BIOINFORMATICS ORIGINAL PAPER Vol. 22 no. 23 2006, pages 2865–2869
doi:10.1093/bioinformatics/btl512

Sequence analysis

Application of a simple likelihood ratio approximant to protein sequence classification

László Kaján[1,†], Attila Kertész-Farkas[2], Dino Franklin[1], Neli Ivanova[1], András Kocsor[2] and Sándor Pongor[1,3,*]

[1]Bioinformatics Group, International Centre for Genetic Engineering and Biotechnology, Padriciano 99, I-34012 Trieste, Italy, [2]Research Group on Artificial Intelligence of the Hungarian Academy of Sciences and University of Szeged, Aradi vértanúk tere 1., H-6720 Szeged, Hungary and [3]Bioinformatics Group, Biological Research Centre, Hungarian Academy of Sciences, Temesvári krt. 62, H-6701 Szeged, Hungary

Received on April 19, 2006; revised and accepted on October 3, 2006
Advance Access publication November 7, 2006
Associate Editor: Martin Bishop

ABSTRACT

Motivation: Likelihood ratio approximants (LRA) have been widely used for model comparison in statistics. The present study was undertaken in order to explore their utility as a scoring (ranking) function in the classification of protein sequences.

Results: We used a simple LRA-based on the maximal similarity (or minimal distance) scores of the two top ranking sequence classes. The scoring methods (Smith–Waterman, BLAST, local alignment kernel and compression based distances) were compared on datasets designed to test sequence similarities between proteins distantly related in terms of structure or evolution. It was found that LRA-based scoring can significantly outperform simple scoring methods.

Contact: pongor@icgeb.org.

Supplementary information: http://www.inf.u-szeged.hu/~kfa/lra06/.

1 INTRODUCTION

Classification of protein sequences is a crucial task in computational genomics. If a newly determined sequence has a close relative in one of the known protein families, classification can be simply carried out using methods of sequence similarity searching, such as BLAST (Altschul *et al*., 1990). Discovery of distant sequence similarities, such as the assignment of a new subgroup of sequences to a known class, usually requires more sophisticated and more computationally intensive methods, like profiles or HMM classifiers (Krogh *et al*., 1994). The goal of the present work is to test the utility of likelihood ratio (LR) as a scoring function in protein classification based on similarity searching using an a priori classified sequence database.

The LR is a familiar concept in statistics for testing the differences between competing models (Hoel, 1962). It is well-known that the Bayes decision rule for binary classification can be reformulated as a log-LR test with a decision threshold (Duda *et al*., 2001). In the framework of protein classification LR can be defined as follows:

$$\mathrm{LR}(x) = \frac{P(x\,|\,+)}{P(x\,|\,-)}, \tag{1}$$

where x is the query protein, the '+' symbol represents a particular protein class, '−' denotes the complementer class, while $P(x\,|+)$ and $P(x\,|-)$ stand for class-conditional probability density functions. Now assuming a preset threshold τ, the following decision rule (i.e. a log-LR test) can be applied:

$$\begin{cases} \text{if } \log \mathrm{LR}(x) > \tau \text{ select class '+'}, \\ \quad \text{otherwise select class '}-\text{'}. \end{cases} \tag{2}$$

In practice, the tuning of threshold τ is often based on a priori information and/or heuristic algorithms. The optimality for the above test follows from the Neyman–Pearson lemma, which guarantees that in simple point hypothesis testing, an LR based test has the most power among all tests of a given size (Hoel, 1962). A variant of this test, with a fixed τ threshold was used in protein structure classification by Røgen and Fain, (2003).

When a query protein sequence is compared to an a priori classified database, the top list of similarities usually contains representatives of several classes, so we have a multiclass problem instead of a two-class problem. The difficulty comes from the fact that the classes are not very well characterized and often very different in terms of size (number of members), sequence length as well as the level of within-group similarities. This situation is not uncommon in machine learning, and the present work was inspired by a simple likelihood ratio approximant (LRA) developed for computer vision applications (Claus and Fitzgibbon, 2004):

$$\mathrm{LRA}(x) \sim \left[\frac{d(x,-)}{d(x,+)}\right], \tag{3}$$

where d is the minimal distance of the query object x taken from the highest ranking members of the first '+' and second '−' classes, respectively. It is noted that Equation (3) implies the assumption that the probability of an object belonging to a class is inversely proportional to its minimal distance from the members of that class (a formal proof is outlined in the Supplementary materials). Second, Claus and Fitzgibbon used Equation (3) as a scoring (ranking)

*To whom correspondence should be addressed.
†Present address: BioInfoBank Institute, 60-744 Poznan, Poland

function, not as a statistical test and for this kind of application it is not necessary to define a threshold τ.

