A pairwise alignment algorithm which favours clusters of blocks.

Elodie Nédélec [*] Thomas Moncion [†‡] Elisabeth Gassiat [*] Bruno Bossard [†§]
Guillemette Duchateau-Nguyen [¶] Alain Denise [†¶] Michel Termier [¶‖]

May 25, 2004

Abstract

Pairwise sequence alignments aim to decide whether two sequences are related or not, and, if so, to exhibit their related domains. Recent works have pointed out that a significant amount of true homologous sequences are missed when using classical comparison algorithms. This is the case when two homologous sequences share several little blocks of homology, too small to lead to a significant score. On the other hand, classical alignment algorithms, when detecting homologies, may fail to recognise all the significant biological signals. The aim of the paper is to give a solution to these two problems. We propose a new scoring method which tends to increase the score of an alignment when "blocks" are detected. This so-called "Block-Scoring" algorithm, which makes use of dynamic programming, is worth being used as a complementary tool to classical exact alignments methods. We validate our approach by applying it on a large set of biological data. Finally, we give a limit theorem for the score statistics of the algorithm.

1 Introduction.

Since the very beginning of Bioinformatics, many efforts have been done to create or to improve tools of sequence comparisons. A particularly useful method for comparing sequences is pairwise alignment. Two sequences may be decided to be homologuous when a high level of similarity is found in their alignment. Obviously, in such cases, there is a strong presumption that they share, at least partly, similar functions. The past works by Needelman and Wunsch [17], Sellers [20] and Smith & Waterman [23] have given the first milestones devoted to this question. However, the literature about alignment is still growing and a complete review is rather impossible. One may find introduction and relevant bibliography in Waterman [27], Durbin *et al.* [10] or Clote and Backhofen [6] for example. Such an effort reveals that many problems remain open about the use of alignment algorithms to answer biological questions. Among the questions in progress, one may quote: the algorithmic complexity, which is of crucial importance at a genomic scale; the choice of the tuning parameters; the underlying

[*]Laboratoire de Mathématiques, Equipe de Probabilités, Statistique et Modélisation, UMR CNRS 8628, Université Paris-Sud, Orsay.

[†]LRI, Equipe Bioinformatique, UMR CNRS 8621, Université Paris-Sud, Orsay.

[‡]LaMI, UMR CNRS 8042, Université d'Evry.

[§]LIMSI, UPR CNRS 3251, Université Paris-Sud, Orsay.

[¶]Bioinformatique des Génomes, Institut de Génétique et Microbiologie (IGM), UMR CNRS 8623, Bât.400, Université Paris-Sud, 91405 Orsay Cedex, France.

[‖]To whom correspondence should be addressed. Email: termier@igmors.u-psud.fr

model and its meaning; the interpretation of results. In all cases it appears that the alignment method must fit to the biological question wich may be, for example, as different as phylogenetic reconstructions or cellular function prediction may be: in phylogeny one has to evaluate the evolution distance between species, as for cellular function prediction one needs to find highly conserved domains.

In some cases, it happens that the alignment score is neither high enough nor low enough to decide wether the compared sequences are related or not. This characterizes the so-called *twilight zone* (see Rost (1999) [19]). This problem is due to multiple reasons. For example, usual parameters (gap penalties and substitution matrices) may not be well fitted for some particular alignments. Thus Blake and Cohen [3] propose to adjust the parameters used in the scoring of the classical algorithms with regard to the evolutionary distance. On the other hand, alignments in the twilight zone may be due also to other parameters. An important one is heterogeneity of the mutation process along the sequences. Indeed, when one observes alignments of related -even distant- sequences, one can see that mutations are not identically distributed. Such an heterogeneity may be attributed to multiple reasons: insertion of translocated sequences (issued from the same genome or from other genomes), differential diverging caused by selection pressure. On biological sequences, it appears that some regions are strongly conserved, such as islands of stability. These conserved "blocks" are likely involved in the active functions of the considered sequence.

Such "blocks" are not taken into account by the classical alignment methods: the weight of relatively short conserved regions may be overwhelmed by the one of numerous unitary identities. Indeed, the current alignment methods are based on the assumptions that mutation events occur homogeneously along adjacent residues. This hypothesis, which greatly simplifies the analytical approach, is also reflecting all the paradigms of the molecular biology at the end of the 60ties. Nevertheless, whatever the alterations affecting the history of the genome, the observation shows that the succession order of amino-acids in proteins is the very basis of their biochemical properties, as, in DNA sequences, the succession order of nucleotides is the basis of the genetic message. Evolution events are thus obviously depending on this order.

The aim of the present work is to propose a solution to this particular problem. We introduce a new alignment algorithm that enhances conserved blocks above the high noise level. Various ideas have been recently proposed to improve block detection. One of them was to build a method taking into account ab initio some "block information" in the alignment. One way to do this is to limit alignments to ungapped blocks with no mutations, called "block-motifs", of the sequence. Liu et al. (1999) ([13] and references therein) developed multiple alignment methods based on this idea. Lam et al. (2003) [14] have proposed an algorithm dividing the whole alignment into segments where residues are independent and segments of pair HMM. One may also use maximum likelihood methods based on evolutionary models such as the TKF model [26]. A quite similar question is the one of "mosaic alignments" where two related sequences have been separated by a long third sequence unrelated to the previous ones. The alignment score is then weaken, inducing false phylogenetic interpretations. Arslan et al. (2001) [2] answered this question with an automatic normalization of alignments.

Our approach consists in the maximization of an alignment score, as done by Smith & Waterman, but with a new scoring function. This new scoring function gives higher weight to what will be called "blocks" than to the same disseminated matches. The optimization may be done via dynamic programing and, in most cases, it is a finite state algorithm, so that the method may be seen as a pair HMM where the increase of the number of states allows to take

the blocks into account.

In section 2, we present our new scoring method. Section 3 describes the dynamic programming algorithm to compute the alignments and the maximum score. Section 4 is devoted to numerical experiments. We show the advantages of our new method: on biological sequences, it allows to detect missed homologies; on random sequences, it does not induce false homologies. In section 5, we state a limit theorem for the score statistics which constitutes the first step in the derivation of asymptotic p-values.

2 New scoring model and block-alignment.

Let $\mathcal{X}$ denote the alphabet of sequences. For any pair of letters $\{a, b\}$ in $\mathcal{X} \times \mathcal{X}$, we denote $\binom{a}{b}$ their alignment, and $s(a, b)$ the score of this alignment, *i.e* the coefficient associated with a and b in the substitution matrix.

