BRIEFINGS IN BIOINFORMATICS. VOL 7. NO 1. 2–24 doi:10.1093/bib/bbk001

Statistical significance in biological sequence analysis

Alexander Yu. Mitrophanov and Mark Borodovsky

Submitted: 11th July 2005; Received (in revised form): 24th November 2005

Abstract

One of the major goals of computational sequence analysis is to find sequence similarities, which could serve as evidence of structural and functional conservation, as well as of evolutionary relations among the sequences. Since the degree of similarity is usually assessed by the sequence alignment score, it is necessary to know if a score is high enough to indicate a biologically interesting alignment. A powerful approach to defining score cutoffs is based on the evaluation of the statistical significance of alignments. The statistical significance of an alignment score is frequently assessed by its *P*-value, which is the probability that this score or a higher one can occur simply by chance, given the probabilistic models for the sequences. In this review we discuss the general role of *P*-value estimation in sequence analysis, and give a description of theoretical methods and computational approaches to the estimation of statistical signifiance for important classes of sequence analysis problems. In particular, we concentrate on the *P*-value estimation techniques for single sequence studies (both score-based and score-free), global and local pairwise sequence alignments, multiple alignments, sequence-to-profile alignments and alignments built with hidden Markov models. We anticipate that the review will be useful both to researchers professionally working in bioinformatics as well as to biomedical scientists interested in using contemporary methods of DNA and protein sequence analysis.

Keywords: *sequence analysis; pairwise alignment; multiple alignment; profile; probabilistic model; statistical significance; P-value; E-value*

INTRODUCTION

Computational methods of biological sequence analysis have become an indispensable part of the modern scientist's research arsenal. In protein studies, the results of sequence similarity searches in databases help generate reasonable hypotheses concerning structural and functional properties of proteins, as well as their evolutionary relationships. On the DNA level, sequence analysis techniques make it possible to identify genes and functional elements in newly sequenced genomes. The abundance of biomolecular sequence information (generated as a result of the ever-increasing number of large-scale sequencing projects), together with a relatively high cost of 'wet lab' experimentation, calls for powerful and efficient computational tools as primary means for high-throughput genomic proteomic investigations. Since sequence comparison and motif analysis methods can be used to predict protein–protein interactions [1, 2] and interactions in transcriptional regularity networks [3], computational sequence analysis also provides a basis for the rapidly developing field of systems biology.

It thus comes as no surprise that searching a biosequence database is probably the first thing a biologist would do once a biological sequence of interest is known. However, the evolution of DNA and proteins in living organisms is influenced by a number of random factors, and observed amino acid or nucleotide patterns may result from the action of these factors, rather than from the selective pressure

Corresponding author. Mark Borodovsky, School of Biology, Georgia Institute of Technology, Atlanta, GA 30332–0230, USA. Tel: +1 404 894 8432; Fax: +1 404 894 0519; E-mail: mark.borodovsky@biology.gatech.edu

Alexander Yu. Mitrophanov is a postdoctoral fellow at the School of Biology, Georgia Institute of Technology. His research interests include applications of probabilistic methods in different areas of bioinformatics and computational biology.

Mark Borodovsky is a Regents' Professor at the School of Biology, Georgia Institute of Technology, and the Wallace H. Coulter Department of Biomedical Engineering at Georgia Institute of Technology and Emory University. His research interests include development of statistical methods for biological sequence analysis and identifying functionally important features of DNA and proteins in the context of cell function and evolution.

maintaining a certain function. Therefore, not all database sequences which resemble the query sequence may be its true homologues. Naturally enough, the biologist is interested in finding biologically relevant targets, namely, those sharing important structural and functional characteristics with the query. The question arises how likely it is that a particular observation of high sequence similarity, a hit, has occurred by chance. This is the starting point for exploring the role of statistical significance in computational analysis of biological sequences.

The strength of an alignment is usually determined by its score, and the statistical significance of the score is assessed by the *P*-value. The term '*P*-value' of an alignment designates the probability of an alignment with this score or higher occurring by chance alone. The exact meaning of 'chance' depends on the method of modeling randomness in the *P*-value estimation procedures. Originally, works on statistical significance analysis [4–8] considered randomly shuffled sequences with preserved compositional properties. Alternatively, sequences can be modeled as sample paths of stochastic models whose parameters are derived from the observed compositional statistics of the real sequences. Such models have the advantages of being analytically tractable in simple settings. This tractability leads to explicit formulas connecting the score distribution to the nucleotide or amino acid composition of the sequence and the parameters of the scoring system, thus simplifying and accelerating the assessment of statistical significance in many practically important situations.

The general fact is that the procedure for determining the *P*-value depends on the sequence analysis problem being solved. Therefore, different types of alignments (global versus local, pairwise versus multiple) require the development of the corresponding *P*-value estimation methods and algorithms. Our goal is to outline some of the important accomplishments in this area. After a broad discussion of the notion of the *P*-value in the context of sequence analysis, we give a description of problems that require a stochastic model for only one sequence. In a sense, this is the simplest class of sequence analysis problems, but it can be viewed as an example illustrating the general difficulties related to the *P*-value estimation. We see that some important results on score distributions obtained for this class can be carried over to the more intricate case of two sequences, i.e. to pairwise alignments. The framework developed for single- and two-sequence methods may be extended further to multiple alignments and sequence-to-profile, as well as to HMM-based alignments.

Although statistically significant patterns are supposed to attract the biologist's attention, it is necessary to realize that statistical significance is not equivalent to biological significance [9–11]. The ultimate way to establish the latter is by experimental studies. The primary purpose of the statistical significance estimation is to highlight the targets for experiments with potentially interesting results. Whether such highlights are (on average) indeed worth further study, depends on the quality of the *P*-value estimation method.

THE USE OF *P*-VALUES IN SEQUENCE COMPARISONS

Definitions and basic methods

The notion of a *P*-value originates in the general statistical methodology of hypotheses testing [12]. Suppose we have a null and an alternative hypotheses, both of which can be used to explain the data produced by an experiment. The hypotheses are mutually exclusive, and we wish to determine which one holds true. Based on a statistical test applied to the data, the null hypothesis can be either accepted or rejected in favor of the alternative hypothesis. To formalize the acceptance/rejection procedure, we introduce the *P*-value and the significance level (α-value) of the test. The *P*-value is defined as the probability of seeing a value of the test statistic at least as extreme as the observed value, assuming that the null hypothesis is true. If the null hypothesis is rejected when the observed statistic has a particular *P*-value, then the probability of making an error, rejecting the null hypothesis when in fact it is true (type I error), is equal to the *P*-value. The predefined level of the type I error selected by the researcher is called the α-value (say, $\alpha = 0.01$). If the *P*-value is smaller than the significance level, then the null hypothesis is rejected and the result is said to be significant (at the given significance level). It is easy to see that the *P*-value is actually all we need; it can be interpreted as the smallest significance level at which we can reject the null hypothesis.

For sequence alignments, the test statistic is usually the alignment score, and the null hypothesis is that the aligned sequences are unrelated. The alternative hypothesis is that the sequences are

biologically related; this relatedness often means homology, i.e. having a common ancestor. In this context, the *P*-value of a score can be viewed as the probability that a prediction of biological relationship with this score or higher is a false positive. The efficiency of score-based binary classification methods is often judged by their sensitivity (Sn) and specificity (Sp). The Sn and Sp are defined as follows:

$$\mathrm{Sn} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \quad \mathrm{Sp} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}},$$

where TP, FP and FN are the numbers of true positives, false positives and false negatives, respectively. Usually, the increase in Sn leads to a decrease in Sp, and it is important to find a score cutoff that would give an optimal, in some sense, balance of Sn and Sp. Choosing *P*-value cutoffs is one of the methods to control Sn and Sp [13, 14].

To calculate the *P*-value for a sequence alignment, we need to rigorously define the null hypothesis by specifying a probabilistic model (the null model) in which the aligned sequences are independent, and estimate the probability that this model generates a score at least as extreme as the observed score. Note that, since usually we are interested in high scores, the score-based tests is a one-sided statistical test, and we should concentrate on the behavior of the right tail of the score distribution. In score-free analyses, such as studies of word occurrences in single or multiple sequences or string matching, the test statistic may be the count of a given word in a sequence, or the distance between words, etc. The *P*-value is again the probability of this statistic taking a value at least as large as actually observed under the assumption that the null hypothesis is true, i.e. by pure chance.

As it was already mentioned, it is possible to implement randomness in the null model either by applying shuffling algorithms or by defining stochastic processes. Yet another approach is to select biomolecular sequences at random from available databases [15, 16]. The motivation for the latter approach is that real sequences may give a more biologically adequate null model, compared to mathematical abstractions or *ad hoc* algorithms. Shuffling methods also use real sequences as raw material, and are oriented toward preserving certain properties of real sequences. For example, the permutation technique devised by Altschul and Erickson [4] can preserve dinucleotide and trinucleotide composition in nucleotide sequences. However, shuffling methods have been criticized for introducing significant bias akin to overfitting in the case of short- and medium-length sequences, and for failing to reflect the natural processes of random nucleotide and amino acid replacement: 'evolution does not scramble sequence' [17]. In the study carried out by Pearson [15], the results of using database-derived 'effectively random' sets of unrelated sequences for the null model parameter estimation were encouraging, however, the null model itself was derived based on analytic properties of probabilistic sequence models. Waterman and Vingron [16] found that, for protein sequences, the estimation results using sequences retrieved from databases are in good agreement with those using random sequence models with independent residues. In general, the use of database sequences for parameter estimation and benchmarking is a promising approach [16, 18].

Having chosen a null model, it is important to have an efficient method for calculating the *P*-value. In the absence of analytical results, *P*-values are usually estimated by simulations; in the case of database-derived sequences, a part of 'simulation' is the process of random selection of the sequences. Suppose we would like to estimate the *P*-value of the value s of the test statistic, e.g. the alignment score. The most straightforward way is to run N simulations for the null model, and calculate the test statistic for each run. If the scores observed in M runs are greater than or equal to s, then $P(s)$ can be estimated by

$$\hat{P}(s) = M/N, \tag{1}$$

and this estimate is consistent. For example, suppose we wish to determine the *P*-value for the alignment score of two sequences of lengths n and n'. In this case, we need to align N pairs of simulated sequences of lengths n and n', determine M, and use (1). An alternative to the naïve direct simulation approach is importance sampling, when the samples are obtained from some *proposal distribution* that gives higher probabilities to the 'more important' points in the sample space (like those having higher scores). This approach can make the simulation more efficient; the bias arising from sampling the proposal distrubtion instead of the target distribution is usually small, and can be corrected for [19]. An example of application of importance sampling for *P*-value estimation is the work of Barash *et al.* [20], who used this method to estimate the statistical significance of newly predicted protein – DNA binding sites.

A more insightful way to estimate *P*-values is to use the (approximate) knowledge of the distribution of the test statistic. If the statistic is continuously distributed, and its probability density function (p.d.f.), $f(x)$, is known, then we can apply the formula

$$P(s) = \int_s^\infty f(x)\,dx. \tag{2}$$

If $f(x)$ (or its analytic approximation) is not known, then instead of $f(x)$ we can plug in (2) an empirical estimate of $f(x)$; such an estimate can be either the smoothed empirical distribution, or some parametric distribution fitted to the simulated data. Analogous statements hold for discrete distributions, with (2) replaced by a summation formula. While sequence alignment scores take discrete values, continuous functions are frequently used to approximate their distributions. We will see that a special role in *P*-value estimation is played by the continuous distribution with the cumulative distribution function (c.d.f.) given by

$$F(x) = \exp\left(-\exp\left[-\left(\frac{x-b}{a}\right)\right]\right), \quad x \in (-\infty, \infty);$$

here b is called the location parameter, and a is the scale parameter. This is the extreme value distribution of type I, also known as the Gumbel distribution [21]; some authors also call it the extreme value distribution. If $b = 0$ and $a = 1$, then the distribution is called standard Gumbel.

