PERGAMON

Computers & Chemistry 23 (1999) 317–331

Computers & Chemistry

Significance of *Z*-value statistics of Smith–Waterman scores for protein alignments

J.P. Comet[a,*], J.C. Aude[a], E. Glémet[a], J.L. Risler[b], A. Hénaut[b], P.P. Slonimski[b], J.J. Codani[a]

[a]*INRIA Rocquencourt, B.P. 105, 78153 Le-Chesnay Cedex, France*
[b]*Centre de Génétique Moléculaire du CNRS, 91198 Gif sur Yvette Cedex, France, et Université de Versailles, 78035 Versailles Cedex, France*

Abstract

The *Z*-value is an attempt to estimate the statistical significance of a Smith–Waterman dynamic alignment score (SW-score) through the use of a Monte–Carlo process. It partly reduces the bias induced by the composition and length of the sequences.

This paper is not a theoretical study on the distribution of SW-scores and *Z*-values. Rather, it presents a statistical analysis of *Z*-values on large datasets of protein sequences, leading to a law of probability that the experimental *Z*-values follow.

First, we determine the relationships between the computed *Z*-value, an estimation of its variance and the number of randomizations in the Monte–Carlo process. Then, we illustrate that *Z*-values are less correlated to sequence lengths than SW-scores.

Then we show that pairwise alignments, performed on 'quasi-real' sequences (i.e., randomly shuffled sequences of the same length and amino acid composition as the real ones) lead to *Z*-value distributions that statistically fit the extreme value distribution, more precisely the Gumbel distribution (global EVD, Extreme Value Distribution). However, for real protein sequences, we observe an over-representation of high *Z*-values.

We determine first a *cutoff value* which separates these overestimated *Z*-values from those which follow the global EVD. We then show that the interesting part of the tail of distribution of *Z*-values can be approximated by another EVD (i.e., an EVD which differs from the global EVD) or by a Pareto law.

This has been confirmed for all proteins analysed so far, whether extracted from individual genomes, or from the ensemble of five complete microbial genomes comprising altogether 16956 protein sequences.

Keywords: Sequence alignment; Dynamic programming; Significance; *Z*-value; Gumbel distribution; Pareto distribution

1. Introduction

This article introduces a method for building intra- and inter-genomic families of proteins from several microbial genomes. The overall stategy, described in Codani et al. (1999), is based on the *Z*-value, which is

* Corresponding author.
E-mail address: Jean-Paul.Comet@inria.fr (J.P. Comet)

0097-8485/99/$ - see front matter
PII: S0097-8485(99)00008-X

an attempt to estimate the statistical significance of a Smith–Waterman dynamic programming alignment score (SW-score) through the use of a Monte–Carlo process. We describe the methods used to adjust the parameters of the Monte–Carlo process, and analyze the Z-value distribution. The present analysis is different from the classical 'sequence against library' approach since our aim is to compare a finite set of sequences (i.e., a complete genome) against itself or against another finite set of sequences (another complete genome or several complete genomes) in all possible pairwise combinations. Indeed, since the Z-values follow a known distribution (*vide infra*), we have been able to perform systematic inter- and intra-genomic comparisons of protein sequences by the Smith–Waterman algorithm, followed by single-linkage clusterings based on the probabilities associated with the Z-values (see the paper by Aude et al. in this issue; Codani et al., 1999; Diaz-Lazcoz et al., 1998; Slonimski et al., 1998).

Using Z-values rather than Smith–Waterman scores obviously leads to different results. In (Codani et al., 1999), we illustrate the quantitative differences observed between scores and Z-values at a whole genome comparison level, and demonstrate the non-correlation between scores and Z-values in the 'twilight zone', i.e., the range of scores in between high scores (sequences are obviously related) and low scores (sequences are obviously non related). The relevance of a pairwise alignment method relies precisely in its ability to provide a reliable criterion concerning the similarity between sequences in the twilight zone.

This article is organized in two sections. The first section defines the Z-value. In the first subsection, we determine the relationship between the computed Z-value, an estimation of its variance and the number of randomizations in the Monte–Carlo process. The second subsection underlines the asymmetric bias, leading to a new definition of the Z-value. The third subsection deals with the influence of the sequence lengths and shows that the Z-values are less dependent on the sequence lengths than the Smith–Waterman scores.

The second section focuses on a statistical analysis of the distribution of Z-values for real proteins, and for sequences of the same length and amino acid composition generated by random shuffling of real sequences (which we shall call henceforth 'quasi-real' sequences). For quasi-real sequences, Extreme Value Distribution (EVD) estimated on *the whole set of alignments* (global EVD) fits well the experimental distribution of Z-values. In contrast, for real proteins, we observe an over-representation of high Z-values as compared to the global EVD model. We determine first a *cutoff value* which separates these overestimated Z-values from those which follow the global EVD. We then show that the 'interesting part' of the tail of distribution of Z-values can be approximated by another EVD (i.e., an EVD which differs from the global EVD) or by a Pareto law.

2. Z-value

2.1. Definitions

The first adaptation of dynamic programming for sequence alignments was due to Needleman and Wunsch (1970) who proposed an efficient algorithm to determine the best gapped *global alignment* between two sequences. The method was later extended to local alignments by Smith and Waterman (1981). Subsequent improvements and extensions were made by Waterman and Eggert (1987) and Miller and Huang (1991). Any alignment of two protein sequences by these algorithms results in a so-called optimal alignment score. Nevertheless, the optimality of the score does not ascertain that the two sequences are indeed related. Numerous reports focus on the expression of a probability that the score could be obtained by chance. For non-gapped alignements, such as those reported by Blast (Altschul et al., 1990), a theoretical model exists. This model does not apply for gapped alignments. One can refer to Mott (1992), where a method is described for estimating the distribution of scores obtained from a databank search using the Smith–Waterman algorithm, that takes into account the length and composition of the sequences in the distribution function. An interesting approach by Waterman and Vingron (1994) gives an estimation of the significance of the score of a gapped alignment. The authors use the Poisson clumping heuristic to describe the behavior of scores: as a result, the probability for a score to be lower than or equal to t is approximately $\exp(-\gamma mnp^t)$, where m, n are the sequence lengths, and γ and p are parameters estimated from the data.

A complementary approach is to use the Z-value. The Z-value relies on a Monte–Carlo evaluation of the significance of the Smith–Waterman score (Lipman et al., 1984; Landès et al., 1992; Slonimski and Brouillet, 1993). The method consists in comparing one of the two sequences with as many as possible randomly shuffled versions of the second one (Lipman et al., 1984). The shuffled sequences share exactly with the initial sequence the same amino acid composition and length. This simulation eliminates in most cases the bias due to the amino acid composition, and partly to the length. It was used, for example in the RDF program (Lipman and Pearson, 1985) and later in other programs such as Bestfit (Devereux, 1989).

Given two sequences A and B, and the Smith–Waterman score $S(A, B)$, the method consists in per-

forming N comparisons between the first sequence A and N shuffled sequences from B, which yields the empirical mean score $\tilde{m}$ and the empirical standard deviation σ. The Z-value Z is then defined as:

$$Z(A,B) = \frac{S(A,B) - \tilde{m}}{\tilde{\sigma}}. \quad (1)$$

For this shuffling process, the 'ideal' number N of *shuffled sequences* is so large that the computation of the mean and standard deviation over all the possible shuffled sequences is not practically feasible. Moreover, the Z-value can depend on the choice of the shuffled sequence (A or B).

2.2. Materials and methods

All the sequences used in this study were obtained by anonymous ftp from the sites that maintain their respective genomic sequences databanks (a list of such sites can be obtained from, e.g., http://www.tigr.org/links).

