
A Brief Tutorial on BLAST

Dirk Husmeier

Biomathematics and Statistics Scotland
at the Scottish Crop Research Institute
Invergowrie, Dundee DD2 5DA, UK

Email: dirk@bioss.ac.uk

<http://www.bioss.ac.uk/~dirk>

- Motivation
- Theory
- Heuristic amendments

Introduction

- **BLAST:** Procedure that (1) searches for high-scoring local alignments between two sequences and then (2) tests for significance of the scores found via P-values.
- **Main application:** Search a database consisting of many sequences for similarity to a single 'query' sequence.
- **Search algorithm:** Uses heuristics to avoid searching through all possible ungapped alignments.
- **P-value calculation:**
 1. Takes into consideration the lengths of the sequences: The longer the sequences, the more likely there is to be a local homology simply by chance.
 2. Allows for the size of the entire database (to correct for multiple testing).
 3. Uses sophisticated approximations to achieve extremely fast calculations.

Preliminaries: Moment generating function

1. Let Y be a random variable with probability distribution $P(Y)$.

2. Moment generating function (mgf):

$$m(t) = \sum_y P(y)e^{ty} = \langle e^{ty} \rangle$$

3. k th moment

$$\langle Y^k \rangle = \lim_{t \rightarrow 0} \frac{d^k}{dt^k} m(t)$$

4. For N iid random variables Y_1, \dots, Y_N :

$$m_{Y_1, \dots, Y_N}(t) = m_Y(t)^N$$

5. Theorem for a discrete random variable Y with mgf $m(t)$. If (1) Y can take at least one negative value and (2) at least one positive value and (3) has nonzero mean: $\langle Y \rangle \neq 0$.

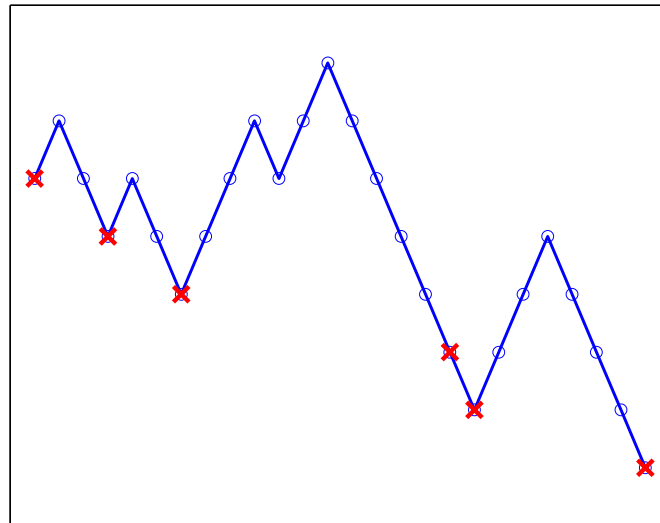
\implies There exists a unique value $\lambda \neq 0$ such that $m(\lambda) = 1$.

Score function and random walk

DNA sequence alignment:

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

Score function $S(i, k)$, example: $S(i, k) = 2\delta_{ik} - 1 = \begin{cases} +1 & \text{if } i = k \\ -1 & \text{if } i \neq k \end{cases}$



Ladder points: Points in the walk lower than any previously reached point.

Excursions

- **Excursion:** Section of the random walk between two consecutive ladder points.
- **Statistic of interest Y_i :** Maximum height of the i th excursion after leaving the i th ladder point and before arriving at the $(i + 1)$ th ladder point.
- **Statistic of interest $Y_{max} = \max\{Y_1, \dots, Y_M\}$**
- **Null hypothesis:**
 1. $P(i, k) = p_i p'_k$
 2. The walk has a negative drift: $\langle S \rangle := \sum_{ik} p_i p'_k S(i, k) < 0$