Instead of using the text statistic itself, researchers frequently use the corresponding *Z*-score, which is defined by

$$Z = (s - \bar{s})/\sigma,$$

where s is the value of the test statistic, $\bar{s}$ is its mean value and σ is the standard deviation. Usually, empirical estimates of the mean and standard deviation are used in the *Z*-score definition instead of the exact values [22]. A general result concerning the sequence alignment *Z*-scores was obtained by Bastien [23], who suggested a theoretical justification to the empirical fact that $Z > 8$ usually corresponds to significant alignments. This result is distribution-free, and follows directly from Chebyshev's inequality (which has been applied in sequence analysis earlier [10, 24]). Note, however, that Chebyshev's inequality is known not to give tight bounds (due to its generality), so this approach may underestimate the *P*-value and fail to detect a large percentage of significant results.

Although obtaining *P*-value estimates of the form (1) or building empirical distribution is always an option, such estimates may fail to give the correct answer for small *P*-values [10, 11, 25]. Furthermore, the applicability of the *P*-values obtained by simulation is frequently restricted to the concrete parameter values for which they were derived [18]. In addition, obtaining good estimates by simulations may be very time consuming [11, 18, 26]. Thus, analytic expressions for the distributions are always desirable. They are especially useful when they give explicit dependence of the *P*-value on some of the parameters, so that this dependence does not have to be estimated by simulations. Moreover, in the case of local gapped alignments analytic results for a simple special case allowed to 'guess' an accurate formula for the alignment score distribution. While simulations are needed to estimate those of the distribution parameters which depend on the sequence composition and the scoring system, the formula gives explicit dependence of the *P*-value on the lengths of aligned sequences, which makes its use quite efficient in practice [16, 18, 27].

Null models

Analytic approximations can only be derived for precisely specified probabilistic models. The question is what probabilistic models are adequate for DNA and protein sequences. For DNA, typical models are homogeneous Markov chains of order $m \geq 0$; the case $m = 0$ accounts for sequences of i.i.d. random variables [7, 28, 29]. Evidence has been obtained that, for protein-coding regions of DNA, non-homogeneous three-matrix Markov chain models of order $m \geq 1$ should be used, while noncoding regions can be modeled by homogeneous Markov chains [30–33]. Markov chain models are built under the assumption that the genome is homogeneous. If the nucleotide composition of a genome varies substantially along its length, it is possible to parse DNA [34] and build Markov models for separate homogeneous segments. It should also be noted that successful modern gene finding methods use hidden Markov or hidden semi-Markov models to represent DNA sequences [35–37]. These models may be best for describing inhomogeneous DNA sequences.

It is apparent that, in general, we should expect some dependence of the *P*-values on the selected

model. However, it has been found that the expression for the mean length of the longest (possibly interrupted) match for two random sequences with i.i.d. letters (nucleotides) can be fitted well to the distribution of the scores of the optimal local (possibly gapped) alignments of unrelated sequences; the variance in these two cases also behaves in a similar way (does not depend on the sequence length) [10, 24]. These findings have been regarded as a justification for the claim that independence models are likely to give accurate enough approximations for the distributions of test statistics in sequence comparisons [11]. For protein sequences, independence models (sequences of i.i.d. random variables) are usually employed [16, 18, 27, 38]; Robinson and Robinson [39] give amino acid frequencies which are considered typical [14, 40]. Numerical experiments show that, for protein sequences, using Markov models instead of independence models gives a relatively small enhancement in accuracy [16, 18, 41]. The study of Goldstein and Waterman [42] also shows that independence models for protein sequences may produce distributions close to those for real sequences. Yet Karlin and Altschul [43] note that independence models should be used only as a reference standard to prove that scores of certain alignments can occur by chance alone. Mott [14] points out that protein sequences consist of segments which differ in composition, and that accurate estimates of statistical significance should take this structure into account. However, the work of Schäffer *et al.* [40] on improving the accuracy PSI-BLAST demonstrated that, while it is important to consider the actual amino acid composition (which may be different from typical), adjustments made to cope with local variations in amino acid composition may not increase the accuracy of the database searches.

In the DNA sequence matching studies by Arratia *et al.* [44], the independence models worked rather well (see also Reich *et al.* [17]). Yet it is known that simple independence models should not be used for describing low complexity and GC-biased regions in DNA sequences [13, 45]. In the studies of transcription factor binding sites, background models in the form of Markov chains are generally considered as more advanced compared to independence models [46]. Thus, the degree of influence of the choice of a probabilistic sequence model on the quality of the statistical significance analysis appears to depend on the specific problem being solved. We shall return to the subject of null model selection when describing specific sequence analysis problems.

SINGLE SEQUENCE ANALYSIS

Occurrence of words in random sequences

Many functional elements in the genome, e.g. DNA restriction sites [42] and transcription factor binding sites [46], may be represented as words, i.e. as strings of letters from an alphabet. The knowledge of probabilistic properties of words in random texts is necessary for the estimation of statistical significance of patterns observed in DNA sequences. In particular, distributional properties of words are important in DNA sequencing by hybridization [28, 47], in the analysis of functional DNA regions [48–50] and in alignment-free comparisons of biological sequences [51].

The abundance of analytic results for Markov chain models makes such models a convenient tool in the studies of statistical properties of words in DNA sequences. Many of the probabilistic properties of words in texts generated by a stationary homogeneous Markov chain of order $m \geq 0$ have been surveyed by Reinert, Schbath and Waterman [28] and Schbath [52]; see also Robin and Schbath [53]. In these works, approximate as well as exact results on the statistical properties of counts of sell-overlapping and non-overlapping words have been discussed. These results allow to assess the probabilities of events related to different word occurrences, thus providing *P*-value estimates for such events. Statistics for both exact and degenerate words, like TAT(A or G)A, have been considered. Word count distributions, distributions of the length of the gaps between words, and occurrences of multiple patterns were analyzed. In asymptotic cases, normal, Poisson and compound Poisson approximations have been derived for word counts; the type of approximation depends on the word length and on the method of counting word occurrences. In particular, if n is the length of the sequence, then, for large n, the distribution of counts of a word can be approximated by the normal distribution; this approximation is good when the length of the word is relatively small compared to the sequence length, and works for both self-overlapping and non-overlapping words. The normal approximation can also be extended to the case of the inhomogeneous three-matrix Markov model [28]. For the distributions of

counts of words for which the expected count is rather small (words with length $O(\ln n)$), Poisson approximation is more appropriate. For periodic words (words consisting of repeated patterns) Poisson approximation is not satisfactory because of possible overlaps, and compound Poisson approximation should be used. Not only limiting distributions, but also bounds on the speed of convergence have been provided for the approximations. Similar results are available for the distributions of joint counts and joint occurrences of multiple patterns. While the Gaussian and Poisson laws are used to approximate the probability that a given word occurs more than a certain number of times, the probability that a given word frequency deviates from its expected value can be approximated using large deviations techniques. The distribution of the length of sequence intervals between two occurrences of a word can (under some conditions) also be approximated by the Poisson law.

The word count statistics could be immediately used in word representation analysis [53]. Its main principle is as follows. Let $N(w)$ be the random variable designating the number of occurrences of word w in a sequence described by a given probabilistic model; $N(w)$ is our test statistic. Denote by $N_{\text{obs}}(w)$ the number of occurrences of w observed in the real sequence. Then the P-value for the observation is defined by

$$P = \Pr\{N(w) \geq N_{\text{obs}}(W)\}.$$

A small P-value corresponds to an overrepresented word w.

Score statistics for random sequences

If a sequence of letters represents the primary structure of a biological macromolecule, certain biologically relevant properties of the monomers (letters) can be described numerically by scores, the numbers associated with each letter. For the letters from an alphabet $\mathcal{A} = \{a_1, \ldots, a_q\}$, a scoring system is defined by the set $\mathcal{S} = \{s_1, \ldots, s_q\}$, where score s_i corresponds to the letter a_i. For example, the score based on amino acid charge can be defined as follows [43]: $s = +2$ for the positively charged amino acids lysine and arginine; for the negatively charged amino acids aspartate and glutamate, $s = -2$; for histidines, $s = 0.04$; for other amino acids, $s = -1$. Other examples of scores include scores derived from target frequencies and scores based on structural alphabets [43, 54, 55]. The aggregate score for a (contiguous) sequence segment is defined as the sum of the scores of constituting letters. Of interest are high-scoring segments, which with an appropriate choice of scoring system correspond to structurally and functionally interesting regions of a biological molecule.

A natural objective of the estimation of statistical significance for scores is to assess the probability of scores higher than a given score for random sequences. The general simulation-based approach to this problem was described above, and here we concentrate on the theoretical estimates for the P-values. One of the most important result in this direction is the theorem of Iglehart–Karlin–Dembo–Kawabata [55, 56], brought into a biological context by Karlin and Altschul [43]. The theorem concerns with maximal scoring segments (MSSs) in sequences of i.i.d. letters with letter probability distribution $(p_1, \ldots, p_q)$. An MSS is a sequence segment with maximal aggregate score. The two main assumptions on the scores are that: (i) at least one of the scores s_i is positive and (ii) the mean score, $\bar{s} := \sum_{i=1}^{q} p_i s_i$, is negative (the case $\bar{s} = 0$ can also be analyzed [43, 55]). The assumption (ii) is to prevent the MSS from extending up to the whole sequence. The conditions (i) and (ii) are naturally satisfied in many situations [43]. Let n be the sequence length, and $s(n)$ be the score of an MSS. Then, for any real number x and for large n,

$$\Pr\{s(n) - \ln n/\lambda^* > x\} \approx 1 - \exp(-K^* e^{-\lambda^* x}), \qquad (3)$$

where K^* and λ^* are positive constants depending on the probabilities p_i and scores s_i. In particular, λ^* is the unique positive root of the equation

$$\sum_{i=1}^{q} p_i \exp(\lambda s_i) = 1,$$

and K^* can be either computed or estimated [43, 55]. The quantity $\ln n/\lambda^*$ in (3) is the asymptotics for the maximal segment score; to account for the growth of $s(n)$ with n, we need to do the substraction in the left-hand side of (3). Setting $s = \ln n/\lambda^* + x$, we can write (3) in the equivalent form:

$$\Pr\{s(n) > s\} \approx 1 - \exp(-K^* n e^{-\lambda^* s}).$$

This expression implies that the distribution of the random variable $\lambda^* s(n) - \ln(K^* n)$ can be approximated by the standard Gumbel distribution. A more rigorous form of (3) is the inequality

$$\Pr\{s(n) > \ln n/\lambda^* + x\} \leq 1 - \exp(-K^+ e^{-\lambda^* x}), \qquad (4)$$

which gives a conservative estimate for the *P*-value [43, 55, 57]. The quantity K^+ in (4) can be easily computed knowing λ^* from the geometrically converging series for K^* (Karlin and Altschul [43]; Karlin, Dembo and Kawabata [55]).

The formula (3) gives estimates for the statistical significance of the score x of an MSS: if the right-hand side equals p, then the score x is significant at the level $\approx p$. This formula can be derived given that the following statement is true [43, 55]: for large n, the distribution of the number, m, of non-intersecting segments with score exceeding $\ln n/\lambda^* + x$ is approximately Poisson with parameter $K^* e^{-\lambda^* x}$. Indeed, the probability that no segment score exceeds $\ln n/\lambda^* + x$ is calculated by setting $m = 0$, and is equal to $\exp(-K^* e^{-\lambda^* x})$. Subtracting this from 1, we obtain (3). This Poisson approximation can be used to assess the statistical significance of several occurrences of high-scoring segments (i.e. distinct segments scoring higher than a given threshold) in a sequence and judge about their possible over-representation. However, while the utility of (3) and the Poisson approximation is obvious, no theoretical bounds on the accuracy of the approximation are available. Thus it is unknown of how large n should be to guarantee a good approximation. Nevertheless, Karlin and Altschul [43] state that, in practical situations, $n \geq 150$ is large enough.

The theory described above concerns with the i.i.d. models of sequences, but the analogue of (3) can be obtained for Markov chain models of the sequence, as was shown by Karlin and Dembo [57]. In that work, the sequence was modeled by a Markov chain with a positive transition probability matrix (the results are extendable to irreducible aperiodic case). For each consecutive pair of letters in the sequence, the scores were represented by random variables associated with the Markov chain transitions, so that the sequence of scores formed a hidden Markov model of a special type. The analogues of the conditions (i) and (ii) were introduced, and formulas similar to (3) and (4) were derived for the distribution of the MSS score.