The program LASSAP (Glémet and Codani, 1997) was used throughout to perform the Smith–Waterman protein sequences comparisons and to post-process their outputs. We consistently used the Dayhoff PAM250 matrix (Schwartz and Dayhoff, 1979) where all the elements had been divided by 3.3 with a gap open penalty $go = 5$ and a gap extension penalty $ge = 0.3$. Admittedly, the gap open value is rather high (corresponding to 16.5 with the original PAM250 matrix) and, with poorly related sequences, will lead to rather short segments of similarity containing few gaps. It is, however, our long-standing experience that the more commonly used value $go = 10$ (as compared to 16.5) is much too permissive and frequently leads to false (longer) alignments. This can be easily checked, for example, with proteins such as cysteinyl-tRNA synthetase and leucyl- or isoleucyl-tRNA synthetases that contain well-defined short segments of high similarity, embedded in long runs of highly divergent sequences. In addition, as stated above, our aim was to build intra- and inter-genomic clusters of proteins, and we wanted them to be robust, that is, containing only—as far as possible—truly related sequences. Our choice of local (Smith–Waterman) rather than global (Needleman–Wunsch) alignments has the same origin. Although Vogt et al. (1995) reported a *slightly* better performance for the global alignments, it is our constant experience that these are too often grossly erroneous when the overall similarity between sequences is weak, and that the biologically significant short segments of higher similarity are frequently missed. Finally, the use of modern similarity matrices such as the Gonnet matrix (Gonnet et al., 1992) or the BLOSUM series (Henikoff and Henikoff, 1992) has been reported to provide improved sensitivity in databank searches and produce more accurate alignments (Pearson, 1995; Vogt et al., 1995; Abagyan and Balatov, 1997). In the present case, however, the use of a high gap creation penalty makes it most probable that the choice of a particular matrix is not that important, as subsequently confirmed by the fact that the PAM250 and BLOSUM62 matrices produce similar results (see below).

2.2.1. Datasets

The present study made use of protein sequences from several genomes: *Escherichia coli* (4286 sequences), *Methanococcus jannaschii* (1735 sequences), *Haemophilus influenzae* (1680 sequences), *Synechocystis sp.* (3168 sequences) and *Saccharomyces cerevisiae* (6087 sequences). The Monte–Carlo process parametrization study has been performed using three sets from the *Saccharomyces cerevisiae* (Yeast) genome, as follows:

2.2.1.1. Real set. This set consisted of 1000 sequences extracted at random from Yeast. We performed all the possible pairwise alignments within this set, except the comparisons of the proteins against themselves. Since the Smith and Waterman algorithm is symmetrical, this resulted in 499,500 alignment scores. Then we calculated the Z-values for each pairwise alignment. As it is not known whether the Z-value has the same property of symmetry as the SW scores, each Z-value was calculated twice, first by shuffling 100 times one of the sequences in the pair and then by shuffling 100 times the other sequence.

2.2.1.2. Quasi-real set. The second set was composed of shuffled versions of the sequences from the real set. We performed the same comparisons as in the case of the real set and obtained two sets of SW-scores and Z-values.

2.2.1.3. High scores set. From all the pairwise Smith and Waterman alignments of the proteins from the yeast genome, we randomly retained 3000 of those pairs whose score S could potentially induce a significant Z-value (threshold arbitrarily set to $S > 30$).

For each of these alignments, we calculated the Z-values for 20, 50, 100, 200 and 500 randomizations. These different sets allowed to study the behavior of the Z-value as a function of the number of randomizations involved. We also calculated two independant Z-values obtained from 2000 and 5000 randomizations, respectively.

2.3. Monte–Carlo process parametrization

Since we use a Monte–Carlo process to estimate the

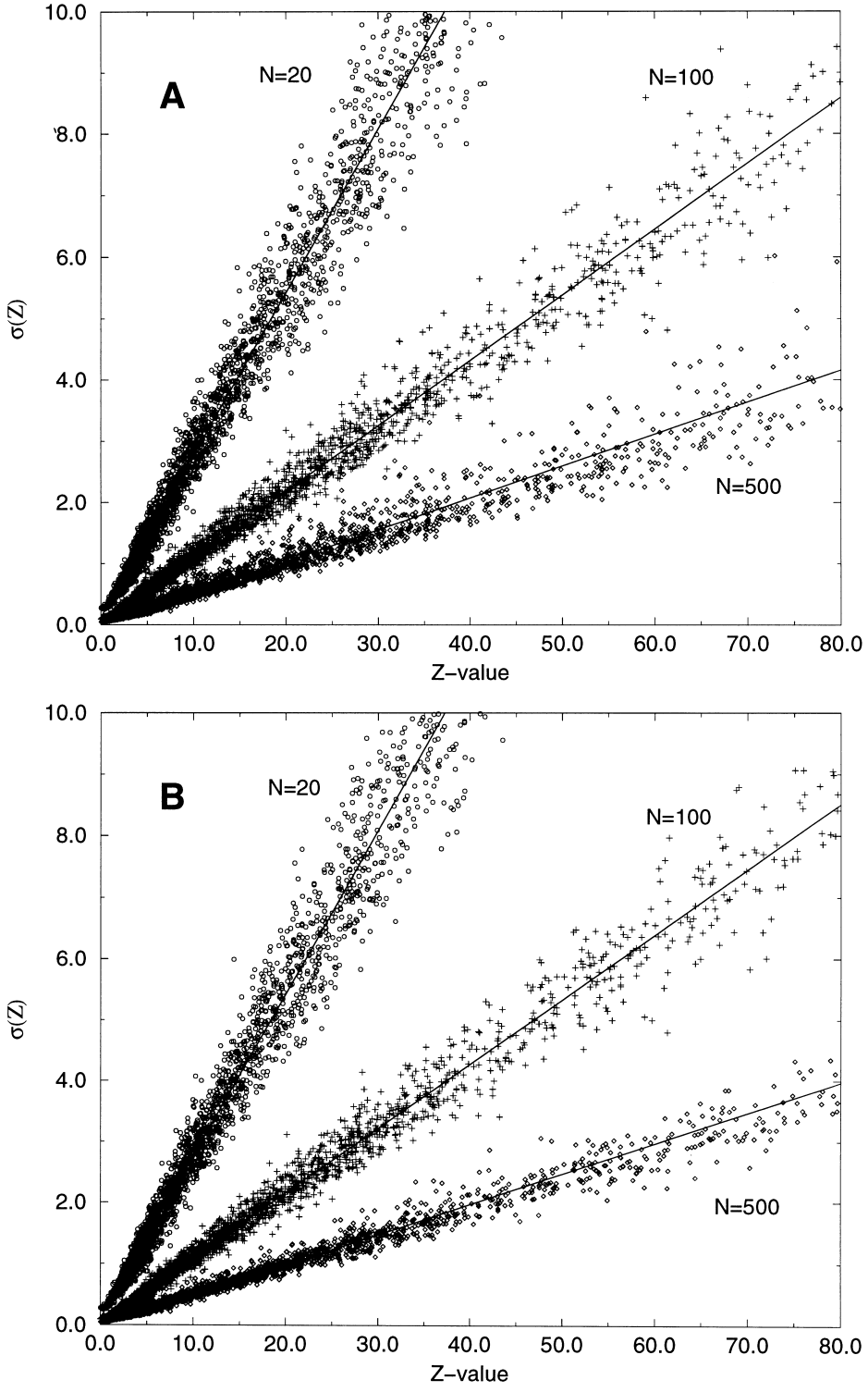


Fig. 1. Variation of $\tilde{\sigma}(Z_N(h_i))$ with Z_N: $N \in \{20,100,500\}$. The straight lines are the linear regressions between Z and the *estimated* standard deviation of $Z(\sigma(Z))$. For each N, and for each h_i (set of 3000 alignments), the reference Z-value has been computed with 2000 randomizations (A) and 5000 randomizations (B). For all these regressions the correlation coefficient is at least 0.98.

