VGIHE2	coronavirus E2 glycoprotein	20	1	1	2.19	2.56	0.99	1.05
VGNZSV	parainfluenza virus cell fusion protein	45	25	26	1.28	1.26	19.13	22.74
VHIV34	influenza virus nucleoprotein	32	5	5	4.61	4.58	2.67	2.01
VPXRWA	rotavirus outer layer protein VP3	35	5	5	3.05	2.92	3.58	3.34
W2WLE	papillomavirus E2 protein	27	26	26	.33	.31	31.1	28.88
W6WL18	papillomavirus E6 protein	29	2	2	3.56	3.55	0.63	0.48
W7WLHS	papillomavirus E7 protein	26	0	0	3.27	3.30	0	0
XHHU3	antithrombin III	25	5	8	2.05	1.78	2.17	4.58
XURT8C	glutathione transferase	106	81	70	1.01	1.13	83.23	71.09

To measure the significance of the difference between two methods, we used the same sign-test as Pearson did. Let N_1 to be a number of families, for which the first method is better, and N_2 to be a number of families, for which the second method is better. The z-value is calculated from the formula:

$$z = [\max(N_1,N_2) - \mu]/\sigma,$$

where $\mu = (N_1 + N_2)P$, $\sigma = [(N_1 + N_2)P(1-P)]^{1/2}$, and $P = 0.5$.

The average results for all three measures are in favor to the normalized similarity score (Table 2).

Table 2. Results of the sign test for equivalence number, S-, and E-scores.

Measure	Number of cases when the normalized score is		z-value	Probability
	better	worse		
Equivalence number	24	11	2.20	$2.8 \cdot 10^{-2}$
S-score	50	17	4.03	$5.5 \cdot 10^{-5}$
E-score	43	20	2.90	$3.8 \cdot 10^{-3}$

4. Conclusion

We believe that the proposed normalization is especially important in the analysis of distantly related sequences. Estmation of a given alignment significance was presented by Waterman (1995), however in this article we propose to use itduring the search for selecting the best alignment. To compare this alignment with any other when searching a database we should also apply extreme value distribution statistics. The correct estimation of the significance of their alignment is crucial for separation related proteins from the others in the twilight zone on a boundary between them. This zone is often saturated by alignments with approximately the same scores and even a small valid score correction can essentially improve the recognition quality. Another possible application of this normalization is a homology modelling, where we have to choose a correct alignment for the related proteins. Selection the right alignment is one of the major problems for many threading algorithms, and our normalization can also lead to better understanding of the relationships between protein sequence, three-dimensional structure and biological function.

Acknowledgments.

We thank Bill Pearson and Phil Green for very helpful discussions.

References

Altschul S.F., Gish W. (1996) Local alignment statistics. In Methods in Enzymology, 266, 460-480.

Altschul S.F., Gish W., Miller W., Mayers E.W., and Lipman D.J. (1990) Basic local alignment search tool. J. Mol. Biol. 215, 403-410.

Arratia R., Morris P., Waterman M.S. (1988) Stochastic scrabble: large deviations for sequences with scores. J. Appl.Probab. 25, 106-119.

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.

Pearson W.R. (1996) Effective protein sequence comparison. In Methods in Enzymology, 266, 227-259.

Sander C., Schneider R. (1991) Database of homology-derived protein structures and the structural meaning of sequence alignment. Proteins 9, 56-68.

Waterman M.S., Vingron M. (1994) Rapid and accurate estimates of statistical significance for sequence data base searches. Proc. Natl. Acad. Sci. USA, 91, 4625-4628.

Waterman M.S. Introduction to computational biology. 1995, Chapman and Hall.