h-tuple Approach to Evaluate Statistical Significance of Biological Sequence Comparison with Gaps,

Afshin Fayyaz Movaghar, Sabine Mercier, Louis Ferré

Statistical Applications in Genetics and Molecular Biology

Volume 6, *Issue* 1 2007 *Article* 22

H-Tuple Approach to Evaluate Statistical Significance of Biological Sequence Comparison with Gaps

Afshin Fayyaz movaghar* Sabine Mercier†

Louis Ferré‡

*Institute of Mathematics of Toulouse, University Toulouse 2, fayyaz@univ-tlse2.fr

†Institute of Mathematics of Toulouse, University Toulouse 2, mercier@univ-tlse2.fr

‡Institute of Mathematics of Toulouse, University Toulouse 2, loferre@univ-tlse2.fr

H-Tuple Approach to Evaluate Statistical Significance of Biological Sequence Comparison with Gaps*

Afshin Fayyaz movaghar, Sabine Mercier, and Louis Ferré

Abstract

We propose an approximate distribution for the gapped local score of a two sequence comparison. Our method stands on combining an adapted scoring scheme that includes the gaps and an approximate distribution of the ungapped local score of two independent sequences of i.i.d. random variables. The new scoring scheme is defined on h-tuples of the sequences, using the gapped global score. The influence of h and the accuracy of the p-value are numerically studied and compared with obtained p-value of BLAST. The numerical experiments emphasize that our approximate p-values outperform the BLAST ones, particularly for both simulated and real short sequences.

KEYWORDS: gapped alignment, local score, p-value

---

*Institute of Mathematics of Toulouse, University Toulouse 2 (Le Mirail). To whom correspondence should be addressed: fayyaz@univ-tlse2.fr We are grateful to the referees for their valuable comments and suggestions which help to clarify our purpose.

1 Introduction

An important step in learning the function of a new biological sequence (DNA or protein) is to compare the new sequence with existing sequences belonging to a database whose biological functions are known. So, finding the database sequences similar to the new sequence can make a guess about its function. This similarity can mean that these sequences are copied from one generation to another, and undergo changes (within any population over the course of many generations), as random mutations, arise and become fixed in the population. Evolutionary theory emphasizes that genes/proteins which have a similar function are likely to have evolved from a common ancestor through mutation. The simplest events that occur during the course of evolution are substitution of one nucleotide/amino acid by another and insertion or deletion (gap). In this paper, we are interested in local alignments and corresponding quality as a similarity factor. It is generally measured from a score calculated by adding substitution scores, s, for each aligned pair of letters and gap penalty, δ, for each gap. Then, any local alignment of sequences can be scored and ranked according to this scoring scheme. The maximum-scoring local alignment is called the **local score**.

The main problem of sequence comparison is to evaluate the statistical significance of sequences showing a particular level of similarity. That is, whether an observed score could have arisen by chance under an appropriate model of random sequence. In this paper, we propose a new approach for assessing the similarity of sequences when gaps are allowed.

Let $\mathbb{A} = A_1, A_2, ..., A_n$ and $\mathbb{B} = B_1, B_2, ..., B_m$ be two independent sequences of i.i.d. random variables on a finite alphabet set $\mathcal{A}$. Let $s : \mathcal{A} \times \mathcal{A} \to \mathbb{Z}$ be a scoring function, and δ be a gap penalty. Let $M_{n,m}$ be the optimal score over all possible choices of two contiguous regions I and J for $I \subset \{A_1, A_2, ..., A_n\}$ and $J \subset \{B_1, B_2, ..., B_m\}$. Formally, the gapped local score is defined (*e.g.*, Waterman, 2000) by

$$M_{n,m} = \max_{I,J} S(I, J) \tag{1}$$

where

$$S(I, J) = \max\{-\delta(\iota - \ell + \tau - \ell) + \sum_{k=1}^{\ell} s(A_{u(k)}, B_{v(k)})\} \tag{2}$$

and the maximum is taken over all global alignments given by two increasing sequences $u(\cdot)$ and $v(\cdot)$ where ℓ is the number of pairs of aligned letters and ι, τ are respectively the lengths of I and J. To apply the affine gap, it is sufficient to

add a term $-\Delta d$ into the maximum of (2), where Δ is the gap opening penalty and d is the number of gaps of the corresponding alignment.

For the ungapped local score with shift defined as

$$H_{n,m} = \max_{\substack{0\leq\ell\leq\min(n,m)-1 \\ 1\leq j\leq m-\ell \\ 1\leq i\leq n-\ell}} \sum_{k=0}^{\ell} s(A_{i+k}, B_{j+k}), \tag{3}$$

the distribution has been asymptotically derived as an extreme-value distribution (Dembo et al., 1994), when $E[s(A, B)]$ is strictly negative while for finite length, Mercier and Daudin (2001) have proposed a method without explicit formula nor constraint on $E[s(A, B)]$ to approximate the ungapped local score of two sequences.

For a long time, for the gapped case, there was a great deal of empirical, rather than theoretical, evidence (Mott, 1992; Altschul and Gish, 1996; Waterman and Vingron, 1994; Spang and Vingron, 1998), indicating that the extreme-value theory underlying the ungapped case carries over, provided the gap penalties are drastic enough. In practice, several methods for estimating the parameters of the p-value of $M_{n,m}$ are used like method of moments (Altschul and Gish, 1996) and maximum likelihood (see, *e.g.*, Bailey and Gribskov, 2002). The time-consuming in the gap case results of the lack of explicit formulae for the parameters. Fortunately, it has been recently improved by means of some conjectures which allow the parameters for gapped local alignment to be estimated from global alignment (Chia and Bundschuh, 2006; Sheetlin et al., 2005; Park et al., 2005; Bundschuh, 2002; Grossmann and Yakir, 2004). A heuristical approximate p-value, using Greedy Extension Model, has been proposed by Mott and Tribe (1999): the authors piece together results for ungapped alignments to obtain useful numerical approximations for gapped alignments. Recently, an approximate p-value has been theoretically derived under certain severe conditions on gaps (Siegmund and Yakir, 2000; 2003). In a way, their approach generalizes both the Dembo et al. (1994) results as well as Mott and Tribe's (1999). However, this approximation involves an infinite sequence of difficult-to-compute parameters even if it is numerically shown that they can finally be reduced to two (Storey and Siegmund, 2001). Many works (see, *e.g.*, Mitrophanov and Borodovsky (2006) for a review) use the Karlin and Altschul (1990) formula, therefore, they are well adapted for long sequences but they are less efficient for short ones. We rather use the Mercier and Daudin's (2001) approach in order to provide a new method that yields appealing results particulary for short sequences.