Now let us define the notions of *block-match* and *block-mismatch*, which are crucial for our purpose. For any letter a, let $T(a)$ be a real number, called the *block-threshold* of a. Block-thresholds must obey to the following property: for any letters a and b, $s(a, b) \geq T(a)$ if and only if $s(a, b) \geq T(b)$. We say that

- $\binom{a}{b}$ is a *block-match* if $s(a, b) \geq T(a)$;
- $\binom{a}{b}$ is a *block-mismatch* if $s(a, b) < T(a)$;
- as usually, $\binom{a}{b}$ is a *gap* if $a =$"-" or $b =$"-".

A *block* is an alignment which only contains block-matches. A *block-score function* is a function β which associates a positive real number to any block, and which is increasing in the following sense: for any block B, for any block-match $\binom{a}{b}$,

$$\beta(B) \leq \beta(B.\binom{a}{b}) \qquad \text{and} \qquad \beta(B) \leq \beta(\binom{a}{b}.B).$$

Similarly, a *block-mismatch-score function* is a function μ which associates a real number to each sequence which only contains block-mismatches.

Finally, a *gap-score function* is a function γ which associates a negative real to each sequence which only contains gaps and which is also *decreasing* in the following sense:

$$\gamma(G) \geq \gamma(G.\binom{a}{b}) \qquad \text{and} \qquad \gamma(G) \geq \gamma(\binom{a}{b}.G)$$

for any sequence G which only contains gaps and for any gap $\binom{a}{b}$.

Now any alignment A can be decomposed as follows:

$$A = A_0.A_1.A_2. \ \ldots \ .A_{q-1}.A_q$$

where each of the A_i's is either a block, or a sequence of block-mismatches, or a sequence of gaps, and where two consecutive A_i's are not of the same kind. Such a decomposition is unique.

We define the score of the alignment A as follows:

$$f(A) = \sum_{i=1}^{q} f(A_i) \quad \text{where} \quad f(A_i) \begin{cases} \beta(A_i) & \text{if } A_i \text{ is a block;} \\ \mu(A_i) & \text{if } A_i \text{ is a sequence of block-mismatches;} \\ \gamma(A_i) & \text{if } A_i \text{ is a sequence of gaps.} \end{cases}$$

In the classical scoring methods, the following property holds: if $\binom{a}{b}$ is not a gap, then, for any alignment A, $f(A.\binom{a}{b}) = f(\binom{a}{b}.A) = f(A) + s(a,b)$. In other words, the contribution of any given block-match or block-mismatch of an alignment to the score is the same whatever the rest of the alignment is. In other words, pointwise additivity of the classical scoring methods do not take the structure of the alignment into account.

The aim of the scoring model that we propose is to give high scores to long blocks. This allows to detect clusters of blocks more efficiently than classical methods. For this purpose, we consider scoring functions where the length and the composition of a block strongly participates in its score. Hence, we define

$$\beta(B) = \sum_{i=1}^{m} g(\begin{array}{c} v_1 v_2 \dots v_i \\ w_1 w_2 \dots w_i \end{array}; i) \qquad \text{for any block } B \begin{pmatrix} v_1 & v_2 & \dots & v_m \\ w_1 & w_2 & \dots & w_m \end{pmatrix}$$

where $g(\cdot; i)$ denotes a positive real function on $(\mathcal{X} \times \mathcal{X})^i$, $i \in \mathbb{N}^*$. The idea here is to choose a function g which is strictly increasing in its second variable (length of the current block). On the other hand, the gap-score function and block-mismatch-score functions are classical: we take $\gamma(G) = -\gamma_o - (|G| - 1) \times \gamma_e$, where γ_o and γ_e denote respectively the gap-opening penalty and the gap-extension penalty and $|G|$ denotes the length of the sequence of gaps G. Regarding block-mismatches, we define $\mu\binom{a}{b} s(a,b)$ and $\mu(M.\binom{a}{b}) = \mu(M) + s(a,b)$ for any sequence of block-mismatches M and any block-mismatch $\binom{a}{b}$.

Example 1 *Take*

$$g\begin{pmatrix} v_1 v_2 \dots v_i \\ w_1 w_2 \dots w_i \end{pmatrix}; i\Big) = \sum_{j=1}^{i} s(v_j, w_j)$$

for any block. This leads to

$$\beta\begin{pmatrix} v_1 v_2 \dots v_m \\ w_1 w_2 \dots w_m \end{pmatrix} = \sum_{j=1}^{m} (m - j + 1) s(v_j, w_j)\,,$$

Hence, taking for example $s(a,a) = 1$ for any symbol a gives $\beta(B) = \frac{m(m+1)}{2}$ for any block B of length m.

Example 2 *A slight modification of the general block scoring method consists in bounding the influence of the length of the block. Given a positive integer K, define for all $i \geq K$:*

$$g\begin{pmatrix} v_1 v_2 \dots v_i \\ w_1 w_2 \dots w_i \end{pmatrix}; i\Big)\, g\begin{pmatrix} v_{i-K+1} v_{i-K+2} \dots v_i \\ w_{i-K+1} w_{i-K+2} \dots w_i \end{pmatrix}; K\Big)\,.$$

When applied in particular to Example 1, this leads to

$$g\begin{pmatrix} v_1 v_2 \dots v_i \\ w_1 w_2 \dots w_i \end{pmatrix}; i\Big) \begin{cases} \sum_{j=1}^{i} s(v_j, w_j) & \text{if } i \leq K \\ \sum_{j=i-K+1}^{i} s(v_j, w_j) & \text{otherwise.} \end{cases}$$

Remark that setting $K = 1$ gives back the usual scoring scheme.

As usually, given two sequences $v^n = v_1 v_2 \cdots v_n$ and $w^m = w_1 w_2 \cdots w_m$, the best global alignment A is the one that maximizes the score $f(A)$ over all possible alignments, and the best local alignment is the one that maximizes f over all possible alignment of subsequences of v^n and w^m. Let us denote by $S(v^n, w^m)$ the maximum score. We will call block-scoring our new method, and abbreviate it to BS, leading to global BS alignments, local BS alignments, BS (global or local) alignment scores.

