A Poisson model for gapped local alignments

Dirk Metzler[1, 2], Steffen Grossmann, and Anton Wakolbinger
Department of Mathematics, Goethe-Universität,
Postfach 11 19 32, D-60054 Frankfurt am Main, Germany

Abstract

High score alignments of DNA sequences give evidence of a common ancestry or function. It is therefore natural to ask whether an observed high score could have arisen by pure chance, and to explore what the high score alignments look like under the null hypothesis of unrelated sequences. We introduce and investigate a simple Poisson model which reflects important features of high score gapped local alignments of independent sequences.

Keywords: sequence alignment, gapped local alignment, Poisson model, high score asymptotics.

[1]corresponding author; e-mail: dmetzler@math.uni-frankfurt.de
[2]supported by the German Science Foundation (DFG) in the program "Interacting stochastic systems of high complexity"

1 Introduction

An important problem in modern genetics is to detect similarities between two DNA (or protein) sequences $(q_1, \ldots, q_n)$ and $(r_1, \ldots, r_m)$ of length n and m, say, which indicate an evolutionary relationship between the two sequences or parts of them. (Here, the q_i and r_k are from a finite alphabet.) Such a similarity could consist in a *gapless local alignment* $\Xi(i, h, k)$, relating two equally long substrings $(q_i, \ldots, q_{i+k})$, $(r_h, \ldots, r_{h+k})$ of the two sequences which have many *matches* $q_{i+l} = r_{h+l}$ and only few *mismatches* $q_{i+l} \neq r_{h+l}$. It will be convenient to think of $\Xi(i, h, k)$ as a diagonal line $((i, h), (i + 1, h + 1), \ldots, (i + k, h + k))$ in the grid $\{1, \ldots, n\} \times \{1, \ldots, m\}$. The similarity of the two sequences along the gapless local alignment is quantified by the *score* $s = \sum_{l=0}^{k} \left(\mathbf{1}_{\{q_{i+l}=r_{h+l}\}} - \mu \mathbf{1}_{\{q_{i+l}\neq r_{h+l}\}}\right)$, with some *mismatch penalty* $\mu > 0$. A high score indicates that the two fragments $(q_i, \ldots, q_{i+k})$ and $(r_h, \ldots, r_{h+k})$ might stem from a common ancestral sequence, where a few positions were affected by substitutions in the course of evolution. However, apart from substitutions also other types of mutations are relevant: the fragments could have been subject to insertions and deletions of shorter pieces. This can be taken into account by considering *gapped local alignments*. A gapped local alignment is a sequence of gapless local alignments $\Xi(i_0, h_0, k_0), \ldots, \Xi(i_j, h_j, k_j)$ with $i_g + k_g < i_{g+1}$, $h_g + k_g < h_{g+1}$, and $(i_g + k_g, h_g + k_g) \neq (i_{g+1} - 1, h_{g+1} - 1)$ for each $g \leq j$.

For two fragments captured by a gapped local alignment, there is evidence for a close relationship if, again, there are many matches and only a few mismatches, and if the number j of gaps is not too high and the gaps are short. Therefore it makes sense to penalize each gap with a *gap open penalty* Δ and a *gap extension penalty* δ multiplied with the *length* $[i_{g+1} - (i_g + k_g) - 1] + [h_{g+1} - (h_g + k_g) - 1]$ of the gap. One obtains the score of a gapped alignment by subtracting the sum of the gap penalties from the sum of the scores of its gapless building blocks. There are efficient algorithms to find the gapped local alignment which maximizes the score for a given pair of sequences and given scoring parameters, even for a more general class of scoring schemes (Smith and Waterman, 1981, Altschul et al., 1990). One of the crucial questions is to find out whether an observed high score could have come up also by mere chance, in other words, one has to study the distribution of high scores under the null hypothesis of independent sequences.