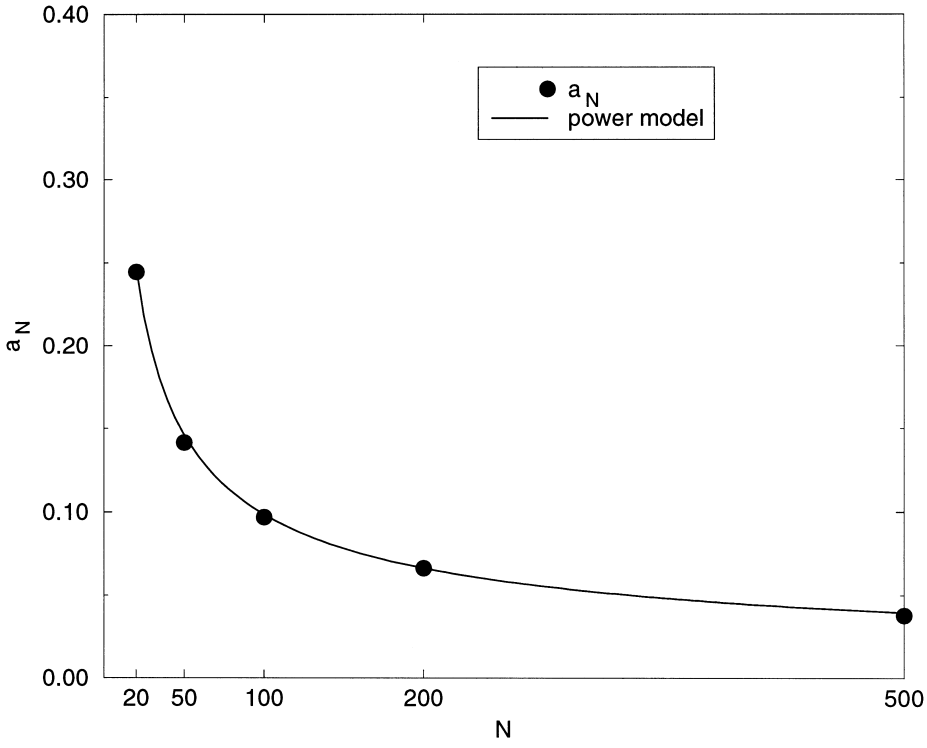


Fig. 2. Power Model: Determination of the coefficients of a global model of the estimated standard deviation of Z-scores. This figure presents the regressions over coefficients a_N (see Eq. 2) involved in the estimated standard deviation $\tilde{\sigma}(Z_N(h_i))$ for the five values $N = 20,50,100,200$ and 500.

Z-value, we introduce a deviation in our estimations around the *exact* Z-value calculated from all the possible sequence permutations (i.e., if we repeat the computation of a Z-value, we observe different values for each trial). Since computation time increases with N, our goal is to find a reasonable value of N that minimizes computation time while leading to Z-values with an acceptable variance (while N increases, the variance of Z decreases).

In order to evaluate the behavior of the variance, we chose 1000 real sequences among all those from the yeast genome, and computed all the SW-scores between all pairs of sequences and all the Z-values with different numbers of permutations. For the pair of sequences $h_i = (h_i^1, h_i^2)$ we computed the SW-score $S(h_i) = S(h_i^1, h_i^2)$, and several Z-values $Z_N(h_i)$ where N was the number of permutations.

Since we cannot compute the *exact* Z-value, we shall only consider an empirical variance $\tilde{\sigma}^2_{\text{observed}}$ defined for each alignment as:

$$\tilde{\sigma}^2_{\text{observed}}(Z_N(h_i)) = \frac{1}{100}\sum_{j=1}^{100}\left(Z^j_N(h_i^1,h_i^2) - \tilde{Z}(h_i^1,h_i^2)\right)^2$$

for $h_i = (h_i^1, h_i^2) \in H$ and $N = \{20,50,100,200,500\}$.

$\tilde{Z}(h_i^1, h_i^2)$ is the reference Z-value using 2000 and 5000 randomizations. As shown in Fig. 1, we observe a strong linear relation between $\tilde{\sigma}(Z_N(h_i))$ and $\tilde{Z}(h_i)$. For each of the five values of N (20, ..., 500), the linear regressions have correlation coefficients greater than 0.99. This computation strengthens the notion that the Monte–Carlo process is justified.

The standard deviation of the Z-value can therefore be estimated by a linear regression for each value of N. We approximate $\tilde{\sigma}(Z_N(h_i))$ by

$$\tilde{\sigma}_{\text{estimated}}(Z_N(h_i)) \approx a_N \cdot \tilde{Z}\left(h_i^1, h_i^2\right) \tag{2}$$

This new expression of the *estimated* standard deviation is not very useful since there is a strong relation between the coefficients a_N and N (see Fig. 1). By testing different regression schemes we chose a power dependence between a_N and N (see Fig. 2). We can express a_N as a function of N:

$$a_N = A \cdot N^B$$

where $A = 1.26$ and $B = -0.53$.

To test the validity of our final model, we calculated the correlation between the *observed* standard deviation of Z and the *estimated* standard deviation calcu-

Table 1
Verification of the model for the variance of Z-values (data from 3000 yeast sequences)[a]

N	20	50	100	200	500	Global
Correlation	0.993	0.992	0.995	0.995	0.992	0.993
delta	0.600	0.310	0.190	0.130	0.100	0.266

[a] Correlation and *delta* value between the *observed* standard deviation of Z and the *estimated* standard deviation calculated according to Eq. (2). N is the number of sequences shufflings.

lated according to Eq. (2). We also denote *delta* the deviation of the observed standard deviation from our model defined as the mean of the absolute difference between $\tilde{\sigma}_{\text{estimated}}$ and $\tilde{\sigma}_{\text{observed}}$. The results are summarized in Table 1.

This model was tested on the 1692 alignments with Smith and Waterman score greater than 30, obtained from the exhaustive comparisons of 578 sequences randomly extracted from *Methanococcus jannaschii*. These Z-values have been computed using $N = 30$, 100 and 300 randomizations. For each value of N, the observed standard deviation is calculated using 30 Z-values, and an independent reference Z-value, $\tilde{Z}$, computed with 2000 randomized sequences. We calculated the correlation coefficient (see Fig. 3A) and the delta index (see Fig. 3B) defined as above, Table 2 summarizes all these results.

Since, for a given value of Z, $\tilde{\sigma}_N$ is a decreasing function of N, formula (2) can be used to implement a method which computes a Z-value with an optimal number of shuffled sequences N. Indeed, for a desired variance $\tilde{\sigma}_d^2$, we can iteratively compute Z_N, by increasing N, until $\tilde{\sigma}_N^2 < \tilde{\sigma}_d^2$ (or $N > N_{\max}$). Note that this method can be refined since the expression (2) of the *estimated* standard deviation of the Z-value allows us to estimate the necessary number of shuffled sequences N_d, for a given variance $\tilde{\sigma}_d^2$ and a Z-value.

For a number of permutations equal to 100, the estimated standard deviation is about 0.1. This is an acceptable level leading to reasonable computing times. Hereafter we shall therefore compute all the Z-values with $N = 100$.

Table 2
Verification of the model for the variance of Z-values (data from 578 *Methanococcus Jannoschii* sequences)[a]

N	30	100	300	Global
Correlation	0.973	0.981	0.980	0.976
delta	0.420	0.230	0.120	0.257

[a] Correlation and *delta* value between the *observed* standard deviation of Z and the *estimated* standard deviation calculated according to Eq. (2). N is the number of sequences shufflings.

2.4. Asymmetry of the Z-value estimation

For the computation of a Z-value between two sequences A and B according to formula (1), only one of the two sequences is generally shuffled. However, shuffling the first sequence ($Z(A, B)$) can lead to Z-values that are markedly different from those obtained by shuffling the second ($Z(B, A)$). This is illustrated in Fig. 4. In this example, one of the sequences (YBR086C) has a normal amino acid composition while the second (YKR092C) is exceptionally rich in serine (48%). As a consequence, shuffling the sequences YBR086C and YKR092C leads to Z-values of 11.5 and 2.7, respectively.