3 Algorithm.

Like the usual ones, our scoring model applies to local as well as global alignments, though we shall apply it to local alignments. The algorithm allows to compute the maximum BS score and to retrieve the best BS alignment. In order to take into account the length of the current block in the computation of the score, we introduce a matrix H which counts, for each pair (v_i, w_j) of letters of two words v and w to be aligned, the length of the maximal block ending with $\binom{v_i}{w_j}$. This matrix is defined as follows:

$$H_{i,j} \begin{cases} H_{i-1,j-1} + 1 & \text{if } i-1 \geq 1, j-1 \geq 1 \text{ and } (v_i, w_j) \text{ is a block-match} \\ 1 & \text{if } (i = 1 \text{ or } j = 1) \text{ and } (v_i, w_j) \text{ is a block-match} \\ 0 & \text{otherwise.} \end{cases}$$

To compute recursively the local maximum score $S_{i,j}$ over all local alignments ending with $\binom{v_i}{w_j}$, we introduce $b_{i,j}$, the length of the block at the end of the local alignment leading to $S_{i,j}$. Now the recurrence for computing the local scores in the matrix of alignment is the following:

$$S_{i,j} = Max \begin{cases} S_{i-1,j-1} + s(v_i, w_j) & \text{if } H_{i,j} = 0 \\ S_{i-1,j} - \delta & \\ S_{i,j-1} - \delta & \\ 0 & \text{(only for local aligment)} \\ S_{i-1,j-1} + g\left(\begin{matrix} v_{i-b_{i-1,j-1}} \cdots v_i \\ w_{j-b_{i-1,j-1}} \cdots w_j \end{matrix} ; b_{i-1,j-1} + 1 \right) & \text{if } H_{i,j} \geq 1 \\ S(i-h, j-(h+1)) - \delta + \beta\binom{v_{i-h+1} \cdots v_i}{w_{j-h+1} \cdots v_j} & \forall\ h \in [b_{i,j}, H_{i,j}] \\ S(i-(h+1), j-h) - \delta + \beta\binom{v_{i-h+1} \cdots v_i}{w_{j-h+1} \cdots v_j} & \forall\ h \in [b_{i,j}, H_{i,j}] \end{cases}$$

where

$$\delta \begin{cases} \gamma_o \text{ in case of gap-opening} \\ \gamma_e \text{ in case of gap-extension} \end{cases}$$

The value $b_{i,j}$ denotes the length of the current block. To update it, set $b_{i,j} = 0$ if $S_{i,j}$ is obtained via the first four lines, $b_{i,j} = b_{i-1,j-1} + 1$ if it is obtained via the fifth line, and $b_{i,j} = h$ if it is obtained via the sixth or seventh line for this particular h.
Computation of the global score is performed by removing the zero in fourth line. The last three lines of the recurrence formula are required because one has to test, when in a block, if the block has to be continued, or if it would be better to insert a gap before its beginning, in order to perform a longer block. This is illustrated in the following example.

	A	C	G	T
A	4	0	3	0
C	0	4	0	3
G	3	0	4	0
T	0	3	0	4

(a) Substitution matrix.

	A	C	T	G	T
A	1	0	0	1	0
C	0	2	1	0	2
G	1	0	0	2	0
T	0	2	0	0	3

(b) Matrix H.

	A	C	T	G	T
A	**4**	0	0	3	0
C	0	**12**	**8**	4	9
G	3	**8**	8	**12**	8
T	0	6	**12**	8	

(c) Penultimate step of the algorithm.

	A	C	T	G	T
A	4	0	0	3	0
C	0	12	**8**	4	9
G	3	8	8	**12**	8
T	0	6	12	8	**21**

(d) Optimal local alignment.

Table 1: An example of block-scoring alignment.

Example 3 *Let* $v =$ `ACTGT` *and* $w =$ `ACGT` *two DNA words to be compared by mean of a local alignment. The substitution matrix, a kind of transition-transversion one, is given in Table 1(a) below. The* $\delta = -4$ *as gap-penalty. Table 1(b) presents the matrix* H*. For any nucleotide* x*, we fix the threshold* $T(x)$ *to* 3 *(see Section 2), so transitions are authorised in blocks. In this toy-example, the key step of the algorithm is the ultimate one, which is illustrated in Table 1, (c) and (d). Table 1(c) shows that, up to that step, there are two optimal local alignments:* $\begin{matrix}\texttt{ACTG}\\ \texttt{AC-G}\end{matrix}$ *and* $\begin{matrix}\texttt{AC-T}\\ \texttt{ACGT}\end{matrix}$ *which both contain the block* $\begin{matrix}\texttt{AC}\\ \texttt{AC}\end{matrix}$ *of length two, one gap and one block of length one, thus they both have score* $(4) + (4 + 4) + (-4) + (4) = 12$*. They are shown in bold font in the alignment matrix of Table 1(c). At first sight, the optimal solution in the final step would consist in making longer the first alignment by matching the two last* `T`*'s; this would give score* $(4)+(4+4)+(-4)+(4)+(4+4) = 20$*. But this is wrong, as shown in Table 1(d). The right solution uses the sixth line of the recurrence; it leads to break the previous alignment by moving the gap one position before and thus creating the new block* $\begin{matrix}\texttt{TGT}\\ \texttt{CGT}\end{matrix}$ *which has score* $(3)+(3+4)+(3+4+4) = 21$*, and constitues the best local alignment. This illustrates the capability of the algorithm to prioritize long blocks in alignments.*

The algorithm runs in $O(|v|\times|w|)$ memory space, which can be reduced to $O(Min(|v|, |w|)\times Max(H_{i,j}))$ if only the score has to be computed. The time complexity is $O(|v| \times |w| \times Max(H_{i,j}))$, but this is a very crude bound.

Pair HMMs. As for the Smith & Waterman alignment algorithm, BS alignment algorithm can be viewed as the Viterbi algorithm of a pair HMM model as described for example in Durbin et al. [10]. Indeed, any "bounded" BS scoring function as in example 2 can be represented by a finite state automaton. The associated HMM model has hidden states "Begin", "End", "Insertion", "Deletion", "Match(1)", ..., "Match(K)", where the state "Match(k)" communicates only with "Match($k + 1$)", "End", "Insertion", "Deletion", for $k < K$, and with itself for $k = K$.

Family	Name	Number of seq.
COG2813	16S RNA G1207 methylase RsmC	22
COG1187	16S rRNA uridine-516 pseudouridylate synthase	64
COG1514	2'-5' RNA ligase	21
COG0621	2-methylthioadenine synthetase	62
COG1670	Acetyltransferases	99
COG0013	Alanyl-tRNA synthetase	47

Table 2: The six COG families on which tests were processed.

In [14], a different use of pair HMMs is proposed to align sequences with conserved blocks and non conserved segments. The method uses a mixture of the classical pair HMM (as used for SW) and a pair HMM where all alignments have a weigth depending only of the numbers of letters. In other words, the underlying probability model is, for each possible alignment, a segmentation of the alignment where the SW pair HMM alternates with a model with independent similarity.

4 Biological validation.

4.1 Criteria for evaluating Block-scoring

Block-scoring was evaluated in its local alignment version. It was compared to the Smith & Waterman algorithm, for the following reasons: Both are *exact* algorithms, *i.e.* they provide the best local alignment according to their respective scoring schemes ; and Smith & Waterman algorithm is the widely acknowledged standard for exact local alignment. In the following, we write BS for "Block-scoring" and SW for "Smith & Waterman". In BS alignments, the symbol "!" denotes a block-match which is not an identity.

