

## Introduction to the statistical theory of local alignment

Suppose  $X_1, X_2, \dots, X_n$  are independently identically distributed with exponential density  $\lambda^{-\lambda x}$ , for  $x > 0$  where  $\lambda > 0$ . Let  $M_n = \max(X_1, \dots, X_n)$

$$\begin{aligned} \text{pr}(M_n \leq x) &= \text{pr}(X_1 \leq x, \dots, X_n \leq x) \\ &= \text{pr}(X_1 \leq x) \times \dots \times \text{pr}(X_n \leq x) \\ &= (1 - e^{-\lambda x})^n \\ &= \left(1 - \frac{ne^{-\lambda x}}{n}\right)^n \\ &\approx \exp(-ne^{-\lambda x}) \quad \text{for large } n \\ &= \exp\left(-\exp\left[-\lambda\left(x - \frac{\log n}{\lambda}\right)\right]\right) \end{aligned}$$

In fact

$$\begin{aligned} E[M_n] &= \frac{1}{n\lambda} + \frac{1}{(n-1)\lambda} + \dots + \frac{1}{\lambda} \\ &= \frac{\log n + \gamma}{\lambda} \quad \gamma = \text{Euler's constant} \approx .577 \end{aligned}$$

$$\begin{aligned} \text{Var}[M_n] &= \frac{1}{(n\lambda)^2} + \dots + \frac{1}{((n-1)\lambda)^2} \\ &\approx \frac{\pi}{6\lambda^2} \end{aligned}$$

Compare this to to the Extreme Value distribution which has cdf  $e^{-e^{-x}}$  for  $-\infty < x < \infty$ , pdf  $e^{-x-e^{-x}}$ , with mean  $\gamma$  and variance  $\frac{\pi^2}{6}$ .

We saw that if  $X_1, X_2, \dots$  were iid  $\text{Exp}(\lambda)$  then

$$\text{pr}\left(\max_{i=1}^n X_i \leq x\right) \approx \exp(-\exp[-\lambda(x - u_n)])$$

where  $U_n = \frac{\log n}{\lambda}$ . This is the easiest way to see the extreme value distribution EV arise

**Corollary** If  $X_1, \dots$  are iid with exponentially decreasing tails, i.e.  $\text{pr}(X_1 \geq x) \approx C \exp(-\lambda x)$  for  $x$  large. The same argument [Exercise!] then shows that

$$\text{pr}\left(\max_{i=1}^n X_i \leq x\right) \approx \exp(-nC \exp[-\lambda(x - u_n)])$$

Note As pointed out, this does not hold for iid  $N(0, 1)$ 's, although the EV still arises with  $u_n \approx \sqrt{\log n}$ . Here  $1 - \Phi(x) \approx \frac{\phi(x)}{x}$ ,  $x$  large.

Now lets turn to two aligned DNA sequences  $p_a, p_g$  etc, then a match occurs with probability

$$p = p_a^2 + p_g^2 + p_c^2 + p_t^2$$

```

g g a g a c t g t a g a c a g c t a a t g c t a t a
  |   |   |||  ||           |||
c a a c g c c c t a g c c a c g a g c c c t t a t c

