Substitution matrices

Need mainly two things:

– For every pair \( a, b \): \( p_{ab} \), the probability of observing \( a \) aligned with \( b \). \( p_{bc} = p_{cb} \)

– For every \( a \): \( p_a \), the probability of observing an \( a \).

Aligned sequences

Related / Unrelated

• Let \( M \) be the model in which \( x \) and \( y \) are related and obtained according to the joint probabilities \( p_{xy} \).

• Let \( R \) be the model in which \( x \) and \( y \) are unrelated and obtained independently at random according to the individual probabilities \( p_x \).
score

- The score of a $x$ and $y$ is the relative likelihood that the sequences are related compared to being unrelated: *odds ratio*

$$\text{score}(x, y) = \frac{p(x, y|M)}{p(x, y|R)} = \prod_i p_{x_i} \prod_j p_{y_j}$$

Intuition

$x = \ldots \ldots a \ldots \ldots$  
$y = \ldots \ldots b \ldots \ldots$  

Let $p(a \rightarrow b) = \text{probability that } a \text{ mutates into } b$ (if $p(b \rightarrow a)$)

Taking the point of view of $a$, the probability that $b$ is there is $p(b \rightarrow a)$.

But there is a chance of $p_0$, the probability that $b$ is there $p(a \rightarrow b)$.

This ratio is $\frac{p(a \rightarrow b)}{p_0}$.

But $p_0 = p_0$, $p(a \rightarrow b)$, therefore we get $\frac{p_{a \rightarrow b}}{p_0}$.

Additive score

- The score is multiplicative
  $$\prod_i \frac{p_{x_i}}{p_{x_i} p_{y_i}}$$

- To make it additive, take the log
  $$\log \prod_i \frac{p_{x_i}}{p_{x_i} p_{y_i}} = \sum_i \log \frac{p_{x_i}}{p_{x_i} p_{y_i}}$$

- Substitution matrix $S$ where $S(a, b) = \log \frac{p_{a \rightarrow b}}{p_{b \rightarrow a}}$
PAM matrices

- Stands for *Point Accepted Mutations*.
- An accepted mutation is a mutation that was positively selected by the environment and did not cause the death of the organism.
- Given a PAM matrix $M, M_{ab} = p[a \to b]$ in a certain *evolutionary time period*.

Unit of Evolution

- It is difficult to capture from statistical data the relation of proteins that are evolutionary very far apart. If $a \to b$, we don't capture the intermediate mutations.
- Define 1 unit of evolution as the amount of evolution that will change 1 in 100 amino acids on average.
- Compute the 1-PAM matrix corresponding to 1 unit of evolution from short time interval statistical data.
- Obtain other $k$-PAM matrices from the first one.

1-PAM matrix

- Compute a matrix $M, M_{ab} = p[a \to b]$ for all $a, b$.
- Scale $M$ such that the expected number of mutations $\sum p_a (1 - M_{ab})$ is 0.01 (1%).
  [this is same as the probability of a mutation]
- Compute $p_a$ for every $a$.
- Then use $S(a,b) = 10 \log_2 (M_{ab}/p_a)$ to obtain an additive score.
1-PAM Computation

- Let $f_{ab}$ = number of times a is aligned with b (both direction).
- Let $f_a = \sum f_{ab}$ (number of a’s).
- Let $f = \sum f_a$ (all characters).
- Estimate $p_{ab}$ as $f_{ab}/f$.
- Estimate $p_a$ as $f_a/f$.
- Then $M_{ab} = p(a \rightarrow b) = p_{ab}/p_a$.
- Note $\Sigma M_{ab} = 1$.

Computation (cont.)

- $M_{ab} = \alpha M_{ab}$ (if $a \neq b$).
- $M_{aa} = \alpha M_{aa} + 1 - \alpha$.
- Note, we still have $\Sigma M_{ab} = 1$.
- $\Sigma \rho_a (1 - M_{ab}) = \alpha \Sigma \rho_a (1 - M_{aa})$.

2-PAM matrix

- $p_a[a \rightarrow b]$ in two units of evolution will be the probability of a mutating to some character c in one unit of evolution and c mutating to b in another unit of evolution.
- $p_a[a \rightarrow b] = \Sigma c p[a \rightarrow c]p[c \rightarrow b] = \Sigma c M_{ac}M_{cb}$
- 2-PAM matrix = $M^2$.
k-PAM matrix

- $k$-PAM = $M^k$
- $S_k(a,b) = 10 \log_{10} \left( \frac{M^k_{ab}}{p_a p_b} \right)$

BLOSUM matrices
(BLOCKS substitution matrices)

- BLOSUM matrices are derived from a database of BLOCKS (the BLOCKS database) where each block is a multiple ungapped alignment of related protein sequences.
- The goal is to obtain a scoring for protein sequences that are evolutionary far apart. How far?
- The sequences from each block are clustered, putting two sequences in the same cluster if they have more than 2% similarity (percentage of aligned matching residues).

BLOSUM computation

- Count number of mutations between distant sequences only, i.e. less than L% similar.
- a and b aligned but end up in different clusters.
- Increment $f_{ab}$ by $1/n_1 n_2$ every time this happens.
Computation (cont.)

- Estimate $p_a$ as $\frac{\sum f_a}{\sum f_a}$
- Estimate $p_{ab}$ as $\frac{f_{ab}}{\sum f_{ab}}$
- BLOSUM-L $(a,b) = \log \frac{p_a}{p_{ab}}$

Example score

- Random Model $R$: Assume $p_a=p_{bc}=p_{c}=1/4$
- Related model $M$: Assume 50% similarity
  - $p_{ba} = \frac{1}{2} \cdot \frac{1}{4} = 1/8$
  - $p_{a} = \frac{1}{2} \cdot \frac{1}{4} (4^4 - 4) = 1/24$

- $m = \log \frac{p_a}{p_{ba}} \cdot \log \frac{1/8}{1/24 / 4/4} = 1$
- $s = \log \frac{p_a}{p_{ba}} \cdot \log \frac{1/24}{1/4/4/4} = -0.585$