BIOINFORMATICS ORIGINAL PAPER

Vol. 00 no. 00 2005, pages 1–6
doi:10.1093/bioinformatics/bti806

Sequence analysis

Application of compression-based distance measures to protein sequence classification: a methodological study

András Kocsor[1,*], Attila Kertész-Farkas[1], László Kaján[2] and Sándor Pongor[2,3]

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

Received on August 30, 2005; revised and accepted on November 27, 2005
Advance Access publication . . .
Associate Editor: Keith A Crandall

ABSTRACT

Motivation: Distance measures built on the notion of text compression have been used for the comparison and classification of entire genomes and mitochondrial genomes. The present study was undertaken in order to explore their utility in the classification of protein sequences.

Results: We constructed compression-based distance measures (CBMs) using the Lempel-Zlv and the PPMZ compression algorithms and compared their performance with that of the Smith–Waterman algorithm and BLAST, using nearest neighbour or support vector machine classification schemes. The datasets included a subset of the SCOP protein structure database to test distant protein similarities, a 3-phosphoglycerate-kinase sequences selected from archaean, bacterial and eukaryotic species as well as low and high-complexity sequence segments of the human proteome, CBMs values show a dependence on the length and the complexity of the sequences compared. In classification tasks CBMs performed especially well on distantly related proteins where the performance of a combined measure, constructed from a CBM and a BLAST score, approached or even slightly exceeded that of the Smith–Waterman algorithm and two hidden Markov model-based algorithms.

Availability: http://www.inf.u-szeged.hu/~kocsor/CBMO5

Contact: kocsor@inf.u-szeged.hu

Supplementary information: http://www.inf.u-szeged.hu/~kocsor/CBMO5

Supplementary Data are not cross-cited in the article except in one place in the abstract. We have inserted Supplementary information para in the abstract based on this. Please confirm whether this article has Supplementary Material.

1 INTRODUCTION

The alignment-based comparison of biological sequences plays a fundamental role in most areas of computational genomics. While exhaustive algorithms (Needleman and Wunsch, 1970; Smith and Waterman, 1981) are computationally expensive for big applications, fast heuristic algorithms such as BLAST (Altschul *et al*., 1990) are among the most extensively used tools in biological computing.

Alignment-free methods of sequence comparison are apparently less popular but are, in fact, quite widely used as pre-selection filters in large-scale sequence processing (Vinga and Almeida, 2003; Vinga *et al*., 2004). In early applications of alignment-free comparisons, sequences were represented in terms of the frequencies of fixed length segments (sequence words, oligones). While very fast when compared with sequence alignment, this approach is complicated by the fact that the word-length and the size of the sequence-alphabet have a critical influence on the sensitivity. A new generation of alignment-free methods is based on a notion borrowed from information theory, the so-called Kolmogorov complexity. Conditional Kolmogorov complexity $K(X|Y)$ is defined as the length of the shortest program computing X on input Y (Li and Vitányi, 1997). The Kolmogorov complexity $K(X)$ of a sequence X is a shorthand notation for $K(X|\lambda)$, where λ is the empty string. The corresponding distance function, (d_1 and d_2), use the relative decrease in complexity or conditional complexity as a measure of sequence similarity (Li *et al*., 2001):

$$d_1(X,Y) = 1 - \frac{K(Y) - K(Y|X)}{K(XY)}, \quad (1)$$

where XY denotes the concatenated X and Y strings. An alternative definition is

$$d_2(X,Y) = \frac{\max\{K(X|Y),\ K(Y|X)\}}{\max\{K(X),\ K(Y)\}}. \quad (2)$$

The distance functions defined above are not metrics, but they satisfy the weak triangle inequality [i.e. $d(X,Y) \leq d(X,Z) + d(Z,Y)$ is fulfilled up to an additive error term] (Li *et al*., 2001, 2004). Kolmogorov complexity is a non-computable notion, and in practical applications it is approximated by the length of the compressed sequence calculated by a compression algorithm like LZW (Lempel and Ziv, 1976) or PPMZ (Bloom, 1998, http://www.cbloom.com/papers/ppmz.zip). The calculation does not require re-solving the sequence with fixed word length segments. Practically speaking, compression-based similarity/distance measures (CBMs) discover a slice of all (subword) similarities which slice is determined by the compressor used (Cilibrasi and Vitányi, 2005).

