1. Types of Alignments

There are many types of alignments:
- Global alignments – the entire lengths of the sequences are aligned
- Local alignments – only subsequences of each sequence are aligned
- Gapped alignments – indels (insertions/deletions) are allowed in the alignments
- Ungapped alignments – indels are not allowed
- Pairwise alignments – two sequences are aligned
- Multiple alignments – more than two sequences are aligned

Note that for a particular type of pairwise alignment, there are many possible such alignments between any two sequences. Good alignments of related sequences are ones that better reflect the evolutionary relationship between them.

In alignments, scores are assigned to each alignment in a meaningful way that reflects the likelihood that this alignment was produced as a consequence of divergence from a common ancestor. As we have known, scores of alignments consist of two main types: similarity scores and distances scores. In similarity scores the higher the score, the more closely related are the two aligned sequences; in the distance measures the opposite is the case.

**Similarity scores**

We denote by \( w(X,Y) \) the score assigned to the aligned pair consisting of the residues \( X \) and \( Y \). The score assigned to a gap (including insertions and deletions) of length \( l \) is usually a function of \( l \), which we denote by \( \delta(l) \). It represents the cost of having a gap of length \( l \) and is therefore zero or negative. The simplest gap penalty model is a linear gap penalty model \( \delta(l) = -ld \) for some nonnegative constant \( d \).

Thus if the alphabet has size \( N \) (\( N = 4 \) for nucleotides and \( N = 20 \) for amino acids), a scoring scheme consists of an \( N \times N \) matrix \( W \) and a gap cost function \( \delta \). The matrix \( W \) is called a substitution matrix and the entry in its \( i \)th row and \( j \)th column is the score of the alignment of the \( i \)th and \( j \)th symbols in the alphabet.

2. BLAST theory

BLAST is the most frequently used method for assessing which DNA or protein sequences in a large database have significant similarity to a given query sequence. Here we describe how the random walk theories are used for this purpose. Various sophisticated versions of BLAST are now available and used widely. Our aim here is to
give an indicator of the nature of BLAST theory by considering a simple version of BLAST, leading to a readily understood statistical analysis.

We start by considering as given an ungapped alignment of two DNA sequences in the simplest setting: we give a score +1 if the two nucleotides in corresponding positions are the same and a score of -1 if they are different. The null hypothesis to be tested is that for each aligned pair of nucleotides, the two nucleotides were generated by independence mechanisms, so that if nucleotide \( j \) occurs at any given position in the first sequence with probability \( p_j \) and nucleotide \( k \) occurs at any given position in the second sequence with probability \( p_k \), the probability that they occur together in a given aligned pair is 

\[
\text{null hypothesis probability of the pair } (j, k) = p_j p_k.
\]

Then the probability of observing a match pair is 

\[
p = \sum_{j=A,C,G,T} p_j p_k,
\]

and the probability of observing a mismatch pair is 

\[
q = 1 - p.
\]

It is required in the theory that \( q > p \).

As we compare the two sequences, starting from the left, the accumulated score performs a random walk.

**Random Walk and Sequence Alignments**

Let us consider the simple case of the two aligned DNA sequences:

```
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
G A A C G C C T A G C C A C G A G C C T T A T C
```

Suppose we give a score +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 1. The filled circles in this figure relate to *ladder points*, the points in the walk lower than any previously reached point (new low).

![Figure 1](image)

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 1 above, 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. So to discuss BLAST it is necessary to consider arbitrary scoring schemes and thus aspects of the general theory of random walks. However, we start by analyzing the simple random walk where the only possible step sizes are +1 and -1.

**Simple Random Walk and Associated Probability Theories**

We first consider a process that starts at some arbitrary point \( h \) and moves, independently of the previous history of the process, a step down every unit with probability \( q \) or a step up with probability \( p \) (\( p+q=1 \)). (We use the notation “\( h \)” throughout to denote the initial position of the walk.) This process is called a simple random walk. The walk is assumed to be restricted to the interval \([a, b]\), where \( a \) and \( b \) are integers with \( a < h < b \), and stops when it reaches either \( a \) or \( b \).

