Université de Lille

Scoring matrices

Sylvain.legrand@univ-lille.fr

Introduction

Definitions

- **Homology**: Property of two sequences that have a shared ancestor → Homology is true or false

- **% Identity**: Percentage of identical residues in an alignment → Used for amino acids or nucleotides

- **% Similarity**: Percentage of amino acid residues in an alignment with a positive substitution score → Not used for DNA

Scoring matrix

- DNA and an amino-acid scoring matrices are **4x4** and **20x20** tables, respectively
- The position X,Y in the table gives **the score of aligning** nucleotide/amino-acid **X** with nucleotide/ amino-acid **Y**
- Involved in all the analyses of **comparison of sequences** (DNA/proteins)
- Alignments are **matrix-dependent**
- Implicitly represent a **theory of evolution** (protein matrices)
- **Understanding** a matrix enables a **good choice** of matrix

DNA matrices

DNA scoring matrices

- ***Ad hoc*** matrix is often given by the software

- The **identity matrix** is the most often used for scoring DNA sequence alignments

	A	C	G	T
A	1			
C	-1	1		
G	-1	-1	1	
T	-1	-1	-1	1

- Not all nucleotide substitutions are equally likely: **transitions** occur about twice as often as **transversions**

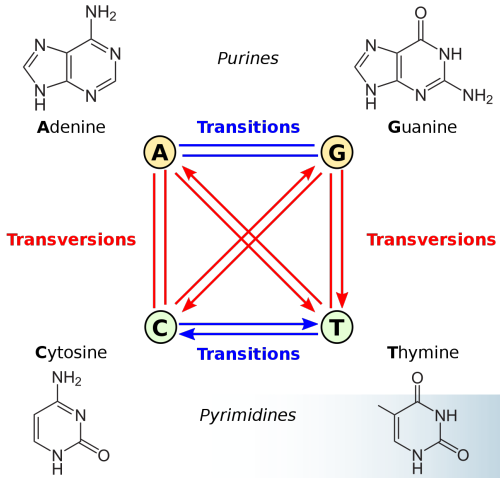


Figure from Wikipedia

	A	C	G	T
A	1			
C	-2	1		
G	-1	-2	1	
T	-2	-1	-2	1

Why are so many DNA scoring matrices?

- Each scoring scheme is to **optimize alignment scoring** to a specific sequence **similarity**

- Ex: the match/mismatch score 1/-4 optimizes the scoring for 100% identical sequences; 1/-1 for 75% identical sequences

- So, you should **choose the scoring scheme** that is **close** to your desired **sequence identity**

- See: https://bioinformaticshome.com/online_software/evaluateDNAscoring/evaluateDNAscoring.html

- There are numerous possibilities to customize scoring depending on **what your goal is**: example, for sequence assembly, you would like to find 100% identical sequences

How to obtain DNA scoring schemes ?

- Use of **log odds ratio**

$$Odds\ for\ match = \frac{P_{Match}}{P_{Random}} \qquad \longrightarrow \qquad Match = \log_2\left(\frac{P_{Match}}{P_{Random}}\right)$$

$$Odds\ for\ mismatch = \frac{P_{Mismatch}}{P_{Random}} \qquad Mismatch = \log_2\left(\frac{P_{Mismatch}}{P_{Random}}\right)$$

- $\mathbf{P_{Match}}$: The probability that two identical nucleotides are aligned by descent

- $\mathbf{P_{Random}}$: The probability that these are aligned by chance

- **Example**:
  - Match A against A
  Match(A-A)=P(A-A)/P(A)xP(A)

  - Mismatch A against C
  Mismatch(A-C)=P(A-C)/P(A)xP(C)

How to obtain DNA scoring schemes ?

