# Random walks for calculations of BLAST score significance

Timo Koski
University of Makerere 2018

## 1 Introduction

In this lecture we study the theory of random walks in order to understand the significance calculations in BLAST (=Best Local Alignment Search Tool). BLAST finds alignments or does sequence similarity search by finding high scoring segments (or words) in sequences. If the sequences (a query sequence and a data base sequence) compared are long, then the probability of finding segments that are similar by chance increases. Hence it is important to be able to compute the significance of the maximal score.

Pearson & Todd (2000) have written what follows: BLAST was developed to to identify distantly related - homologous- sequences based on sequence similarity. Two sequences are said to be homologous, if we believe that the two sequences diverged from a common ancestor in the past. Almost without exception, if two (protein) sequences share statistically significant similarity, they will share significant structural similarity.

This lecture is based on sections 7.1, 7.4, 7.5, 7.6, of Ewens and Grant. We start by simple properties of random walks and generalize them as needed in the BLAST calculations. For details of BLAST we refer to a handout of appropriate pages of (Mount 2001). Another clear presentation of the main ideas is by the developers of BLAST, Karlin and Altschul (1990).

## 2 Definitions for Random Walks

Let for $n = 1, \ldots,$

$$S_i = \begin{cases} 1 & \text{with probability } p \\ -1 & \text{with probability } q = 1 - p. \end{cases}$$

Hence $S_i$ are (transformed) I.I.D. Bernoulli R.V's. We take $W_0 = h$. Let

$$W_n = h + \sum_{i=1}^{n} S_i.$$

This is a random motion of a particle that inhabits one of the integer points of the real line. We record the motion of the particle as the sequence $\{(n, W_n)|n =, 1, \ldots, \}$, if these are joined by solid lines between neighbors, the graph is called the *path* of the particle.

We call this the **simple random walk**. The **classical simple random**

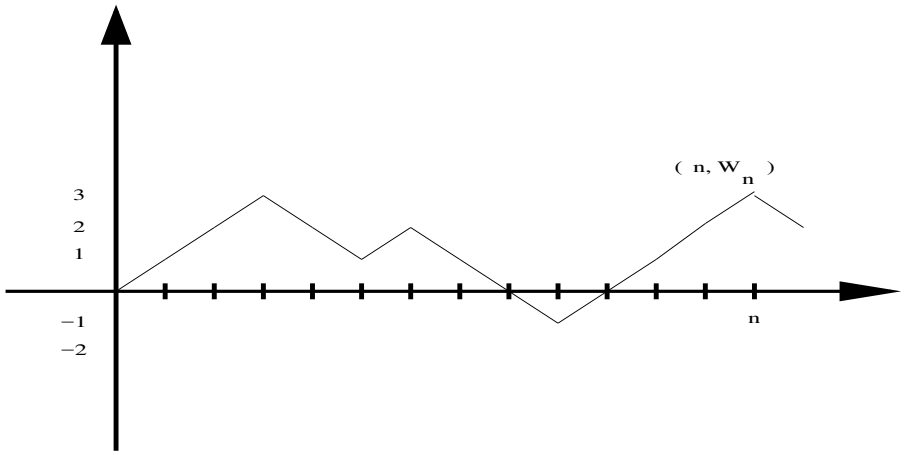


Figure 1: A random walk

**walk** is defined by $p = q = 1/2$.

One interpretation of random walk in terms of bioinformatics is as an accumulated similarity score for two local segments of sequences:

$$W_n(= \overline{s}\,(\mathbf{x}[j:l], \mathbf{y}[j:l])) = \sum_{i=j}^{l} s\,(x_i, y_i)\,.$$

where

$$s(x, y) = \begin{cases} +1 & \text{if } x = y \\ -1 & x \neq y. \end{cases}$$

There are other applications of random walks in bioinformatics or in molecular biology, e.g., (Berg 1993, Paxia et.al. 2002).

# 3 Properties of the simple random walk

**Lemma 3.1** The simple random walk is spatially homogeneous, that is

$$P\left(W_n = j \mid W_0 = h\right) = P\left(W_n = j + b \mid W_0 = h + b\right).$$

*Proof:*

$$P\left(W_n = j \mid W_0 = h\right) = P\left(\sum_{i=1}^{n} S_i = j - h\right)$$

$$P\left(W_n = j + b \mid W_0 = h + b\right) = P\left(\sum_{i=1}^{n} S_i = j - h\right).$$

■

**Lemma 3.2** The simple random walk is temporally homogeneous, that is

$$P\left(W_n = j \mid W_0 = h\right) = P\left(W_{n+m} = j \mid W_m = h\right).$$

*Proof:*

$$P\left(W_n = j \mid W_0 = h\right) = P\left(\sum_{i=1}^{n} S_i = j - h\right) =$$

$$P\left(\sum_{i=m+1}^{m+n} S_i = j - h\right) = P\left(W_{n+m} = j \mid W_m = h\right).$$

**Lemma 3.3** The simple random walk has the Markov property, that is,

$$P\left(W_{n+1} = j \mid W_0, W_1, \ldots, W_n\right) = P\left(W_{n+1} = j \mid W_n\right).$$

*Proof:*

$$P\left(W_{n+1} = j \mid W_0, W_1, \ldots, W_n = a\right) =$$

$$= P\left(S_{n+1} = j - a\right) = P\left(S_{n+1} = j - a \mid W_n = a\right) = P\left(W_{n+1} = j \mid W_n = a\right)$$

■

Hence we are dealing with a Markov chain, whose state space is the set of integers and $\{\ldots, -2, -1, 0, 1, 2, \ldots\}$.

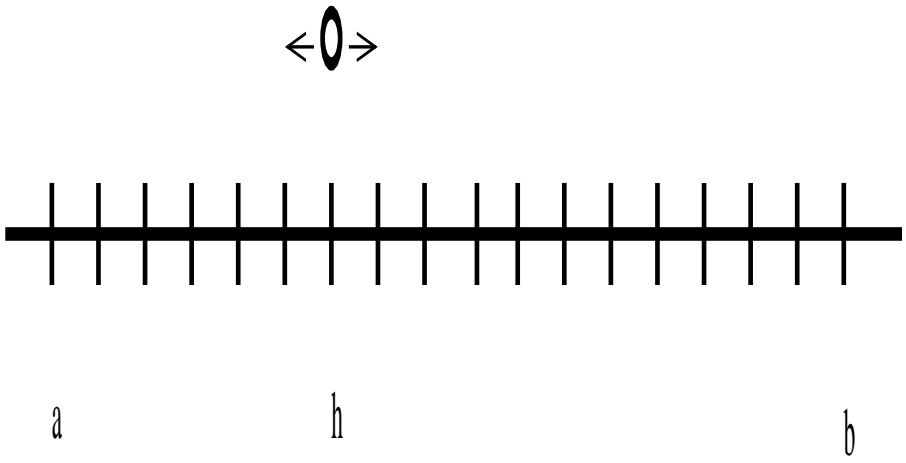


Figure 2: Absorbed random walk

# 4 Absorption probabilities: the m.g.f. method

Let $a$ and $b$ be two integers, $a < b$. We take these points as barriers of absorption in the sense that when the simple random walk hits either of these points it will terminate there.

