UNIVERSITE D'EVRY
VAL D'ESSONNE

Département Informatique

LaMI

Laboratoire de Méthodes Informatiques

Approximated asymptotic law of Z-value and applications

J.-N. Bacro and J.-P. Comet

e-mail : bacro@inapg.inra.fr, comet@lami.univ-evry.fr

Rapport de Recherche n^o 53-2000

Octobre 2000

Université d'Evry–Val d'Essonne
Cours Monseigneur Roméro
F–91025 Evry Cedex France

Approximated asymptotic law of Z-value and applications

J.-N. Bacro
INA PG, Dpt OMIP,
U.M.R. INAPG/INRA
16, rue Claude Bernard 75231 Paris cedex 05, France.
e-mail : bacro@inapg.inra.fr
tel : (33) 1.44.08.72.71 - fax : (33) 1.44.08.16.66

J.-P. Comet
LaMI,
Université d'Evry-Val d'Essone,
Cours Monseigneur Roméro, 91025 Evry Cedex, France.
e-mail : comet@lami.univ-evry.fr
tel : (33) 1.69.47.74.53 - fax : (33) 1.69.47.74.72

October 2000

Running title : Asymptotic law of Z-value

keywords: Dynamic programming sequence alignment, Significance, Z-value, Asymptotic distribution, Gumbel distribution.

Abstract

The Z-value is an attempt to estimate the statistical significance of a Smith & Waterman dynamic alignment score (H-score) through the use of a Monte-Carlo procedure.

In this paper, we give an approximation for the Z-value law deduced from the Poisson clumping heuristic developped by Waterman and Vingron (Waterman and Vingron, 1994) in the case of i.i.d. sequences comparison. As for non-gapped alignment scores, our approximation is of Gumbel type but with parameters which are sequence independent. This result makes clear the related experimental results mentionned by Comet et al. (Comet et al., 1999).

Using "quasi-real" *sequences (i.e. randomly shuffled sequences of the same length and amino acid composition as the real ones) we investigate the revelance of our approximation result. Since the Monte-Carlo approach we use generates a bias for the Gumbel decay parameter estimation, a correction procedure is proposed.*

Applications to real sequences are considered and we show how our results can be used to detect the potential biological relationships between real sequences.

Introduction

Sequences comparison has become a central notion in modern molecular biology. To evaluate the similarity between two sequences, a lot of indices are now available, allowing global alignments and gapped or ungapped local alignments. The Smith and Waterman algorithm (Smith and Waterman, 1981) answers exhaustivly to the question of the search of the alignments with the best score. Most of other approaches are based on heuristics. The Smith and Waterman algorithm finds the best local gapped alignments between two sequences, leading to an alignment score that can be used as a basis for determining a possible homology. The statistical significance of such a score, however, is a crucial problem. In this respect, two ways of research have been explored in the last years and may be briefly summarized as follows: the first one is based on known results concerning non-gapped alignments (Altschul et al., 1990), looking for possible extensions that mimic these results (Waterman and Vingron, 1994) or exhibiting relevant score approximation whose properties are related to the ungapped case (Mott and Tribe, 1999); the second one is based on simulation results, using a shuffling procedure and a particular statistics called Z-value (Lipman et al., 1984; Landes et al., 1992; Slonimski and Brouillet, 1993). In a recent paper Comet et al. (Comet et al., 1999) proposed an experimental study of the Z-value statistics. In particular, these authors surmised that the high Z-value distribution differs for randomly shuffled sequences and for real sequences respectively. In the first case, they showed that a Gumbel law fits well the data, but it seems that in the second case, the same law fits poorly. As a consequence the introduction of another extreme value distribution was suggested leading to a biological interpretation of the associated cutoff value (see Comet and al. (1999) for details). The aim of the present paper is to precise and to highlight these experimental results. For i.i.d random sequences, using the Waterman and Vingron approach (Waterman and Vingron, 1994), we first show that the asymptotic distribution of the Z-values can be approximated by a particular Gumbel law, with fixed parameters. For randomly shuffled sequences, we characterise the bias introduced by the shuffling method and we propose a correction procedure allowing to interpret the associated Z-value on the basis of the Waterman and Vingron approach. We show that the empirical data based on shuffled sequences fit well the proposed model.

In the case of real sequences, the Z-value asymptotic distribution appears to be of the same type as that for shuffled sequences (Gumbel law) but with other parameters. In other words, the Poisson clumping heuristic does not explain completely the observed distribution of Z-value for real sequences.

This article is organised as follows. The first section defines the Z-value variable, following Comet et al. (Comet et al., 1999). The second section is devoted to the asymptotic approximation for the Z-values distribution under the hypothesis of random sequences comparison. The third section focuses on testing the approximation law for shuffled sequences. A correction procedure for the parameters estimation is proposed in order to take into account the shuffling induced bias. This procedure is then applied to real sequences. Section fourth gives an overview of the advantages related to the Z-value approach.

1 The Z-value statistics

Let $\mathbf{X}$ and $\mathbf{Y}$ be two sequences and consider the corresponding maximum local alignment score $H(\mathbf{X}, \mathbf{Y})$ based on the Smith and Waterman algorithm (Smith and Waterman, 1981). We suppose here that the penality function for consecutive gaps has been well chosen in order to characterize aligning subsequences which have more similarity than random sequences. Such a kind of score is usually refered as score *with parameters in the logarithmic* *region* : for details see (Arratia and Waterman, 1994; Waterman and Vingron, 1994). In order to evaluate a p-value for the $(\mathbf{X},\mathbf{Y})$ comparison, we consider the corresponding Z-value variable :

$$Z(\mathbf{X},\mathbf{Y}) = \frac{H(\mathbf{X},\mathbf{Y}) - E(H(\mathbf{X},\mathbf{Y}))}{\sigma_{H(\mathbf{X},\mathbf{Y})}}$$

where $E(H(\mathbf{X},\mathbf{Y}))$ and $\sigma_{H(\mathbf{X},\mathbf{Y})}$ stand respectively for the expectation and the standard deviation of $H(\mathbf{X},\mathbf{Y})$.