Our validation tests took the following criteria into account:

1. *Detection of homologies.* Naturally, the first criterion consisted in verifying if BS can detect biologically relevant information in cases where SW cannot. For this purpose, we processed and compared alignments in large sets of homologous sequences. This is detailed in subsection 4.2

2. *Searching for false positives.* Since BS tends to give longer alignments and higher scores than SW, we had to test if BS did not result in too many false positives, with regard to SW. This was done by comparing Z-scores of SW and BS alignments on both biological and random sequences, and also by measuring the lengthening of alignments by BS on *ad hoc* random sequences. This is developed in subsection 4.3

4.2 Detection of homologies

We compared the results of SW and BS on a large set of homologous while distant sequences, on an evolutionary point of view. We took sequences from the COG database (Cluster of Orthologous Groups [24]), which is dedicated to families of orthologous proteins and gives their phylogenetic relationships. We focused on the six families given in Table 2. We computed BS and SW alignments for each couple of sequences within each of these families, using several

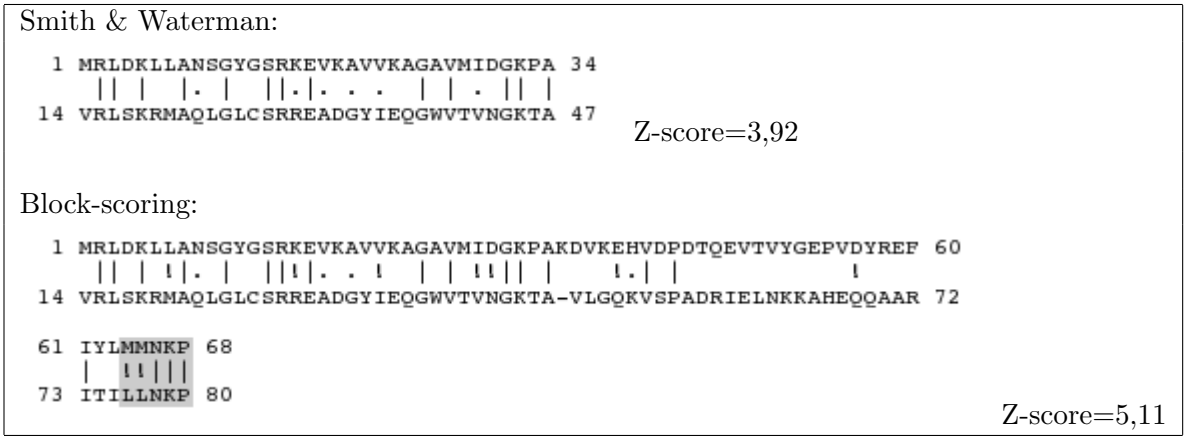

Figure 1: SW and BS best local alignments of sequences BS_ytzG and NMA1016 of the COG1187 family. (PAM 250, γ_o=15, γ_e=5, K=20).

combinations of substitution matrices and gap penalties. When significant differences were found, we tested whether the alignments given by BS where relevant by comparing them with known motifs and profiles in the BLOCKS [11] and PROSITE [12] databases, or with a multiple alignment of the family of sequences. When possible, confirmation was searched in the literature.

Several combinations of values for gap-opening penalty γ_o and gap-extension penalty γ_e were tested. We observed that BS requires high values of these parameters in order to give relevant results. This was expected since BS gives higher scores than SW as soon that there are blocks in an alignment. In the following, we set $\gamma_o = 15$ and $\gamma_e = 5$. PAM 250 was chosen as substitution matrix and we set $K = 20$. These values seem to be rather optimal for BS in our experiments. In the cases where BS results were better than SW ones, SW alignments were recomputed using smaller values of γ_o and γ_e, in order to improve them.

The results are as follows. Generally, the alignments given by both algorithms lie in the same region of the sequences. As expected, BS alignments are longer than SW ones. Two cases have to be considered.

In the very most frequent case (about 90% of the alignments), the SW alignment is exactly included in the BS one, but the latter goes further. This case is illustrated in Figures 1 and 2 by alignments of proteins from the family COG1187, *16S rRNA uridine-516 pseudouridylate synthase* proteins, whose function is catalysation of pseudouridine synthesis in position number 516 of 16S RNA during the assembly of the 30S ribosomal subunit. These proteins contain three main domains [22]: a N-terminal one which allows the protein to bind to the rRNA, a central one which contains the catalytic site, and a C-terminal one that corresponds to a *ferredoxin-like* folding. In the 2/3 of the considered alignments, both algorithms align the N terminal domain only. However, in several cases, BS adds significant information, as illustrated in Figure 1: BS aligns a five amino acids block further which is precisely the core of the binding-site of the protein [22]. In the major part of the remaining alignments, the aligned region is the central and C terminal one, but in some cases SW aligns only one domain whereas BS align both of them. Figure 2 presents a typical illustration of this situation: both SW and BS align well the catalytic site (surrounded in the figure), but only BS aligns the C-terminal motif, after a rather long sequence of gaps. Once more, the additional information