- **Example** with **75% identity**

  - $\mathbf{P_{Match}}$: the total match probability is 0.75; so the probability for a specific base: 0.75/4=0.1875

  - $\mathbf{P_{Mismatch}}$: the total mismatch probability is 1-0.75=0.25 Probability for a specific mismatch : 0.25/12=0.0208333 (mismatches for A : A-C; A-G; A-T)

  - $\mathbf{P_{Random}}$: we assume an equal random probability for each base, then the probability of a random base is 0.25 and the probability of a base pair matching is 0.0625 (0.25x0.25)

  - **Odds match**=0.1875/0.0625=3 → **Match**=$\log_2(3)$=1.6

  - **Odds mismatch**=0.333333 → **Mismatch**=$\log_2(0.333333)$=-1.6

Source
https://bioinformaticshome.com/bioinformatics_tutorials/sequence_alignment/DNA_scoring_matrices.html

Protein matrices

Different kind of matrices

- **Identity matrix**: Exact matches receive one score and non-exact matches a different score

- **Mutation data matrix**: a scoring matrix compiled based on observation of protein mutation rates → some mutations are observed more often then other (PAM, BLOSUM)

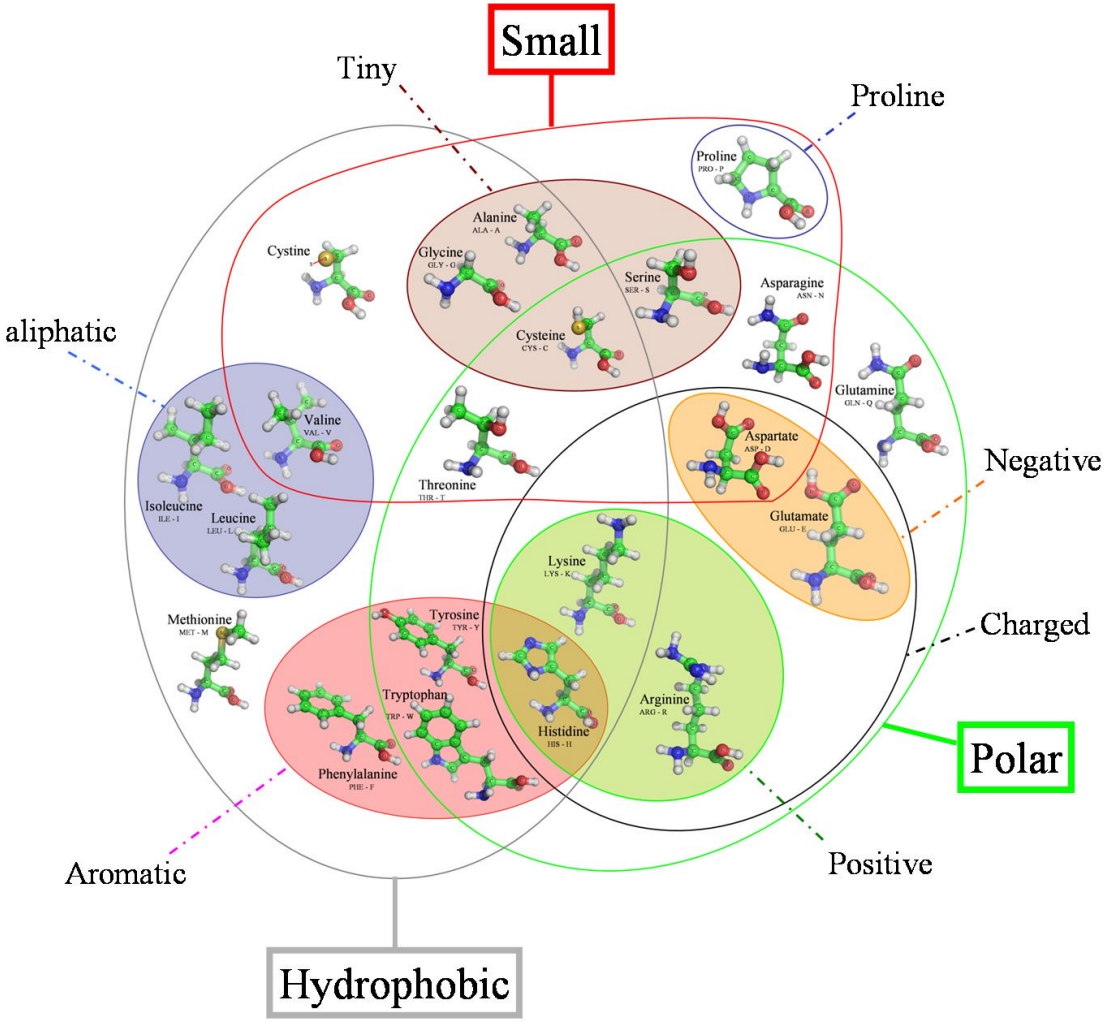
- **Physical properties matrix**: amino acids with with similar biophysical properties receive high score

