

## A Brief Tutorial on BLAST

Dirk Husmeier

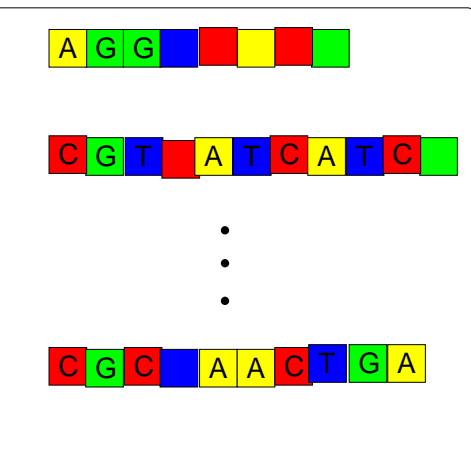
Biomathematics and Statistics Scotland  
at the Scottish Crop Research Institute  
Invergowrie, Dundee DD2 5DA, UK

Email: [dirk@bioss.ac.uk](mailto:dirk@bioss.ac.uk)  
<http://www.bioss.ac.uk/~dirk>

## Motivation

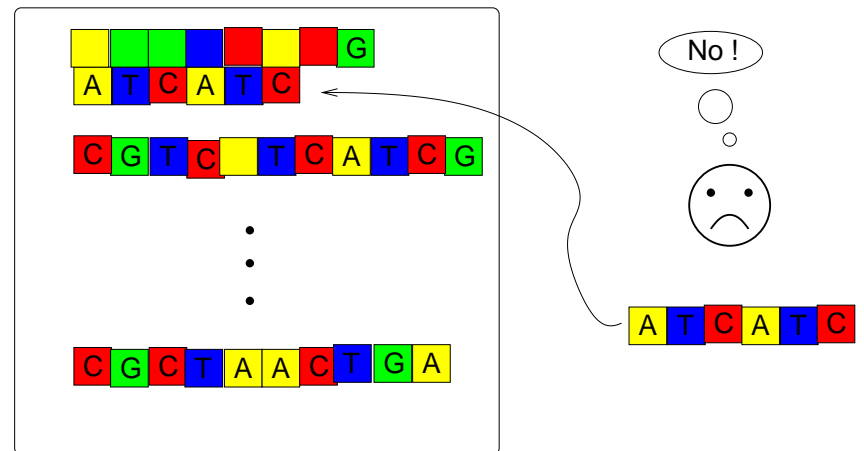
- Gene involved in **diabetes** .
- **Pharma company**: Design drug to target **regulatory sequences** → Prevent expression of this gene.
- Only have **2000** bases, need the entire gene ( **100 000** ) bases.
- Homology search of the human **databank GenBank** with **BLAST** . New sequences are added to GenBank daily.
- 4 months later (June 2000), they get a **match** → **entire gene** .
- Plug the entire **human gene** into the **mouse-genome database** . **BLAST** homology search.
- Find **five matches** in the noncoding area.
- DNA involved in **gene regulation** is well conserved during evolution → These region are putative targets for drugs.