This is a Markov chain, whose state space is the set of integers $= \{a, \ldots, b\}$ and the transition probability matrix of which is of the form

$$\begin{pmatrix} 1 & 0 & 0 & 0 & \ldots & 0 \\ q & 0 & p & 0 & \ldots & 0 \\ 0 & q & 0 & p & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \ldots & q & 0 & p \\ 0 & 0 & 0 & \ldots & 0 & 1 \end{pmatrix}.$$

We will be interested in two problems

(i) finding the probability that the walk stops at $b$ rather than at $a$.

(ii) finding the mean number of steps until the walk terminates.

We use the moment generating functions for both of these problems. The m.g.f. of any single step $S$ is obviously

$$m(\theta) = qe^{-\theta} + pe^{\theta}.$$

We recall a theorem from page 35 in the textbook to show that there exists $\theta^*$ such that

$$m(\theta^*) = 1.$$

In fact this is a quadratic equation solved by

$$\theta^* = \log\left(\frac{q}{p}\right).$$

Let

$$N = \min\{n \geq 0 \mid W_n = a \quad \text{or} \quad W_n = b\}.$$

Then $N$ is the random number of steps until the simple random walk terminates and is an example of a **Markov time**. The m.g.f. of the total displacement

$$T_N = \sum_{i=1}^{N} S_i$$

after $N$ steps is, since single steps are I.I.D., by a previous result

$$m(\theta)^N = \left(qe^{-\theta} + pe^{\theta}\right)^N$$

Hence

$$m(\theta^*)^N = \left(qe^{-\theta^*} + pe^{\theta^*}\right)^N = 1.$$

We get also

$$w_h = P(W_N = b) = P(T_N = b - h),$$

and

$$u_h = P(W_N = a) = P(T_N = a - h) = 1 - w_h.$$

The moment generating function of $T_N$ is thus

$$\begin{aligned} m_{T_N}(\theta) &= w_h \cdot e^{(b-h)\theta} + u_h e^{(a-h)\theta} \qquad (4.1) \\ &= w_h \cdot e^{(b-h)\theta} + (1 - w_h)e^{(a-h)\theta}. \end{aligned}$$

Then also

$$m_{T_N}(\theta^*) = m(\theta^*)^N = 1,$$

which gives

$$w_h \cdot e^{(b-h)\theta^*} + (1 - w_h)e^{(a-h)\theta^*} = 1$$

$\Leftrightarrow$

$$w_h \cdot \left(e^{b\theta^*} - e^{a\theta^*}\right) = e^{h\theta^*} - e^{a\theta^*}$$

$\Leftrightarrow$

$$\underline{w_h = \frac{e^{h\theta^*} - e^{a\theta^*}}{(e^{b\theta^*} - e^{a\theta^*})}}.$$

Hence we have found the answer to (i) above.

**Proposition 4.1** The probability of absorption at $b$ is

$$w_h = \frac{e^{h\theta^*} - e^{a\theta^*}}{(e^{b\theta^*} - e^{a\theta^*})}. \tag{4.2}$$

■

In order to compute the mean number of steps until the walk terminates, i.e., $E[N]$, we need a result known as Wald's identity:

$$E\left[m(\theta)^{-N} e^{\theta T_N}\right] = 1 \tag{4.3}$$

for all $\theta$ that the m.g.f. exists. The identity (4.3) will not be proved, since the proof requires a lot of groundwork in martingale theory.

Next we differentiate Wald's identity with respect to $\theta$.

$$\frac{d}{d\theta} E\left[m(\theta)^{-N} e^{\theta T_N}\right] = E\left[\frac{d}{d\theta} m(\theta)^{-N} e^{\theta T_N}\right] = 0,$$

where

$$\frac{d}{d\theta} m(\theta)^{-N} e^{\theta T_N} = -N m(\theta)^{-N-1} \frac{d}{d\theta} m(\theta) e^{\theta T_N} + m(\theta)^{-N} T_N e^{\theta T_N}$$

Hence

$$E\left[-N m(\theta)^{-N-1} \frac{d}{d\theta} m(\theta) e^{\theta T_N} + m(\theta)^{-N} T_N e^{\theta T_N}\right] = 0.$$

If we insert $\theta = 0$ in this equation we get

$$E[-NE(S) + T_N] = -E[N] \cdot E[S] + E[T_N] = 0,$$

which gives

$$E[T_N] = E[N] \cdot E[S].$$

Hence

$$E[N] = \frac{E[T_N]}{E[S]}.$$

The expected time to absorption is thus the ratio of the expected displacement at absorption to the expected step size. But we know that

$$E[T_N] = w_h(b-h) + u_h(a-h),$$

and

$$E(S) = p - q.$$

Hence we have found the solution to (ii) above.

**Proposition 4.2** The mean time to absorption is

$$E[N] = \frac{w_h(b-h) + u_h(a-h)}{p-q}. \tag{4.4}$$

■

# 5 An Asymptotic Case

Let us suppose $h = 0$, $a = -1$ and let $b = y$. We are going to let $y \to +\infty$. By this we want to compute

(iii) the distribution of the maximum value of the walk before it hits $a = -1$.

(iv) the mean number of steps until the walk eventually terminates at $a = -1$.

We are going to assume that we are dealing with the case in the theorem from page 35 in the textbook showing that there exists $\theta^*$ such that

$$m(\theta^*) = 1$$

with $\theta^* > 0$: Then we have

$$0 < \theta^* = \log\left(\frac{q}{p}\right). \tag{5.1}$$

$$\Leftrightarrow$$

$$1 < \frac{q}{p} \Leftrightarrow p < q \Leftrightarrow 0 < q - p,$$

which means that the random walk has a negative drift.