In the area of biological seqeunce comparison, CBMs were used primarily in the comparison of DNA sequences, including mitochondrial and full genomes (Cilibrasi and Vitányi, 2005; Li *et al*., 2001, 2004; Otu and Sayood, 2003). It has been suggested that CBMs—like other alignment-free measures—might be useful as pre-selection filters for alignment-based querying in large-scale applications (Li *et al*., 2001; Vinga and Almeida, 2003). It was also shown that compression methods can be used to speedup

*To whom correspondence should be addressed.

probabilistic profile matching (Freschi and Bogliolo, 2005). A method for the complexity-based comparison of protein structures is described in Krasnogor and Pelta (2004).

The goal of the present study was to explore the properties of CBMs in classifying protein sequences, using support vector machines (SVM) and nearest neighbour (1NN) classification schemes. We found that their performance—especially on shorter sequences such as protein domains—falls short of alignment-based methods, but a combination of the BLAST score and a CBM can approach, even exceed the performance of the Smith–Waterman algorithm.

2 ALGORITHMS AND DATASETS

2.1 Sequence comparison methods

The formula for calculating compression-based similarity measures using the length values of compressed strings was derived from Equation (2) (Cilibrasi and Vitányi, 2005) as

$$\mathrm{CBM}(X,Y) = \frac{C(XY) - \min\{C(X), C(Y)\}}{\max\{C(X), C(Y)\}}, \qquad (3)$$

where $C(.)$ denotes the length of a compressed string, compressed by a particular compressor C, such as the LZW or the PPMZ algorithm. 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).

Alignment-based sequence comparisons were made using version 2.2.4 of the BLAST program (Altschul *et al*., 1990) with a cutoff score of 25 and with the Smith–Waterman algorithm as implemented in MATLAB (MathWorks, 2004). The BLOSUM 62 matrix (Henikoff *et al*., 1999) was used in each case. The Smith–Waterman comparison data on Dataset I (below) were taken from Liao and Noble (Liao and Noble, 2003).

2.2 Classifier algorithms

1NN classification (Duda *et al*., 2001) is a simple technique whereby a 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 like an alignment-based or compression-based measure.

SVMs are now widely used in the classification of protein sequences (Noble, 2004; Yang, 2004). In most of the early applications, the sequences to be classified were represented in an intermediate between the vector pairs, and a distance calculated from the vectors was used to train the classifiers. Another class of methods uses sequence alignment scores for the training of the classifiers (Liao and Noble, 2003; Vlahovicek *et al*., 2005). We chose one of the latter methods, the so-called pairwise SVM approach (Liao and Noble, 2003), where a sequence X is represented by a feature vector $F_X = f_{x1}, f_{x2}, \ldots, f_{xn}$, n is the total number of proteins in the training set and f_{xi} is a similarity/distance score, such as the BLAST score, the Smith–Waterman score or a CBM between sequence X and the i-th sequence in the training set. For the SVM experiments we used the SVM$^{\mathrm{light}}$ software package (Joachims, 1999). The tests were done using the procedure published by Liao and Noble (2003).

2.3 Datasets

Dataset I. The evaluation of classification performance was tested on a sequence dataset designed to test distant protein similarities (Liao and Noble, 2003). This set consists of 4352 protein domain sequences (whose lengths range from 20 to 994 amino acids) selected from the SCOP database (Andreeva *et al*., 2004). The sequences of this dataset belonging to 54 superfamilies were divided into positive and negative examples, training and test sets in such a way that the test set consisted of members of a protein family that was not represented in the training set, i.e. there is a low degree of sequence similarity and no guaranteed evolutionary relationship between the two sets. The evaluation was carried out by standard ROC (receiver operator curve) analysis (Gribskov and Robinson, 1996) like that described in (Liao and Noble, 2003).

