Bernoulli **10**(5), 2004, 829–845

Large deviations for global maxima of independent superadditive processes with negative drift and an application to optimal sequence alignments

STEFFEN GROSSMANN[1] and BENJAMIN YAKIR[2]

[1]*Department of Computational Molecular Biology, Max-Planck-Institute for Molecular Genetics, Ihnestrasse 63–73, D-14195 Berlin, Germany. E-mail: steffen.grossmann@molgen.mpg.de*
[2]*Department of Statistics, Hebrew University, Jerusalem, Israel. E-mail: msby@mscc.huji.ac.il*

We examine the distribution of the global maximum of an independent superadditive process with negative drift. We show that, under certain conditions, the distribution's upper tail decays exponentially at a rate that can be characterized as the unique positive zero of some limiting logarithmic moment generating function. This result extends the corresponding one for random walks with a negative drift. We apply our results to sequence alignments with gaps. Calculating *p*-values of optimal gapped alignment scores is still one of the most challenging mathematical problems in bioinformatics. Our results provide a better understanding of the tail of the optimal score's distribution, especially at the level of large deviations, and they are in accord with common practice of statistical evaluation of optimal alignment results. However, a complete mathematical description of the optimal score's distribution remains far from reach.

Keywords: large deviations; sequence alignment; superadditive processes

1. Introduction

Consider a random walk $(S_n)_{n\geqslant 0}$ with negative drift but a positive probability of having positive steps. Much is known about the behaviour of the global maximum $G := \max_{n\geqslant 0} S_n$ of such a random walk – see such classic textbooks as Feller (1968; 1971) and Spitzer (1976). One basic result from the world of large deviations says that the upper tail of the distribution of G decays exponentially at a rate that can be characterized via the logarithmic moment generating function $\Lambda(\theta) := \log \mathbb{E}[\exp(\theta S_1)]$ of the random walk's increments: As long as Λ is not degenerate, it has a unique positive zero θ^* and

$$\lim_{t\to\infty} -\frac{1}{t}\log \mathbb{P}(G \geqslant t) = \theta^*. \tag{1}$$

We will extend this result from random walks to the class of *independent superadditive processes.* These processes play a crucial role in describing the statistics of the scores of optimal local sequence alignments with gaps – a model that nowadays is widely applied in bioinformatics.

Observe that the random walk result (1) has already been applied to *gapless* alignments. Gapless alignments have a certain additivity property, and therefore random walks are the right 'building blocks' for the model. For a complete description of the statistics of optimal ungapped alignments, see Dembo *et al.* (1994a; 1994b).

In *gapped* alignments, this additivity is no longer present, but is weakened to superadditivity combined with a certain independence structure. The basic building blocks therefore are now independent superadditive processes, and this is why we deal with them.

The rest of the paper is organized as follows. In the next section we state our results for independent superadditive processes and give the proofs. How the results apply to optimal gapped alignments will be shown in Section 3. We discuss the practical implications of our results in Section 4.

2. Global maxima of independent superadditive processes

Let $(T_{i,j})_{0\leqslant i\leqslant j}$ be an independent superadditive process, i.e. we assume the following:

(i) For all $i \leqslant j \leqslant k$, we have $T_{i,k} \geqslant T_{i,j} + T_{j,k}$.
(ii) $(T_{i,j})_{0\leqslant i\leqslant j}$ has the same joint distribution as $(T_{i+1,j+1})_{0\leqslant i\leqslant j}$.
(iii) $\mathbb{E}[T_{0,1}^-] < \infty$.
(iv) For any increasing sequence $0 = i_0 < i_1 < i_2 < \ldots < i_k$, the random variables $(T_{i_{j-1},i_j})_{1\leqslant j\leqslant k}$ are independent.

Of course, this is more than enough for a superadditive ergodic theorem to hold, i.e. we have (see, for example, Durrett 1996, Section 6.6)

$$\lim_{n\to\infty} \frac{T_{0,n}}{n} = \gamma \qquad \text{almost surely,} \tag{2}$$

where γ is the *growth constant*

$$\gamma := \lim_{n\to\infty} \frac{\mathbb{E}[T_{0,n}]}{n} = \sup_{n>0} \frac{\mathbb{E}[T_{0,n}]}{n} \in (-\infty, \infty]. \tag{3}$$

An important large-deviations result for independent superadditive processes was proved in Hammersley (1974):

Theorem 1. *For each n, define the logarithmic moment generating function Λ_n of $T_{0,n}$ on $\mathbb{R}_+$ by*

$$\Lambda_n(\lambda) := \log \mathbb{E}[\exp(\lambda T_{0,n})].$$

Then the limits

$$\Lambda(\lambda) = \lim_{n\to\infty} \frac{1}{n}\Lambda_n(\lambda) = \sup_{n>0} \frac{1}{n}\Lambda_n(\lambda) \tag{4}$$

and

$$r(q) = \lim_{n\to\infty}\left\{-\frac{1}{n}\log \mathbb{P}(T_{0,n} > qn)\right\} = \inf_{n>0}\left\{-\frac{1}{n}\log \mathbb{P}(T_{0,n} > qn)\right\} \tag{5}$$

exist in $\overline{\mathbb{R}} := [-\infty, \infty]$ *for* $\lambda \geqslant 0$ *and* $q \in \mathbb{R}$. *Moreover,* Λ *and* r *are convex functions and are related by*

$$r(q) = \sup_{\lambda\geqslant 0}\{\lambda q - \Lambda(\lambda)\} \tag{6}$$

and

$$\Lambda(\lambda) = \sup_{q\in\mathbb{R}}\{q\lambda - r(q)\} =: r^*(\lambda), \tag{7}$$

for all q *in the interior of* $\mathcal{D}_r := \{q' : r(q') < \infty\}$ *and* $\lambda \geqslant 0$ *in the interior of* $\mathcal{D}_{r^*} := \{\lambda' : r^*(\lambda') < \infty\}$.

Remark 1. In Hammersley (1974) the result was originally stated in the more general setting of superconvolutive families of distribution functions, for which independent superadditive processes provide natural examples. The statement of Theorem 1 is also given in Kingman (1975, Theorem 3.4, p. 214). Although Kingman's argument for proving (7) is only valid for $\mathcal{D}_\Lambda := \{\lambda' : \Lambda(\lambda') < \infty\}$, which might in general be strictly smaller than $\mathcal{D}_{r^*}$, the result does hold in the generality stated in Theorem 1. More details concerning the proofs of Hammersley and Kingman and how they relate to the more modern Gärtner–Ellis theorem can be found in Grossmann (2003).