In this paper, we insert gaps in the scoring function, in such a way that the p-value of gapped alignment can be derived from the one of ungapped case, proposed by Mercier and Daudin (2001). Then, we numerically assess the performance

and quality of the proposed p-value. The theoretical evaluation of this new approximate p-value is a challenge which deserves serious attention but it is out of the scope of this paper.

In Section 2, we present our approach. We recall, in Section 2.1, the method for achieving the approximate p-value in the ungapped case presented by Mercier and Daudin (2001) for two independent sequences of i.i.d. random variables. In Section 2.2, we suggest a new gapped local score, denoted $\mathfrak{M}_{n,m}$, that approximates $M_{n,m}$ in (1) by calculating the ungapped local score of h-tuples of the two sequences $\mathbb{A}$ and $\mathbb{B}$ for a given h. This local score is obtained by using the scoring scheme based on the gapped global score of the h-tuples of $\mathbb{A}$ and $\mathbb{B}$ derived from (2). Its approximate p-value is derived by using the approximate distribution of ungapped local score (Mercier and Daudin, 2001). Section 3 is devoted to simulations. In Subsection 3.1, we numerically verify that the assumption of independence among shifts in the ungapped case used in Mercier and Daudin (2001) can be circumvented without any damage. Then, in Section 3.2, we focus on the accuracy of $\mathfrak{M}_{n,m}$ to approximate $M_{n,m}$ through a comparison with the approximated gapped local score proposed by Zhang (1995). The new p-value is then compared with an empirical one in order to find an appropriate value for h, the length of h-tuples, in Section 3.3. In Section 3.4 and 3.5, we assess our approximate p-value via a comparison with BLAST on both simulated and real (SCOP 1.53) database and we finally conclude in Section 4.

2 Statistical Significance

In this section, we suppose, without loss of generality, that $m \geq n$, where n and m are respectively the lengths of the i.i.d. sequences $\mathbb{A}$ and $\mathbb{B}$.

2.1 Ungapped alignments

As mentioned in the introduction, our approach in the gapped case requires the computation of an approximated p-value derived by Mercier and Daudin (2001). So, we briefly recall how this p-value is obtained. The p-value of ungapped local score (with shift), $P(H_{n,m} \geq a)$ where $H_{n,m}$ is defined in (3), is approximated in (Mercier and Daudin, 2001) by

$$p_u(a) = 1 - \left[\prod_{i=1,\ldots,n-1} P(H_i < a)^2\right] P(H_n < a)^{m-n+1} \tag{4}$$

where H_i is the local score for any two continuous subsequences $(E_1, \ldots, E_i) \subset \mathbb{A}$ and $(F_1, \ldots, F_i) \subset \mathbb{B}$ with same length i, *i.e.*,

$$H_i((E_1, \ldots, E_i), (F_1, \ldots, F_i)) = \max_{1 \leq k \leq l \leq i} \sum_{t=k}^{l} s(E_t, F_t). \tag{5}$$

The p-value of H_i is derived from the one sequence case in Mercier and Daudin (2001). This latter method is based on Markov chain theory, particulary Lindley process. In this case, the distribution is independent of the sign of the expected score of residues, unlike Karlin and Altschul (1990), and it is shown that

$$P[H_n < a] = 1 - P_1 \Pi^n P'_{a+1} \tag{6}$$

where $H_n = \max_{1 \leq i \leq j \leq n} \sum_{k=i}^{j} \widetilde{s}(A_k)$ is the local score of sequence $\mathbb{A}$ with $\widetilde{s}$ defined on $\mathcal{A}$ for the one sequence case, a is the observed local score of the studied sequence, P_i is a vector $1 \times (a+1)$ whose ith element is one and zero elsewhere and the $(a+1) \times (a+1)$ matrix Π is filled by using the letter score distribution as follows

$$\Pi = \left(\begin{array}{c|ccc|c} F(0) & p(0) & \ldots & p(a-1) & 1 - F(a-1) \\ \vdots & & \vdots & & \vdots \\ F(-t) & \ldots & p(\ell - t) & \ldots & 1 - F(a-t-1) \\ \vdots & & \vdots & & \vdots \\ F(1-a) & p(2-a) & \ldots & p(0) & 1 - F(0) \\ \hline 0 & 0 & \ldots & 0 & 1 \end{array} \right)$$

where $F(k) = P[\widetilde{s}(A) \leq k]$ and $p(k) = P[\widetilde{s}(A) = k]$ and $0 \leq t, \ell \leq a$. Note that $\widetilde{s}$ takes its values in $\mathbb{Z}$.

This approximated p-value, $p_u(a)$, relies on the independence assumption among shifts. Of course, dependence exists particulary among close shifts and this issue will be numerically addressed in Section 3.1.

2.2 Gapped alignments

Our aim is to propose a new theoretical approximate p-value, when gaps are allowed, which stands on a method different from the previous studies (see, *e.g.*, Mott and Tribe, 1999; Siegmund and Yakir, 2000; 2003). We extend (4), obtained for the ungapped alignment, by using the idea of Zhang (1995) who incorporates gaps through the choice of a scoring function. This scoring function is defined on the pairs of h-tuples of letters by using the gapped global score S defined in (2): we

insert the gaps through the scoring function and approximate the exact gapped local score of two sequences by the ungapped local score of the h-tuples produced from all possible shifts. We investigate an approach close to the one of Zhang who uses this scoring function to find an almost sure limit for the gapped local score, $M_{n,m}$. The differences will be highlighted below. Note first that, our work is not asymptotic, unlike others on this subject, see, *e.g.*, Karlin and Altschul (1990), Dembo et al. (1994) and Siegmund and Yakir (2000; 2003).