- **Genetic code matrix**: amino acids are scored based on similarities in the coding triple

BLOSUM62

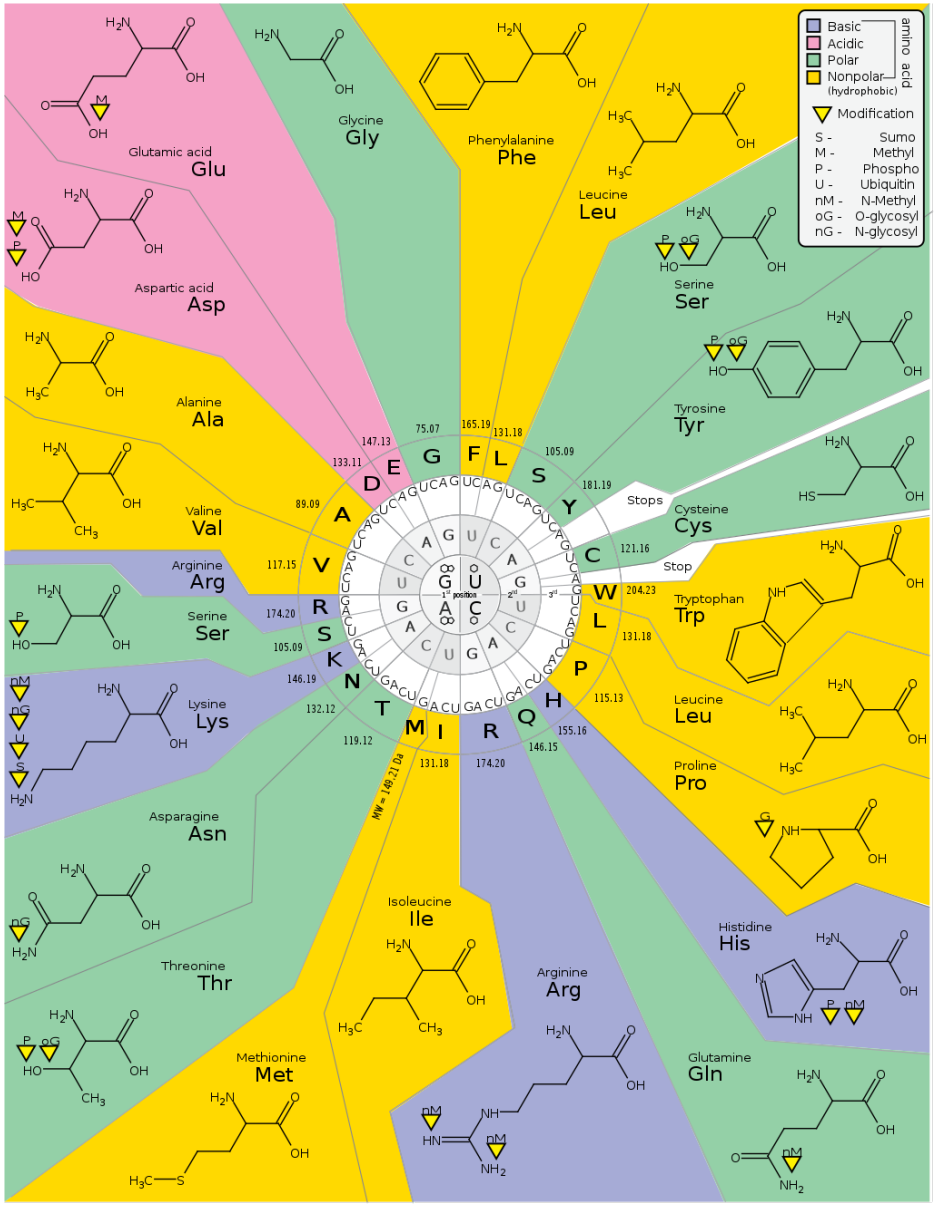
```
#  Matrix made by matblas from blosum62.iij
#  * column uses minimum score
#  BLOSUM Clustered Scoring Matrix in 1/2 Bit Units
#  Blocks Database = /data/blocks_5.0/blocks.dat
#  Cluster Percentage: >= 62
#  Entropy =  0.6979, Expected =  -0.5209
   A  R  N  D  C  Q  E  G  H  I  L  K  M  F  P  S  T  W  Y  V  B  Z  X  *
A  4 -1 -2 -2  0 -1 -1  0 -2 -1 -1 -1 -1 -2 -1  1  0 -3 -2  0 -2 -1  0 -4
R -1  5  0 -2 -3  1  0 -2  0 -3 -2  2 -1 -3 -2 -1 -1 -3 -2 -3 -1  0 -1 -4
N -2  0  6  1 -3  0  0  0  1 -3 -3  0 -2 -3 -2  1  0 -4 -2 -3  3  0 -1 -4
D -2 -2  1  6 -3  0  2 -1 -1 -3 -4 -1 -3 -3 -1  0 -1 -4 -3 -3  4  1 -1 -4
C  0 -3 -3 -3  9 -3 -4 -3 -3 -1 -1 -3 -1 -2 -3 -1 -1 -2 -2 -1 -3 -3 -2 -4
Q -1  1  0  0 -3  5  2 -2  0 -3 -2  1  0 -3 -1  0 -1 -2 -1 -2  0  3 -1 -4
E -1  0  0  2 -4  2  5 -2  0 -3 -3  1 -2 -3 -1  0 -1 -3 -2 -2  1  4 -1 -4
G  0 -2  0 -1 -3 -2 -2  6 -2 -4 -4 -2 -3 -3 -2  0 -2 -2 -3 -3 -1 -2 -1 -4
H -2  0  1 -1 -3  0  0 -2  8 -3 -3 -1 -2 -1 -2 -1 -2 -2  2 -3  0  0 -1 -4
