Algorithms for Bioinformatics - lect 4

Francesco Penasa

March 4, 2020

`2020 03 04`

Till now, we have talked about global sequence alignment (NW algorithm) and local sequence Alignment (SW algorithm). In both cases we want to minimize a score, $M_{i,j}$.

Until now the alphabet we have seen is limited to the DNA (ATCG), $match = 1$ $mismatch = -1$ independently from the actualt letters. Both algorithm that we have seen so far are correct, we actually find the maximum score.

1. solve pratical problems (so far we are here, NW and SW)
2. modeling a biological phenomenom
3. understanding the computation done by the biological system.

1 Modeling a biological phenomenom

`http://xaktly.com/GeneticCode.html` here we can see the abbrevietions for the aminoacid, this means that so far we use only one alphabet but we could use even the aminoacid one.

AUG AUU GGU
AUG AUU GGC

this seems different but due to the fact that the last triplet represents the same aminoacid the actual difference is 0. With this alphabet we don't expect that the match-mismatch model works in the same optimal way.

1.1 Substitution matrices

1. PAM (1978) `https://en.wikipedia.org/wiki/Point_accepted_mutation` Percent Accepted Mutation.
2. BLOSUM

AAU
AAA

1. Generation

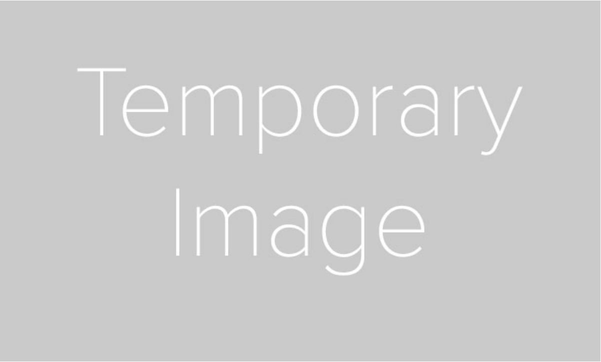

Figure 1: For each aminoacid we have the probability of be substituted by another aminoacid, basically a markov chain.

Table 1: accepted point mutation

	A	B	C	D	G	H	I	J
A			1	1				
B			1	1				
C	1	1						
D	1	1						
G								
H								
I								
J								

2. Conservation (acceptation)

Let's assume we have this kind of sequence

ACGH DBGH ADIJ CBIJ
(ACGH)−ABGH−(DBGH) (ADIJ)−ABIJ−(CBIJ)
(ABGH)−AB**−(ABIJ)

$$A_{jk} = \frac{1}{nT} \sum_T A^T_{jk}$$

$$P^{(1)}_{jk} = m_j \frac{A_j k}{\sum_{h \neq j} A_{jh}}$$

$$P^{(1)}_{jj} = 1 - m_j$$

$$m_j = \frac{1}{np_j z} \frac{\sum l \neq j A_j l}{\sum_h \sum_{l \neq h} A_{hl}}$$

1. m_j = the mutability of the amminoacid j

2. z is a normalizing factor s.t.$= \sum_{ji}(P_j \mu_j) = \frac{1}{100}$

The steps we needed were, find probabilities and find the scores.

$$q_{jk}^{(n)} = P_j p_{jk}^{(n)}$$

$$S_{[j,k]}^{n} = \lambda log \frac{q_{jk}}{P_j P_k} = \lambda log \frac{P_{jk}^{(n)}}{P_k}$$

j and k are aminoacid, p their propabilities, q is the probability that j mutated in a certain way, the score is the comparison.

$$S_{[j,k]}^{(n)} \neq S_{[k,j]}^{(n)}$$

the PAM100 matrix represents 100% change, the PAM250 matrix represents 250% change. They were determined by the global alignment of sequences that differ by less than 85%. One PAM represents a 1% change in all residues or one Point Accepted Mutation per 100 residues

1.1.1 BLOSUM

While PAM works with trees, BLOSUM works with **BLOCKS!**
BLOCKS: sequences with a level of similarity put togheter through clustering techniques and then we build some kind of statistic of elements in the same cluster. The similarity is give with a percentage (62 is the at-least 62 similarity used for the clusters.)

Table 2: relation between results of BLOSUM and PAM

	PAM	**BLOSUM**
20%	250	45
30%	160	62
40%	120	80