Dataset II. A total of 131 sequences of the essentially ubiquitous glycolytic enzyme, 3-phosphoglycerate kinase (3PGK, 358–505 residues in length)—obtained from 15 archaean, 83 bacterial and 33 eukaryotic species—was used as an example of evolutionarily related sequences (Pollack *et al*., 2005). Since a distance matrix of sequence comparison measure is not useful for seeing how well the sequence groups are separated, for better visualization it is necessary to map the sequences onto a 2D plane while preserving as possible the distance metric relations. Methods such as multidimensional scaling and locally linear embedding (LLE) are suitable for this task (Roweis and Sul, 2000). After the LLE transformation of datasets, the efficiency of the classification was evaluated by separating those sequences belonging to the three taxonomic groups using a multidimensional counterpart of the Fisher separation ratio (Fukunaga, 1990). This ratio is defined as follows:

$$J = \frac{|S_{\mathrm{B}}|}{|S_{\mathrm{w}}|}, \qquad (4)$$

where S_{B} and S_{W} are between- and within-class scatter matrices and $|.|$ denotes matrix determinant. Here the between-class scatter matrix shows the scatter of the class mean vectors around the overall mean vector, while the within-class scatter matrix represents the weighted average scatter of the covariance matrices of the sample vectors belonging to each class.

Dataset III. This dataset consists of artificial sequences designed to study the behaviour of CBMs. The effect of domain deletions/insertions and rearrangements was studied on the Complement C1s subcomponent precursor (Swiss-Prot ID: C1S_HUMAN, AC: P09871), a multidomain protein of 688 residues consisting of a signal peptide (A), two CUB domains (B, B′), an EGF domain (C), two SUSHI domains (D, D′) and a trypsin-like catalytic domain (E) that is post-translationally cleaved from the precursor. The domain architecture of the native protein can be written as ABCB′DD′E, and a hypothetical circular permutant can be written as DD′EABCB′. A series of truncated and rearranged sequences were produced (Table 3) and compared with the native sequence.

Finally, each of the 4352 sequences in Dataset I were random-shuffled using the shuffleseq program of the EMBOSS package (Rice *et al*., 2000) and compared with its original courterpart (4352 pairwise comparisons) using CBMs or the BLAST score.

Dataset IV. A dataset of high and low complexity sequences was produced by taking human proteins from the KOG database (Koonin *et al*., 2004), processing them with the SEG program of J. Wootton (Wootton, 1994), using the parameter values of window length = 45, trigger complexity = 3.25 and extension complexity = 3.55, as recommended by the author. The segments with length in the range of 20–1000 amino acids were chosen for further analysis.

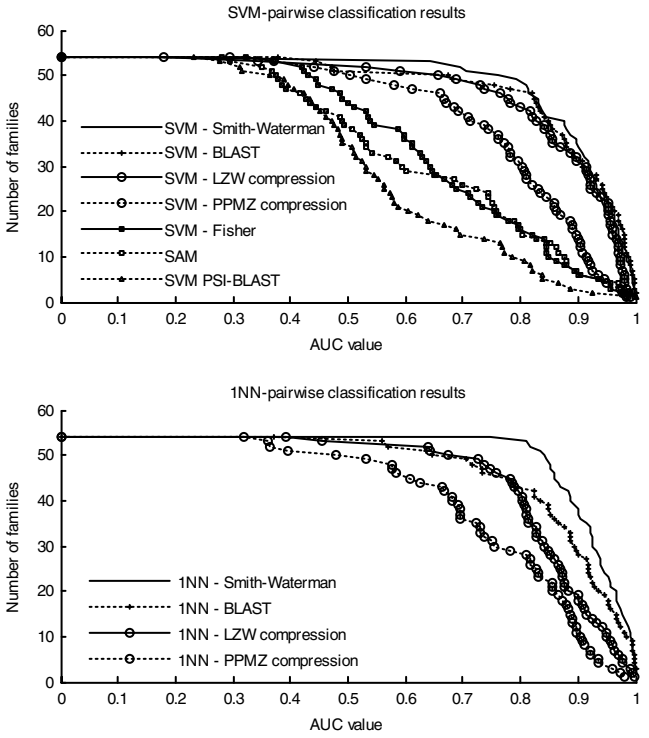

Fig. 1. Relative classification performance of the various similarity/distance measures in SVM (**a**) and Nearest Neighbour (**b**) classifiers as determined on 54 SCOP domain groups (Dataset I). Each graph plots the total number of families for which the integral of the ROC curve (AUC) exceeds a score threshold indicated on the *x*-axis. A higher curve corresponds to more accurate classification performance. The data for SVM-Fisher (Jaakkola *et al*., 1999), SAM profile HMM (Krogh *et al*., 1994) and PSI-BLAST (Altschul *et al*., 1997) were taken from Liao and Noble (2003).

From a total of 163 473 high-complexity and 53 849 low-complexity regions, we randomly selected 8859 and 3772 sequences, respectively, for an analysis whose results are shown in Figure 5.