2 Asymptotic approximation for the Z-value distribution

Suppose that $\mathbf{X} = X_1 \ldots X_n$ and $\mathbf{Y} = Y_1 \ldots Y_m$ are two random sequences where X_i and Y_j are independant and identically distributed. Waterman and Vingron (Waterman and Vingron, 1994) proposed a practical procedure to assign statistical significance for the $\mathbf{X}$ and $\mathbf{Y}$ comparison based on H, which can be summarized as follows : an approximated p-value for the $\mathbf{X}$ and $\mathbf{Y}$ comparison can be achieved using $1 - e^{-\gamma mnp^{H(\mathbf{X},\mathbf{Y})}}$ where γ and p are two parameters to be estimated.
The Waterman and Vingron result (Waterman and Vingron, 1994) is based on the approximation :

$$P\left(H(\mathbf{X},\mathbf{Y}) < t = \frac{\log nm}{|\log p|} + c\right) \simeq e^{-\gamma mnp^t} \tag{1}$$

which extends the Poisson approximation presented by Karlin and Altschul for general scoring scheme without indels (Karlin and Altschul, 1990).

The approximation (1) has been obtained as a result of the following two stages :

(a) A Poisson approximation for the optimal local score distribution using the Aldous clumping heuristic (Aldous, 1989) : for m and n sufficiently large

$$P\left(H(\mathbf{X},\mathbf{Y}) < t = \frac{\log mn}{|\log p|} + c\right) \simeq e^{-\alpha p^t} \tag{2}$$

where $\alpha \equiv \alpha(\mathbf{X},\mathbf{Y})$ and $p \equiv p(\mathbf{X},\mathbf{Y})$ are two positive parameters (this correspond to the assumptions (A1) and (A2) of the Waterman and Vingron approach).

(b) A normalization related to the different lengths of the sequences by setting $\alpha = \gamma mn$.

Now, from relation (2) we deduce that, for m and n sufficiently large

$$P\left(H(\mathbf{X},\mathbf{Y}) - \frac{\log nm}{|\log p|} < c\right) \simeq \exp\left(-\exp\left(-|\log p|\left(c + \frac{\log\left(\frac{mn}{\alpha}\right)}{|\log p|}\right)\right)\right)$$

which states that the distribution of $H(\mathbf{X},\mathbf{Y}) - \frac{\log nm}{|\log p|}$ can be approximated, for m and n sufficiently large, by a Gumbel distribution with parameters $\frac{-\log\left(\frac{mn}{\alpha}\right)}{|\log p|}$ and $\frac{1}{|\log p|}$, say

$$H(\mathbf{X},\mathbf{Y}) - \frac{\log nm}{|\log p|} \overset{\mathcal{D}}{\approx} G\left(-\frac{\log\left(\frac{mn}{\alpha}\right)}{|\log p|}, \frac{1}{|\log p|}\right).$$

Using well known results related to the Gumbel distribution we can deduce the two following approximations :

$$E(H(\mathbf{X},\mathbf{Y})) \simeq \frac{K+\log\alpha}{|\log p|} \tag{3}$$

where $K = 0,57721..$ denotes the Euler's constant and

$$\sigma^2_{H(\mathbf{X},\mathbf{Y})} \simeq \frac{\pi^2}{6(\log p)^2} \tag{4}$$

It is then straightforward to obtain an approximation for the law of the $Z(\mathbf{X},\mathbf{Y})$ variable : for m and n sufficiently large, and under assumption (a), we have :

$$\frac{H(\mathbf{X},\mathbf{Y}) - E(H(\mathbf{X},\mathbf{Y}))}{\sigma_{H(\mathbf{X},\mathbf{Y})}} \overset{\mathcal{D}}{\approx} \frac{\sqrt{6}|\log p|}{\pi} G(-\frac{K}{|\log p|}, \frac{1}{|\log p|}) \tag{5}$$

which can be stated as

$$\frac{\pi}{\sqrt{6}} Z(\mathbf{X},\mathbf{Y}) \overset{\mathcal{D}}{\approx} G(-K,1). \tag{6}$$

In other words our approximation is sequence independent : in (6), the approximation of the Z-value distribution does not depend on sequences length and composition. It is well known that such a property is not verified when dealing with the H-score (Comet, 1998, and references therein). While the length dependency of alignment scores has been extensively discussed in the literature (Arratia and Waterman, 1989; Arratia et al., 1986; Arratia et al., 1989; Arratia et al., 1990; Arratia and Waterman, 1994; Dembo and Karlin, 1991a; Dembo and Karlin, 1991b; Karlin and Altschul, 1990; Karlin et al., 1990; Karlin and Dembo, 1992; Goldstein and Waterman, 1992; Goldstein and Waterman, 1994; Waterman, 1994b; Waterman, 1994a; Waterman and Vingron, 1994), there are no results yet available concerning the sequence composition dependency. Note that the normalization described above, eq. (b), which is an attempt to take into account the different lengths of the considered sequences, seems to be poorly fitted in most of the practical situations (Waterman and Vingron, 1994), leading to conservative p-values. From these different facts, the Z-value is clearly of interest. But the difficulty now comes from a practical point of view : how can we obtain a direct evaluation of the Z-values ? The idea is to use a shuffling procedure as presented in Comet et al. (Comet et al., 1999) which seems to be well adapted to simulate random sequences with the same amino acid composition than the initial ones. We compute two different Z-values $\widehat{Z_1}(\mathbf{X},\mathbf{Y})$ and $\widehat{Z_2}(\mathbf{X},\mathbf{Y})$ by shuffling the first and second sequences respectively, and therefore choose the minimum to estimate $Z(\mathbf{X},\mathbf{Y})$, which correponds to a conservative approach.

