Application of protein lossless compression for local similarity detection

ANDREA HATEGAN
TAMPERE University of Technology
Institute of Signal Processing
P.O. Box 553, FIN-33101, Tampere
Finland

IOAN TABUS
TAMPERE University of Technology
Institute of Signal Processing
P.O. Box 553, FIN-33101, Tampere
Finland

Abstract: As more biological data become available, the need for specific tools to analyze the data is more pressing. The methods for protein comparisons, which can asses if two given proteins are similar or not, are amongst the most important tools for analyzing biological data. These methods are very useful when a new protein is discovered, because finding similar proteins about which something is already known can help in evaluating the function and structure of the new protein. The classical methods for finding similarities at the sequence level involve a substitution matrix and a sequence alignment algorithm for aligning the two sequences. The alignment algorithm can be global - when the sequences are aligned from one end to the other - and local - when only isolated regions of the two proteins are aligned. In this paper, we present a new method to compute the local similarity score using a lossless compression algorithm.

Key–Words: protein lossless compression, local similarity score, substitution matrix, compositional adjustment

1 Introduction

The most reliable way to determine the structure and the function of a protein is by experimental methods such as x-ray crystallography and nuclear magnetic resonance (NMR). But since it is more easier to determine the sequence of a protein than it is to experimentally determine its structure and function, there is a strong motivation for the development of tools that are able to determine the similarity of two proteins at the sequence level. In that case, for a newly discovered protein, the function and the structure can be evaluated by finding similar proteins for which the function and the structure are already known.

Protein sequences accumulated lots of insertions, deletions and substitutions during the evolutional process and for deciding if two proteins are similar or not, one first tries to find an alignment between them. Generally, to produce an alignment between two sequences, a dynamic programming algorithm is used [1] [2][3]. The alignment algorithm can produce a global alignment, that is the two sequences are aligned from one end to the other [1][2] or it can produce a local alignment when only subsequences of the two proteins are aligned [3]. Despite the fact that these algorithms produce the optimal alignment between two sequences, they cannot be used in everyday life for searching protein databases because of the time complexity. To overcome this problem, heuristic algorithms for pairwise sequence comparison [4][5][6][7] and multiple sequence comparison were developed[8].

All of these algorithms make use of a key component, that is a substitution matrix, to produce the alignment of the two sequences and the associated alignment score. Generally, the most likely biological alignment has to have the greatest alignment score and then the scoring matrix has to reflect the evolutionary history, three dimensional structures and other properties that constrain the evolution of the primer structure of proteins. Among the most widely used substitution matrices are the PAM family[9][10] and BLOSUM matrices [11]. It has been showed that in the context of local ungapped alignments, the elements of every such substitution matrix are log odds ratios of the probability of the two amino acids being related and of the probability of the two amino acids being unrelated [12][13].

In [15], a compression algorithm for proteome sequences was presented. This algorithm makes also use of a substitution matrix which is adaptively built for every proteome sequence that is to be compressed. Generally, when a compression scheme succeeds to compress a given sequence by making use of a statistical model, it means that the model was relevant for that sequence. The matrix in this algorithm is collected from the regions where the statistics will de-

crease the description length with respect to the raw model. A measure of the relatedness of two proteomes was defined as the difference between the average codelengths obtained when the first proteome is encoded using only the statistics of its own sequence and the average codelength obtained when it is conditionally encoded knowing the other proteome. The fact that plausible biological phylogenetic trees can be built using this measure of relatedness for the full proteome sequences, made us raise the question if this measure can also be used to compare different pairs of proteins.

The rest of the paper is organized as follows. In the next section, some recent observations concerning the compositional adjustment of substitution matrices are presented [14]. In the third section, some modifications of the compression algorithm from [15] are presented so that the algorithm is suited for comparison of two proteins. In the forth section, the results obtained when comparing different pairs of proteins using compositional adjusted matrices and our method are presented. Finally, some conclusions are drawn.