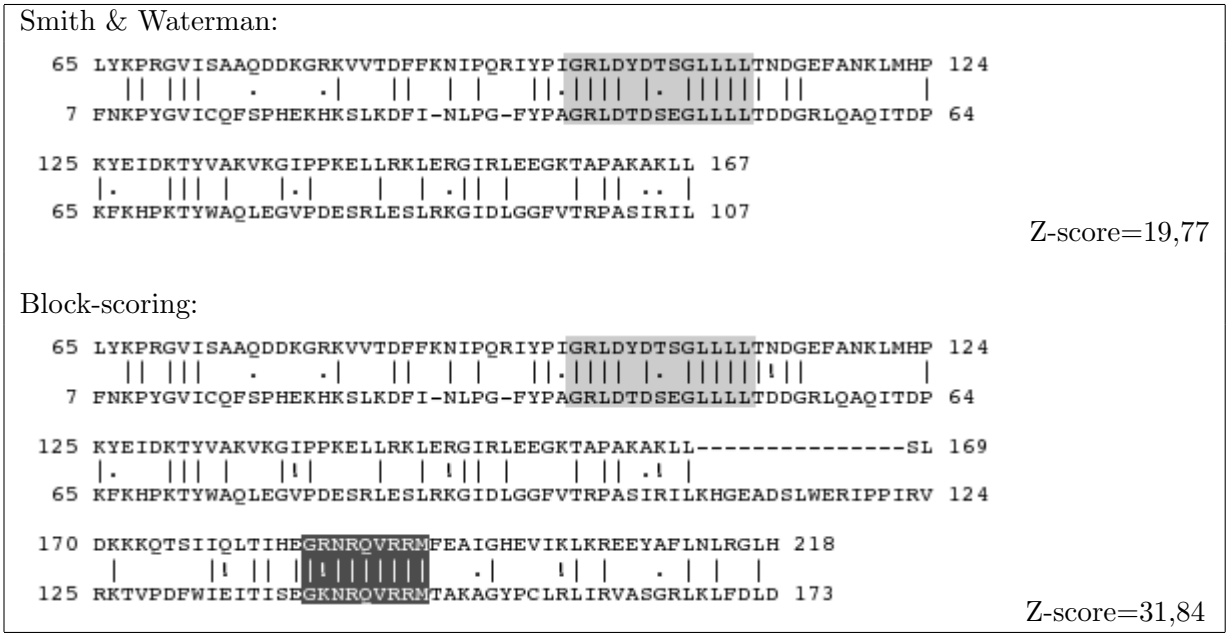

Figure 2: SW and BS best local alignments of sequences BS_ypuL and NMB1496 of the COG1187 family. (PAM 250, γ_o=15, γ_p=5, K=20.)

given by BS is relevant.

The other case (about 10% of the alignments) groups alignments which are generally very different, while the subsequences aligned by SW are always included in those aligned by BS. We picked up and studied thoroughly some of these alignments in four families, attempting to determine which one was the more relevant, biologically speaking. For this purpose, we seeked for motifs in PROSITE and BLOCKS and performed multiple alignments with CLUSTALW [25]. Confirmation was searched in the literature. In each case, it turned out that BS provided the more biologically relevant alignment. And in most cases, changing γ_o and γ_e for SW did not cause it to fit better. An example is illustrated in Figure 3. It concerns two alanyl tRNA synthetase proteins from the COG0013 family. The two alignments are totally different: for example, see the two motifs in grey, which are aligned by SW but lie far away in BS. Request on PROSITE, as well as multiple alignment of the whole set of alanyl tRNA synthetases of COG0013, show that the right alignment is given by BS. In particular, the two motifs in grey in the BS alignment are confirmed by the multiple alignment. Shortly, the other alignments that we have examined are the following. We just write here some indications. The corresponding alignments and supplementary results can be read at the following address: `http://www.igmors.u-psud.fr/BioInfo/fr/align/add_data`.

- In the COG1187 family, SW alignments of slr0361 and NMA1016, slr0361 and TP0459, yjbc and NMB0806 respectively, miss the important motif NKP, while BS aligns it well. Results were confirmed with a multiple alignment of the whole family. PROSITE and [22] show that the N-terminal domain belongs to a family of RNA-binding domains and that the NKP motif is highly conserved in this family.

- In the COG2813 family, a similar situation occurs: for at least four alignments, BS

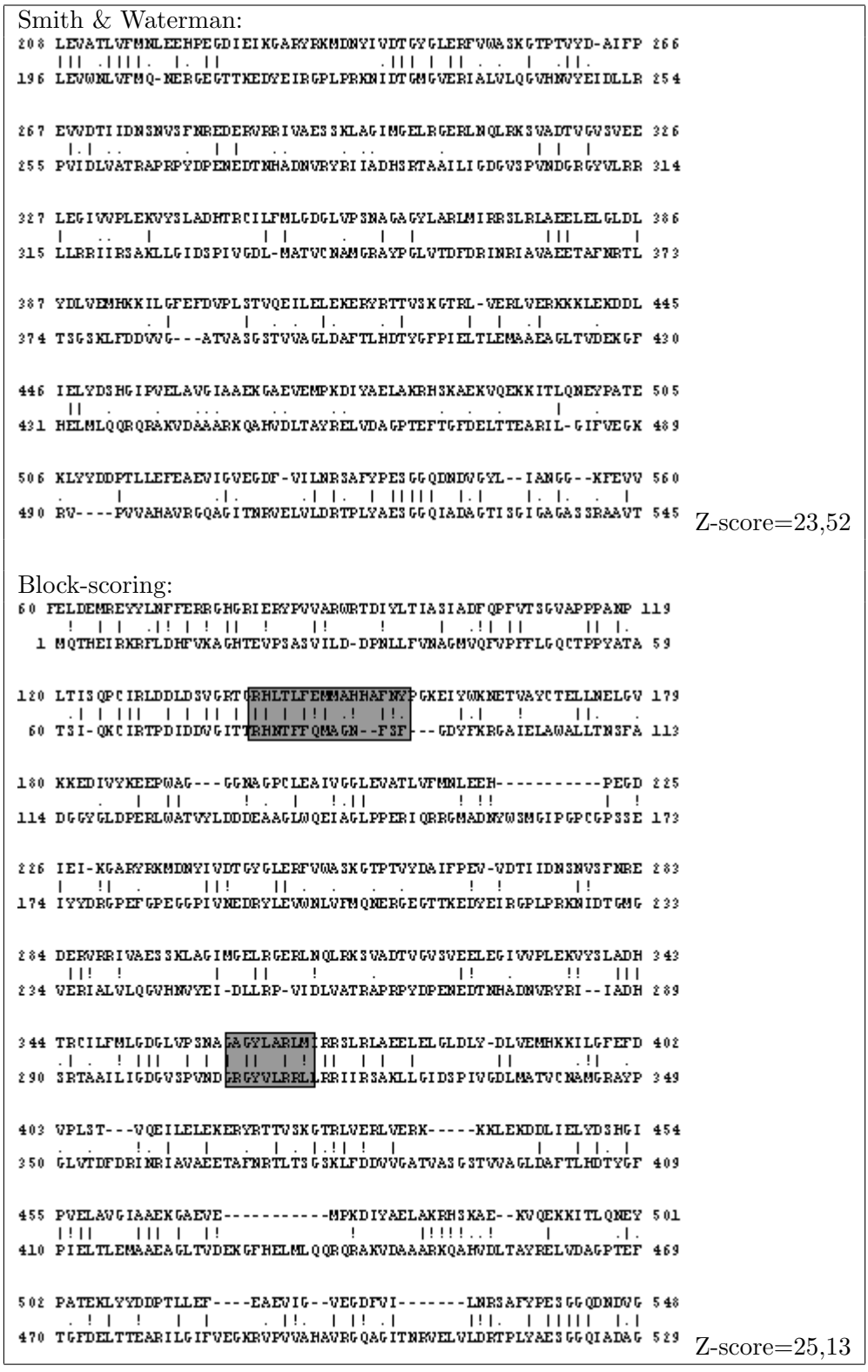

Figure 3: SW and BS best local alignments of sequences AF2255 and ML0512 of the COG0013 family. (PAM 250, γ_o=15, γ_e=5, K=20.)

aligns well the important NPP catalytic site [4], while SW does not.

