HAL
open science

New approximate P-value of gapped local sequence alignments,

Afshin Fayyaz Movaghar, Sabine Mercier, Louis Ferré, Claudie Chabriac

To cite this version:

Afshin Fayyaz Movaghar, Sabine Mercier, Louis Ferré, Claudie Chabriac. New approximate P-value of gapped local sequence alignments,. Comptes rendus de l'Académie des sciences. Série I, Mathématique, 2007, 346, pp.87-92. hal-00937496

HAL Id: hal-00937496

https://hal.science/hal-00937496v1

Submitted on 28 Jan 2014

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

New Approximate P-value of Gapped Local Sequence Alignments

Afshin M. FAYYAZ [a], Sabine MERCIER [a] Louis FERRE [a] Claudie HASSENFORDER [a]

[a] *Institute of mathematics of Toulouse, University Toulouse Le Mirail, 31058 Toulouse cedex 9, France*

Received *****; accepted after revision +++++

Presented by £££££

Abstract

We propose a new method to approximate the significativity of gapped local sequence alignments. We focus on short sequences for which standard methods are known to be less accurate since they have been developed under asymptotics. Our approach combines an approximate distribution of ungapped local score of two sequences and a special scoring scheme that allows the insertion of gaps. For a positive integer h, the scoring scheme is defined on h-tuples of the components of the sequences and corresponds to the gapped global score. The influence of h and the accuracy of the p-value are numerically studied. *To cite this article: A. Name1, A. Name2, C. R. Acad. Sci. Paris, Ser. I 340 (2005).*

Résumé

Nouvelle P-valeur Approchée d'Alignements Locaux de Séquences avec Gaps.

Nous proposons une nouvelle méthode pour estimer la signification statistique des alignements locaux de deux séquences avec gaps. On s'intéresse plus particulièrement aux séquences courtes pour lesquelles les méthodes standards sont moins efficaces étant donné leur aspect asymptotique. Notre approche combine une distribution approchée du score local sans gaps de deux séquences et une fonction de score spécifique qui permet d'introduire les gaps. Soit h un entier positif, la fonction de score est définie sur les h-uplets des composants des séquences et correspond au score global avec gap. L'influence de h et la qualité de la p-valeur sont ensuite étudiées numériquement.*Pour citer cet article : A. Name1, A. Name2, C. R. Acad. Sci. Paris, Ser. I 340 (2005).*

Email addresses: `fayyaz@univ-tlse2.fr` (Afshin M. FAYYAZ), `mercier@univ-tlse2.fr` (Sabine MERCIER), `loferre@univ-tlse2.fr` (Louis FERRE), `chabriac@univ-tlse2.fr` (Claudie HASSENFORDER).

Version française abrégée

1. Introduction

L'un des grands objectifs de la génomique consiste à trouver des régions de similarité communes à deux séquences biologiques pouvant être issues d'un ancêtre commun. On utilise pour cela les alignements locaux qui prennent en compte les mutations et éventuellement les insertions et/ou les délétions de composants de séquences, appelées **indels** ou **gaps**. Les scores d'alignements locaux sont usuellement calculés en sommant les scores de substitutions, s, pour chaque paire de composants (ou lettres alignés) et une pénalité pour chaque gap, δ. Le plus grand score d'alignements locaux est appelé **score local**.

Soient $\mathbb{A} = A_1, A_2, ..., A_n$ et $\mathbb{B} = B_1, B_2, ..., B_m$ deux séquences indépendantes de variables i.i.d. qui prennent leurs valeurs dans un alphabet fini. Soient I et J deux régions contiguës telles que $I \subset \{A_1, A_2, ..., A_n\}$ et $J \subset \{B_1, B_2, ..., B_m\}$ et de longueurs respectives i et j. Le score local de Smith-Waterman (SW) qui inclut les gaps est défini, par exemple dans [10], par $M_{n,m} = \max_{I,J} S(I, J)$ où

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

est le score global avec gaps de I et J et où le maximum est pris sur tous les alignements, chacun étant défini par une suite croissante $u(\cdot)$ et $v(\cdot)$ des indices des ℓ paires de lettres alignées.