2 Compositional adjustment of substitution matrices

One of the long standing problems concerning the widely used substitution matrices, PAM and BLOSUM, is that the substitution score is of the form of log-odds ratio of the target frequencies and of the background frequencies derived from accurate alignments of closely related proteins. These matrices are then appropriate for comparison of protein sequences for which the amino acid composition is close to the background frequencies used to construct them. Unfortunately, the standard substitution matrices are also used when comparing protein sequences with very different background frequencies. To overcome this problem, a method for adjusting the implicit target frequencies of the substitution matrix used for comparison is presented in [14].

Given a set of target frequencies q_{ij} that sum to 1, where q_{ij} is the probability of observing any pair of amino acids, and two sets of background frequencies p_i and p'_j, defined as the marginal sums of q_{ij} :

$$\begin{aligned} p_i &= \sum_j q_{ij}; \\ p'_j &= \sum_i q_{ij}, \end{aligned} \tag{1}$$

a substitution matrix defined as

$$s_{ij} = \frac{1}{\lambda} \ln(\frac{q_{ij}}{p_i p'_j}) \tag{2}$$

is called valid in the context of p_i and p'_j, where λ is a scaling factor. Although PAM and BLOSUM matrices are valid in the context of their background frequencies, these matrices are used to compare sequences with different background frequencies P_i and P'_j. If the expected score $\sum_{ij} P_i P'_j s_{ij}$ remains negative, the substitution matrix can be written in the form of log-odds ratios:

$$s_{ij} = \frac{1}{\Lambda} \ln[\frac{Z_{ij}}{P_i P'_j}] \tag{3}$$

Then, in the new background context, s_{ij} remains a log-odds matrix with a new set of target frequencies Z_{ij} and a new scale factor Λ. The problem is that P_i and P'_j are not the marginal probabilities of Z_{ij} and then s_{ij} is not valid in the context of P_i and P'_j.

To adjust any given log-odds matrix to a nonstandard context, they seek for a new set of target frequencies Q_{ij} that is as close as possible to the original set of target frequencies q_{ij} and satisfies the consistency conditions:

$$\begin{aligned} P_i &= \sum_j Q_{ij} \\ P'_j &= \sum_i Q_{ij} \end{aligned} \tag{4}$$

For finding the new target frequencies Q_{ij} the Kullback-Leibler distance is used:

$$D(Q, q) = \sum_{ij} Q_{ij} \ln(\frac{Q_{ij}}{q_{ij}}) \tag{5}$$

Compositional adjusted matrices have shown an improved performance in detecting biological appropriate sequences.

3 A compression algorithm for measuring the local similarity of two proteins

In a recent work [15], we presented a lossless compression algorithm for compression of proteome sequences, i.e. the ensemble of all proteins in a given organism. The algorithm can also be used for conditional coding, i.e. compressing a proteome given another proteome. Using the compressibility results, we defined a measure of "relatedness" of the two proteomes and using this measure we constructed a phylogenetic tree for the proteomes of several organisms.

The fact that biological plausible phylogenetic trees can be built using this measure leads to the conclusion that our compression algorithm manages to capture the biological meaningful features of the data.

In this paper, we investigate if our method of defining "relatedness" at a macro scale can also be used at a micro scale, i.e to compute the "relatedness" of two proteins in different organisms and compare the results obtained with our method to those obtained when using compositional adjusted matrices.

The algorithm introduced in [15] is a two pass algorithm. In the first pass, the proteome sequence which is to be compressed is split into non-overlapping blocks and for each block a regressor block is found in the already seen sequence. If the proteome is conditionally encoded, the regressor may be found in the other proteome. The block and the regressor block have the same length and for each pair of blocks the number of matches is computed, i.e. the number of positions in which the same amino acid is found. A substitution frequency matrix is collected from the pairs of blocks for which the number of matches is greater than a given threshold. The pair of amino acids that have the same rank specifies the row and the column for the cell value that is updated each time. The amino acids are numbered in alphabetical order from one to twenty and the amino acid in the regressor block specifies the row in the matrix and the amino acid in the current block specifies the column. For each conditional distribution obtained from each row of the matrix, a Huffman code was built. In the second step, a given block can be encoded using the Huffman codes designed in the first pass, if the number of matches is greater than the specified threshold, or using an adaptive first order Markov model and the arithmetic encoding, if the number of matches is less than the threshold. In order for the decoder to be able to decode the message received, the Huffman codes have to be sent as a prefix of the message. In [15], we have also presented a version where a fixed substitution matrix is used. Finally, the "relatedness" of two organisms, based on the sequence of their proteomes, was defined as the difference between the average codelength obtained when a proteome is encoded based on the statistics observed only within its sequence and the average codelength obtained when the proteome is conditionally encoded based on the other proteome.