- Proteins of the COG1514 family belong to the 2H phosphoesterase superfamily [15](phosphoesterase with two conserved histidines). Proteins of this superfamily share a common active site characterized by two conserved motifs where a histidine is present in each of them. At least three SW alignments miss one of the conserved motif while BS aligns it well.

4.3 Searching for false positives

4.3.1 Comparison of Z-scores

Since BS gives generally longer local alignments and higher scores than SW, the possibility of false positives had to be studied. Hence we compared Z-scores of alignments given by SW and BS. The classical formula for the Z-score is:

$$Z(v,w) = \frac{S(v,w) - E}{\sigma}$$

where E and σ stand, respectively, for the expected score and the standard deviation of alignments of v with a random sequence which has the same numbers of occurrences of residues than w. Since no formula is available for E and σ, and in order to avoid problems of 'asymmetry' of Z-scores, we followed [7] and computed:

$$\text{Z-score} = \min(Z'(v,w), Z'(w,v))$$

where $Z'(v,w)$ stands for the Z-score measured experimentally by generating n random sequences having the same numbers of occurrences of residues than w.

In order to search for possibly false-positives, we generated random aminoacids sequences (with Bernoulli probabilities) of three different kinds:

- sequences where all aminoacids were equally distributed;
- sequences where some kinds of aminoacids were overrepresented, according to their properties (e.g. hydrophobicity, polarity);
- sequences where some kinds of aminoacids were overrepresented, according to their ability to form block-matches (high-scoring aminoacids).

We computed Z-scores of about 3000 alignments of these three kinds of sequences. Each Z-score was computed on the basis on 500 random sequences of length 300. We set to 10 the Z-score significance threshold, as usually. None false-positive was detected when using BS, comparatively to SW. Moreover, in some cases the Z-score for BS was lower than for SW.

We also computed Z-scores of alignments of clearly homologous biological sequences. Obviously, in all alignments both Z-scores were significant. In most cases, Z-score for BS was 5 to 50% higher than for SW.

4.3.2 Lengthening of BS alignments

In order to measure the rate of lengthening of BS alignments with respect to SW ones, we generated random sequences as follows.

The COG2813 family contains 8 sequences for which SW and BS algorithm give exactly the same alignments (using PAM 250 with γ_o=15, γ_e=5, and K=20.) Hence we took these 8 sequences and generated from them two kinds of sets of randoms ones:

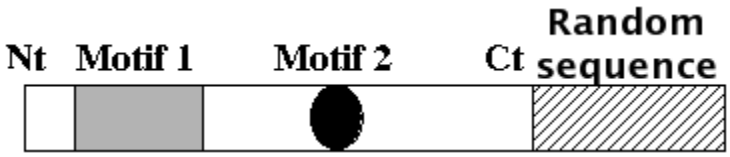

Figure 4: First set of random sequences generated from COG2813 proteins

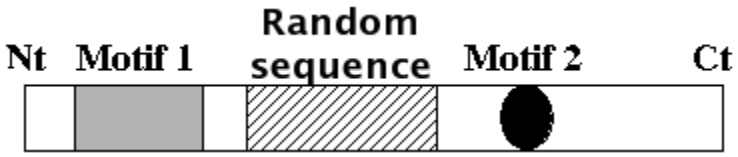

Figure 5: Second set of random sequences generated from COG2813 proteins

- by adding random stretches of aminoacids (of lengths from 10 to 50) just after their C-terminal extremities (Figure 4);
- by adding random sequences of aminoacids (of lengths from 10 to 120) between two known relevant motifs (Figure 5);

For each length of stretch, we performed alignments of any pair of "random" proteins of the first set. Table 3 shows the results. The lengthening due to BS grows significantly faster comparatively to SW. Then we performed alignments between each original protein and each sequence in the second set. We present in Table 4 the results. Up to about 20 nucleotides, SW as well as BS align both motifs. Then BS still aligns them in a non-negligible rate far beyond.

5 Asymptotics for the detected score.

The aim of a local alignment algorithm is twofold: find parts of the sequences that seem to be related, and decide whether this similarity comes from true homology or is only due to randomness. In section 4, we investigated the first point on biological examples and we have shown how the BS algorithm can find interesting homologies between distant sequences.

stretch	10	20	30	40	50
SW	13 / 5	20 / 10	21 / 14	20 / 18	18 / 11
BS	21 / 7	27 / 15	35 / 15	32 / 21	30 / 30

Table 3: Lengthening of alignments according to the number of nucleotides added after the C-terminal extremities. The first number of each cell denotes the number of alignments (over 64) which are lengthened, the second number denotes the average lengthening. (PAM 250, γ_o=15, γ_e=5, K=20.)

#residues	10	20	30	40	50	60	70	80	90	100	110	120
SW	56	56	41	18	13	2	2	0	0	0	0	0
BS	56	56	56	49	50	33	32	26	22	20	16	14

Table 4: Lengthening of alignments according to the number of nucleotides added between the two motifs. The number in each cell denotes the number of alignments (over 56) which contain both motifs. (PAM 250, γ_o=15, γ_e=5, K=20.)

For the second point, one has to compare numerically scores obtained for truly homologous sequences and scores obtained when the sequences are not related. Let $v^n = v_1 v_2 \cdots v_n$ and $w^m = w_1 w_2 \cdots w_m$ be the sequences to be locally aligned, and let $S(v^n, w^m)$ be their BS alignment score. Since homologous subsequences come from higher similarity, we forecast a higher score. Thus to decide whether the aligned sequences are related, one compares the obtained score to a threshold $t_{n,m}$: if $S(v^n, w^m) \leq t_{n,m}$, one decides that the sequences are not related, and if $S(v^n, w^m) \geq t_{n,m}$, one decides that the sequences are related. The statistical formulation of such a decision's rule is a test between the hypotheses H_0: "the sequences are independent" and H_1: "the sequences are related". In other words, one assumes that the words $v_1 v_2 \cdots v_n$ and $w_1 w_2 \cdots w_m$ are the realizations of random processes $V^n = V_1 V_2 \cdots V_n$ and $W^m = W_1 W_2 \cdots W_m$. To set the threshold $t_{n,m}$, one has to know the distribution of the (random) score $S(V^n, W^m)$ under H_0, that is when the sequences V^n and W^m are independent; $t_{n,m}$ is then some chosen quantile of the distribution. Numerical evaluation of quantiles by simulation may be used (or bootstrap methods) since the determination of the exact (or asymptotic) distribution is a difficult task in general. For the Smith & Waterman alignment score, the asymptotic distribution was known till recently only for ungapped alignments, see Dembo et al. [8]. Asymptotic approximations for p-values in the general case have been proposed heuristically and by simulations. A recent paper gives asymptotic approximations for the p-values of the local alignment score of Smith & Waterman, see Siegmund and Yakir [21], see also Chan [5], where more detailed literature on the subject of p-values evalutions may be found.

