BLAST

Basic Local Alignment Search Tool

Used to find nucleic acid or protein sequences in a large database that has significant similarity to a given query sequence.

2 Types of BLAST

NCBI
Wash Univ
The Comparison of 2 Aligned Sequences

Consider ungapped alignment of 2 segs of length $N$

Seq 1: TQLAAW

Seq 2: RHLDSW

$p_j = P(AA_j \text{ occurs in 1st seq})$

$p'_k = P(AA_k \text{ occurs in 2nd seq})$

$H_0: P(J, k \text{ occur as aligned pairs}) = p_j p'_k$

$H_a: P(J, k \text{ occur as ordered pairs}) = \gamma(J, k)$

The BLAST RM

<table>
<thead>
<tr>
<th>Position</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Seq 1</td>
<td>T</td>
<td>Q</td>
<td>L</td>
<td>A</td>
<td>A</td>
<td>W</td>
</tr>
<tr>
<td>Seq 2</td>
<td>R</td>
<td>H</td>
<td>L</td>
<td>D</td>
<td>S</td>
<td>W</td>
</tr>
</tbody>
</table>

$S_{(i,j)} = -1 1 5 -2 1 11 < \text{step size}$

$S_{(i,j,k)}\ \text{BRASSO}(2, 1) = 238$

$\sum S_{(i,j,k)} = -1 0 5 3 4 15 \rightarrow \text{random walk}$
Under $H_0$, $E_{H_0}(S(i,k))$

(mean score) = $\sum_{j,k} S(i,k)j_k' < 0$  \hspace{1cm} (negative drift)

And least one $S(i,k) > 0$

$C_r C_D$ of $S(i,k) = 1$

$Y_i, Y_2, \ldots$ max height attained.

If excursion,

$Y_{\text{max}} = \text{max}(Y_i, Y_2, \ldots)$

\[ \theta \underset{\text{MLE}}{\rightarrow} \frac{1}{\theta^*} \]

\[ \text{from } 7.34 \sum_{j=-c}^{7} x_j e^{ij\theta^*} = 1 \text{ for a unique } +ve \theta \]

Under $H_0$:

$\sum_{j,k} b_j b_k' \theta^* S(i,k) = 1$

$\Rightarrow \sum_{j,k} b_j b_k' \theta^* S(i,k) = 1$

Use numerical method to find value of $\lambda$.
Test statistic $Y_{\text{max}} = \max \{ Y_j \}$ discard

$$F_{Y_{\text{max}}}(y) = 1 - \left[ F_Y(y-d) \right]^n \rightarrow (\Delta)$$

To calculate (\Delta), we need $C$, $\theta^\circ$ ($\in \lambda$), and the mean $\# \text{ of ladder points in the BLAST Walk}$ $(n)$.

$$n = \frac{N}{A_0} = \frac{\text{Length of saga}}{E(\text{dist between ladderph})}$$

$$A_0 = \frac{1}{2} \sum_{j=1}^{r} j R_{-j}$$

$$\sum_{j=1}^{d} \sum_{k=j}^{z-c} j R_{-j}$$

$R_{-j}$ is the probability that the walk finishes at $-j$.

$$S = \begin{cases} +1 & b \\ -2 & 1-b \end{cases}$$

Ladder point reached in the is at a dist $1$ or $2$ below the inner one.
\[ R_{-1} = P \quad (\text{dark is 1 below the dark}) \]
\[ R_{-2} = PC \quad (\text{dark is 2 below the dark}) \]
\[ R_{-2} = P(\text{walk goes to } -2) + PC \quad \text{goes to } n+1, \text{and} \]
\[ \text{then } 0, \text{the } -2 \]
\[ = q + p(1-R_{-2})R_{-2} \]
\[ \Rightarrow R_{-2} = q + \frac{4q^2 + q^2}{2p} \]

**Choice of Score**

Calculated: \[ S(i, k) \rightarrow \text{PAM} \rightarrow \text{BLOSUM} \]

**Score Statistic**

\[ S_{1,0}(y) = \log \frac{P(y_j | \xi_0)}{P(y_j | \xi_1)} \quad \text{(sequential analysis)} \]

\[ S(i, k) = \log \frac{q(i, k)}{b_j b_k} \]

\[ S(i, k) \propto \log \frac{q(i, k)}{b_j b_k} \quad \text{(will also do)} \]

\[ = x^{-1} \log \frac{q(i, k)}{b_j b_k} \]
\[ q_{j,k} = b_j b_k' e^{\lambda s(j,k)} \]

\[ \sum_{j,k} q(j,k) = 1 \]