For the experiment in this paper, we used a slightly different version of the algorithm presented in [15]. Because in this experiment we compare only pairs of proteins in different organisms, the substitution matrix is not collected for each pair of proteins because there is not enough statistics at this level and only the cost of transmitting the matrix may be greater than the cost of encoding the whole sequence of amino acids. Then, for each pair of organisms, a substitution frequency matrix is collected at the proteome level and the associated Huffman codes are used to compute the "relatedness" of proteins from that organisms.

For computing the "relatedness" of two proteins, the same idea as in [15] is used. For a given pair of proteins, one is considered to be encoded using only the statistics of its own sequence and conditioned to the other protein and then the relatedness is the difference between the two average codelengths obtained. The protein is also split in non-overlapping blocks of a certain length. In the first case, the regressor block is searched only in the already seen sequence and in the second case, the regressor is searched also in the other protein. The pair of blocks having the number of matches greater than a fixed threshold are encoded conditional on their regressor using the Huffman codes designed at the proteome level, while the blocks with less number of matches than the threshold are encoded in clear using $\log_2(20)$ bits/amino acid. Using this method to encode the amino acids in the blocks with number of matches less than the threshold has an interesting interpretation when computing the "relatedness" of the two proteins.

Let $X = x_1, \ldots, x_{N_x}$ be the protein which is to be encoded and $Y = y_1, \ldots, y_{N_y}$ the conditioning protein. For each block $x^k = x_{(i-1)k+1}, \ldots, x_{(i-1)k+L}$, where L is the length of the block, a regressor block is found $r_p^k = r_1, \ldots, r_L$ where depending on the value of p we have two cases: $r_1, \ldots, r_L = x_t, \ldots, x_{t+L-1}$ if $p = 1$, which means that the regressor is found before the current block in the protein sequence which is to be encoded, or $r_1, \ldots, r_L = y_t, \ldots, y_{t+L-1}$ if $p = 2$, which means that the regressor is found in the other protein. If $p = 1$, then $t \leq (i-2)k + L$ and if $p = 2$, then $t \leq N_y$. The "relatedness" of the two proteins is computed the same way as in [15]:

$$R(X, Y) = \mathcal{L}(X) - \mathcal{L}(X|Y) \tag{6}$$

where $\mathcal{L}(X)$ is the average codelength obtained when the protein X is encoded without knowing the protein Y and $\mathcal{L}(X|Y)$ is the average codelength obtained when the protein X is encoded conditionally on protein Y.

Let $N_b = \lfloor \frac{N_x}{L} \rfloor$ be the number of non-overlapping blocks in the protein X, then (6) becomes:

$$R(X, Y) = \sum_{k=1}^{N_b} [\mathcal{L}(x^k|r_1^k) - \mathcal{L}(x^k|r_p^k)] =$$

$$= \sum_{i=1}^{N_{b1}} [\mathcal{L}(x^i|r_1^i) - \mathcal{L}(x^i|r_p^i)] +$$
$$+ \sum_{j=1}^{N_{b2}} [\mathcal{L}(x^j|r_1^j) - \mathcal{L}(x^j|r_2^j)] =$$
$$= \sum_{j=1}^{N_{b2}} [\mathcal{L}(x^j|r_1^j) - \mathcal{L}(x^j|r_2^j)] \quad (7)$$

where N_{b1} is the number of blocks for which the number of matches is less then the fixed threshold and N_{b2} is the number of blocks for which at least in the conditional case the number of matches is greater than the fixed threshold.