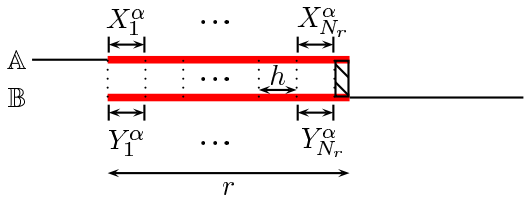

Figure 1: The h-tuples of a fixed shift α.

First, we define an approximate gapped local score and its p-value will be derived afterwards. Let α be an integer which assigns a number between $1-n$ and $m-1$ to each shift: when A_1 (respectively B_1) is aligned with $B_i,\ i=1,\ldots,n$ (resp. $A_j,\ j=1,\ldots,m$), α takes the value $1-i$ (resp. $j-1$). Let $r=r(\alpha)$ be the length of the opposite segments of the two sequences for α. For $i=1,\ldots,r$, let A_i^α (resp. B_i^α) be the ith letter of the segment of $\mathbb{A}$ (resp. $\mathbb{B}$) relative to shift α. Given a positive integer h, let $N_r=[\frac{r}{h}]$ be the number of created h-tuples on the opposite subsequences and let $\mathbb{X}^\alpha=\{X_i^\alpha\}_{i=1,\ldots,N_r}$ and $\mathbb{Y}^\alpha=\{Y_j^\alpha\}_{j=1,\ldots,N_r}$ with $X_i^\alpha=(A^\alpha_{(i-1)h+1},\ldots,A^\alpha_{ih})$ and $Y_j^\alpha=(B^\alpha_{(j-1)h+1},\ldots,B^\alpha_{jh})$. These definitions are illustrated in Figure 1 and examples are given in Figure 2. Note that $\mathbb{X}^\alpha$ and $\mathbb{Y}^\alpha$ are independent sequences of h-tuples in $\mathcal{A}^h$, since $\mathbb{A}$ and $\mathbb{B}$ are independent. It is clear that for a fixed α, the X_i^α's are independent and so are the Y_j^α's.

Let H_{N_r} be the ungapped local score without shift of the sequences $\mathbb{X}^\alpha$ and $\mathbb{Y}^\alpha$ defined as follows

$$H_{N_r}=\max_{1\le i\le j\le N_r}\sum_{k=i}^{j}S(X_k^\alpha,Y_k^\alpha). \tag{7}$$

The latter is a summation on the scoring function which is based on the gapped global score (2). We approximate the gapped local score of $\mathbb{A}$ and $\mathbb{B}$, $M_{n,m}$ in (1), by

$$\mathfrak{M}_{n,m}=\max_{\alpha}H_{N_r},\qquad \text{for}\qquad 1-n\le\alpha\le m-1. \tag{8}$$

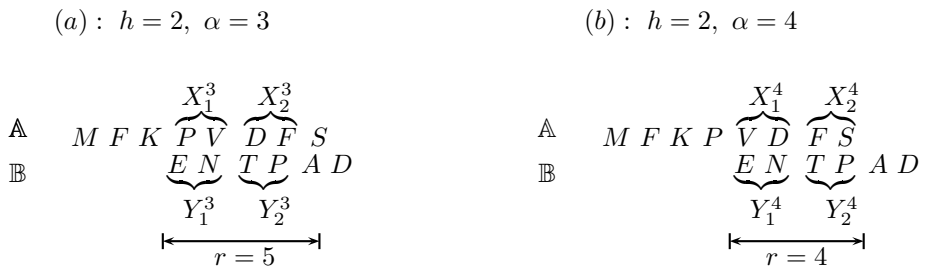

Figure 2: These examples indicate how we specify h-tuples of each shift for $h = 2$. (a) shows the shift corresponding to $\alpha = 3$ with opposite subsequences length $r = 5$. We have two pairs of h-tuples, *i.e.* $N_5 = 2$. As it can be observed the letters pair (S,A) is not used in the calculation of the relative score shift obtained by (7). (b) similarly corresponds to the shift $\alpha = 4$ with length $r = 4$ and $N_4 = 2$. In this shift, all letters are taken into account in the score shift calculation.

For all h, we clearly have, $\mathfrak{M}_{n,m} \leq M_{n,m}$.

Remark: Our local score differs from the one of Zhang (1995) by the way that the shifts are introduced. Indeed, in his case, the shifts are relative to the h-tuples (*i.e.* for $\mathbb{X}^0$ and $\mathbb{Y}^0$) while, in our work, the shifts are associated to letters (*i.e.* for $\mathbb{A}$ and $\mathbb{B}$). In other words, we define the h-tuples for each shift α whereas Zhang considers the shifts on the h-tuples of the initial sequences. Formally, the approximate local score of Zhang, denoted by $\mathcal{M}_{n,m}$, is defined as $\mathcal{M}_{n,m} = \max_{i,j} \sum_{1 \leq i \leq j \leq N_{r(0)}} S(\mathbb{X}_i^0, \mathbb{Y}_j^0)$ that is H_{N_r} for $\alpha = 0$. From another point of view, $\mathcal{M}_{n,m}$ corresponds to the maximum in (8) restricted to $(1 - n) \leq \alpha = \pm wh \leq (m - 1)$ where w is a nonnegative integer. It clearly gives that $\mathcal{M}_{n,m} \leq \mathfrak{M}_{n,m} \leq M_{n,m}$.

Our approximate p-value of the gapped local alignments is obtained by adapting (4) for (8). For a given h, the p-value of $\mathfrak{M}_{n,m}$ (and the p-value of $M_{n,m}$, $P(M_{n,m} \geq a)$) is estimated by

$$p_h(b_a) = 1 - \left[\left(\prod_{r=h,2h,\ldots,(K/h-1)h} P(H_{N_r} < b_a)^{2h} \right) P(H_{N_K} < b_a)^{2(n-K)} \right] \times P(H_{N_n} < b_a)^{m-n+1}, \quad (9)$$

where $K = [\frac{n-1}{h}]h$ and b_a is the observed value of $\mathfrak{M}_{n,m}$. The first and third terms in the products in (9) derive straightforwardly from (4). However, the values of

index r are multiples of h because h consecutive shifts are necessary to move from N_r to $N_r \pm 1$ (the effect of these h shifts occurs in the power of the probabilities in the product). In addition, as the common length of the opposite subsequences, r, is not necessarily a multiple of h, the second term in the products appears in this way to take into account the end of these opposite segments. The right hand side of (9) is calculated by using (6) where the matrix Π is filled by replacing $\widetilde{s}(\cdot)$ with scoring scheme $S(\cdot,\cdot)$ defined in (2).