3 RESULTS

Synthetic classification experiments. From the large number of machine learning algorithms available today we chose two for testing the classification efficiency of the compression-based distance measures. The first is (1NN) classification, which is often used as a reference classifier in comparative studies, while the second method is that of SVM (Cristianini and Shawe-Taylor, 2000). From the latter we chose the 'pairwise SVM' architecture so that our comparisons could be directly compared with published data (Liao and Noble, 2003).

The classification performance of the CBMs was first evaluated on protein domain sequences taken from the SCOP database (Dataset I). This is a carefully selected dataset designed to evaluate distant sequence similarities. The results displayed in Figure 1 reveal that alignment-based methods Smith–Waterman and BLAST perform better than the compression-based methods (LZW and PPMZ) for both the SVM classifier and the (1NN) classifier. The results (Smith–Waterman, BLAST, LZW, PPMZ) are quantitatively compared in Table 1.

Table 1. Classification performances[a] of various similarity/distance measures on protein domain sequences (Dataset I)

Similarity/distance measure	Classification method	
	SVM	Nearest neighbor
Smith–Waterman *P*-value	48.66	50.22
BLAST Score	47.71	47.33
LZW [Equation (3)]	46.97	45.73
PPMZ [Equation (3)]	42.50	41.25
LZW-BLAST [Equation (5)]	49.00	37.18
PPMZ-BLAST [Equation (5)]	47.71	35.23
SVM-Fisher[b]	36.84	n.a.
SAM[b]	35.49	n.a.
PSI-BLAST[a]	31.82	n.a.

[a]Expressed as the integral of the AUC curve in Figure 1. As Dataset I contains 54 families, the integral has a maximum value of 54. This value is attained if all the 54 groups are classified without errors.
[b]SVM-Fisher denotes the method described in (Jaakkola *et al*., 1999), SAM denotes the profile HMM classifier of Krogh's team (Krogh *et al*., 1994) while PSI-BLAST represents the algorithm proposed in (Altschul *et al*., 1997); The values of the latter three were calculated from the data published in (Liao and Noble, 2003).

Table 2. Classification performance of various similarity/distance measures on 3-phosphoglycerate kinase sequences from different taxa (Dataset II)

Measure	Fisher separation ratio
Smith–Waterman *P*-value	1.629
BLAST Score	0.428
LZW [Equation (3)]	0.670
PPMZ [Equation (3)]	0.994
LZW-BLAST [Equation (5)]	0.751
PPMZ-BLAST [Equation (5)]	0.803

The results (Smith–Waterman, BLAST, LZW, PPMZ) obtained on the 3PGK sequences (Dataset II) are shown in Table 2 and Figure 2. In contrast to the members of Dataset II, 3PGK proteins are closely related to each other and fulfill the same biological function so they are a good example of homologous proteins. The sequences in Dataset II were carefully selected so as to provide a wide taxonomic coverage (Pollack *et al*., 2005). The performance of the Smith–Waterman algorithm is the best, but the performance of the compression-based distances exceeds that of the BLAST algorithm.

From these preliminary results it appears that the performance of each CBM falls somewhat short of that of a rigorous sequence-alignment algorithm such as the Smith–Waterman algorithm, but can approach or even match the performance of a heuristic algorithm such as BLAST.

Classification using CBMs combined with BLAST. The second experiment was designed in order to find out whether CBMs can be used in a mixed-feature setting. In particular we were interested in finding out if a combination of the BLAST score and the compression-based measures could approach the performance of