Remark 2. A formulation equivalent to (6) and (7) is that $\mathrm{cl}(r)$ and $\overline{\Lambda}$ are a pair of convex conjugate functions, where $\mathrm{cl}(r)$ denotes the closure of the convex function r and

$$\overline{\Lambda}(\lambda) := \begin{cases} \mathrm{cl}(\Lambda(\lambda)), & \lambda \geqslant 0, \\ \infty, & \lambda < 0. \end{cases}$$

For more details on convex conjugacy, see Rockafellar (1979).

We now study the global maximum of an independent superadditive process. In the following we assume the process $(T_{i,j})_{0\leqslant i\leqslant j}$ to

(i) have a negative growth constant $\gamma < 0$;
(ii) have a positive probability of being positive, i.e.

$$\mathbb{P}(T_{0,1} > 0) > 0;$$

(iii) be linearly bounded, i.e. there is a constant c_u such that

$$T_{0,n} \leqslant c_u n.$$

We define the global maximum of the process $(T_{0,n})_{n\geqslant 0}$ as

$$G := \max_{n\geqslant 0} T_{0,n}. \tag{8}$$

Observe that (i), (iii) and (2) together imply that $G < \infty$ a.s.

Our main result in the general setting of independent superadditive processes is the following characterization of the large deviations of G.

Theorem 2. *Assume that the equation $\Lambda(\lambda) = 0$ has a unique positive solution θ^*. Then*

$$\lim_{t\to\infty} -\frac{1}{t}\log \mathbb{P}(G > t) = \theta^*.$$

Remark 3. It is not the case that the simple assumption $\gamma < 0$ implies the existence of a unique positive zero θ^* of Λ. In fact, it is only clear that $\Lambda'(0+) \geqslant \gamma$ (where $\Lambda'(0+)$ denotes the right derivative of the convex function Λ in 0). But observe that since we assume (ii) and (iii), the existence of a unique positive zero θ^* of Λ follows as soon as $\Lambda'(0+) < 0$.

Before proving the theorem we need a couple of lemmas.

Lemma 1. *Assume that $\Lambda(\lambda) = 0$ has a unique positive solution θ^*. Then*

$$\theta^* = \inf_{q>0} \frac{r(q)}{q} = \frac{r(q^*)}{q^*},$$

for any value q^* such that $r(q^*) < \infty$ and

$$\Lambda'(\theta^*-) \leqslant q^* \leqslant \Lambda'(\theta^*+).$$

Proof. By convexity and since $r(q) = 0$ for $q \leqslant \lambda < 0$, it is clear that r is non-negative and non-decreasing and that we can define

$$\bar\theta := \max\{\theta \geqslant 0 | \forall q > 0 : \theta q \leqslant r(q)\}, \tag{9}$$

which is the slope of the unique increasing straight line that goes through the origin and is tangential to r. With this definition of $\bar\theta$ we clearly have

$$\bar\theta = \inf_{q>0} \frac{r(q)}{q}.$$

Next, observe that it follows from (7) and the definition of θ^* that, for all $q > 0$, we have $\theta^* q \leqslant r(q)$, from which $\bar\theta \geqslant \theta^* > 0$ follows. By the assumptions on Λ, this implies $\Lambda(\bar\theta) \geqslant \Lambda(\theta^*) = 0$. On the other hand, it follows from (7) and (9) that $\Lambda(\bar\theta) \leqslant 0$ and therefore $\bar\theta = \theta^*$. The characterization of θ^* via q^* follows directly from convex conjugation. □

Lemma 2. *Assume that $\Lambda(\lambda) = 0$ has a unique positive solution θ^*. For all n, all λ such that $\Lambda_n'(\lambda-) > 0$ and all q with*

$$\Lambda_n'(\lambda-) \leqslant qn \leqslant \Lambda_n'(\lambda+),$$

we have

$$\theta^* \leqslant \lambda - \frac{(1/n)\Lambda_n(\lambda)}{q}.$$

Proof. Clear from Figure 1. □

Lemma 3. *Assume that* $\Lambda(\lambda) = 0$ *has a unique positive solution* θ^*. *Then*

$$\lim_{t\to\infty} \min_{n\geqslant 0} -\frac{1}{t} \log \mathbb{P}(T_{0,n} > t) = \theta^*.$$

Proof. We begin by showing that $\mathbb{P}_0(T_{0,n} > t) \geqslant \mathrm{e}^{-t\theta^*}$. To this end, for given t and n, define $q_{t,n} = t/n$ and $\lambda_{t,n}$ as the unique value with

$$\frac{1}{n}\Lambda_n'(\lambda_{t,n}-) \leqslant q_{t,n} \leqslant \frac{1}{n}\Lambda_n'(\lambda_{t,n}+). \tag{10}$$

(Of course, if we denote $\sigma_h := \sup_{\lambda\geqslant 0}\Lambda'(\lambda+)$ and chose t and n such that $q_{t,n} > \sigma_h$, (10) cannot be fulfilled, since we have, for all $\lambda \geqslant 0$,

$$\frac{1}{n}\Lambda_n'(\lambda+) \leqslant \sigma_h.$$

But then $0 = \mathbb{P}(T_{0,n} > t) \leqslant \mathrm{e}^{-t\theta^*}$ is a trivial conclusion.)