**Under \( H_0 \):** The freq of \((j,k)\) arising in high scoring excursions for score like (10.7) \( \rightarrow q(j,k) \)

\[ I^{-1} \log \frac{q(j,k)}{b_j b_k'} \]

A support given by \( obs(c_{j,k}) \)

in favour of \( H_A \) over \( H_0 \)

**Under \( H_A \):** The mean support is given as

\[ I_t = \sum_{j,k} q(j,k) \log \frac{q(j,k)}{b_j b_k'} \]

\[ = \sum_{j,k} q(j,k) \lambda s(j,k) \]

\[ = \lambda E(s(j,k)) \]

\[ E_{H_1}(s(j,k)) \rightarrow \chi^2 \quad (\text{Used in BLAST}) \]

(inc a high scoring segment)

Convergence is slow
Bounds and Approximation for the BLAST p-value

Bounds for the P-value in BLAST:

\[ 1 - e^{-KN} \leq P(Y_{max} > Y_{max}) \leq 1 - e^{-KN e^{-\lambda}} \]

\[ K = \frac{C}{A} e^{-\lambda} \]

\[ C = 1 + R_1 e^{-\lambda} + R_2 e^{-2\lambda} \]

The P-value used in BLAST implementation uses the lower bound of (10.22).

Calculation of \( \lambda, K, C, A \)

(i) \( \lambda \) use (10.3)

(ii) For \( K \) calculate \( C \) & \( A \) then use them.

Particular case

\[ \max \{ SC(i,j) \} + 1 \Rightarrow K = (e^{x} - e^{2x}) E(Se_{i,j}) \]

\[ \min \{ SC(i,j) \} - 1 \Rightarrow K = (e^{x} - e^{2x}) \frac{[E(S)]^2}{E(Se_{i,j})} \]
The Normalized and Bit Scores

Normalized score \( S' = \lambda \text{max} - \log(NK) \)

Using (10.20), replacing \( S \)

Substituting \( S' \) in (10.20) we get

\[ e^{-e^{-s}} \leq P(S' \leq s) \leq e^{-e^{-s}} \]

\[ \Rightarrow \text{P-value} \quad P(S' > s) = 1 - e^{-e^{-s}} \]

for an observed \( s' = \lambda \text{max} - \log(NK) \)

\[ \text{P-value} \approx 1 - e^{-e^{-s'}} \quad \text{(centered to lower bound / 10.20)} \]

\[ \approx e^{-s'} \quad \text{if } s' \text{ is large} \]

\[ E_{H_0}(Y_{\text{max}}) \approx \chi^2(\log(KN) + V) \approx X^2(\log(KN)) \]

since \( V \ll KN \)

\[ V_{H_0}(Y_{\text{max}}) \approx \frac{\chi^2}{(6\chi^2)} \]
BLAST print out they give "bit" score.

\[ \text{bit score} = \frac{A \times \text{max} - \log K}{\log 2} \]

\[ = \frac{A \times \text{max}}{\log 2} \quad \text{(old formula)} \]

Invariant to any multiplicative constant A.

A distribution S' is known whatever may be the substitution matrix since \[10.27\] is free of any parameter.

Similarly if N is given the bit score distribution is known.

(Add the dist of S')

\[ BS = S' + \left( \log \frac{NK^2}{K} \right) \frac{1}{\log 2} \]
# 06 High Scoring Excursions

$Y_1, Y_2, \ldots$ max of excursions from
labeled $p$ to $ind$

$N \approx \# \text{excursions} \approx \frac{N}{A}$

$P(Y_i \geq y) \approx c \ e^{-\lambda y}$

$U = \# \text{of excursions } h-1 \geq y$
$= \# \text{of highest scoring pairs } Z+SPs$

$U \sim \text{Bin} \left( \frac{N}{A}, c \ e^{-\lambda y} \right)$

$E(U) = \frac{N}{A} c \ e^{-\lambda y} = N K \ e^{-\lambda y+1}$

$1 \leq \frac{c}{A} \ e^{-\lambda y} \approx N K \ e^{-\lambda y}$

(Blas)