Remember that the basic assumption here is that $\mathbf{X}$ and $\mathbf{Y}$ are both i.i.d. random sequences. The most natural way to test our approximation law would be to generate a lot of i.i.d. random sequences in order to work with. Since our approximation is obtained as a particular consequence of the well-known Waterman and Vingron result (Waterman and Vingron, (1994)), but under the only assumptions (A1) and (A2), it seems reasonable to think that our result would be validate for i.i.d. sequences comparison. From a practical point of view, the i.i.d. assumption is clearly unrealistic (and that is why only very small p-value are considered to characterize significant H-score values). But there are no theoritical results allowing to appreciate how robust is the Waterman and Vingron approach or how robust is our Gumbel approximation with regards to this i.i.d. assumption. Even if we know that a deviation from the Gumbel approximation is systematic for the Z-value when working on real sequences, we also may hope that the deviation remains still slight in the case of sequences which do not exhibit particular structure similarity, as for the i.i.d. case. A lack of robustness for our approximation result regarding to the i.i.d. assumption would clearly be a major drawback for practical applications. In order to appreciate the robustness of our result we decide to test our approximation on shuffling sequences build from real ones. Such sequences are not i.i.d. but do not exhibit any particular structure effect and do not represent any more any biological phenomenon. In a certain sense, such type of sequences allows to mimic properties related to alignment scores for i.i.d. sequences. For an easy implementation in practice, it is fundamental that our approximation remains still valid in the case of shuffling sequences comparisons because it ensures a possibility to build, from our approximation, a discrimination test between real sequences which present significant similarities from real sequences which do not have stronger similarities than i.i.d. ones. That is why we decide to test the validity of our approximation law on sequences deduced from real ones by shuffling. Since no biological links are present in these sequences, we clearly hope that our approximation fits well with the related Z-value observations.

In the sequel we consider two sets of sequences described in (Comet et al., 1999): the set of real sequences and the set of "quasi-real" sequences which designate sequences obtained by shuffling real ones. Appart from its amino-acid composition which corresponds to a real case, no particular structure is introduced in "quasi-real" sequences. "Quasi-real" sequences will be shuffled many times to evaluate the Z-value leading to a set of results for quasi-real sequence alignments. We shall see first that for such a set a direct application of our approximation leads to a bad fit. A correction procedure taking account the bias induced by the shuffling approach, will be proposed. Having then a good fit for such sequences closed to random sequences, we will apply the whole procedure on the set of real sequences.

3 Testing the approximation on *quasi-real* and real data sets

3.1 Parameters estimation for the Gumbel law

The distribution function of a Gumbel $G(\lambda, \delta)$ variable (say T) is given by :

$$P(T \leq x) = exp\left(-exp\left(-\left(\frac{x-\lambda}{\delta}\right)\right)\right), \quad x \in \mathbb{R}$$

Usually the first parameter is called the *decay parameter* and the second one the *characteristic value*.

To evaluate the relevance of our Gumbel $G(-K\sqrt{6}/\pi, \sqrt{6}/\pi)$ approximation (eq. 6), we consider three differents Z-value samples described below. Parameter estimations will be performed using the maximum likelihood method (see e.g. Johnson and Kotz, 1970) on different samples.

Data Description: A first databank of 16 956 sequences is built from five completely sequenced genomes (see (Comet et al., 1999) for details). Then we build a "*quasi-real*" sequence databank containing the shuffled versions of each of the real sequences. We compute the Z-value between the first sequence of this databank and the second one, between the second one and the third one and so on. We obtain 16 955 Z-values. But in such a sample, there are some dependencies. To break them we divide this previous sample into two smaller samples:

- The first sequence against the second one, the third against the fourth and so on. This sample has 8 478 Z-values.

- The second sequence against the third, the fourth against the fifth and so on. This sample has 8 477 Z-values.

The table 1 gives the values of the maximum likelihood estimators for these two samples. Another smaller sample is considered in order to appreciate the possible effect of the sample size. This one is built from *Saccharomyces cerevisiae*: we chose 1000 sequences at random and shuffled each of them. In the same way we computed the Z-values between the first sequence and the second one, between the third one and the fourth one and so on.

Tab. 1

Results: The results seem to be slightly different from those expected, especially for the decay parameter λ. Apart from the bias resulting from the maximum likelihood estimation, two possible explanations for these somewhat disappointing results may be explored : the first one deals with the quality of the Gumbel distribution approximation and the second one concerns the direct evaluation of the Z-values, in other words the role of the shuffling process.

Since the approximation (5) is nothing more than a simple consequence of the earlier Waterman and Vingron approach (Waterman and Vingron, 1994), there are no particular reasons to call it into question. However, the shuffling method may have a particular effect on the required estimations of $E(H(\mathbf{X},\mathbf{Y}))$ and $\sigma^2_{H(\mathbf{X},\mathbf{Y})}$. A detailed study is presented below.

