Random Walks (RW)

Random Walks are used for finding high scoring local alignments and then finding P-values (BLAST).

BLAST is used to search a database consisting of a number of sequences for similarity to a single "query" sequence.

Suppose we give a score of +1 if the two nucleotides in corresponding positions are the same and a score of -1 if they are different. As we compare the two sequences, starting from the left, the accumulated score performs a random walk. In the above example, the walk can be depicted graphically as in Figure 7.1. The filled circles in this figure relate to ladder points that is to points in the walk lower than any previously reached point.

The part of the walk from a ladder point until the highest point attained before the next ladder point is called an excursion. BLAST theory focuses on the maximum heights achieved by these excursions. In Figure 7.1 these maximum heights are, respectively, 1, 1, 4, 0, 0, 0, 3. (If the walk moves from one ladder point immediately to the next, the corresponding height is taken as 0.)
In practice, BLAST theory relates to cases that are much more complicated than this simple example. It is often applied to the comparison of two protein sequences and uses scores other than the simple scores +1 and 1 for matches and mismatches. These scores are described by the entries in a substitution matrix such as those given in the BLOSUM62 substitution matrix shown in Table 6.7. These scores determine the upward or downward movement of the random walk describing the score for that protein comparison. For example, if the score for any amino acid comparison is given by the appropriate entry in the BLOSUM62 substitution matrix then for the alignment

\[
T Q L A A W C R M T C F E I E C K V \\
R H L D S W R R A S D D A R I E E G
\]

(7.2)

the scores are \(-1, 1, 5, -2, 1, 15, -1, 7, -1, 2, -4, \) etc..., and therefore the graph of the accumulated score goes through the points

\[
(1, -1), (2, 0), (3, 5), (4, 3), (5, 4), (6, 19), (7, 15), (8, 22), \text{ etc..} \quad (7.3)
\]

The graph of the accumulated score for this walk is depicted in Figure 7.2.
Simple Random Walk (RM)

- Set up with lines
  \( p \)
  \( h \)
  \( z \)
  \( \text{RM restricted on } [a, b] \)
  \( a < h < b \)
  \( \text{ie steps also reach } a \text{ or } b. \)

- Mover independently of the previous step

- Mean \# of steps until the walk stops

- \( P \left[ \text{walk ends at } b \text{ rather than } a \right] \)
  \( \text{end} = \text{absorbed}. \)

\[ \text{whz } P \left[ \text{absorption at } b \text{ rather than } a \mid \text{initial height } h \right] \]

\[ = P \left[ E \mid h \right] \]

\[ = P \left( E \text{ and 1st step up } \mid h \right) + P \left( E \& \text{and 1st move down} \mid h \right) \]

\[ = P \left( E \mid h \uparrow \right) P \left( \uparrow \right) + P \left( E \mid h \downarrow \right) P \left( \downarrow \right) \]

\[ = P \left( E \mid \text{start at lower} \right) P \left( \uparrow \right) + P \left( E \mid \text{start at higher} \right) Q \]

\[ = \omega_{h+1} p + \omega_{h-1} q \rightarrow (7.4) \]
\[ (7.6) \]

\[ w_0 = \theta \quad \text{it stays at } a \quad \text{it remains at } a \quad \rightarrow \quad 7.1. \]

\[ w_0 = 1 \quad \text{it stays at } b \quad \text{it remains at } b. \]

Using the difference equation approach on (7.4) and (7.5) as the boundary conditions a solution of the homogeneous difference equation (7.4) is obtained as: \( w_h = e^{\theta h} \)

where \( \theta \) is a fixed constant

\[ \theta = 0 \text{ or } \log \left( \frac{a}{b} \right) \quad (a \neq b) \]

\[ \rightarrow \quad w_h = 1 \quad \text{and} \quad w_h = e^{\theta_x h} \]

\[ \theta = \log \left( \frac{b}{a} \right) \]

The general solution is given by

\[ w_h = C_1 + C_2 e^{\theta_x h} \]

\[ = \frac{e^{\theta_x h} - e^{\theta x a}}{e^{\theta x b} - e^{\theta x a}} \]

\[ w_h = 2 \text{ the above calculation does not hold.} \]

In BLAST we do not encounter the \( \theta = 2 \) case.
\[ U_n = P \left\{ R \text{ exits at } a \text{ rather than } b \mid \text{ start at } h \right\} \]

\[ U_h = e^{\theta b} - e^{\theta h} \]

\[ e^{\theta h} - e^{\theta a} \]

\[ \omega + U_h = 1 \]

\[ \text{Mean \# of steps taken until } \text{walk stops} \]

\[ m_h = E \left[ \text{\# steps until walk stops} \mid \text{start at } h \right] \]

\[ = E(N) \Theta \]

\[ \Theta \left( E(N) \right) = E(N) - 1 \]

\[ = P(\theta) E(N | h=1) + P(\theta') E(N | h=0) \]

\[ z \]

\[ m_{h-1} = 1 \times m_{h-1} + 2 \times m_{h-1} \rightarrow \)

\[ m_0 = 0 \]

\[ \]
The general solution to this homogeneous equation is

\[ m = \frac{u_n (b-h) + u_n (a-h)}{b-q} \]