Our general goal is to develop LRAs for protein classification and to use them as scoring functions in similarity searching. For the comparison of two sequences most of the popular programs (BLAST, Smith–Waterman, etc.) use similarity measures that are maximal for identical and zero or minimal for different sequences. The similarity measures are thus inversely related to the distance or dissimilarity measures, that are zero for identical sequences and very large for distantly related sequences. Taking this inverse relationship into consideration we suggest the following formula for an LRA scoring function:

$$\mathrm{LRA}(x) \sim \left[\frac{s(x,+)}{s(x,-)}\right], \tag{4}$$

where s gives the similarity taken between the query sequence x and the highest ranking members of the first '+' and second '−' classes, respectively. The goal of this paper is to develop a family of LRAs based on Equations (3 and 4) and to test them in various scenarios of protein classification. We are particularly interested in the analysis of distant similarities, especially in testing a classifier's ability to recognize a new subgroup in a given protein class. It is noted, that Equation (4) assumes a reciprocal relationship between s and d, which may not necessarily be the case. In this work we will use two additional monotone decreasing functions to interconvert s and d into each other.

The rest of the paper is structured as follows. The definitions of the LRAs are described in section 2. Then section 3 describes the elements of the test system, including (1) the datasets used to represent various examples of distant protein similarities, (2) the sequence comparison methods; (3) the classification algorithms and (4) the performance measures used. Finally, section 4 contains the results and the discussion.

2 DEFINITION OF LRA

If we have a sequence similarity score, such as an alignment score computed with the BLAST, Smith–Waterman, Needleman Wunsch algorithms, Equation (4) can be used to construct the LRA function. If we have a sequence distance measure, such as a compression based distance or a Euclidean distance of features, we can use Equation (3). If the original suggestion of Claus and Fitzgibbon is valid also in protein classification, we can expect better results when the LRA measures computed according Equations (3 or 4) are used instead of the original sequence similarity or dissimilarity scores, respectively.

Let us now suppose that d and s are related to each other by some monotone decreasing function f so that $d \sim f(s)$ and $s \sim f(d)$. In other terms, we will try other functions than the simple reciprocal dependence implied by Equations (3 and 4). In particular we will use

$$f_1(x) = e^{-\beta x} \tag{5}$$

$$f_2(x) = \frac{1}{\log(2)}\left[-\beta d + \log\left(1 + e^{\beta x}\right)\right], \tag{6}$$

where x is either s or d, while β is a positive, tunable parameter. f_1 and f_2 map to [0,1] and the β parameter regulates the slope. The f functions transform similarity measures to dissimilarity measures and vice versa. In particular, if x is a similarity measure, $f(x)$ will be a dissimilarity measure, that can be used to compute an LRA according to Equation (3). And vice versa, if x is a dissimilarity measure, $f(x)$ will be a similarity measure that can be used in conjunction with Equation (3).

3 DATASETS AND ALGORITHMS

3.1 Datasets

Dataset A. The sequence dataset of Liao and Noble was designed to test distant protein similarities (Liao and Noble, 2003). This set consists of 4352 protein domain sequences (lengths ranging from 20 to 994 amino acids) selected from the SCOP database (Andreeva *et al.*, 2004). A total of 54 classification tasks are defined on this dataset, each of them represented by positive and negative examples divided into training and test sets. In a given classification experiment, the members of a superfamily are the positive set from which the members of one particular protein family are the test set while the rest of the superfamily is the training set. This setup ensures that there is a little sequence similarity and no guaranteed evolutionary relationship between the test and the training sets. The same principle was used to construct the other two datasets.

Dataset B. The dataset was constructed from evolutionarily related sequences of a ubiquitous glycolytic enzyme, 3-phosphoglycerate kinase (3PGK, 358 to 505 residues in length). A total of 131 3PGK sequences were selected, which represent various species of the archaean, bacterial and eukaryotic kingdoms (Pollack *et al.*, 2005). Ten classification tasks were defined on this dataset as follows. The positive examples were taken from a given kingdom. One of the phyla (with at least five sequences) was the test set while the remaining phyla of the kingdom were used as the training set. The negative set contained members of the other two kingdoms subdivided in such a way that members of one phylum could be either test or train. The division is shown in Table 1.