L'objectif est de déterminer si un score observé est ou non le fruit du hasard sous un modèle approprié de séquences aléatoires.

Dans le cas sans gap, la distribution asymptotique du score local (avec décalages) est donnée par une *distribution des valeurs extrêmes* [2], quand $E[s(A, B)] < 0$. Le résultat est également mis en évidence de manière empirique dans le cas du score local avec gaps, $M_{n,m}$ (*cf., e.g.*, [1]). Récemment, une p-valeur asymptotique a été obtenue sous certaines conditions sévères sur les gaps ([8], [9]).

Les méthodes usuelles reposent sur une approche asymptotique et ne donnent pas de résultats pertinents pour des séquences courtes. Nous présentons ici une nouvelle p-valeur approchée d'alignements avec gaps. Notre travail repose sur la p-valeur du score local sans gaps [5] et la prise en compte des gaps par le biais de la fonction de scores. Dans la Section 2, nous proposons la nouvelle p-valeur dont nous étudions ensuite numériquement les performances.

2. P-valeur approchée d'alignements avec gaps

Afin d'étudier une p-valeur approchée du score local usuel de SW, nous commençons par définir un nouveau score local qui prend en compte les gaps (mais d'une manière détournée) et dont nous donnons, par la suite, une p-valeur approchée.

Soient A_i^α et B_i^α, respectivement, les $i^{\text{ème}}$ lettres des segments en vis-à-vis de $\mathbb{A}$ et $\mathbb{B}$, dans un décalage fixé α. Le nouveau score local avec gap est construit sur des blocs de h-uplets des séquences initiales. Pour un nombre entier positif donné h, soient $\mathbb{X}^\alpha = \{X_i^\alpha\}$ et $\mathbb{Y}^\alpha = \{Y_j^\alpha\}$ deux séquences indépendantes où $X_i^\alpha = (A^\alpha_{(i-1)h+1}, ..., A^\alpha_{ih})$ et $Y_j^\alpha = (B^\alpha_{(j-1)h+1}, ..., B^\alpha_{jh})$ avec $i, j = 1, \ldots, N_r$, $r = r(\alpha)$ étant la longueur de la partie vis-à-vis des deux séquences pour un décalage fixé α et $N_r = [\frac{r}{h}]$ (*cf.* Figure 1). Soit $H_{N_r} = \max_{1 \le i \le j \le N_r} \sum_{k=i}^{j} S(X_k^\alpha, Y_k^\alpha)$, le score local sans gaps appliqué à la séquence des h-uplets $(X_i^\alpha, Y_i^\alpha)_{1 \le i \le N_r}$ avec, pour fonction de scores S, définie en (1). Nous proposons comme estimation du score local $M_{n,m}$, le score local définie par $\mathfrak{M}^h_{n,m} = \max_\alpha H_{N_r}$. Ce score local bien qu'étant à la base sans gap, prend toutefois en compte les indels par l'intermédiaire de la fonction de scores, S. Cette manière d'introduire les gaps, nous permet d'utiliser le résultat de [5] pour étudier la p-valeur de $\mathfrak{M}^h_{n,m}$ en l'adaptatant aux cas des séquences des h-uplets.

Alors, pour $m \ge n$, nous avons :

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

où $K = [\frac{n-1}{h}]h$ et b_a est la valeur observée de $\mathfrak{M}^h_{n,m}$. La probabilité exacte de H_{N_r} découle de Mercier et Daudin [5] : elle repose sur la puissance N_r (la longueur de la séquence couple $(\mathbb{X}^\alpha, \mathbb{Y}^\alpha)$) d'une matrice carrée de taille $(b_a + 1)$ construite à partir de la distribution de la nouvelle fonction de scores S.

Remarques :

* L'étude de la pertinence de $\mathfrak{M}^h_{n,m}$ pour l'estimation de $M_{n,m}$ est effectuée dans [3].
* La p-valeur approchée de deux séquences proposée dans [5] s'appuie sur l'hypothèse de décalages indépendants. Cependant, la dépendance entre décalages existe, en particulier pour des décalages proches, mais elle n'est usuellement pas considérée. Des simulations (*cf.* [3], [7]) mettent en évidence que cette hypothèse ne joue pas un rôle essentiel dans le calcul de la p-valeur. C'est la raison pour laquelle nous utiliserons également cette hypothèse ici.

