BLOSUM & PAM matrices

INFOF434 - 23/11/18

Blosum

- **Blo**cks **Su**bstitution **M**atrix. Scores for each position are obtained frequencies of substitutions in blocks of local alignments of protein sequences (Henikoff & Henikoff, 1992) .

- For example BLOSUM62 is derived from sequence alignments with no more than 62% identity.

BLOSUM overview

1. Look for conserved (gapless, <=62% identical) regions in alignments.

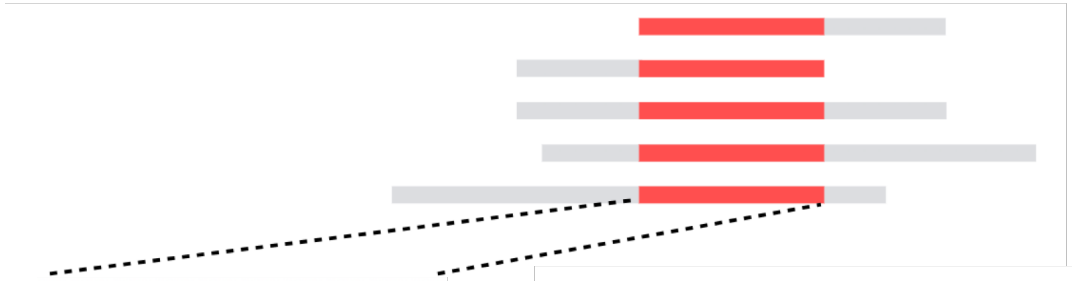
2. Count all pairs of amino acids in each column of the alignments.