3 Numerical Results

3.1 Study of the assumption of shift independence in the ungapped case

We numerically verify that relaxing the assumption of independence of shifts does not affect the results in practice. These simulations complete the results of Park and Spouge (2002). We independently generate a sample $\mathcal{S}$ of size $N = 10,000$ pairs of sequences, $\{(\mathbb{A}_k, \mathbb{B}_k)\}_{k=1,\ldots,N}$, for a given distribution of letters and for different given lengths, n and m. We calculate the exact ungapped local score of each pair, $H^k_{n,m} = H_{n,m}(\mathbb{A}_k, \mathbb{B}_k)$, using BLOSUM62. The empirical p-values, $p_e(a)$ are calculated by the ratio of the pairs whose ungapped local scores are greater or equal to a, *i.e.*,

$$p_e(a) = \frac{N_a}{N} = \frac{\#\{(\mathbb{A}_k, \mathbb{B}_k) : H^k_{n,m} \geq a\}}{N}. \tag{10}$$

For large values of N, the stability of the empirical distribution $p_e(a)$ is almost reached in practice.

In order to compare $p_u(a)$ of (4) with $p_F(a)$, the p-value obtained by FASTA, we consider a subsample of $\mathcal{S}$ of size $\mathcal{N}$ lower or equal to 60, denoted $\{(\mathbb{A}_\tau, \mathbb{B}_\tau)\}_{\tau=1,\ldots,\mathcal{N}}$. This limitation is imposed by the way FASTA is used here (*i.e.* to compare couples of sequences instead of a sequence with a database as usually). The FASTA program can be found at the following address: `www.infobiogen.fr/services/analyseq/cgi-bin/fasta_in.pl`. They are compared with the corresponding $p_e(a)$ by means of a χ^2 measure, *i.e.*,

$$\chi^2(p_u) = \sum_{\tau=1}^{\mathcal{N}} \frac{[p_e(a_\tau) - p_u(a_\tau)]^2}{p_e(a_\tau)} \tag{11}$$

where a_τ is the observed value of $H_{n,m}$ of the τth pair of the subsample. Table 1 shows that $\chi^2(p_u)$ is smaller than $\chi^2(p_F)$ for all of the different lengths and particulary for the short sequences.

Table 1: COMPARISON BETWEEN THE P-VALUE p_u AND FASTA ONE p_F

Lengths	m	40	57	85	106	561	422
	n	25	51	80	92	57	368
$\chi^2(\cdot)$	p_u	**0.651**	**0.428**	**1.045**	**0.423**	**0.269**	**0.793**
	p_F	17.741	10.240	8.958	3.912	7.121	10.939
PSE$(\cdot)$	p_u	**0.2468**	**0.2307**	**0.2412**	**0.2099**	**0.0075**	**0.4114**
	p_F	0.7875	0.9909	0.5231	0.6826	0.4405	0.6726

χ^2 and PSE values for the theoretical p-values p_u [see (4)] and the ones given by FASTA, p_F, using BLOSUM62.

Similar conclusions are reached by P-value **S**lope **E**rror (PSE), introduced by Bailey and Gribskov (2002). This metric allows the accuracy of different local score distribution methods to be compared. To get PSE, the weighted least-squares regression line of p-values logarithm, ***i.e.***,

$$\log(p_u) = r_u \log(p_e) + b \tag{12}$$

is calculated. The slope gives an indication of the direction and magnitude of the errors in the p-values. So, the p-value slope error metric is defined as

$$PSE(p_u) = 1 - r_u. \tag{13}$$

If the PSE value is close to zero, the points $(\log(p_u), \log(p_e))$ lie approximately along the line $x = y$, in other words, the estimates of the local score distribution is accurate. Note that, the PSE is only accurate when the coefficient of determination is close to 1 and the intercept to 0.

Table 1 also shows the absolute values of the PSE for p_u and p_F that confirm the results obtained by (11). In addition, to study the behavior of the p-value of ungapped alignments, p_u, and FASTA one, p_F, we plot the logarithm of the p-values, p_u and p_F, against the ones of p_e (see, ***e.g.***, Figure 3). A large dispersion among $(\log(p_F), \log(p_e))$'s also indicates the weak performance of FASTA (the largest correlation coefficient is 0.82 for the sequence lengths $m = 106$ and $n = 92$), although $(\log(p_u), \log(p_e))$'s can be properly explained as a line which is close to the line $x = y$ and passes through the origin (the smallest correlation coefficient is 0.94 for the sequence lengths $m = 85$ and $n = 80$).

These results clearly show the accuracy of p_u for approximating the empirical p-value and then, they indicate that the assumption of independence, used to derive $p_u(\cdot)$, does not play a crucial role in its calculation. So, it will be used below.

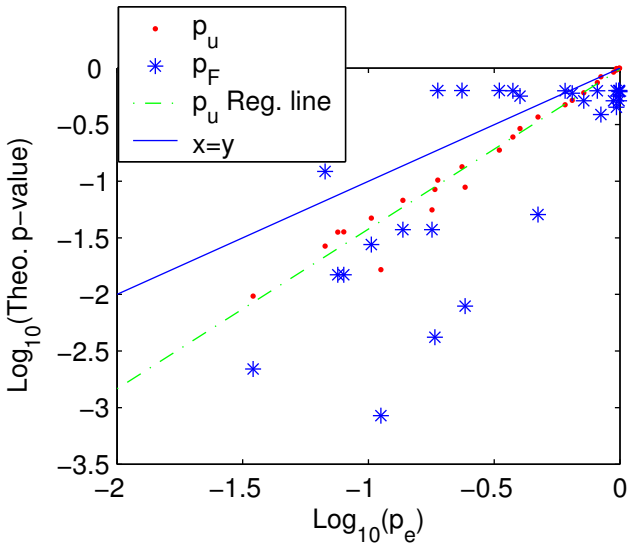

Figure 3: Behavior of p_u and the p-value obtained by FASTA, p_F, are compared with the empirical p-value for the sequences of lengths $m = 422$ and $n = 368$. This figure is based on the logarithm of the p-values. The correlation coefficients are 0.99 and 0.68 for $\log(p_u)$ and $\log(p_F)$, respectively.