Dataset C. This dataset is a subset of the COG database of functionally annotated orthologous sequence clusters (Tatusov *et al.*, 2003) In the COG database, each COG cluster contains functionally related orthologous sequences belonging to unicellular organisms, including archaea, bacteria and unicellular eukaryotes. For a given COG group, the positive test set included the sequences from three unicellular eukaryotic genomes, while the positive training set was compiled from the rest of the sequences in the group. Of the over 5665 COGs we selected 117 that contained at least 8 eukaryotic sequences (positive test group) and 16 additional prokaryotic sequences (positive training group). This dataset contained 17 973 sequences. The negative training/test sets were obtained by randomly assigning sequences from the remaining COGs to the two groups (see Supplementary material) Table 2.

3.2 Sequence comparison methods

Alignment-based sequence comparisons were carried out using version 2.2.4 of the BLAST program (Altschul *et al.*, 1990) with a cutoff score of 25. The Smith–Waterman algorithm (Smith and Waterman, 1981) was used as implemented in MATLAB (MathWorks, 2004), the program implementing local alignment

Table 1. 3-phosphogycerate kinase (3PGK) sequences (Dataset B)

Group	Subgroup	Positive		Negative	
		Test	Train	Test	Train
Archaea	Crenarchaeota	4	11	18	98
Archaea	Euryarchaeota	11	4	18	98
Bacteria	Actinobacteridae	5	68	33	25
Bacteria	Firmicutes	35	38	33	25
Bacteria	Proteobacteria	30	43	33	25
Eukaryota	Alveolata	4	39	15	73
Eukaryota	Euglenozoa	5	38	15	73
Eukaryota	Fungi	10	33	15	73
Eukaryota	Metazoa	12	31	15	73
Eukaryota	Viridaeplantae	8	35	15	73
Total:		131 sequences			

[a]The definition of taxonomic groups and subgroups is taken from (Pollack *et al*., 2005). See Supplementary materials for further details.

Table 2. Distribution of orthologous sequences in Dataset C.

COG No.	Name	Positive[a]		Negative	
		Train	Test	Train	Test
COG0631	Serine/threonine protein phosphatase	40	15	1456	1603
COG0697	DMT drug transporters superfamily	372	14	1550	1544
COG0814	Amino_acid_permeases	61	14	1492	1521
COG0847	DNA_polymerase III_epsilon_subunit	86	9	1498	1555
Total of 117 COGs		17973 sequences			

[a]The positive test sequences included eukaryotic proteins from *S. cerevisiae*, *S. pombe* and *E. cuniculi* in a given COG, the negative sequences were randomly chosen from the other COGs. See Supplementary Materials for further details.

kernel algorithm (Saigo *et al*., 2004) was obtained from the authors of the method.

The BLOSUM 62 matrix (Henikoff *et al*., 1999) was used in each case. The Smith–Waterman comparison data on Dataset A was taken from Liao and Noble (Liao and Noble, 2003). The BLAST scores on Dataset C were taken from the COG database (Tatusov *et al*., 2003).

Alignment-free sequence comparisons were carried out using compression based distance measures (CBMs). The formula for calculating CBMs using the length values of compressed strings was adapted from (Cilibrasi and Vitányi, 2005) and can be written in the following general form:

$$\mathrm{CBM}(X,Y) = \frac{C(XY) - \min\{C(X), C(Y)\}}{\max\{C(X), C(Y)\}}, \qquad (7)$$

where X and Y are sequences to be compared and $C(.)$ denotes the length of a compressed string, compressed by a particular compressor C, such as the LZW or the PPMZ algorithm (Kocsor *et al*., 2005). In this study the LZW algorithm was implemented in MATLAB while the PPMZ2 algorithm was downloaded from Charles Bloom's homepage (http://www.cbloom.com/src/ppmz.html).