Sometimes it may be necessary to analyze not only the MSS, but also the second-, third-, ..., rth highest-scoring segments. Such a necessity arises when, in a biomolecule, there exist several high-scoring domains with a common property of biological interest. In this case, the search for the MSS only will ignore other important information. Let $s_1, \ldots, s_r$ be the scores of the r distinct highest-scoring segments. Write $s'_k = \lambda^* s_k - \ln(K^* n)$; this normalization is introduced for convenience. For large n, the joint distribution for $s_1, \ldots, s_r$ is well approximated by the p.d.f.

$$f(x_1, \ldots, x_r) = \exp\left(-e^{-x_r} - \sum_{k=1}^{r} x_k\right), \qquad (5)$$

where $x_1 \geq x_2 \geq \cdots \geq x_r$ (Karlin and Altschul [58] and Karlin [59]). If $r = 1$, then (5) reduces to (3). From (5) we can compute the distribution of any function of $s_1, \ldots, s_r$. One interesting statistic derived from (5) is the score distribution of the rth highest-scoring segment, which can also be obtained from the Poisson approximation to the distribution of the number of high-scoring segments [58]. But if we analyze a sequence with all of the r highest-scoring segments having relatively large *P*-values, the statistics for the separate segments will not indicate significant regions. Here, we would need a statistic which gives a collective description of the segments, and may be useful in the analysis of groups of highest-scoring segments which do not have considerably small *P*-values when analyzed separately. Such a statistic is the distribution of the sum of scores for the r highest-scoring segments [58]. Using (5), it can be shown [58, 59] that the distribution of $T_r = \sum_{k=1}^{r} s'_k$ for large n can be approximated as follows:

$$P\{T_r > s\} \approx \frac{1}{r!(r-2)!} \int_s^\infty e^{-x} \int_0^\infty y^{r-2} \exp(e^{(y-x)/r})\, dy\, dx; \qquad (6)$$

if s is sufficiently large, then

$$P\{T_r > s\} \approx \frac{e^{-s} s^{r-1}}{r!(r-1)!}. \qquad (7)$$

STATISTICAL SIGNIFICANCE OF PAIRWISE ALIGNMENTS

Global alignments

Two sequences can be aligned in many different ways. Optimal alignments, those with the best scores, are of great practical interest. Global optimal alignments are optimized along the whole length of the two sequences. The dynamic programming algorithm for constructing the optimal global pairwise alignment was proposed by Needleman and Wunsch [60]. A more efficient version was later devised by Gotoh [61].

So far, the statistical significance of global alignment scores has been assessed via simulations;

no theoretical results on the score distribution are known. Reich *et al*. [17] investigated score distributions for the DNA global alignments. Their scoring system was defined by three numbers: a match score (equal to +1), a mismatch score (equal to 0) and a linear gap penalty. The work was centered around the case of zero gap penalties. Using independence models for DNA sequences, the authors found that the score distribution for alignments of such sequences is quite close to that for alignments of DNA sequences randomly retrieved from a database. The authors used the Z-score as the statistical measure for the extremity of the scores. As is well known, if the score s is normally distributed, then Z-scores such that $Z > 3$ are highly significant; however, in the case of massive screening in databases, Z-score cutoffs $Z = 5$ or $Z = 6$ should be used [17]. Reich *et al*. argued that, although in reality the score distribution may display significant deviations from normal, the approach to P-value estimation based on the use of Z-scores may work well in practice. For sequences of equal length and uniform four-letter distribution, aligned with zero gap penalties, Reich *et al*. provided empirical formulas for the dependence of $\bar{s}$ and σ on the sequence length. Corrections to these formulas for the cases of non-zero gap penalties, non-uniform letter distributions and non-equal sequence lengths have also been discussed.

Even if the empirical rule that $Z \geq 6$ corresponds to significant results holds in some cases, it should be used with caution in the absence of knowledge about the tail behavior of the score distribution. An example of another possible approach to assessing P-values for global optimal alignments is the estimation of estimation of statistical significance for evolutionary distances between sequences (derived from global alignments) carried out by Altschul and Erickson [4]. In that work, the authors devised a special shuffling procedure (mentioned in the section of the use of P-values in sequence comparisons), and proposed to use it for P-value estimation. Since computations showed that the distribution is slightly non-Gaussian, the authors preferred to make no assumptions concerning the tail behavior, and used an expression similar to (1) for P-value estimation.

The above-cited works considered global alignments of DNA sequences. Webber and Barton [62] investigated the Z-score distribution for global protein alignments with length-independent gap penalties (the affine gap penalty with gap extension coefficient equal to zero). The protein sequences to be aligned were selected from existing databases and modified by shuffling. Having tried different distributions (including the normal distribution and the extreme value distribution) for fitting the data, the authors came to the conclusion that the best fit was given by the gamma distribution with density defined by

$$f(x) = \frac{(x+\beta)^{\alpha-1}}{\Gamma(\alpha)\lambda^{\alpha}} e^{-(x+\beta)/\lambda}, \tag{8}$$

where $0 \leq x + \beta < \infty$, $\alpha > 0$ is a shape parameter, $\lambda > 0$ is a scale parameter, and $\Gamma(\cdot)$ is the gamma function. These parameters account for the length variability for the studied proteins. The authors stated that the choice of this distribution was justified by the quality of fit. In the paper, the authors provided the values of α, β, and λ for several frequently used scoring matrices (PAM [63], BLOSUM [64], and Gonnet [65]) and length-independent gap penalties. Using the formula (8), the P-value of a given Z-score can be computed via (2). Webber and Barton provided P-values for Z-scores ranging from 0 to 11.

Local alignments

In many practical situations we are interested in the highest-scoring alignment between the subsequences of two given sequences, the best local alignment. The local alignment is called optimal if (a) the aligned subsequences form a high-scoring segment pair (HSP) (with score that cannot be increased by extending or shortening the aligned subsequences) and (b) all other HSPs have the same or smaller alignment score. The dynamic programming algorithm for finding optimal local alignment was proposed by Smith and Waterman [66]. Gotoh [61] developed a faster version for the case of affine gap scores.

The problem of P-value estimation for local alignments has been intensively studied. We first describe the theoretical results available for simplified cases, and then move on to a description of different corrections and modifications necessary for the P-value estimates to work in practically important situations. Consider a local ungapped optimal alignment with score matrix $S = (s_{ij})$. Such an alignment can be obtained by applying the Smith–Waterman algorithm with sufficiently large gap costs. The score s_{ij} is the score of aligning the letter i in one sequence with the letter j in the other sequence. The fundamental result in the P-value estimation for such

alignments is given by the extension of the single sequence formula (3) to the case of two sequences. The possibility of such an extension was stated by Karlin and Altschul [43], and the corresponding approximation is known as the Karlin–Altschul statistic. Consider sequences of letters from the alphabet $\mathcal{A} = \{a_1, \ldots, a_q\}$. For two independence sequence models with letter distributions $(p_1, \ldots, p_q)$ and $(p'_1, \ldots, p'_q)$, respectively, we assume that: (i) $p_i p'_j s_{ij} > 0$ for some i, j and (ii) $\sum_{i,j=1}^{q} p_i p'_j s_{ij} < 0$. Assume also that the letter distributions (p_i) and (p'_i) are not too dissimilar, and that the sequence lengths n and n' grow at roughly equal rates. In this case, the two-sequence analogue of (3) is

$$\Pr\{s(nn') > \ln(nn')/\lambda_u + x\} \approx 1 - \exp(-K_u e^{-\lambda_u x}), \quad (9)$$

or, equivalently,

$$\Pr\{s(nn') > s\} \approx 1 - \exp(-K_u nn' e^{-\lambda_u s}).$$

Here, $s(nn')$ is the local alignment score (whose expected value is $\ln((nn')/\lambda^u)$, λ_u is the unique positive root of the equation

$$\sum_{i,j=1}^{q} p_i p'_j \exp(s_{ij}) = 1, \quad (10)$$

and K_u is to be computed (or estimated) in the same way as K^* in (3) for a single sequence (the subscript 'u' stands for 'ungapped'). In the single sequence analysis, K^* depends on the sequence composition; to calculate K_u, we can use either of the compositions, since we have assumed that they do not differ much. Note that it is more important to accurately compute λ_u than K_u, because of the double-exponential dependence on λ_u in (9). If x is large, then, using (9) with the expression $\exp(a) = 1 - a + a^2/2 + o(a^2)$, we obtain that

$$\Pr\{s(nn') > \ln(nn')/\lambda_u + x\} \leq K_u e^{-\lambda_u x};$$

this inequality appears in Karlin and Altschul [43]. Notice also the connection of (9) with the extreme value statistics: it follows from (9) that the distribution of the random variable $\lambda_u s(nn') - \ln(K_u nn')$ is close to the standard Gumbel distribution.

In general, condition (ii) needs to be checked for a given scoring system and sequence compositions, but there is a special case when (ii) is satisfied automatically. This is the case when s_{ij} are log-odds scores, i.e. by definition

$$s_{ij} = \ln \frac{p_{ij}}{p_i p_j},$$

where p_{ij} is the probability that the letters i and j form a pair in the alignment (thus, $p_{ij} = p_{ji}$) [67, 68]; here we assume that $p_i = p'_i$ for all i. The scores given by PAM [63] and BLOSUM [64] matrices are examples of log-odds scores. For log-odds scores, (ii) is satisfied unless $p_{ij} = p_i p_j$ for all i, j; see Durbin *et al.* [68]. Also, it is easy to check that $\lambda_u = 1$ for log-odds scores.

The expression (9) was rigorously proved by Dembo, Karlin and Zeitouni [69] for the special case $n = n'$, under an additional assumption on the sequence compositions and the scoring system (see also Karlin [59]). This assumption is satisfied if $p_i = p'_i$, $s_{ij} = s_{ji}$ for all i, and s_{ij} are not of the form $s(i) + s(j)$. In fact, they proved that the number of distinct HSPs with score larger than $\ln n^2/\lambda_u + x$ has Poisson distribution with mean $K_u \exp(-\lambda_u x)$, which implies (9) if $n = n'$. In the general case, this Poisson approximation and the expression (9) can be considered intuitive. Yet (9) has been shown to work quite well in practice [16, 17]. The formula (6) for the score distribution of the rth best HSP could also be adapted to the case of local ungapped pairwise alignments: it describes the p.d.f. of the statistic $T_r^* + \ln r!$, where T_r^* is the greatest value attainable by the sum of the normalized scores from r distinct and consistently ordered segment pairs [58]. The approximation (7) in the case of pairwise alignments becomes

$$P\{T_r^* > s\} \approx \frac{e^{-s} s^{r-1}}{(r-1)!}. \quad (11)$$

The knowledge of the behavior of the score distribution for ungapped alignments provides insights into what to expect of gapped alignments. In this more general situation, frequently used approach has been to fit the score distribution via an extreme value distribution of the form (9), where the constants K_g and λ_g (the 'gapped' analogues of K_u and λ_u) are to be estimated from computations with random sequences or with sequences randomly retrieved from databases [15, 16, 18, 70, 71]. In general, the obtained approximations are quite accurate. Altschul and Gish [27] came to the conclusion that 'the statistical theory for ungapped alignments carries over essentially unchanged to gapped alignments'. The currently available theoretical results support the assumption that such approximations are appropriate. In particular, a theorem of Siegmund and Yakir [72, 73] gives the following result. Consder two independence

sequences of lengths n and n', and suppose that the given scoring system (the set of matching scores) has a rather general form (the scores occurring with positive probabilities cannot be represented as a sequence of increasing numbers with constant increment). These sequences are locally aligned with gaps, using affine gap penalties. Suppose also that, for the gap opening penalty, $\Delta(s)$, we have the formula $\lambda_u\Delta(s) = \ln(\lambda_u s) + C$, where s is a real number representing the alignment score, and C is some constant. If, in addition, $nn' \exp(-\lambda_u s)$ converges to a finite, positie limit as $s, n, n' \to \infty$, then, for large n, n', s,

$$\Pr\{s(nn') \geq s\} \approx 1 - \exp(-Knn'e^{\lambda_u s});$$

where K is a computable constant [72]. Since the gap penalty, as defined above, grows with the score, this result in fact is relevant to what can be called 'asymptotically ungapped' alignments. Currently, no rigorous mathematical theory is available for the general case of local gapped pairwise alignments.