3.2 The difference between approximate and exact gapped local score as a function of h

To check the quality of $\mathfrak{M}_{n,m}$, we generate several databases. Each database contains 1,000 pairs of sequences, independently generated with given lengths and a fixed letter probability which is calculated from the letters frequencies of some Homo sapiens (human) sequences, common for all databases. According to the lengths, the databases are classify to three categories: (i) "short" where $\max(n,m) < 100$, (ii) "medium" for $100 \leq n,m \leq 500$ and (iii) "long" where $\min(n,m) > 500$. There are 4 short, 6 medium and 3 long databases. We obtain $M_{n,m}$ and $\mathfrak{M}_{n,m}$ by using the scoring scheme BLOSUM62 with the penalty of gap opening and extension -11 and -2, respectively. For $h = 2,3,4,6,9$ and 11, the accuracy of $\mathfrak{M}_{n,m}$ is measured by

$$NMSE_h = \sum_{k=1}^{1000} \left(\frac{M_{n,m}^k - \mathfrak{M}_{n,m}^k}{M_{n,m}^k} \right)^2 \tag{14}$$

where $\mathfrak{M}_{n,m}^k$ and $M_{n,m}^k$ are, respectively, the estimated and the exact local score of the kth pair of sequences in our database. Table 2 shows that the $NMSE_h$ increases

Table 2: COMPARISON OF $\mathfrak{M}_{n,m}$ WITH $\mathcal{M}_{n,m}$ RELATIVELY TO THE EXACT LOCAL SCORE

	$NMSE_h$ of $\mathfrak{M}_{n,m}$			$NMSE_h$ of $\mathcal{M}_{n,m}$		
h \ length category	short	medium	long	short	medium	long
2	**0.007**	**0.010**	**0.010**	**0.089**	**0.057**	**0.048**
3	0.017	0.015	0.015	0.153	0.093	0.071
4	0.027	0.021	0.019	0.211	0.126	0.090
6	0.050	0.032	0.030	0.314	0.185	0.131
9	0.087	0.048	0.041	0.500	0.248	0.180
11	0.113	0.058	0.048	0.692	0.293	0.210

Average of $NMSE_h$ defined in (14), for two approximate local scores, $\mathfrak{M}_{n,m}$ defined in (8) and $\mathcal{M}_{n,m}$, proposed by Zhang (1995). The categories "short", "medium" and "long", respectively, are applied for the pairs of sequences whose lengths satisfy $\max(n,m) < 100$, $100 \leq n, m \leq 500$ and $\min(n,m) > 500$.

with h. In other words, $M_{n,m}$ is more underestimated for the largest values of h (recall that we have always $\mathfrak{M}_{n,m} \leq M_{n,m}$). However, it seems that $\mathfrak{M}_{n,m}$ is a near ideal fit to the exact local score $M_{n,m}$.

On the other hand, $NMSE_h$ is computed for the approximate local score proposed by Zhang (1995), $\mathcal{M}_{n,m}$. Table 2 indicates that $\mathcal{M}_{n,m}$ has a behavior similar to $\mathfrak{M}_{n,m}$. However, $\mathfrak{M}_{n,m}$ outperforms Zhang's one, $\mathcal{M}_{n,m}$, for all h and this is consistent with the fact that $\mathcal{M}_{n,m} \leq \mathfrak{M}_{n,m} \leq M_{n,m}$.

Note that, as we are not concerned by asymptotics, our work takes place out of this framework by considering short sequences as illustrated by the simulations in which $\max(n,m) < 850$. This can explain the weakness of $\mathcal{M}_{n,m}$ which has been developed for large values of n, m and h.

3.3 The influence of h on the p-value for different sequence lengths

To compare $p_h(b_a)$, defined in (9), and the empirical p-value, $p_e(a)$, $\chi^2(\overline{p}_h)$ is calculated for three different values of $h = 2, 3$ and 4, with $p_e(a)$ defined in (10), but computed for the exact gapped local score $M_{n,m}$ and

$$\overline{p}_h(a) = \frac{\sum_{C_a} p_h(b_a)}{\#C_a} \tag{15}$$

with $C_a = \{(\mathbb{A}_k, \mathbb{B}_k) : M^k_{n,m} = a\}$ where $\{(\mathbb{A}_k, \mathbb{B}_k)\}$ are the kth pair of sequences of generated database as in the previous subsection. As it leads to

Table 3: INFLUENCE OF h ON THE NEW P-VALUE

	$\chi^2(\overline{p}_h)$			$PSE(\overline{p}_h)$		
h \ length category	short	medium	long	short	medium	long
2	**99.30**	**581.61**	441.93	**0.09**	**0.30**	**0.26**
3	112.25	597.07	431.44	0.10	0.31	0.27
4	164.34	654.30	**360.91**	0.11	0.33	0.28

Average of $\chi^2(\overline{p}_h)$ and PSE($\overline{p}_h(a)$) defined in (11) and (13), respectively. These results are calculated for gapped case over all databases in each category. The categories "short", "medium" and "long", respectively, are applied for the pairs of sequences whose lengths satisfy $\max(n,m) < 100$, $100 \leq n,m \leq 500$ and $\min(n,m) > 500$.

stable results for the empirical p-value, we consider a database of 10,000 pairs of sequences. In addition, to verify the stability of χ^2, 10 databases of 10,000 sequence pairs have been simulated. Value of $\chi^2(\overline{p}_h)$ is computed for each one of 10 databases corresponding to a sequence length. We define three categories "short", "medium" and "long" in the same way as subsection 3.2. Table 3 shows the average of $\chi^2(\overline{p}_h)$ values calculated over all sequence lengths (each length in 10 databases) in each of the three categories of sequence length. In average, the minimum of χ^2 is reached for $h = 2$ for the short and medium sequences. For the longest sequences, $h = 4$ gives the minimum error, in other words, the more accurate p-values, but the corresponding running time is substantially longer.