Dembo, Karlin and Zeitouni (1994) computed the tail of the distribution of the maximal score of *gapless* local alignments of two i.i.d. sequences over a finite alphabet. They considered gapless local alignments which cannot be improved by extending or shortening in forward or backward direction, and showed that the probability that a fixed pair of positions is the starting point of a segment which is in this sense locally optimal and whose score exceeds b is asymptotically between $k_1 e^{-\lambda b}$ and $k_2 e^{-\lambda b}$ as $b \to \infty$. Thereby, μ is reqired to be so large that the expected score of an alignment of unrelated sequences is negative. In what follows, this will always be assumed. Also, in Dembo et al. (1994) the rate λ as well as the bounds k_1 and k_2 are computed in terms of the parameters of the scoring scheme. The constants k_1 and k_2 are in general different from each other if the possible scores lie on a lattice. We will ignore these lattice effects and assume $k_1 = k_2 =: k$.

Recently, Siegmund and Yakir (2000) obtained a tail asymptotics for the score distribution of gapped local alignments. One of our aims is to get a better understanding of the structure of their result by considering a simpler model which concentrates on the essential features. Intuitively, for long sequences the "good" gapless alignments should be scattered in a Poissonian way as diagonal pieces into the rectangular grid spanned by the two sequences, and the high score *gapped* alignment should correspond to certain finite configurations of such diagonal lines (see also Waterman and Vingron, 1994) using essential parts of the locally optimal gapless local alignments (see Figure 1.) To avoid effects of discreteness we replace the grid $\{1, \ldots, n\} \times \{1, \ldots, m\}$ by the continous rectangle $[0, n] \times [0, m]$.

2 A Poisson point model for local alignment

Let Φ be a homogeneous Poisson process on $\mathbb{R}_+ \times \mathbb{R}_+$ with intensity k, and attach independent marks to the points of Φ which are exponentially distributed with parameter λ. (The meaning of k and λ is decribed in the introduction.)

We write Ψ for the marked point process arising in this way. In the following we will refer to the points together with their marks as "good pieces", since they mimic the gapless local alignments that reach a certain score.

We say that $\mathbf{x} = (x_0, \ldots, x_j)$ is a *path* in Φ *with j steps* (or a j-step path for short) if $x_i \in \Phi$ and $x_{i-1} \leq x_i$ for $i = 1, \ldots, j$, where $\leq$ denotes component-wise ordering in $\mathbb{R}^2$. Also, we call $\mathbf{a} = ((x_0, r_0), \ldots, (x_j, r_j))$ a *(marked) path* in Ψ if $(x_0, \ldots, x_j)$ is a path in Φ and r_i is the mark attached to x_i.

Now we have to translate the alignment scoring scheme into our Poisson model. The mismatch penalty μ is implicit in k and λ. The gap open penalty enters directly into the Poisson model: each step has a basic cost Δ. We propose to describe the additional cost for a step from x_{i-1} to x_i by the nonnegative random variable $R_{x_{i-1}}(x_i - x_{i-1})$, where, given Φ, $\{R_x : x \in \Phi\}$ is a family of i.i.d. copies of a random function $R : \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R}_+$ which has the following scaling property: For some constant $\omega > 0$, and all $l > 0$,

$$\text{the expected area of } \{y : R(y) \leq l\} \text{ equals } \omega l^2/2 \tag{1}$$

The intuition behind this scaling property is as follows: From (the end point of) any ungapped local alignment one can reach a 2-dimensional range of points for net costs of at most 1. The range which can be reached for costs of at most l should have a similar shape, scaled with l, and therefore it should have approximately the l^2-fold area. The parameter ω ocurring in (1) plays a crucial role in the whole picture; we will come back to this later. For the moment let us only mention that a simple (non-random) example for R would consist in the Euclidean norm $R(y) := ||y||$, which leads to $\omega = \pi/2$.

The score of a marked path $\mathbf{a} = (\mathbf{x}, \mathbf{r})$ in Ψ is now defined as

$$S(\mathbf{a}) := \sum_{i=0}^{j} r_j - \sum_{i=1}^{j} R_{x_{i-1}}(x_i - x_{i-1}) - j \cdot \Delta \tag{2}$$

We call the three summands appearing in (2) the *point quality, distance penalty* and *step penalty* of the marked path $\mathbf{a}$, and note that $\omega \ l \ dl \cdot k^2 \ dx_0$ is the expected number of one-step paths starting in dx_0 and with distance penalty in dl.