We discuss (iii), i.e., the maximum value of the walk ever reaches before hitting $-1$. In (4.2) we have

$$w_h = \frac{e^{h\theta^*} - e^{a\theta^*}}{(e^{b\theta^*} - e^{a\theta^*})},$$

which with $h = 0$, $a = -1$ and $b = y$ yields

$$w_0 = \frac{1 - e^{-\theta^*}}{(e^{y\theta^*} - e^{-\theta^*})}.$$

Hence we can see that for large $y$, since $\theta^* > 0$ is assumed in (5.1),

$$w_0 \propto \left(1 - e^{-\theta^*}\right) e^{-y\theta^*}.$$

Hence, if $Y$ is the maximum of the walk,

$$\text{Prob}\left(Y \geq y\right) \sim \left(1 - e^{-\theta^*}\right) e^{-y\theta^*}$$

as $y \to \infty$, this is a geometric-like probability, where we put

$$C \stackrel{\text{def}}{=} \left(1 - e^{-\theta^*}\right). \tag{5.2}$$

For (iv) we let

$$A \stackrel{def}{=} E\left[N\right]. \tag{5.3}$$

Then with $a = -1$, $h = 0$ and $b = y$

$$A = \frac{w_h(b - h) + u_h(a - h)}{p - q}$$

becomes

$$A = \frac{w_0 y - u_0}{p - q} = \frac{u_0 - w_0 y}{q - p}.$$

In the expression above $w_0 y \to 0$, as $y \to \infty$ and $u_0 = 1 - w_0 \to 1$, since $w_0 \to 0$ Hence

$$A = \frac{1}{q - p}$$

as $y \to \infty$.

$A$ and $C$ will be the quantities we shall concentrate upon in the sequel in more general random walks, since these are needed in BLAST calculations.

# 6 General Random Walks

## 6.1 Assumptions on the Random Walk

Next we study a renewal theory for general random walks in order to prove the basic asymptotic probability results for significance calculations in BLAST Let $S$ be a random variable with a finite number of discrete integer values, the possible step sizes in a random walk, which are

$$-c, -c+1, -c+2, \ldots, 0, \ldots, d-1, d$$

with the respective probabilities

$$p_{-c}, p_{-c+1}, p_{-c+2}, \ldots, p_0, \ldots, p_{d-1}, p_d$$

We assume that

(a)

$$p_{-c} > 0, p_d > 0. \tag{6.1}$$

(b)

$$E(S) = \sum_{j=-c}^{d} j p_j < 0. \tag{6.2}$$

(c) the greatest common divisor of the step sizes that have non-zero probability is 1.

The m.g.f. is

$$m(\theta) = \sum_{j=-c}^{d} p_j e^{j\theta}.$$

By (a) we have, by virtue of the theorem on page 35 in Ewens and Grant, that there is $\theta^*$ for which

$$\sum_{j=-c}^{d} p_j e^{j\theta^*} = 1. \tag{6.3}$$

In view of assumption (b) we have that

$$\theta^* > 0.$$

Let $S_i$ be independent, identically $S_i \stackrel{\mathrm{d}}{=} S$ - distributed random variables. We consider the general random walk

$$W_n = \sum_{i=1}^{n} S_i, W_0 = 0.$$

As earlier, we interpret this as the score in a pairwise alignment or in a similarity search.

## 6.2 Ladders, Excursions, and Maxima

We define the **ladder variables** (or ladder times) as

$$K_0 = 0, K_v = \min_k \{k | k \geq K_{v-1} + 1, W_k - W_{K_{v-1}} < 0\}, v = 1, 2, \ldots, \quad (6.4)$$

see Figure 3. A **ladder point** is thus a point $(n, W_n)$, which is lower than than any previously reached point. Since the drift is negative, the random

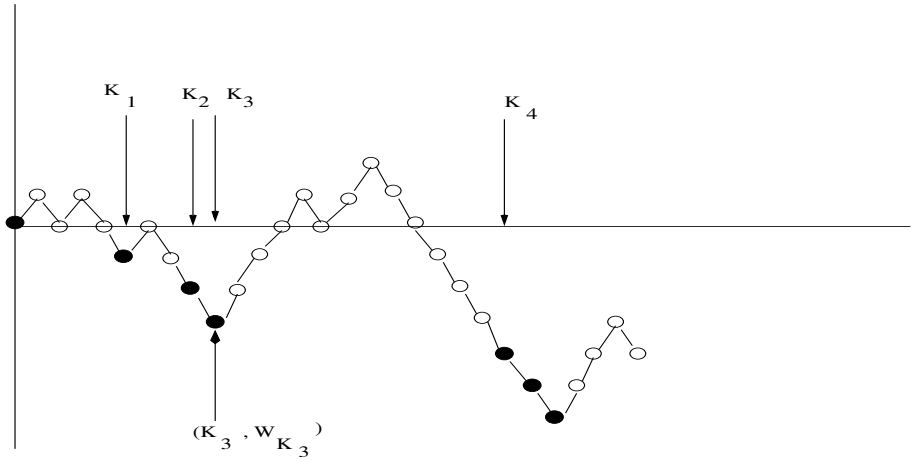


Figure 3: Ladder times

variables $K_v - K_{v-1}$ are positive integer-valued i.i.d. random variables.

The time frame $K_{v-1}+1$ to $K_v$ is called the $v$**th excursion** of the process $\{W_n\}$, see Figure 4.

The variable

$$D_v = \max_{K_{v-1} \leq l < K_v} \left( W_l - W_{K_{v-1}} \right) \quad (6.5)$$

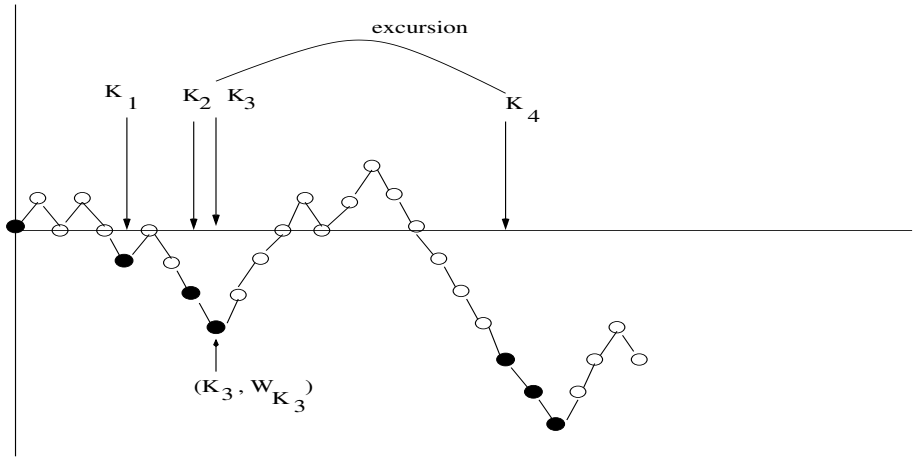


Figure 4: Excursions

evaluates the **maximal displacement attained during the $v$th excursion**, see Figure 5. Obviously $D_1, D_2, \ldots,$ are non-negative i.i.d. random variables. We define

$$M(n) = \max_{0 \le k \le l \le n} (W_l - W_k) . \tag{6.6}$$

Thus $M(n)$ **corresponds to a segment of the sequence $S_1, S_2, \ldots, S_n$ with maximal accumulated score**. The probability of such a segment is what we want to compute.

