ELSEVIER

Computers and Chemistry 25 (2000) 401–410

Computers & Chemistry

www.elsevier.com/locate/compchem

# Sequence alignment: an approximation law for the $Z$-value with applications to databank scanning

J.N. Bacro [a,*], J.P. Comet [b,1]

[a] *INA PG, Dpt OMIP, U.M.R. INAPG/INRA, 16, rue Claude Bernard, 75231 Paris Cedex 05, France*
[b] *LaMI, Université d'Evry-Val d'Essonne, Cours Monseigneur Roméro, 91025 Evry Cedex, France*

Received 16 June 2000; received in revised form 14 January 2001; accepted 16 January 2001

**Abstract**

The $Z$-value is an attempt to estimate the statistical significance of a Smith and Waterman dynamic programming 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 developed by Waterman and Vingron (Stat. Sci. 9 (1994) 367) in the case of independent and identically distributed sequences comparison. As for non-gapped alignment scores, our approximation is of Gumbel type but with parameters that are sequence independent. This result makes clear the related experimental results mentioned by Comet et al. (Comput. Chem. 23 (1999) 317). Using 'quasi-real' sequences (i.e. randomly shuffled sequences of the same length and amino acid composition as the real ones) we investigate the relevance 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. 

*Keywords:* Dynamic programming sequence alignment; Significance; $Z$-value; Approximated distribution; Gumbel distribution

## 1. Introduction

Sequence comparison has become a central notion in modern molecular biology. To evaluate the similarity between two sequences, many indices are now available, allowing global alignments and gapped or ungapped local alignments. The algorithm of Smith and Waterman (1981) answers exhaustively 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) or developing a structural relatedness procedure based on extended non-gapped alignment results (Abagyan and Batalov, 1997). The second one is based on simulation results, using a shuffling procedure and a particular statistics called

* Corresponding author. Tel.: +33-1-44-08-7271; fax: +33-1-44-08-1666.
[1] Corresponding co-author. Tel.: +33-1-69-47-7453; fax: +33-1-69-47-7472
*E-mail addresses:* bacro@inapg.inra.fr (J.N. Bacro), comet@lami.univ-evry.fr (J.P. Comet).

0097-8485/01/$ - see front matter 
PII: S0097-8485(01)00074-2

$Z$-value (McLachlan, 1971; Lipman et al., 1984; Landès et al., 1992; Slonimski and Brouillet, 1993). In a recent paper, 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 the data well, 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 et al., 1999 for details). The aim of the present paper is to precise and to give a new light on these experimental results.

For independent and identically distributed (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. From a practical point of view, the $Z$-values for real sequence comparisons are usually evaluated through a shuffling procedure. Focusing at first on shuffled sequences comparison (that means alignment of two sequences coming from a shuffling of two real ones), we characterize 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 the proposed model well.

In the case of real sequences, the $Z$-value asymptotic distribution appears to be of the same type as the one for shuffled sequences (Gumbel law) but with other parameters. The experimental results of Comet et al. (1999) then become clear: there is only one type of law for the $Z$-value distribution approximation and the only change from shuffled sequences comparison to real ones is in fact the change of scale for the approximation distribution. This result can be used to characterize the statistical significance of a $Z$-value when looking for similarity between real sequences.

This article is organized as follows: Section 2 defines the $Z$-value variable following Comet et al. (1999). Section 3 is devoted to the asymptotic approximation for the $Z$-value distribution under the hypothesis of a sequence comparison between two i.i.d. random sequences. Section 4 focuses on testing the approximation law for shuffled sequences comparisons. A correction procedure for the parameter estimation is proposed in order to take into account the shuffling induced bias. This procedure is then applied to real sequences. Section 5 gives a general conclusion concerning the use of $Z$-value for gapped alignments.

## 2. 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 penalty 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 referred as score *with parameters in the logarithmic region* (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 $H(\mathbf{X},\mathbf{Y})$.

## 3. 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 independent and identically distributed. 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 $1 - \mathrm{e}^{-\gamma mnp^{H(\mathbf{X},\mathbf{Y})}}$ where $\gamma$ and $p$ are two parameters to be estimated.