I -1 -3 -3 -3 -1 -3 -3 -4 -3  4  2 -3  1  0 -3 -2 -1 -3 -1  3 -3 -3 -1 -4
L -1 -2 -3 -4 -1 -2 -3 -4 -3  2  4 -2  2  0 -3 -2 -1 -2 -1  1 -4 -3 -1 -4
K -1  2  0 -1 -3  1  1 -2 -1 -3 -2  5 -1 -3 -1  0 -1 -3 -2 -2  0  1 -1 -4
M -1 -1 -2 -3 -1  0 -2 -3 -2  1  2 -1  5  0 -2 -1 -1 -1 -1  1 -3 -1 -1 -4
F -2 -3 -3 -3 -2 -3 -3 -3 -1  0  0 -3  0  6 -4 -2 -2  1  3 -1 -3 -3 -1 -4
P -1 -2 -2 -1 -3 -1 -1 -2 -2 -3 -3 -1 -2 -4  7 -1 -1 -4 -3 -2 -2 -1 -2 -4
S  1 -1  1  0 -1  0  0  0 -1 -2 -2  0 -1 -2 -1  4  1 -3 -2 -2  0  0  0 -4
T  0 -1  0 -1 -1 -1 -1 -2 -2 -1 -1 -1 -1 -2 -1  1  5 -2 -2  0 -1 -1  0 -4
W -3 -3 -4 -4 -2 -2 -3 -2 -2 -3 -2 -3 -1  1 -4 -3 -2 11  2 -3 -4 -3 -2 -4
Y -2 -2 -2 -3 -2 -1 -2 -3  2 -1 -1 -2 -1  3 -3 -2 -2  2  7 -1 -3 -2 -1 -4
V  0 -3 -3 -3 -1 -2 -2 -3 -3  3  1 -2  1 -1 -2 -2  0 -3 -1  4 -3 -2 -1 -4
B -2 -1  3  4 -3  0  1 -1  0 -3 -4  0 -3 -3 -2  0 -1 -4 -3 -3  4  1 -1 -4
Z -1  0  0  1 -3  3  4 -2  0 -3 -3  1 -1 -3 -1  0 -1 -3 -2 -2  1  4 -1 -4
X  0 -1 -1 -1 -2 -1 -1 -1 -1 -1 -1 -1 -1 -1 -2  0  0 -2 -1 -1 -1 -1 -1 -4
* -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4  1
```