It follows from definition of $D_v$ that

$$M(n) = \max \left( D_1, D_2, \ldots, D_{R(n)}, D_n^* \right), \tag{6.7}$$

where $R(n)$ is the number of complete excursions over the time frame from 1 to $n$ and $D_n^*$ is the residual maximum during the last incomplete excursion up to time $n$.

We shall use the following notations

$$F(y) = P\left(D_1 \le y\right) \tag{6.8}$$

and

$$A = E\left[K_1\right]. \tag{6.9}$$

We shall prove the following lemma

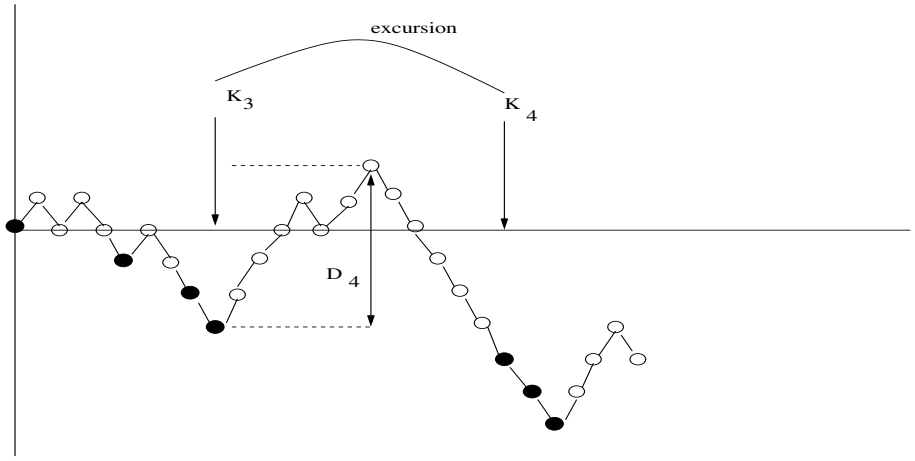


Figure 5: Maximum level during an excursion

**Lemma 6.1** *Under the assumptions* (a) - (c) *above, we have*

$$\lim_{y\to\infty} [1 - F(y)]\, e^{\theta^* y} = C \cdot e^{-\theta^*}. \tag{6.10}$$

*where $\theta^*$ is given in (6.3).*

■

Using this lemma we shall establish the following main theorem, which is the cornerstone of significance computations in BLAST.

**Theorem 6.2** *Under the assumptions* (a) - (c) *above, we have*

$$\lim_{n\to\infty} P\left(M(n) - \frac{\log n}{\theta^*} \le x\right) = e^{-Ke^{-\theta^* x}} \tag{6.11}$$

*where*

$$K = \frac{C^*}{A} e^{-\theta^*} \tag{6.12}$$

■

Various expressions for $K, C^*, A$ and $C$ in (6.10) will be given later. The proof is originally due to (Iglehart 1972, Karlin and Dembo 1992).

# 7 Proof of the lemma

We need a renewal equation and want then to apply the renewal theorem.

Where do we get this from ? This is more or less a standard idea in textboooks in stochastic processes. Let us recall the following well known textbook definition of the renewal process. A renewal process is a non-negative integervalued stochastic process that registers the successive occurrences of an event, where the time durations between consecutive events are positive, independent and identically distributed random variables. Examples of renewal processes are

$$K_v - K_{v-1}, v = 1, 2, \dots,$$

and

$$W_{K_v} - W_{K_{v-1}}, v = 1, 2, \dots, .$$

A renewal process has a regenerative nature, so that we may evaluate expected number of renewals by decomposing the paths of the process by the first time of occurrence of an event. This is the '**renewal argument**', which yields the renewal equation for the expected number of renewals. The renewal theorem gives the limit of the solution of the renewal equation.

## 7.1 Step 1: The renewal equation

Since $E(S) < 0$ the random walk will eventually drift down to $-\infty$. Before doing so, it might visit various positive values. The first step is to find an equation for the probabilities $Q_k$, where

$$Q_k = \text{Prop}\,(\text{ the walk visits } k > 0 \text{ before any other positive value })$$

We define by convention

$$Q_0 = 0.$$

We introduce

$$\sigma^+ = \min\{l \geq 1 | W_l > 0\}, \tag{7.1}$$

so that

$$Q_k = \text{Prob}\,(W_{\sigma^+} = k)\,. \tag{7.2}$$

Since the largest positive step size is $d$ we have $Q_k = 0$ for $k > d$. We have here that

$$\sum_{k=1}^{d} Q_k < 1,$$

since the random walk $W_l$ might always stay in the negative axis, We need an auxiliary result obtained by introducing artificial barriers.

### 7.1.1 Artificial barriers

The random walk is absorbed if it hits $-1$ or any smaller negative integer and an artifical upper termiantion point $y$ or larger. Let us set

$$P_k = \text{Prob}\,(\text{the walk is terminated at k})$$

Let again $N$ be the time of termination and

$$T_N = \sum_{i=1}^{N} S_i.$$

We have again Wald's identity

$$E\left[m\,(\theta)^{-N}\,e^{\theta T_N}\right] = 1. \tag{7.3}$$

If we evaluate this for $\theta = \theta^*$ we have, since $T_N$ has the values $-c, -c+1, \ldots, -1$ and $y, y+d-1$, that

$$\sum_{k=-c}^{-1} P_k e^{k\theta^*} + \sum_{k=y}^{y+d-1} P_k e^{k\theta^*} = 1 \tag{7.4}$$

If we now take the lower artificial boundary at $-L$, where $L >> 0$ and the uper boundary at $+1$, then we obtain from (7.4)

$$\sum_{k=-L-c+1}^{-L} Q_k(L) e^{k\theta^*} + \sum_{k=1}^{d} Q_k(L) e^{k\theta^*} = 1$$

where $Q_k(L)$ is the probability (depending on $L$) that the walk stops at the value $k$. If the walk stops at $k > 0$, then $k$ is the first positive value reached, as $+1$ is the boundary. Therefore

$$\lim_{L\to\infty} Q_k(L) = Q_k$$

Since $\theta^* > 0$ and $Q_k(L) < 1$, we have that

$$\lim_{L\to\infty} \left[ \sum_{k=-L-c+1}^{-L} Q_k(L) e^{k\theta^*} + \sum_{k=1}^{d} Q_k(L) e^{k\theta^*} \right]$$

$$= \sum_{k=1}^{d} Q_k e^{k\theta^*} = 1.$$

The auxiliary result sought for and obtained in this subsection is thus

$$\sum_{k=1}^{d} Q_k e^{k\theta^*} = 1. \tag{7.5}$$

### 7.1.2 Probability for maximum excursion

Next we want to find an equation for (6.8) using $Q_k$ or

$$F(y) = P\left(D_1 \leq y\right). \tag{7.6}$$

The event $\{D_1 \leq y\}$ is the union of several non-overlapping events:

$$\{D_1 \leq y\} = A_- \cup_{k=1}^{y} A_k$$

The $A_-$ is the event that the maximal level never reaches positive values, and has the probability

$$\text{Prob}\left(A_-\right) = \overline{Q} = 1 - Q_1 - Q_2 - \ldots - Q_d. \tag{7.7}$$