A systematic study of the differences between $Z(A, B)$ and $Z(B, A)$, calculated on a great number of sequence pairs from various genomes, indicated that large differences are far from seldom. We therefore used systematically a *conservative* approach, taking a new formula for the Z-value:

$$Z'(A,B) = \min(Z(A,B),Z(B,A)) \tag{3}$$

This allowed to eliminate high Z-values resulting artificially from sequences of abnormal amino acid composition.

Note: hereafter we will always refer to this formulation of the Z-value in our computations.

2.5. Length dependency

It is well known that the longer the sequences, the higher the SW-scores (Waterman and Vingron, 1994). As a consequence, two long sequences which are not biologically related can have a higher score than two short biologically-related sequences. Therefore, one cannot set easily a cut-off value in order to decide whether an alignment is significant or not.

Since the Z-value is essentially the number of standard deviations exceeding the mean of scores from sequences with the same amino acid composition and length, the Z-value decreases the bias due to the sequence length. In Fig. 5 are depicted the relationships between the lengths of the sequences and the scores (SW-scores and Z-values), based on 1.5 millions pairwise comparisons on the proteins from the *M. jannaschii* genome.

It shows clearly that the Z-values are much less dependent on the lengths of the sequences than are the SW-scores.

3. Statistical analysis of the distribution of *Z*-values

The aim of this study is to find a law of probability that the experimental *Z*-values will follow. The idea is that if we can compute a probability, then we can infer a dissimilarity index between two sequences. In such a case, we can build families of related sequences, and apply classification algorithms on each of them.

A preliminary study indicated that the scores and the *Z*-values seem to follow the extreme value distribution (type *I*), more precisely the Gumbel distribution, one of the three possible distributions of extreme values (Johnson and Kotz, 1970; Gumbel, 1958). This has to be correlated with studies by Karlin, Altschul and coworkers (Karlin and Altschul, 1990; Karlin et al., 1990; Altschul et al., 1990), which have shown that the distribution of BLAST scores for sequences of independent identically distributed letters follows the Extreme Value Distribution (EVD, type *I*). Briefly, for two random sequences $A = a_1a_2,\ldots,a_n$ and $B = b_1b_2,\ldots,b_m$, given the distribution of individual residues, and given a scoring matrix, the probability of finding an ungapped segment pair with a score greater than or equal to s is:

$$P(X \geq s) = 1 - \exp(-K \cdot m \cdot n \cdot e^{-\lambda s})$$

where λ and K can be calculated from the scoring matrix and sequence compositions.

In order to determine whether the *Z*-values follow the same law, we have to find two parameters, the characteristic value θ and the decay value ξ, such as:

$$P(Z \geq z) = 1 - \exp\left(-e^{-(z-\xi)/\theta}\right)$$

where z is the observed *Z*-value.

For θ and ξ two estimators exist:

- *Estimators based on sample moments* ($\tilde{\theta}$ *and* $\tilde{\xi}$): Let μ and σ be the empirical mean and standard deviation of the sample. These estimators are simply obtained from the following formulas (Johnson and Kotz, 1970):

$$\tilde{\theta} = \left(\frac{\sqrt{6}}{\pi}\right)\sigma$$

$$\tilde{\xi} = \mu - \gamma\,\tilde{\theta}$$

  where γ is the Euler's constant: $\gamma = 0.5772$.
- *Maximum Likelihood Estimators* ($\hat{\theta}$ *and* $\hat{\xi}$): θ is the solution of the following equation (Johnson and Kotz, 1970):

$$\hat{\theta} - \mu + \left[\sum_{j=1}^{n} X_j e^{-X_j/\hat{\theta}}\right] \times \left[\sum_{j=1}^{n} e^{-X_j/\hat{\theta}}\right]^{-1} = 0. \quad (4)$$

$$\hat{\xi} = \mu - \gamma\,\hat{\theta} \quad (5)$$

  $\hat{\theta}$ and $\hat{\xi}$ are biased estimators of θ and ξ, respectively (Johnson and Kotz, 1970), but better than the sample moments estimators. For calculating these estimators, we initialize the maximum likelihood estimator θ with the value of the estimator based on sample moments θ, and we use a Newton–Raphson procedure to find the root of Eq. (4). Then ξ can be calculated.

The two parameters θ and ξ, have been estimated with the Maximum Likelihood Estimators (Johnson and Kotz, 1970), using large datasets of real sequences extracted from the yeast genome (Real set: 499500 *Z*-values) and *quasi-real* ones (quasi-real set: 499500 *Z*-values). We then performed χ^2 and Kolmogorov–Smirnov tests on both the distributions of SW-scores and *Z*-values as well as on many subsets of SW-scores. In the case of SW-scores, the hypothesis of fitting the extreme value distribution was consistently rejected, due to the length dependancy of the Smith–Waterman scores.

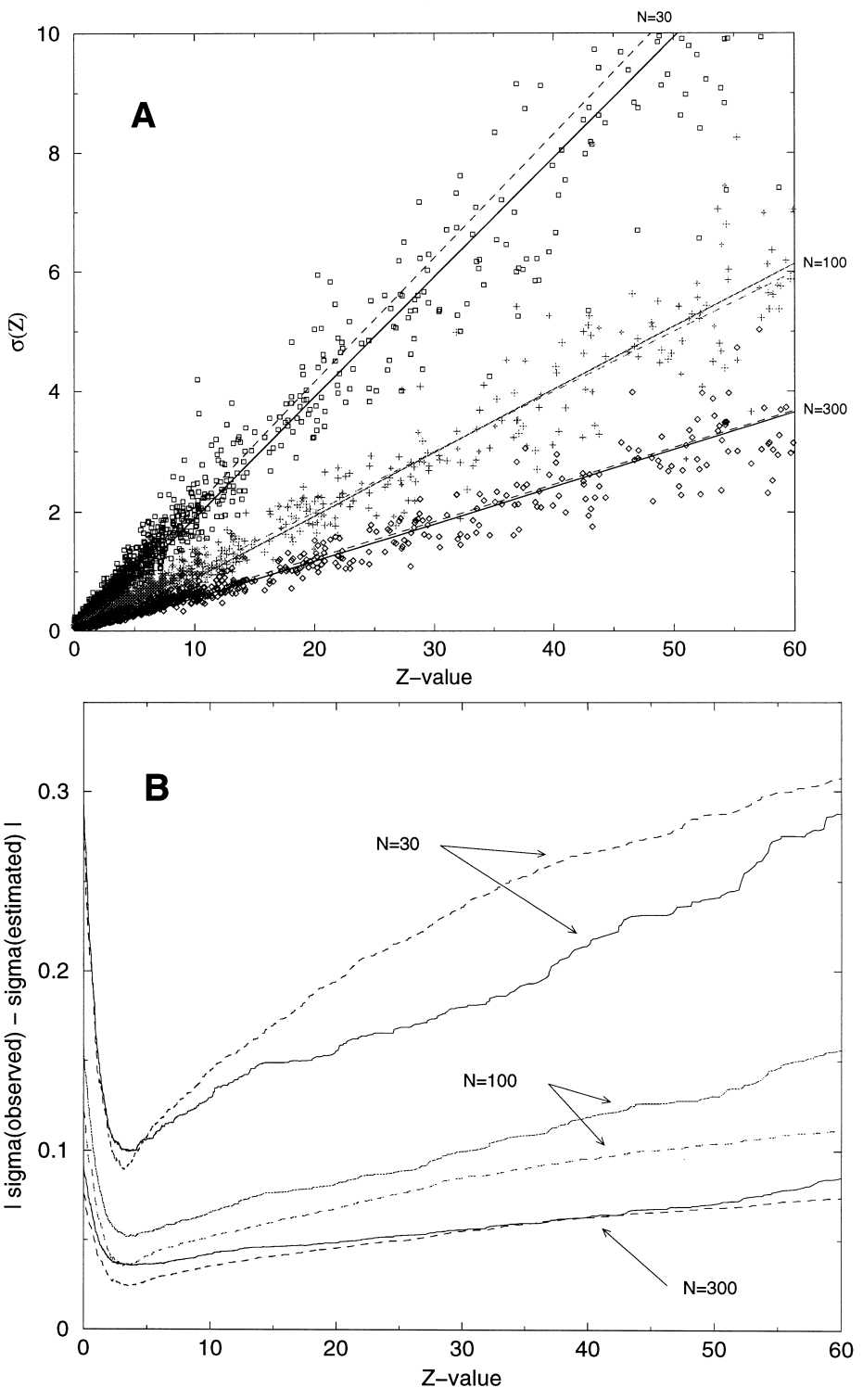
The behavior of the *Z*-values is different and we observed that:

- in the case of 'quasi-real' sequences, the EVD model is a good estimation of the observed distribution (Fig. 6A).
- in contrast, for real protein sequences, the EVD model is not satisfactory for high *Z*-values, in which case there are about 1 out of 1000 over-represented *Z*-values (Fig. 6B). This over-representation of high *Z*-values can lead to wrong values of their significance (i.e., the probability $P(Z \geq z_0)$ that one could obtain a *Z*-value greater than or equal to a value z_0).