3.2 Shuffling process and estimation bias

The two parameters $\alpha \equiv \alpha(\mathbf{X},\mathbf{Y})$ and $p \equiv p(\mathbf{X},\mathbf{Y})$ considered in the Poisson approximation (eq. 1) are of different nature. In the i.i.d. case, the p parameter does depend on the letter positions in each sequence, which is clearly related to the sequence compositions. At the opposite the α parameter seems to be dependent not only on the lengths but also on the structure of the sequences. Since the shuffling procedure breaks down the structures but saves the sequence compositions, it seems natural to consider that a possible effect of the shuffling procedure should particularly affect the α parameter. As a consequence, if we suppose that the shuffling process is applied to $\mathbf{Y}$, for all comparisons $(\mathbf{X},\mathbf{Y_i})_{i=1..N}$, the $p(\mathbf{X},\mathbf{Y_i})$ parameters can be considered as a constant p while the role of the $\alpha(\mathbf{X},\mathbf{Y_i})$ parameters must be taken into account.

For a particular sequence comparison $(\mathbf{X},\mathbf{Y}_i)$, under (5), we then have

$$H(\mathbf{X},\mathbf{Y}_i) \overset{\mathcal{D}}{\approx} \frac{\log \alpha(\mathbf{X},\mathbf{Y}_i)}{|\log p|} + \frac{\Lambda}{|\log p|}$$

where Λ is a Gumbel $G(0,1)$ variable.

It follows that

$$|\log p| \;\; E(\overline{H}_2(\mathbf{X},\mathbf{Y})) \simeq K + \frac{1}{N}\log\left(\prod_{i=1}^{N} \alpha(\mathbf{X},\mathbf{Y}_i)\right)$$

where $\overline{H}_2(\mathbf{X},\mathbf{Y}) = \frac{1}{N}\sum_{i=1}^{N} H(\mathbf{X},\mathbf{Y}_i)$.
Using (3), we obtain

$$|\log p| \left(E(H(\mathbf{X},\mathbf{Y})) - E(\overline{H}_2(\mathbf{X},\mathbf{Y}))\right) \simeq \log \alpha(\mathbf{X},\mathbf{Y}) - \frac{1}{N}\log\left(\prod_{i=1}^{N} \alpha(\mathbf{X},\mathbf{Y}_i)\right) \qquad (7)$$

which characterises the bias estimation for the mean when shuffling $\mathbf{Y}$.
Since

$$\frac{\sqrt{6}}{\pi}|\log p|\left(H(\mathbf{X},\mathbf{Y}) - \overline{H}_2(\mathbf{X},\mathbf{Y})\right) = \frac{\sqrt{6}}{\pi}|\log p|\left(H(\mathbf{X},\mathbf{Y}) - E(H(\mathbf{X},\mathbf{Y}))\right) + \frac{\sqrt{6}}{\pi}|\log p|\left(E(H(\mathbf{X},\mathbf{Y})) - \overline{H}_2(\mathbf{X},\mathbf{Y})\right) \quad (8)$$

we deduce from (7) that for N large enough

$$\widehat{Z_2} \approx Z + \frac{\sqrt{6}}{\pi}\log\frac{\alpha(\mathbf{X},\mathbf{Y})}{\left(\prod_{i=1}^{N}\alpha(\mathbf{X},\mathbf{Y}_i)\right)^{1/N}} \equiv Z + a_2 \quad (9)$$

where a_2 designates a constant value. Note that if $\forall i\ \alpha(\mathbf{X},\mathbf{Y}_i) \equiv \alpha(\mathbf{X},\mathbf{Y})$, then $a_2 = 0$.

When shuffling the sequence $\mathbf{X}$, the same type of result holds and we finally have:

$$\widehat{Z} = \min(\widehat{Z_1},\widehat{Z_2}) \approx Z + a \quad (10)$$

where

$$a = \frac{\sqrt{6}}{\pi}\min\left(\log\frac{\alpha(\mathbf{X},\mathbf{Y})}{\left(\prod_{i=1}^{N}\alpha(\mathbf{X},\mathbf{Y}_i)\right)^{1/N}},\ \log\frac{\alpha(\mathbf{X},\mathbf{Y})}{\left(\prod_{i=1}^{N}\alpha(\mathbf{X}_i,\mathbf{Y})\right)^{1/N}}\right).$$

Clearly the observed lack of fit between our Gumbel model and the results of our approximation may be simply related to the shuffling process. This problem is analyzed in the following and a bias reduction procedure is proposed.

3.3 Bias reduction

Consider the probability integral transform:

$$U = exp(-exp(-\frac{Z-\lambda_0}{\delta_0})) \quad (11)$$

where $\lambda_0 = -K\sqrt{6}/\pi$ and $\delta_0 = \sqrt{6}/\pi$. U is then uniformly distributed on $[0,1]$.

Quasi-real sequences: From data on "*quasi-real*" sequences the probability integral transform allows us to estimate the bias on decay parameter λ_0 using a QQ-plot approach. Ordering all probability integral transformed points $U_{(1)} \leq U_{(2)} \leq \cdots \leq U_{(N)}$, we have $E(U_{(i)}) = i/N+1$. Let us consider the $\left(\frac{i}{N+1}, U_{(i)}\right)$ points. These points should be accumulated close to the first bisecting line. To increase the resolution we use the log log transformation, and we consider

$$\left(-\delta_0\log\left(-\log\left(\frac{i}{N+1}\right)\right) + \lambda_0, \widehat{Z}_{(i)}\right) \quad (12)$$

If our approximation is correct, all points are expected to be close to the line $y = x$. If the slope of the QQ-plot is near 1, the intercept of linear regression gives an approximation a_0 for the bias a. If the slope is far from 1, our approximation (5) should be called into question.

We present below the QQ-plot for only the first sample composed of 8478 alignments (see fig. 1-A). Similar graphics are observed with the second and third samples.