La p-valeur approchée de $\mathfrak{M}^h_{n,m}$ est alors utilisée pour estimer celle du score local avec gaps, $M_{n,m}$. L'étude de l'influence de h sur la qualité de la p-valeur proposée montre que, pour les séquences courtes, $h = 2$ donne les résultats satisfaisants (*cf.* Table 1). Par ailleurs, une comparaison avec BLAST (en utilisant comme référence une p-valeur empirique calculée sur 10000 couples de séquences) révèle que nos résultats surpassent ceux de BLAST pour les séquences courtes (*cf.* Table 2).

3. Conclusion

L'approximation de la p-valeur du score local avec gaps que nous proposons diffère de celles fréquemment utilisées et donne des résultats numériques satisfaisants pour résoudre le problème de la signification statistique du score local avec gaps dans le cas des séquences courtes.

---

1. Introduction

One of the broad objectives of genomic is to find the common regions of similarity of two protein or DNA sequences that can be the consequence of a common ancestor. For this, we use the local alignments which take into account the mutations in certain regions. Gaps may be added to either segment of sequences to represent insertions or deletions. Local alignment quality is generally measured by a score calculated by adding substitution scores, s, for each aligned pair of letters and gap penalty, δ, for each insertion and deletion. Any local alignment of sequences can be scored and ranked according to this scoring scheme. The maximum-scoring local alignment is called **local score**.

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 $\mathcal{A}$. We suppose $m \geq n$ without loss of generality. Let $M_{n,m}$ be the usual local score with gaps *i.e.* Smith-Waterman score: the optimal aligned 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, we have (see, *e.g.*, [10]) $M_{n,m} = \max_{I,J} S(I, J)$ where

$$(3) \qquad S(I, J) = \max\{-\delta(i - \ell + j - \ell) + \sum_{k=1}^{\ell} s(A_{u(k)}, B_{v(k)})\}$$

and the maximum is taken over all alignments, given by two increasing sequences $u(\cdot)$ and $v(\cdot)$ where ℓ is the number of the pairs of aligned letters and i, j are the lengths of I and J, respectively. Note that, S is the gapped global score of I and J.

We want to know whether an observed score could have arisen by chance under an appropriate model of sequence randomness. In other words, how high may a score be expected to occur purely by chance?

The distribution of ungapped local score with shift has been asymptotically derived as an extreme-value distribution [2], when $E[s(A,B)] < 0$. Whereas for the gapped local score, $M_{n,m}$, there is a great deal of empirical evidence (see, *e.g.*, [1]) indicating that the extreme-value theory underlying the ungapped case carries over, provided the gap penalties are severe enough. Another heuristic approximate p-value was proposed using Greedy Extension Method [6]. Recently, an approximate p-value was theoretically derived under certain severe conditions on gaps ([8], [9]). These methods rely on asymptotics whereas we are concerned by short sequences in this note.

For instance, BLAST (which is based on [2]) is not enough accurate for short sequences. So, we propose hereafter a new approximate p-value of the gapped local score for short sequences. It stands on the p-value of the ungapped case in [5] and an approximate gapped local score based on S in (3) by working with h-tuples rather than letters. In Section 2, new gapped local score is defined and the approximate p-value is derived. We mainly focus on applications (the theoretical aspects are out of the scope of this paper), then, after having numerically studied on the effect of h, we compare on simulations our p-value with the one obtained by BLAST, in Section 3.

2. Approximate p-value of gapped alignment

The new p-value stands on the ungapped p-value of [5] and a new approximate gapped local score based on an adaptation of the proposition of Zhang (see [10]). Les us first define this new local score.