```

Figure 1: Two aligned DNA sequences

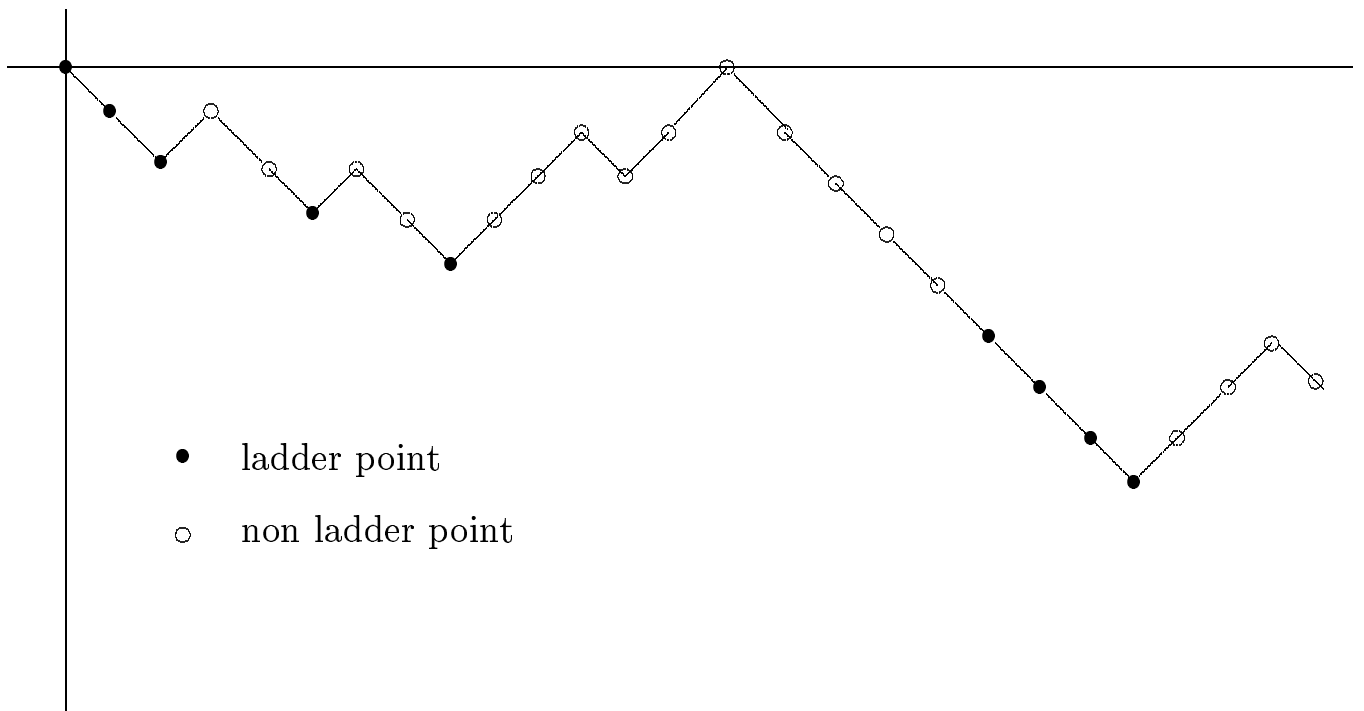


Figure 2: Random walk associated with two aligned sequences

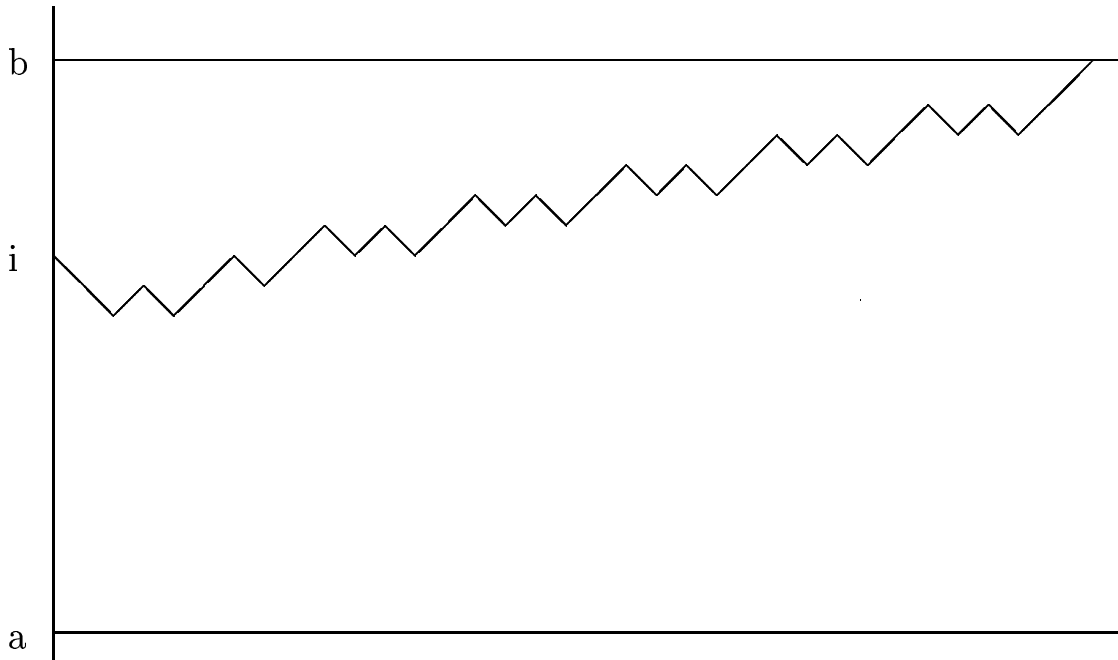


Figure 3: A simple random walk

and a mismatch occurs with probability  $q = 1 - p$ . We can consider this alignment a  $p \nearrow, q \searrow$  random walk as illustrated for our example alignment in figure 2.