Example:

$$p := \sum_i p_i p'_i < 0.5, \quad S(i, k) = \begin{cases} +1 & \text{if } i = k \\ -1 & \text{if } i \neq k \end{cases}$$

$$\sum_{ik} p_i p'_k S(i, k) = \sum_i p_i p'_i - \sum_{i \neq k} p_i p'_k = \sum_i p_i p'_i - \left(1 - \sum_i p_i p'_i\right) = 2p - 1 < 0$$

Distribution of Y under the null hypothesis

Start a random walk at a ladder point L_{i-1} . The random walk will terminate at ladder point $L_i < L_{i-1}$, that is, with a final vertical displacement $D_i = L_i - L_{i-1}$. The maximum upwards excursion of this walk is Y_i .

- Single-step mgf: $m_S(t) = \sum_{ik} p_i p'_k e^{tS(i,k)}$; $m_S(\lambda) = 1 \rightarrow \lambda$
 - Multi-step mgf: $m_{S_1, \dots, S_N}(t) = m_S(t)^N \implies m_{S_1, \dots, S_N}(\lambda) = 1$
 - Final-displacement mgf: $m_D(t) = \sum_D P(D) e^{Dt}$
 - For $N \rightarrow \infty$, the walk will eventually terminate. This implies that $m_{S_1, \dots, S_N}(t) \rightarrow m_D(t)$ and, consequently, that $m_D(\lambda) = 1$.
 - Distribution of interest: $P(Y \geq y)$, the probability that the maximum upwards excursion is greater or equal y .
 - From $m_D(\lambda) = 1 \implies$ For $y \rightarrow \infty$, $P(Y \geq y) = C e^{-\lambda y}$.
 - This means that under the null hypothesis, the maximum height Y has (asymptotically) a geometric-like distribution.
-

Example

-
- Two step sizes: $S \in \{1, -1\}$, $P(S = 1) = p < 0.5$
 - Single-step mgf: $m_S(t) = pe^t + (1 - p)e^{-t}$
 - $m_S(\lambda) = 1 \implies pe^{2\lambda} - e^\lambda + (1 - p) = 0 \implies e^\lambda \in \left\{1, \frac{p}{1 - p}\right\}$

Unique nonzero solution: $\lambda = \ln\left(\frac{p}{1 - p}\right)$

- Start a random walk at a ladder point. The walk will stop at the next ladder point, which has the displacement $D = -1$. This is the $y \rightarrow \infty$ limit of a random walk that stops either at $D = -1$, with probability $P_{(D=-1)}$, or at $D = y$, with probability $P_{(D=y)}$. Note: $P_{(D=y)} + P_{(D=-1)} = 1$ and $P_{(D=y)} = P_{(Y \geq y)}$
- $m_D(t) = P_{(D=-1)}e^{-t} + P_{(D=y)}e^{yt} = e^{-t} + P_{(D=y)}(e^{yt} - e^{-t})$
- $m_D(\lambda) = 1 \implies e^{-\lambda} + P_{(Y \geq y)}(e^{y\lambda} - e^{-\lambda}) = 1$
- $\implies P_{(Y \geq y)} = \frac{1 - e^{-\lambda}}{e^{y\lambda} - e^{-\lambda}}$. For $y \rightarrow \infty$: $P_{(Y \geq y)} = [1 - e^{-\lambda}]e^{-\lambda y}$.
- Null hypothesis, $y \rightarrow \infty$: Geometric-like distribution: $P_{(Y \geq y)} = Ce^{\lambda y}$, with $C = 1 - e^{-\lambda}$