The point qualities in the Poisson model mimic the scores of locally optimal ungapped local alignments (LOULAs). Note that, as depicted in Figure 1, the diagonal pieces in the optimal gapped local alignment in general do not perfectly coincide with the LOULAs they intersect with. Near its gaps the optimal alignment might extend along a diagonal beyond a LOULA or miss a part of a LOULA. The distance penalty R in the Poisson model is thought to capture also these extra costs.

For $j = 0, 1, \ldots$ and $B \subset \mathbb{R}_+$ we write $M(B, j)$ for the expected number of j-step paths starting per unit volume and with score in B, and $M(B) := \sum_{j=0}^{\infty} M(B, j)$. The following proposition gives the expected number of paths of a certain score.

Proposition *For $j \geq 1$,*

$$M(ds, j) = ds \frac{k\lambda\, e^{-\lambda s}}{j!\,(2j-1)!} \left(\frac{\omega k\, e^{-\lambda\Delta}}{\lambda} \right)^j \int_0^\infty l^{2j-1}(s + j\Delta + l/\lambda)^j e^{-l} dl. \quad (3)$$

Proof The expected number of j-step paths starting per unit volume and having a distance penalty $\sum l_i$ in dl is

$$\frac{\partial}{\partial l} k \cdot \int_0^l \int_0^{l-l_1} \cdots \int_0^{l-l_1-\cdots-l_{j-1}} (k \cdot \omega)^j \cdot \prod_{i=1}^{j} l_i \, dl_j \cdots dl_1 \cdot dl = \frac{k^{j+1}\omega^j l^{2j-1}}{(2j-1)!}\, dl$$

Since the point quality of a of j-step path is Gamma($j+1$,λ)-distributed, the expected number of j-step paths starting per unit volume, having a point quality in $ds + j\Delta + l$ and a distance penalty in dl is

$$\frac{\lambda^{j+1}}{j!} e^{-\lambda(s+j\Delta+l)} (s + j\Delta + l)^j ds \frac{k^{j+1}\omega^j l^{2j-1}}{(2j-1)!}\, dl \quad (4)$$

Integrating (4) over l we arrive at (3). □

Expanding the term $(s+j\Delta+l/\lambda)^j$ on the r.h.s. of (3), we see that the leading term in the integral contributes $s^j(2j-1)!$, whereas all the other summands are $O(s^{j-1}\Delta^j)$, which by assumption is uniformly $o(s^j)$ as $s \to \infty$. This yields the following

Corollary *For fixed $j \in \{0, 1, \ldots\}$,*

$$M(ds, j) \sim k\lambda\, e^{-\lambda s} ds\, (\omega k s\, e^{-\lambda\Delta}/\lambda)^j/j! \quad \text{as } s \to \infty \quad (5)$$

uniformly in $s \geq \eta e^{\lambda\Delta}$, where η is an arbitrary (but fixed) positive number. □

3 High score asymptotics in the Poisson model

The next result renders the asymptotics of the expected number of paths starting per unit volume and with score exceeding b. The Theorem covers the cases that the number of steps in the path is kept small. More precisely this means that the gap open penalty Δ is assumed to grow with b, in a way such that the proportion of paths with a high number j of steps is small among all paths with a score higher than b. Similar as in Siegmund and Yakir (2000) we require that

$$\beta := \lim_{b\to\infty} b^{\gamma} e^{-\lambda\Delta} \text{ exists and is strictly positive.} \tag{6}$$

The implication of (6) for j will become clearer in section 4. Note that Δ remains constant in the case $\gamma = 0$.