We call a point a ladder point if it is the first time we have reached a point that far down. Figure 2 illustrates this for our sequence.

Let  $Y$  be the number of matching bases following a non-matching pair. Then

$$\text{pr}(Y = y) = qp^y \quad y = 0, 1, \dots$$

If  $Y_{\max}$  is the max of  $n$  iid such, then

$$\begin{aligned} \text{pr}(Y_{\max} \leq y) &= (1 - p^y)^n \quad \text{can do exactly, or} \\ &\approx \exp(-n \exp[y \log p]) \end{aligned}$$

if  $p = \frac{1}{4}$ , and  $y = 10$  that this is  $\approx .07$ . When  $y = 12$  it is  $\approx .0045$

In practice we score matches and mismatches and do not go for long runs.

## Some results concerning the simple random walk

### Absorption probabilities

Consider the random walk as illustrated in figure 3. It has probability  $p$  of moving  $\nearrow +1$  and probability  $q$  of moving  $\searrow -1$ .

Let  $w_i$  = probability of hitting  $b$  first starting at  $i$   $w_a = 0, w_b = 1$   $w_i = pw_{i+1} + qw_{i-1}, i \neq a, b$

Try  $w_i = s^i$  get  $s = 1$  or  $s = q/p$  if  $p \neq q$ . Thus  $w_i = A + B (q/p)^i$  and we get to

$$w_i = \frac{\left[ \left(\frac{q}{p}\right)^i - \left(\frac{q}{p}\right)^a \right]}{\left[ \left(\frac{q}{p}\right)^b - \left(\frac{q}{p}\right)^a \right]}$$

Is this unique? Lets examine some values. First considering  $a = 0, b = 100$  we get the following table

$p$	$w_{40}$
.5	.40
.51	.81
.52	.96

and then considering  $a = 0, b = 200$

$p$	$w_{80}$
.5	.40
.51	.96
.52	.998

when  $p = q = \frac{1}{2}$ ,  $w_i = \frac{i-a}{b-a}$

### Mean time to absorption

$$m_i = pm_{i+1} + qm_{i-1} + 1$$

with  $m_a = m_b = 0$ . Inhomogeneous: need to solve homogeneous and get a PS. One such PS is  $m_i = ki$  with  $k = \frac{1}{q-p}$  (need  $p \neq q$ ). So  $m_i = \frac{i}{q-p} + A + B \left(\frac{q}{p}\right)^i$ . thus get

$$m_i = \frac{w_i (b - i) + (1 - w_i) (a - i)}{p - q}$$

*Exercise*  $m_i$  when  $p = \frac{1}{2}$  PS of  $-i^2$ ?

### Alternative approach

If  $S_0 = i$ ,  $S_n = S_{n-1} + \text{increment}$ ,  $Z_n = \frac{\exp \theta S_n}{m(\theta)^n}$  is a martingale (MG) (give proof). Here

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

is the moment generating function of the increment. Solve  $m(\theta) = 1$  for  $\theta$  and we get  $\theta = 0$  and  $\theta^* = \log \left(\frac{q}{p}\right)$ . Suppose  $q > p$ , that is a drift down. Then  $\theta^* > 0$ . (see figure 4) and  $m(\theta)$  is convex,  $= 1$  at zero, slope  $< 0$  at 0.

We use MG stopping time theorem (aka Wald's identity) to get formula for  $w_i$ . if  $N = \#$  of steps until hit hit a or b then

$$\mathbb{E} \left[ e^{\theta^* S_N} | S_0 = i \right] = e^{\theta^* i}$$

Also get  $m_i$  from Wald.