Distribution of Y_{max} under the null hypothesis

- Number of excursions: n
 - Y_1, \dots, Y_n are iid random variables, $P_{(Y \geq y)} \approx Ce^{-\lambda y}$
 - $Y_{max} = \max\{Y_1, \dots, Y_n\}$
 - Distribution of interest: $P_{(Y_{max} \geq y)}$
 - Extreme value distribution:
 - $P_{(Y_{max} \leq y)} = P_{(Y_1 \leq y \wedge Y_2 \leq y \wedge \dots \wedge Y_n \leq y)} = \prod_{i=1}^n P_{(Y_i \leq y)} = [P_{(Y \leq y)}]^n$
 - $P_{(Y_{max} \geq y)} = 1 - P_{(Y_{max} \leq y-1)} = 1 - [P_{(Y \leq y-1)}]^n = 1 - [1 - P_{(Y \geq y)}]^n$
 - From $P_{(Y \geq y)} \approx Ce^{\lambda y} \implies P_{(Y_{max} \geq y)} = 1 - [1 - Ce^{\lambda y}]^n$
 - Expected number of excursions with $Y_{max} \geq y$: $E = nP_{(Y_{max} \geq y)} \approx nCe^{\lambda y}$
 - Approximation to the extreme value distribution (without derivation):
 $P_{(Y_{max} \geq y)} \approx 1 - e^{-E}$
-

Average number of high-scoring excursions E

- Average number of high-scoring excursions with $Y \geq y$: $E = nCe^{\lambda y}$
- We need n , the average total number of excursions .
- Average step size : $\langle S \rangle = \sum_{ik} p_i p'_k S(i, k)$
- Average final displacement : $\langle D \rangle = \sum_D P_D D$
- Average length of an excursion : $A = \langle D \rangle / \langle S \rangle$
- Average number of excursions : $n = N/A$, where N is the length of the alignment.

Example: Two step sizes, $S \in \{1, -1\}$, $P(S = 1) = p < 0.5$

- Average step size: $\langle S \rangle = (1)p + (-1)(1 - p) = 2p - 1$
 - Average final displacement: $P_{(D=-1)} = 1 \implies \langle D \rangle = -1$
 - Average length of an excursion: $A = \frac{\langle D \rangle}{\langle S \rangle} = \frac{(-1)}{2p - 1} = \frac{1}{1 - 2p}$
 - Average number of excursions: $n = \frac{N}{A} = (1 - 2p)N$
-

Summary (so far)

- Start with some score function $S(i, k)$ (to be discussed shortly).
 - This defines a random walk from the DNA sequence alignment.
 - Find the ladder points, which segments the walk into excursions.
 - For each excursion: Find the maximum height, Y .
 - Compute the extreme value distribution $P(Y_{max} \geq y)$ under the null hypothesis, that is, the P-value. This depends on λ , C , and n .
 1. Obtain λ from the single-step mgf: $m_S(\lambda) = 1$.
 2. Get C from the final-displacement mgf: $m_D = 1$.
 3. Estimate the average number of excursions $n = \langle D \rangle / \langle S \rangle$ from the average displacement $\langle D \rangle$ and the average step size $\langle S \rangle$.
 4. Compute the average number of high-scoring excursions $E(y) = nCe^{-\lambda y}$
 5. From this, get $P(Y_{max} \geq y) = 1 - e^{-E(y)}$
 - Identify excursions with small P-values, e.g. $P(Y_{max} \geq y) \leq 0.01$. The matching sequence pair is the part of the alignment between the corresponding ladder point and the site where the maximum upward excursion from the ladder point is reached.
-

Edge effects and multiple testing

- Comparison of two unaligned sequences:

- Given two sequences of lengths N_1 and N_2 without a specific alignment. Goal: Find the significance of high-scoring segment pairs between all possible local alignments .
- Approximation: Apply the previous theory with the replacement: $N \rightarrow N_1 N_2$

- Database search: Have a query sequence , search an entire database of many sequences for those with significant similarity to the query sequence.

→ Two problems:

- Edge effects

- Previous derivation: asymptotic result , valid for infinite sequence length: $N \rightarrow \infty$.
- A high-scoring random walk excursion induced by the comparison of two sequences might be cut short at the end of a sequence match. Consequence: The height of high-scoring excursions, and the number of such excursions, tend to be less than predicted by the asymptotic theory.