Two questions are interesting to us:
1. What is the probability that eventually the walk finishes at \( b \) rather than \( a \)?
2. What is the mean number of steps taken until the walk stops?

**Absorbing probabilities**

Let \( w_h \) be the probability that the simple random walk eventually finishes at, or is “absorbed” at, the point \( b \) rather than at the point \( a \), given that the initial point is \( h \). Then comparing the situation just before and just after the first step of the walk, we get

\[
    w_h = pw_{h+1} + qw_{h-1}. \tag{1}
\]

Furthermore,

\[
    w_a = 0, \quad w_b = 1. \tag{2}
\]

Equation (1) is a homogeneous difference equation, with boundary conditions (2). We can solve the equation (1) with condition (2) using general tricks in difference equations:

\[
    w_h = pw_{h+1} + qw_{h-1} \quad \Rightarrow \quad (p + q)w_h = pw_{h+1} + qw_{h-1} \\
    \quad \Rightarrow \quad p(w_{h+1} - w_h) = q(w_h - w_{h-1}) \\
    \quad \Rightarrow \quad \frac{w_{h+1} - w_h}{w_h - w_{h-1}} = \frac{q}{p}. \tag{3}
\]

(3) is true for any \( h \) satisfying \( a < h < b \). So for any \( h \) with \( a < h < b \) we have
\[
\begin{align*}
\frac{w_{h+1} - w_h}{w_{a+1} - w_a} &= \left(\frac{q}{p}\right)^{h-a} \\
\Rightarrow w_{h+1} - w_h &= \left(\frac{q}{p}\right)^{h-a} w_{a+1} - w_a \\
\Rightarrow w_{h+1} - w_h &= \left(\frac{q}{p}\right)^{h-a} (w_{a+1} - w_a) \\
\Rightarrow w_{h+1} - w_h &= \left(\frac{q}{p}\right)^{h-a}. \\
\end{align*}
\]

Then
\[
\begin{align*}
(w_h - w_{h-1}) + \ldots + (w_{a+2} - w_{a+1}) + (w_{a+1} - w_a) &= w_{a+1} \left(\frac{q}{p}\right)^{h-a} + \ldots + (\frac{q}{p}+1) \\
&= w_{a+1} \left(\frac{q}{p}\right)^{h-a} \frac{1-\left(\frac{q}{p}\right)^{h-a}}{\frac{p}{1-\left(\frac{q}{p}\right)}} \quad \text{when } p \neq q \quad (5) \\
&= w_{a+1}(h-a) \quad \text{when } p = q \quad (6) \\
\end{align*}
\]

and
\[
\begin{align*}
(w_b - w_{b-1}) + (w_{b-1} - w_{b-2}) + \ldots + (w_{h+1} - w_h) + \ldots + (w_{a+2} - w_{a+1}) + (w_{a+1} - w_a) &= w_{a+1} \left(\frac{q}{p}\right)^{b-a-1} + \ldots + \left(\frac{q}{p}+1\right) \\
&= w_{a+1} \left(\frac{q}{p}\right)^{b-a} \frac{1-\left(\frac{q}{p}\right)^{b-a}}{\frac{p}{1-\left(\frac{q}{p}\right)}} \quad \text{when } p \neq q \quad (7) \\
&= w_{a+1}(b-a) \quad \text{when } p = q \quad (8) \\
\end{align*}
\]