3.1. Cutoff values

The *Z*-value distribution curve for real sequences diverges from the curve for '*quasi-real*' *Z*-values beyond a certain *Z*-value *c*. This means that *Z*-values above *c* are not obtained by chance. This value, *c*, will be called the *cutoff value*. It is shown in Fig. 7 that we can adopt the value 8.0 as a conservative estimation of the cutoff.

This purely formal conclusion has an obvious biological interpretation. Real protein sequences result from evolution where gene duplications, mutations,



(*Caption opposite*).

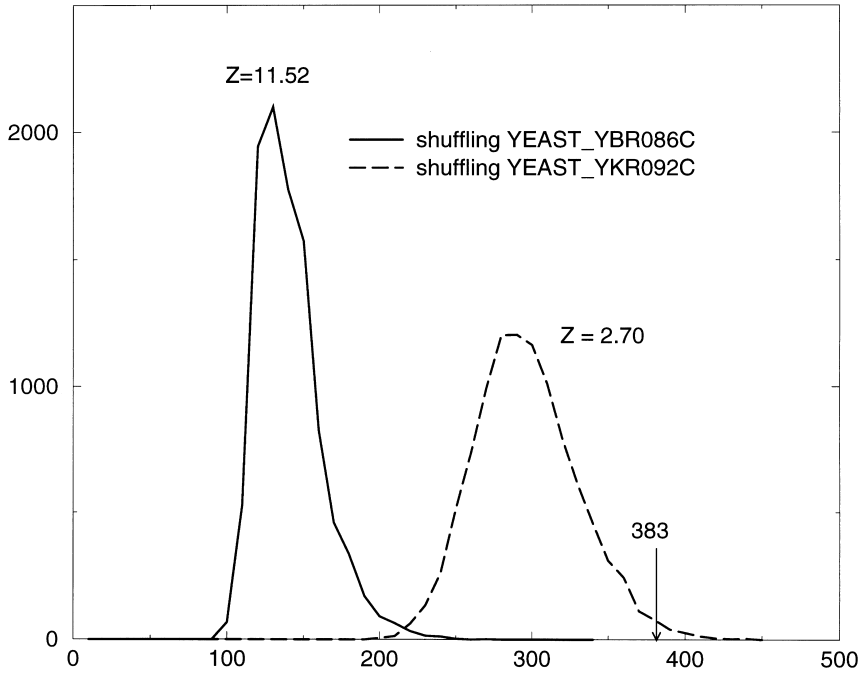


Fig. 4. Distributions of Smith–Waterman scores between YEAST_YBR086C and YEAST_YKR092C upon shuffling one of the two sequences. The initial score is 383 (matrix PAM250, gapo = 16.5, gape = 1). Shuffling YEAST_YKR092C leads to a Z-value $Z_1 = 2.70$, shuffling YEAST_YBR086C leads to a Z-value $Z_2 = 11.52$. In both cases 10,000 shuffling processes have been performed.

fusions and recombinations take place continuously as major forces conserving sequence similarities and generating sequence diversities. It should be kept in mind that the real protein sequences, those that do exist actually, and those that did exist during life's history, represent an infinitely small fraction of all the possible random permutations of an average length of 300 with 20 different amino acids (20^{300}). The real protein space is a microcosm within the macrocosm of quasi-real sequence space.

Figs. 6 and 7 also show that the EVD model is adequate for Z-values within the twilight zone ($5<Z<8$). It may be noted that, although the twilight zone represents a small fraction of all the pairwise alignments, the fraction of proteins involved in it may be quite large. For example, the all vs. all comparisons of the 1735 sequences from the *M. jannaschii* genome produced 1504245 pairs. Among them, 1500153 pairs had a Z-value lower than 5 and 1295 pairs a Z-value greater than 8. Thus, only 2797 pairs (0.18%) fell within the twilight zone. However, 1583 sequences (as much as 91% of the 1735 genomic sequences) were involved in those pairs.

Fig. 3. Experimental and calculated variance of the Z-values. (A) From a set of 578 sequences extracted from *M. jannaschii*, all the pairwise Z-values were calculated 30 times using $N = 30{,}100$ and 300 sequence shufflings. For each sequence pair and for each value of N, the 'observed' standard deviation of Z was obtained from these 30 values and are reported in the figure (squares, crosses and circles for $N = 30{,}100$ and 300, respectively). The solid lines represent the linear regressions over the experimental points. Using Eq. 2, we then obtained directly the 'estimated' standard deviation for each point in the set $N = 100$ using the reference Z-value from the yeast training set. The 'estimated' standard deviations in the set $N = 30$ (respectively $N = 300$) were obtained by interpolation from the values calculated with $N = 20$ and $N = 50$ (respectively $N = 200$ and $N = 500$), using the same reference Z-values as above. The dotted lines represent the linear regressions over the estimated values. The close correspondence between the observed and estimated regression lines indicate that the model given by Eq. 2 is adequate. (B) Variation of the mean difference between the observed and the estimated variance of Z for all Z-values between zero and Z, as a function of Z. Data are obtained from 1000 sequences from Yeast (dotted lines) and 578 sequences from *M. jannaschii* (solid lines). One observes only a slight variation of the differences as Z increases, which indicates that the model (Eq. (2)) is practically independent of Z.

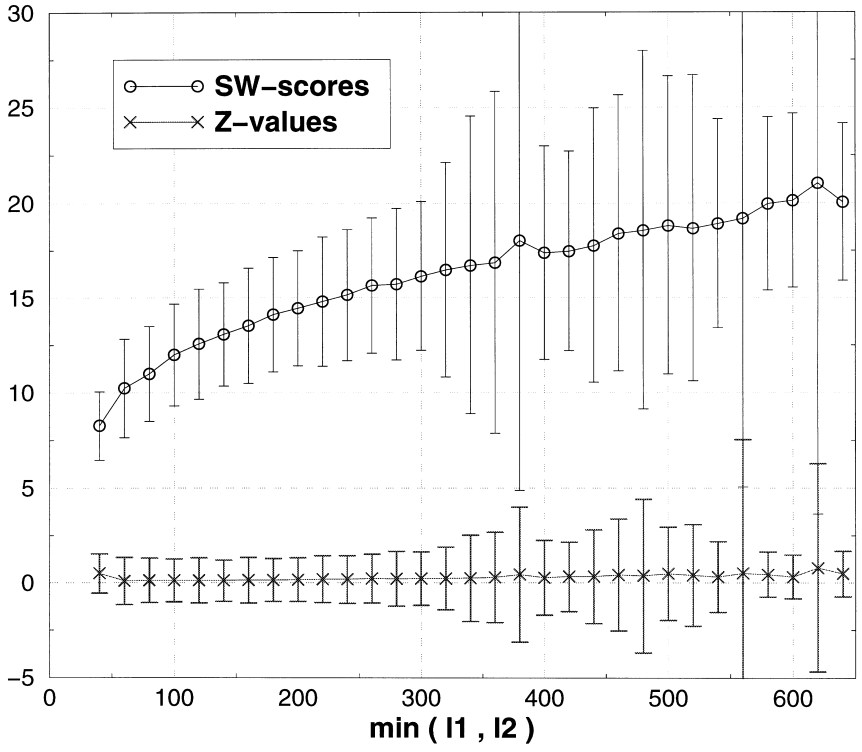


Fig. 5. Length dependency. Influence of the lengths of the sequences on the SW-scores and the Z-values. All the possible pairwise alignments were performed on all the protein sequences from *Methanococcus jannaschii*. All the SW-scores and Z-values were calculated and classified by intervals of the smallest sequence length in each pair. Within each interval of size 20, we calculated the mean and the standard deviation of the SW-scores and Z-values, which are reported in the figure. It is clear that the SW-scores increase together with the lengths of the sequences and have a high variance. In contrast, the Z-values are practically independent of the sequence lengths.

