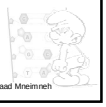
Computational Biology

Lecture 8



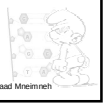
Saad Mneimneh

Substitution matrices

Need mainly two things:

– For every pair ***a***, ***b***: p_{ab}, the probability of observing ***a*** aligned with ***b***. $p_{ab} = p_{ba}$

– For every ***a***: p_a, the probability of observing an ***a***.



Saad Mneimneh

Aligned sequences Related / Unrelated

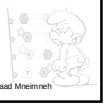
- Let M be the model in which x and y are related and obtained according to the joint probabilities p_{ab}.
- Let R be the model in which x and y are unrelated and obtained independently at random according to the individual probabilities p_a.

Saad Mneimneh

score

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

score(*x*, *y*) = $\frac{p(x,y \mid M)}{p(x,y \mid R)} = \frac{\prod_i p_{x_i y_i}}{\prod_i p_{x_i} \prod_i p_{y_i}} = \prod_i \frac{p_{x_i y_i}}{p_{x_i} p_{y_i}}$



Saad Mneimneh

Intuition

x = ……***a***……..
y = ……***b***……..

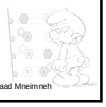
Let $p[a \rightarrow b]$ = probability that ***a*** mutates into ***b*** ($\neq p[b \rightarrow a]$)

Taking the point of view of ***a***, the probability that ***b*** is there is $p[a \rightarrow b]$.

But there is a chance of p_b for a random occurrence of ***b***.

This ratio is: $\frac{p[a \rightarrow b]}{p_b}$

But $p_{ab} = p_a.p[a \rightarrow b]$, therefore we get $\frac{p_{ab}}{p_a p_b}$



Saad Mneimneh

Additive score

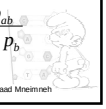
- The score is multiplicative

$$\prod_i \frac{p_{x_i y_i}}{p_{x_i} p_{y_i}}$$

- To make it additive, take the log

$$\log \prod_i \frac{p_{x_i y_i}}{p_{x_i} p_{y_i}} = \sum_i \log \frac{p_{x_i y_i}}{p_{x_i} p_{y_i}}$$

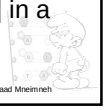
- Substitution matrix *S* where $S(a,b) = \log \frac{p_{ab}}{p_a p_b}$



Saad Mneimneh

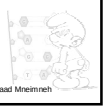
PAM matrices

- 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, $M_{ab} = p[a \rightarrow b]$ 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 $\boldsymbol{a} \rightarrow \boldsymbol{b}$, 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.



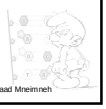
1-PAM matrix

- Compute a matrix M, $M_{ab} = p[a \rightarrow b]$ for all $\boldsymbol{a}$,$\boldsymbol{b}$.
- Scale M such that the expected number of mutations $\Sigma_a p_a.(1 - M_{aa})$ is 0.01 (1%).
  [this is same as the probability of a mutation]
- Compute p_a for every $\boldsymbol{a}$.
- Then use

  scale to reduce discrepancy with the integer approximation

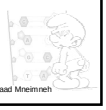
  $$S(a,b) = 10 \log_{10} (M_{ab}/p_b)$$

  to obtained an additive score.



1-PAM Computation

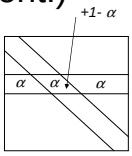
- Let f_{ab} = number of times a is aligned with b (both direction).
- Let $f_a = \Sigma_b f_{ab}$ (number of a's)
- Let $f = \Sigma_a f_a$ (all characters)
- Estimate p_{ab} as f_{ab}/f
- Estimate p_a as f_a/f
- Then $M_{ab} = p[a \rightarrow b] = p_{ab}/p_a$
- Note $\Sigma_b M_{ab} = 1$



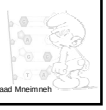
Saad Mneimneh

Computation (cont.)

- $M'_{ab} = \alpha M_{ab}$ (if $a \neq b$)
- $M'_{aa} = \alpha M_{aa} + 1 - \alpha$



- Note, we still have $\Sigma_b M'_{ab} = 1$.
- $\Sigma_a p_a(1 - M'_{aa}) = \alpha \Sigma_a p_a(1 - M_{aa})$



Saad Mneimneh

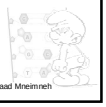
2-PAM matrix

- $p_2[a \rightarrow b]$ 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.
- $p_2[a \rightarrow b] = \Sigma_c p[a \rightarrow c].p[c \rightarrow b] = \Sigma_c M_{ac}.M_{cb}$
- 2-PAM matrix = M^2

Saad Mneimneh

k-PAM matrix

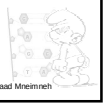
- k-PAM = M^k
- $S_k(a,b) = 10 \log_{10} (M^k_{ab}/p_b)$



Saad Mneimneh

BLOSUM matrices (BLOCKS substitution matrices)

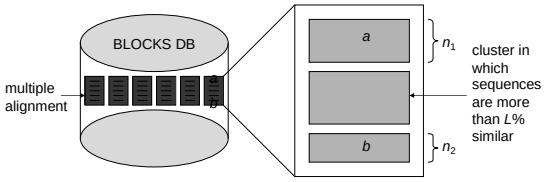
- BLOSUM matrices are derived from a database of BLOCKS (the BLOCKS database) where each block is a multiple ungapped alignment of related protein sequences.
- The goal is to obtain a scoring for protein sequences that are evolutionary far apart. How far?
- The sequences from each block are clustered,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



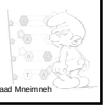
Saad Mneimneh

BLOSUM computation

- Count number of mutations between distant sequences only, i.e. less than L% similar.



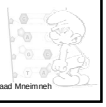
- ***a*** and ***b*** aligned but end up in different clusters.
- Increment f_{ab} by $1/n_1n_2$ every time this happens.



Saad Mneimneh

Computation (cont.)

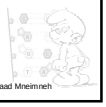
- Estimate p_a as $\frac{\sum_b f_{ab}}{\sum_{c,d} f_{cd}}$
- Estimate p_{ab} as $\frac{f_{ab}}{\sum_{c,d} f_{cd}}$
- BLOSUM-L (a,b) = $\log \frac{p_{ab}}{p_a p_b}$



Saad Mneimneh

Example score

- Random Model R : Assume $p_A=p_G=p_C=p_T=1/4$
- Related model M: Assume 50% similarity
  - p_{aa} = ½ . ¼ = 1/8
  - P_{ab} = ½. 1/(4^2 – 4) = 1/24
- m = $\log \frac{P_{aa}}{p_a p_a} = \log \frac{1/8}{1/4.1/4} = 1$
- s = $\log \frac{P_{ab}}{p_a p_b} = \log \frac{1/24}{1/4.1/4} = -0.585$



Saad Mneimneh