The rest of the events are

$$A_k = \{W_{\sigma^+} = k, \max_{\sigma^+ \leq l < K_1} \left(W_l - W_{\sigma^+} \leq y - k\right).$$

A generic $A_k$ is thus the event that the first positive value obtained during an excursion is $k$, $k = 1, 2, \ldots, y$ and that then, starting form the first positive value, the walk achieves a further displacement falling below $y - k$. The events $A_-, \{A_k\}_{k=1}^{y}$ are disjoint. Elementary probability gives

$$F(y) = P\left(D_1 \leq y\right) = \text{Prob}\left(A_-\right) + \sum_{k=1}^{y} \text{Prob}\left(A_k\right).$$

Then, again by elementary conditional probability,

$$\text{Prob}\left(A_k\right) =$$

$$\text{Prob}\left(W_{\sigma^+} = k\right) \cdot \text{Prob}\left(\max_{\sigma^+ \leq l < K_1} \left(W_l - W_{\sigma^+} \leq y - k | W_{\sigma^+} = k\right)\right).$$

Here

$$\text{Prob}\left(\max_{\sigma^+ \leq l < K_1} \left(W_l - W_{\sigma^+} \leq y - k | W_{\sigma^+} = k\right)\right)$$

$$= \text{Prob}\left(\max_{\sigma^+ \leq l < K_1} \left(W_l - W_{\sigma^+}\right) \leq y - k\right),$$

since whatever happens after the random walk has reached the level $W_{\sigma^+} = k$ is independent of things that happened up and prior to $\sigma^+$ (actually the so-called strong Markov property). Hence

$$\text{Prob}\left(A_k\right) = \text{Prob}\left(W_{\sigma^+} = k\right) \text{Prob}\left(\max_{\sigma^+ \leq l < K_1} \left(W_l - W_{\sigma^+}\right) \leq y - k\right),$$

where by previous definition in (7.2)

$$\text{Prob}\left(W_{\sigma^+} = k\right) = Q_k.$$

Now,

$$\text{Prob}\left(\max_{\sigma^+ \le l < K_1} \left(W_l - W_{\sigma^+}\right) \le y - k\right) = \text{Prob}\left(\max_{0 \le l < K_1} \left(W_l - W_0\right) \le y - k\right) =$$

$$\text{Prob}\left(\max_{0 \le l < K_1} \left(W_l\right) \le y - k\right) = F(y-k).$$

This is what we mean by **renewal argument**: in terms probability the random walk starts afresh (like from zero again) after the time $\sigma^+$.

Thus we have

$$F_Y(y) = \overline{Q} + \sum_{k=0}^{y} Q_k F(y-k) \tag{7.8}$$

This is a **renewal equation**. It turns out that we need to rewrite it in order to use the renewal theorem.

## 7.2 Step 2: The renewal equation rewritten

Let us introduce

$$V(y) \stackrel{\text{def}}{=} \left(1 - F(y)\right) e^{y\theta^*}$$

or

$$F(y) = 1 - V(y)e^{-y\theta^*}. \tag{7.9}$$

Then we get from (7.8)

$$1 - V(y)e^{-y\theta^*} = \overline{Q} + \sum_{k=0}^{y} Q_k \left(1 - V(y-k)e^{-(y-k)\theta^*}\right)$$

and reorganize this as

$$V(y) = e^{y\theta^*}\left[1 - \overline{Q} - \sum_{k=0}^{y} Q_k + \sum_{k=0}^{y} Q_k V(y-k)e^{-(y-k)\theta^*}\right].$$

Here from (7.7)

$$1 - \overline{Q} - \sum_{k=0}^{y} Q_k = Q_1 + Q_2 + \ldots + Q_d - \sum_{k=0}^{y} Q_k,$$

and for $y < d$ this gives, since $Q_0 = 0$,

$$V(y) = e^{y\theta^*}\left(Q_{y+1} + Q_{y+2} + \ldots + Q_d\right) + \sum_{k=0}^{y} e^{k\theta^*} Q_k V(y-k), \qquad (7.10)$$

and for when $y \geq d$

$$V(y) = \sum_{k=0}^{d} e^{k\theta^*} Q_k V(y-k). \qquad (7.11)$$

The equations (7.10) and (7.11) are in a form that permits application of a **renewal theorem** (see Appendix A below).

## 7.3 Step 3: Application of the renewal theorem

In this theorem we choose

$$u_y = V(y),$$
$$f_k = e^{k\theta^*} Q_k,$$

and

$$b_y = e^{y\theta^*}\left(Q_{y+1} + Q_2 + \ldots + Q_d\right),$$

if $y < d$ and $b_y = 0$ if $y \geq d$.

In the renewal theorem we find

$$B = \sum_y b_y = \sum_{k=0}^{d} e^{k\theta^*}\left(Q_{k+1} + Q_{k+2} + \ldots + Q_d\right).$$

If we multiply this equation by $e^{\theta^*} - 1$, we have

$$(e^{\theta^*}-1)B = \sum_{k=0}^{d} e^{(k+1)\theta^*}\left(Q_{k+1} + Q_{k+2} + \ldots + Q_d\right) - \sum_{k=0}^{d} e^{k\theta^*}\left(Q_{k+1} + Q_{k+2} + \ldots + Q_d\right).$$

- Here

$$\sum_{k=0}^{d} e^{(k+1)\theta^*}\left(Q_{k+1} + Q_{k+2} + \ldots + Q_d\right) - \sum_{k=0}^{d} e^{k\theta^*}\left(Q_{k+1} + Q_{k+2} + \ldots + Q_d\right) =$$

$$e^{\theta^*}\left(Q_1 + Q_2 + \ldots + Q_d\right) + e^{2\theta^*}\left(Q_2 + Q_3 + \ldots + Q_d\right) + e^{3\theta^*}\left(Q_3 + Q_4 + \ldots + Q_d\right) + \ldots$$
$$- \left(Q_1 + Q_2 + \ldots + Q_d\right) - e^{\theta^*}\left(Q_2 + Q_3 + \ldots + Q_d\right) - e^{2\theta^*}\left(Q_3 + Q_4 + \ldots + Q_d\right) - \ldots =$$
$$e^{\theta^*}Q_1 + e^{\theta^*}Q_2 + \ldots e^{2\theta^*}Q_2 + e^{2\theta^*}Q_3 + \ldots + e^{3\theta^*}Q_3 + e^{3\theta^*}Q_4 + \ldots$$
$$- \left(Q_1 + Q_2 + \ldots + Q_d\right) - e^{\theta^*}Q_2 - e^{\theta^*}Q_3 -$$
$$- e^{2\theta^*}Q_3 - e^{2\theta^*}Q_4 - \ldots$$