The coefficient of variation (standard deviation/mean) has also been computed over 10 values of χ^2 corresponding to 10 databases of each sequence length. Then, for each category, the mean of coefficients of variation has been calculated over all sequence lengths involved in the category. In average, for the longest sequences, the smallest value is equal to 0.11 and it is realized for $h = 4$, while for the medium sequences, it is equal to 0.29 (respectively 0.33 and 0.44) for $h = 3$ (resp. 2 and 4). For the short sequences, these values are smaller than 0.16 and $h = 2$ gives the minimum value 0.14.

We also apply the PSE metric and the results are given in Table 3. They confirm the ones obtained by the χ^2's, except for the longest sequence for which the smallest value of the mean of PSE is obtained for $h = 2$. However, the values of PSE are very close, so these results do not really contradict the χ^2 ones.

These simulations seem to indicate that finally, the choice of h weakly depends on the lengths of sequences and that, even when $h = 4$ returns the optimal value, $h = 2$ is still a feasible solution. In the following, we will nevertheless work with the optimal choice, that is, $h = 2$ or 4 according to the sequence lengths.

Note, moreover, that for sequences with $n \ll m$, simulations (not reported here) seem to emphasize that the smallest length plays the main role.

3.4 Comparison between the new approximate p-value and BLAST on simulated database

The previous sections were devoted to the comparison of two sequences. Now, in order to be more realistic and to compare our approach to a classical software, BLAST, we aim at comparing a sequence to a database. Then, we compare the approximate p-value, p_h, to the one of BLAST, denoted p_B (this latter is obtained by a heuristics based on extreme-value distribution (Altschul et al., 1997)). We measure the differences between the above theoretical p-values and the empirical one on a simulated database.

We consider 10 different sequence lengths (50, 80, 104, 206, 305, 370, 480, 550, 630 and 820) and, for each given sequence length, we simulate a database of 100 sequences with the same distribution of letter as in Subsection 3.2. Let $\mathcal{D}^l$, for $l = 1$ to 10, be the database corresponding to each length mentioned above and let $\mathcal{D} = \bigcup_{l=1}^{10} \mathcal{D}^l$, the essential database of our study. Then, $\mathcal{D}$ involves 1,000 independent sequences built with 10 different lengths. Also, six queries $\{\mathbb{Q}_j\}_{j=1,\ldots,6}$ of different lengths 82, 150, 307, 485, 638 and 827 are independently generated with the same distribution as the database $\mathcal{D}$.

On the one hand, the approximate and BLAST p-values, p_h and p_B respectively, are computed by aligning the queries to the database $\mathcal{D}$. For $i = 1$ to 100, $l = 1$ to 10 and $j = 1$ to 6, this alignment leads us to compute a_{ilj} and $b_{a_{ilj}}$, the observed values of $M_{n,m}$ and $\mathfrak{M}_{n,m}$ of the the jth query when it is compared with the ith sequence of the database $\mathcal{D}^l$. Our p-value, p_h, is computed by using (9). Note that according to the discussion of the previous subsection, we choose $h = 2$ when the length of the shortest sequence of comparison is less than 500 and $h = 4$ otherwise. For BLAST, the p-value, p_B, is approximately obtained as the e-value, given by BLAST, multiplied by n/N, where n is the length of the database sequence and N is the edge-corrected cumulative length of the database sequences, as reported in the program output (Altschul and Gish, 1996; Altschul et al., 2001; Webber and Barton, 2003). The version of BLAST program that we use is BLAST 2.2.13 and it can be found at the following address: `www.infobiogen.fr/services/analyseq/cgi-bin/blast2_in.pl`. Note that all the parameters of local score computation, scoring scheme and gap penalty, are identical to the previous subsection.

On the other hand, to obtain the empirical p-value, for each $l = 1$ to 10, a database of 10,000 sequences, denoted by $D^l = \{\mathbb{A}_k^l\}_{k=1,\ldots,10000}$, is generated with

Table 4: COMPARISON BETWEEN THE NEW P-VALUE p_h AND THE BLAST ONE p_B

	χ^2_j		Coefficient of determination	
Query length	p_h	p_B	R^2_h	R^2_B
82	**39.98**	121.74	**0.77**	0.05
150	**72.64**	171.34	**0.61**	0.15
307	**55.54**	307.94	**0.64**	0.35
485	**105.46**	193.96	**0.71**	0.35
638	**66.84**	92.57	**0.61**	0.56
827	176.64	**129.37**	**0.59**	0.43

$\chi^2_j(p_h)$, $\chi^2_j(p_B)$ and the coefficient of determination for aligning different queries to the database where p_h is the new approximate p-value defined in (9) and p_B is the p-value obtained by BLAST.

the same distribution as $\mathcal{D}^l$. The above queries $\{\mathbb{Q}_j\}_{j=1,\ldots,6}$, are aligned to each database D^l in order to calculate the empirical p-values. Thus, from the comparison of the jth query, $\mathbb{Q}_j$, with the database D^l, the empirical p-value $p_e(a)$ is defined as follows

$$p_e^{lj}(a) = \frac{N_a}{10000} = \frac{\#\{(\mathbb{A}_k^l, \mathbb{Q}_j) : M_{n,m} \geq a\}}{10000} \tag{16}$$

where $\mathbb{A}_k^l$ is the kth sequence of D^l and $M_{n,m}$ is the local score of $\mathbb{A}_k^l$ and $\mathbb{Q}_j$.

For each query $\mathbb{Q}_j$, $j = 1$ to 6, the accuracy of our p-value, p_h, is measured by χ^2_j defined as follows

$$\chi^2_j(p_h) = \sum_{l=1}^{10} \sum_{i=1}^{100} \frac{[p_e^{lj}(a_{ilj}) - p_h(b_{a_{ilj}})]^2}{p_e^{lj}(a_{ilj})}. \tag{17}$$

The accuracy of the p-value obtained by BLAST, p_B, is calculated in the same way by substituting $p_h(b_{a_{ilj}})$ with $p_B(a_{ilj})$.