The basic calculation is

$$\begin{aligned}\mathbb{P}_0(T_{0,n} > t) &= \mathbb{E}_0[1;\ T_{0,n} > t] \\ &= \exp\left\{-t\left[\lambda_{t,n} - \frac{1}{n}\Lambda_n(\lambda_{t,n})/q_{t,n}\right]\right\}\mathbb{E}_{\lambda_{t,n}n}[\exp(-\lambda_{t,n}(T_{0,n} - t));\ T_{0,n} > t],\end{aligned} \tag{11}$$

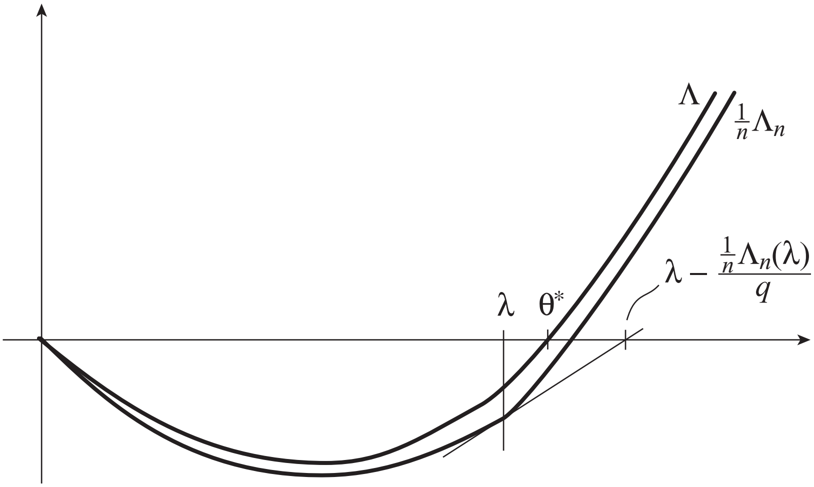

Figure 1. Proof of Lemma 2

where the measure $\mathbb{P}_{\lambda,n}$ is defined via

$$\frac{\mathrm{d}\mathbb{P}_{\lambda,n}}{\mathrm{d}\mathbb{P}}(\omega) := \exp(\lambda T_{0,n}(\omega) - \Lambda_n(\lambda)).$$

Since the expectation in (11) is bounded from above by one, it follows for arbitrary n from Lemma 2 that

$$\mathbb{P}_0(T_{0,n} > t) \leqslant \exp\left\{-t\left[\lambda_{t,n} - \frac{1}{n}\Lambda_n(\lambda_{t,n})/q_{t,n}\right]\right\} \leqslant \mathrm{e}^{-t\theta^*}. \qquad (12)$$

Maximizing over n, taking logarithms and dividing by $-1/t$ gives

$$\lim_{t\to\infty} \min_{n\geqslant 0} -\frac{1}{t}\log \mathbb{P}(T_{0,n} > t) \geqslant \theta^*.$$

For the reverse inequality choose q^* such that

$$\Lambda'(\theta^* -) \leqslant q^* \leqslant \Lambda'(\theta^* +)$$

and define $n_t^* := \lceil t/q^* \rceil$. Then it is clear that

$$\mathbb{P}(T_{0,n_t^*} > t) \geqslant \mathbb{P}(T_{0,n_t^*} > n_t^* q^*)$$

and

$$\begin{aligned}\lim_{t\to\infty} \min_{n\geqslant 0} -\frac{1}{t}\log \mathbb{P}(T_{0,n} > t) &\leqslant \lim_{t\to\infty} -\frac{1}{t}\log \mathbb{P}(T_{0,n_t^*} > t)\\ &\leqslant \frac{1}{q^*}\lim_{n_t^*\to\infty}\left\{\frac{n_t^* q^*}{t}\cdot\left[-\frac{1}{n_t^*}\log \mathbb{P}(T_{0,n_t^*} > n_t^* q^*)\right]\right\}\\ &= \frac{r(q^*)}{q^*} = \theta^*.\end{aligned}$$

□

Observe that from (12) we obtain the following:

Corollary 1.

$$\max_{n\geqslant 0} \mathbb{P}(T_{0,n} > t) \leqslant \mathrm{e}^{-t\theta^*}.$$

Proof of Theorem 2. The proof is based on the fact that

$$\max_{n\geqslant 0} \mathbb{P}(T_{0,n} > t) \leqslant \mathbb{P}\left(\max_{n\geqslant 0} T_{0,n} > t\right) \leqslant \sum_{n=0}^{\infty} \mathbb{P}(T_{0,n} > t). \qquad (13)$$

The first inequality directly gives that

$$\lim_{t\to\infty} -\frac{1}{t}\log\mathbb{P}\left(\max_{n\geqslant 0} T_{0,n} > t\right) \leqslant \lim_{t\to\infty}\min_{n\geqslant 0} -\frac{1}{t}\log\mathbb{P}(T_{0,n} > t) = \theta^*,$$

whereas the second gives

$$\lim_{t\to\infty} -\frac{1}{t}\log\mathbb{P}\left(\max_{n\geqslant 0} T_{0,n} > t\right) \geqslant \lim_{t\to\infty} -\frac{1}{t}\log\sum_{n=0}^{\infty}\mathbb{P}(T_{0,n} > t).$$

So all that it remains to show is that

$$\lim_{t\to\infty} -\frac{1}{t}\log\sum_{n=0}^{\infty}\mathbb{P}(T_{0,n} > t) \geqslant \theta^*.$$

Choose the value q^* such that

$$\Lambda'(\theta^*-) \leqslant q^* \leqslant \Lambda'(\theta^*+).$$

Observe that since we assume that θ^* is the unique positive solution of $\Lambda(\lambda) = 0$, there must be values $\bar{\theta} < \theta^*$ and $\bar{q} < q^*$ with

$$\Lambda(\bar{\theta}) < 0 \quad \text{and} \quad \Lambda'(\bar{\theta}-) \leqslant \bar{q} \leqslant \Lambda'(\bar{\theta}+),$$

since otherwise we would have $\Lambda(\lambda) = 0$ for all $\lambda \in [0, \theta^*]$. Using $\bar{q}$, we define $\bar{n}_t := \lfloor t/\bar{q} \rfloor$.

