Homework #8

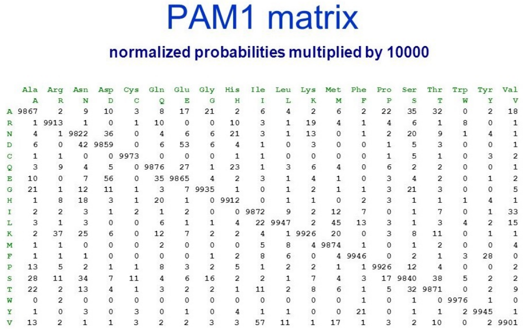
Lupașcu Marian
Syntactic Modeling of Biological Systems
University of Bucharest

May 9, 2020

Context

A substitution matrix is a stochastic matrix that contains the probability that a nucleotide or an amino acid will change into another nucleotide, respectively another amino acid. They have dimensions of 4 by 4, in the case of nucleotide alignment matrices and 20 by 20 in the case of amino acid alignment matrices. In addition, they are symmetrical matrices, so it makes sense to retain either the information above the main diagonal, including the diagonal, or below the diagonal. Because a codon (a 3-nucleotide sequence) is associated with an amino acid then an amino acid substitution matrix is more relevant and encompasses more information than one per nucleotide. In the following we will only talk about amino acid substitution matrices.[3]

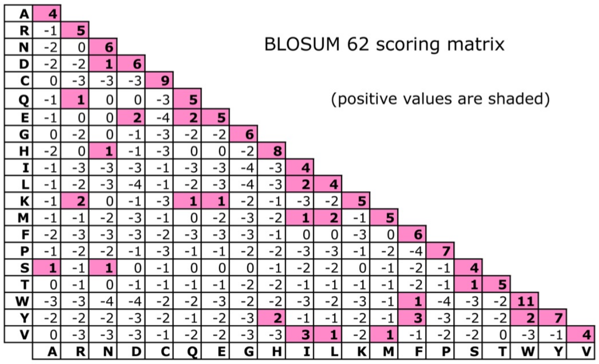
The score in a substitution matrix can be calculated according to several principles: either based on the physico-chemical properties of amino acids (size, hydrophobicity, polarization, etc.), or based on an evolutionary model. That is, a phytocenosis tree is inspected and genetic changes are observed from an evolutionary point of view (which amino acid has changed into another amino acid). If we refer to the physico-chemical properties, for example it is more likely that Alanine (hydrophobic amino acid, without benzyl radicals) will be replaced by Valine (also hydrophobic and without benzyl groups), to the detriment of Histidine (hydrophilic with benzyl radicals) or even Phenylalanine (hydrophobic with benzyl radical). There are a number of constructions for such substitution matrices. Among them are the identity matrix (this allows the evolution change of an amino acid in itself and that's it, which is not true), PAM, BLOSUM, etc.[1,2,3]



PAM1 matrix

normalized probabilities multiplied by 10000

	Ala A	Arg R	Asn N	Asp D	Cys C	Gln Q	Glu E	Gly G	His H	Ile I	Leu L	Lys K	Met M	Phe F	Pro P	Ser S	Thr T	Trp W	Tyr Y	Val V
A	9867	2	9	10	3	8	17	21	2	6	4	2	6	2	22	35	32	0	2	18
R	1	9913	1	0	1	10	0	0	10	3	1	19	4	1	4	6	1	8	0	1
N	4	1	9822	36	0	4	6	6	21	3	1	13	0	1	2	20	9	1	4	1
D	6	0	42	9859	0	6	53	6	4	1	0	3	0	0	1	5	3	0	0	1
C	1	1	0	0	9973	0	0	0	1	1	0	0	0	0	1	5	1	0	3	2
Q	3	9	4	5	0	9876	27	1	23	1	3	6	4	0	6	2	2	0	0	1
E	10	0	7	56	0	35	9865	4	2	3	1	4	1	0	3	4	2	0	1	2
G	21	1	12	11	1	3	7	9935	1	0	1	2	1	1	3	21	3	0	0	5
H	1	8	18	3	1	20	1	0	9912	0	1	1	0	2	3	1	1	1	4	1
I	2	2	3	1	2	1	2	0	0	9872	9	2	12	7	0	1	7	0	1	33
L	3	1	3	0	0	6	1	1	4	22	9947	2	45	13	3	1	3	4	2	15
K	2	37	25	6	0	12	7	2	2	4	1	9926	20	0	3	8	11	0	1	1
M	1	1	0	0	0	2	0	0	0	5	8	4	9874	1	0	1	2	0	0	4
F	1	1	1	0	0	0	0	1	2	8	6	0	4	9946	0	2	1	3	28	0
P	13	5	2	1	1	8	3	2	5	1	2	2	1	1	9926	12	4	0	0	2
S	28	11	34	7	11	4	6	16	2	2	1	7	4	3	17	9840	38	5	2	2
T	22	2	13	4	1	3	2	2	1	11	2	8	6	1	5	32	9871	0	2	9
W	0	2	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	9976	1	0
Y	1	0	3	0	3	0	1	0	4	1	1	0	0	21	0	1	1	2	9945	1
V	13	2	1	1	3	2	2	3	3	57	11	1	17	1	3	2	10	0	2	9901