**MGF approach to derive** \( w_n \)

\[ S_i = \begin{cases} +1 & \text{if } i = 1, \ldots, N \\ -1 & \text{if } i = a \text{ or } b \end{cases} \]

\( S_i \) i.i.d.

\[ T_N = \sum_{i=1}^{N} S_i = \begin{cases} a-h & \text{will hold if } 1 = q \\ b-h & \text{will hold if } a = w_n \end{cases} \]

**MGF of** \( S_i \)

\[ \text{m}_{S_i}(\theta) = E[e^{\theta S_i}] = \frac{q e^{-\theta} + b e^\theta}{q + b} \]

(usually we derivate MGF \( k \).)
Th 1.1: For any r.v. \( X \) with \( E(X) \neq 0 \) and some \( a \geq 0 \), \( \exists \) a unique nonzero value \( \theta^* \) (for all \( \theta \) for which \( m(\theta) = 1 \)

**Wald's Identity**

\[
E \left( m(\theta)^N \; e^{\theta^T X} \right) = 1
\]

Using Th 1.1 in Wald's identity:

\[
E \left( m(\theta^*)^{-N} \; e^{\theta^* T X} \right) = 1
\]

\[
\Rightarrow E \left[ e^{\theta^T X} \right] = 1
\]

\[
\Rightarrow (1 - w_n) \; e^{\theta^*(a-b)} + w_n \; e^{\theta^*(b-a)} = 1
\]

\[
\Rightarrow w_n = \frac{e^{\theta^*(b-a)} - e^{\theta^*(a-b)}}{e^{\theta^*(b-a)} - e^{\theta^*(a-b)}}
\]
Taking the derivative of both sides of the
Wold's equation and exchanging the
expectation and differentiation we get

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

but \( \theta = 0 \)

\[ \Rightarrow E(CT_N) = E(S) m_k \]

\[ E(S) = (\gamma_1, g + \gamma_1) = k - g \]

\[ E(\sum_{i=1}^{N} S_i) = E \left( \sum_{i=1}^{N} \right) \]

\[ = (a - b) (1 - w) + (b - h) w \]

\[ = (a - b) U_h + (b - h) w \]

\[ \therefore m_k = \frac{(a - b) U_h + (b - h) w}{k - g} \]
Asymptotic Case

In BLAST

\[ h = 0 \]
\[ a = -1 \]
\[ b = \text{no upper boundary.} \]

\[ E(S) < 0 \]

Under BLAST, a walk is destined eventually to reach \(-1\).

(i) \( \mathbb{P}(\text{p.d.f. of } \max(\sum_{j=1}^{N} S_j)) \)

(ii) \( E(N) = \text{mean of steps before it reaches } -1 = \mu_0 \)

put an artificial stopping bound \( y \).

\[ \therefore \quad w = \frac{1 - e^{-\theta^*}}{e^{-\theta^*} - e^{-\theta^*}} \]

\[ a \rightarrow \infty \quad \text{whence } (1 - e^{-\theta^*}) e^{\theta^*} y \]

\[ y = \text{max height} \]

\[ P[Y > y] \sim (1 - e^{-\theta^*}) e^{\theta^*} y \]
Similarly \( Az_{x+z} \frac{u_0 - yw_0}{z-b} \)

\[ a \to x \quad yw_0 \to 0 \]

since \( u_0 \to 0, \quad u_0 \to 1 \)

\[ \lim_{y \to a} \lim_{y \to 0} \frac{u_0 - yw_0}{z-b} = \frac{1}{z-b} \]

**General Walks**

Possible step sizes \(-c, -c+1, \ldots, 0, \ldots, d-1, d\).

prob \( k_c, \ldots, 1, k_d, \ldots \)

Assume

(i) \( b-c > 0 \) and \( bd > 0 \)

(ii) \( E(S) < 0 \)

(iii) \( \gcd \) of the step sizes that have non-zero probability is 1.

\[ P_y = P(Y \geq y) \sim C e^{-y\Theta^*} \]

**Y** max height obtained in the walk \( \Theta^* = \log \left( \frac{2 + \sqrt{4b^2 + 2g}}{2b} \right) \)

\[ R_{-1} = \lim_{n \to \infty} P_{-1}, \quad R_{-2} = \lim_{n \to \infty} P_{-2} \]
Unrestricted Walks

No boundaries are placed on the walk.

\[ V(y) = \left[1 - F_{\text{unr}}(y)\right] e^{y \theta^*} \]

\( F_{\text{unr}}(y) \) = prob that in the unrestricted
the max upward excursion is \( y \) or less for any \( y > 0 \).

\( V_2 = \lim_{y \to +\infty} V(y) \)

\[ = \frac{\bar{Q}}{(e^{\theta^*} - 1) \left( \sum_{k=1}^{d} k \bar{Q}_k e^{k \theta^*} \right)} \]

\( \bar{Q} = 1 - \sum_{i=1}^{d} Q_i \) = Prob max excursion never reaches the values.

\( \bar{Q}_k \) = Prob that the walk visits the tre value \( k \) before reaching any tre value.