We now split the sum in (13) into two parts:

$$\sum_{n=0}^{\infty}\mathbb{P}(T_{0,n} > t) = \sum_{n=0}^{\bar{n}_t}\mathbb{P}(T_{0,n} > t) + \sum_{n=\bar{n}_t+1}^{\infty}\mathbb{P}(T_{0,n} > t). \tag{14}$$

For the first sum on the right we have

$$\sum_{n=0}^{\bar{n}_t}\mathbb{P}(T_{0,n} > t) \leqslant (\bar{n}_t + 1)\max_{n\geqslant 0}\mathbb{P}(T_{0,n} > t),$$

so taking logarithms and dividing by $-t$ gives

$$-\frac{1}{t}\log\sum_{n=0}^{\bar{n}_t}\mathbb{P}(T_{0,n} > t) \geqslant -\frac{\log(t/\bar{q} + 2) = \max_{n\geqslant 0}\log\mathbb{P}(T_{0,n} > t)}{t} \underset{t\to\infty}{\longrightarrow} \theta^*.$$

Turning to the second sum on the right of (14), we define $q_{t,n} := t/n$ and $\lambda_{t,n}$ as in (10). Recall the basic equality (11). For convenience we set

$$\mu_{t,n} := \lambda_{t,n} - \frac{(1/n)\Lambda_n(\lambda_{t,n})}{q_{t,n}},$$

so that we can write

$$\mathbb{P}(T_{0,n} > t) \leqslant \mathrm{e}^{-t\mu_{t,n}}. \tag{15}$$

Furthermore, it is clear that for the above chosen $\bar{\theta}$ and $\bar{q}$ we have

$$\bar{\mu} := \bar{\theta} - \frac{\Lambda(\bar{\theta})}{\bar{q}} \geqslant \theta^*.$$

The definition of $\bar{n}_t$ gives that

$$q_{t,\bar{n}_t} = \frac{t}{\bar{n}_t} = \frac{t}{\lfloor t/\bar{q} \rfloor} \geqslant \bar{q},$$

which implies that

$$\frac{1}{\bar{q}} \geqslant \frac{\bar{n}_t}{t}. \tag{16}$$

Also we have for $n > \bar{n}_t$ that $q_{t,n} \leqslant \bar{q}$, and it follows from convexity considerations which are clarified in Figure 2 that for such n we have

$$\mu_{t,n} = \lambda_{t,n} - \frac{(1/n)\Lambda_n(\lambda_{t,n})}{q_{t,n}} \geqslant \bar{\theta} - \frac{\Lambda(\bar{\theta})}{q_{t,n}} =: \bar{\mu}_{t,n}. \tag{17}$$

Expressions (16) and (17) together imply that, for $n > \bar{n}_t$,

$$\mu_{t,n} - \bar{\mu} \geqslant \frac{\Lambda(\bar{\theta})}{t}(\bar{n}_t - n)$$

or

$$-t(\mu_{t,n} - \bar{\mu}) \leqslant \Lambda(\bar{\theta})(n - \bar{n}_t). \tag{18}$$

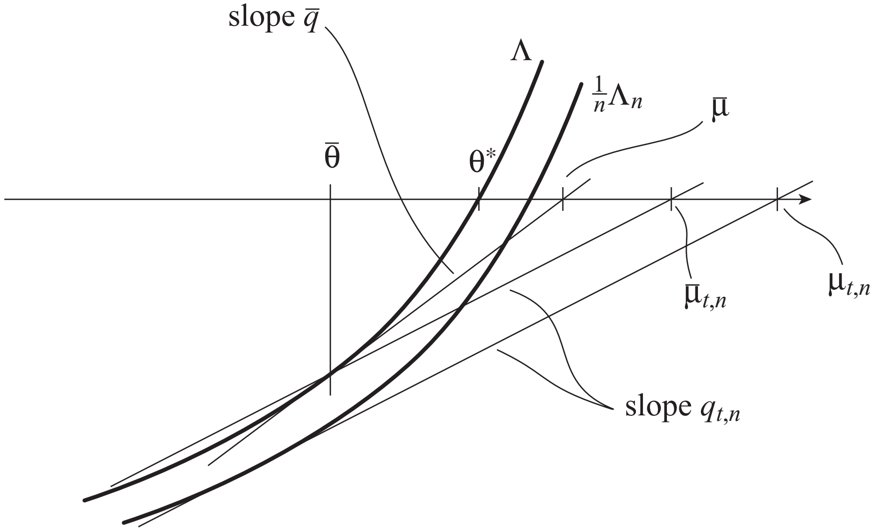

Figure 2. Illustrating (17)

Finally, this gives

$$\begin{aligned}\sum_{n=\bar{n}_t+1}^{\infty} \mathbb{P}(T_{0,n} > t) &\leqslant \mathrm{e}^{-t\bar{\mu}} \sum_{n=\bar{n}_t+1}^{\infty} \mathrm{e}^{-t(\mu_{t,n}-\bar{\mu})} \\ &\leqslant \mathrm{e}^{t\bar{\mu}} \underbrace{\sum_{n=\bar{n}_t+1}^{\infty} \mathrm{e}^{\Lambda(\bar{\theta})(n-\bar{n}_t)}}_{=:c_5} \\ &\leqslant c_5 \mathrm{e}^{-t\theta^*}.\end{aligned}$$

□

Again from the proof it is obvious that we have the following upper bound.

Corollary 2. *There exists some constant $c > 0$ such that for t large enough we have*

$$\mathbb{P}(G > t) \leqslant ct\mathrm{e}^{-t\theta^*}.$$

3. Optimal gapped alignments

Consider two sequences $\mathbf{X} = (X_1, X_2, \ldots)$ and $\mathbf{Y} = (Y_1, Y_2 \ldots)$ of independent and identically distributed letters drawn from some finite alphabet $\mathcal{A}$ according to some distribution μ. Take two substrings $X_{i_1+1}, \ldots, X_{j_1}$ and $Y_{i_2+1}, \ldots, Y_{j_2}$ of the sequences, where $i_1 \leqslant j_1$ and $i_2 \leqslant j_2$. An *alignment* of these two substrings is simply a way of writing them one above the other such that some letter pairs are aligned vertically. Also we are allowed to insert *gaps* at appropriate points in the sequences in order to deal with those letters that are not aligned to others.

Our aim is to define the *optimal global scores* $S(X_{i_1+1}, \ldots, X_{j_1}; Y_{i_1+1}, \ldots, Y_{j_2})$ and the *optimal local alignment scores* $M(X_{i_1+1}, \ldots, X_{j_1}; Y_{i_2+1}, \ldots, Y_{j_2})$, which we will do by first defining a score for each possible alignment and then maximizing over certain sets of alignments.

To define the score of an alignment fix a *scoring matrix* $F : \mathcal{A} \times \mathcal{A} \to \mathbb{R}$ which assigns a real value to each possible combination of two aligned letters. The usual scoring matrices have positive entries for combinations of equal letters and negative entries otherwise. For the rest of the paper we assume that the scoring matrix has positive entries at least for some letter combination that has a positive probability under $\mu \times \mu$, i.e. that

$$\mathbb{P}_{\mu\times\mu}(F(X, Y) > 0) > 0. \tag{H}$$

The score of the alignment is obtained by summing the score values over the aligned letter pairs and subtracting a penalty for the gaps in the alignment. One popular gap penalty scheme is the *affine-linear gap penalty*, where a sequence of k successive gaps is penalized by $g(k) = \Delta + \delta k$, with $\delta, \Delta \geqslant 0$.