The aim of this section is to give a preliminary asymptotic result on the behaviour of the BS alignment score, namely the asymptotic equivalent of the expectation, in particular situations: for blocks of maximal length K (as in example 2) and with no gaps allowed.

Assume that $V = (V_i)_{i \geq 1}$ and $W = (W_i)_{i \geq 1}$ are independent sequences of independent random variables with values in $\mathcal{X}$. Assume also that $m = n$, so that $V^n = V_1 V_2 \cdots V_n$ and $W^m = W_1 W_2 \cdots W_n$ are the words to be aligned. Define the K dimensional random vectors $X_i = (V_i, \ldots, V_{i+K-1})$ and $Y_i = (W_i, \ldots, W_{i+K-1})$. Let F be some real function on $\mathcal{X}^K \times \mathcal{X}^K$ giving a positive probability to positive values:

$$P\left(F(X_1, Y_1) > 0\right) > 0, \tag{1}$$

and such that the expectation

$$E\left(F(X_1, Y_1)\right) < 0. \tag{2}$$

Define

$$S_n = \max_{l\geq 1,\ 1\leq i,j\leq n-l-K+2} \sum_{k=0}^{l-1} F(X_{i+k}, Y_{j+k}).$$

Notice that assumption (2) is required, as usual, to be in the so-called logarithmic scale. In the opposite situation where (2) does not hold, S_n grows linearly with n and does not lead to a consistent test of hypothesis of H_0 against H_1.
The asymptotic result is the following.

Theorem 1 *Under the assumptions (1) and (2), there exists a real number γ^* (see below) such that almost surely*

$$\lim_{n\to\infty} \frac{S_n}{\log n} = \gamma^*.$$

Application to BS alignment score. For particular choices of the function F, Theorem 1 gives the asymptotic leading term of the BS alignment score. The idea is that, knowing X_i and Y_j, one can tell the length of the block ending with V_{i+K-1} and W_{j+K-1} in an alignment where X_i and Y_j are aligned, in case this length is not bigger than K, or tell that this block has length at least K. In other words, with the notations of Section 3, $\inf\{K, H_{i+K-1,j+K-1}\}$ is a function of (X_i, Y_j). It is thus possible to define

$$F(X_i, Y_j) = \begin{cases} g\begin{pmatrix} X_i \\ Y_j \end{pmatrix}; \inf\{K, H_{i+K-1,j+K-1}\}\Big) & \text{if} H_{i+K-1,j+K-1} \geq 1 \\ s(V_{i+K-1}, W_{j+K-1}) & \text{otherwise} \end{cases}$$

Doing so, it appears that, taking $\delta = +\infty$ (that is allowing no gaps), $S_n \approx S(V^n, W^n)$ for big n. Indeed, ignoring boundary effects S_n is the score of the local BS alignment of V^n and W^n when no gaps are allowed. More precisely, for each local alignment

$$A(i,j,l) = \begin{matrix} V_iV_{i+1}\dots V_{i+K+l-2} \\ W_jW_{j+1}\dots W_{j+K+l-2} \end{matrix}$$

the BS score is

$$S(A(i,j,l)) = f(A(i,j,0)) + \sum_{k=0}^{l-1} F(X_{i+k}, Y_{j+k}),$$

and asymptotically the BS local alignment score is equivalent to S_n.

Sketch of proof of Theorem 1. The proof follows closely that of Dembo et al. [8], and is detailed in [16]. The main points are the following:

- Use a large deviations result on additive functions of Markov chains to obtain that the typical length of the alignment of maximal score is of order $\log n$ at most.
- Use the fact that the Markov chain $((X_i, Y_i))$ is K-dependent to be able to use the same arguments (using the method of types) as [8] for proving the upper and the lower bound for S_n.

The value of the limit. Let μ_V be the distribution of the variables V_i, and μ_W be the distribution of the variables W_i. Now, (X_i) is clearly a Markov-chain, with transition matrix Π_X which is irreductible and has stationary distribution $\mu_X = \mu_V^{\otimes K}$, that is the distribution of K independent random variables with the same distribution μ_V. Also, (Y_i) is a Markov-chain, with transition Π_Y which is irreductible and has stationary distribution $\mu_X \mu_W^{\otimes K}$. Thus, $((X_i, Y_i))$ is also an irreductible Markov chain with transition $\Pi_{(X,Y)}$ given by

$$\Pi_{(X,Y)}((i,j)|(k,l)) = \Pi_X(i|k)\Pi_Y(j|k),$$

and stationary distribution $\mu_X \otimes \mu_Y$.

Let ν be some probability measure on some finite set $(\Omega)^2$, and Π some transition matrix on Ω. Denote by ν_1 and ν_2 the marginal distributions of ν. Define the Kullback-Leibler information divergence between ν and Π by

$$D(\nu||\Pi) = \sum_{x,y \in \mathcal{X}} \nu(x,y) \log \frac{\nu(x,y)}{\nu_1(x)\Pi(y|x)},$$

with $\log 0 = -\infty$, $\log \frac{a}{0} = +\infty$ if $a > 0$, $0 \log 0 = 0$ and $\log \frac{0}{0} = 0$. Define also for any ν on $(\mathcal{A}^K \times \mathcal{A}^K)^2$

$$D^*(\nu) = \max \left\{ \frac{1}{2} D(\nu||\Pi_{(X,Y)}) \, ; D(\nu_1||\Pi_X) \, ; D(\nu_2||\Pi_Y) \right\},$$

and

$$J(\nu) = \frac{E_{\nu_1}(F(X,Y))}{D^*(\nu)}.$$

Then the constant γ^* is the maximum value of J over the distributions having the same marginal distributions:

$$\gamma^* = \sup \left\{ J(\nu) \ : \ \nu_1 = \nu_2 \right\}.$$

Since the distributions of X and Y are unknown, they have to be estimated empirically, that is on the sequences, to obtain an estimate of γ^*.

Comments. Obviously, gaps have to be allowed in applications, and are allowed in the algorithm. One could follow the ideas in Zhang [28] to obtain the asymptotic equivalent of the expectation when gaps are allowed, and the ideas in Siegmund and Yakir [21] to evaluate approximations of the distribution and the quantiles of the local BS alignment score.