We can further write (7) as:

$$R(X,Y) = \sum_{i=1}^{N_{b21}} [L\log_2(20) - \mathcal{L}(x^i|r_2^i)] +$$
$$+ \sum_{j=1}^{N_{b22}} [\mathcal{L}(x^i|r_1^i) - \mathcal{L}(x^i|r_2^i)] \quad (8)$$

where N_{b21} is the number of blocks for which only in the conditional case the number of matches exceeds the fixed threshold and N_{b22} is the number of blocks for which in both cases the number of matches is greater than the fixed threshold. It turns out that our method to define the relatedness of two proteins is in fact a measure of the local similarities of the two proteins because the regions where the proteins are not similar are discarded.

4 Experimental results

To test the ability of our method in measuring the local similarities of two proteins, we used the same data set as the one in [14]. This data set is formed of three sets of pair of proteins from organisms with very biased AT or GC genomes. The three pairs of organisms considered are: (i)*Clostridium tetani* (AT-rich) and *Mycobacterium tuberculosis* (GC-rich) with contrasting strong biases; (ii) *Bacillus subtilis* and *Lactococcus lactis* both with relatively unbiased genomes; and (iii) *Mycobacterium tuberculosis* and *Streptomyces coelicolor* with strong biases in the same GC direction. For each pair of organisms, two sets of pairs of proteins were considered, one for positive control which contains only orthologous pairs of proteins and one for negative control which contains all possible pairs of proteins in the positive control set excluding the orthologous pair.

The results for the orthologous pairs of protein for the three pairs of organisms considered are listed in Table 1. In this table, the values in the columns denoted by "Relatedness" are computed with our method, while the others are taken from [14]. In the third column, for each pair of organisms, the mean of the local alignment bit score obtained for each pair of orthologous proteins when using a scaled version of the BLOSUM 62 substitution matrix is listed. In the fifth and sixth columns, the median change in bit score with respect to BLOSUM 62 when using composition-adjusted matrices is listed. For the composition-adjusted matrices, the background frequencies were adjusted for proteome frequencies (the column denoted by "Organism") and for the frequencies of the two sequences considered (the column denoted by "Sequence"). In the last two columns, the median changes in bit score when using our method to compute the local similarity score with respect to the values obtained when using the scaled version of the BLOSUM 62 substitution matrix and composition adjusted matrices with background frequencies adjusted for the frequencies of the two sequences compared are presented . In Table 3, the bit scores and the median changes in bit score for all the orthologous pairs for *Bacillus subtilis* and *Lactococcus lactis* are presented. The values in the last column in this table are computed with respect to the best values obtained in [14] (the values in the column "Sequences"). The values in the columns denoted by "Relatedness" are computed using our method.

With our method to compute the local similarity of two proteins, we obtained better results for the last two pairs of organisms with respect to values obtained when the scaled version of the BLOSUM 62 substitution matrix is used and when composition adjusted matrices with background frequencies adjusted for the two sequences compared are used.

For the set of unrelated sequences, we did not have access to the values computed for all pairs of sequences and in this way we cannot report the median change in bit score with respect to values obtained when using the BLOSUM 62 substitution matrix or values obtained when using composition-adjusted matrices. The only value we can report is the mean bit score for the three pair of organisms. The values are presented in Table 2. The third column contains the mean bit score obtained when comparing the unrelated protein pairs using the scaled version of the BLOSUM 62 substitution matrix and the last column presents the mean bit score when comparing the unrelated pairs of proteins using our method. Using our method to compute the local similarity, a decrease for the mean bit score was observed for the last two pairs of organisms.

5 Discussions and conclusions

We have tested if our method of defining the relatedness at a macro scale, by comparing different proteomes, can also be used at a micro scale to compare different pairs of proteins. Using a slightly different algorithm introduced in [15], the relatedness of two proteins can be seen as a measure of the local similarity of the two proteins. To test the ability of our algorithm to measure the local similarity of two proteins, we used the same data set as in [14] where compositional adjusted matrices were used to score all the protein pairs in the data set. Using our method to measure the local similarity of two proteins, we obtained an increased value for the median change in bit score for two of the data sets for the orthologous pairs and for the same two data sets, a decreased value of the mean bit score for the unrelated pairs, which means that our method manages to capture the biological meaningful features of the sequences involved.