## Motivation



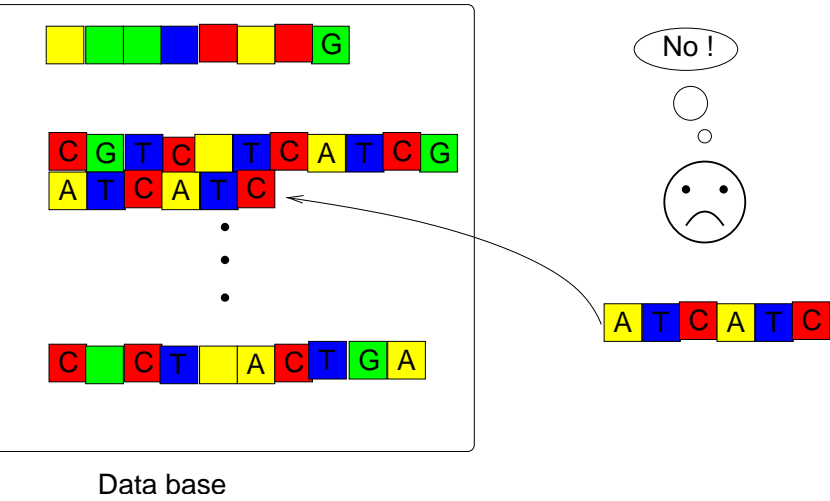
Data base

## Motivation

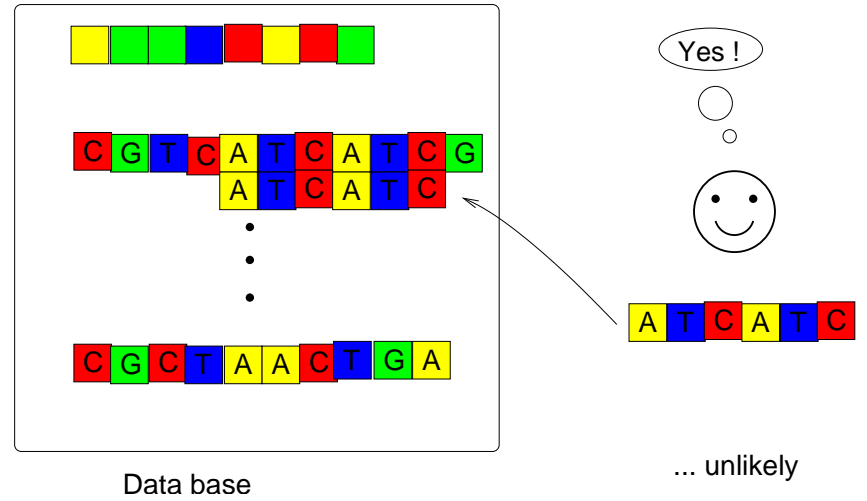


Data base

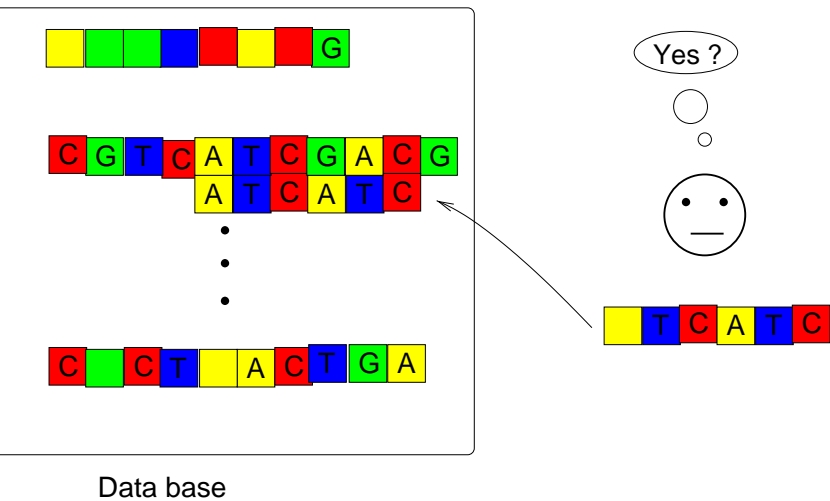
Motivation



Motivation



Motivation



Preliminaries: Moment generating function

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

- Moment generating function (mgf):

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

- $k$ th moment

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

- For  $N$  iid random variables  $Y_1, \dots, Y_N$ :

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

Proof:  $m_{Y_1+\dots+Y_N}(t) = \sum_{y_1, \dots, y_N} P(y_1, \dots, y_N) e^{t(y_1+\dots+y_N)} = \sum_{y_1, \dots, y_N} P(y_1) \dots P(y_N) e^{t(y_1+\dots+y_N)}$

$$= \prod_{i=1}^N \sum_{y_i} P(y_i) e^{ty_i} = \left[ \sum_y P(y) e^{ty} \right]^N = m_Y(t)^N$$



## Distribution of peak height $Y$ under the null hypothesis: Example

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

Single-step mgf:  $m_S(t) = pe^t + (1-p)e^{-t}$

$$m_S(\lambda) = 1 \implies pe^{2\lambda} - e^\lambda + (1-p) = 0 \implies e^\lambda \in \left\{1, \frac{1-p}{p}\right\}$$

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

Multi-step mgf:  $m_{S_1+\dots+S_N}(\lambda) = [m_S(\lambda)]^N = 1$   
 $\implies$  Same root  $\lambda$ .

## Distribution of the peak height $Y$ under the null hypothesis: Example

- 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  $P_{(D=y)} = P_{(Y \geq y)}$
- Displacement mgf:  $m_D(t) = \sum_D e^{Dt} p(D)$
- $m_D(t) = P_{(D=-1)}e^{-t} + P_{(D=y)}e^{yt} = e^{-t} + P_{(D=y)}(e^{yt} - e^{-t})$
- $m_D(\lambda) = 1 \implies e^{-\lambda} + P_{(Y \geq y)}(e^{y\lambda} - e^{-\lambda}) = 1$
- $\implies P_{(Y \geq y)} = \frac{1 - e^{-\lambda}}{e^{y\lambda} - e^{-\lambda}}$ . For  $y \rightarrow \infty$ :  $P_{(Y \geq y)} = [1 - e^{-\lambda}]e^{-\lambda y}$ .
- Null hypothesis: **Geometric-like distribution**:  $P_{(Y \geq y)} \rightarrow Ce^{-\lambda y}$ , with  $C = 1 - e^{-\lambda}$

## Distribution of the peak height $Y$ under the null hypothesis: Generalisation

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  $y \rightarrow \infty$ ,  $P(Y \geq y) = Ce^{-\lambda y}$ .
- This means that under the null hypothesis, the maximum height  $Y$  has (asymptotically) a **geometric-like distribution**.