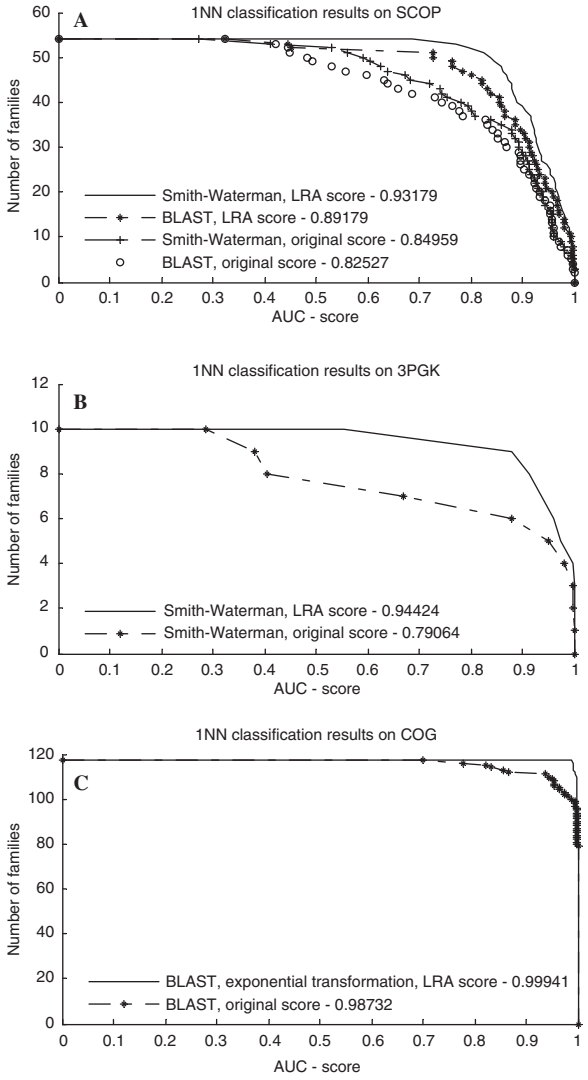

Fig. 1. Comparison of the scoring performance of various scoring methods by ROC analysis. (**A**) Dataset A (SCOP sequences), (**B**) Dataset B (Taxonomic classification of 3PGK sequences) and (**C**) (COG sequences). The Y-value represents the number of the families scoring better than the AUC value indicated on the *x*-axis. In this representation the curves running higher represent better performances.

Classifier algorithms Nearest neighbour (1NN) classification (Duda *et al*., 2001) is a technique whereby a query sequence is assigned to the a priori known class of the database entry that was found most similar to it in terms of a distance/similarity measure. <3NN> denotes a related method where the average of the top three scores, obtained by comparing a query to members of a given class, is used, and the query is assigned to the class with the best such value. Of the top three scores, only the non-zero values were included in the average calculation.

3.3 Performance evaluation

The evaluation was carried out via standard receiver operator characteristic (ROC) analysis that characterizes the performance of learning algorithms under changing conditions, such as

Table 3. Comparison of LRA scoring with simple scoring, using nearest neighbour classification

Algorithm used for sequence similarity/ distance calculations[a]	1NN				<3NN>			
	Original score	LRA scoring Without transformation Equations (3 or 4)	After transformation by Equation (5)[b]	Equation (6)[b]	Original score	LRA scoring Without transformation Equations (3 or 4)	After transformation by Equation (5)[b]	Equation (6)[b]
(A) Fold classification (Dataset A)								
Smith–Waterman	0.8496	0.9318	0.9309	0.9304	0.8636	**0.9447**	0.9440	0.9439
Local alignment kernel	0.8223	0.9509	0.9393	0.9393	0.8315	**0.9601**	0.9428	0.9428
BLAST	0.8253	0.8918	0.9197	0.9193	0.8296	0.9019	**0.9314**	0.9312
LZW	0.8060	0.8468	0.8464	0.8464	0.8264	**0.8627**	0.8623	0.8623
PPMZ	0.6625	0.7641	0.7778	**0.7779**	0.6793	0.7571	0.7707	0.7708
(B) Taxonomic classification (Dataset B)								
Smith-Waterman	0.7906	0.9442	0.9169	0.9164	0.7778	**0.9540**	0.9142	0.9139
BLAST	0.7922	0.9411	0.9157	0.9144	0.7780	**0.9514**	0.9129	0.9124
Local alignment kernel	0.7862	0.9393	0.9204	0.9200	0.7765	**0.9423**	0.9139	0.9137
LZW	0.7370	0.8658	0.8642	0.8642	0.7625	**0.8926**	0.8915	0.8915
PPMZ	0.7523	0.9004	0.9030	0.9030	0.7588	0.9024	**0.9051**	**0.9051**
(C) Functional classification based on BLAST scores (Dataset C)								
COG0631	0.8659	**0.9961**	**0.9961**	**0.9961**	0.8659	0.8663	**0.9961**	**0.9961**
COG0697	0.8566	0.8568	**0.9901**	**0.9901**	0.8565	0.8568	0.9891	0.9891
COG0814	0.8208	0.8212	**0.9909**	**0.9909**	0.8207	0.8212	0.9906	0.9906
COG0847	0.7773	**0.9919**	**0.9919**	**0.9919**	0.7773	0.7773	**0.9919**	**0.9919**
Average of 117 COGs	0.9873	0.9874	**0.9994**	**0.9994**	0.9873	0.9874	**0.9994**	**0.9994**