In order to make this definition of alignment scores more transparent, it is convenient to have a special graphical representation of alignments. To this end we introduce the *lattice* $\mathbb{L}$ with *vertex set* $V_{\mathbb{L}} := \mathbb{N}_0 \times \mathbb{N}_0$. Denote vertices in $V_{\mathbb{L}}$ by bold lower-case letters ($\mathbf{i}, \mathbf{j}, \mathbf{k}, \ldots$) and use italic lower-case letters to refer to their coordinates ($\mathbf{i} = (i_1, i_2)$, $\mathbf{j} = (j_1, j_2), \ldots$). The *edge set* is then given by

$$E_{\mathbb{L}} := \{e = (\mathbf{i}, \mathbf{j}) \in V_{\mathbb{L}} \times V_{\mathbb{L}} : \mathbf{j} - \mathbf{i} \in \{(1, 0), (0, 1), (1, 1)\}\}.$$

Any alignment of the substrings $(X_{i_1+1}, \ldots, X_{j_1})$ and $(Y_{i_2+1}, \ldots, Y_{j_2})$ can now be represented by an *increasing path* z that starts at $\mathbf{i}$ and ends at $\mathbf{j}$. *Increasing* means that the path visits the vertices in increasing order, where we use the usual partial ordering on $\mathbb{N}_0 \times \mathbb{N}_0$.

The representation goes as follows. If for some $i_1 < k \leqslant j_1$ and $i_2 < l \leqslant j_2$ the path z follows the edge $((k-1, l-1), (k, l))$, this means that the letters X_k and Y_l are aligned. If z uses the horizontal edge $((k-1, l), (k, l))$ this means that X_k is aligned with a gap, and if it uses the vertical edge $((k, l-1), (k, l))$ it means that Y_l is aligned with a gap.

The score of an alignment can now easily be calculated using this path representation. We start with the case $\Delta = 0$. Assign values to each edge in $E_{\mathbb{L}}$. To all horizontal or vertical edges we simply assign the value $-\delta$. To the diagonal edge $((k-1, l-1), (k, l))$ we assign the (random) value $F(X_k, Y_l)$. The score of an alignment is then given by the sum of the edge values along the path. In the case $\Delta > 0$ we have to consider another small complication. Denote by ξ_z the number of horizontal or vertical stretches in the path z (each block of successive gaps in one of the sequences is a stretch). From the score value calculated so far, we then have to subtract $\xi_z\Delta$ to obtain the final score of the alignment. We denote this path score by S_z.

We can now define the optimal scores. For two vertices $\mathbf{j} \geqslant \mathbf{i}$ of the lattice, denote by $\mathcal{Z}_{\mathbf{i},\mathbf{j}}$ the set of all increasing paths that start at $\mathbf{i}$ and end at $\mathbf{j}$. The *optimal global score* can then be defined as

$$S_{\mathbf{i},\mathbf{j}} := S(X_{i_1+1}, \ldots, X_{j_1}; Y_{i_2+1}, \ldots, Y_{j_2}) := \max_{z \in \mathcal{Z}_{\mathbf{i},\mathbf{j}}} S_z,$$

whereas we define the optimal local score as

$$M_{\mathbf{i},\mathbf{j}} := M(X_{i_1+1}, \ldots, X_{j_1}; Y_{i_2+1}, \ldots, Y_{j_2}) := \max_{\mathbf{i} \leqslant \mathbf{k} \leqslant \mathbf{l} \leqslant \mathbf{j}} S_{\mathbf{k},\mathbf{l}}.$$

We also introduce the shorthand notation $S_{i,j} := S_{(i,i),(j,j)}$ and $M_{i,j} := M_{(i,i),(j,j)}$.

This definition of optimal *global* scores is obviously reminiscent of the definition of first-passage times in percolation theory. However, the concept of optimal *local* scores has no counterpart in classical first-passage percolation theory.

Many basic properties of $S_{0,n}$ and $M_{0,n}$ were discussed in Arratia and Waterman (1994) (see also Zhang 1995). However, Arratia and Waterman did not connect the fact that the process $(S_{i,j})_{0 \leqslant i \leqslant j}$ forms an independent superadditive process to Hammersley's result which we stated in Theorem 1. As we will see, only the combination of the results from Arratia and Waterman (1994) and Theorem 1 gives additional insight into the gapped alignment model. The rest of our paper is devoted to this special independent superadditive process, and all the objects (Λ, r, etc.) that we introduced in the context of general independent supperadditive processes are henceforth tacitly understood to be defined on the alignment model.

From the results in Arratia and Waterman (1994) it follows directly that $(S_{i,j})_{0\leqslant i\leqslant j}$ has some advantageous properties. Whereas it is in general only true that $\Lambda'(0+) \geqslant \gamma$ (cf. Remark 3), equality holds for the alignment model. The basis for this is Theorem 2 in Arratia and Waterman (1994), where, using an Azuma–Hoeffding–type concentration inequality, it is shown that

$$r(q) > 0, \qquad \text{for all } q > \gamma. \tag{19}$$

Suppose now that $q_0 := \Lambda'(0+) > \gamma$. Then one directly obtains from (6) that $r(q_0) = 0$ – a contradiction to (19). Therefore Λ has a unique positive zero θ^* as soon as $\gamma < 0$.

The central result in Arratia and Waterman (1994) is that, depending on the sign of γ, the mean behaviour of $M_{0,n}$ exhibits a phase transition in the parameter space. When $\gamma > 0$ the optimal *local* score $M_{0,n}$ behaves asymptotically as its global counterpart $S_{0,n}$, i.e. we have

$$\lim_{n\to\infty} \frac{M_{0,n}}{n} = \gamma \qquad \text{a.s.,}$$

whereas when $\gamma < 0$ there exists a constant b such that

$$\lim_{n\to\infty} \frac{M_{0,n}}{2\log n} = b \qquad \text{in probability.}$$

These two phases are usually referred to as the *global* and the *local phase*, repectively. It should be clear that local similarities can only be detected by using parameter values from the local phase, where long alignments typically get a large negative score. This all fits nicely together: when $\gamma < 0$ we are in the regime of practical interest and also our results from the previous section hold. The rest of our paper therefore is restricted to this local phase.