Let A_i^α (respectively B_i^α) be the ith letter of the segment of $\mathbb{A}$ (resp. $\mathbb{B}$) which participates in the fixed shift α. Given a positive integer h, let $r = r(\alpha)$ the length of the opposite parts of the two sequences for a fixed shift α, and $N_r = [\frac{r}{h}]$. Now, let $\mathbb{X}^\alpha = \{X_i^\alpha\}_{i=1,\dots,N_r}$ and $\mathbb{Y}^\alpha = \{Y_j^\alpha\}_{j=1,\dots,N_r}$ be two independent sequences of h-tuples in $\mathcal{A}^h$ with $X_i^\alpha = (A^\alpha_{(i-1)h+1}, \dots, A^\alpha_{ih})$ and $Y_j^\alpha = (B^\alpha_{(j-1)h+1}, \dots, B^\alpha_{jh})$ (Figure 1). It is clear that for a fixed α, the X_i^α's are independent and so are the Y_j^α's.

For the fixed shift α, the ungapped local score applied to the h-tuple sequence $(X_i^\alpha, Y_i^\alpha)_{1\le i\le N_r}$ is defined by

$$H_{N_r} = \max_{1\le i\le j\le N_r} \sum_{k=i}^{j} S(X_k^\alpha, Y_k^\alpha) \tag{4}$$

where the scoring scheme, S, is the gapped global score of (X_k^α, Y_k^α) defined in (3). Now, we approximate $M_{n,m}$ by

$$\mathfrak{M}^h_{n,m} = \max_\alpha H_{N_r}. \tag{5}$$

Note that, this local score is initially defined without gaps which are actually taken into account through the scoring function.

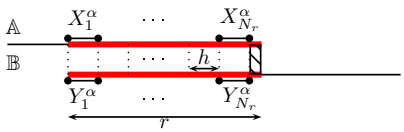

Figure 1. The h-tuples of a fixed shift α. (Les h-uplets d'un décalage fixé α.)

Now, we approximate the p-value of $\mathfrak{M}^h_{n,m}$ (and thus $P(M_{n,m} \ge a)$, one of the exact local score) by

$$p_h(b_a) = 1 - \left[\left(\prod_{r=h,2h,\dots,(K/h-1)h} P(H_{N_r} < b_a)^{2h}\right) P(H_{N_K} < b_a)^{2(n-K)}\right] P(H_{N_n} < b_a)^{m-n+1}, \tag{6}$$

where $K = [\frac{n-1}{h}]h$ and b_a is the observed value of $\mathfrak{M}^h_{n,m}$. Indeed, since H_{N_r} is the ungapped local score of two h-tuples, we can apply the results of [5] in the ungapped case and we have

$$p_h(b_a) = 1 - \left[\prod_{r=1,\dots,n-1} P(H_{N_r} < b_a)^2\right] P(H_{N_n} < b_a)^{m-n+1}.$$

But, the terms into brackets can be rewritten as follows

$$(\prod_{r=h,\dots,2h-1} P(H_1 < b_a)^2) \times \cdots \times (\prod_{r=K-h,\dots,K-1} P(H_{K/h-1} < b_a)^2) \times (\prod_{r=K,\dots,n-1} P(H_{K/h} < b_a)^2)$$

$$= P(H_1 < b_a)^{2h} \times \cdots \times P(H_{K/h-1} < b_a)^{2h} \times P(H_{K/h} < b_a)^{2(n-K)}$$

which yields (6).

Now, we can use the one sequence case [5] which is based on Markov chains theory, particulary Lindley process, to achieve the distribution of H_{N_r} :

$$P[H_{N_r} < b_a] = 1 - P_1 \Pi^{N_r} P'_{b_a+1} \tag{7}$$

where P_i is a vector $1 \times (b_a+1)$ whose ith element is one and zero elsewhere and the matrix $\Pi_{(b_a+1)\times(b_a+1)}$ is filled by using the distribution of S, the new scoring function. Note that, the distribution (7) is independent of the sign of the expected score of residues (unlike [4]).

Remarks:

* For all h, we clearly have, $\mathfrak{M}^h_{n,m} \leq M_{n,m}$. The quality of $\mathfrak{M}^h_{n,m}$ is studied in [3] and the results show that the accuracy of $\mathfrak{M}^h_{n,m}$ decreases with enhancement of h.
* The proposed p-value for ungapped case in [5] relies on the assumption of independent shifts. Of course, there is dependence among the shifts, particulary among close shifts, but, simulations (see [3], [7]) have emphasized relaxing this assumption does not affect the results in practice. For this reason, this assumption is also considered to build our p-value.