Venn diagram of amino acid properties



https://96954219-a-62cb3a1a-s-sites.googlegroups.com/site/apodtele/

Genetic code



Wikipedia

Mutation data matrices

- **log odds ratio** matrices

$$S_{ij} = \log \frac{q_{ij}}{p_i p_j}$$

- expresses the ratio between:
  - The frequency that two residuals *i* and *j* are aligned by descent
  - The probability that these are aligned by chance

- Explanation:
  - ***qij*** = the frequency that the alignment of *i* and *j* is observed in homologous sequences
  - ***pi*** and ***pj***= the frequency of occurrence of *i* and *j*, respectively

  - a score is > 0 if the probability of a significant match is > to the probability of a random match

⇒ **PAM** and **BLOSUM** matrices

PAM matrices

- In 1978, **Margret Dayhoff** performed **global** protein sequence alignments (71 protein groups of closely related proteins (85% identities), and counts the number of substitutions between each pair of amino acids

- She obtained the counts shown in the matrix below

		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y
A	Ala																			
R	Arg	30																		
N	Asn	109	17																	
D	Asp	154	0	532																
C	Cys	33	10	0	0															
Q	Gln	93	120	50	76	0														
E	Glu	266	0	94	831	0	422													
G	Gly	579	10	156	162	10	30	112												
H	His	21	100	226	43	10	243	23	10											
I	Ile	66	30	36	13	17	8	35	0	3										
L	Leu	95	17	37	0	0	75	15	17	40	253									
K	Lys	57	477	322	85	0	147	104	60	23	43	39								
M	Met	29	17	0	0	0	20	7	7	0	57	207	90							
F	Phe	20	7	7	0	0	0	0	17	20	90	167	0	17						
P	Pro	345	67	27	10	10	93	40	49	50	7	43	43	4	7					
S	Ser	772	137	432	98	117	47	86	450	26	20	32	168	20	40	269				
T	Thr	590	20	169	57	10	37	31	50	14	129	52	200	28	10	73	696			
W	Trp	0	27	3	0	0	0	0	0	3	0	13	0	0	10	0	17	0		
Y	Tyr	20	3	36	0	30	0	10	0	40	13	23	10	0	260	0	22	23	6	
V	Val	365	20	13	17	33	27	37	97	30	661	303	17	77	10	50	43	186	0	17
		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y
		Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val

M. O. Dayhoff , R. M. Schwartz, A model of evolutionary change in proteins (1978)

- Then she used this count matrix to derive “point matrices accepted mutations” (PAM)

PAM matrices

- **Scores** are calculated as **log-odds**

$$S_{ij} = \log \frac{q_{ij}}{p_i p_j}$$

- **Positive values** reflect **frequent** substitutions ("accepted by natural selection"), *i.e.* substitutions observed more frequently than expected by chance

- **Negative values** reflects **rare** mutations, *i.e.* those that are observed less frequently than expected by chance

- The diagonal reflects residue conservation

	C	S	T	P	A	G	...
C	11.5						...
S	0.1	2.2					...
T	-0.5	1.5	2.5				...
P	-3.1	0.4	0.1	7.6			...
A	0.5	1.1	0.6	0.3	2.4		...
G	-2.0	0.4	-1.1	-1.6	0.5	1.6	...
...	...	...	...	...	...	...	...

PAM matrices

- The alignments carried out by Margret Dayhoff in 1978 had an **average identity of ~85%**

- However, **the frequency of substitutions depends on the degree of divergence** between sequences, as the number of substitutions increases over time.