Fig. 1

In order to test our $G(\lambda_0, \delta_0)$ we then consider the $\widetilde{Z}$-value defined by a correction on the shuffling estimations: $\widetilde{Z} = \widehat{Z} - a_0$. As shown in figure 1-B, the Gumbel distribution $G(\lambda_0, \delta_0)$ seems graphically to be a good approximation of the law of the Z-value.

Table 2 gives the maximum likelihood estimation results when using the corrected Z-value estimations.

Tab. 2

The results now obtained are close to the expected values, which supports the validity of our asymptotic approximation.

Real sequences : The Gumbel approximation concerns the comparisons between *i.i.d. random* sequences, that is, without an intrinsic structure. As already noted when considering real sequences, this underlying hypothesis will never be strictly satisfied, and in practical situations, deviations from the Gumbel law may be observed even for real sequences that have no biological relationships. As a consequence the same approach as the one used for quasi-real sequences should be irrelevant.

- A first way is to consider that the bias value a_0 obtained from quasi-real sequences can be used for real sequences comparisons. In such a case there are two possibilities : one can use an "universal" value for a estimated on a very large set of quasi-real sequences or one can implement for the real sequences under consideration the whole procedure which first build the associated quasi-real databank on which a_0 will be computed. In both cases the variable will be: $\widetilde{Z} = \widehat{Z} - a_0$.

- A second way may be to consider that the bias value a cannot be correctly estimated : the only information we have is given by the $\widehat{Z}$-values. But if the shuffling number is large enough, we have $a_0 \leq 0$. The reason is that the $\alpha(\mathbf{X}, \mathbf{Y})$-function decreases as a function of the $\mathbf{X}$ and $\mathbf{Y}$ similarity : under the null hypothesis of *iid* sequences, the closer $\mathbf{X}$ and $\mathbf{Y}$ are, the lower is the P-value. Using Poisson approximation (2) one expects that $\alpha(\mathbf{X}, \mathbf{Y_i}) \geq \alpha(\mathbf{X}, \mathbf{Y})$ for each i. In such a case our approximation leads to conservative conclusions.

In the sequel we will consider that the bias a is well approximated and we will compute $\widetilde{Z}$ with the value $a = a_0$.

Databank scanning : Several new challenges arise when a query sequence is used to scan a databank. All general databanks are build up of sequences that are widely different in length. These databanks include some sequences of the same family, and even duplicated sequences. Certainly, the iid model for real sequences fails. To remove the effects of duplication of sequences we constructed a protein database which included only one representative sequence from each protein family. The input data were taken from the databank described in Park and Teichman (Park and Teichmann, 1998)[1] retaining only one sequence from each cluster built from E. Coli. This bank contains 618 non-redondant sequences.

We choose now one of these sequences (EC1003) and compare it against all other sequences computing all $\widetilde{Z}$-values. The QQ-plot of these data is shown in figure 2. The model fits well with the empirical data on real sequences although the $\widetilde{Z}$-values for a databank scanning does not constitute a sample since the query sequence is shared by all alignments. This sequence represents the link between each alignment.

Fig. 2

[1] http://www.mrc-lmb.cam.ac.uk:80/genomes/

Global genome analysis : Now that many complete genomes have been sequenced, one extensive research domain deals with the classification of sequences from the same or from different genomes.

In such cases we are looking for biological links which are due to the duplication phenomenom. The hypothesis of independent sequences cannot be verified. To build clusters of sequences the first stage is to compute all the pairewise comparison indices, and to induce a dissimilarity matrix. Since the number of sequences is too large to simply apply classical classification methods, one often separate sequences in a first level of clusters by single linkage clustering. In each cluster a hierarchical analysis can be performed.
For such a goal the most important point is to have a global index which does not depend on individual sequences, especially on individual sequence length. In this problematics the Z-value can be useful.

From the complete genome from *Saccharomyces cerevisiae* we randomly chose 1000 sequences. This database has been shuffled to build a quasi-real sequences databank. On both sets of sequences (quasi-real and real) all pairewise comparisons have been performed and all pairewise $\widehat{Z}$-values computed and corrected. Figure 3 shows the probability integral transform for both sets of non independent $\widetilde{Z}$-values.

Fig. 3

Despite the dependency between the $H(\mathbf{X}, \mathbf{Y})$ scores, the Gumbel distribution fits well in the case of quasi-real sequence comparisons. In the case of real sequences one notice a totally different behavior : the observed $\widetilde{Z}$-values significatively deviate from the Gumbel law as earlier noticed in (Comet et al., 1999). For smaller values the Gumbel model seems to be valid. The cut-off value v may be related to the 0.9999 quantile of the $G\left(-\frac{K\sqrt{6}}{\pi}; \frac{\sqrt{6}}{\pi}\right)$ distribution which is about 6.7. Note that this threshold supports the empirical threshold used by biologists : in practice the value 8 allows them to determine if an alignment is biologically significant or not.

4 Conclusion

This article gives a frame to justify the use of simulations to evaluate the significance of gapped alignments. It is well known that the Smith-Waterman score law depends on length and amino acid composition of sequences. This study shows that the asymptotic law of the Z-value is sequence independent, which is fundamental particularly when analyzing complete genomes.

In practical applications, one can observe a deviation of the Z-values from the initial Gumbel distribution. This divergence from the asymptotic approximation law highlights the biological links : if an empirical Z-value is greater than a cutoff [2], the null hypothesis of random sequences is rejected, which means that we may conclude to the existence of a biological link.
In other words all conclusions based on simulations are interpretable, since the asymptotic law of Z-value is independent of sequences. Only the shuffling process can introduce a bias, which is evaluated by the exposed method. This frame gives a new view on the 20 years old method for achieving the significance of gapped alignment.