Approximations by parametric distributions of the form (9) work not for all scoring systems, gap penalties and sequence compositions. It can be shown that, in the case of affine gap costs and symmetric scores, there are two types of regions for scoring parameters: *linear* and *logarithmic* [16, 18, 74]. In the logarithmic region, the growth of the expected local alignment score is proportional to the logarithm of the product of the sequence lengths; in the linear region, the expected local alignment score grows proportionally to the sequence lengths. This result holds both for independence and Markov sequence models [74]. The precise definitions of the logarithmic and linear regions are as follows. Let $s_g(n^2)$ be the optimal *global* alignment score of two random sequences of length n, and $Es_g(n^2)$ be its expected value. It can be shown that the limit $\omega = \lim_{n\to\infty} Es_g(n^2)/n$ exists. If $\omega < 0$, then the *local* alignment with this scoring system is in the logarithmic region; if $\omega > 0$, then it is in the linear region [74]. The requirement that the local alignments be in the logarithmic region is the extension of condition (ii) for ungapped alignments to alignments with gaps. The parallel with the ungapped case becomes obvious if we remember that in that case the expected score is $\ln(nn')/\lambda_u$, which exactly corresponds to our description of the logarithmic region. In the linear region, the penalty for poorly aligned letters and indels is too low, so the limiting expected global alignment score per aligned letter pair is positive. In this case, high-scoring local alignments may have regions of poorly aligned letters. Consequently, using local alignments with scores falling in the linear region is not productive in biological sequence analysis.

Waterman and Vingron [16, 18] showed that (9) can be extended to gapped local alignments in the logarithmic region, but not in the linear region. They used Poisson clumping heuristic to model the distinct HSPs. The basic idea of the approximation is that the number of nonintersecting segment pairs with score greater than or equal to s is approximately Poisson distributed with mean $K_g nn' e^{-\lambda_g s}$, where n, n' are the sequence lengths, and K_g, λ_g are the parameters to be estimated from simulations (notice the similarity with the Poisson approximation in the single sequence case). In parallel to the single sequence case, this Poisson approximation also yields the distribution of the rth best alignment score s_r (see Waterman and Vingron [16]):

$$\Pr\{s_r \geq s\} \approx 1 - \exp(K_g nn' e^{-\lambda_g s}) \sum_{j=0}^{r-1} (K_g nn' e^{-\lambda_g s})^j / j!.$$

This probability is simply the Poisson probability of finding at least r distinct HSPs each scoring at least s. Such a proability can be useful in reporting a combined assessment of statistical significance of a number of HSPs [27]. The simulations and fitting were performed for both independence and Markov sequence models. For parameter estimation, Waterman and Vingron used the maximum likelihood method (see also Mott [26]), linear regression on transformed data, and a specially designed method based on the evaluation of the expected number of local alignments scoring higher than a given threshold. All these methods were shown to be quite efficient.

Another study of the quality of approximation of the score distribution for gapped alignments by (9) was undertaken by Altschul and Gish [27], who estimated K_g, λ_g via the method of moments for independence sequence models. One of the important facts that they discovered was that (6) gave a good approximation for the sum of normalized scores distribution for gapped local pairwise alignments. Another fact was the necessity of using corrected sequence lengths when extending (9) to gapped alignments. Such a necessity arises because local alignments starting near the end of either sequence are likely to run out of sequence before they can attain a sufficiently large score. All sequences must therefore be considered as having

an effective length shorter than the actual one. Altschul and Gish [27] propose using the corrected lengths

$$n_c = n - l, \qquad n'_c = n' - l,$$

where l is the length of a typical optimal local alignment. These edge effects become negligible for $n, n' \to \infty$ due to which they do not arise in theoretical considerations which provide asymptotics. When l is a sizable part of either n or n', the edge effects cannot be discarded, and asymptotic theory loses accuracy. It has also been noticed that gapped alignments are generally longer than ungapped alignments, so finite-length effects and, respectively, the importance of the finite-length corrections are greater in the gapped case [75, 76]. The formula for l, suggested by Altschul and Gish [27], is as follows:

$$l \approx \ln(K_g nn')/H, \tag{12}$$

where H is the relative entropy of the scoring system [67]. Finite-length corrections were considered in a number of studies, thus becoming standard [14, 15, 70]. Some other approaches, besides (12), were proposed to estimate l. In particular, Altschul *et al.* [70] state that, empirically, the mean length of random optimal local alignments with sufficiently large score s depends linearly on s:

$$l \approx \alpha s + \beta.$$

The constants α and β can be estimated from simulations. A theoretical investigation of finite-length corrections for ungapped alignments was carried out by Spouge [77]. However, the correction (12) was shown [75] to be generally more suitable for applications than the one developed by Spouge [77].

The importance of finite-length corrections for a given pair of sequences depends on how good an approximation is provided by (9) for sequences of such lengths. Mott [14] has shown that the quality of this approximation depends on the scoring system and on the sequence composition, for both gapped and ungapped alignments. In the case of protein sequences with typical composition, relatively accurate approximation is achieved at lengths about $n = n' = 250$. For sequences with biased compositions, even $n = n' = 1000$ may be insufficient. Since many real protein sequences have skewed composition, and the average protein sequence length is about 330 residues, finite-length corrections are as a rule necessary.

The search for computationally efficient methods of estimating the parameters of the score distribution for gapped alignments continues. Several methods for computing the parameters were compared by Pearson [15]. Among recently proposed methods, we have to mention the 'islands' method [70] for estimating K_g and λ_g, which is being used in the latest modifications of BLAST [40]. 'Islands' are in fact distinct local alignments defined in a special way; this definition leads to an effective algorithm for finding such alignments. Altschul *et al.* [70] approximated the distribution of the number of such alignments by the Poisson distribution analogous to that used by Waterman and Vingron [16, 18]. This gave an explicit expression for the expected number of such alignments, which allowed to estimate the parameters K_g, λ_g by fitting this expression to the islands score data obtained by simulations. An importance sampling based method for estimating λ_g was developed by Bundschuh [71], and theoretically justified by Grossman and Yakir [78]. Bailey and Gribskov [79] treated H in (12) as a parameter to be estimated, and devised a maximum likelihood method for simultaneous estimation of K_g, λ_g and H. Mott and Tribe [80] proposed an ingenious method of estimating K_g, λ_g, which uses a suboptimal approximation to the Smith–Waterman algorithm with simpler score statistics. For this approximation, termed the Greedy Extension Model (GEM), the score distribution was made dependent on the gap penalty function through a parameter α, and K_g, λ_g are estimated as

$$K_g \approx K_u \kappa(\alpha), \quad \lambda_g \approx \lambda_u \theta(\alpha),$$

where κ, θ depend only on α. Note that the method of Mott and Tribe allows to predict the transition point between the linear and algorithmic region. Subsequently, Mott [14] further developed and improved this approach.

Note that we need to know sequences' lengths and composition to use the Karlin–Altschul statistic (9). While this statistic explicitly shows the dependence of the score distribution on the lengths, the dependence on sequence composition is implicit. It is therefore desirable to have a composition-free measure of statistical significance. Bacro and Comet [81] have shown that the Z-score of an optimal local alignment is such a measure. Under the assumptions used by Waterman and Vingron [16], Bacro and Comet demonstrated that the Z-score has

approximately extreme value distribution with parameters independent of sequence lengths or compositions. The Z-values can be obtained from simulations, using the shuffling procedure described by Comet *et al*. [22].

Local pairwise alignments and database searches

One of the frequently used types of database searches is the search for sequence similarity implemented as a series of local pairwise alignments of the query sequence to all the sequences in the database. Current databases are quite large, and the Smith–Waterman full dynamic programming algorithm is too slow to be practical in this context. This is why fast heuristic algorithms have been designed for database searches. The best-known and most widely used heuristic algorithm are BLAST [38, 40, 82] and FASTA [83]. Though fast, the heuristic local alignment algorithms are not as sensitive as the Smith–Waterman algorithm.

Since results of database searches are best local alignments (hits), the Karlin–Altschul statistics is applicable for the estimation of their statistical significance. However, the specific nature of database searches makes it necessary to modify the straightforwardly obtained Karlin–Altschul estimates. Before describing these modifications, we introduce the notion of the E-value. As was mentioned above, the number of HSPs for gapped alignments with score not lower than s has approximately the Poisson distribution with parameter $K_g nn' \exp(-\lambda_g s)$. Since the parameter of the Poisson distribution is also its mean value, the expected number of HSPs with score not less than s is given by the formula

$$E = K_g nn' e^{-\lambda_g s}.$$

This quantity is called the E-value for the score s. For ungapped alignments, the definition is fully analogous. The E-value is a convenient measure of the statistical significance of database hits; both the BLAST and FASTA programs return E-values. Note the relationship between the P-value and the E-value (which comes from the Poisson approximation):

$$P = 1 - e^{-E}.$$

Therefore, for small P, $P \approx E$. In the context of sequence similarly detection by massive screening of local alignments to database sequences, the E-value of an alignment score can be interpreted as the expected number of false positives having this score or a higher one.

The approach to the evaluation of the statistical significance of database hits depends on how to consider the set of sequences in the database. In this respect, BLAST and FASTA differ significantly. In BLAST, all the sequences are treated as one concatenated sequence [38, 84]. Thus, if a query sequence of length n hits a database sequence with score s, then the corresponding E-value is

$$E = K_g nL e^{-\lambda_g s},$$

where L is the sum of the lengths of all the sequences in the database. This approach is based on the assumption that long sequences are more likely to have high-scoring hits, which is true for the probabilistic models of the sequences. To give a formal justification, let $s_{\max} = \max_i\{s_i\}$, where s_i is the score obtained from the ith entry in the database. Assuming that s_i are independent random variables, we have that

$$\begin{aligned} \Pr\{s_{\max} < s\} &= \Pr\left\{\bigcap_i \{s_i < s\}\right\} = \prod_i \Pr\{s_i < s\} \\ &\approx \exp\left(\sum_i -K_g nn_i e^{-\lambda_g s}\right) \end{aligned},$$

where n_i is the length of the ith database entry (Spang and Vingron [76]). Therefore,

$$\Pr\{s_{\max} < s\} \approx \exp(-K_g nL e^{-\lambda_g s}). \qquad (13)$$

Spang and Vingron showed that, for real databases, instead of L one should use in (13) the *effective size* of the database.

In evaluation of the FASTA hit it is assumed that all database sequences, independent of their length, *a priori* have equal chances to score high when compared to a given query sequence [15, 84]. This can be justified by the assumption that a significant similarity to the query sequence can occur in both short and long database sequences. In this case, the E-value of a database hit with score s is calculated as

$$E = PN,$$

where P is the P-value for the score s of the alignment of the query sequence to the database

sequence, and N is the number of sequences in the database [15].

STATISTICAL SIGNIFICANCE OF MULTIPLE ALIGNMENTS

Global alignments

The exact dynamic programming algorithm for constructing a multiple alignment is known but is impractical for more than a few sequences, therefore, heuristic methods, such as progressive alignment, are usually used [68, 85]. The question of how to estimate the statistical significance of such alignments is even more obscure than it is for pairwise alignments. The papers describing well-known multiple alignment algorithms such as CLUSTAL W [86] and T-COFFEE [87], and newer methods such as MUSCLE [88] and MAFFT [89], say nothing about statistical significance of the produced alignment. Of course, it is possible to use simulations and distribution curve fitting for *P*-value estimations, but methods based on pure simulation may fail in the case of small *P*-values [25]. Thus, analytic approaches should be developed. When estimating statistical significance, it is natural to take into account the specific features of a multiple alignment algorithm. For example, if a method uses a sequence of pairwise alignments, then it may be possible to utilize the estimates for the pairwise alignment score statistics to assess the *P*-value for the multiple alignment. This approach is implemented in DIALIGN-T [90], which builds multiple alignments from *ungapped* pairwise local alignments, called fragments, involving pairs from the whole set of sequences. The score of a multiple alignment produced by DIALIGN-T is the sum of the scores of the constituting fragments, while a fragment score is defined as a negative logarithm of the *P*-value of global (Needleman–Wunsch) pairwise alignment for the fragment. Thus, an optimal multiple alignment is a collection of fragments with minimal product of pairwise *P*-values, and the score of a multiple alignment is the negative logarithm of an estimate of its *P*-value. The probabilistic sequence models for both DNA and protein alignments are independence models with uniform letter distribution; the global pairwise score *P*-values are estimated via simulations combined with heuristic formulas. Although such an approach seems reasonable, it was argued that DIALIGN-T over-estimates the probabilities of random occurrences of alignments with high scores [90].