The Waterman and Vingron (1994) result is based on the approximation:

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

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

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

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

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

where $\alpha \equiv \alpha(\mathbf{X},\mathbf{Y})$ and $p \equiv p(\mathbf{X},\mathbf{Y})$ are two positive parameters (this corresponds to 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

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

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

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

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

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

where $K = 0.57721$ denotes 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:

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

which can be stated as

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

In other words our approximation is sequence independent: in Eq. (6), the approximation of the $Z$-value distribution does not depend on sequence lengths and compositions. 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, 1994; Arratia et al., 1986, 1989, 1990; Dembo and Karlin, 1991a,b; Karlin and Altschul, 1990; Karlin et al., 1990; Karlin and Dembo, 1992; Goldstein and Waterman, 1992, 1994; Waterman, 1994a,b; Waterman and Vingron, 1994), there are no results yet available concerning the sequence composition dependency. In particular, the normalization described by the first equation in (b) is an attempt to take into account the different lengths of the considered sequences, but seems to be poorly fitted in most of the practical situations (Waterman and Vingron, 1994). Note that approximations (2) and (6) are both obtained in the case of i.i.d. random sequences comparison. But our results do not need any particular normalization concerning the lengths and the amino acid compositions of the involved sequences since both are indeed taken into account through the evaluation process of the $Z$-value. From these different facts, the $Z$-value is clearly relevant. 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. (1999) which seems to be well adapted to simulate random sequences with the same amino acid composition and length than the initial ones. Following Comet et al. (1999), we compute two different $Z$-values $\hat{Z}_1(\mathbf{X},\mathbf{Y})$ and $\hat{Z}_2(\mathbf{X},\mathbf{Y})$ by shuffling the first and second sequence, respectively, and use the minimum value to estimate $Z(\mathbf{X},\mathbf{Y})$. The choice of the minimum value as an estimator for $Z(\mathbf{X},\mathbf{Y})$ is argued in Comet et al. (1999) and leads to a conservative approach for the test comparison.

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 (1994) result, but under the only assumptions (A1) and (A2), it seems reasonable to think that our result would be validated 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 theoretical 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 not i.i.d. 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 the i.i.d. assumption would clearly be a really major drawback for practical applications. In order to appreciate the robustness of our result we decide to test our approximation on shuffling sequences built from real ones. Such sequences are not i.i.d. but do not exhibit any particular structure effect and do not present any more biological feature. Since no biological links are present in these sequences, we clearly hope that our approximation fits well with the related $Z$-value observations. A deviation from our Gumbel law in practical applications would then indicate significant similarities between the considered sequences.

In the sequel, we consider two sets of sequences, fully described in Comet et al. (1999): the set of real sequences and the set of 'quasi-real' sequences which designates the set of sequences obtained by shuffling real ones. Apart 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 then shuffled many times to evaluate the $Z$-value leading to a set of values for quasi-real sequence alignments. We shall see first that for such a set, a direct application of our approximation leads to a poor fit. Showing that the shuffling approach induces an estimation bias, we will propose a correction procedure. Having then a good fit for such sequences close to random sequences, we will apply the whole procedure on the set of real sequences.

## 4. Testing the approximation on quasi-real and real data sets

### 4.1. Parameter estimation for the Gumbel law

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

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

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 different $Z$-value samples described below. Parameter estimations will be performed using the maximum likelihood method (see, e.g. Johnson and Kotz, 1970) on different samples.

#### 4.1.1. Data description

A first databank of 16956 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. For each alignment score, we use the PAM 250 substitution matrix and the values 5 and 0.3 for gap-open and gap-extend penalties. We obtain 16955 $Z$-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 8478 $Z$-values.
- The second sequence against the third, the fourth against the fifth and so on. This sample has 8477 $Z$-values.

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.

Table 1
Gumbel maximum likelihood estimations [a]

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

[a] $\lambda$ and $\delta$ are the decay parameter and the characteristic value of the Gumbel law.

#### 4.1.2. Results

The results seem to be slightly different from those expected, especially for the decay parameter $\lambda$. 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 approximation (5) is nothing more than a simple consequence of the earlier Waterman and Vingron (1994) approach, 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 in the following paragraphs.