3. Use amino acid pair frequencies to derive “score” for a mutation/replacement

Step 1 : conservated regions

- Add sequences of block with a least x%

Step 2 : count the pair frequences

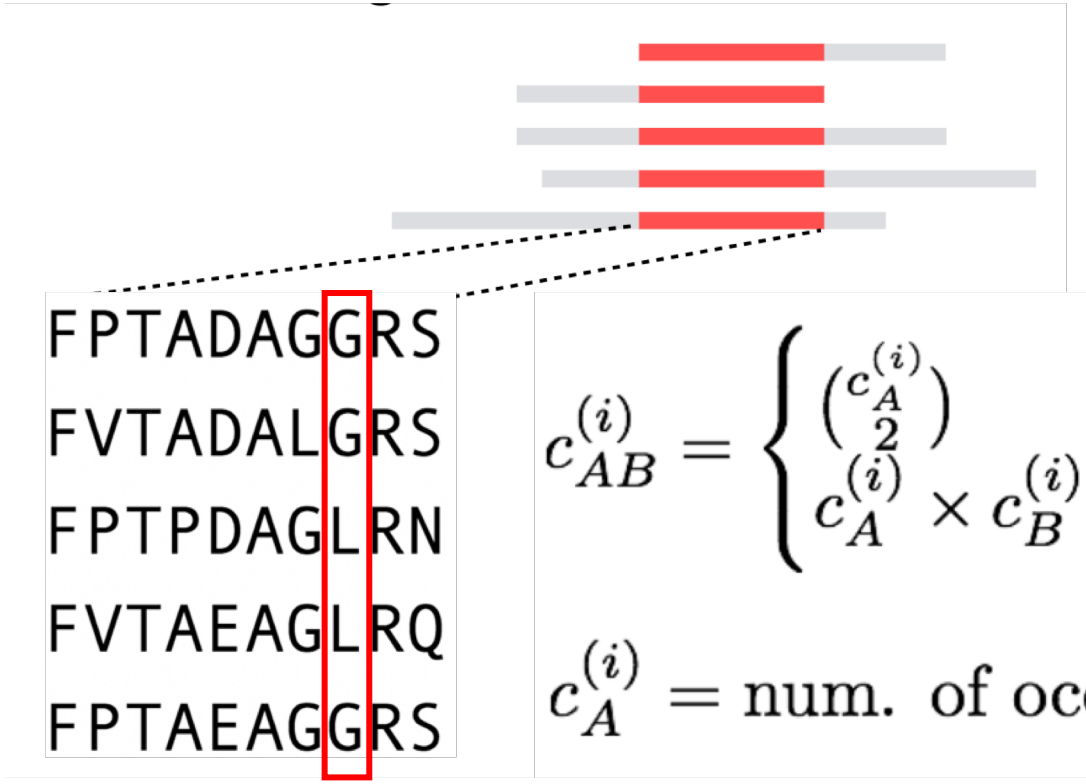


FPTADAGGRS
FVTADALGRS
FPTPDAGLRN
FVTAEAGLRQ
FPTAEAGGRS

$$c_{AB}^{(i)} = \begin{cases} \binom{c_A^{(i)}}{2} & \text{if } A = B \\ c_A^{(i)} \times c_B^{(i)} & \text{otherwise} \end{cases}$$

$$c_A^{(i)} = \text{num. of occurrences of } A \text{ in column } i$$

Step 2 : count the pair frequences



$$c_{AB}^{(i)} = \begin{cases} \binom{c_A^{(i)}}{2} & \text{if } A = B \\ c_A^{(i)} \times c_B^{(i)} & \text{otherwise} \end{cases}$$