Theorem *Assume 6.*
a) *For fixed* $j \in \{0, 1, \ldots\}$ *and* $\gamma \geq 0$,

$$M((b,\infty), j) \sim k e^{-\lambda b} \left(\omega k \beta b^{1-\gamma}/\lambda\right)^j / j! \quad \textit{as } b \to \infty \tag{7}$$

b) *For* $\gamma = 1$ *there holds*

$$M((b,\infty)) \sim k\, e^{-\lambda b} \exp\left(\omega k \beta/\lambda\right) \quad \textit{as } b \to \infty. \tag{8}$$

Proof
a) Because of the Corollary it suffices to show that

$$\lim_{b\to\infty} \lambda e^{\lambda b} \left(b^{\gamma-1}\right)^j \int_b^\infty e^{-\lambda s} (s\, e^{-\lambda\Delta})^j \, ds = \beta^j. \tag{9}$$

Indeed because of (6) we have

$$\begin{aligned}\lim_{b\to\infty} \lambda e^{\lambda b} \left(b^{\gamma-1}/\beta\right)^j \int_b^\infty e^{-\lambda s} (s\, e^{-\lambda\Delta})^j \, ds &= \lim_{b\to\infty} \lambda \int_b^\infty e^{-\lambda(s-b)} (s/b)^j \, ds \\ &= \lim_{b\to\infty} \lambda \int_0^\infty e^{-\lambda u} (1 + u/b)^j \, du = 1.\end{aligned}$$

b) Because of a) and dominated convergence it suffices to show the existence of a summable sequence (c_j) such that for suitably large Δ_0 and j_0,

$$e^{\lambda b} M((b,\infty), j) \leq c_j, \quad \Delta \geq \Delta_0,\ j \geq j_0. \tag{10}$$

Combining the Proposition with the Lemma below we have for all $j \geq j_0$

$$\begin{aligned}e^{\lambda b} M((b,\infty), j) \ \leq\ & 2k\lambda \left(\frac{2\omega k}{\lambda}\right)^j \left[\frac{1}{j!} \int_b^\infty e^{-\lambda(s-b)} (s\, e^{-\lambda\Delta})^j \, ds \right.\\ & \left. + \left(e^{1-\lambda\Delta}(\Delta + 3/\lambda)\right)^j \int_b^\infty e^{-\lambda(s-b)} ds\right].\end{aligned}$$

Writing $C := 2\omega k/\lambda$, we note that, for all sufficiently large Δ,

$$e^{1-\lambda\Delta}(\Delta + 3/\lambda) \leq 1/(2C) \tag{11}$$

and

$$b(\Delta) e^{-\lambda\Delta} \leq 2\beta. \tag{12}$$

Thus we have for all $j \geq j_0$, all sufficiently large Δ and $b = b(\Delta)$ satisfying (6)

$$e^{\lambda b} M((b,\infty), j) \leq 2k\,(2\beta C)^j \frac{1}{j!} \int_b^\infty \lambda e^{-\lambda(s-b)} (s/b)^j ds + 2k\lambda 2^{-j} \lambda^{-1}. \qquad (13)$$

The integral in (13) equals $\int_0^\infty e^{-u}(1+u/(\lambda b))^j du$ and hence is bounded by $2^j(1 + (\lambda b)^{-j} j!)$. Thus it suffices to choose Δ so large that besides (11) and (12) also $b(\Delta) > 5\beta C/\lambda$ is valid. □

To complete the proof of the Theorem we need the following Lemma.

Lemma *For sufficiently large j_0 and all $j \geq j_0$,*

$$\begin{aligned} \frac{1}{(2j-1)!} \int_0^\infty l^{2j-1} (s + j\Delta + l/\lambda)^j e^{-l} dl &\leq 2(s + j(\Delta + 3/\lambda))^j \\ &\leq 2^{j+1}(s^j + j!\, e^j (\Delta + 3/\lambda)^j). \end{aligned}$$

Proof We set $a := s + j\Delta$ and split the range of integration at $3j$. First we observe that

$$\int_0^{3j} l^{2j-1}(a + l/\lambda)^j e^{-l} \mathrm{d}l \leq \int_0^\infty l^{2j-1}(a + 3j/\lambda)^j e^{-l} \mathrm{d}l = (a + 3j/\lambda)^j \cdot (2j-1)! \qquad (14)$$