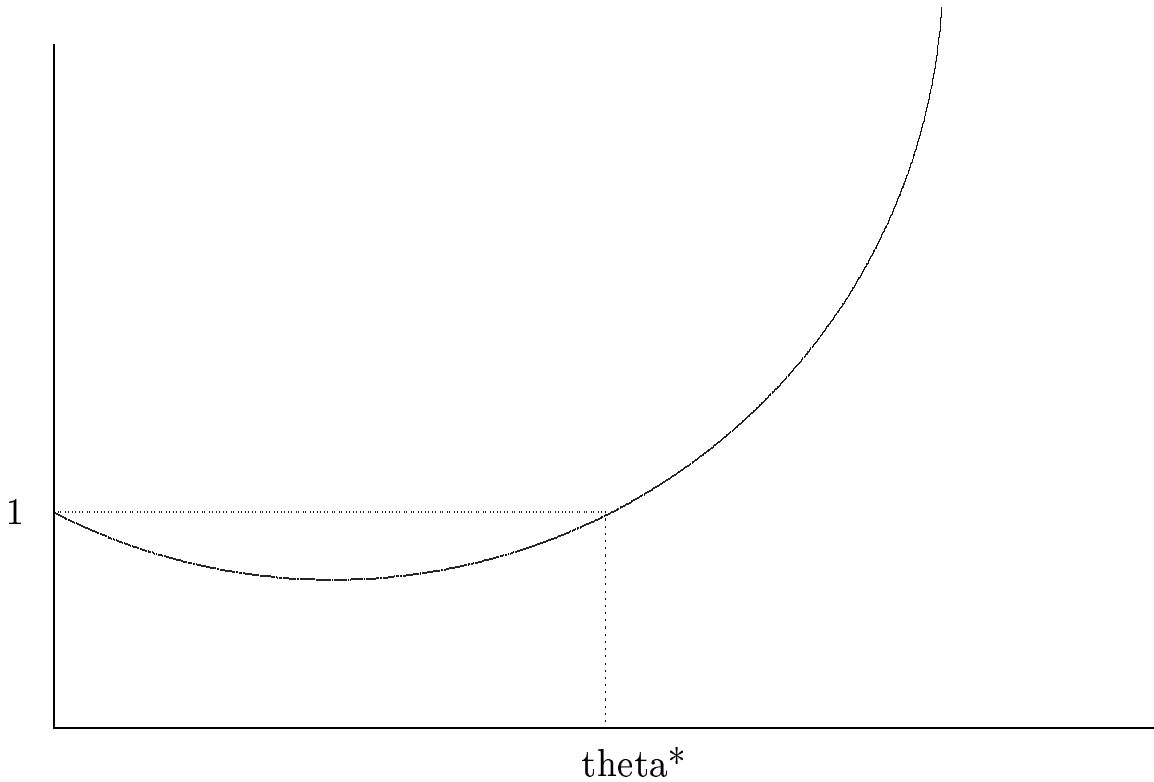


Figure 4:  $m(\theta)$

### Martingale Stopping theorem

For any stopping time  $N$ , Martingale  $Z_i$ .

$$E(Z_N) = E(Z_0)$$

Case  $b \rightarrow \infty$

Suppose we are interested in the case  $b \rightarrow \infty, a = -1, i = 0$ . If  $q > p$ , we find that  $w_0 = 0$ . ie we hit  $-1$  with probability 1 and the mean time until doing so is  $\frac{1}{q-p}$ . Is this obvious? Again use Wald.

Now lets ask for the probability  $p(y)$  that the walk stops  $\leq y$  before hitting  $-1$ . How? Put  $b = y + 1, a = -1$  and find

$$p(y) = \frac{\left[ \left(\frac{q}{p}\right)^0 - \left(\frac{q}{p}\right)^{-1} \right]}{\left[ \left(\frac{q}{p}\right)^{y+1} - \left(\frac{q}{p}\right)^{-1} \right]} \sim \frac{q-p}{q} \left(\frac{p}{q}\right)^{y+1}$$

Equivalently,

$$\begin{aligned} p(y) &= \text{pr}(\text{max of walk} \leq y) \\ &\sim 1 - \frac{q-p}{q} \left(\frac{p}{q}\right)^{y+1} \end{aligned}$$