$$c_A^{(i)} = \text{num. of occurrences of } A \text{ in column } i$$

$$c_{GG}^{(i)} = \binom{3}{2} = 3$$

$$c_{GL}^{(i)} = 3 \times 2$$

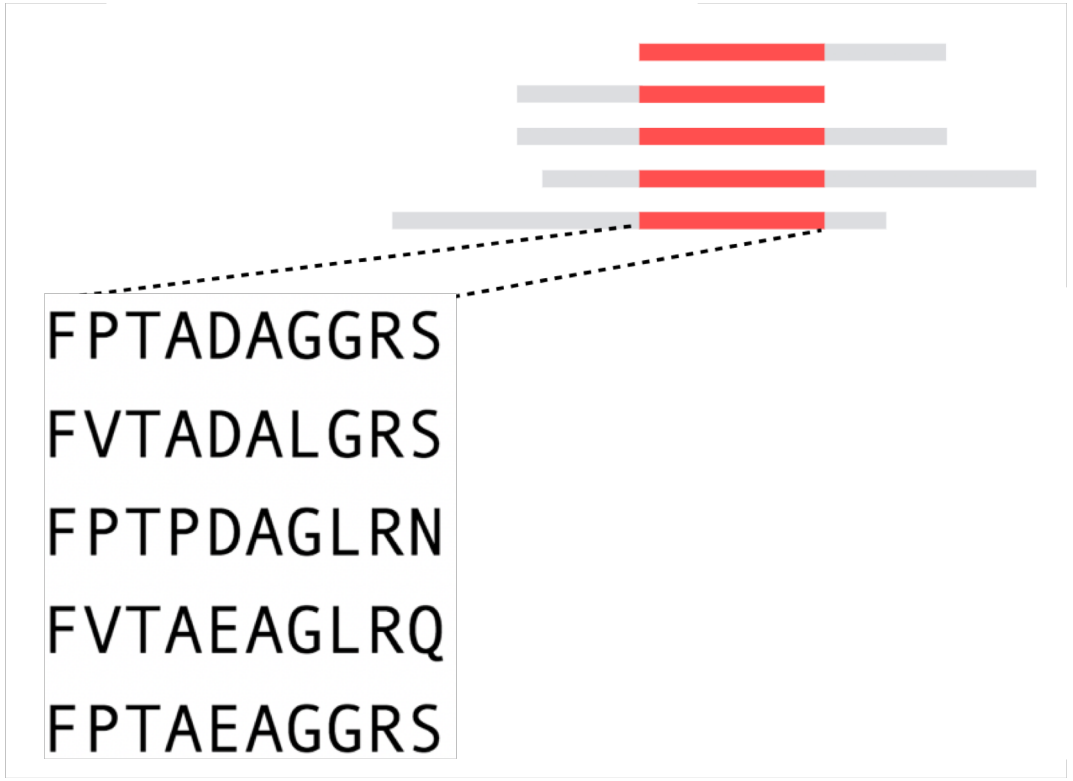
$$c_{LL}^{(i)} = \binom{2}{2} = 1$$

Step 3 : compute the scores

Total # of potential align. between A & B: $c_{AB} = \sum_i c_{AB}^{(i)}$

Total number of pairwise char. alignments: $T = \sum_{A \geq B} c_{AB}$

Normalized frequency of aligning A & B: $q_{AB} = \frac{c_{AB}}{T}$



In our example, we get

$$q_{GL} = \frac{0+0+0+0+0+0+0+4+6+0+0}{10\frac{(5)(4)}{2}} = \frac{10}{100}$$

Step 3 : compute the scores

Probability of occurrence of amino acid A in any {A,B} pair:

$$p_A = q_{AA} + \sum_{A \neq B} \frac{q_{AB}}{2}$$

Expected likelihood of each {A,B} pair, assuming independence:

$$e_{AB} = \begin{cases} (p_A)(p_B) = (p_A)^2 & \text{if } A = B \\ (p_A)(p_B) + (p_B)(p_A) = 2(p_A)(p_B) & \text{otherwise} \end{cases}$$

Step 3 : from probabilities to score

Recall the original idea (likelihood → scores)

$$\text{score} = \log \text{ odds ratio} = s_{AB} \propto \log\left(\frac{\text{observed}}{\text{expected}}\right)$$