Now we turn to $l > 3j =: l_0$. Using the convexity of log we obtain:

$$\begin{aligned} l^{2j-1}(a + l/\lambda)^j e^{-l} \quad \leq \quad & \exp\Big((2j-1)\log l_0 + (l - l_0)\frac{2j-1}{l_0} + j \log(a + l_0/\lambda) \\ & + (l - l_0)\frac{j}{a\lambda + l_0} - l\Big) \end{aligned}$$

This implies:

$$\begin{aligned} & \int_{3j}^\infty l^{2j-1}(a + l/\lambda)^j e^{-l} dl \\ & \leq \quad l_0^{2j-1}\,(a + l_0/\lambda))^j\, e^{-l_0} \int_0^\infty \exp\left(l\left(\frac{2j-1}{l_0} + \frac{j}{a\lambda + l_0} - 1\right)\right) \mathrm{d}l \\ & = \quad \frac{(3j)^{2j-1}(s + j\Delta + 3j/\lambda)^j e^{-3j}}{1 - (2j-1)/(3j) - j/(\lambda b + j(\lambda\Delta + 3))}. \end{aligned}$$

The denominator is bounded away from zero uniformly in j. From Stirling's formula it follows that for sufficiently large j the numerator is smaller than

$$(2j-1)! \cdot (s + j\Delta + 3j/\lambda)^j \cdot \frac{2}{3}\left(\frac{4e}{9}\right)^{-j}$$

Combining this with (14) we immediately arrive at the first inequality of the Lemma for j large enough. The second inequality follows from Stirling's formula and the simple inequality $(n+m)^j < (2n)^j + (2m)^j$. □

4 What does a high score path look like?

What is the distribution of the number of steps, the distance penalties and the point qualities in high score paths?

The number of steps: A first observation is that in the case of $\gamma = 1$ considered in part (c) of the above theorem, it follows from part (b) and (c) of the theorem that the number of steps in a path of score $> b$ is asymptotically Poisson-distributed with parameter $\omega k\beta/\lambda$ as b tends to ∞. In the case $\gamma < 1$ the number of steps in a path of score $> b$ grows to ∞ for $b \to \infty$, and in the case $\gamma > 1$ the distribution of j will be asymptotically concentrated on 0 as $b \to \infty$.

The distance penalties: For fixed j, the density of the joint distribution of the distance penalties $(l_1, \cdots, l_j)$ in the j-step paths of score s is

$$\frac{k\lambda^{j+1}e^{-\lambda(s+j\Delta+\sum l_i)}(s + j\Delta + \sum l_i)^j(\omega k)^j l_1 \cdots l_j \, dl_1 \cdots dl_j \, ds}{M(ds, j)}$$
$$= \frac{\lambda^{2j}e^{-\lambda\sum l_i}\left(1 + \left(j\Delta + \sum l_i\right)/s\right)^j l_1 \cdots l_j \, dl_1 \cdots dl_j}{\frac{1}{(2j-1)!}\int_0^\infty l^{2j-1}\left(1 + \left(j\Delta + l/\lambda\right)/s\right)^j e^{-l}dl} \tag{15}$$

If we let s tend to infinity, the fraction (15) converges to

$$\lambda^{2j} \cdot \prod_{i=1}^{j} e^{-\lambda l_i} l_i \, dl_i$$

uniformly on $s \geq \eta e^{\lambda\Delta}$, where η is a fixed positive number. In other words, the distance penalties the high-score j-step paths are asymptotically distributed like independent Gamma(2, λ)-random variables.

The point qualities: The joint distribution of the point qualities in the j-step paths of score s and with distance penalties $l_1, \ldots, l_j$ obviously equals the joint distribution of the fragment lengths that one obtains by breaking the intervall $[0, s + j\Delta + \sum l_i]$ in j independently uniformly chosen points.

5 Simulation studies and comparison with known asymptotics

We will compare our results from section 3 (in particular formula (8)) with the asymptotics recently obtained by Siegmund and Yakir (2000) for the distribution of high scores in the local alignment of long independent sequences. This comparison will, for given scoring parameters, fix the numerical value of the parameter ω which appears in our Section 3. We will then check by simulation whether this value meets the interpretation of ω from which we started in (1).