3.2. Law of high Z-values distribution

As explained before, one aim of this work was to find whether the Z-values follow a known probability distribution and, if so, to build families of related sequences by a single linkage algorithm. In most cases, it will be desirable to build 'robust' families, that is, to use a threshold Z-value beyond the twilight zone. Hence, we now try to estimate the law of the Z-values distribution for Z-values greater than 8.

For most of the very high Z-values, the sequences involved are strongly related (e.g., more than 80% of identities on their whole lengths). From a statistical point of view, these values represent the tail of the Z-values distribution. If one wants to estimate a law of distribution for (or from) these values, it must be kept in mind that this sample is biased, and therefore, estimating methods must be used carefully.

In our case, we considered the Z-values in the range [8, 50]. First, in the same way as (Waterman and Vingron, 1994), we used the probability integral transform in order to estimate the ξ and θ parameters of an EVD fitting these Z-values. The results are satisfactory (see Table 3). In a second run, we used linear regression techniques for fitting the distribution curve in the range [8, 50]. In that case, the retained model is the Pareto law (Zajdenweber, 1996). The density function of the Pareto law is

$$f(z) = A \cdot z^{-(1+\alpha)} \quad \text{if } z \geq 8.0$$

$$f(z) = 0 \quad \text{otherwise}$$

with $\alpha \geq 0$

The coefficient A is just a normalisation coefficient and is not informative. α is called the Pareto index.

Table 2 displays the estimated parameters when using a Gumbel model and a Pareto model, for five complete microbial genomes, as well as the genomes taken all together. One can observe that for both models, the estimated parameters are independent of the genome size and of the similarity matrix used in

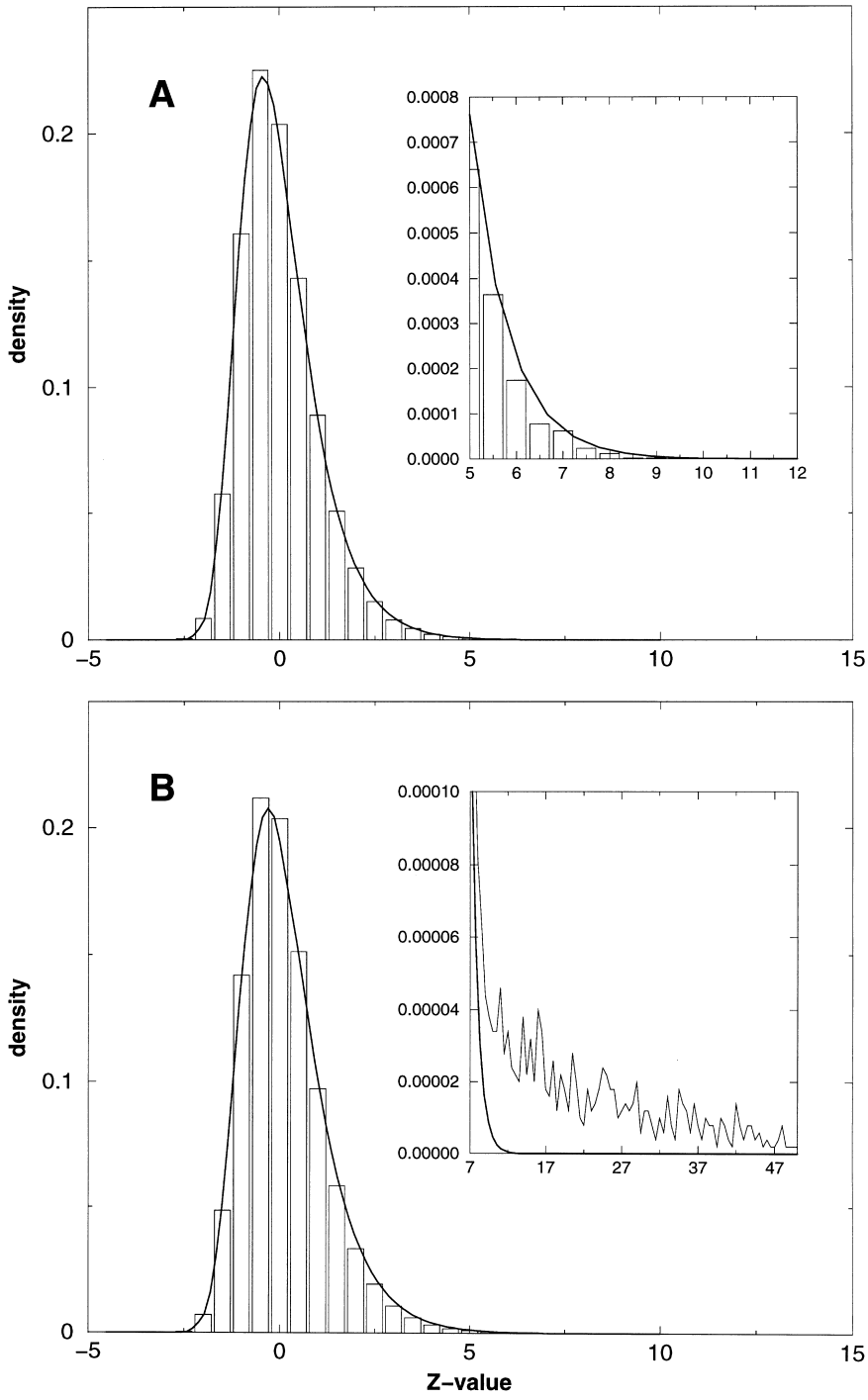


Fig. 6. Distribution of Z-values: (A) empirical distribution (rectangles) and Gumbel model (solide line) for quasi-real sequences. (Insert) the Gumbel model fits the experimental distribution for high Z-values. (B) empirical and Gumbel model for real sequences. (Insert) the Gumbel model (thick line) does not fit the experimental distribution (thin line) for high Z-values.