[2]For details, see (Comet et al., 1999)

References

Aldous, D. (1989). *Probability Approximations via the Poisson Clumping Heuristic.* Springer, New York.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.*, 215:403–410.

Arratia, R., Gordon, L., and Goldstein, L. (1989). Two moments suffice for Poisson approximations : the Chen-Stein method. *Ann. Prob.*, 17:9–25.

Arratia, R., Gordon, L., and Waterman, M. S. (1986). An extreme value theory for sequence matching. *Ann. Stat.*, 14:971–993.

Arratia, R., Gordon, L., and Waterman, M. S. (1990). The Erdos-Renyi law in distribution for coin tossing and sequence matching. *Ann. Stat.*, 18:539–570.

Arratia, R. and Waterman, M. S. (1989). The Erdos-Renyi strong law for pattern matching with a given proportion of mismatches. *Ann. Prob*, 17:1152–1169.

Arratia, R. and Waterman, M. S. (1994). A phase transition for the score in matching random sequences allowing deletions. *Ann. appl. Prob.*, 4:200–225.

Comet, J.-P. (1998). *Programmation dynamique et alignements de séquences biologiques.* PhD thesis, Université de Compiègne, France.

Comet, J.-P., Aude, J.-C., Glémet, E., Risler, J.-L., Hénaut, A., Slonimski, P., and J.-J., Codani (1999). Significance of Z-value statistic of Smith-Waterman scores for protein alignments. *Comput. Chem.*, 23:317–331.

Dembo, A. and Karlin, S. (1991a). Strong limit laws of empirical functionals for large exeedences of partial sums of i. i.d. variables. *Ann. Prob.*, 19:1737–1755.

Dembo, A. and Karlin, S. (1991b). Strong limit theorems of empirical distributions for large segmental exceedances of partial sums of Markov variables. *Ann. Prob.*, 19:1756–1767.

Goldstein, L. and Waterman, M. S. (1992). Poisson, compound Poisson and process approximations for testing statistical significance in sequence comparisons. *Bull. Math. Biol.*, 54(5):785–812.

Goldstein, L. and Waterman, M. S. (1994). Approximations to profile score distributions. *J. Comput. Biol.*, 1(2):93–104.

Johnson, N. L. and Kotz, S. (1970). *Distribution in statistics : Continuous univariate distributions - 1.* The Houghton Mifflin Series in statistics.

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

Karlin, S. and Dembo, A. (1992). Limit distributions of maximal segmental score among Markov-dependent partial sums. *Ann. Appl. Prob.*, 24:113–140.

Karlin, S., Dembo, A., and Kawabata, T. (1990). Statistical composition of high-scoring segments from molecular sequences. *Ann. Stat.*, 18:571–581.

Landes, C., Hénaut, A., and Risler, J.-L. (1992). A comparaison of several similarity indices bused in the classification of protein sequences: a multivariate analysis. *Nucl. Acids Res.*, 20(14):3631–3637.

Lipman, D. J., Wilbur, W. J., Smith, T. F., and Waterman, M. S. (1984). On the statistical signifiance of nucleic acid similarities. *Nucl. Acids Res.*, 12:215–226.

Mott, R. and Tribe, R. (1999). Approximate Statistics of Gapped Alignment. *J. Comput. Biol.*, 6(1):91–112.

Park, J. and Teichmann, S. A. (1998). Divclus: an automatic method in the geanfammer package that finds homologous domains in single- and multi-domain proteins. *Bioinformatics*, 14:144–150.

Slonimski, P. P. and Brouillet, S. (1993). A database of chromosome III of saccharomyces cerevisiae. *Yeast*, 9:941–1029.

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

Waterman, M. S. (1994a). Estimating statistical significance of sequence alignments. *Phil. Trans. R. Soc. Lond. B*, 344:383–390.

Waterman, M. S. (1994b). The statistical significance of local alignement score. Technical report, University of Southern California, USA.

Waterman, M. S. and Vingron, M. (1994). Sequence comparison significance and Poisson approximation. *Stat. Sci.*, 9(3):367–381.

	λ	δ
8478 Z-values	-0.549	0.789677
8477 Z-values	-0.527	0.789987
500 Z-values	-0.535	0.796190
Gumbel model	$-K\sqrt{6}/\pi = -0.45$	$\sqrt{6}/\pi = 0.7797$

Table 1: **Gumbel maximum likelihood estimations.** λ and δ are the decay parameter and the characteristic value of the Gumbel law.

	λ	δ
8478 Z-values	-0.454	0.789668
8477 Z-values	-0.432	0.789974
500 Z-values	-0.441	0.796196
Gumbel model	$-K\sqrt{6}/\pi = -0.45$	$\sqrt{6}/\pi = 0.7797$

Table 2: **Gumbel maximum likelihood estimations - corrected $\widehat{Z}$-values.** λ and δ are the decay parameter and the characteristic value of the Gumbel law.