### 4.2. 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. (2)) 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. Now, consider some i.i.d. sequences with the same length and a common amino acid composition. For all the alignments of such sequences, the $p$ parameter can be considered as a unique constant. But some of the corresponding $H$-scores may be quite different from the other ones. As a consequence, when comparing two sequences, it seems that the $\alpha$ parameter used in approximation (2) could be dependent not only on the lengths but also on the structures of the sequences. Since the shuffling procedure breaks down the structures but respects the sequence compositions, it seems natural to consider that a possible effect of the shuffling procedure should particularly affect the $\alpha$ parameter.

Consequently, 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 Eq. (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 $\Lambda$ is a Gumbel $G(0,1)$ variable.

It follows that

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

where

$$\bar{H}_2(\mathbf{X},\mathbf{Y}) = \frac{1}{N}\sum_{i=1}^{N} H(\mathbf{X},\mathbf{Y}_i).$$

Using Eq. (3), we obtain

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

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

Since

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

we deduce from Eq. (7) that for $N$ large enough

$$\hat{Z}_2 \approx Z + \frac{\sqrt{6}}{\pi}\log\frac{\alpha(\mathbf{X},\mathbf{Y})}{\left(\prod_{i=1}^{N}\alpha(\mathbf{X},\mathbf{Y}_i)\right)^{1/N}} \equiv Z + a_2 \quad (9)$$

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

$$\hat{Z} = \min(\hat{Z}_1,\hat{Z}_2) \approx Z + a \quad (10)$$

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 section and a bias reduction procedure is proposed.

### 4.3. Bias reduction

Consider the probability integral transform:

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

where $F \equiv F_{\lambda_0,\delta_0}$ designates the $G(\lambda_0,\delta_0)$ distribution function with $\lambda_0 = -K\sqrt{6}/\pi$ and $\delta_0 = \sqrt{6}/\pi$. $U$ is then uniformly distributed on [0,1]. Defining the ordered sample $U_{(1)} \le \cdots \le U_{(N)}$, we have $U_{(i)} = F(Z_{(i)})$ and $E(U_{(i)}) = i/(N+1)$. The graph $\{(U_{(i)},(i/(N+1))),i = 1,\ldots,N\}$ is approximately linear and it is common to plot the graph $\{(Z_{(i)},F^{-1}(i/(N+1))),i = 1,\ldots,N\}$, often called the quantile plot (or QQ-plot), to test whether the sample $Z_1,\ldots,Z_N$ comes from the $F$ distribution (see Shorack and Wellner, 1986 for details). Briefly, if the data were generated from the $F$ distribution, the plot should look close to the line $y = x$; moreover, the change of the reference distribution by a linear transformation simply transforms the QQ-plot by the same transformation.

Below we use a QQ-plot approach to test our Gumbel approximation for quasi-real sequences.

#### 4.3.1. Quasi-real sequences

From data on quasi-real sequences a QQ-plot approach is used to test graphically the $F$ distribution as the reference distribution for the $Z$-value data set.

We consider the graph

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

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

In order to test our $G(\lambda_0,\delta_0)$ law we then consider the $\tilde{Z}$-value defined by a correction on the shuffling estimations: $\tilde{Z} = \hat{Z} - a_0$. As shown in Fig. 1B, 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.