Local alignments

Similar to the case of global multiple alignments, the full dynamic programming solution to the problem of local multiple alignments is currently unfeasible. Practically efficient approaches include heuristic block analysis [91], expectation–maximization [92], Gibbs sampling [93] and Eulerian path approach [94]. Also, the program DIALIGN-T can construct local alignments, and has been shown to perform quite well (the *P*-value estimation procedure of DIALIGN-T was outlined above). As is the case with pairwise alignments, some analytic results are available which can increase the efficiency of the *P*-value estimation for local multiple alignments. The choice of a method for *P*-value estimation depends on the nature of the alignment algorithm and on scoring system.

Formulas for the *P*-value estimation are available only for ungapped local multiple alignments. The basic idea behind the *P*-value estimation is the conjecture that expression (9) is extendable to the case of multiple sequences [43, 69]. Let $S_1, \ldots, S_r$ be independent sequences consisting of i.i.d. letters from the alphabet $\mathcal{A} = \{a_1, \ldots, a_q\}$. Let $(p_i^{(j)})_{i=1}^{q}$ be the letter probabilities for the sequence S_j. We specify a length and choose one segment of this length from each of the sequences; the set of such segments is called a block. The score of this block is defined as the sum of scores for the block columns. If there is a block whose score cannot be increased by shifting any of its borders, then this block is called a high-scoring block. A high-scoring block with the maximal score corresponds to the optimal local ungapped alignment of the sequences $S_1, \ldots, S_l$. Suppose that: (i) some column score is positive with positive probability; (ii) the average column score is negative; (iii) column scores do not change upon permuting the block's rows. These assumptions are valid if the column scores are defined as the SP(sum-of-pairs)-scores [91]. The SP-score of a column is the sum of all the pairwise scores for the letters it comprises, with each pair being counted only once. For example, if the letters denote amino acids, then the pairwise substitution scores can be the usual log-odds scores defined by a PAM matrix.

SP-scores are frequently used scores for multiple sequence alignments [68].

Suppose now that the lengths $n_1, \ldots, n_l$ of the sequences $S_1, \ldots, S_l$ are large and do not differ much from each other; the sequences are assumed to have similar sequence composition. Then, for the *P*-value of the multiple sequence alignment with score $s(n_1 \cdots n_l)$, we have the expression

$$\Pr\{s(n_1 \cdots n_l) > s\} \approx 1 - \exp\left(-K_m \prod_{i=1}^{l} n_i e^{-\lambda_m s}\right). \quad (14)$$

Here, K_m is calculated as in the case of two sequences, and λ_m satisfies

$$\sum_{i_1,\ldots,i_l=1}^{q} \prod_{k=1}^{l} p_{i_k}^{(k)} \exp(s_{i_1 \ldots i_l}) = 1,$$

where $s_{i_1 \ldots i_l}$ is the score of a column having letter i_1 in the first row, letter i_2 in the second row, etc. If $l = 2$, then (14) reduces to (9). This approach to the *P*-value estimation was implemented in the local multiple alignment program MACAW [91]. It should also be noted that the Poisson clumping heuristic, successfully used for local pairwise alignments as described earlier, was generalized for the case of local multiple alignments by Zhang and Waterman [94]. Though the generalization (14) and the Poisson approximation seem to be reasonable, we emphasize that they have not been proven mathematically, and we are unaware of any systematic studies of the quality of these approximations.

Another method of the estimation of *P*-values in local ungapped multiple alignments of sequences of letters from alphabet $\mathcal{A}$ was proposed by Hertz and Stormo [25]. They used the information content of a block of width w as a test statistic; the block itself might be derived by any alignment method (e.g. Gibbs sampling). The main assumption is that interesting alignments are those whose letter frequencies differ significantly from the *a priori* letter probabilities. The information content of the block is defined by

$$I = \sum_{i=1}^{q} \sum_{j=1}^{w} f_{ij} \ln\left(\frac{f_{ij}}{p_i}\right),$$

where f_{ij} is the frequency of the letter a_i occurring in the column j of the block, and p_i is the *a priori* probability for the letter a_i. The *a priori* probability of a letter might be the frequency of the letter within all sequences of the database or the frequency within a subset of sequences. The information content measure works well for DNA alignments. For proteins, this measure may not be sufficient to characterize the alignment, because it does not take into account physicochemical similarities between different amino aicds [95]. Note that the Hertz and Stormo model treats the aligned sequences as independent sequences with i.i.d. letters. Therefore, to ensure applicability of the method to real DNA sequences, models of higher order might be considered.

Hertz and Stormo describe two methods for estimating the *P*-value of a test statistic for a block (e.g. the information content of the block). One of the methods combines analytic large deviations techniques with computations; the other one is entirely computational (the computational method was recently improved by Keich and Nagarajan [96]). The first method is founded on the use of moment-generating function for the distribution of the test statistic; the authors devised an efficient method of computing this moment-generating function for the case of the information content statistic. The computational method uses the probability generating function approach, which is related to that of Staden [97] as described next. Both methods have their strengths and weaknesses, and the choice of the method should depend on the situation. Furthermore, Hertz and Stormo proposed a greedy algorithm for finding alignment blocks having maximal information content; this algorithm was implemented in the program called CONSENSUS [25]. Hertz and Stormo also argued that the *P*-value of an individual alignment might not be sufficient to assess the statistical significance of that alignment, since we are typically interested in the overall best alignment, given a large number of possible alignments of some fixed width. If P_{mat} is the *P*-value for an individual alignment matrix, then the overall *P*-value, P_{overall}, can be estimated [25] by

$$\begin{aligned} P_{\text{overall}} &= 1 - (1 - P_{\text{mat}})^A \\ &\approx 1 - \exp(-AP_{\text{mat}}) \\ &\approx AP_{\text{mat}}, \end{aligned} \quad (15)$$

where A is the number of possible alignments of the specified width w. This estimate is based on the assumption that different alignments are independent, which does not hold in reality. Thus, the obtained *P*-value should be used for benchmarking purposes only. Also, the approximation in (15) assumes that $P^{\text{mat}} \ll 1$. Using this approach, Hertz and Stormo obtained the *P*-value for the particular

type of extreme value distribution approximating the weight matrix score distribution (Claverie [98]).

In the two sections aforesaid we described two general methods applicable for statistical significance estimation of local ungapped multiple alignments. The methods are quite different in nature. While the choice of the specific method depends on the nature of sequences being aligned, there may exist situations when both methods are applicable. It would be interesting to compare their relative accuracy and performance. Also, it would be interesting to explore the applicability of (14) to gapped local alignments, and devise efficient methods for estimating the constants K_m and λ_m from simulated alignments.

ALIGNING TO POSITION-SPECIFIC SCORING MATRICES

Single PSSMs

A position-specific scoring matrix (PSSM), also called a position-specific weight matrix or a profile (in more general case), is frequently used to model different evolutionary conserved regions situated within protein and nucleotide sequences [99–102]. PSSMs are usually built from multiple alignments. The general purpose of PSSMs is to summarize the information contained in a multiple alignment, describing the propensities of different letters (nucleotides or amino acids) to occur in different positions (columns) of the alignment. For sequences of letters from the alphabet $\mathcal{A} = \{a_1, \ldots, a_q\}$, a PSSM of width w is usually a matrix $W = (w_{ij})$ with q rows and w columns The columns correspond to the positions in the multiple alignment, and the rows correspond to letters from the alphabet $\mathcal{A}$. When this matrix is aligned to a sequence without gaps, the score for this alignment is calculated as follows. If a letter of type a_i in the sequence is aligned with the jth column of the PSSM, then the score for this position is w_{ij}. The total alignment score is the sum of the scores over all the aligned PSSM positions. There are different approaches to deriving amino acid or nucleotide PSSMs given a set of aligned protein or DNA sequences (some of the approaches are discussed by Claverie and Audic [13, 98]). It is also possible to consider gapped sequence-to-PSSM alignments [100]; in this case, the PSSM would have to have an additional row specifying gap costs for each position.

The two major types of sequence-to-PSSM alignments are as follows. The first one is to align all of the PSSM's positions to the sequence with no gaps allowed in either PSSM or the sequence [13, 102–104]. The second one uses a local pairwise alignment algorithm to find the optimal local alignment with gaps [38, 100, 101]. The difference between this alignment and the conventional sequence alignment lies in the scoring system: for sequence-to-PSSM alignments, the position-specific scores for the PSSM are defined by the PSSM itself.

For the independence model for the sequence and ungapped sequence-to-PSSM alignments, it has been shown that, as the length of the sequence and the PSSM's width tend to infinity, the normalized score distribution for an individual alignment (at a fixed position) tends to the normal distribution, and the optimal alignment score distribution tends to the Gumbel distribution [105]. The latter result can be stated as follows. Consider a sequence of i.i.d. letters with length $n + w_n - 1$. The integer w_n depends on the integer n so that $\lim_{n\to\infty} w_n = \infty$. For every n, consider $W^{(n)}$, a PSSM having width w_n. We assume that the sequence $\{W^{(n)}\}$ is 'well-behaved', that is, satisfies a set of natural regularity conditions of Goldstein and Waterman [105]. Denote by $X_j^{(n)}$ the score of the ungapped alignment of $W^{(n)}$ to the target sequence such that the first position of $W^{(n)}$ is matched to the jth position of the target sequence. Let $Z_j^{(n)}$ corresponding normalized score having mean zero and variance 1:

$$Z_j^{(n)} = (X_j^{(n)} - \mu_n)/\sigma_n, \tag{16}$$

where μ_n and σ_n^2 are the mean and variance of $X_j^{(n)}$. Clearly, $Z_j^{(n)}$ is just the Z-score for $X_j^{(n)}$. Set $M_n = \max_{1\le j\le n} Z_j^{(n)}$; this is the normalized maximal alignment score, whose distribution we would like to approximate. With introduction of the quantities

$$a_n = (2\ln n)^{1/2},$$
$$c_n = (2\ln n)^{1/2} - (2\ln n)^{-1/2}(\ln\ln n + \ln(4\pi))/2,$$

Goldstein and Waterman [105] showed that, for any score s,

$$\Pr\{a_n(M_n - c_n) \le s\} \to \exp(-e^{-s}) \quad \text{as} \quad n \to \infty. \tag{17}$$

Thus, the asymptotic behavior of the normalized scores is the same for all the PSSMs obeying the regularity conditions, meaning that we do not need to estimate any PSSM- or sequence-specific constants for the limiting distribution. This situation resembles the use of Z-scores for pairwise local alignments, where the corresponding asymptotic results are also length- and composition-free [22, 81]. However, the convergence to the

Gumbel distribution in (17) may be rather slow [106]. The computational experiments of Claverie and Audic [13] showed that the Z-score distributions of best PSSM matches to nucleic acid and protein sequences resemble the extreme value distribution in practical situations (e.g. for a nucleotide sequence of length 250 and a PSSM of width 30).

Although the theoretical asymptotic behavior of the distribution of the best ungapped alignment score has been clarified in the limiting case, the problem of efficient numerical P-value estimation in practical situations is still open. As was mentioned earlier, the P-values for PSSM matches could be computed via the Monte Carlo importance sampling techniques [20]. A different approach to this problem was suggested by Claverie and Audic [13], who considered independence models. First, using an efficient iterative procedure, they compute the approximate distribution function, $f(s)$, for the score distribution of individual alignments. Integrating, they obtained the c.d.f., $F(s)$. Note that $1-F(s)$ is the P-value for an individual PSSM hit having score s. The probability that no individual sequence-to-PSSM match has a score $\geq s$ is given by

$$p(s) = (F(s))^{n-w+1}, \tag{18}$$

where n is the length of the sequence, and w is the width of the PSSM. Clearly, the P-value for the score s is then given by

$$P(s) = 1 - p(s).$$

Formula (18) implies that the PSSM alignment scores for different sequence positions are considered as independent, while overlapping matches do occur. Certainly, in reality these scores for overlapping matches will be dependent. However, it has been observed that treating possibly overlapping (and thus dependent) PSSM matches as independent does not introduce significant errors into the estimates [103, 106].