$$s_{AB} = \text{Round}\left(2\log_2\left(\frac{q_{AB}}{e_{AB}}\right)\right)$$

Example

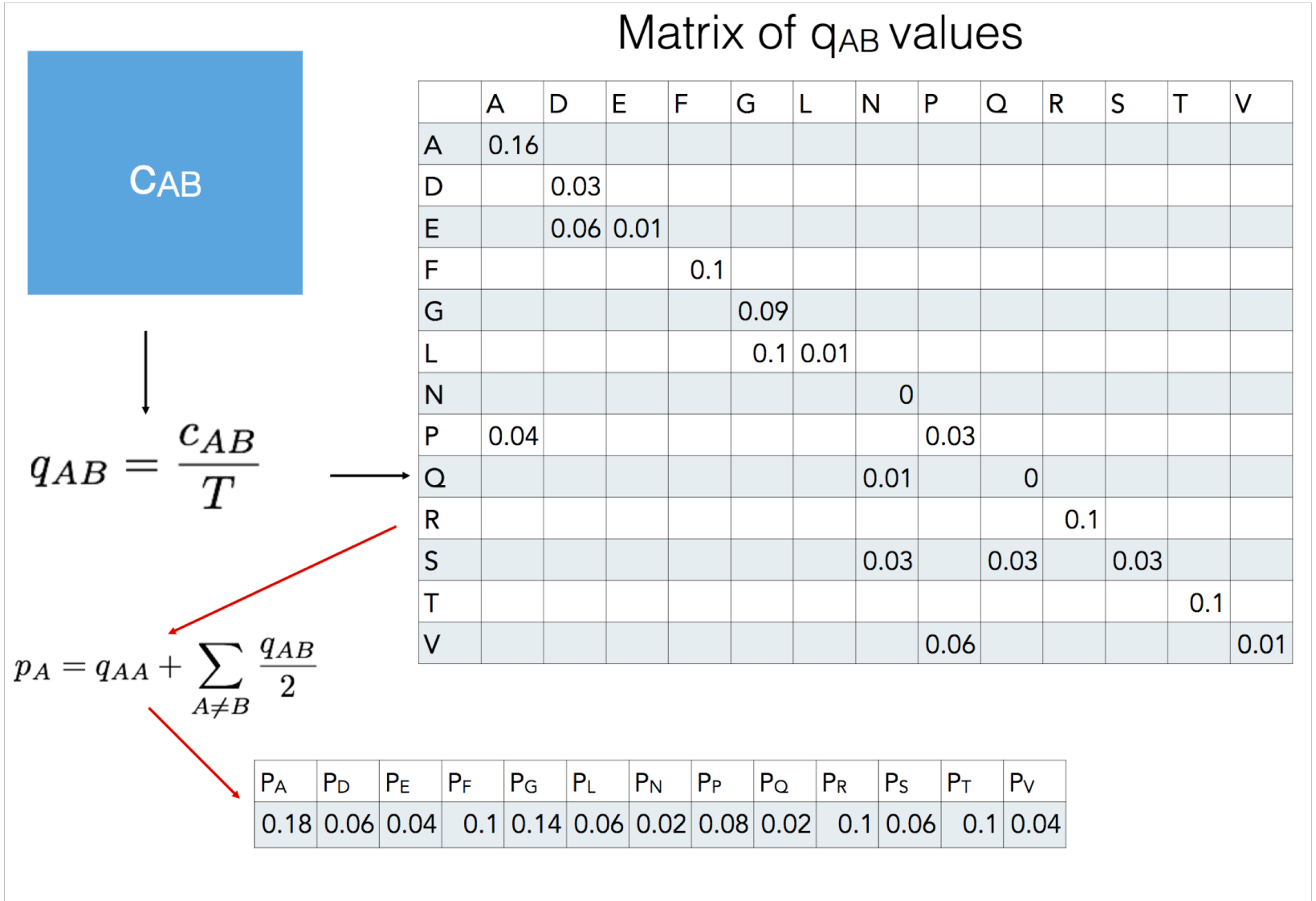
FPTADAGGRS
FVTADALGRS
FPTPDAGLRN
FVTAEAGLRQ
FPTAEAGGRS

$$c_{AB} = \sum_i c_{AB}^{(i)}$$

Matrix of c_{AB} values

	A	D	E	F	G	L	N	P	Q	R	S	T	V
A	16												
D		3											
E		6	1										
F				10									
G					9								
L					10	1							
N							0						
P	4							3					
Q							1		0				
R										10			
S							3		3		3		
T												10	
V								6					1

Example



Matrix of q_{AB} values

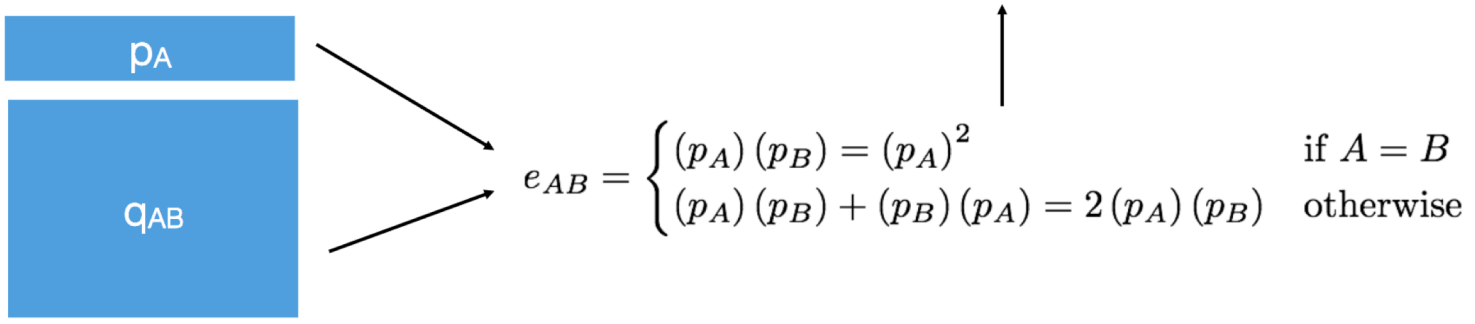
	A	D	E	F	G	L	N	P	Q	R	S	T	V
A	0.16												
D		0.03											
E		0.06	0.01										
F				0.1									
G					0.09								
L					0.1	0.01							
N							0						
P	0.04							0.03					
Q							0.01		0				
R										0.1			
S							0.03		0.03		0.03		
T												0.1	
V								0.06					0.01