Furthermore, from (5) and (6), we have
\[
\begin{align*}
w_h &= w_{a+1} \left(\frac{q}{p}\right)^{h-a} \quad \text{when } p \neq q, \quad \text{and } w_h = w_{a+1}(h-a) \quad \text{when } p = q, \quad (9) \\
\end{align*}
\]

and from (7) and (8), we have
\[
\begin{align*}
(w_b - w_a) &= w_{a+1} \left(\frac{q}{p}\right)^{b-a} \quad \text{when } p \neq q, \quad \text{and } (w_b - w_a) = w_{a+1}(b-a) \quad \text{when } p = q \\
\end{align*}
\]
Above all, combining (9) and (10), we have

\[
w_h = \frac{1 - \left(\frac{q}{p}\right)^{b-a}}{1 - \left(\frac{q}{p}\right)^{b-a}} \quad \text{when } p \neq q, \quad \text{and } w_h = \frac{1}{b-a} \quad \text{when } p = q.
\]

In some applications interest focuses on the probabilities \( u_h \) that the work finishes at \( a \) rather than at \( b \). The probabilities \( u_h \) satisfy the same difference equation (1) as \( w_h \), but with the boundary conditions \( u_a = 1, \ u_b = 0 \). This leads to

\[
u_h = \frac{\left(\frac{q}{p}\right)^b - \left(\frac{q}{p}\right)^a}{\left(\frac{q}{p}\right)^b - \left(\frac{q}{p}\right)^a} \quad \text{when } p \neq q, \quad \text{and } u_h = \frac{b-h}{b-a} \quad \text{when } p = q.
\]

**Mean number of steps taken until the walk stops**

We continue to assume that the walk starts at \( h \) and ends when one of \( a \) or \( b \) is reached, \( a < h < b \). We define \( N \) as the (random) number of steps taken until the walk stops. The probability distribution of \( N \) depends on \( h, a, \) and \( b \). We are interested in the mean value \( m_h \) of this probability distribution, that is, the expectation of \( N \). We will derive \( m_h \) without computing the complete probability distribution of \( N \). Instead, we use difference equation methods similar to those in above section, which lead directly to the value of the mean.

Considering the situation after one step of the walk in two different ways, we get

\[
m_h - 1 = pm_{h+1} + qm_{h-1}
\]

with \( m_a = m_b = 0 \). (11) is an inhomogeneous difference equation. The tricks used above can be applied to solve (11). The details will not be shown here.

The solution of (11) is

\[
m_h = \frac{w_h(b-h) + u_h(a-h)}{p-q}.
\]

**Distribution of the maximum height of an excursion \( (Y_i) \)**

Let \( Y_1, Y_2, \ldots, Y_m \) be the respective maximum heights of the excursions of this walk after leaving one ladder point and before arriving at the next, and let \( Y_{\text{max}} \) be the maximum of
these maxima. \( Y_{\text{max}} \) is in effect the test statistic used in BLAST, so it is necessary to find its null hypothesis distribution.

By previous notes, we know that for a simple random walk starting at \( h \) and stopping at \( a \) or \( b \) with \( a < h < b \), the probability that the walk finishes at \( b \) rather than \( a \) is

\[
 w_h = \frac{1 - (q/p)^{b-a}}{1 - (q/p)^{b-a}} = \frac{(q/p)^b - (q/p)^a}{(q/p)^b - (q/p)^a}
\]

(1)

And the mean number of steps taken until walk stops is

\[
m_h = \frac{w_h(h-h) + u_h(a-h)}{p-q}
\]

(2)

The theory of BLAST concerns those walks that start at \( h = 0 \), where there is a lower boundary at \( a = -1 \) but no upper boundary, and where the mean step size \( E(S) = p - q < 0 \). Such a walk is destined eventually to reach -1. BLAST theory requires the calculation of two quantities:

1. The probability distribution of the maximum value \( \left(Y_i\right) \) that the walk ever achieves before reaching -1, and
2. The mean number of steps until the walk eventually reaches -1.