BLOSUM 62 scoring matrix

(positive values are shaded)

A	**4**																			
R	-1	**5**																		
N	-2	0	**6**																	
D	-2	-2	**1**	**6**																
C	0	-3	-3	-3	**9**															
Q	-1	**1**	0	0	-3	**5**														
E	-1	0	0	**2**	-4	**2**	**5**													
G	0	-2	0	-1	-3	-2	-2	**6**												
H	-2	0	**1**	-1	-3	0	0	-2	**8**											
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	**4**										
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	**2**	**4**									
K	-1	**2**	0	-1	-3	**1**	**1**	-2	-1	-3	-2	**5**								
M	-1	-1	-2	-3	-1	0	-2	-3	-2	**1**	**2**	-1	**5**							
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	**6**						
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	**7**					
S	**1**	-1	**1**	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	**4**				
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	**1**	**5**			
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	**1**	-4	-3	-2	**11**		
Y	-2	-2	-2	-3	-2	-1	-2	-3	**2**	-1	-1	-2	-1	**3**	-3	-2	-2	**2**	**7**	
V	0	-3	-3	-3	-1	-2	-2	-3	-3	**3**	**1**	-2	**1**	-1	-2	-2	0	-3	-1	**4**
	A	**R**	**N**	**D**	**C**	**Q**	**E**	**G**	**H**	**I**	**L**	**K**	**M**	**F**	**P**	**S**	**T**	**W**	**Y**	**V**

PAM

When constructing a PAM or Point Accepted Mutations matrix, we start from amino acid sequences that are very similar (at most 15% different amino acids). And it is based on the principle of moving to a fixed point that is not fatal (ie it is accepted). It was first proposed in 1978, and M (a, b) represents the probability that a will change to b or vice versa at a stage of evolution. PAM_1 is defined as the substitution matrix with an average number of amino acid substitutions of 1%, in a single step. For example PAM_m presented the substitution matrix in m stages. The PAM_1 matrix is used as a calculation basis for other PAM matrices. Currently, more thean 500 such PAM matrices are calculated.

Because it is assumed that a minor change in amino acids from a few generations and a major change in amino acids comes from a larger number of generations we have that: when aligning similar amino acid sequences is used in small m, usually 30, and for alignment amino acid sequences with a high degree of dissimilarity use a large m (usually 70). In addition, if the alignment of divergent amino acid sequences is desired, then the PAM substitution matrices do not give very good results, as they are based on the evolutionary model, and from an evolutionary model in a maximum number of 250 generations no sequences can appear. divergent amino acids. To solve this case, BLOSUM matrices were introduced.[1]

BLOSUM

BLOSUM matrices or BLocks of Amino Acid Substitution Matrix were introduced in 1992 to solve a number of problems and restrictions of PAM matrices. The idea behind these matrices is the notion of block, ie identical amino acid sequences that have been found in multiple proteins. These sequences are of major functional importance so they are preserved from one generation to the next. The construction of BLOSUM matrices involves a clustering of proteins (according to a certain threshold, given by the index of the BLOSUM matrix) and implicitly of amino acid sequences.

For example, the $BLOSUM_{62}$ matrix which is set threshold at 62% assumes a clustering of amino acid sequences so that these sequences have a degree of similarity (identical amino acids in identical positions) of at least 62%. Within the same cluster the alignment score is lower. The $BLOSUM_{62}$ matrix has the power to detect similarities in distant sequences, something that PAM matrices cannot detect. In addition to this matrix, $BLOSUM_{62}$ is used in applications such as BLAST or others.

Conversely, in the case of PAM matrices for an index, it assumes that the amino acid sequences are strongly correlated and a small index assumes that the sequences have a low degree of similarity. That is, if we study two sequences that are supposed to have a high degree of similarity, then a PAM_{30} makes sense compared to a PAM_{70}, and a $BLOSUM_{62}$ makes more sense than a $BLOSUM_{30}$.[2,3]

References

[1] Scoring matrix, Bioinformatics
https://www.bioinformatics.org/wiki/Scoring$_matrix$

[2] Supratim Choudhuri, Sequence Alignment and Similarity Searching in Genomic Databases, Bioinformatics for Beginners, 2014
https://www.sciencedirect.com/science/article/pii/B9780124104716000062

[3] Arcady R.Mushegian Finding Sequence Similarities, Foundations of Comparative Genomics, 2007
https://www.sciencedirect.com/science/article/pii/B9780120887941500021