3. Numerical results

To compare the approximate p-value, p_h, and an empirical p-value, p_e, we independently generate 10000 pairs of sequences, $\{(\mathbb{A}_k, \mathbb{B}_k)\}$, from the letters frequencies of some Homo sapiens proteins of human. The scoring scheme is BLOSUM62 with the opening gap penalty -11 and its extension -2. Let $\overline{p}_h(a) = \left(\sum_{C_a} p_h(b_a)\right)/|C_a|$ with $C_a = \{(\mathbb{A}_k, \mathbb{B}_k) : \ M^k_{n,m} = a\}$ and $p_e(a) = |\bigcup_{t \geq a} C_t|/10000$. A measure, $\chi^2(\overline{p}_h) = \sum_{i=1}^{10000} (p_e(a) - \overline{p}_h(a))^2/p_e(a)$, is calculated for three different values of h : $h = 2, 3$ and 4. The results are given in Table 1. In average, $h = 2$ reaches the minimum χ^2 for the small and medium sequences for which we are interested. For the long sequences, $h = 4$ gives the minimum error, in other words, the more accurate p-values.

$h\backslash$ length	< 100	[100,500]	> 500
2	**99.30**	**581.61**	441.93
3	112.25	597.07	431.44
4	164.34	654.30	**360.91**

Table 1
Average of $\chi^2(\overline{p}_h)$ over generated databases in each category "< 100", "$[100, 500]$" and "> 500". (Moyenne de $\chi^2(\overline{p}_h)$ sur les bases de données dans chaque catégorie "< 100", "$[100, 500]$" et "> 500".)

In order to assess the validity of our approach in practice, we perform the comparison with BLAST. Let p_B be the BLAST p-value. Five generated queries $\{\mathbb{Q}_d\}$ are compared with a database involving 1000 independently generated sequences $\{\mathbb{A}_k\}$ and the p-values p_h and p_B are computed. Lengths of database sequences vary from 50 until 820. In order to compare the two methods, we compute the target p-value: the empirical p-value p_e. For each pairwise comparison $(\mathbb{Q}_d, \mathbb{A}_k)$ with the gapped local score a, we generate 10000 sequences of the same length of $\mathbb{A}_k$ and calculate $M_{n,m}$ by aligning $\mathbb{Q}_d$ to the sequences of the latter database. Now, $p_e(a)$ is the frequency of the sequences with $M_{n,m} \geq a$ divided in 10000. The values of $\chi^2(p_h)$ and $\chi^2(p_B)$ are given Table 2: they show that the p-value p_h outperforms the BLAST one in this short sequences framework. Moreover, Figure 2 illustrates the behaviors of p_h and p_B on an example: clearly, the values of p_h are more concentrated around the line $x = y$ than the ones of p_B and this highlights the superiority of our approach over BLAST in this context.

Query length	82	150	307	485	638
$\chi^2(p_h)$	**39.98**	**72.64**	**55.54**	**105.46**	**66.84**
$\chi^2(p_B)$	121.74	171.34	307.94	193.96	92.57

Table 2
$\chi_j^2(p_h)$ and $\chi_j^2(p_B)$ for aligning different queries to the database where p_h is the new approximate p-value defined in (6) and p_B is the p-value obtained by BLAST. ($\chi_j^2(p_h)$ et $\chi_j^2(p_B)$ pour alignement des requêtes différentes à la base de données où p_h est définie dans (6) et p_B est obtenue de BLAST.)

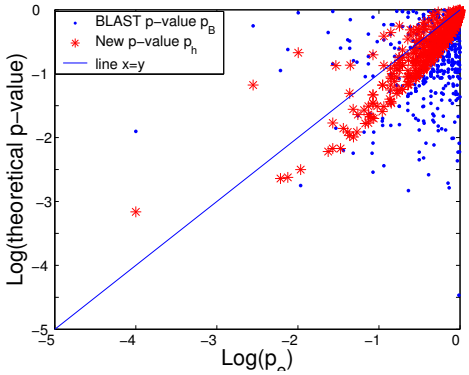

Figure 2. The log-log plot of the p-value p_h and the BLAST one p_B vs the empirical p-value for the query of length 82. (La p-valeur p_h et celle de BLAST en fonction de la p-valeur empirique pour la requête de longueur 82 en échelle logarithmique.)