Checking the pattern of terms that cancel out in the expression above we get that

$$(e^{\theta^*} - 1)B =$$

$$\sum_{k=1}^{d} Q_k e^{k\theta^*} - (Q_1 + Q_2 + \ldots + Q_d) = 1 - (Q_1 + Q_2 + \ldots + Q_d) = \overline{Q}.$$

in view of (7.5). Hence

$$B = \frac{\overline{Q}}{e^{\theta^*} - 1}.$$

## 7.4 Step 4: verification of the conditions in the renewal theorem

For the intended application of the renewal theorem, proposition A.1 in Appendix A below, we should verify the conditions of the theorem. These are that

$$B = \sum_{k=0}^{\infty} b_k < +\infty,$$

which we already know, and

$$\sum_{k=0}^{\infty} f_k = 1, \sum_{k=0}^{\infty} k f_k < +\infty,$$

where we have taken

$$f_k = e^{k\theta^*} Q_k,$$

and

$$b_y = e^{y\theta^*} \left(Q_{y+1} + Q_2 + \ldots + Q_d\right),$$

The condition

$$\sum_{k=0}^{\infty} f_k = 1$$

is true by (7.5) above, since $Q_k = 0$ for $k > d$. Then also

$$\sum_{k=0}^{\infty} k f_k = \sum_{k=0}^{\infty} k e^{k\theta^*} Q_k = \sum_{k=0}^{d} k e^{k\theta^*} Q_k < +\infty,$$

and the renewal theorem as stated in the Appendix A can be used.

### 7.5 Step 5: The limit in the lemma

The renewal theorem implies now that

$$V = \lim_{y\to\infty} V(y),$$

and (see (A.1)) that

$$V = \frac{\overline{Q}}{(e^{\theta^*}-1)\sum_{k=0}^{d} kQ_k e^{k\theta^*}} \tag{7.12}$$

From

$$V(y) = (1-F(y))\, e^{y\theta^*}$$

we have

$$F(y) = 1 - V(y)e^{-y\theta^*}. \tag{7.13}$$

or

$$[1-F(y)]\, e^{\theta^* y} = V(y).$$

Hence

$$\lim_{y\to\infty} [1-F(y)]\, e^{\theta^* y} = V = Ce^{-\theta^*},$$

where from (7.12)

$$C = \frac{\overline{Q}}{(1-e^{-\theta^*})\sum_{k=0}^{d} kQ_k e^{k\theta^*}}.$$

This proves lemma 6.1, as claimed, and gives an expression for the constant $C$. ∎

# 8 Proof of the Main Theorem

We now want to prove

$$\lim_{n\to\infty} P\left(M(n) - \frac{\log n}{\theta^*} \le x\right) = e^{-Ke^{-\theta^* x}}, \tag{8.1}$$

where

$$K = \frac{C^*}{A} e^{-\theta^*} \tag{8.2}$$

and

$$M(n) = \max_{0 \le k \le l \le n} (W_l - W_k) . \tag{8.3}$$

By our definitions we have

$$M(n) = \max \left( D_1, D_2, \ldots, D_{R(n)}, D_n^* \right), \tag{8.4}$$

where $R(n)$ is the number of complete excursions over the time frame from 1 to $n$ and $D_n^*$ is the residual maximum during the last incomplete excursion up to time $n$.

Let us consider the random walk from 0 to finite number $(m)$ of full excursions, which in view of the definition of $K_m$ in (6.4) gives

$$M\left(K_m\right) = \max\left(D_1, D_2, \ldots, D_m\right). \tag{8.5}$$

Since $D_1, D_2, \ldots, D_m$ are i.i.d. reandom variables, we have by a basic elementary result, see e.g. (2.89) on page 82 of Ewens and Grant, that

$$P\left(M\left(K_m\right) - \frac{\log m}{\theta^*} \le x\right) = \left[P\left(D_1 - \frac{\log m}{\theta^*} \le x\right)\right]^m,$$

and from (7.6), $F(y) = P\left(D_1 \le y\right)$,

$$= \left[F\left(x + \frac{\log m}{\theta^*}\right)\right]^m.$$

We write the right hand side as

$$\left[F\left(x + \frac{\log m}{\theta^*}\right)\right]^m = e^{m \log F\left(x + \frac{\log m}{\theta^*}\right)},$$

where from (7.13)

$$e^{m \log F\left(x + \frac{\log m}{\theta^*}\right)} = e^{m \log\left[1 - V\left(x + \frac{\log m}{\theta^*}\right) e^{-x\theta^* - \log m}\right]}.$$

We use the series expansion (valid for $-1 \le x < 1$)

$$\log(1+x) = x - \frac{1}{2}x^2 + \ldots$$

with the result

$$m \log\left[1 - V\left(x + \frac{\log m}{\theta^*}\right) e^{-x\theta^*} e^{\log 1/m}\right] =$$

$$= -m\frac{1}{m}V\left(x + \frac{\log m}{\theta^*})\right) e^{-x\theta^*} - \frac{1}{2}m\frac{1}{m^2}G(m) + O\left(\frac{1}{m}\right)\ldots$$

where $G(m)$ is a bounded function of $m$ and $O\left(\frac{1}{m}\right)$ represents the omitted terms that turn to zero as $m \to +\infty$. Hence we have

$$\lim_{m\to\infty} m \log\left[1 - V\left(x + \frac{\log m}{\theta^*})\right) e^{-x\theta^*} e^{\log 1/m}\right]$$

$$= -Ve^{-x\theta^*} = -Ce^{-\theta^*}e^{-x\theta^*}.$$

In other words we have proved that

$$\lim_{m\to\infty} P\left(M\left(K_m\right) - \frac{\log m}{\theta^*} \leq x\right) = e^{-Ce^{-\theta^*}e^{-x\theta^*}}.$$

The property that we can write $C = C^*/A$ for an explicit $C^*$ is not expanded upon in a special case later.

The result just shown is valid for $K_{\lfloor n\rfloor/A}$. A law of large numbers proves that

$$K_{\lfloor n\rfloor/A}/n \to 1,$$

and from this the result in the theorem follows in the general case by continuity and monotonicity. ∎

# 9 A Restricted Walk

## 9.1 The Main Theorem

The constants $K$, $C$, $A$ are specific to the random walk. Next we compute some of these for a restricted walk having a stopping boundary at $-1$. We assume that the walk starts at 0. Let now $F(y)$ from (6.8) above

$$F(y) = P\left(D_1 \leq y\right)$$

be denoted by $F_{\text{unr}}(y)$, this is the excursion maximum distribution of the unrestricted random walk in the preceding, and let

$$F^*_{\text{unr}}(y) = 1 - F_{\text{unr}}(y)$$

Also

$$F^*_{\text{restr}}(y) = 1 - F_{\text{restr}}(y)$$

holds for the distribution function of the excursion maximum distribution of the restricted random walk to be treated next, i.e., with a stopping boundary at $-1$.