[a]The values represent the integral of the cumulative AUC of the 54 ROC curves calculated for the data (examples shown in Figure 1). The maximal values in each row are indicated in bold. 'Original score': the score of the original method was used for ranking. 'Without transformation' refers to ratio calculation according to Equation (4) for similarity scores and Equation (3) for distance measures. The references are given in the text.
[b]$\beta = 0.0001$

misclassification costs or class distributions (Egan, 1975). This method is especially useful for protein classification as it includes both sensitivity and specificity, based on a ranking of the objects to be classified (Gribskov and Robinson, 1996). In our case the ranking variable was the NN similarity or distance value (1NN or <3NN>) obtained between a sequence and the members of the positive training set. Briefly, the analysis was carried out by plotting sensitivity versus 1 − specificity at various threshold values, then the resulting curve was integrated to give an 'area under curve' or AUC value. We note that AUC = 1.0 for a perfect ranking, while for random ranking AUC = 0.5 (Egan, 1975). If the evaluation procedure contains several ROC experiments, one can draw a cumulative distribution curve of the AUC values (examples shown in Fig. 1). The integral of this cumulative curve, divided by the number of the classification experiments is in [0,1] interval, with the higher values representing the better performances (Liao and Noble, 2003). These values were used as 'AUC scores' (Table 3).

4 RESULTS AND DISCUSSION

The datasets and the classification scenarios chosen for the evaluation were selected so as to represent different types of distant protein sequence similarities. In Dataset A (a subset of SCOP), we attempt to recognize a protein family based on other families of the same superfamily. The domain sequences included in this dataset are variable in terms of length and often there is relatively little sequence similarity between the protein families. The situation is quite different with Dataset B. Here the sequences are uniform in length and are highly related to each other, i.e. there is a strong similarity both between the members of a given group (phylum) and between the various groups. In Database C the recognition tasks are designed to answer the question: can we annotate genomes of unicellular eukaryotes based on prokaryotic genomes? In this dataset the members of a given group have high similarity to each other while there is relatively little similarity between the various groups.

The ranking performance of the various methods is summarized in Figure 1 and Table 3. The columns of Table 3 shaded in gray contain the results obtained with the original scores; the white columns contain the results obtained with the various combinations of LR scoring. It is apparent that the LRA scoring substantially improves the ranking efficiency as indicated by the high cumulative AUC values.

It is also apparent that the performances of the various LRA schemes do not substantially differ among each other i.e. the choice of the monotone decreasing function does not seem to be critical. This is shown by the fact that the data obtained 'without transformation' [i.e. using Equation (3) for distance measures and Equation (4) for similarity measures] are nearly identical with those obtained after transformations employing Equations (5 or 6). This fact is also shown by the dependence of the results on the β parameter in Equations (5 and 6): In general there is a large range where the classification performance is independent of β (Fig. 2 and Supplementary materials).

As the datasets and the algorithms are quite different in nature, so we tend to believe that the consistent performance increase is due to LRA scoring. On the other hand, the results are dataset-dependent: they vary from group to group within each of the three datasets (see Supplementary materials for details). For instance it is conspicuous

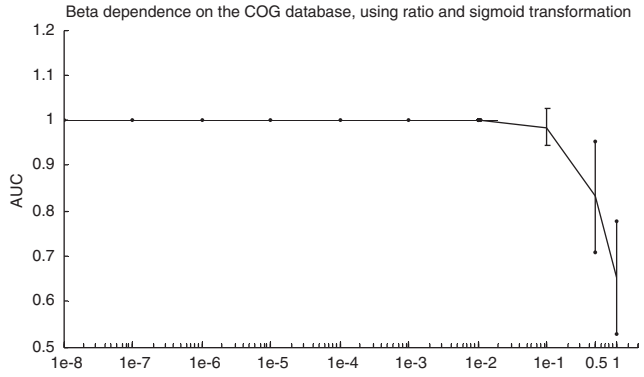

Fig. 2. Dependence of scoring performance (cumulative AUC) on the β parameter. Dataset C (COG sequences). The error bars denote standard deviation.

that Dataset C contain many groups that are perfectly separated (79 out of the 117 COGs have an AUC of 1.000 using BLAST without LRA scoring), as shown by the long vertical region of the curves in Figure 1C. Nevertheless, the increase caused by LRA scoring is apparent on the less-perfect COG groups of which we included four in Table 3, (C).