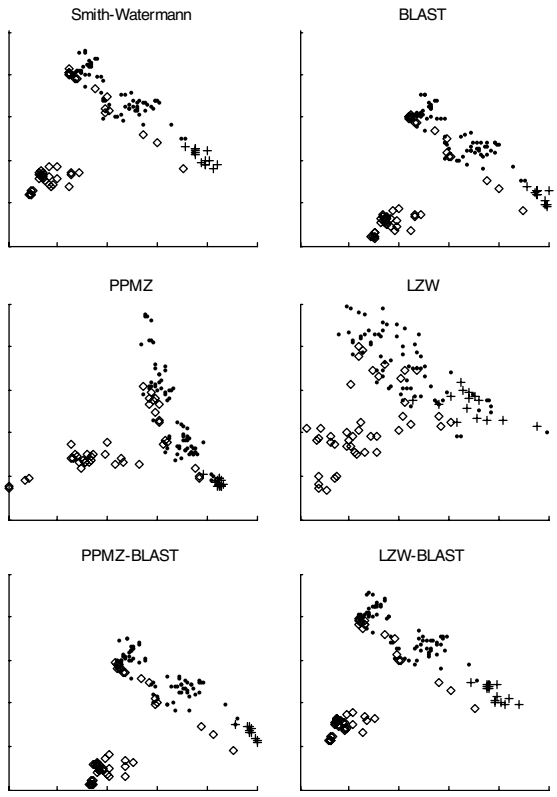

Fig. 2. Separation of Eukaryotic (diamond), bacterial (dot) and archæan (cross) 3-phosphoglycerate kinase sequences (Dataset II). The pairwise similarity/distance values were subjected to LLE.

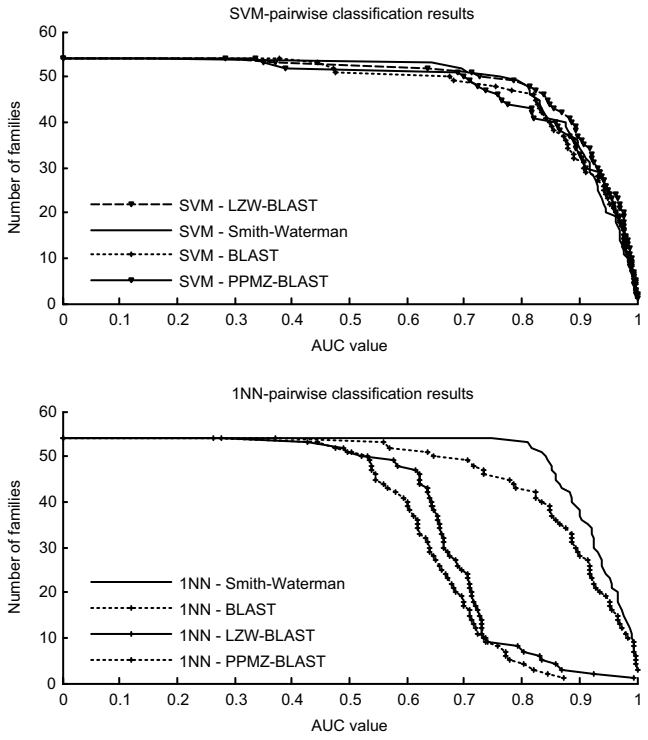

Fig. 3. The relative performance of the mixed distance measures for the 54 SCOP superfaimilies in Dataset I (see the legend of Fig. 1 for explanations).

the Smith–Waterman algorithm. Although numerous combination schemes are available in the literature (cf. Fuzzy theory: Klir, 1997) here the combination of distance measures was carried out using the multiplication rule:

$$F(X,Y) = (1 - \frac{S(X,Y)}{S(X,X)})\mathrm{CBM}(X,Y), \qquad (5)$$

where $S(X,Y)$ is a BLAST score computed between a query X and a subject Y, $S(X,X)$ is the BLAST score of the query compared with itself. The term in parenthesis is used to transform the BLAST score into a normalized distance measure that ranges between zero and one. Equation (5) is a straightforward method for combining CBMs with more specialized methods (in this paper normalized alignment score). The rationale is that an application specific bias may improve the performance of the applied independent compression method. The performance of this combined measure (LZW-BLAST, PPMZ-BLAST) was in fact close to and, in some cases, even slightly superior to that of the Smith–Waterman algorithm, both on protein domains (Fig. 3) and on the 3PGK sequences (Table 2). We consider this result encouraging since Equation (5) does not contain any optimized parameter.

In Figure 1 we also plotted the performance of two methods based on hidden markov models (HMMs). HMMs are currently the most popular tools for detecting remote protein homologies (Durbin *et al*., 1999). The SAM algorithm (Krogh *et al*., 1994) is a profile-based HMM, while the SVM-Fisher method (Jaakkola *et al*, 1999, 2000) couples an iterative HMM training scheme with an SVM-based classification and is reportedly one of the most accurate methods for detecting remote protein homologies. The performance data of these methods were calculated on Dataset I and are taken from Liao and Noble (2003). Here the performance of both the HMM-based methods seems to fall short of the best-performing CBMs. In quantitative terms the AUC integral value of the SAM algorithm is 35.49, that of SVM-Fisher is 36.84, while the CBMs have AUC integral values >42 (Table 1). Even though the results may depend on the database used and should be confirmed by a more extensive statistical analysis, we find it encouraging that the combination of two, low time-complexity measures—CBM and a BLAST score—can compete with HMMs in terms of classification accuracy.