Interestingly, an efficient method for calculating $F(s)$ was proposed by Staden [97, 107]. He developed an algorithm which uses probability generating functions for exact (up to machine precision) calculation of the score distribution. The generating function, $G(x)$, of a discrete random variable X taking values in the set $\{0, 1, 2, \ldots\}$ with distribution (g_i), is defined by

$$G(x) = \sum_{i=1}^{\infty} g_i x^i.$$

A well-known property is that the generating function of a sum of independent discrete random variables is the product of the individual generating functions. Suppose that a position in an i.i.d. sequence is aligned to the jth column of the PSSM W with weights w_{ij}. To apply the generating function method, the weights in the PSSM should be scaled to positive integers. With this done, the alignment score becomes a discrete random variable with generating function

$$G_j(x) = \sum_{i=1}^{q} p_i x^{w_{ij}},$$

where (p_i) is the letter distribution for the sequence. Since all letters in the i.i.d. sequence are independent, the aggregate PSSM match score has generating function

$$G(x) = \prod_{j=1}^{w} G_j(x) = \prod_{j=1}^{w} \sum_{i=1}^{q} p_i x^{w_{ij}}. \tag{19}$$

The probability of getting the score s given by the multiplier of x^s in this expression unfolded. Thus, all we need is an algorithm for multiplying polynomials (more precisely, for multiplying the polynomial coefficients). Such an algorithm was also proposed by Staden [97].

Huang *et al.* [108] generalized Staden's approach to Markov (of order $m \geq 1$) models of sequences. This effort was undertaken in the context of algorithm development for predicting transcription factor binding sites in DNA sequences. One of the features of this project was the use of Markov chains as local background models. After scanning the sequence with a PSSM, the top 0.1% hits were selected. For each hit, the first order Markov model was built from a rather short (e.g. 1000 nt) subsequence centered around the hit. This model was used with the Markov generalization of (19) to assess the significance of the hit. Huang *et al.* reported that their method was 1000 times faster than assessment of statistical significance by direct simulation.

We now consider the second definition of the sequence-to-PSSM alignment, the one that allows gaps. The statistical significance of local sequence-to-PSSM alignments with gaps was empirically investigated by Altschul *et al.* [38] upon implementation of the program PSI-BLAST [38]. For the special way of choosing w_{ij} that Altschul *et al.* [38] considered, they hypothesized that if the protein sequence is described by an independence sequence

model, then formula (9) with the parameters K_g and λ_g can give fairly good approximations to the score distribution. A comparison with the simulated score distribution showed that such an approximation was indeed reasonably accurate. Because it would be too time-consuming to reestimate the gapped parameters for every new PSSM, Altschul *et al.* [38] proposed the following strategy. For a PSSM and a target sequence, λ_u is calculated using an extension of (10) (such an extension is discussed by Mott [14]). By scaling the PSSM elements (multiplying by a constant and rounding to the nearest integer), this parameter can be made equal to λ_u tabulated for pairwise alignments of protein sequences having typical amino acid composition using one of the conventional substitution matrices. It is then conjectured that, if gaps are allowed and gap costs stay the same, then λ_g for sequence-to-PSSM alignments for the scaled PSSM will be the same as λ_g for the pairwise sequence alignments estimated previously (see the section on the extension of the Karlin–Altschul statistic to gapped alignments [38, 40]). The values of λ_g estimated in such a way differ by <2% from the values λ_g known for pairwise alignments; the less critical parameter K_g can also be accurately estimated [38].

An effective approach to increase the quality of the *P*-value estimates is to take into account the actual amino acid composition of the target sequence, which may differ from the assumed average composition [40]. Such a compositional correction can be performed by rescaling the PSSM. This approach was taken by Schäffer *et al.* and implemented in the software package IMPALA [84]. To increase the search speed, the algorithm filtered out candidates for in-depth analysis and performed PSSM rescaling only if nearly-significant PSSM-sequence alignment was obtained in the first iteration. Mott [14] extended his GEM formulas for the parameters K_g and λ_g to sequence-to-profile alignments, using an explicit expression for the overall probability, $h(x)$, that the score between a randomly chosen profile position and sequence residue is x. Mott's approximation appears to work quite well, supporting the general claim that the score distribution of local sequence-to-PSSM alignments can be approximated by the Karlin–Altschul statistic.

Groups of PSSMs

If a sequence family is characterized by several simultaneously occurring motifs, then the fact that a sequence belongs to the family can be established by scanning the sequence using the PSSM models of the motifs. Even if the individual PSSM hits have relatively large *P*-values, the hits occurring simultaneously may provide sufficient evidence that the sequence indeed belongs to the specified family. To assess the importance of such combined evidence, it is necessary to estimate the statistical significance of simultaneous hits of several PSSMs. Computations show that using multiple-PSSM queries indeed gives better database search results [103, 106].

Here we consider only ungapped sequence-to-PSSM alignments. There exist several approaches to statistical significance estimation for multiple PSSM hits. The most straightforward is to estimate the distribution function of the sum of the best scores for each of the motifs [103]. This approach is directly related to the idea of using the statistic T_r (see formulas (6) and (7)) to assess the significance of occurrence of several high-scoring segments in single sequence analysis. As we know, this idea also works for pairwise alignments. Bailey and Gribskov [103] suggested the following algorithm to compute the distribution of the sum of the best scores. Consider a set of N PSSMs which define a sequence family. Denote by f_i, $i = 1, \ldots, N$, the best hit score for the ith PSSM, and put $g = \sum_{i=1}^{N} f_i$; note that all f_i are discrete random variables. For every i, the range of the hit scores is $r_i \le f_i \le R_i$. The distribution of g can be estimated in the following three steps.

- Estimate $C_i(x) = \Pr\{f_i \le x\}$, the c.d.f. of f_i, using the method of Staden [107].
- Estimate the distribution of f_i by

$$D_i(x) = \Pr\{f_i = x\} = \begin{cases} C_i(x) - C_i(x-1), & x > r_i, \\ C_i(r_i), & \text{otherwise.} \end{cases}$$

  (Note that the formula for $D_i(x)$ given by Bailey and Gribskov [103] is different; it holds true if $C_i(x) = \Pr\{f_i \ge x\}$.)
- Estimate $P(x)$, the c.d.f. for g, as follows. Denote by $p^{(k-1)}(x)$ the distribution of the sum of hit scores for the first $k-1$ PSSMs. Assuming that all scores are independent, the probability distribution of the sum of hit scores of the first k motifs is

$$p^{(k)}(x) = \Pr\left\{\sum_{i=1}^{k} f_i = x\right\} = \sum_{i=r_k}^{R_k} p^{(k-1)}(x-i) D_k(i).$$

The recursion starts with $p^{(1)}(x) = D_1(x)$. The c.d.f., $P(x)$, is computed by summing the elements of $p^{(n)}(x)$.

Another approach was also suggested by Bailey and Gribskov [103, 106], and is closely related to the first one. It is based on calculating the distribution of the sum of the scores, but this time the hit scores are the normalized scores. Like the normalization (16), the normalization of Bailey and Gribskov leads to standard Gumbel random variables. The two normalizations differ in that the latter one uses parameter values obtained from the fitting of the expression for the expected best PSSM hit score to the observed data. The normalization of Bailey and Gribskov was shown to give a better approximation to the empirical score distribution for PSSMs of finite widths [106]. The density of the standard Gumbel distribution is given by the formula

$$f(x) = d(\exp(-e^{-x}))/dx = \exp(-x - e^{-x}).$$

Integrating the product of such densities, we can obtain the probability, $\Pr\{C \geq x\}$, that the sum, C, of r independent Gumbel random variables, is not less than x:

$$\Pr\{C \geq x\} = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} e^{-y-z} \times (1 - \exp(-e^{y-x}))dx_{r-1} \ldots dx_1,$$

where $y = \sum_{i=1}^{r-1} x_i$ and $z = \sum_{i=1}^{r-1} e^{-x_i}$ (see Bailey and Gribskov [101]). For large s, and, therefore, for small P-values, this probability can be approximated by

$$\Pr\{C \geq s\} \approx \frac{e^{-s}s^{r-1}}{(r-1)!}.$$

Interestingly, the right-hand side of this formula coincides with that of (11).

The third method is based on combining the P-values for individual hits rather than on computing the distribution for the sum of hit scores. This approach was originally suggested by Claverie and Audic [13] (see also Staden [97, 107]), who defined the test statistic for the group of PSSMs as the product of the P-values for individual PSSM hits:

$$P(s) = \prod_{i=1}^{N} P_i(s).$$

Claverie and Audic proposed to determine the statistical significance of combined PSSM hits by comparing the value of $P(s)$ with a pre-selected threshold p (note that here we are interested in *small* values of the test statistic). Bailey and Gribskov [103, 104] further developed this approach, and considered the P-values for individual best PSSM hits, $P_i(s)$, as random variables which are (approximately) independent and uniformly distributed on [0, 1]. In this setup, the statistical significance for the values of $P(s)$ can be assessed via the expression

$$\Pr\{P(s) \leq p\} \approx p \sum_{i=0}^{N-1} \frac{(-\ln p)^i}{i!}.$$

Bailey and Gribskov [104] also suggested an algorithm for calculating P-values for the best individual PSSM hits.

The first (sum of scores) and the third (product of P-values) methods described above are quite general and are applicable with any sequence model, while the second one (sum of normalized scores) is based on analytic results for independence models. The three methods were compared by Bailey and Gribskov [103], who utilized independence models for protein sequences. They showed that the product of P-values produced the highest accuracy when used in the context of assignment of a protein to a given family. Also, the method gave high accuracy of P-value estimation, judging by the comparison of the predicted distribution of the P-values for the scores of multiple hits with the observed distribution of the same type P-values. This method was implemented in the well-known program MAST [104]. The sum of normalized scores gives slightly less accurate classification results, whereas the performance of the sum of scores method is significantly worse [103]. Interestingly, the P-value estimation accuracy of the sum of scores is almost as good as for the product of P-values, whereas the accuracy for the sum of normalized scores is lower. The better performance of the product of P-values is promising and suggests further development of this approach. An algorithmic method of combining individual P-values was proposed by Johansson *et al.* [109] and implemented in the software tool MSCAN.

A different approach to statistical significance estimation for multiple PSSMs, based on a stochastic model of a sequence of motifs, was suggested by Frith *et al.* [45]. They assumed that motif locations in nucleotide sequences are Poisson distributed, and each motif has a score drawn independently from some distribution. The score of a motif cluster was defined as the sum of PSSM scores for the motifs minus a linear 'gap penalty' for the spacers between motifs. In this setup, the cluster scores have extreme value distribution, and the authors give explicit expressions for the parameters. The authors considered three random models for the sequence: the

independence model, the fifth-order Markov chain and the independence model in which the DNA sequence is broken into fragments with different nucleotide compositions. The latter model appeared to give the best agreement between the estimated motif group *E*-values and the expected number of observed motif groups in real sequences. This approach to statistical significance estimation was implemented in the program named COMET [45].

STATISTICAL SIGNIFICANCE ESTIMATION FOR SEQUENCE ANALYSIS WITH HIDDEN MARKOV MODELS

Hidden Markov models and their generalizations are efficient and frequently used tools in bioinformatics [68]. They can be applied to problems ranging from gene finding [35, 37, 110, 111] to protein domain modeling [112, 113]. Different HMM packages may use different methods for the evaluation of statistical significance, and subsequently we describe three of such approaches.