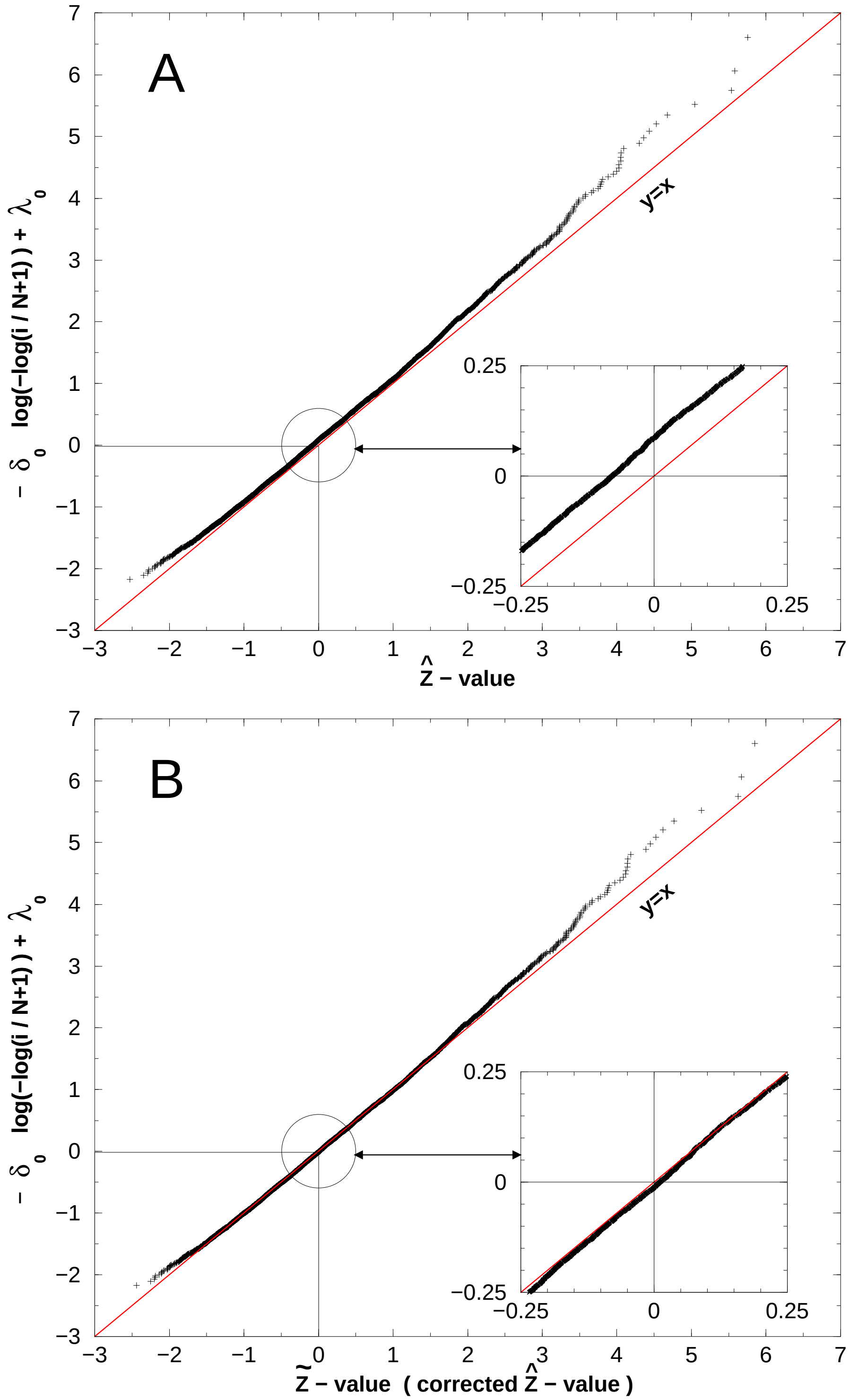

Figure 1: **QQ-plot on *quasi-real* sequences (first sample : 8478 alignments):**
Figure A**:** QQ-plot of $\widehat{Z}$-values
Figure B**:** QQ-plot of the corrected $\widehat{Z}$-values : $\widetilde{Z} = \widehat{Z} - a_0$
The graphic A allows to approximate the correction a_0 induced by the shuffling procedure (see text).