To discuss the maximum value that the walk ever reaches before eventually reaching -1, we install an artificial stopping boundary at the value \( y \), where \( y \geq 1 \). Then if \( h = 0 \) and \( a = -1 \), equation (1) shows that the probability that the unrestricted walk finishes at the artificial boundary \( y \) rather than at the value \( a = -1 \) is

\[
 \frac{(q/p)^y - (q/p)^0}{(q/p)^y - (q/p)^{b-a}} = \frac{1 - (q/p)^{-1}}{1 - (q/p)^{-1}} = 1 - \frac{1}{e^{-\lambda}} 
\]

(3)

where \( \lambda = \ln(q/p) \). Since \( q > p \), \( \lambda \) is positive and then the term \( e^{ky} \) dominates the denominator in (3) when \( y \) is large. So the probability (3) is asymptotic to \( (1 - e^{-\lambda})e^{-ky} \). Thus if \( Y \) is the maximum height achieved by the walk,

\[
\Pr(Y \geq y) \sim (1 - e^{-\lambda})e^{-ky}.
\]

(4)

Note that (3) gives the probability that the walk ever achieves \( y \), which is equivalent to the probability that the maximum height achieved by the walk is \( \geq y \).

Turning to question 2, the expression for the mean number \( A \) of steps before that walk finishes at -1 or \( b \) is found by first making the substitution \( h = 0 \) and \( a = -1 \) in equation (2), to obtain
\[ A = \frac{w_0(b-0) + u_0(-1-0)}{p-q} = \frac{u_0 - bw_0}{q-p}. \]  \hspace{1cm} (5)

Further since
\[ \lim_{b \to \infty} bw_0 = \lim_{b \to \infty} b \frac{1 - e^{-\lambda}}{e^{db} - e^{-\lambda}} = \lim_{b \to \infty} b \frac{1 - e^{-\lambda}}{e^{db} - e^{-\lambda}} = 0 \]
and
\[ \lim_{b \to \infty} u_0 = \lim_{b \to \infty} \frac{e^{db} - 1}{e^{db} - e^{-\lambda}} = 1, \]
taking the limit \( b \to \infty \) in equation (5) shows that \( A = 1/(q-p) \).

**Distribution of \( Y_{\text{max}} \)**

Ignoring end effects, \( Y_1, Y_2, ..., Y_m \) can be taken as independently and identically distributed. Then
\[
\Pr(Y_{\text{max}} \geq y) = 1 - \Pr(Y_{\text{max}} < y) = 1 - \Pr(Y_1 < y, Y_2 < y, ..., Y_m < y) = 1 - \prod_{i=1}^{m} \Pr(Y_i < y) \approx 1 - (1 - (1 - e^{-\lambda})e^{-\lambda}y)^m.
\]

**BLAST theory in general scoring schemes**

In the alignment of protein sequences, more complex scoring schemes are often used. Below we give a brief discussion on the distribution of the maximum height (\( Y_i \)) the walk achieves under a general scoring scheme in an excursion.

We are given an ungapped global alignment of two protein sequences, both of length \( N \). We number the positions in the alignment from left to right as positions 1, 2, ..., \( N \). A score \( S(j,k) \) is allocated to each position where the aligned amino acid pair \((j, k)\) is observed, where \( S(j,k) \) is the \((j, k)\) element in the substitution matrix chosen. It is required in the theory, and is assumed throughout, that at least one element in the substitution matrix be positive and that the null hypothesis mean score \( \mu(j, k) \) be negative. An accumulated score at position \( i \) is calculated as the sum of the scores for the various amino acid comparisons at positions 1, 2, ..., \( i \). As \( i \) increases, this accumulated score undergoes a general random walk (generally possible step sizes in a random walk are \(-c, -c+1, ..., 0, ..., d-1, d\)).
Here we give the distribution of $Y_i$ directly and skip the details on how to derive it in the generalized case:

$$P(Y_i > y) \sim C e^{-\lambda y},$$

where $\lambda$ can be determined by the equation $\sum_{j,k} p_j p_k e^{\lambda (j,k)} = 1$. The calculation of $C$ is much more involved. We have known the value of $C$ in one special case: $C = 1 - e^{-\lambda}$ in a simple random walk.