6 Conclusion.

Our experiments show that block-scoring effectively detects relevant similar blocks in cases where the sequences, while homologous, are too distant for the classical alignment algorithm to be fully accurate. Moreover, block-scoring does not give raise to false-positive in comparison with Smith & Waterman. We think that block-scoring is worth being used complementarily to the Smith & Waterman algorithm when precise block information has to be detected.

The algorithm runs in time $O(n^3)$, but this is a very crude bound. Experimentally the program runs much faster, particularly when sequences are distant, because $H_{i,j}$'s are small is this case. A theoretical study of the complexity has yet to be done.

Attention must be paid to the values of gap-opening and gap-extension penalties: they have to be much higher for block-scoring than for Smith & Waterman; and block-scoring

algorithm is very sensitive to variations of these values. On the other hand, since the scoring scheme of block-scoring takes into account not only one-to-one residues substitutions but dependencies within blocks, probably designing new *ad hoc* substitution matrices would improve even more its results. In addition, the scoring model that we experimented here is a very particular and "basic" case of the general scheme given in Section 2. Other more sophisticated models are worth to be studied.

The mathematical part of the paper constitutes a first step to the analytical study of the Z-score of block-scoring alignments, aiming to determine precise thresholds for deciding if two sequences are related or not.

Finally, and thanks to David Abergel, the program is available on-line at the following address: **http://www.igmors.u-psud.fr/BioInfo/fr/align**.

Acknowledgements

We thank David Abergel for his help, and Francis Quetier for his encouragements and kind suggestions.

References

[1] Altschul, S.F., Gish, W., Miller, M., Myers, E.W. and Lipman, D.J. (1990): *Basic local alignment search tool*. Journal of Molecular Biology, 215, pp. 403-410.

[2] Arslan, A.N., Egecioglu,O., and Pevzner, P.A. (2001) *A new approach to sequence comparison: Normalized local alignment.* Bioinformatics, 17(4):327-337.

[3] Blake, J.D.and Cohen, F.E. (2001): *Pairwise sequence alignment below the twilight zone*. Journal of Molecular Biology, 307, pp. 721-735.

[4] Bujnicki, J.M., Rychlewski, L. (2002): *RNA:(guanine-N2) methyltransferase RsmC/RsmD and their homologs revisited - bioinformatic analysis and prediction of the active site based on the uncharacterized Mj0882 protein structure*. BMC Bioinformatics. 2002 Apr 3;3(1):10.

[5] Chan, H.P. (2003): *Upper bounds and importance sampling of p-values for DNA and protein sequence alignment*. Bernoulli, 9 (2) , pp. 183-199.

[6] Clote, P. and Backofen, R. (2000): *Computational Molecular Biology. An Introduction*. Wiley series in mathematical and computational biology.

[7] Comet J.P., Aude J.C., Glémet E., Risler J.L., Hénaut A., Slonimski P.P., Codani J.J. (1999): *Significance of Z-value statistic of Smith-Waterman scores for protein alignments*. Computers & Chemistry 23, pp. 317-331

[8] Dembo, A., Karlin, S. and Zeitouni, O. (1994): *Critical phenomena for sequence matching with scoring*. Annals of Probability 22, pp. 1993-2021.

[9] Dembo, A., Karlin, S. and Zeitouni, O. (1994): *Limit distribution of maximal non-aligned two-sequence segmental score*. Annals of Probability 22, pp. 2022-2039.

[10] Durbin, R., Eddy, S., Krogh, A. and Mitchison, G. (1998): *Biological Sequence Analysis: probabilistic models of proteins and nucleic acids*. Cambridge University Press.

[11] Henikoff, S., Henikoff, J.G., and Pietrokovski, S. (1999) *Blocks+: A non-redundant database of protein alignment blocks dervied from multiple compilations* Bioinformatics 15(6):471-479.

[12] Hulo N., Sigrist C.J.A., Le Saux V., Langendijk-Genevaux P.S., Bordoli L., Gattiker A., De Castro E., Bucher P., Bairoch A. *Recent improvements to the PROSITE database.* Nucl. Acids. Res. 32:D134-D137(2004)

[13] Liu, J.S., Neuwald, A.F. and Lawrence, C.E. (1999): *Markovian structures in biological sequence alignments*. Journal of the American Statistical Association 94, pp. 1-15.

[14] Lam, F., Alexandersson, M. and Pachter, L. (2003): *Picking Alignments from Steiner Trees*. Journal of Computational Biology 10, pp. 509-520.

[15] Mazumder, R., Iyer, L.M., Vasudevan, S. and Aravind, L. (2002): *Detection of novel members, structure-function analysis and evolutionary classification of the 2H phosphoesterase superfamily*. Nucleic Acids Research vol 30 no 23, pp. 5229-5243.

[16] Nédélec, E. (2000): *A new statistic for sequences alignment based on blocks of matches*. Manuscript.

[17] 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*. Journal of Molecular Biology 48, pp. 443-453.

[18] Pearson, W.R. and Lipman, D.J. (1988): *Improved tools for biological sequence comparison*. Proceedings of the National Academy of Science USA 85, pp. 2444-2448.

[19] Rost, B. (1999): *Twilight zone of protein sequence alignments*. Protein Engineering 12, pp. 85-94.

[20] Sellers, P.H. (1974): *On the theory and computation of evolutionary distances*. SIAM J. Appl. Math. 26, 787-793.

[21] Siegmund, D. and Yakir, B. (2000): *Approximate p-values for local sequence alignments*. The Annals of Statistics 28, pp. 657-680 .

[22] Sivaraman et al. (2002): *Structure of the 16S rRNA pseudouridine synthase RsuA bound to uracil and UMP*. Nature Structural Biology vol 9 num 5, pp. 353-358.

[23] Smith,T.F. and Waterman, M.S. (1981): *Identification of common molecular subsequences*. Journal of Molecular Biology 147, pp. 195-197.

[24] Tatusov, R.L., Natale D.A., Gargavtsev, I.V., Tatusova, T.A., Shankavaram, U.T., Rao, B.S., Kiryutin, B., Galperin, M.Y., Fedorova, N.D. and Koonin, E.V. (2001): *The COG database: new developments in phylogenetic classification of proteins from complete genomes*. Nucleic Acids Research vol 29 no 1, pp. 22-28.

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

[26] Thorne,J.L., Kishino, H. and Felsenstein, J. (1991): *An evolutionary model for maximum likelihood alignment of DNA sequences.* Journal of Molecular Evolution 33, pp. 114-124.

[27] Waterman, M.S. (1995): *Introduction to computational biology.* Chapman & Hall, New-York.

[28] Zhang, Y. (1995): *A limit theorem for matching random sequences allowing deletions.* Annals of applied Probability 5, pp. 1236-1240.