- To account for the rate of divergence, Margret Dayhoof calculated a **series of score matrices**, each **reflecting** a certain **substitution rate**

  - **PAM001** substitution rate between amino acids at the end of an evolutionary time resulting in ~1% substitutions per position.
  - **PAM050** idem with 50% mutations /position
  - **PAM250** idem with 250% mutations/position (note: the same position can be subject to several successive mutations)

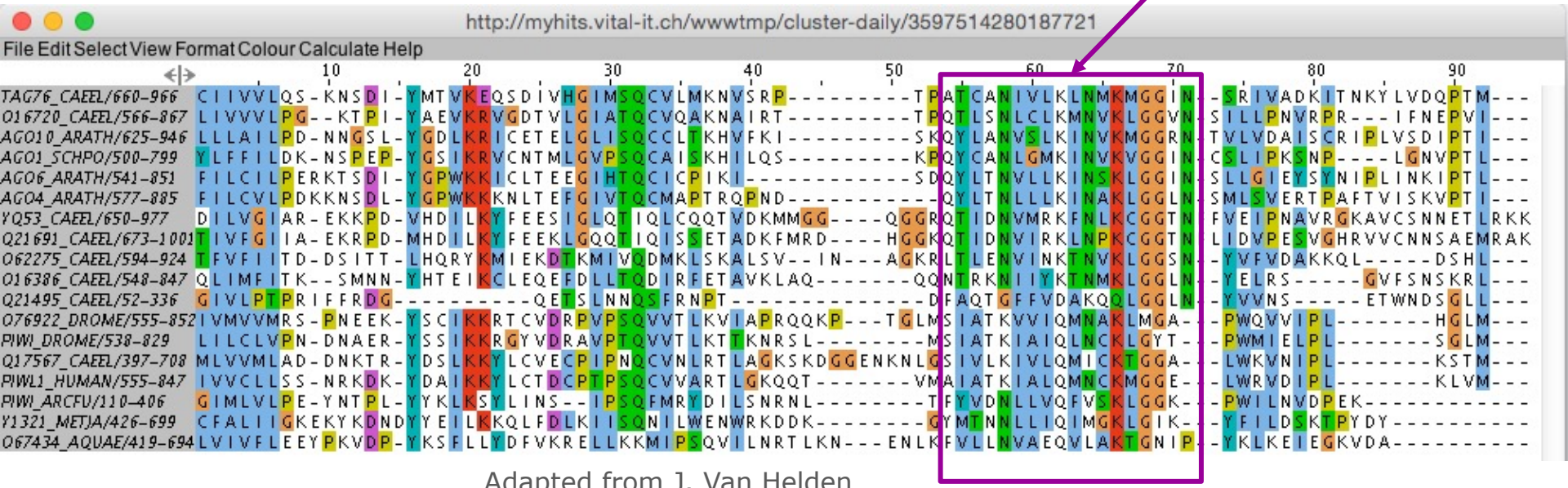
- When making an alignment, one **must choos**e one of the matrices of this series, taking into account the rate of difference between the two sequences that one wants to align.

BLOSUM matrices

- **Henikoff and Henikoff** (1992) analysed the frequencies of substitutions in blocks from multiple alignments generated from a large number of protein families (blocks)

- They derived the "BLOSUM" series of matrices, which correspond to different rates of evolutionary conservation between sequences.

Multiple alignement block



Adapted from J. Van Helden

BLOSUM matrices

- Examples
  - The **BLOSUM62** matrix was calculated from blocks of ≥62% identity
  - The **BLOSUM80** matrix was calculated from blocks of ≥80% identity

- When aligning sequences, **the most appropriate matrix** should always be **chosen**, based on the percentage of similarity

- The problem is that before the alignment is carried out, this percentage is not known. **How can this circularity be resolved?**

  - A first alignment is carried out with a "medium" matrix (BLOSUM62).
  - The % identity is observed in this alignment
  - The matrix whose index is closest to this % is then chosen.
  - Alignment is redone with the new matrix

Summary

- **Different substitution scoring matrices** have been established
  - Residue categories (Phylip)PAM (Dayhoff, 1979).
  - PAM means “Percent Accepted Mutations”
  - BLOSUM (Henikoff & Henikoff, 1992) BLOSUM means “Block sum”