The use of so-called profile HMMs [112–114] is a standard method for describing protein domains. For instance, protein domain profile HMMs are implemented in the two well-known packages, HMMer (unpublished) and SAM [115]. Although the modeling principles are the same, the procedures used in HMMer and SAM to estimate the statistical significance of high scores are completely different. Before describing them, we give the definition for the HMM score. An HMM score for a sequence is defined as

$$S = \log_z \frac{\Pr\{seq|HMM\}}{\Pr\{seq|null\}}. \qquad (20)$$

In this definition, $\Pr\{\text{seq}\,|\,HMM\}$ is the probability that the sequence has been generated by the HMM, and $\Pr\{seq\,|\,null\}$ is the probability that the sequence has been generated by the null model. The null model is usually a very simple (one hidden state) HMM, which is equivalent to the random independence model. The log base, z, can be any real number >1; HMMer uses $z=2$ (binary logs), and SAM uses $z=e$ (natural logs). Thus the HMM score is a log-odds score.

The definition of the score used by HMMer is in fact more complex [116], but the basic idea stays the same. HMMer approximates the distribution of the log-odds score by the extreme value distribution, using the maximum likelihood method. While it has been noticed that the fit is never perfect, the approximation for the right tail is satisfactory [116]. HMMer returns *E*-values as the estimates of statistical significance.

SAM uses an algorithmic significance test of Miloslavljević and Jurka [117, 118] to estimate statistical significance. According to this method, the probability of getting a score larger than s is not larger than z^{-s} (z is the logarithm base), if the null model is reasonably accurate [117, 118]. To account for multiple hits during a database search, Barrett *et al.* [117] introduced the parameter N, the number of 'individual placements' of the model. There exist several approaches to choosing N. For instance, Barrett *et al.* [117] suggest the following assignment:

$$N = \sum_{i=1}^{S} \max(n_i - h, 1),$$

where S is the number of sequences in the database, n_i is the length of the ith sequence and h is the length of the profile HMM. For given N,

$$\Pr\{s_i \geq s \text{ for some } i\} \leq \sum_{i=1}^{N} \Pr\{s_i \geq s\} \leq Nz^{-s},$$

where s_i is the score for the ith placement. Thus, for a certain level of statistical significance σ, a score s such that $\sigma \geq Nz^{-s}$ will indicate significance of the hit.

The third approach to *P*-value estimation is implemented in the gene finder EasyGene [111]. This program builds an HMM for prokaryotic genes using an automatically extracted training set, and scores putative genes (ORFs) with this HMM. The score, β, is a log-odds score similar to (20). The authors show that this score can be transformed into what they call a standard score, Γ, with approximately standard normal distribution. For a random sequence of a given length, the expected number of ORFs of length l' can be written as

$$N(l') = \exp(A - Bl'),$$

where the constants A and B can be found from linear regression of the logarithm of the number of ORFs against the ORF length (which is measured in codons) [111]. Therefore, the expected number, $C(l', \Gamma)$, of ORFs of length l' with scores higher than Γ can be estimated by

$$C(l', \Gamma) = \exp(A - Bl')(1 - \Phi(\Gamma)), \qquad (21)$$

where Φ is the c.d.f. of the standard normal distribution. The formula for the total expected number of ORFs of any length with scores higher than Γ can be derived from (21) by summing over l'.

CONCLUSION

Although many results have been obtained in the area of assessment of statistical significance of biosequence analysis, many important questions remain open. Even in the 'classic' area of local pairwise alignments, the search for efficient parameter estimation methods continues, and the complete rigorous theory for gapped alignments is yet to be developed. With global alignments, the picture is fuzzy both for pairwise and for multiple alignments, and general efficient methods are in fact absent. As for the local multiple alignments, the existing approaches need to be compared and evaluated, and new ideas will probably emerge. In pattern matching and sequence-to-PSSM alignments, further research is needed to investigate better choices for the null model. Non-PSSM motif models will apparently require creation of specific methods for the statistical significance estimation of meaningful hits. Similarly, comparison of the existing approaches and development of generally applicable methods of statistical significance estimation would be very desirable in the area of sequence analysis using hidden Markov sequence models. The problem of how to combine different pieces of evidence of homology in a theoretically sound manner using the notion of statistical significance remains largely open. However, some promising approaches have been developed, such as the methods of the *P*-value estimation for multiple PSSM hits. The methods based on the sum of scores and the product of *P*-values are by their nature quite general, and may be successfully used in other contexts.

As we have mentioned in the introduction, each type of sequence analysis problem requires specific methods of statistical significance assessment. Of course, in this review we were unable to touch upon all the techniques that have been developed for a variety of different contexts. But the selected topics discussed above illustrate the major problem settings and directions of research, which may provide useful starting points for approaching yet unsolved problems. The common features of the *P*-value estimation techniques presented in this article may be summarized as follows.

- Parametric continuous distributions can often be fitted well to the discrete score distributions encountered in practice. This is convenient, since analytic functions are easier to analyze than discrete distributions, studied in many cases only by simulations.
- The extreme value (Gumbel) distribution gives a good approximation in many, but not all, situations.
- A good way of obtaining high-quality approximations to distributions of scores is to undertake a rigorous study of a simple special case, and then conjecture a reasonable extension of the initial formula to the general situation. Examples: certain results from single-sequence and pairwise studies may be generalized to the multiple sequence case; some results for sequences can be extended to profiles.
- Development of efficient computational procedures for the statistical significance estimation is always a valuable research topic. The methods which work for a variety of problems are of special importance.

The goal of the future studies in the filed of *P*-value estimation in biosequence analysis will be to further develop the theoretical foundations of practical methods of statistical significance assessment, as well as to suggest new and improve the existing methods. The studies will naturally concentrate on the open or partially solved problems which are of primary importance, such as statistical significance estimation for profile-to-profile alignments and for combinations of homology evidence obtained from different sources. But even for the well-studied problems, there is room for improvement, and much work is needed to bring the notion of statistical significance as close as possible to biological significance, thus maximizing its utility in the structural, functional and evolutionary studies of biological macromolecules.

Key Points

- The primary goal of statistical significance estimation is to identify candidate sequence segments for in depth theoretical and experimental analysis.
- Practically important statistical significance estimation techniques in sequence analysis frequently combine rigorous mathematical approaches with empirical "rules of thumb".
- Choice of the *P*-value cutoff determines the ratio of false positives and false negatives in candidate selection.

Acknowledgements
This work was supported in part by the grant awarded to MB by the US National Institutes of Health.

References

1. Dandekar T, Snel B, Huynen M, *et al*. Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem Sci* 1998;**23**:324–28.
2. Enright AJ, Iliopoulos I, Kyrpides NC, *et al*. Protein interaction maps for complete genomes based on gene fusion events. *Nature* 1999;**402**:86–90.
3. Wagner A. Genes regulated cooperatively by one or more transcription factors and their identification in whole eukaryotic genomes. *Bioinformatics* 1999;**15**:776–84.
4. Altschul SF, Erickson BW. Significance of nucleotide sequence alignments: a method for random sequence permutation that preserves dinucleotide and codon usage. *Mol Biol Evol* 1985;**2**:526–38.
5. Karlin S, Burge C, Campbell AM. Statistical analyses of counts and distributions of restriction sites in DNA sequences. *Nucleic Acids Res* 1992;**20**:1363–70.
6. Karlin S, Ghandour G. Comparative statistics for DNA and protein sequences: single-sequence analysis. *Proc Natl Acad Sci USA* 1985;**82**:5800–4.
7. Karlin S, Ost F, Blaisdell BE. Patterns in DNA and amino acid sequences and their statistical significance. In: Waterman MS (ed). *Mathematical Methods for DNA Sequences*. Boca Raton: CRC Press, 1989;133–57.
8. Lipman DJ, Wilbur WJ, Smith TF, *et al*. On the statistical significance of nucleic acid similarities. *Nucleic Acids Res* 1984;**12**:215–26.
9. Altschul SF, Boguski MS, Gish W, *et al*. Issues in searching molecular sequence databases. *Nature Genet* 1994;**6**:119–29.
10. Waterman MS. Sequence alignments. In: Waterman MS (ed). *Mathematical Methods for DNA Sequences*. Boca Raton: CRC Press, 1989;53–92.
11. Waterman MS. Consensus patterns in sequences. In: Waterman MS (ed). *Mathematical Methods for DNA Sequences*. Boca Raton: CRC Press, 1989;93–115.
12. Lehmann EL. *Testing Statistical Hypotheses*. New York: Wiley, 1986.
13. Claverie JM, Audic S. The statistical significance of nucleotide position-weight matrix matches. *CABIOS* 1996;**12**:431–39.
14. Mott R. Accurate formula for *p*-values of gapped local sequence and profile alignments. *J Mol Biol* 2000;**300**: 649–59.
15. Pearson WR. Empirical statistical estimates for sequence similarity searches. *J Mol Biol* 1998;**276**:71–84.
16. Waterman MS, Vingron M. Sequence comparison significance and Poisson approximation. *Statist Sci* 1994;**9**: 367–81.
17. Reich JG, Drabsch H, Däumler A. On the statistical assessment of similarities in DNA sequencies. *Nucleic Acids Res* 1984;**12**:5529–43.
18. Waterman MS, Vingron M. Rapid and accurate estimates of statistical significance for sequence database searches. *Proc Natl Acad Sci USA* 1994;**91**:4625–28.
19. Liu JS. *Monte Carlo Strategies in Scientific Computing*. New York: Springer, 2001.
20. Barash Y, Elidan G, Kaplan T, *et al*. CIS: compound importance sampling method for protein-DNA binding site *p*-value estimation. *Bioinformatics* 2005;**21**:596–600.
21. Coles S. *An Introduction to Statistical Modeling of Extreme Values*. London: Springer, 2001.
22. Comet JP, Aude JC, Glémet E, *et al*. Significance of *Z*-value statistics of Smith-Waterman scores for protein alignments. *Computers Chem* 1999;**23**:317–31.
23. Bastien O, Aude JC, Roy S, *et al*. Fundamentals of massive automatic pairwise alignments of protein sequences: theoretical significance of *Z*-value statistics. *Bioinformatics* 2004;**20**:534–37.
24. Smith TF, Waterman MS, Burks C. The statistical distribution of nucleic acid similarities. *Nucleic Acids Res* 1985;**13**:645–56.
25. Hertz GZ, Stormo GD. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics* 1999;**15**:563–77.
26. Mott R. Maximum-likelihood estimation of the statistical distribution of Smith–Waterman local sequence similarity scores. *Bull Math Biol* 1992;**54**:59–75.
27. Altschul SF, Gish W. Local alignment statistics. *Meth Enzymol* 1996;**266**:460–80.
28. Reinert G, Schbath S, Waterman MS. Probabilistic and statistical properties of words: an overview. *J Comp Biol* 2000;**7**:1–46.
29. Tavaré S, Giddins BW. Some statistical aspects of the primary structure of nucleotide sequences. In: Waterman MS (ed). *Mathematical Methods for DNA Sequences*. Boca Raton: CRC Press, 1989;117–33.
30. Borodovskii MY, Sprizhitskii YA, Golovanov EI, *et al*. Statistical patterns in the primary structures of functional regions in the genome of *E. coli*. 3. Computer recognition of coding regions. *Mol Biol* 1986;**20**:1144–50.
31. Borodovskii MY, Sprizhitskii YA, Golovanov EI, *et al*. Statistical patterns in the primary structures of functional regions in the genome of *E. coli*. 1. Frequency characteristics. *Mol Biol* 1986;**20**:826–33.
32. Borodovskii MY, Sprizhitskii YA, Golovanov EI, *et al*. Statistical patterns in the primary structures of functional regions in the genome of *E. coli*. 2. Nonuniform Markov models. *Mol Biol* 1986;**20**:833–40.
33. Tavaré S, Song B. Codon preference and primary sequence structure in protein-coding regions. *Bull Math Biol* 1989; **51**: 95–115.
34. Braun JV, Müller HG. Statistical methods for DNA sequence segmentation. *Statist Sci* 1998;**13**:142–62.
35. Burge C, Karlin S. Prediction of complete gene structures in human genomic DNA. *J Mol Biol* 1997;**268**:78–94.
36. Krogh A, Mian IS, Haussler D. A hidden Markov model that finds genes in *E. coli* DNA. *Nucleic Acids Res* 1994;**22**: 4768–78.
37. Lukashin AV, Borodovsky M. GeneMark.hmm: new solutions for gene finding. *Nucleic Acids Res* 1998;**26**: 1107–15.
38. Altschul SF, Madden TL, Schäffer AA. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 1997;**25**:3389–3402.
39. Robinson AB, Robinson LR. Distribution of glutamine and asparagine residues and their near neighbors in peptides and proteins. *Proc Natl Acad Sci USA* 1991;**88**:8880–84.