Arratia and Waterman (1994) characterized the logarithmic growth constant b using the function r which we have already encountered in Theorem 1. They show that

$$b := \sup_{q>0} \frac{q}{r(q)},$$

from which we directly obtain

$$\theta^* = b^{-1}$$

by our Lemma 1.

Our main result concerning optimal local alignments is the following characterization of the large deviations of M.

Theorem 3. *Let m, n and t tend to infinity in such a way that $g = o(\min(m, n))$ and $\log(mn) = o(t)$. Then*

$$\lim_{m,n,t\to\infty} -\frac{1}{t} \log \mathbb{P}(M_{(0,0),(m,n)} > t) = \theta^*.$$

The key to proving this theorem is the following two-dimensional generalization of Theorem 2.

Theorem 4. *Define*

$$G := \max_{\mathbf{j} \geqslant (0,0)} S_{(0,0),\mathbf{j}}. \tag{20}$$

Then

$$\lim_{t\to\infty} -\frac{1}{t} \log \mathbb{P}(G > t) = \theta^*.$$

Proof. The statement can essentially be proved by following the line of argument for Theorem 2. We will only point out how to handle the problems that arise from the bidimensionality.

To this end we first define, for all $\mathbf{i} \in V_{\mathbb{L}}$,

$$\Lambda_{\mathbf{i}}(\lambda) := \log \mathbb{E}[\mathrm{e}^{\lambda S_{(0,0),\mathbf{i}}}].$$

In this notation the defining equation (4) for Λ becomes

$$\Lambda(\lambda) = \lim_{n\to\infty} \frac{\Lambda_{(n,n)}(\lambda)}{\|(n,\, n)\|_1/2} = \sup_{n>0} \frac{\Lambda_{(n,n)}(\lambda)}{\|(n,\, n)\|_1/2}.$$

This can be extended to the statement that, for all $\mathbf{i} \in V_{\mathbb{L}}$, we have

$$\frac{\Lambda_{\mathbf{i}}(\lambda)}{\|\mathbf{i}\|_1/2} \leqslant \Lambda(\lambda). \tag{21}$$

Indeed, if we define $\bar{\mathbf{i}}$ as the vertex obtained by interchanging the coordinates of $\mathbf{i}$, then $\mathbf{i} + \bar{\mathbf{i}}$ is a vertex on the main diagonal with $\|\mathbf{i} + \bar{\mathbf{i}}\|_1 = 2\|\mathbf{i}\|_1$. Thus we have

$$\begin{aligned}\Lambda_{\|\mathbf{i}\|_1}(\lambda) &= \log \mathbb{E}[\mathrm{e}^{\lambda S_{(0,0),\mathbf{i}+\bar{\mathbf{i}}}}] \\ &\geqslant \log \mathbb{E}[\mathrm{e}^{\lambda (S_{(0,0),\mathbf{i}} + S_{\mathbf{i},\mathbf{i}+\bar{\mathbf{i}}})}] \\ &= 2 \log \mathbb{E}[\mathrm{e}^{\lambda S_{(0,0),\mathbf{i}}}] \\ &= 2\Lambda_{\mathbf{i}}(\lambda)\end{aligned}$$

by superadditivity and independence. This proves (21).

From this point on everything proceeds as in Section 2. The analogues to Lemmas 2 and 3 can be proved: in particular, the basic equation (11) holds in the adapted formulation, and we also have

$$\mathbb{P}(S_{(0,0),\mathbf{i}} > t) \leqslant \mathrm{e}^{-t\theta^*}$$

for all $\mathbf{i} \in V_{\mathbb{L}}$. The only thing that changes is the number of summands in the analogue of the first sum in (14), which is of the order $O(t^2)$. □

The statement corresponding to Corollary 2 reads as follows:

Corollary 3. *There exists some constant* $c > 0$ *such that, for* t *large enough, we have*

$$\mathbb{P}(G > t) \leqslant ct^2 \mathrm{e}^{-t\theta^*}.$$

Theorem 3 can now be proved.

Proof of Theorem 3. In analogy to (20) we define, for every $\mathbf{i} \in V_{\mathbb{L}}$,

$$G_{\mathbf{i}} := \max_{\mathbf{j} \geqslant \mathbf{i}} S_{\mathbf{i},\mathbf{j}}.$$

It is clear that

$$\mathbb{P}(M_{(0,0),(m,n)} > t) \leqslant \mathbb{P}\left(\max_{(0,0) \leqslant \mathbf{i} \leqslant (m,n)} G_{\mathbf{i}} > t \right) \leqslant mn \cdot \mathbb{P}(G > t). \tag{22}$$

Taking logs and dividing by $-t$ gives

$$\lim_{t \to \infty} -\frac{1}{t} \log \mathbb{P}(M_{(0,0),(m,n)} > t) \geqslant \theta^*$$

by Theorem 4 and since we assumed $\log(mn) = o(t)$.

For the reverse inequality we can simply repeat the second part of the proof of Lemma 3. Observe that in the notation introduced there we have

$$\mathbb{P}(T_{0,n_t^*} > t) \leqslant \mathbb{P}(M_{(0,0),(m,n)} > t),$$

since $t = o(\min(m, n))$. It follows that

$$\lim_{t \to \infty} -\frac{1}{t} \log \mathbb{P}(M_{(0,0),(m,n)} > t) \leqslant \lim_{t \to \infty} -\frac{1}{t} \log \mathbb{P}(T_{0,n_t^*} > t) = \theta^*.$$

□

It is obvious that from (22) we can give an upper bound for $\mathbb{P}(M_{(0,0),(m,n)} > t)$ which is in the same spirit as Corollary 3.

