Bioinformatique M1: Lecture 3

P. Derreumaux

MATRICES DE SUBSTITUTION:

1. Definitions and Generalities

Sequence Comparison

- Generally, sequence determines structure and structure determines function
- By studying sequence similarity, we hope to find correlations between our sequence and other sequences with known structure or function
- This approach is often successful, however, many molecules have low sequence similarity, yet still share similar structure or function

Terminology

- Homology
  - Common ancestor
  - Homology is not a measurable quantity even if % identity ≥ 30% accepted
- Identity
  - Objective and well defined
  - % = $\frac{\text{nbre matches}}{\text{nbre total}}$
- Similarity
  - f(algorithms, scoring matrix)

The Problem

Sequence Matching

Dynamic Programming Alignment

- Calculation of alignment score using affine gap model

```
BOVGH  G G T G C C A C - - T C C C A - - - - - C T G
         : : : : : : : :   : :   :             : : :
R02321 A G T G C C A C C   C C C A A T G C C G C T G
      -3+4+4+4+4+4+4+4-12-3+4+4-3+4  -12-4x4 +4+4+4
```

matches	52
mismatches	-9
gaps	-40
Overall Score	3

	A	C	G	T
A	4			
C	-3	4		
G	-3	-3	4	
T	-3	-3	-3	4

Gap opening = -12
Gap extension = -4

Scoring Matrices

- They are used to assign a score to each comparison of a pair of characters
- positive scores for identical or similar character pairs; negative scores for dissimilar pairs.
- the matrix is symmetric

- Score

S = Total Score

$S(i,j)$ = similarity matrix for aligning i and j

Sum is carried out over all aligned i and j

n = number of gaps

P = gap penalty

$$S = \sum_{i,j} S(i,j) + \sum_{k=1}^{n} P$$

$$P = a + bN$$

a : cost of opening a gap

bN : cost of extending the gap of length N

N.B. $S(i,j)$, a, bN ?

- **Remarks**

1. Total score $S = f$ (number characters)

2. $S = f$ (gap, matrix)

3. Understanding theories underlying scoring matrix is essential

4. Keep in mind.
   evolution rate = f (species)
   = f (regions for each protein / ADN)

Next: Where do matrices come from??

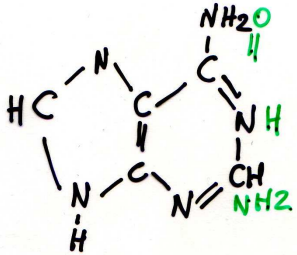
2. Nucleic Matrices

Use protein sequence rather than nucleotide sequence when possible (redundancy in genetic code with up to six codons translated into the same amino acid)

TTA
TTG
CTT
CTC
CTA
CTG

→ Leu (idem for Arg)

Purine bases



Adenine (Guanine)

Fig.1 Two DNA sequences

CGCCGT-AT

CG-CC-TA-

Fig.2 DNA similarity matrices

1. Identity matrix (similarity)

	A	T	C	G
A	1	0	0	0
T	0	1	0	0
C	0	0	1	0
G	0	0	0	1

2. Transition/Transversion Matrix

	A	T	C	G
A	1	0	-5	-5	1
T	-5	1	0	1	-5
C	-5	1	1	0	-5
G	1	-5	-5	1	0

transition : ring number conserved (cycle)

A → G

C → T

3. Nucleic Acid PAM 50 Scores, 3 to 1 Transition to Transversion Ratio

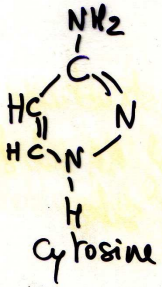
A	1.36			
G	-0.37	1.36		
C	-1.60	-1.60	1.36	
T	-1.60	-1.60	-0.37	1.36

transversion : ring number not conserved

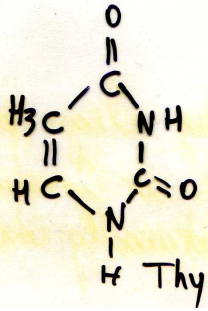
A → C or A → T

etc...

Pyrimidine bases



Cytosine



Thymine

3. MATRICES PROTEIQUES LIEES AU CODE GENETIQUE et AUX PROPRIETES CHIMIQUES.

- principe de la dégénérescence du code génétique (FITCH) $AA_i \longrightarrow AA_j$ number of nucleotide changes

- f (propriétés physico-chimiques) (FENG, Mc Lachlan) hydrophobe / hydrophile

J. Mol. Evol. 21 : 112 (1985)

4. MATRICES PROTEIQUES DERIVEES DES STRUCTURES

3D

- Risler et al. 32 protéines J. Mol. Biol. 204: 10 (1988)
- Johnson et Overington. 235 protéines
  - manually align protein structures and look at frequency of AA substitutions at structurally constraint sites

5. EMPIRICAL EVOLUTIONARY MATRICE

- pioneered by Margaret Dayhoff (1970)
- Henikoff and Henikoff proposed an alternative model in 1992
- Both models assume that gaps are not allowed and compute log-odds ratio

$$s(a,b) = \frac{1}{\lambda} \log_2 \left(\frac{freq(O)}{freq(E)} = \frac{p_{ab}}{f_a f_b} \right)$$

Common ancestor hypothesis vs. random hypothesis

λ : scaling factor (BLOSUM 62's original $\lambda = 0.347$

O : observed exchanges

E : observed exchanges

p_{ab} : target frequency : probability to observe residues a and b aligned in homologous sequences alignments.

f_a : probability to observe amino acid a on average in any protein sequence

$$S(a,b) = \frac{1}{\lambda} \log \left(\frac{freq(O)}{freq(E)} \right)$$

Why log-odds ratio ?