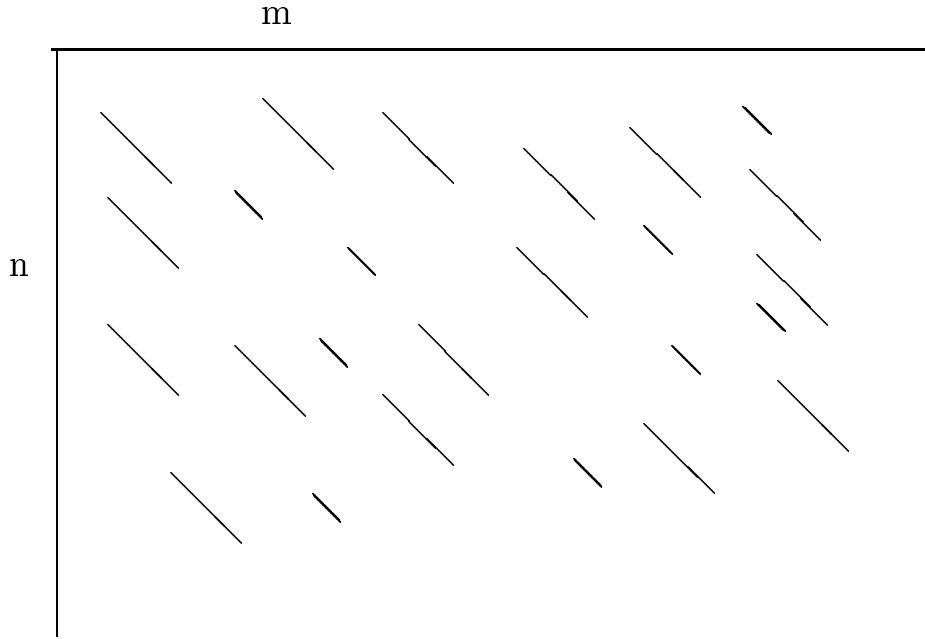


Figure 5: A Dotplot

Now write  $\theta^* = \log\left(\frac{q}{p}\right)$  The above becomes

$$p(y) \sim 1 - e^{-\theta^*} (1 - e^{-\theta^*}) e^{-\theta^* y}$$

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

next consider  $n$  such walks, reaching their maximum at  $y_1, y_2, \dots$ . The max of these will have tail probability

$$\text{pr}(M \leq y) \sim 1 - K e^{-\theta^* y}$$

with  $K = nC^*$ .

To connect this to local alignment, consider the dotplot (figure 5) generated by aligning a sequence of length  $m$  to one of length  $n$ .  $m, n$  large and  $m \ll n$ . Ignoring edge effects, there are about  $n$  diagonals each of length about  $m$ . along each diagonal there will be about  $m \frac{1}{q-p}$  ladder points. If we ignore the slight dependence across the diagonals and fluctuations in the no. of ladder points, our best local alignment with scores  $+1, -1$  having probabilities  $p, q > p$  will be as above with  $K = mn \frac{1}{q-p}$ . Perhaps quite surprisingly, this can be made rigorous quite generally.

Generally, we use a score matrix  $S = (s_{ab})$  for  $a, b$  being either the 4 bases or the 20 amino acids, and composition vectors  $(P_a), (q_b)$  (usually the same) such that

$$\sum_{a,b} p_a q_b s_{ab} < 0$$

Under this assumption, there is a unique positive solution to

$$\sum_{a,b} p_a q_b e^{t s_{ab}} = 1$$

typically denoted by  $\lambda$ . The theoretical analysis proceeds as with the simple random walk, but in calculating asymptotic expressions for probabilities of maxima of walks, we need to take into account overshooting the boundaries.

The best score along a given diagonal in a gap-free alignment corresponds to the maximum height of an excursion relative to the ladder point from which it started, before reaching the next ladder point.

In figure 2 positive scoring excursions are from the ladder points  $(2, -2), (5, -3), (8, -4), \dots$ , having sizes of  $1, 1, 4, 3, \dots$  respectively. Of course there are many zero scoring excursions.

### How many excursions?

The mean number of steps between consecutive ladder points is the mean  $m_0$  when  $b \rightarrow \infty$  and  $a = -1$  that is  $\frac{1}{q-p}$ . Thus an alignment of  $N$  base pairs or residues can be expected to have about  $\frac{N}{\frac{1}{q-p}} = N(q-p)$  ladder points, equivalently, excursions. We can put this into our formula and infer that

$$\text{pr}(\text{max excursion score} > y) \sim C e^{-\theta^* y}$$

where  $C = C^* N(q-p)$  and  $\theta^* = \log\left(\frac{q}{p}\right)$ .

Note: we are ignoring fluctuations about this mean, and BLAST also takes max across weakly dependent alignments.