Then from (7.9)

$$F_{\text{unr}}(y) = V(y)e^{-y\theta^*}$$

and

$$\lim_{y\to\infty} F_{\text{unr}}(y)e^{y\theta^*} = V. \tag{9.1}$$

The size of an excursion of of the unrestricted walk can exceed the value $y$ either before or after the random walk has reached negative values (starting at 0). In the latter case the first negative value reached by the random walk is one of $-1, -2, \ldots, -c$. Let the probability of first hitting $-j$ be $R_j$. Then by the same intutition about renewals that was used above to find (7.8)

$$F^*_{\text{unr}}(y) = F^*_{\text{restr}}(y) + \sum_{j=1}^{c} R_{-j} F^*_{\text{unr}}(y+j)$$

(Yes, up till the time that the unrestricted random walk hits $-1, -2, \ldots, -c$, the unrestricted and restricted random walks are identical).

Then we multiply by $e^{y\theta^*}$ throughout in this equation and we get from (9.1), since

$$F_{\text{unr}}(y+j) = V(y+j)e^{-(y+j)\theta^*},$$

that

$$V = \lim_{y\to\infty} F^*_{\text{restr}}(y) + V\sum_{j=1}^{c} R_{-j}e^{-j\theta^*}$$

and from this

$$\lim_{y\to+\infty} F^*_{\text{restr}}(y) = V\left[1 - \sum_{j=1}^{c} R_{-j}e^{-j\theta^*}\right].$$

From (7.12)

$$\lim_{y\to\infty} F^*_{\text{restr}}(y) = \frac{\overline{Q}\left(1 - \sum_{j=1}^{c} R_{-j}e^{-j\theta^*}\right)}{(e^{\theta^*} - 1)\sum_{k=0}^{d} kQ_k e^{k\theta^*}}.$$

This we write as

$$\lim_{y\to\infty} F^*_{\text{restr}}(y)e^{y\theta^*} = C \cdot e^{-\theta^*},$$

where

$$C = \frac{\overline{Q}\left(1 - \sum_{j=1}^{c} R_{-j} e^{-j\theta^*}\right)}{(e^{\theta^*} - 1)\sum_{k=0}^{d} k Q_k e^{k\theta^*}}$$

Then the limit

$$\lim_{n\to\infty} P\left(M(n) - \frac{\log n}{\theta^*} \leq x\right) = e^{-Ke^{-\theta^* x}}, \tag{9.2}$$

can (?) be found as above. Here

$$K = \frac{C^*}{A} e^{-\theta^*}.$$

Next we give formulas for $C^*$ and $A$.

## 9.2 The constants

From the equation

$$A = \frac{E\left[T_N\right]}{E\left[S\right]}$$

(obtained from Wald's identity) we get

$$A = \frac{\sum_{j=1}^{c} j R_{-j}}{\sum_{j=1}^{c} j p_j}.$$

Karlin & Dembo (1992) have shown that

$$C = \frac{C^*}{A},$$

where

$$C^* = \frac{\left(1 - \sum_{j=1}^{c} R_{-j} e^{-j\theta^*}\right)^2}{(1 - e^{-\theta^*}) E\left[S e^{\theta^* S}\right]}.$$

# 10 Practical Approximations

## 10.1 Null hypothesis on aligned pairs and the scoring random walk

We consider next the random walk as the score in a pairwise alignment, which we read from from left to right along the positions $1, 2, \ldots, N$. The

generic term is ascore $S_{j,k}$, e.g., an element in a substitution matrix (PAM(n), BLOSUM), which is assigned to each position, where aligned pair $(j,k)$ (of amino acids) is observed.

$$W_N = \sum_{l=1}^{N} S_{j,k}(l), W_0 = 0.$$

We use the **null hypothesis** of independence

$$\text{Prob}\left((j,k)\right) = p_k p_j.$$

The m.g.f. is then

$$m\left(\theta\right) = \sum_{j,i} p_k p_j e^{S(j,k)\theta}.$$

We assume that the scoring system (substitution matrix) is such that there is $\theta^*$ for which

$$m\left(\theta^*\right) = 1, \tag{10.1}$$

and that

$$\theta^* > 0.$$

This is obtained if for there is positive probability for positive and negative scores and

$$\sum_{j,i} p_k p_j S(j,k) < 0.$$

Here we write

$$\theta^* \leftrightarrow \lambda. \tag{10.2}$$

## 10.2 The asymptotic distribution and the constants therein

Let $M(n)$ be the score of the maximally scoring subsequence, or subalignment. We have shown

$$\lim_{n\to\infty} P\left(M(n) - \frac{\log n}{\lambda} \leq x\right) = e^{-Ke^{-\lambda x}} \tag{10.3}$$

The pertinent constant $K$ is

$$K = \frac{C^*}{A} e^{-\lambda} \tag{10.4}$$

and $C^*$ and $A$ have been given in subsection 9.2.

## 10.3 Expressions for $K, C, C^*, A$

From the equation

$$A = \frac{E[T_N]}{E[S]}$$

(obtained from Wald's identity) we get

$$A = \frac{\sum_{j=1}^{c} jR_{-j}}{\sum_{j=1}^{c} jp_j}.$$

It has been shown show that

$$C = \frac{C^*}{A},$$

where

$$C^* = \frac{\left(1 - \sum_{j=1}^{c} R_{-j} e^{-j\lambda}\right)^2}{(1 - e^{-\lambda} E[Se^{\lambda S}]}.$$

and the meaning of $R_{-j}$ is explained in a previous lecture.

Since $A$ is also the mean distance between the ladder variables (times), and the sequences to be compared are of length $N$, then

$$n = \frac{N}{A}$$

is a meaningful and intuitive formula for the mean number of ladder points.

## 10.4 Choice of $S_{j,k}$ on statistical grounds

If the pair (of amino acids) $(j, k)$ is observed at any position, then the alternative hypothesis for the probability of this pair is denoted by

$$q_{j,k}.$$

Then, if we insist that $S_{j,k}$ should be the log likelihood ratio, we should use

$$S_{j,k} = \log \frac{q_{j,k}}{p_k p_j}.$$

In fact Karlin and Altschul (as well as Ewens and Grant) use an alternative argument:

Let us define $q_{j,k}$ as

$$q_{j,k} = p_j m_{j,k}^{(n)},$$

where $m_{j,k}^{(n)}$ the array $(j,k)$ in a $PAM(n)$ matrix. Then

$$\log \frac{q_{j,k}}{p_k p_j} = \log \frac{m_{j,k}^{(n)}}{p_k},$$

and we take $\lambda$ from (10.2). If we set

$$S_{j,k} = \frac{1}{\lambda} \log \frac{q_{j,k}}{p_k p_j}$$

Then

$$q_{j,k} = p_k p_j e^{S(j,k)\lambda}$$

is a probability distribution by our choice of $\lambda$ (see (10.1) and (10.2).