Corollary 4. *There exists some constant* $c > 0$*, such that, for* t *large enough, we have*

$$\mathbb{P}(M_{(0,0),(m,n)} > t) \leqslant cmnt^2 \mathrm{e}^{-t\theta^*}.$$

Convergence results implied by sub- or superadditivity arguments, as have frequently appeared in this paper, immediately raise the question of the corresponding rates of convergence. Whereas it is a common opinion (see, Steele 1997, p. 13) that in general one cannot say much about sub- or superadditive convergence rates, interesting results have meanwhile been given in a number of concrete cases (see, for example, Alexander 1993, 1994, 1997). One such case is the alignment model, where we show that:

Lemma 4. *There exists a constant c (depending on the scoring scheme) such that, for all n large enough, we have*

$$0 \leqslant n\Lambda(\theta) - \Lambda_n(\theta) \leqslant \theta c + \log(2n+1).$$

Proof. We start by defining

$$\overline{\Lambda}_n(\lambda) := \log \sum_{h=-n}^{n} \mathrm{e}^{\Lambda_{n-h,n+h}(\lambda)}.$$

It is clear that

$$0 \leqslant \overline{\Lambda}_n(\lambda) - \Lambda_n(\lambda) \leqslant \log(2n+1)$$

and therefore also that $(1/n)\overline{\Lambda}_n(\lambda)$ converges to $\Lambda(\lambda)$. In fact we show that, for some constant c,

$$n\Lambda(\lambda) - \overline{\Lambda}_n(\lambda) \leqslant \lambda c, \tag{23}$$

from which the statement follows.

To show (23) we first need some more notation. Define the secondary diagonals in distance n and $2n$ as

$$D_{l,n} := \{\mathbf{k} : \mathbf{k} \geqslant (0,0), \|\mathbf{k}\|_1 = 2ln\}, \qquad l = 1, 2.$$

The key calculation for a $\mathbf{j} \in D_{2,n}$ is

$$\begin{aligned} \mathrm{e}^{\Lambda_{\mathbf{j}}(\lambda)} &\leqslant \mathrm{e}^{\lambda c} \sum_{\substack{\mathbf{i} \in D_{1,n} \\ \mathbf{i} \leqslant \mathbf{j}}} \mathrm{e}^{\Lambda_{\mathbf{i}}(\lambda) + \Lambda_{\mathbf{j}-\mathbf{i}}(\lambda)} \mathbb{E}_{((0,0),\mathbf{i},\mathbf{j}),\lambda} \left[\frac{\max_{\mathbf{i}' \in D_{1,n}, \mathbf{i}' \leqslant \mathbf{j}} \mathrm{e}^{\lambda(S_{(0,0),\mathbf{i}'} + S_{\mathbf{i}',\mathbf{j}})}}{\sum_{\mathbf{i}'' \in D_{1,n}, \mathbf{i}'' \leqslant \mathbf{j}} \mathrm{e}^{\lambda(S_{(0,0),\mathbf{i}''} + S_{\mathbf{i}'',\mathbf{j}})}} \right] \\ &\leqslant \mathrm{e}^{\lambda c} \sum_{\substack{\mathbf{i} \in D_{1,n} \\ \mathbf{i} \leqslant \mathbf{j}}} \mathrm{e}^{\Lambda_{\mathbf{i}}(\lambda) + \Lambda_{\mathbf{j}-\mathbf{i}}(\lambda)}, \end{aligned}$$

where the first inequality comes from Lemma 5 and the second inequality comes from the fact that the expectation of the maximum over the sum is bounded from above by one. Summing over all $\mathbf{j} \in D_{2,n}$ gives, after a simple rearrangement of the summands,

$$\begin{aligned} \mathrm{e}^{\overline{\Lambda}_{2n}(\lambda)} &\leqslant \mathrm{e}^{\lambda c_2} \sum_{\mathbf{j} \in D_{2,n}} \sum_{\substack{\mathbf{i} \in D_{1,n} \\ \mathbf{i} \leqslant \mathbf{j}}} \mathrm{e}^{\Lambda_{\mathbf{i}}(\lambda) + \Lambda_{\mathbf{j}-\mathbf{i}}(\lambda)} \\ &= \mathrm{e}^{\lambda c_2} \sum_{\mathbf{i} \in D_{1,n}} \mathrm{e}^{\Lambda_{\mathbf{i}}(\lambda)} \sum_{\substack{\mathbf{j} \in D_{2,n} \\ \mathbf{j} \geqslant \mathbf{i}}} \mathrm{e}^{\Lambda_{\mathbf{j}-\mathbf{i}}(\lambda)} = \mathrm{e}^{\lambda c} \mathrm{e}^{2\overline{\Lambda}_n(\lambda)} \end{aligned}$$

from which it follows, by taking logarithms, that

$$\overline{\Lambda}_{2n}(\lambda) \leqslant \lambda c + 2\overline{\Lambda}_n(\lambda).$$

Applying Lemma 6 to this inequality gives (23). □

We conclude this section with the two lemmas cited in the above proof. The first is a result on the splitting of optimal paths.

Lemma 5. *Let* $\mathbf{j} \geqslant \mathbf{i}$. *Fix some integer* d *with* $0 \leqslant d \leqslant \|\mathbf{j}-\mathbf{i}\|_1$. *Then there is a constant* c *(depending on the scoring scheme) such that*

$$S_{\mathbf{i},\mathbf{j}} - c \leqslant \max_{\substack{\mathbf{k}:\mathbf{i}\leqslant\mathbf{k}\leqslant\mathbf{j}\\ \|\mathbf{k}-\mathbf{i}\|_1=d}} \{S_{\mathbf{i},\mathbf{k}} + S_{\mathbf{k},\mathbf{j}}\} \leqslant S_{\mathbf{i},\mathbf{j}}.$$

Proof. The second inequality is clear by superadditivity. The first inequality comes from the fact that the optimizing path for $S_{\mathbf{i},\mathbf{j}}$ has to hit or pass close by one of the vertices from the set $\{\mathbf{k} : \mathbf{i} \leqslant \mathbf{k} \leqslant \mathbf{j},\ \|\mathbf{k}-\mathbf{i}\|_1 = d\}$. An optimizing path for $S_{\mathbf{i},\mathbf{j}}$ that does not hit one of those vertices can easily be modified to a path that does so and still gets a score that differs from $S_{\mathbf{i},\mathbf{j}}$ by a scoring-scheme-dependent amount of at most c. □

The second lemma derives a bound for the difference between the elements of a convergent series and their limit from a bound for the difference between the elements. The proof is by elementary algebra and will be omitted.