Experiments on rearranged and randomized sequences. Protein evolution often includes rearrangements such as the gain or loss of domains and circular permutations. Then the question arises of whether or not CBMs can detect the differences between the products. The results in Table 3 show that a reshuffling of the domains does not substantially affect CBMs. For example, comparing the C1S precursor, a multidomain protein of 688 amino acids with its reshuffled counterparts in which the domain-order is reversed, gives a PPMZ distance of 0.067 while the C1S compared with itself gives a value of 0.020 (The respective BLAST score values are 1435 and 3789, i.e. there is a substantial difference). Also, circular permutation of a long sequence has a smaller effect on the compression distance than on the BLAST score. This property of CBM may be useful when similarities between rearranged sequences are to be detected.

Sequence length and sequence complexity. It is generally known that short sequence similarities are more difficult to detect than long

Table 3. Effect of sequence rearrangements on the various similarity/distance measures

Comparison of human C1S precursor[a] (domain structure: ABCB′DD′E) with	BLAST Score	%[b]	LZW Score	%	PPMZ Score	%	LZW-BLAST Score	%	PPMZ-BLAST Score	%
ABCB′DD′E (itself)	3789	0	0.646	0	0.02	0	0.000	0	0.000	0
ABCCB′DD′E (duplication of domain C)	3085	18.8	0.657	8.0	0.031	1.2	0.122	15.9	0.005	0.5
ABB′DD′E (deletion of domain C)	2815	26.0	0.657	8.2	0.092	7.6	0.169	22.0	0.023	2.5
ABDD′E (deletion of domains C, B′)	2289	40.0	0.689	32.7	0.251	24.5	0.273	35.5	0.099	10.4
D′E ABCB′D (circular permutation)	2440	36.0	0.648	1.6	0.034	1.5	0.231	30.1	0.013	1.4
E D′D B′CBA (reverse domain order)	1435	62.8	0.657	8.2	0.067	5.0	0.408	53.2	0.045	4.7
2 × ABCCB′DD′E (duplication)	3789	0.0	0.743	74.4	0.019	0.1	0.000	0.0	0.000	0.0
Random shuffled[c] human C1S precursor	42.3	100.0	0.776	100.0	0.961	100.0	0.768	100.0	0.951	100.0

[a]Swiss-Prot AC = P09871, A = Signal peptide (res.1–15) B, B′ = CUB domains (res. 16–130, 175–290, respectively), C = EGF domain (res. 131–172), D, D′ = Sushi domains (res. 292–356, 357–423, respectively), E = Peptidase SI (res. 438–688)
[b]Score of CIS with itself = 0%, score of CIS with random shuffled CIS = 100%
[c]Scores are the average of comparing CIS to 50 random shuffled CIS sequences.

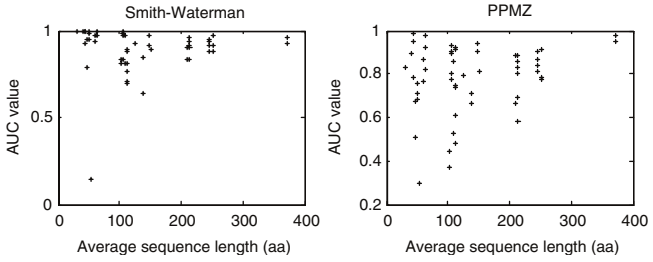

Fig. 4. Dependence of classification performance on sequence length in Database I (see also http://www.inf.u-szeged.hu/~kocsor/CBM05).

ones. This tendency also seems to be valid for the two CBMs analysed here. For example, if we plot the classification performance measure AUC for the 54 SCOP superfamilies as a function of the average sequence length, the low values are predominantly found in the shorter superfamilies (Fig. 4). This tendency is quite similar for the Smith–Waterman algorithm, CBMs and the BLAST algorithm (www.inf.u-szeged.hu/~kocsor/CBM05).