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

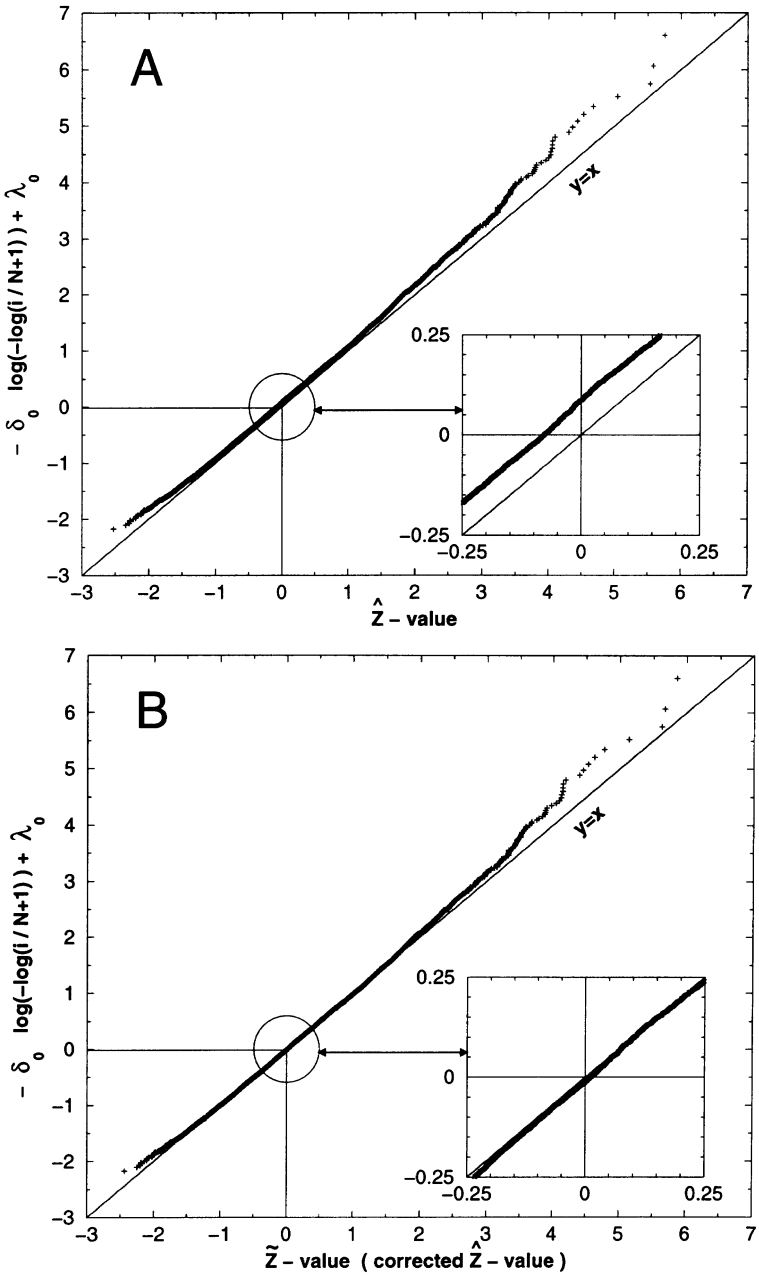


Fig. 1. QQ-plot on *quasi-real* sequences (first sample: 8478 alignments): (A) QQ-plot of $\hat{Z}$-values; (B) QQ-plot of the corrected $\hat{Z}$-values: $\tilde{Z} = \hat{Z} - a_0$. The graphic A allows to approximate the correction $a_0$ induced by the shuffling procedure (see text).

*4.3.2. Real sequences*

The Gumbel approximation concerns the comparisons between *random* sequences, that is, without any intrinsic structure. As already noted when considering real sequences, this underlying hypothesis will never be strictly satisfied, and in real practical situations, deviations from the Gumbel law may be observed even for real sequences that have no biological relationships. Consequently, a direct approach such as the one used for quasi-real sequences should be erroneous and has to be modified as follows:

- 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 se-

Table 2
Gumbel maximum likelihood estimations — corrected $\hat{Z}$-values [a]

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

[a] $\lambda$ and $\delta$ are the decay parameter and the characteristic value of the Gumbel law.

quences 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: $\tilde{Z} = \hat{Z} - a_0$.

- 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 $\hat{Z}$-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 i.i.d. 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 $\tilde{Z}$ with the value $a = a_0$.

#### 4.3.3. Databank scanning

Several new challenges arise when a query sequence is used to scan a databank. All general databanks are built up with sequences that are widely different in length. These databanks include some sequences of the same family, and even duplicated sequences. Certainly, the i.i.d. model for real sequences fails. To remove the effects of duplication of sequences we constructed a protein database which includes only one representative sequence from each protein family. The input data were taken from the databank described in Park and Teichmann (1998) (http://www.mrc-lmb.cam.ac.uk:80/genomes/) 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 $\tilde{Z}$-values. The QQ-plot of these data is shown in Fig. 2. The model fits well with the empirical data on real sequences although the $\tilde{Z}$-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.