Siegmund and Yakir consider a gap penalty of the form $\delta u + \Delta j$, where u is the number of unaligned positions and j is the number of gaps in the alignment. For transparency we will restrict to the case of a match/mismatch score taking two values only. Theorem 3 in Siegmund and Yakir (2000) then implies that the probability of the maximal score exceeding a level b converges, in the limit of long sequence lengths $m,\ n$, to $1 - e^{-Q}$, where

$$Q = k\alpha \exp(c\lambda\beta), \tag{16}$$

provided that

$$m\,n\,e^{-\lambda b} \to \alpha$$

and

$$b\,e^{-\lambda\Delta} \to \beta.$$

Here, the constant λ is the same as the one obtained for the ungapped case by Dembo et al. (1994), and c is a function of the scoring parameters μ and δ. (The proof of Theorem 3 of Siegmund and Yakir (2000), which is slightly incomplete, has recently been amended by the authors in a correction note.) Siegmund and Yakir concentrate on nonlattice match/mismatch scoring schemes, for which the bounds k_1, k_2 mentioned in Section 1 coincide (and equal k in (16)). A comparison of (16) and (8) then suggests to set $\omega = \lambda^2 c/k$ in order to make our Poisson model fit to the gapped alignment picture. In the lattice case we assume that Siegmund and Yakir's asymptotics still work if the lattice width is not too large, and thus obtain two pairs (k_1, ω_1) and (k_2, ω_2). We conjecture that clumping effects of high scored paths are small enough to allow a Poisson approximation for their number. Then, we obtain from (8) that the probability for a score exceeding b approximately equals

$$\Pr(\text{maximal score} \geq b) \approx 1 - \exp(-mnke^{-\lambda b}\exp(\omega k b e^{-\lambda\Delta}/\lambda)). \tag{17}$$

We have performed an empirical test concerning this fit. We generated two independent sequences of length 1000, both uniform over a four-letter alphabet, and used a scoring scheme with match reward equal to 1, mismatch penalty $\mu = 1.1$, gap open penalty $\Delta = 3.07$ and gap extension penalty $\delta = 0.52$. In this case we get $(k_1, \omega_1) = (0.218, 11.37)$, $(k_2, \omega_2) = (0.245, 10.13)$, and $\lambda = 1.15$. Figure 2 shows that (17) fits quite well in the case of these scoring parameters, even for the relatively small Δ of 3.07. For the dashed line we used the parameter values $(k_3, \omega_3) = (0.23, 8.6)$. k_3 is the rounded mean of k_1 and k_2, and ω_3 will be explained below in the context of Figure 3.

As promised let us now turn to our interpretation (1) of ω. We simulated a pair of i.i.d. sequences of length 1000 and considered all those gaps between two "good diagonal pieces" where the best alignment bridging this gap did not touch another "good diagonal piece". Out of these gaps we counted all those whose total gap costs were $\Delta + l'$, with $l' \in [l, l+h]$ for $l, h > 0$, denoting their number as $f(l) \cdot h$. (We made the special choice of $h = 1/6$ and $l = i/6$, $i = 1, 2, \ldots$.) As "good diagonal pieces" we took all LOULAs whose score exceeds the value 3. Let $\bar{k}$ be the observed mean number per unit area of these diagonal pieces in our simulation run. Note that $f(l)dl/\bar{k}$ corresponds to the expected area of $\{y\ :\ R(y) \in dl\}$ in the Poisson model. An equivalent assumption to (1) is that the expected value of this area equals $\omega l dl$. In Figure 3 we compare the linear functions $l \mapsto \omega_i l$ for $i \in \{1, 2, 3\}$ to a histogram displaying $l \mapsto f(l)/\bar{k}$. A substantial part of the variation seen in the bar chart stems from lattice effects inherent in the scoring scheme. The dashed line is a linear fit to the histogram, weighted with the asymptotic intensities (dotted line) of the distance penalty l in high scored paths (see Section 4). We define ω_3 as the slope of this line – remarkably, the fit in Figure 2 becomes even better with this choice of ω.