40. Schäffer AA, Aravind L, Madden TL, *et al*. Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic Acids Res* 2001;**29**:2994–3005.
41. Waterman M. Estimating statistical significance of sequence alignments. *Philosophical Transactions: Biological Sciences* 1994; **344**:383–90.
42. Goldstein L, Waterman MS. Poisson, compound Poisson and process approximations for testing statistical significance in sequence comparisons. *Bull Math Biol* 1992;**54**:785–812.
43. Karlin S, Altschul SF. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc Natl Acad Sci USA* 1990;**87**: 2264–68.
44. Arratia R, Gordon L, Waterman MS. The Erdös-Rényi law in distribution, for coin tossing and sequence matching. *Ann Statist* 1990;**18**:539–70.
45. Frith MC, Spouge JL, Hansen U, *et al*. Statistical significance of clusters of motifs represented by position specific scoring matrices in nucleotide sequences. *Nucleic Acids Res* 2002;**30**: 3214–24.
46. Pavesi G, Mauri G, Pesole G. *In silico* representation and discovery of transcription factor binding sites. *Brief Bioinform* 2004;**5**:217–36.
47. Arratia R, Martin D, Reinert G, *et al*. Poisson process approximation for sequence repeats, and sequencing by hybridization. *J Comp Biol* 1996;**3**:425–63.
48. Mariño-Ramírez L, Spouge JL, Kanga GC, *et al*. Statistical analysis of over-represented words in human promoter sequences. *Nucleic Acids Res* 2004;**32**:949–58.
49. Sinha S, Tompa M. Discovery of novel transcription factor binding sites by statistical overrepresentation. *Nucleic Acids Res* 2002;**30**:5549–60.
50. Sumazin P, Chen GX, Hata N, *et al*. DWE: discriminating word enumerator. *Bioinformatics* 2005;**21**:31–38.
51. Vinga S, Almeida J. Alignment-free sequence-comparison—a review. *Bioinformatics* 2003;**19**:513–23.
52. Schbath S. An overview on the distribution of word counts in Markov chains. *J Comp Biol* 2000;**7**:193–201.
53. Robin S, Schbath S. Numerical comparison of several approximations of the word count distribution in random sequences. *J Comp Biol* 2001;**8**:349–59.
54. Karlin S, Brendel V. Chance and statistical significance in protein and DNA sequence analysis. *Science* 1992;**257**: 39–49.
55. Karlin S, Dembo A, Kawabata T. Statistical composition of high-scoring segments from molecular sequences. *Ann Statist* 1990;**18**:571–81.
56. Iglehart DL. Extreme values in the GI/G/1 queue. *Ann Math Statist* 1972;**43**:627–35.
57. Karlin S, Dembo A. Limit distributions of maximal segmental score among Markov-dependent partial sums. *Adv Appl Probab* 1992;**24**:113–40.
58. Karlin S, Altschul SF. Applications and statistics for multiple high-scoring segments in molecular sequences. *Proc Natl Acad Sci USA* 1993;**90**:5873–77.
59. Karlin S. Statistical studies of biomolecular sequences: score-based methods. *Philosophical Transactions: Biological Sciences* 1994;**344**:391–402.
60. Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol* 1970;**48**:443–53.
61. Gotoh O. An improved algorithm for matching biological sequences. *J Mol Biol* 1982;**162**:705–8.
62. Webber C, Barton GJ. Estimation of *P*-values for global alignments of protein sequences. *Bioinformatics* 2001;**17**: 1158–67.
63. Dayhoff MO, Schwartz RM, Orcutt BC. A model of evolutionary change in proteins. In: Dayhoff MO (ed). *Atlas of Protein Sequence and Structure*; **Vol. 5**. Washington, DC: National Biomedical Research Foundation, 1978:345–58.
64. Henikoff S, Henikovv JG. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci USA* 1992; **89**:10915–19.
65. Gonnet GH, Cohen MA, Benner SA. Exhaustive matching of the entire protein sequence database. *Science* 1992;**256**: 1443–45.
66. Smith TF, Waterman MS. Identification of common molecular subsequences. *J Mol Biol* 1981;**147**:195–97.
67. Altschul SF. Amino acid substitution matrices from an information theoretic perspective. *J Mol Biol* 1991;**219**: 555–65.
68. Durbin R, Eddy SR, Krogh A, *et al*. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge: Cambridge University Press, 1991.
69. Dembo A, Karlin S, Zeitouni O. Limit distribution of maximal non-aligned two-sequence segmental score. *Ann Probab* 1994;**22**:2022–39.
70. Altschul SF, Bundschuh R, Olsen R, *et al*. The estimation of statistical parameters for local alignment score distributions. *Nucleic Acids Res* 2001;**29**:351–61.
71. Bundschuh R. Rapid significance estimation in local sequence alignment with gaps. *J Comp Biol* 2002;**9**: 243–60.
72. Siegmund D, Yakir B. Approximate *p*-values for local sequence alignments. *Ann Statist* 2000;**28**:657–80.
73. Siegmund D, Yakir B. Correction: approximate *p*-values for local sequence alignments (vol 28, pg 657, 2000). *Ann Statist* 2003;**31**:1027–31.
74. Arratia R, Waterman MS. A phase transition for the score in matching random sequences allowing deletions. *Ann Appl Probab* 1994;**4**:200–25.
75. Park Y, Spouge JL. The correlation error and finite-size correction in an ungapped sequence alignment. *Bioinformatics* 2002;**18**:1236–42.
76. Spang R, Vingron M. Statistics of large-scale sequence searching. *Bioinformatics* 1998;**14**:279–84.
77. Spouge JL. Finite-size corrections to Poisson approximations of rare events in renewal processes. *J Appl Probab* 2001; **38**:554–69.
78. Grossmann S, Yakir B. Large deviations for global maxima of independent superadditive processes with negative drift and an application to optimal sequence alignments. *Bernoulli* 2004;**10**:829–45.
79. Bailey TL, Gribskov M. Estimating and evaluating the statistics of gapped local-alignment scores. *J Comp Biol* 2002; **9**:575–93.
80. Mott R, Tribe R. Approximate statistics of gapped alignments. *J Comp Biol* 1999;**6**:91–112.
81. Bacro JN, Comet JP. Sequence alignment: an approximation law for the *Z*-value with applications to databank scanning. *Computers Chem* 2001;**25**:401–10.

82. Altschul SF, Gish W, Miller W, *et al*. Basic local alignment search tool. *J Mol Biol* 1990;**215**:403–10.
83. Pearson WR, Lipman DJ. Improved tools for biological sequence comparison. *Proc Natl Acad Sci USA* 1988;**85**:2444–48.
84. Schäffer AA, Wolf YI, Ponting CP, *et al*. IMPALA: matching a protein sequence against a collection of PSI-BLAST-constructed position-specific score matrices. *Bioinformatics* 1999;**15**:1000–11.
85. Feng DF, Doolittle RF. Progressive alignment of amino acid sequences and construction of phylogenetic trees from them. *Meth Enzymol* 1996;**266**:368–82.
86. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* 1994;**22**: 4673–80.
87. Notredame C, Higgins DG, Heringa J. T-Coffee: a novel method for fast and accurate multiple sequence alignment. *J Mol Biol* 2000;**302**:205–17.
88. Edgar RC. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 2004;**5**, Art. No. 113.
89. Katoh K, Kuma K, Toh H, *et al*. MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res* 2005;**33**:511–18.
90. Subramanian AR, Weyer-Menkhoff J, Kaufmann M, *et al*. DIALIGN-T: an improved algorithm for segment-based multiple sequence alignment. *BMC Bioinformatics* 2005;**6**, Art. No. 66.
91. Schuler GD, Altschul SF, Lipman DJ. A workbench for multiple alignment construction and analysis. *PROTEINS* 1991;**9**:180–90.
92. Bailey TL, Elkan C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. In: Altman R, Brutlag D, Karp P, Lathrop R, Searls D (eds). CA: AAAI Menlo Park, 1994;28–36.
93. Lawrence CE, Altschul SF, Boguski MS. Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science* 1993;**262**:208–14.
94. Zhang Y, Waterman MS. An Eulerian path approach to local multiple alignment for DNA sequences. *Proc Natl Acad Sci USA* 2005;**102**:1285–90.
95. Thompson JD, Plewniak F, Ripp R, *et al*. Towards a reliable objective function for multiple sequence alignments. *J Mol Biol* 2001;**314**:937–51.
96. Keich U, Nagarajan N. A fast reliable algorithms to estimate the p-value of the multinomial llr statistic. *Lect Notes Comp Sci* 2004;**3240**:111–22.
97. Staden R. Methods for calculating the probabilities of finding patterns in sequences. *CABIOS* 1989;**5**:89–96.
98. Claverie JM. Some useful statistical properties of position weight matrices. *Computers Chem* 1994;**18**:287–94.
99. Gribskov M, Lüthy R, Eisenberg D. Profile analysis. *Meth Enzymol* 1990;**183**:146–59.
100. Gribskov M, McLachlan AD, Eisenberg D. Profile analysis: detection of distantly related proteins. *Proc Natl Acad Sci USA* 1987;**84**:4355–58.
101. Gribskov M, Veretnik S. Identification of sequence patterns with profile analysis. *Meth Enzymol* 1996;**266**: 198–212.
102. Stormo GD. DNA binding sites: representation and discovery. *Bioinformatics* 2000;**16**:16–23.
103. Bailey TL, Gribskov M. Methods and statistics for combining motif match scores. *J Comp Biol* 1998;**5**: 211–21.
104. Bailey TL, Gribskov M. Combining evidence using p-values: application to sequence homology searches. *Bioinformatics* 1998;**14**:48–54.
105. Goldstein L, Waterman MS. Approximations to profile score distribution. *J Comp Biol* 1994;**1**:93–104.
106. Bailey TL, Gribskov M. Score distributions for simultaneous matching to multiple motifs. *J Comp Biol* 1997;**4**: 45–59.
107. Staden R. Searching for patterns in protein and nucleic acid sequences. *Meth Enzymol* 1990;**183**: 193–211.
108. Huang HY, Kao MCH, Zhou XH, *et al*. Determination of local statistical significance of patterns in Markov sequences with application to promoter element identification. *J Comp Biol* 2004;**11**:1–14.
109. Johansson Ö, Alkema W, Wasserman WW, *et al*. Identification of functional clusters of transcription factor binding motifs in genome sequences: the MSCAN algorithm. *Bioinformatics* 2003;**19**:i169–76.
110. Azad RK, Borodovsky M. Probabilistic methods of identifying genes in prokaryotic genomes: connections to the HMM theory. *Brief Bioinform* 2004;**5**: 118–30.
111. Larsen TS, Krogh A. EasyGene—a prokaryotic gene finder that ranks ORFs by statistical significance. *BMC Bioinformatics* 2003;**4**, Art. No. 21.
112. Baldi P, Chauvin Y, Hunkapiller T, *et al*. Hidden Markov models of biological primary sequence information. *Proc Natl Acad Sci USA* 1994;**91**:1059–63.
113. Krogh A, Brown M, Mian IS, *et al*. Hidden Markov models in computational biology. Applications to protein modeling. *J Mol Biol* 1994;**235**:1501–31.
114. Eddy SR. Profile hidden Markov models. *Bioinformatics* 1998;**14**:755–63.
115. Hughey R, Krogh A. Hidden Markov models for sequence analysis: extension and analysis of the basic method. *CABIOS* 1996;**12**:95–107.
116. ftp://ftp.genetics.wustl.edu/pub/eddy/hmmer/CURRENT/Userguide.pdf. Last accessed 13 January 2006.
117. Barrett C, Hughey R, Karplus K. Scoring hidden Markov models. *CABIOS* 1997;**13**:191–99.
118. Milosavljević A, Jurka J. Discovering simple DNA sequences by the algorithmic significance method. *CABIOS* 1993;**9**:407–11.