- Multiple testing

Heuristic amendments to P-value calculations

Several **heuristics** are introduced when **correcting the P-values for edge effects and multiple testing**. Quotes from the textbook 'Statistical Methods in Bioinformatics', by W.J.Ewens and G.R. Grant:

- BLAST calculations allow for edge effects, and do this by [...] **The justification for this is largely empirical.** (→ p.283)
 - While this formula is used in BLAST, there appears to be **no publication justifying its validity.** (→ p.285)
 - Due to multiple testing, an amendment to formal P-value calculations is necessary. Unfortunately, there is **no rigorous theory** available to deal with this issue. (→ p.286, top)
 - Some details of these amendments appear not to be mentioned in BLAST documentation, and **only become clear by careful reading of the code.** (→ p.286, middle)
 - It might be a matter of concern that various somewhat arbitrary constants enter the above calculations. This concern is reinforced by the fact that [...] the calculated **P-values are quite sensitive to the somewhat arbitrary numerical values of these constants.** (→ p.291, middle)
-

The score function for proteins

- So far: $S(i, k) = 2\delta_{ik} - 1 = \begin{cases} +1 & \text{if } i = k \\ -1 & \text{if } i \neq k \end{cases}$
- Better choice: **PAM matrices** $S(i, k) = \log \frac{q_\tau(i, k)}{p_i p_k}$, where $q_\tau(i, k)$ is the probability of observing the amino acid pair (i, k) . Motivation: **likelihood ratio test**.
- τ represents evolutionary time. The larger τ , the larger the **evolutionary distance** between the sequences.

$$\begin{aligned} \tau \rightarrow 0 &\implies q_\tau(i, k) \rightarrow \delta_{ik} \\ \tau \rightarrow \infty &\implies q_\tau(i, k) \rightarrow p_i p_k \end{aligned}$$

- Two **sequences are related**, and we know the correct value of $\tau \longrightarrow$ **Average steps size $\langle S \rangle$ positive**. $\langle S \rangle =$ **mutual information** between the sites.

$$\langle S \rangle = \sum_i \sum_k q_\tau(i, k) S(i, k) = \sum_i \sum_k q_\tau(i, k) \log \frac{q_\tau(i, k)}{p_i p_k} > 0$$

- Two **sequences are unrelated** \longrightarrow **Average steps size $\langle S \rangle$ negative**:

$$\langle S \rangle = \sum_i \sum_k p_i p_k S(i, k) = \sum_i \sum_k p_i p_k \log \frac{q_\tau(i, k)}{p_i p_k} = - \sum_i \sum_k p_i p_k \log \frac{p_i p_k}{q_\tau(i, k)} < 0$$

Problem with the score function

- PAM score matrices obtained from a database of aligned protein sequences. **Statistics and degree of divergence might be different** from (1) the query sequence and (2) the database of interest.
- Assume two sequences are related, but we use the **wrong evolutionary time τ'** rather than correct value τ .

$$\langle S \rangle = \sum_i \sum_k q_\tau(i, k) S_{\tau'}(i, k) = \sum_i \sum_k q_\tau(i, k) \log \frac{q_{\tau'}(i, k)}{p_i p_k}$$

- If $\tau' \gg \tau \rightarrow q_{\tau'}(i, k)$ is more similar to $p_i p_k$ than to $q_\tau(i, k)$:

$$\langle S \rangle \rightarrow \sum_i \sum_k q_\tau(i, k) \log \frac{p_i p_k}{p_i p_k} = 0$$

- If $\tau' \ll \tau \rightarrow q_\tau(i, k)$ is more similar to $p_i p_k$ than to $q_{\tau'}(i, k)$:

$$\langle S \rangle \rightarrow \sum_i \sum_k p_i p_k \log \frac{q_{\tau'}(i, k)}{p_i p_k} = - \sum_i \sum_k p_i p_k \log \frac{p_i p_k}{q_{\tau'}(i, k)} < 0$$

- Consequence: We don't get any positive hits \rightarrow **large type II error**.
-