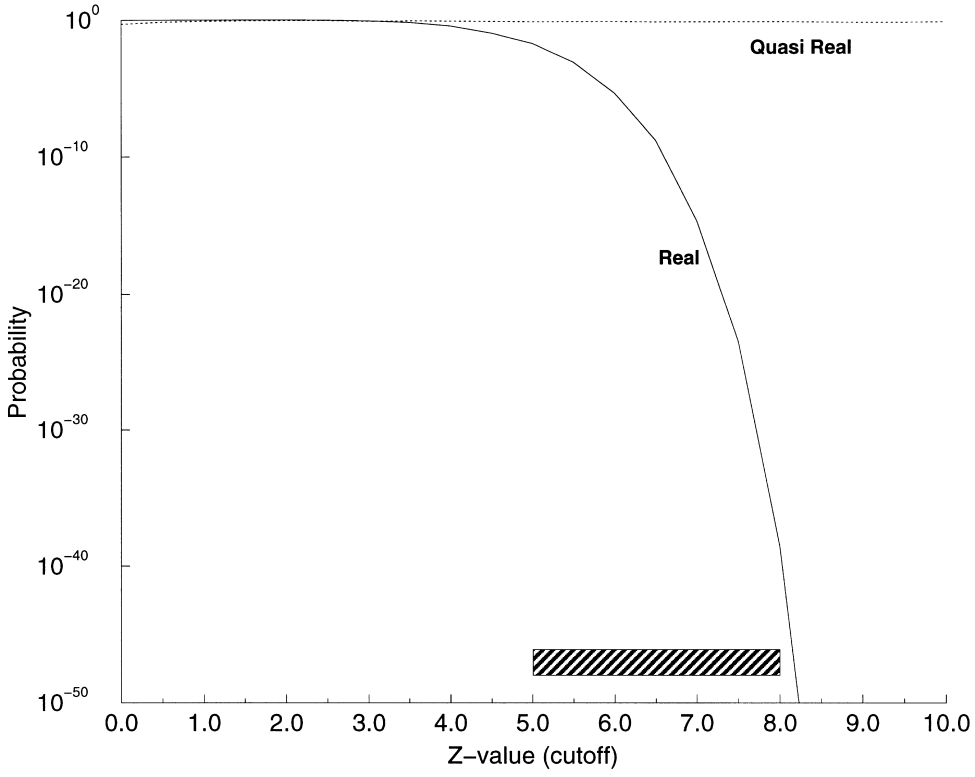


Fig. 7. Cutoff value: Estimation of the cutoff value for splitting the EVD-like Z-values from high Z-values. Let $X \sim B(N,p_c)$ be a binomial variable, where N is the number of observed Z-values and p_c the probability that the EVD variable Z is greater or equal to c. X is the expected number of Z-values greater or equal to c. This figure shows the variation of the probability $P(X > N^c_{\text{obs}})$ as a function of c, where N^c_{obs} is the observed number of Z-values greater than c. The observed distribution for real protein sequences diverges from the EVD for $c > 5.0$ and the probability becomes practically zero at 8.0. This study has been carried out for both the *Haemophilus* and *Methanococcus* genomes and the results are basically the same. Solid line: real set; Dotted-line: quasi-real set (see Section 2.2).

Table 3
Parameters of the curves fitting the tail of the Z-value distribution[a]

	Number of pairwise comparisons	Pareto α	χ^2	EVD		χ^2
				ξ	θ	
Subset of *S. cerevisiæ*	499500	1.20	28.66	−122.687	19.938	27.83
S. cerevisiæ	18,522,741	0.90	17.66	−162.46	23.8	30.00
Escherichia coli	9,182,755	1.26	18.33	−119.889	18.501	15.33
Haemophilus influenzae						
PAM250, go = 5, ge = 0.3	1,410,360	1.63	25.66	−93.436	14.448	12.66
BLOSUM62, go = 10, ge = 0.6	1,410,360	1.66	44.66	−90.571	14.392	13.00
Methanococcus jannaschii	1,504,245	1.66	23.33	−127.938	18.682	30.33
Synechocystis	5,016,528	1.05	17.66	−135.259	21.498	20.33
all vs. all	143,744,490	1.16	11.00	−136.921	19.085	23.66

[a] This table shows that both the Pareto and Gumbel laws are good models for high Z-values, whatever the size of the genome. All the indices have been computed using the PAM250 matrix divided by 3.3, gap open = 5, gap extend = 0.3. The *Haemophilus influenzae* genome has been recomputed using the BLOSUM62 matrix, gap open = 10, gap extend = 0.6. Based on the relative entropies of the PAM250 and BLOSUM62 matrices (Henikoff and Henikoff, 1992) the two gap open values are equivalent.

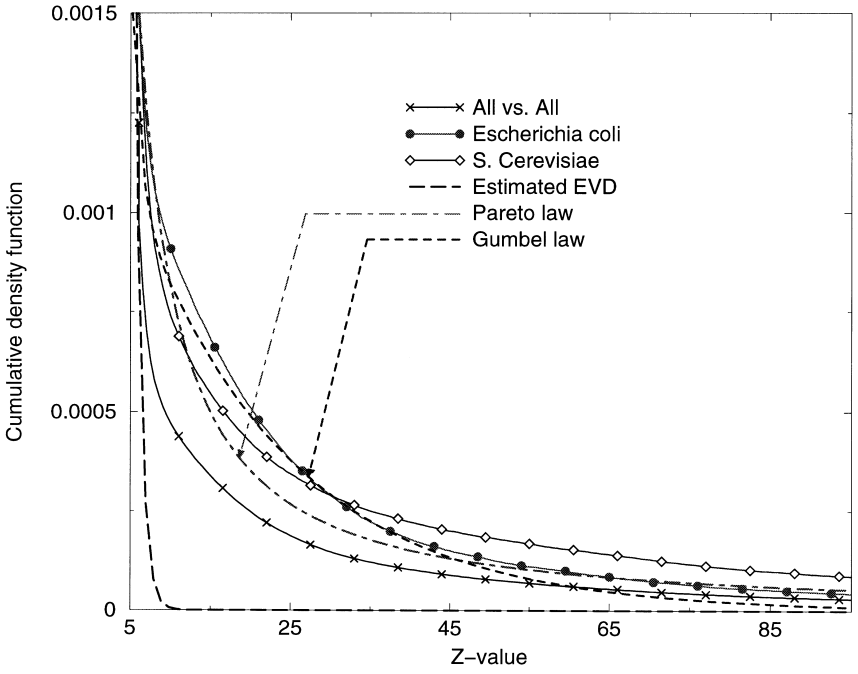


Fig. 8. Density of Z-values: For the five genomes studied here, the Z-value density has a non-negligible tail that differs from the Gumbel distribution. The observed distributions of two genomes (*E. coli* and *S. cerevisiæ*) are shown here, as well as the observed distribution of the five genomes taken all together (All vs. All curve). These distributions are similar and can be fitted by a Pareto law. The Pareto index α is taken as the mean of estimates for five genomes (see Table 2). One can see again that the EVD strongly deviates from the experimentally observed one.

the alignments. In Fig. 8 the experimental distribution of the Z-values together with the Pareto curve are displayed. An interesting feature can be noticed in this figure, namely the distribution curves for *E. coli* and *S. cerevisiae* crosses at about $Z = 30$ where the yeast distribution jumps over that of *E. coli*. This suggests that the yeast clusters are more robust or, in other words, that the proportion of highly similar paralogous sequences is greater in yeast than in *E. coli*. This is indeed what is observed, as shown in Fig. 9. For a threshold Z-value of 100, the proportion of sequences that are still grouped into clusters is 9 and 18% for *E. coli* and yeast, respectively. We have no clear explanation at the moment for this observation and further work is clearly needed.

4. Discussion

The results presented here show that Z-values can be used to estimate probabilities for gapped alignments of real protein sequences.

Once a probability has been calculated for a pairwise alignment, one can induce a dissimilarity index between two sequences. It is therefore possible to build clusters of related sequences with different probability thresholds, and apply classification algorithms on each of them. That is, the sequences can be grouped in 'connective clusters' by a transitive closure algorithm such that in any given connective cluster, any sequence shares a Z-value greater than a given threshold (or shares a probability lower than a given threshold) with at least another sequence in the same cluster. Indeed, it has been shown recently (Gerstein, 1998; Park et al., 1997) that the sensitivity of sequence comparisons is improved by transitive sequence matching techniques as compared with more classical direct matching methods (see also Tatusov et al., 1997; Codani et al., 1999). Using LASSAP (Glémet and Codani, 1997), the computation as well as the analysis of large sets of sequence data can be performed efficiently. Therefore, complete intra- and inter-genome comparisons and classifications can be carried out as soon as genomes are sequenced which can lead to intriguing and challenging biological implications (Slonimski et al., 1998).