#### 4.3.4. Global genome analysis

Now that many complete genomes have been sequenced, one extensive research domain deals with the

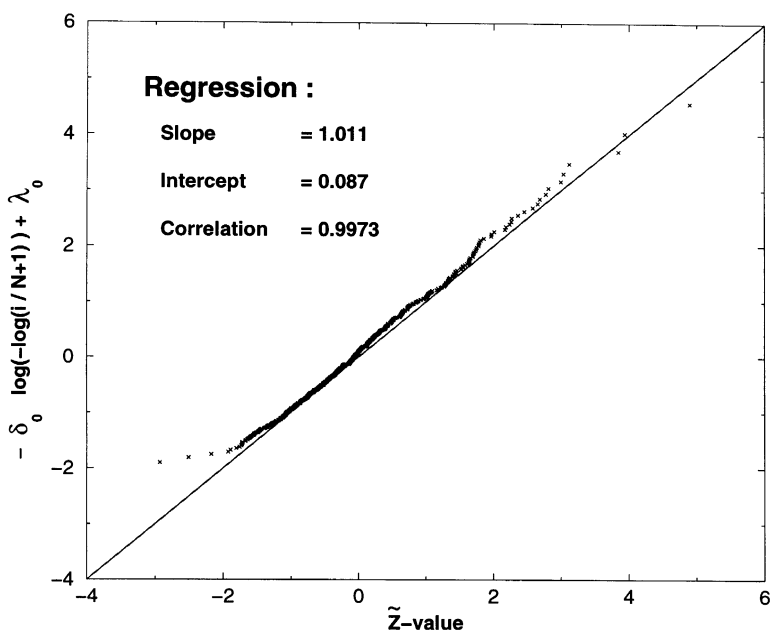


Fig. 2. QQ-plot of $\tilde{Z}$-values obtained during a databank scanning: comparison of one sequence against a non-redundant databank of sequences.