4. Conclusion

The p-value proposed here is an appealing candidate to solve the problem of significativity of gapped local score for short sequence alignment. The p-value p_h outperforms the BLAST one, for all the simulations proposed here. These features are also confirmed by comparison on real sequences given in [3]. Concerning the choice of h, $h = 2$ yields in average a good estimates for the small and medium sequences, while $h = 4$ achieves the minimum error for long sequences. Note that, for long sequences, we have restricted our study to $h \leq 4$ because of the computation time required: then, for long sequences, BLAST has still the advantage of speed. At the contrary, for short sequences, our approach is more precise while requiring computation time comparable to BLAST.

References

[1] S.F. Altschul, W. Gish, Local alignment statistics, Methods Enzymol 266 (1996) 460-480.

[2] A. Dembo, S. Karlin, O. Zeitouni, Limit distribution of maximal non-aligned two-sequences segmental score, Ann. Prob. 24 (1994) 2022-2039.

[3] A. Fayyaz, S. Mercier and L. Ferré, h-tuple Approach to Evaluate Statistical Significance of Biological Sequence Comparison with Gaps, submitted for publication.

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

[5] S. Mercier, J.J. Daudin, Exact distribution for the local score of one i.i.d. random sequence, J. Comp. Biol. 8 (2001) 373-380.

[6] R.F. Mott, R. Tribe, Approximate statistics of gapped alignments, J. Comp. Biol. 6 (1999) 91-112.

[7] Y. Park, J.L. Spouge, The correlation error and finite-size correction in an ungapped sequence alignment, Bioinformatics. 18 (2002) 1236-1242.

[8] D. Siegmund, B. Yakir, Approximate p-values for local sequence alignments, Ann. Stat. 28 (2000) 657-680.

[9] D. Siegmund, B. Yakir, Correction: Approximate p-values for local sequence alignments, Ann. Stat. 31 (2003) 1027-1031.

[10] Y. Zhang, A limit theorem for matching random sequences allowing deletions, Ann. Appl. Prob. 5 (1995) 1236-1240.

Réponse des auteurs à la note CRAS pma06254 :

"New approximate P-value of gapped local sequence alignments"
by A. Fayyaz, S. Mercier, L. Ferré and C. Hassenforder

Nous tenons, avant toute chose, à remercier le référé pour ses commentaires judicieux qui nous ont permis, du moins le croyons nous, d'améliorer substantiellement notre note. En effet, comme suggéré par ses commentaires, nous avons totalement repris la réécriture de notre article afin de rendre celui plus compréhensible en mettant maintenant l'accent sur les éléments vraiment importants (ce qui ne semblait pas vraiment le cas dans la version initiale et était vraisemblablement responsable de confusions).

Ainsi, avons nous concentré notre propos sur le cas des alignements avec gaps et surtout, ce qui fait l'originalité de notre approche, sur le traitement des séquences courtes. Nous avons donc, dans un premier temps, détaillé la méthode que nous proposons en montrant très précisément comment notre p-valeur était obtenue. Dans la mesure où les p-valeurs sont notre centre d'intérêt, la place accordée antérieurement au score local $\mathfrak{M}_{nm}^h$ a été réduite. De la même façon, concernant l'hypothèse d'indépendance des shifts, nous avons supprimé nos simulations en les remplaçant par des références d'articles qui traitent de ce problème plus en détail (notons que cette question est indépendante de l'utilisation de $h-$uplets et que ceux-ci n'ont pas vocation à y répondre).

Puis, nous nous intéressons à l'influence du paramètre h au travers de simulations et, enfin, nous montrons, point nouveau de cette version, par le biais d'une comparaison numérique, que notre p-valeur donne des résultats plus précis que BLAST pour les courtes séquences. Il est également à noter que la taille des simulations a été portée de 1000 à 10 000.

Nous avons également pris en compte les remarques de détail à l'exception de la notation N_r qui ne nous a pas semblé problématique (nous pensons que, même si N_r dépend de h, lui rajouter cette lettre alourdirait la notation).