While it is beyond the scope of this paper to give detailed results on genomic sequence comparisons, it may be of interest to provide some observations about the clusters that have been built by the Z-value method. One often asked question is 'are there many false positives or false negatives'? As far as we can see, the answer is 'no'. For example, the study of aminoa-

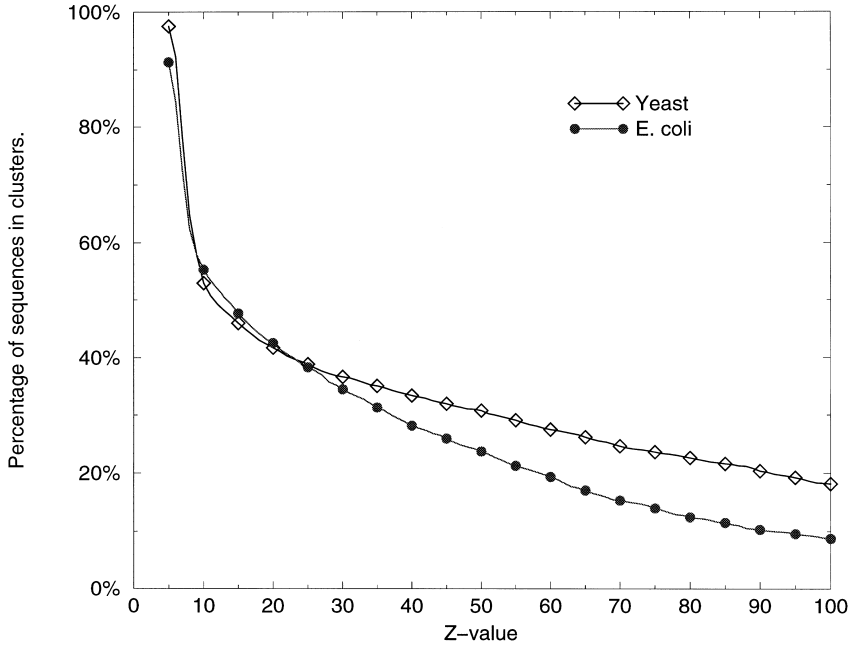


Fig. 9. Proportion of sequences (total number of sequences in clusters divided by the total number of coding sequences in the genome) that are included into clusters in Yeast and *E. coli* as a function of the threshold *Z*-value. The proportion of highly similar paralogous sequences in Yeast is clearly greater than in *E. coli*.

cyl-tRNA synthetases (aaRS) can be considered a test case since these proteins are well known for being extremely diverse, yet sharing some well defined characteristics. Here we can compare our results (Diaz-Lazcoz et al., 1998; Aude et al., this series) with those of Tatusov et al. (1997) see also http://ncbi.nlm.nih.-gov/COG. These authors built clusters (called COGs) on the basis of pairwise Blast scores (as opposed to SW *Z*-values). The results for aaRS are rather different: (i) Tatusov et al. obtained 21 different COGs while we got only 10 clusters; (ii) five of our clusters grouped synthetases with different specificities, that is, (Cys, Ile, Leu, Met, Val), (Gly, His, Ser, Pro, Thr), (Asn, Asp, Lys), (Trp, Tyr) and (Gln, Glu) while only two COGs are composite: (Asn, Asp, Lys) and (Gln, Glu); (iii) the PheRS alpha and beta chains are grouped into one single cluster together with the monomeric yeast mitochondrial PheRS while the last protein belong to the PheRS alpha COG, the PheRS beta chains making another COG; (iv) all the Asn-, Asp- and LysRS are grouped into one single cluster while the AspRS are split into two different COGs. These differences most probably arise from the fact that the sets of proteins were different in the two studies (we used much more aaRS sequences, hence we benefited from more pairwise links). It is also probable that the Blast score threshold used by Tatusov et al. is more stringent than our *Z*-value threshold that was set to 14. In any case, the clusters built by Diaz-Lazcoz et al. from 465 aminoacyl-tRNA sequences do not show any false positive. In addition, all the groupings of heterologous aaRS in one cluster (such as Cys, Ile, Leu, Met, Val) are totally consistent with our current biochemical knowledge of these proteins.

Another clustering procedure has been used by Teichmann et al. (see http://www.mrc-lmb.cam.a-c.uk:80/genomes/MG_fams.html). Here the clusters resulting from various inter- and intra-genomic comparisons were built on the basis of pairwise Smith–Waterman *E*-values. Then a thorough study of multiple domains or duplication modules was undertaken (Park and Teichmann, 1998), leading to the creation of new clusters containing only related duplication modules. This procedure is obviously efficient and sensitive since, in the case of *M. genitalium* aminoacyl- tRNA synthetases, all the class I aaRS were grouped together into one single cluster through their catalytic domain. We think, however, that splitting sequences into different clusters according to their modular nature may lead to a loss of information. For example, such a procedure would fail to show immediately that the monomeric yeast phenylalanyl-tRNA synthetase is a composite of the alpha- and beta-chains of the standard dimeric PheRS (Diaz-Lazcoz et al., 1998).

References

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

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.

Codani, J.-J., et al., 1999. Automatic analysis of large scale pairwise alignments of protein sequences. In: Methods in Microbiology, vol. 28. Academic Press, New York.

Devereux, J., 1989. The GCG sequence analysis software package. Package, Version 6.0, Genetics Computer Group, Inc., University Research Park, 575 Science Drive, Suite B, Madison, Wisconsin, 53711, USA.

Diaz-Lazcoz, Y., et al., 1998. Evolution of genes, evolution of species: the case of aminoacyl-tRNA synthetases. Mol. Biol. Evol 15(11), 1548–1561.

Gerstein, M., 1998. Measurement of the effectiveness of transitive sequence comparison, through a third intermediate sequence. Bioinformatics 14, 707–714.

Glémet, E., Codani, J.-J., 1997. LASSAP: a large scale sequence comparison package. Comp. Appl. BioSci 13 (2), 137–143.

Gonnet, G.H., Cohen, M.A., Benner, S.A., 1992. Exhaustive matching of the entire protein sequence database. Science 256, 1433–1445.

Gumbel, E.J., 1958. Statistics of Extremes. Columbia University Press, Columbia.

Henikoff, S., Henikoff, J.G., 1992. Amino acid substitution matrices from protein blocks. Proc. Natl. Acad. Sci. USA 89, 10915–10919.

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

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., 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 comparaison of several similarity indices used in the classification of protein sequences: a multivariate analysis. Nucl. Acids. Res 20 (14), 3631–3637.

Lipman, D.J., Pearson, W.R., 1985. Rapid and sensitive protein similarity searches. Science 227, 1435–1441.

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.

Miller, W., Huang, X., 1991. A time-efficient, linear-space local similarity algorithm. Adv. Appl. Math 12, 337–357.

Mott, R., 1992. Maximum-likelihood estimation of the statistical distribution of Smith–Waterman local sequence similarity scores. Bull. Math. Biol 54 (1), 59–75.

Needleman, S.B., Wunsch, C.D., 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. Mol. Biol 48, 443–453.

Park, J., Teichmann, S.A., Hubbard, T., Chothia, C., 1997. Intermediate sequences increase the detection of homology between sequences. J. Mol. Biol 273, 349–354.

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.

Pearson, W.R., 1995. Comparison of methods for searching protein sequence databases. Protein Sci 4, 1145–1160.

Schwartz, R.M., Dayhoff, M.O., 1979. Matrices for detecting distant relationships. vol. 5, suppl 3, Nat. Biomed. Res. Found. Washington DC, Washington, pp. 353–358.

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

Slonimski, P.P., et al., 1998. The first laws of genomics. Comparative and Microbial Genomics 3 (2), 46.

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

Tatusov, R.L., Koonin, E.V., Lipman, D.J., 1997. A genomic perspective on protein families. Science 278, 631–637.

Vogt, G., Etzold, T., Argos, P., 1995. An assessment of amino acid exchange matrices in aligning protein sequences: the twilight zone revisited. J. Mol. Biol 249, 816–831.

Waterman, M.S., Eggert, M., 1987. A new algorithm for best subsequence alignments with application to tRNA-tRNA comparisons. J. Mol. Biol 197, 723–728.

Waterman, M.S., Vingron, M., 1994. Sequence comparison significance and poisson approximation. Statistical Science 9 (3), 367–381.

Zajdenweber, D., 1996. Extreme value in business interruption insurance. J. Risk and Insurance 63, 95–110.