Lemma 6. *Let* $(a_n)_{n\geqslant 0}$ *be a sequence for which* $\alpha := \lim_{n\to\infty} a_n/n$ *exists and for which*

$$2a_n - l(2n) \leqslant a_{2n} \leqslant 2a_n + u(2n),$$

for all n *and for some non-negative functions* $l(\cdot)$ *and* $u(\cdot)$. *Then*

$$-\sum_{i=1}^{\infty} 2^{-i} l(2^i n) \leqslant n\alpha - a_n \leqslant \sum_{i=1}^{\infty} 2^{-i} u(2^i n).$$

4. Discussion

The calculation of p-values of optimal gapped sequence alignments has been one of the major statistical problems motivated by bioinformatics. Practitioners are comfortable with the conjecture that optimal gapped alignments behave qualitatively in the same way as the optimal ungapped alignments which were thoroughly analysed in Dembo *et al.* (1994a; 1994b). At least since the paper of Waterman and Vingron (1994) it has therefore been assumed that

$$\mathbb{P}(M_{(0,0),(m,n)} \geqslant t) \sim Kmn \exp(-\theta t), \tag{24}$$

for large m, n and t, where the two parameters K and θ depend on the chosen scoring scheme. Various methods, from naive simulations to the formulation of complex functional relationships (as in Mott 2002), have been proposed to describe this dependency.

Exact formulae for K and θ were given in Siegmund and Yakir (2000) in the related case where gaps are allowed, but the gap-open penalty Δ is required to grow logarithmically in t. This growth has the effect that the gaps which appear in optimal alignments happen to follow an essentially Poisson-distributed number of gap intervals as $t \to \infty$. This is in sharp contrast to the case treated here, where the number of gap intervals cannot be expected to be bounded for growing t. Formally, this difference is reflected in the fact that in the model from Siegmund and Yakir (2000) θ is given by the corresponding gapless rate (and therefore does not depend on the gap penalties), whereas in our full model θ differs substantially from the gapless rate.

Returning to more practical matters, we wish to point out that our results give a theoretical justification for one special method to calculate θ that was proposed in Bundschuh (2002). It is clear that our Theorem 3 directly implies that the unknown parameter θ in (24) is equal to θ^*, i.e. the unique positive zero of Λ. Bundschuh conjectured this and showed how to estimate θ^* in practice. Of course, since a limiting procedure is involved, Λ cannot be calculated directly, and neither can θ^*. But if we define θ_n^* as the unique positive solution of the equation

$$\Lambda_n(\theta) = 0,$$

Lemma 4 gives that $\theta_n^* = \theta^* + O(\log n/n)$. Bundschuh showed that this recipe for estimating θ^* works well in so far as it suffices to first estimate θ_n^* for a small set of different, only moderately large values of n, and then to extrapolate to obtain an estimate for θ^*. Of course, our results say nothing about the second parameter K, but is has been pointed out by practitioners (see Mott 2002) that p-values depend more crucially on the parameter θ.

From a mathematical point of view, conjecture (24) seems a long way from being proved rigorously. Our results are a first step towards a more precise understanding of the model. We give a rigorous characterization of the leading rate of the exponential decay of $\mathbb{P}(M_{(0,0),(m,n)} \geqslant t)$ that extends the gapless case in a consistent way. However, a finer description of the asymptotic behaviour of the p-values remains an open and apparently hard problem.

Acknowledgements

We thank the Erwin Schrödinger Institute for financial support during a stay in Vienna in January 2003. B.Y. also thanks the United States-Israel Binational Science Foundation for support that partially resulted in this work.

References

Alexander, K.S. (1993) A note on some rates of convergence in first-passage percolation. *Ann. Appl. Probab.*, **3**(1), 81–90.

Alexander, K.S. (1994) The rate of convergence of the mean length of the longest common subsequence. *Ann. Appl. Probab.*, **4**(4), 1074–1082.

Alexander, K.S. (1997) Approximation of subadditive functions and convergence rates in limiting-shape results. *Ann. Probab.*, **25**(1), 30–55.

Arratia, R. and Waterman, M.S. (1994) A phase transition for the score in matching random sequences allowing deletions. *Ann. Appl. Probab.*, **4**(1), 200–225.

Bundschuh, R. (2002) Rapid significance estimation in local sequence alignment with gaps. *J. Comput. Biol.*, **9**, 243–260.

Dembo, A., Karlin, S. and Zeitouni, O. (1994a) Critical phenomena for sequence matching with scoring. *Ann. Probab.*, **22**(4), 1993–2021.

Dembo, A., Karlin, S. and Zeitouni, O. (1994b) Limit distributions of maximal non-aligned two-sequence segmental score. *Ann. Probab.*, **22**(4), 2022–2039.

Durrett, R. (1996) *Probability: Theory and Example*, 2nd edn. Belmont, CA: Duxbury Press.

Feller, W. (1968) *An Introduction to Probability Theory and Its Applications*, Vol. I, 3rd edn. New York: Wiley.

Feller, W. (1971) *An Introduction to Probability Theory and Its Applications*, Vol. II, 2nd edn. New York: Wiley.

Grossmann, S. (2003) Statistics of optimal sequence alignments. Doctoral thesis, Johann Wolfgang Goethe University, Frankfurt. Available at http://publikationen.stub.uni-frankfurt.de/volltexte/2003/292/.

Hammersley, J.M. (1974) Postulates for subadditive processes. *Ann. Probab.*, **2**(4), 652–680.

Kingman, J.F.C. (1975) Subadditive processes. In P.L. Hennequin (ed.), *École d'Été de Probabilités de Saint-Flour V*, Lecture Notes in Math. 539, pp. 167–223. Berlin: Springer-Verlag.

Mott, R. (2002) Accurate formula for *p*-values of gapped local sequence and profile alignments. *J. Mol. Biol.*, **300**, 649–659.

Rockafellar, R.T. (1979) *Convex Analysis*. Princeton, NJ: Princeton University Press.

Siegmund, D. and Yakir, B. (2000) Approximate *p*-values for local sequence alignments. *Ann. Statist.*, **28**(3), 657–680.

Spitzer, F. (1976) *Principles of Random Walks*, 2nd edn. Springer, New York: Springer-Verlag.

Steele, J.M. (1997) *Probability Theory and Combinatorial Optimization.* Philadelphia: Society for Industrial and Applied Mathematics.

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

Zhang, Y. (1995) A limit theorem for matching random sequences allowing deletions. *Ann. Appl. Probab.*, **5**(4), 1236–1240.

Received October 2003 and revised March 2004