References:

[1] S. B. Needleman and C. D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, **48**, pp. 443-453, 1970.

[2] O. Gotoh, An improved algorithm for matching biological sequences, *J. Mol. Biol.*, **162**, pp. 705-708, 1982.

[3] T. Smith and M. Waterman, Identification of Common Molecular Sequences, *J. Mol. Biol.*, **147**, pp. 195-197, 1981.

[4] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, Basic local alignment search tool, *J. Mol. Biol.*, **215**, pp. 403-410, 1990.

[5] S. F. Altschul et al, Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Res.*, vol.25, no.17, pp. 3389-402, 1997.

[6] D. Lipman and W. Pearson, Rapid and sensitive protein similarity searches. *Science*, **227**, pp. 1435-1441, 1985.

[7] D. Lipman and W. Pearson, Improved tools for biological sequence comparison, *Proc. Natl. Academy Science*, **85**, pp. 2444-2448, 1988.

[8] J.D. Thompson, D.G. Higgins and T.J. Gibson, CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice, *Nucleic Acids Res.*, **22**, pp.4673-4680, 1994.

[9] M.O. Dayhoff, R.M. Schwartz and B. Orcutt, A model of evolutionary change in proteins, *Atlas of Protein Sequence and Structure*, vol. 5, pp. 345-352, 1978.

[10] R.M. Schwartz and M.O. Dayhoff, Matrices for detecting distant relationships, *Atlas of Protein Sequence and Structure*, vol. 5, suppl. 3, pp. 353-358, 1978.

[11] S. Henikoff and J. G. Henikoff, Amino acid substitution matrices from protein block, *Proceedings of the National Academy of Science USA*, vol. 89, no. 22, pp. 10915-10919, 1992.

[12] S.F. Altschul, Amino acid substitution matrices from an information theoretic perspective, *J. Mol. Biol.*, **219**, pp. 555-665, 1991.

[13] S. Karlin and S. F. Altschul, Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes, *Proc. Natl. Acad. Sci. USA*, **87**, pp.2264-2268, 1990.

[14] Yi-Kuo Yu, John C. Wootton and Stephen F. Altschul, The compositional adjustment of amino acid subtitution matrices, *PNAS*, vol. 100, no. 26, pp. 15688-15693, 2003.

[15] A. Hategan and I. Tabus, Protein is compressible, *Proceedings of the 6th Nordic Signal Processing Symposium, NORSIG 2004*, pp. 192-195, 2004.

Sequence pairs	Organisms compared	No. of sequences	Mean BLOSUM 62 bit score	Median change in bit score with respect to BLOSUM 62			Median change in bit score with respect to Sequence
				Organism [1]	Sequence [1]	Relatedness [2]	Relatedness [2]
Related	*C.tetani* and *M.tuberculosis*	40	68.3	+1.6	**+2.3**	+0.6	-3.5
	B.subtilis and *L.lactis*	37	59.8	+1.1	+2.1	**+10.9**	+7.5
	M.tuberculosis and *S. coelicolor*	34	58.6	+1.4	+2.7	**+4.6**	+1.79

Table 1: The relatedness computed for the orthologous pairs in the three sets. [1] Values taken from [14]. [2] Values computed with our method.

Sequence pair	Organism compared	No. of sequences	Mean bit score	
			BLOSUM 62 [1]	Relatedness [2]
Unrelated	*C. tetani* and *M. tuberculosis*	1,560	**16.7**	17.05
	B.subtilis and *L. lactis*	1,332	15.7	**12.26**
	M. tuberculosis and *S. coelicolor*	1,122	16.4	**14.33**

Table 2: The mean bit score for the unrelated pairs of proteins in the three sets. [1] Values taken from [14]. [2] Values computed with our method.