This measure is used to compare the new approximate p-value p_h to the BLAST one p_B and the results are given in Table 4. For each query, p_h outperforms p_B, except for the longest length 827. The behavior of χ^2 for this latter query is probably due to the sensitivity of χ^2 when the empirical p-value is much smaller than the approximate one. Indeed, for this query, we only have one point such that the p-value p_h is much larger than p_e and which leads to a value of χ^2 substantially larger than the BLAST one: removing this sequence leads to $\chi^2_6(p_h) = 130.22$ that is a value close to the one of BLAST $\chi^2_6(p_B) = 129.36$. Moreover, if we focus

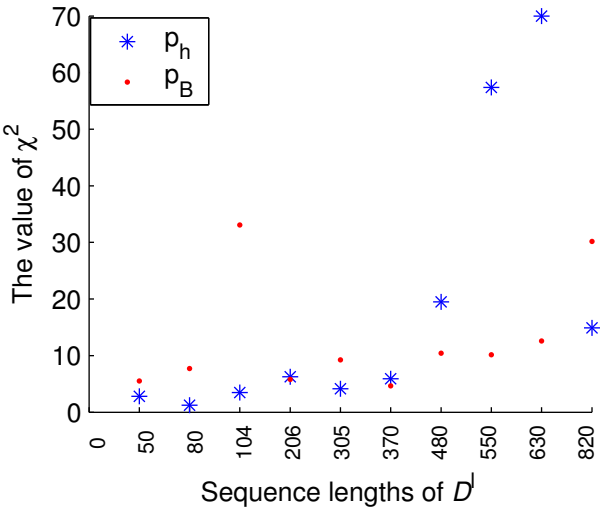

Figure 4: Behavior of $\chi^2(p_h)$ and $\chi^2(p_B)$ according to $\mathcal{D}^l$'s, $l = 1, \ldots, 10$, which present the databases with the different sequence lengths. This figure corresponds to the longest query, $\mathbb{Q}_6$, with the length equal to 827.

on the behavior of the query with length 827 through each term l of the sum (17), for $l = 1$ to 10, we find interesting results displayed in Figure 4. It emphasizes two cases: the databases with sequence lengths 550 and 630, which lead to the large value of $\chi^2_6(p_h)$. Globally, for the longest query where BLAST is known to be relevant, our p-value and the BLAST one have a similar behavior. For shorter sequences, the p-value p_h, appears more accurate than the BLAST one, p_B.

Here, PSE is not an appropriate criterion to compare our approach with BLAST. Indeed, for BLAST p-value, measure of PSE is not meaningful since the regression of $\log(p_B)$ over $\log(p_e)$ does not satisfy a linear model as shown by the coefficient of determination given in Table 4.

3.5 New approximate p-value and the BLAST one obtained on real sequences

In this section, we compare the accuracy of p_h and p_B on real sequences, using the **S**tructural **C**lassification of **P**roteins (**SCOP**) (Murzin et al., 1995) version 1.53. Sequences are selected using the Astral database (Brenner et al., 2000), removing similar sequences using an e-value threshold of 10^{-25}. This procedure yields 4352

Table 5: COMPARISON BETWEEN THE P-VALUE p_h AND BLAST ONE p_B ON THE REAL DATABASE SCOP 1.53

Query Length	48	80	103	175	255	309
$\chi^2_j(p_h)$	**41.34**	**84.56**	**125.18**	**52.04**	**141.79**	**66.02**
$\chi^2_j(p_B)$	57.74	161.83	163.44	213.62	171.35	273.45

χ^2_j values of the p-values p_h defined in (9) and the one given by BLAST, p_B, relative to different queries.

distinct sequences, $D = \{\mathbb{A}_k\}_{k=1,\ldots,4352}$, grouped into families and superfamilies [1].

Six sequences of different lengths are randomly selected from the above database D defining the queries $\{\mathbb{Q}_j\}_{j=1,\ldots,6}$. They are successively compared with the remaining sequences $D \setminus \{\mathbb{Q}_j\}$. Let a_{kj} be the observed exact local score corresponding to $(\mathbb{Q}_j, \mathbb{A}_k)$, $j = 1$ to 6 and $k = 1$ to 4351. These local scores are obtained over the scoring scheme BLOSUM62 and the penalty of gap opening and extension -11 and -1, respectively. The p-value p_h is calculated from (9) and p_B from BLAST 2.2.15 [2]. To find the empirical p-value, for each sequences pair $(\mathbb{Q}_j, \mathbb{A}_k)$, $j = 1$ to 6 and $k = 1$ to 4351, a database $\{(\mathbb{Q}^l_j, \mathbb{A}^l_k)\}_{l=1,\ldots,10000}$ (with the same lengths as $(\mathbb{Q}_j, \mathbb{A}_k)$) is independently generated from the letter empirical distribution of the database D. The empirical p-value $p_e(a)$ is computed as

$$p_e(a) = \frac{N_a}{10000} = \frac{\#\{(\mathbb{Q}^l_j, \mathbb{A}^l_k) : M_{n,m} \geq a\}}{10000} \tag{18}$$

where $M_{n,m}$ is the local score of $(\mathbb{Q}^l_j, \mathbb{A}^l_k)$.

Similarly to the latter subsection, only the χ^2 measure is considered since we have verified (by computing the coefficient of determination R^2) that PSE is not appropriate for both methods. The χ^2 is defined as follows

$$\chi^2_j(p_h) = \sum_{k=1}^{d} \frac{[p_e(a_{jk}) - p_h(b_{a_{jk}})]^2}{p_e(a_{jk})} \tag{19}$$

where d is the number of sequences for which BLAST calculates their e-value. We set the expectation value of BLAST program to 4000 and we get $d \simeq 500$ which varies according to different queries. The χ^2 values for BLAST are obtained by replacing $p_h(b_{a_{jk}})$ with $p_B(a_{jk})$.

[1] Available from `www.cs.columbia.edu/compbio/svm-pairwise`.

[2] The latter address, mentioned in Subsection 3.4, is no more accessible and we then use the version available from `www.ncbi.nlm.nih.gov/BLAST/download.shtml`.

Table 5 shows the values of χ^2 for the two p-values p_h and p_B. As it is seen, the χ^2 values of p_h are smaller than the BLAST ones which explains the accuracy of the new approximate p-value p_h. In addition, the variation of $\chi^2(p_B)$'s is larger than $\chi^2(p_h)$'s. Note that these values and the ones of Table 4 for the sequences of lengths 82, 150 and 307 are nearby. Then, our method appears as a relevant solution to compute the statistical significance in order to compare a query to a database in practice.