- Substitution matrices **allow to detect similarities between more distant proteins** than what would be detected with the simple identity of residues

- The matrix **must be chosen carefully**, **depending on the expected rate of conservation** between the sequences to be aligned

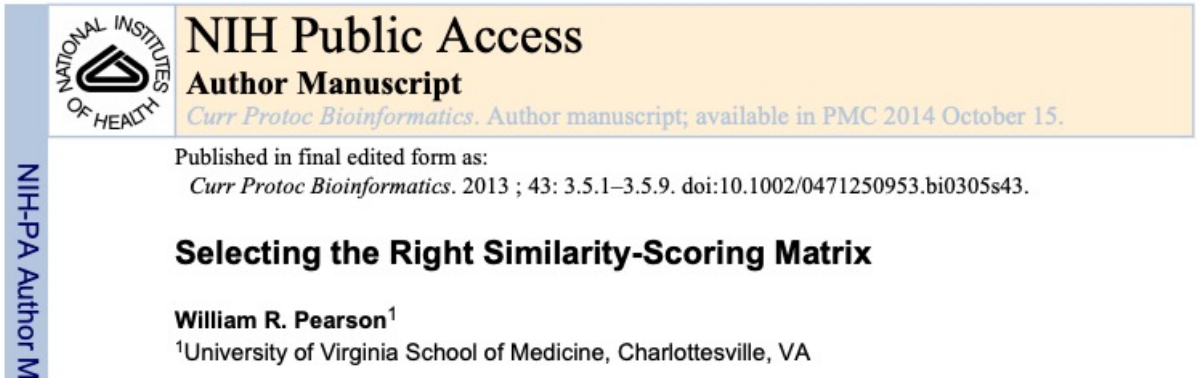
Summary

- With **PAM** matrices, the score indicates the percentage of substitution per position -> **higher numbers** are appropriate for **more distant** proteins

- With **BLOSUM** matrices, the score indicates the percentage of conservation -> **higher numbers** are appropriate for **more conserved** proteins

	% ≠ observé	dist. évolutive PAM
BLOSUM-80	1	1
PAM-1	5	5
faible divergence	10	11
	15	17
	20	23
	25	30
	30	38
BLOSUM-62	34	47
PAM-120	40	56
	45	67
	50	80
	55	94
	60	112
	65	133
BLOSUM-45	70	159
PAM-250	75	195
forte divergence	80	246
	85	328

References


NIH Public Access
Author Manuscript
Curr Protoc Bioinformatics. Author manuscript; available in PMC 2014 October 15.

Published in final edited form as:
Curr Protoc Bioinformatics. 2013 ; 43: 3.5.1–3.5.9. doi:10.1002/0471250953.bi0305s43.

Selecting the Right Similarity-Scoring Matrix

William R. Pearson[1]
[1]University of Virginia School of Medicine, Charlottesville, VA

https://bioinformaticshome.com/bioinformatics_tutorials/sequence_alignment/DNA_scoring_matrices.html

Course from Jacques Van Helden:
http://pedagogix-tagc.univ-mrs.fr/courses/bioinfo_intro/pdf_files/03.01.matrices_de_substitution_fr_6ppf.pdf

http://tbb.bio.uu.nl/BDA/2017/20170221_quantifying_sequence_similarity.pdf

PAM
Dayhoff, M. O., Schwartz, R. M. & Orcutt, B. (1978). A model of evolutionary change in proteins. Atlas of Protein Sequence and Structure 5, 345--352.

BLOSUM
Henikoff, S. & Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci U S A 89, 10915-9

Sylvain Legrand
Maître de Conférences
UMR CNRS 8198 EVO-ECO-PALEO
Evolution, Ecologie et Paléontologie
Université de Lille - Faculté des Sciences et Technologies
Bât SN2, bureau 208 - 59655 Villeneuve d'Ascq

sylvain.legrand@univ-lille.fr | http://eep.univ-lille.fr/
Tél. +33 (0)3 20 43 40 16