$Y \sim \text{Poisson} \left( N K e^{-\lambda y} \right)$

(Evolve) BLAST

$E = \text{expected } \# \text{of excursions corresponding to the maximal value } y\text{max}$

$= N K \ e^{-\lambda y\text{max}}$
\[ S' = \log E' \]

\[ R\text{ value} = P(S' > s) \approx 1 - e^{-s'} = 1 - e^{-e} \]

**Kalinin-Altschul Sum Statistic**

A statistic that uses information of the 2nd largest, 3rd largest, etc. encumbrances also in the R.M. along with the largest encumbrance.

\[ S_i' \] (Score calculated using i'th largest encumbrance)

\[ T_r = S_1' + \ldots + S_r' \]

\[ P(T_r > t) \approx \frac{e^t + r - 1}{r! (r - 1)!} \]
Comparison of 2 Unaligned Sequences

We have 2 sequences of length $N_1$ and $N_2$. Find the significance of the highest scoring segment-pair between all possible ungapped local alignments. The highest scoring pair is known as maximal-scoring segment pair (MSP).

Total # of alignments $= N_1 + N_2 - 1$

Seq 1: 
Seq 2: 

Seq 1: 
Seq 2: 

Seq 1: 
Seq 2: 

Same as in 10.2.2 but replace

$N = N_1 N_2$

$Y_{max} = \text{max score achieved in } X \text{ RM Company } X$

Seq. = MSP
MSP starts at a ladder point in BLAST R M and finishes the first time that the max excursion from this ladder is reached.

\[ Y_{\text{max}} \sim \text{geometric like} \]

will \( \lambda \), C, n.

The mean \# of ladder points = \( \frac{N_1 N_2}{A} \)

Under null hypothesis

\[ 1 - \bar{e} K e^{-\frac{z}{2}} \leq P(Y_{\text{max}} > x) \leq 1 - \bar{e} K e^{-2(z-1)} \]

\[ S_{\text{max}} \leq \tilde{Y}_{\text{max}} - \log(N_1 N_2 K) \]

\[ E' = N_1 N_2 K e^{-\frac{1}{2} Y_{\text{nom}}} \]

\[ E(Y_{\text{nom}}) = x'(\log(N_1 N_2 K) + x) \]

\[ P - \text{rel} \equiv 1 - \bar{e} E' \]
Edge Effects

Overlap = edge factor effect

\[ y \] = height achieved by a high scoring excursus.

\[ L \] = length of the excursus.

\[ E(L | y) = \frac{\lambda y}{H} \]

\[ N_1' = N_1 - E(L) \], \[ N_2' = N_2 - E(L) \]

\[ \lambda Y_{max} = \log(N_1'N_2'K) \]

\[ N_1' = N_1 - \frac{\lambda Y_{max}}{H} \]

\[ N_2' = N_2 - \frac{\lambda Y_{max}}{H} \]

Expected # of excursus scores \( y \) or higher

\[ = N_1'N_2'K e^{\frac{-\lambda y}{H}} \]

\[ N_1' = N_1 - \frac{\lambda y}{H} \]

\[ N_2' = N_2 - \frac{\lambda y}{H} \]

\[ E' = N_1'N_2'K e^{\frac{-\lambda y}{H}} \]
Karlin Altschul sum statistic $T_{*}$ with raw edge effect correction $\lambda \left( \sum_{i=1}^{\gamma} Y_i \right)
\frac{B_z}{H}

N_2' = N_2 - E(L)
N_1' = N_1 - E(L)

\therefore S_i' = \lambda Y_i - \log \left( N_1', N_2', K \right)

K-A sum statistic

\[ T_{*} = \sum_{i=1}^{\gamma} S_i' \]
Multiple Testing

No choice for the value of $\gamma$. Consider all $\gamma = 1, 2, 3, \ldots$.

Chose the set of HSPs for which $P(T_0 > t)$ is smallest, i.e., one test for each $\gamma$, then involve multiple testing.

Adjusted P-value: $\frac{P(T_0 > t)}{(1 - \gamma)^{\gamma-1}}$

$\gamma = 0.5$

$E' = \frac{E'(1035)}{(1 - \gamma)}$ when $\gamma = 1$

$E' = \varepsilon$

$E' = 2 N_1 N_c \left| 1 - \varepsilon \right|^b$

P-value $\Rightarrow 1 - \varepsilon E$