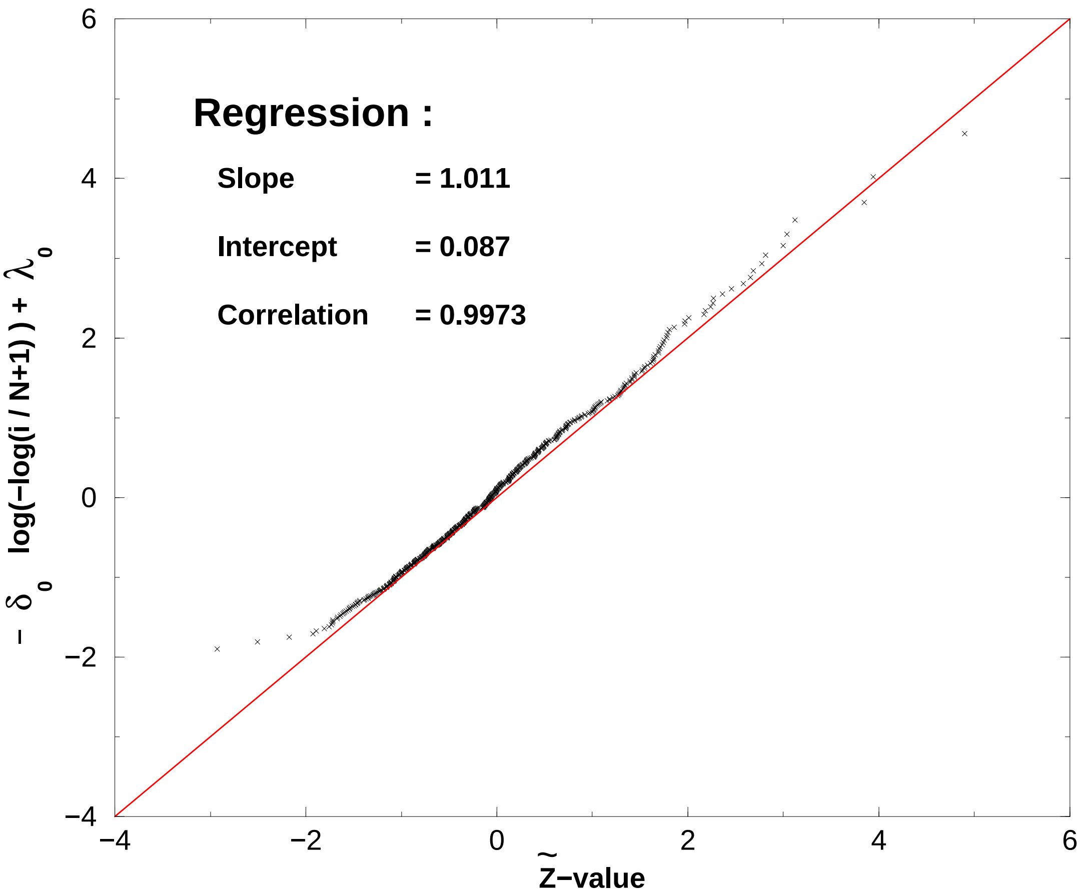

Figure 2: **QQ-plot of $\widetilde{Z}$-values obtained during a databank scanning :** Comparison of one sequence against a non redondant databank of sequences.

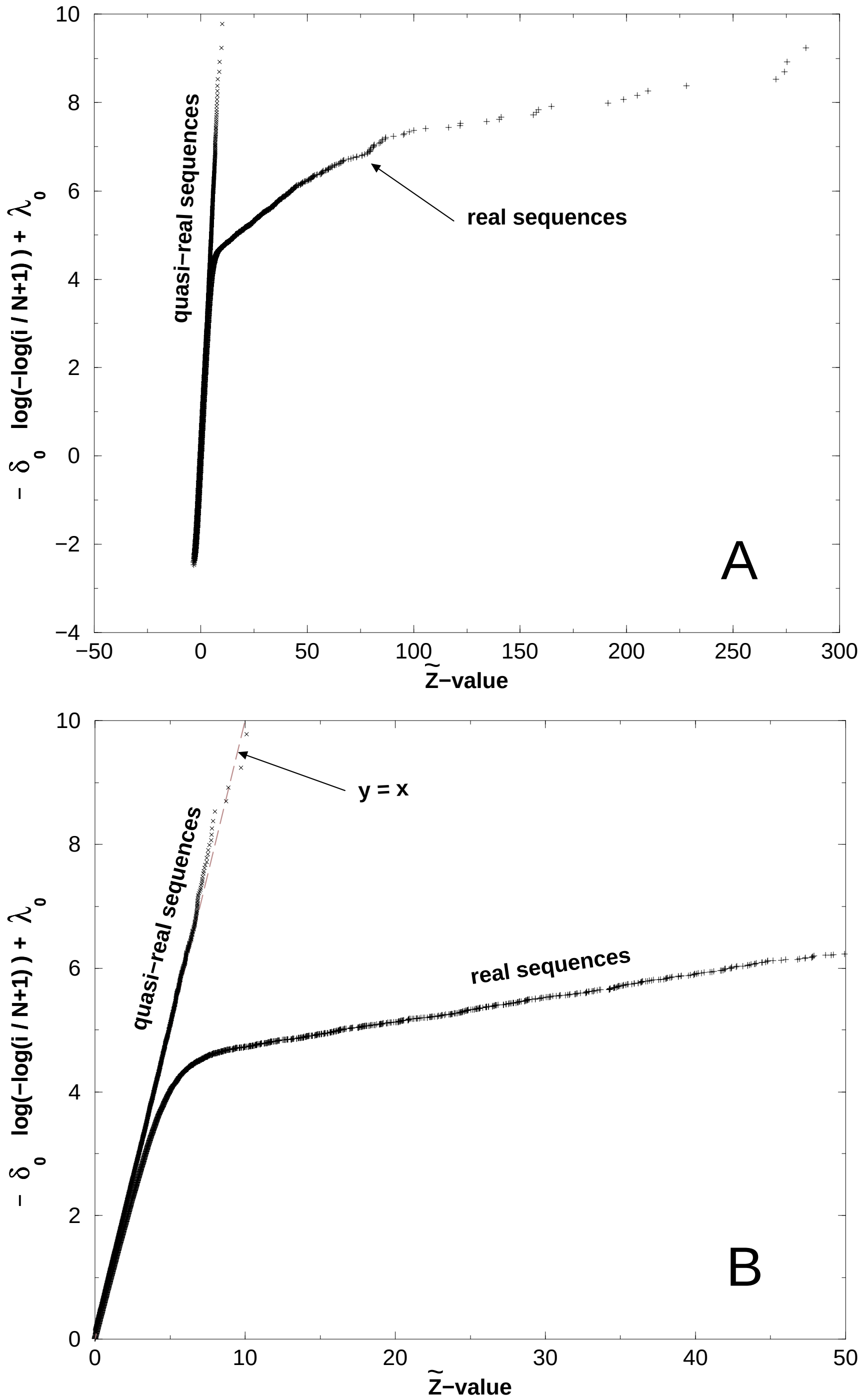

Figure 3: **QQ-plot on two sets built on Yeast genome (figure B is a zoom of figure A).**
Real sequences: 1000 real sequences have been choosen at random in the complete genome of *Saccharomyces cerevisiae*. All pairewise Z-values have been computed (499500 Z-values).
Quasi real sequences: 1000 quasi real sequences have been built by shuffling the above real sequences. The 499500 Z-values have been computed. For real sequences we observed a behavior different from that for quasi-real ones. For real sequences 94 Z-values are greater than 50.