It will be interesting to further explore the role of the parameter ω. Also, other questions remain to be clarified, the most challenging being how more slowly growing (or even constant) gap open penalties Δ change the asymptotics. This will be explored in forthcoming work.

Acknowledgments

For helpfull dicussions we thank Amir Dembo, Benny Yakir, Jochen Geiger, Frank den Hollander, Matthias Birkner, Brooks Ferebee, Götz Kersting, and Lars Kauffmann. For financial support we thank the German Science Foundation (DFG).

References

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

Arratia, R., Waterman, M. S. (1991) A phase transition for the score in matching random sequences allowing deletions. *Ann. Appl. Prob.* **4**, 200–225.

Dembo, A., Karlin, S., Zeitouni, O. (1994), Limit distribution of maximal non-aligned two-sequence segmental score. *Ann. Probab.* **22**, 2022–2039.

Siegmund, D., Yakir, B. (2000), Approximate p-values for Local Sequence Alignments. *Annals of Statistics* **28**, 657–680.

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

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

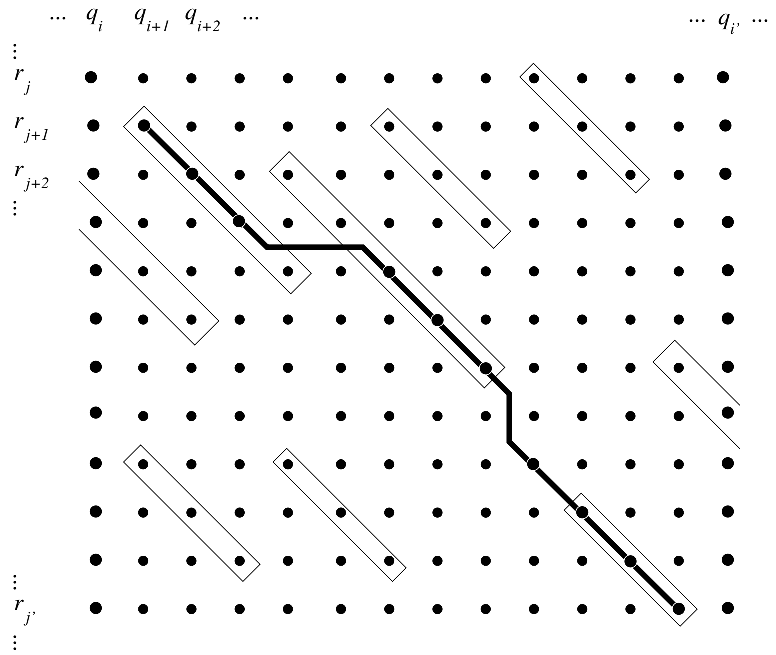

Figure 1: The best gapped local alignment (line) of the sequences $(q_1, \ldots, q_n)$ and $(r_1, \ldots, r_n)$ uses essential parts of locally optimal ungapped local alignments (LOULAs) (boxes).