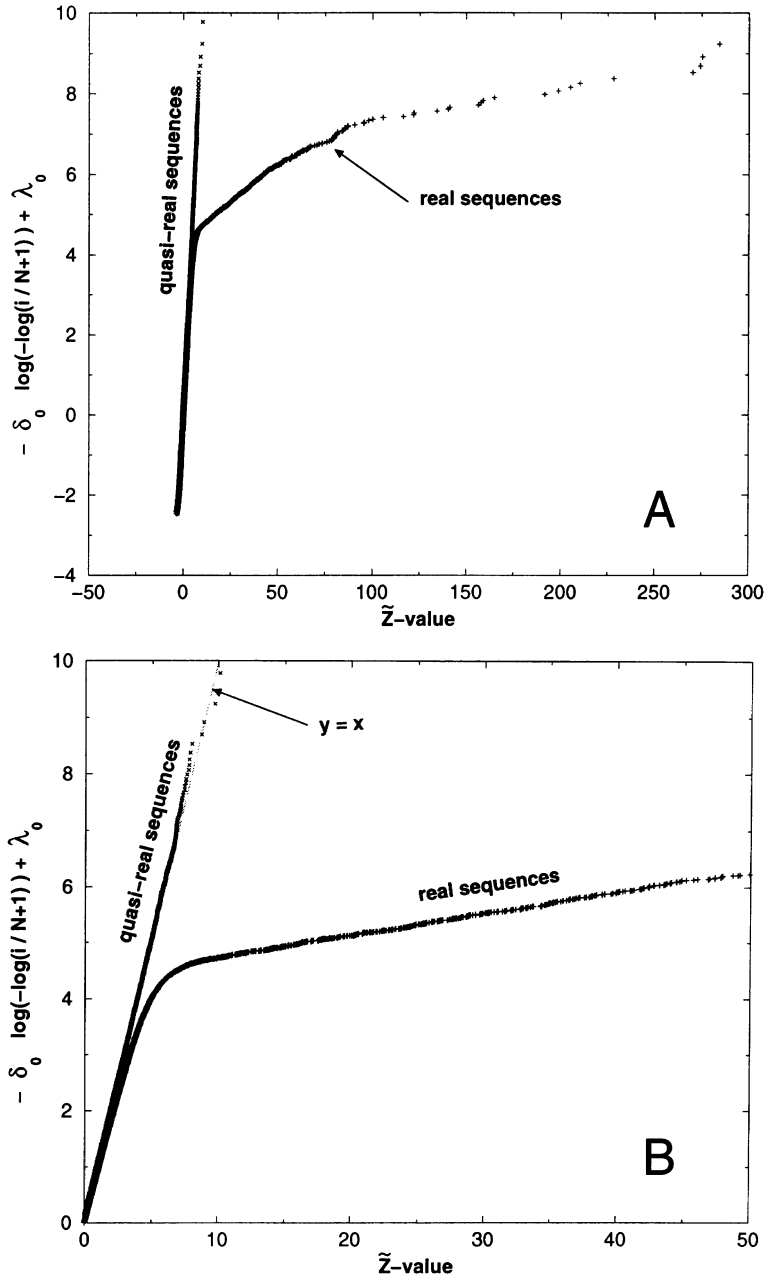


Fig. 3. QQ-plot on two sets built on yeast genome ((B) is a zoom of (A)). Real sequences: 1000 real sequences have been chosen at random in the complete genome of *S. cerevisiae*. All pairwise $\hat{Z}$-values have been computed (499 500 $\hat{Z}$-values). Quasi-real sequences: 1000 quasi-real sequences have been built by shuffling the above real sequences. The 499 500 $\hat{Z}$-values have been computed. For real sequences we observed a behavior different from that for quasi-real ones. For real sequences 94 $\hat{Z}$-values are greater than 50.

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 phenomenon. The hypothesis of independent sequences cannot be verified. To build clusters of sequences the first stage is to compute all the pairwise comparison indices, and to induce a dissimilarity matrix. Since the number of sequences is too large to simply apply classical classification methods, one often separates 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 such problematics the $Z$-value can be useful.

From the complete genome from *S. 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 pairwise comparisons have been performed and all pairwise $\hat{Z}$-values computed and corrected. Fig. 3 shows the QQ-plots for both sets of non-independent $\tilde{Z}$-values.

Despite the dependency between the $\tilde{Z}(\mathbf{X},\mathbf{Y})$ scores, the Gumbel distribution fits well in the case of quasi-real sequences. In the case of real sequences one notices a totally different behavior: the observed $\tilde{Z}$-values significantly 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 $G(-K\sqrt{6}/\pi;\ \sqrt{6}/\pi)$ 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.

## 5. Conclusions

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 and 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, 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 alignments.

## Acknowledgements

We thank the anonymous referee for many constructive suggestions that improved substantially the presentation of the paper.

## References

Abagyan, R.A., Batalov, S., 1997. Do aligned sequences share the same fold? J. Mol. Biol. 273 (1), 355–368.

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

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

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

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

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

Arratia, R., 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., 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., Codani, J.-J., 1999. Significance of $Z$-value statistic of Smith–Waterman scores for protein alignments. Comput. Chem. 23, 317–331.

Dembo, A., Karlin, S., 1991a. Strong limit laws of empirical functionals for large exceedances of partial sums of i.i.d. variables. Ann. Prob. 19, 1737–1755.

Dembo, A., 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., 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., Waterman, M.S., 1994. Approximations to profile score distributions. J. Comput. Biol. 1 (2), 93–104.

Johnson, N.L., Kotz, S., 1970. Distribution in statistics: continuous univariate distributions — 1, The Houghton Mifflin Series in Statistics.

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

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

Landès, C., Hénaut, A., Risler, J.L., 1992. A comparison of several similarity indices used 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., Waterman, M.S., 1984. On the statistical significance of nucleic acid similarities. Nucl. Acids Res. 12, 215–226.

McLachlan, A.D., 1971. Tests for comparing related amino-acid sequences. Cytochrome c and cytochrome c 551. J. Mol. Biol. 61 (2), 409–424.

Mott, R., Tribe, R., 1999. Approximate statistics of gapped alignment. J. Comput. Biol. 6 (1), 91–112.

Park, J., 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.

Shorack, G.R., Wellner, J.A., 1986. Empirical processes with applications to statistics. Wiley, New York.

Slonimski, P.P., Brouillet, S., 1993. A database of chromosome III of *Saccharomyces cerevisiae*. Yeast 9, 941–1029.

Smith, T.F., 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. Philos. Trans. R. Soc. London, B 344, 383–390.

Waterman, M.S., 1994b. The statistical significance of local alignment score. Technical Report, University of Southern California, USA.

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