The group specific behavior is also apparent in how the results depend on the value of the β parameter. In general, we find that there is a large range where the classification performance is independent of β, but as the value approaches to one, there is a substantial variation between the groups. Typically (cf. Fig. 2), there is a sometimes substantial decrease in the AUC value, but in some cases there is a slight increase (details in Supplementary materials).

Summarizing we can conclude that LRA scoring provided a consistent performance increase in the protein sequence classification tasks analyzed in this work. The extent of the improvement seems to depend on the protein group as well as on the database. We note that the AUC value characterizes the ranking performance of a variable (in our case the LR score), but it does not describe the actual performance of any particular classifier that can be built using that variable. For practical purposes one can build classifiers using threshold values, such as the empirical score or *E*-value thresholds used in conjunction with BLAST (Altschul *et al*., 1990), or by using any of the database-dependent optimization techniques (Duda *et al*., 2001). Finally we mention that the principle described here can be easily implemented, and the fact that it applies to a wide range of scoring methods and classification scenarios makes us hope that it will be applicable also to other areas of protein classification.

ACKNOWLEDGEMENTS

A.K. was supported by the János Bolyai fellowship of the Hungarian Academy of Sciences. D.F. is on leave from Computer Science Department, Campus de Catalão and Federal University of Goiás, Brazil. The work at the Szeged Biological Centre was partly supported by grants of the Hungarian Ministry of Education (OMFB-01887/2002, OMFB-00299/2002). Work at ICGEB was supported in part by grants from the Ministero dell'Università e della Ricerca (D.D. 2187, FIRB 2003 (art. 8), ''Laboratorio Internazionale di Bioinformatica'').

Conflict of Interest: none declared.

REFERENCES

Altschul,S.F. *et al*. (1990) Basic local alignment search tool. *J. Mol. Biol*., **215**, 403–410.

Andreeva,A. *et al*. (2004) SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Res*., **32**, D226–D229.

Cilibrasi,R. and Vitányi,P.M.B. (2005) Clustering by compression. *IEEE Trans. Inform. Theory*, **51**, 1523–1542.

Claus,D. and Fitzgibbon,A. (2004) Reliable fiducial detection in natural scenes. In Pajdla,T. and Matas,J. (eds), *Computer Vision ECCV 2004 Proceedings of the 8th European Conference on Computer Vision*. Prague, pp. 469–480.

Duda,R.O., Hart,P.E. and Stork,D.G. (2001) *Pattern Classification*. John Wiley and Sons, Inc., NY.

Egan,J.P. (1975) *Signal Detection theory and ROC Analysis*. NY.

Gribskov,M. and Robinson,N.L. (1996) Use of receiver operating characteristic (ROC) analysis to evaluate sequence matching. *Comp. Chem*., **20**, 25–33.

Henikoff,S. *et al*. (1999) Blocks+: a non-redundant database of protein alignment blocks derived from multiple compilations. *Bioinformatics*, **15**, 471–479.

Hoel,P.G. (1962) Likelihood Ratio Tests. In: *Introduction to Mathematical Statistics*. John Wiley & Sons, NY, pp. 220–228.

Kocsor,A. *et al*. (2005) Application of compression-based distance measures to protein sequence classification: a methodological study. *Bioinformatics*.

Liao,L. and Noble,W.S. (2003) Combining pairwise sequence similarity and support vector machines for detecting remote protein evolutionary and structural relationships. *J. Comput. Biol*., **10**, 857–868.

MathWorks,T. (2004) *MATLAB*. The MathWorks, Natick, MA.

Pollack,J.D. *et al*. (2005) Taxonomic utility of a phylogenetic analysis of phosphoglycerate kinase proteins of Archaea, Bacteria and Eukaryota: insights by Bayesian analyses. *Mol. Phylogenet. Evol*., **35**, 420–430.

Røgen,P. and Fain,B. (2003) Automatic classification of protein structure by using Gauss integrals. *Proc. Natl Acad. Sci. USA*, **100**, 119–124.

Saigo,H. *et al*. (2004) Protein homology detection using string alignment kernels. *Bioinformatics*, **20**, 1682–1689.

Smith,T.F. and Waterman,M.S. (1981) Identification of common molecular subsequences. *J. Mol. Biol*., **147**, 195–197.

Tatusov,R.L. *et al*. (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, **4**, 41.