## 10.5 Relative Entropy

The relative entgropy or Kullback distance is

$$H = \sum_{j,k} q_{j,k} \log \frac{q_{j,k}}{p_k p_j} = \lambda \sum_{j,k} q_{j,k} S(j,k) = \lambda E\left[S\right]$$

Thus, by a piece of mathematics due Karlin and Altschul and others, the mean score of the high-scoring segments is asymptotically

$$\frac{1}{\lambda} H.$$

This is used in BLAST printouts.

## 10.6 Normalized and Bit Scores, P-values

Karlin and Altschul (1993) call the expression $Sc$ defined by

$$Sc \stackrel{\text{def}}{=} \lambda M(n) - \log NK \tag{10.5}$$

a normalized score. Then

$$P\left(\lambda M(n) - \log NK \leq s\right) = P\left(\lambda M(n) - \log N \leq s + \log K\right) =$$

$$= P\left(M(n) - \frac{1}{\lambda}\log N \leq \frac{1}{\lambda}s + \frac{1}{\lambda}\log K\right)$$

and from (6.11)

$$\begin{aligned} &\approx e^{-Ke^{-\lambda\left[\frac{1}{\lambda}s+\frac{1}{\lambda}\log K\right]}} \\ &= e^{-Ke^{-\lambda\left[\frac{1}{\lambda}s+\frac{1}{\lambda}\log K\right]}} \\ &= e^{-Ke^{-s-\log K}} = e^{-e^{-s}}. \end{aligned}$$

Hence we get

$$P\left(\lambda M(n) - \log NK \geq s\right) \approx 1 - e^{-e^{-s}}.$$

**The P-value corrersponding to an observed value of the normalized score $\lambda m(n) - \log NK$ is thus**

$$\mathrm{P-value} = 1 - e^{-e^{-\lambda m(n) - \log NK}}. \tag{10.6}$$

From an earlier study of extreme value theory ( see problem 2.15 in Ewens and Grant) we get thus that

$$E\left[M(n)\right] = \lambda^{-1}\left(\log NK + \kappa\right)$$

where $\kappa$ is Euler's constant.

The normalized score has an absolute intrepretation, which is independent of the substitution matrix underlying $S(j,k)$, since $e^{-e^{-s}}$ is independent of all the parameters.

## 10.7 The number of high-scoring excursions

We know from the preceding analysis that for any high scoring segment, the corresponding score, $D_v$ has the distribution function

$$F(y) = P\left(D_v \leq y\right) \tag{10.7}$$

and that

$$1 - F(y) \approx C \cdot e^{-\lambda y}. \tag{10.8}$$

Then the mean number of excursions with height higher than $y$ is approximately equal to

$$\frac{N}{A}e^{-\lambda y}.$$

This is in BLAST approximated by

$$NKe^{-\lambda y},$$

where $K = \frac{C^*}{A}e^{-\lambda}$. Since $D_1, D_2, \ldots,$ are i.i.d. (a renewal process), then the number of excursions having height more than $y$ or more is binomial with mean $NKe^{-\lambda y}$. We approximate this, as in any grundkurs, by a Poisson distribution.

Thus, the actual number of **high segment pairs** (HSP) exceeding $y$ can be compared with the tail of a Poisson distribution. The expected value of the number of excursions corrseponding to an observed maximal score, $m(n)$, is found by replacing the arbitrary number $v$ in $NKe^{-\lambda y}$ by $m(n)$. The expected value is denoted by $E^{'}$, so that

$$E^{'} = NKe^{-\lambda m(n)}.$$

Then the normalized score $Sc$ in (10.5) is

$$Sc = -\log E^{'}$$

and the relation between $P$-value in (10.6) and $E^{'}$ is

$$P\text{-value} \approx 1 - e^{-E^{'}}, E^{'} = -\log(\ P\text{-value}).$$

These are used in BLAST printouts.

# 11 Statistical composition of high-scoring segments

Karlin, Dembo and Kawabata (1990) formulate the follwoing statement, which is the rationale for the parameter calculationsin the preceding section.

**Proposition 11.1** *If the probability of $(j,k)$ is $p_j p_k$, then the frequency of $(j,k)$ in the maximal (in any sufficiently high scoring) segment is approaching*

$$p_j p_k e^{\lambda S_{j,k}}.$$

∎

# A The Renewal Theorem

**Proposition A.1** Suppose the three sequences $\{b_j\}_{j=0}^{\infty}$, $\{f_j\}_{j=0}^{\infty}$ and $\{u_j\}_{j=0}^{\infty}$ of nonnegative constants satisfy the equation

$$u_y = b_y + \left( \sum_{k=0}^{y} u_{y-k} f_y \right)$$

for all $y$. Suppose further that

$$B = \sum_{k=0}^{\infty} b_k < +\infty,$$

$$\sum_{k=0}^{\infty} f_k = 1,$$

and

$$\mu = \sum_{k=0}^{\infty} k f_k < +\infty,$$

and that the greatest common divisor of $\{f_j\}_{j=0}^{\infty}$ is one. Then

$$u_y \to \frac{B}{\mu} \tag{A.1}$$

as $y \to +\infty$

■

# B References


H.C. Berg (1993): *Random Walks in Biology* . Princeton Univ. Press, Princeton.

D. Iglehart (1972): Extreme values in the $GI/G/1$ queue. *Annals of Mathematical Statistics*, 43, pp. 627−635.

S. Karlin and S. Altschul (1990): Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Procedings of the National Academy of Sciences, USA* , 87, pp. 2264−2268.

S. Karlin and S. Altschul (1990): Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes.

*Procedings of the National Academy of Sciences, USA* , 87, pp. 2264−2268.
S. Karlin and A. Dembo (1992): Limit distributions of maximal segmental score among Markov-dependent partial sums. *Advances in Applied Probability*, 24, pp. 113−140.
D.W. Mount (2000): *Bioinformatics.* Cold Spring Harbor Press. Cold Spring Harbor, New York.
S. Paxia, A.Rudra and B. Mishra (2002): A Random Walk Down the Genomes: DNA Evolution in Valis. *IEEE Spectrum* , July, pp. 11−17.
W.R. Pearson and T.C. Todd (2000): Statistical Significance in Biological Sequence Comparison. *somewhere in the www*