## Average number of high-scoring excursions $E$

- What is the **average 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.
- Average **number of high-scoring excursions** with  $Y \geq y$ :  
 $E = nP(Y \geq y) = nCe^{-\lambda y}$

## Average number of high-scoring excursions $E$

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

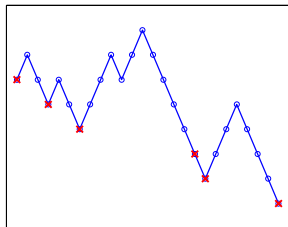
- Average **step size**:  $\langle S \rangle = (1)p + (-1)(1 - p) = 2p - 1$
- Average **final displacement**:  $P_{(D=-1)} = 1 \implies \langle D \rangle = -1$
- Average **length of an excursion**:  $A = \frac{\langle D \rangle}{\langle S \rangle} = \frac{(-1)}{2p - 1} = \frac{1}{1 - 2p}$
- Average **number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

$$E = nP(Y \geq y) = nCe^{-\lambda y}, \quad C = 1 - e^{-\lambda}$$

$$= (1 - 2p)(1 - e^{-\lambda})Ne^{-\lambda y}$$

## Example

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



core function:  $S(i, k) = \begin{cases} +1 & \text{if } i = k \\ -1 & \text{if } i \neq k \end{cases}, \quad p = P(S = 1) = 0.25$

average **number of high-scoring excursions** with  $Y \geq y$ :  $E = (1 - 2p)(1 - e^{-\lambda})Ne^{-\lambda y}$

from **single-step mgf**:  $\lambda = \ln \frac{1-p}{p} = \ln \frac{0.75}{0.25} = \ln 3$

$$C(y) = 0.5 \left(1 - \frac{1}{3}\right) N \left(\frac{1}{3}\right)^y = N \left(\frac{1}{3}\right)^{y+1}$$

$$N = 25, Y_{max} = 4 \rightarrow E = 25 \left(\frac{1}{3}\right)^5 = 0.103$$

$$p\text{-value} = 1 - e^{-E} = 0.098$$

## 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**:

$$- P_{(Y_{max} \leq y)} = P_{(Y_1 \leq y \wedge Y_2 \leq y \wedge \dots \wedge Y_n \leq y)} = \prod_{i=1}^n P_{(Y_i \leq y)} = \left[P_{(Y \leq y)}\right]^n$$

$$- P_{(Y_{max} \geq y)} = 1 - P_{(Y_{max} \leq y-1)} = 1 - \left[P_{(Y \leq y-1)}\right]^n = 1 - \left[1 - P_{(Y \geq y)}\right]^n$$

$$- \text{From } P_{(Y \geq y)} \approx Ce^{-\lambda y} \implies P_{(Y_{max} \geq y)} = 1 - \left[1 - Ce^{-\lambda y}\right]^n$$

- Approximation to the extreme value distribution (without derivation):

$$P_{(Y_{max} \geq y)} \approx 1 - e^{-E(y)}$$

where  $E(y)$  = expected number of excursions with  $Y_{max} \geq y$  :

$$E(y) = nP_{(Y_{max} \geq y)} \approx nCe^{-\lambda y}$$

## 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:  $N \rightarrow N_1 N_2$

- **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**.

- **Database search**: Have a **query sequence** , search an **entire database** of many sequences for those with significant similarity to the query sequence.  $\rightarrow$  Correct for **multiple testing**

## 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**.
- $\tau$  represents evolutionary time. The larger  $\tau$ , the larger the **evolutionary distance** between the sequences.

$$\tau \rightarrow 0 \implies q_\tau(i, k) \rightarrow \delta_{ik}$$

$$\tau \rightarrow \infty \implies q_\tau(i, k) \rightarrow p_i p_k$$

- Two **sequences are related**, and we know the correct value of  $\tau \rightarrow$  **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**  $\rightarrow$  **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  $\tau'$**  rather than correct value  $\tau$ .

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

- 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  $q_\tau(i, k)$  than to  $p_i p_k$ :

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

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