In order to find out whether there are systematic tendencies in the length dependence, we plotted the values of the self-similarities as a function of the sequence length using the 4352 SCOP sequences (Fig. 5). In principle, this value should be close to zero. In practice we found non-zero values that gradually decreased as a function of the sequence length. Furthermore, we compared each sequence with its random shuffled counterpart and plotted the resulting CBM value as a function of the sequence length. These values were expected to be close to one. In practice this was found only in the case of longer sequences, the shorter ones giving values well below one. It is worth noting that these tendencies are quite clear, for both cases as shown by the log–log plots.

The two panels on the right of Figure 5 show the results obtained on low and high complexity segments of human proteins. Low-complexity segments correspond to non-globular regions in proteins characterized by a biased amino acid composition and/or a repetitive amino acid sequence. Such regions are well-known to obscure the detection of biologically important similarities (Wootton, 1994). We had two particular reasons for testing CBMs on low-complexity regions: (1) repetitive character-sequences can be, by definition,

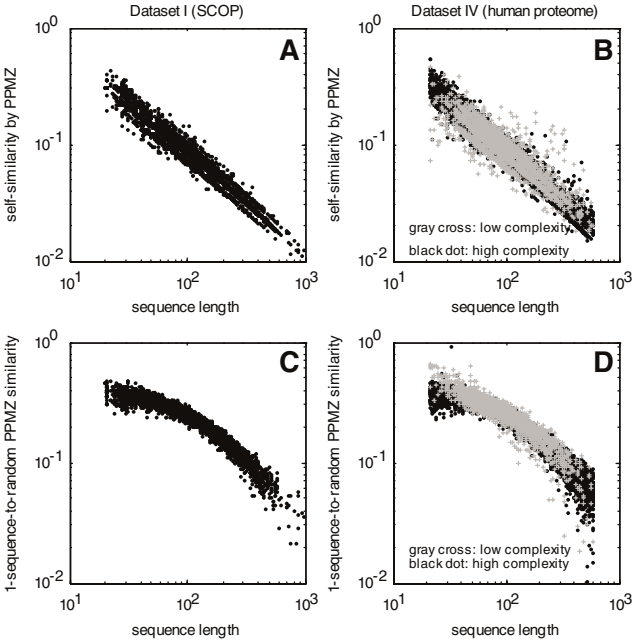

Fig. 5. Dependence of the PPMZ distance as a function of sequence length for protein domains of Dataset I (**A**), and high (black) and low-complexity (grey) segments of human proteins (**B**). The distance of a sequence versus itself is supposed to give a value of 0.00. The actual values display a length dependence shown in panels A and B (log–log plot). A comparison of native sequences with their random shuffled counterparts in Dataset I (B) and high (black) and low-complexity (grey) segments of human proteins (**D**). The PPMZ distance of a sequence from its random-shuffled conterparts is supposed to give a high value, a value close to 1.00. The length dependence of the actual value (i.e. its difference from 1.00) is shown in the lower panel (log–log plot) (see also http://www.inf.u-szeged.hu/~kocsor/CBM05).

Note that the figure caption does not mention part C. Please provide an appropriate citation for part C in the figure legend.

better compressed than the average and (2) our datasets of natural sequences (Datasets I and II) are composed of proteins that are predominantly globular so they are expected to contain few low-complexity regions. Contrary to our expectations, we found that the distributions of CBM values on low and high complexity proteins were not markedly different (Fig. 5B and D).

4 CONCLUSIONS

The original reason for this study was to gain an insight into the behaviour of CBMs in protein classification tasks. We found that a combination of two, low time-complexity measures (BLAST score and CBMs) can approach or even exceed the classification performance of such computationally intensive methods as the Smith–Waterman algorithm or HMM methods. Summarizing we conclude that CBMs may be a useful auxiliary tool for increasing the classification performance of heuristic protein sequence alignment. It is apparent on the other hand, that short sequence similarities are not efficiently detected by CBMs and future work is needed to clarify if this property can be corrected. Nevertheless we feel that, in some applications such as the comparison of full-length proteins CBMs may offer a cost-effective alternative to probabilistic modeling methods such as HMMs.

ACKNOWLEDGEMENTS

The authors would like to thank Profs J. Dennis Pollack for the 3-phosphoglycerate kinase sequences, Jack A.M. Leunissen for help and advice with the Smith–Waterman calculations and John Wooton for his suggestions regarding sequence complexity calculations. A.K. was supported by the János Bolyai fellowship of the Hungarian Academy of Sciences. The Bioinformatics Group of the Szeged Biological Center was partly supported by grants of the Hungarian Ministry of Education (OMFB-01887/2002, OMFB-00299/2002).