P_A	P_D	P_E	P_F	P_G	P_L	P_N	P_P	P_Q	P_R	P_S	P_T	P_V
0.18	0.06	0.04	0.1	0.14	0.06	0.02	0.08	0.02	0.1	0.06	0.1	0.04

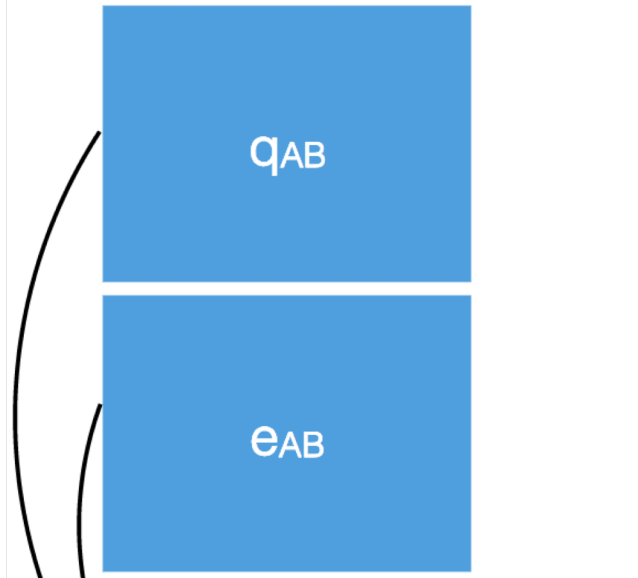
Example

Matrix of e_{AB} values

	A	D	E	F	G	L	N	P	Q	R	S	T	V
A	0.0324												
D	0.0216	0.0036											
E	0.0144	0.0048	0.0016										
F	0.0360	0.0120	0.0080	0.0100									
G	0.0504	0.0168	0.0112	0.0280	0.0196								
L	0.0216	0.0072	0.0048	0.0120	0.0168	0.0036							
N	0.0072	0.0024	0.0016	0.0040	0.0056	0.0024	0.0004						
P	0.0288	0.0096	0.0064	0.0160	0.0224	0.0096	0.0032	0.0064					
Q	0.0072	0.0024	0.0016	0.0040	0.0056	0.0024	0.0008	0.0032	0.0004				
R	0.0360	0.0120	0.0080	0.0200	0.0280	0.0120	0.0040	0.0160	0.0040	0.0100			
S	0.0216	0.0072	0.0048	0.0120	0.0168	0.0072	0.0024	0.0096	0.0024	0.0120	0.0036		
T	0.0360	0.0120	0.0080	0.0200	0.0280	0.0120	0.0040	0.0160	0.0040	0.0200	0.0120	0.0100	
V	0.0144	0.0048	0.0032	0.0080	0.0112	0.0048	0.0016	0.0064	0.0016	0.0080	0.0048	0.0080	0.0016



Example



$$_{AB} = \text{Round}\left(2\log_2\left(\frac{q_{AB}}{e_{AB}}\right)\right)$$

Matrix of scores

	A	D	E	F	G	L	N	P	Q	R	S	T	V
A	5												
D		6											
E		7	5										
F				7									
G					4								
L					5	3							
N													
P	1							4					
Q							7						
R										7			
S							7		7		6		
T												7	
V								6					5

PAM

PAM, or Point Accepted Mutation, defines the probability of a replacement of one amino acid by another through natural selection.

The index on the PAM matrix defines the number of mutations per 100 amino acids.

We give you the PAM1 matrix. How would you reach the PAMx knowing that x represents the number of mutations events ?

Write a function *create_PAM(file,x)* reading the PAM1 matrix file and returning the PAMx matrix.