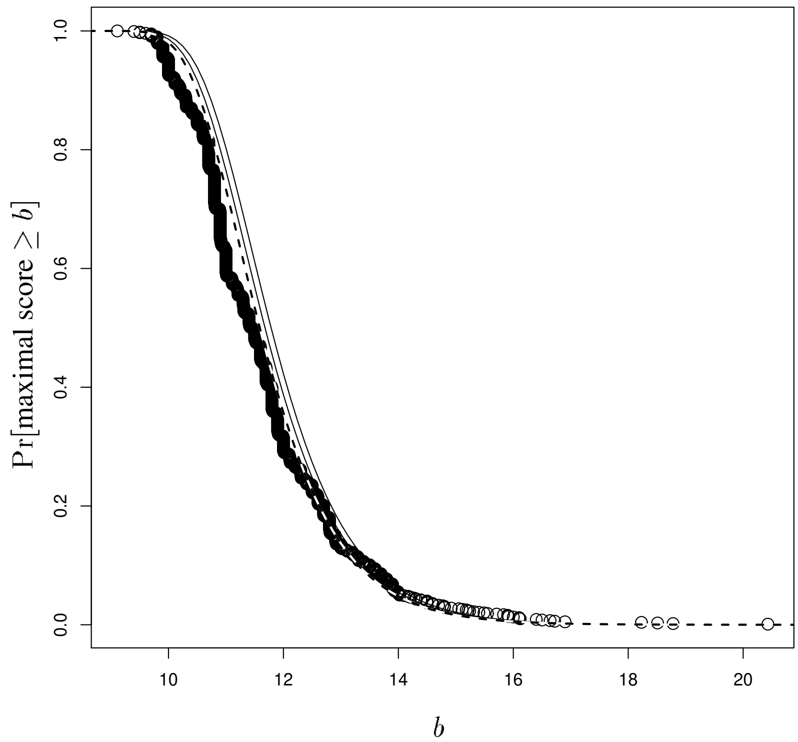

Figure 2: Comparison of the empirical density function of the optimal alignment scores from 1000 simulated pairs of unrelated sequences of length 1000 to the asymptotic result (17). The parameter values are $\mu = 1.1$, $\Delta = 3.07$, $\delta = 0.52$. The solid lines correspond to (k_1,ω_1)=(0.218,11.37) and (k_2,ω_2)=(0.245,10.13), the dashed line corresponds to (k_3,ω_3)=(0.23,8.6).

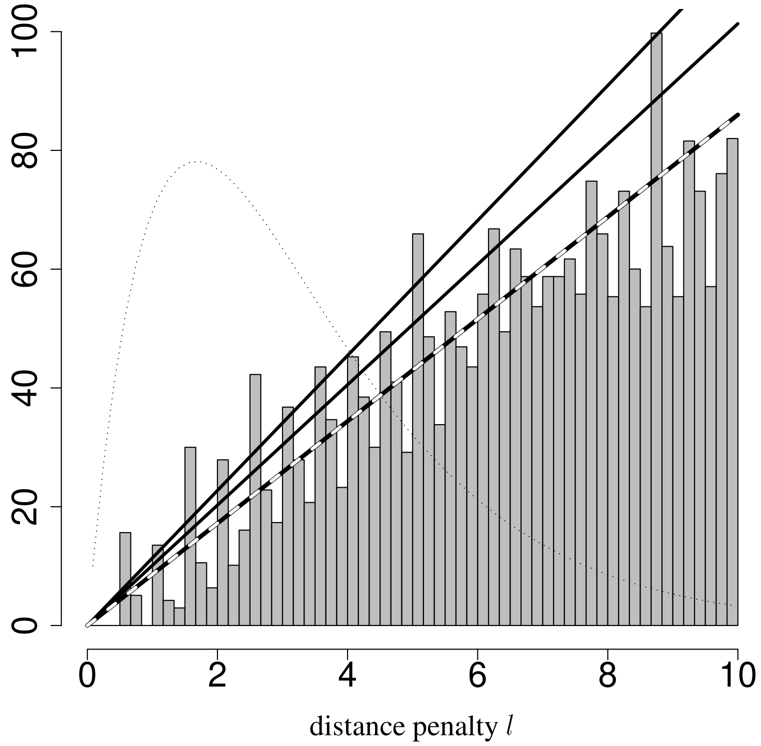

Figure 3: The histogram displays the intensity of LOULAs that could be reached from a given LOULA for total gap costs of $\Delta + l$, estimated from a simulated pair of unrelated sequences of length 1000. We only counted LOULAs with score higher than 3 and divided the mean number of such LOULAs reachable from a given one with total gap costs around $\Delta + l$ by the observed frequency of LOULAs with score > 3 per unit volume. The dotted curve is (a factor times) the density of the asymptotic Gamma distribution of the gap length distribution in high score alignments found in section 4. The dashed line $l \mapsto \omega_3 l$ was linearly fitted to the histogram (with weights given by the dotted line) and the solid lines show the predictions from ω_1 and ω_2.