4 Conclusion and Outlook

In this work, we have introduced a new method for comparing sequences in the gapped case, based on the statistical significance of gapped alignments. It relies on h-tuples and h is a crucial parameter that has to be selected in practice. While the theoretical issue is difficult to address, we have carried out some numerical studies to outline some tracks to help in the choice of h. The conclusion is that the p-values are not significantly influenced by the value of h. However, $h = 2$ seems to be particularly appealing for short sequences while larger h's have to be considered for large sequences. In any case, our method yields appealing results. Nevertheless, for the large values of h ($h \geq 4$), the computation time can be long (this explains the limitation of our study in this case) and we guess that the choice of a practitioner will be motivated by a balance between accuracy and computation time.

But, our work focuses on short and medium sequences ($n, m \leq 500$). The comparison with BLAST method, in this case, on both simulated and real database, shows that our approach is an appropriate alternative to the p-value of BLAST: while our method is more accurate than BLAST method, we achieve a comparable computation time (recall that we use $h = 2$ in this case). This latter point is improved by using the direct p-value algorithm presented by Nuel (2006), which computes the exact value of p_h. Then, the proposed method can be considered as a relevant solution to sequence comparison problems.

The next step will be to classify sequences into predefined families by using p-values: a small p-value is related to an exceptional similarity and reveals the closeness to the family candidate. It offers the opportunity to compare the classification derived from both our approach and the BLAST one. The quality of classification is evaluated, for instance, via the ROC score (Gribskov and Robinson, 1996) which is based on the true and false positive rates.

References

Altschul, S., Bundschuh, R., Olsen, R. and Hwa, T. (2001), 'The estimation of statistical parameters for local alignment score distributions.', *Nucleic Acids Res.* **29**, 351–361.

Altschul, S. and Gish, W. (1996), 'Local alignment statistics.', *Methods Enzymol* **266**, 460–480.

Altschul, S., Stefan, F., Madden, T., Schäffer, A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D. (1997), 'Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.', *Nucleic Acids Res.* **25**, 3389–3402.

Bailey, T. and Gribskov, M. (2002), 'Estimating and evaluating the statistics of gapped local alignment scores.', *J. Comp. Biol.* **9**, 575–593.

Brenner, S., Koehl, P. and Levitt, M. (2000), 'The ASTRAL compendium for sequence and structure analysis.', *Nucleic Acids Res.* **28**, 254–256.

Bundschuh, R. (2002), 'Rapid significance estimation in local sequence alignment with gaps.', *J. Comp. Biol.* **9**, 243–260.

Chia, N. and Bundschuh, R. (2006), 'A practical approach to significance assessment in alignment with gaps.', *J. Comp. Biol.* **13**, 429–441.

Dembo, A., Karlin, S. and Zeitouni, O. (1994), 'Limit distribution of maximal non-aligned two-sequences segmental score.', *Ann. Prob.* **24**, 2022–2039.

Gribskov, M. and Robinson, N. (1996), 'Use of Receiver Operating Characteristc (ROC) analysis to evaluate sequence matching.', *Computers and Chemistry* **20**, 25–33.

Grossmann, S. and Yakir, B. (2004), 'Large deviations for global maxima of independent superadditive processes with negative drift and an application to optimal sequence alignments.', *Bernoulli* **10**, 829–845.

Karlin, S. and Altschul, S. (1990), 'Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes.', *Proc. Natl. Acad. Sci. USA* **87**, 2264–2268.

Mercier, S. and Daudin, J. (2001), 'Exact distribution for the local score of one i.i.d. random sequence.', *J. Comp. Biol.* **8**, 373–380.

Mitrophanov, A. and Borodovsky, M. (2006), 'Statistical significance in biological sequence analysis.', *Briefings in Bioinformatics* **7**, 2–24.

Mott, R. (1992), 'Maximum-likelihood estimation of the statistical distribution of Smith-Waterman local sequence similarity scores.', *Bull. Math. Biol.* **54**, 59–75.

Mott, R. and Tribe, R. (1999), 'Approximate statistics of gapped alignments.', *J. Comp. Biol.* **6**, 91–112.

Murzin, A., Brenner, S., Hubbard, T. and Chothia, C. (1995), 'SCOP: A structural classification of proteins database for the investigation of sequences and structures.', *J. Comp. Biol.* **247**, 536–540.

Nuel, G. (2006), 'Effective p-value computations using Finite Markov Chain Imbedding (FMCI): application to local score and to pattern statistics.', *Algo. Mol. Biol.* **1**, 5.

Park, Y., Sheetlin, S. and Spouge, J. (2005), 'Accelerated convergence and robust asymptitic regression of the gumbel scale parameter for gapped sequence alignment.', *J. of Physics A:MATHEMATICAL AND GENERAL* **38**, 97–108.

Park, Y. and Spouge, J. (2002), 'The correlation error and finite-size correction in an ungapped sequence alignment.', *Bioinformatics* **18**, 1236–1242.

Sheetlin, S., Park, Y. and Spouge, J. (2005), 'The Gumbel pre-factor k for gapped local alignment can be estimated from simulations of global alignment.', *Nucleic Acids Res.* **33**, 4987–4994.

Siegmund, D. and Yakir, B. (2000), 'Approximate p-values for local sequence alignments.', *Ann. Stat.* **28**, 657–680.

Siegmund, D. and Yakir, B. (2003), 'Correction: Approximate p-values for local sequence alignments.', *Ann. Stat.* **31**, 1027–1031.

Spang, R. and Vingron, M. (1998), 'Statistics of large-scale sequence searching.', *Bioinformatics* **14**, 279–284.

Storey, J. and Siegmund, D. (2001), 'Approximated p-value for local sequence alignments: numerical studies.', *J. Comp. Biol.* **8**, 549–556.

Waterman, M. (2000), *Introduction to computational biology.*, Chapman & Hall.

Waterman, M. and Vingron, M. (1994), 'Sequence comparison significance and poisson approximation.', *Stat. Sci.* **9**, 367–381.

Webber, C. and Barton, J. (2003), 'Increased coverage obtained by combination of methods for protein sequence database searching.', *Bioinformatics* **19**, 1397–1403.

Zhang, Y. (1995), 'A limit theorem for matching random sequences allowing deletions.', *Ann. Appl. Prob.* **5**, 1236–1240.