Conflict of Interest: none declared.

REFERENCES

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

Altschul,S.F. *et al*. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*., **25**, 3389–3402.

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

Bloom,C. (1998) Solving the Problems of Context Modelling. California Institute of Technology. 1–11.

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

Cristianini,N. and Shawe-Taylor,J. (2000) *An Introduction to Support Vector Machines*. Cambridge University Press, Cambridge.

Duda,R.O. *et al*. (2001) *Pattern Classification*. John Wiley and Sons, Inc., NY.

Durbin,R. *et al*. (1999) *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, Cambridge.

Freschi,V. and Bogliolo,A. (2005) Using sequence compression to speed up probabilistic profile matching. *Bioinformatics*, **21**, 2225–2229.

Fukunaga,K. (1990) *Introduction to Statistical Pattern Recognition*. Academic Press, San Diego, NY, Boston.

Gribskov,M. and Robinson,N.L. (1996) Use of receiver operating characteristic (ROC) analysis to evaluate sequence matching. *Comput. 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.

Jaakkola,T. *et al*. (1999) Using the Fisher kernel method to detect remote protein homologies. *Proc. Int. Conf. Intell. Syst. Mol. Biol*., 149–158.

Jaakkola,T. *et al*. (2000) A discriminative framework for detecting remote protein homologies. *J. Comput. Biol*., **7**, 95–114.

Joachims,T. (1999) Making large-scale SVM learning practical. In Schoelkopf,B., Burges,C. and Smola,A. (eds), *Support Vector Learning*. MIT Press, Boston, MA.

Klir,G. *et al*. (1997) *Fuzzy Set Theory: Foundations and Applications,* Prentice Hall PTR.

Please provide publisher's location for the reference Klir,G. et al. (1997).

Koonin,E.V. *et al*. (2004) A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. *Genome Biol*., **5**, R7.

Krasnogor,N. and Pelta,D.A. (2004) Measuring the similarity of protein structures by means of the universal similarity metric. *Bioinformatics*, **20**, 1015–1021.

Krogh,A. *et al*. (1994) Hidden Markov models for detecting remote protein homologies. *Bioinformatics*, **14**, 846–856.

Lempel,A. and Ziv,J. (1976) On the complexity of finite sequences. *IEEE Trans. Inf. Theory*, **22**, 75–81.

Li,M. *et al*. (2001) An information-based sequence distance and its application to whole mitochondrial genome phylogeny. *Bioinformatics*, **17**, 149–154.

Li,M. *et al*. (2004) The similarity metric. *IEEE Trans. Inf. Theory*, **50**, 3250–3264.

Li,M. and Vitányi,P.M.B. (1997) *An Introduction to Kolmogorov complexity and its Applications*. Springer Verlag, NY.

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.

Needleman,S.B. and 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.

Noble,W.S. Support vector machine applications in computational biology. In Schoelkopf,B., Tsuda,K. and Vert,J.P. (eds), *Kernel Methods in Computational Biology*. MIT Press, Boston, MA, pp. 71–92.

Otu,H.H. and Sayood,K. (2003) A new sequence distance measure for phylogenetic tree construction. *Bioinformatics*, **19**, 2122–2130.

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.

Rice,P. *et al*. (2000) EMBOSS: the European Molecular Biology Open Software Suite. *Trends. Genet*., **16**, 276–277.

Roweis,A. and Sul,L. (2000) Nonlinear dimensionality reduction by locally linear embedding. *Science*, **290**, 2323–2326.

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

Vinga,S. and Almeida,J. (2003) Alignment-free sequence comparison—a review. *Bioinformatics*, **19**, 513–523.

Vinga,S. *et al*. (2004) Comparative evaluation of word composition distances for the recognition of SCOP relationships. *Bioinformatics*, **20**, 206–215.

Vlahovicek,K. *et al*. (2005) The SBASE domain sequence resource, release 12: prediction of protein domain-architecture using support vector machines. *Nucleic Acids Res*., **33**, D223–D225.

Wootton,J.C. (1994) Non-globular domains in protein sequences: automated segmentation using complexity measures. *Comput. Chem*., **18**, 269–285.

Yang,Z.R. (2004) Biological applications of support vector machines. *Brief Bioinform*., **5**, 328–338.