1. $+ \longrightarrow$ More likely than random
   $0 \longrightarrow$ At random base rate
   $- \longrightarrow$ less likely than random

2. $\log xyz = \log x + \log y + \log z$
   adding is easier than multiplying.

Fig.3 PAM 250 Amino Acid Similarity Matrix

	Cys	Gly	Pro	Ser	Ala	Thr	Asp	Glu	Asn	Gln	His	Lys	Arg	Val	Met	Ile	Leu	Phe	Tyr	Trp
Cys	12																			
Gly	-3	5																		
Pro	-3	-1	6																	
Ser	0	1	1	1																
Ala	-2	1	1	1	2															
Thr	-2	0	0	1	1	3														
Asp	-5	1	-1	0	0	0	4													
Glu	-5	0	-1	0	0	0	3	4												
Asn	-4	0	-1	1	0	0	2	1	2											
Gln	-5	-1	0	-1	0	-1	2	2	1	4										
His	-3	-2	0	-1	-1	-1	1	1	2	3	6									
Lys	-5	-2	-1	0	-1	0	0	0	1	1	0	5								
Arg	-4	-3	0	0	-2	-1	-1	-1	0	1	2	3	6							
Val	-2	-1	-1	-1	0	0	-2	-2	-2	-2	-2	-2	-2	4						
Met	-5	-3	-2	-2	-1	-1	-3	-2	0	-1	-2	0	0	2	6					
Ile	-2	-3	-2	-1	-1	0	-2	-2	-2	-2	-2	-2	-2	4	2	5				
Leu	-6	-4	-3	-3	-2	-2	-4	-3	-3	-2	-2	-3	-3	2	4	2	6			
Phe	-4	-5	-5	-3	-4	-3	-6	-5	-4	-5	-2	-5	-4	-1	0	1	2	9		
Tyr	0	-5	-5	-3	-3	-3	-4	-4	-2	-4	0	-4	-5	-2	-2	-1	-1	7	10	
Trp	-8	-7	-6	-2	-6	-5	-7	-7	-4	-5	-3	-3	2	-6	-4	-5	-2	0	0	17

- $S_{ij} \times 10$
- S_{ii}

- alignement global
- 34 familles de protéines, 1572 substitutions
- Matrice 1PAM (1 Percent Accepted Mutations) = 99% identity ⟹ identical function
- PAM 250 corresponds to seq ~ 20% identical

Sources of error in PAM model:

1. Replacement is not equally probable over entire sequence.
2. Many sequences depart from average composition.
3. Errors in 1PAM are magnified in the extrapolation to 250 PAM.
4. Rare replacements were observed too infrequently to resolve relative probabilities accurately (
   36 pairs no replacements were observed!).

Ce constat a conduit à une réactualisation de la matrice (PET91, An updated Dayhoff matrix) Jones et al., 1992 en considérant 16 130 séquences issues de la version 15 de Swissprot, ce qui correspond à 2 621 familles de protéines. Cette étude a permis de prendre davantage en compte substitutions qui étaient mal représentées en 1978.

Gonnet et al. have also designed a new matrix with a slightly different (but theoretically equivalent) method.

Fig.4 Blosum 45 Amino Acid Similarity Matrix

	Gly	Pro	Asp	Glu	Asn	His	Gln	Lys	Arg	Ser	Thr	Ala	Met	Val	Ile	Leu	Phe	Tyr	Trp	Cys
Gly	7																			
Pro	-2	9																		
Asp	-1	-1	7																	
Glu	-2	0	2	6																
Asn	0	-2	2	0	6															
His	-2	-2	0	0	1	10														
Gln	-2	-1	0	2	0	1	6													
Lys	-2	-1	0	1	0	-1	1	5												
Arg	-2	-2	-1	0	0	0	1	3	7											
Ser	0	-1	0	0	1	-1	0	-1	-1	4										
Thr	-2	-1	-1	-1	0	-2	-1	-1	-1	2	5									
Ala	0	-1	-2	-1	-1	-2	-1	-1	-2	1	0	5								
Met	-2	-2	-3	-2	-2	0	0	-1	-1	-2	-1	-1	6							
Val	-3	-3	-3	-3	-3	-3	-3	-2	-2	-1	0	0	1	5						
Ile	-4	-2	-4	-3	-2	-3	-2	-3	-3	-2	-1	-1	2	3	5					
Leu	-3	-3	-3	-2	-3	-2	-2	-3	-2	-3	-1	-1	2	1	2	5				
Phe	-3	-3	-4	-3	-2	-2	-4	-3	-2	-2	-1	-2	0	0	0	1	8			
Tyr	-3	-3	-2	-2	-2	2	-1	-1	-1	-2	-1	-2	0	-1	0	0	3	8		
Trp	-2	-3	-4	-3	-4	-3	-2	-2	-2	-4	-3	-2	-2	-3	-2	-2	1	3	15	
Cys	-3	-4	-3	-3	-2	-3	-3	-3	-3	-1	-1	-1	-2	-1	-3	-2	-2	-3	-5	12

BLOSUM (BLOcks SUbstitution Matrix)

- local multiple alignments of related seq without gaps (PROSITE)
- Derivation of the BLOCKS database
- Sequences similar at the some x% are clustered

↳ q_{ij}

↳ BLOSUM-x $= \log_2 \left(\frac{q_{ij}}{P_i P_j} \right)$

(BLOSUM 60 ≠ PAM 60) mean opposite

```
AABCDA...BBCDA
DABCDA.A.BBCBB
BBBCDABA.BCCAA
AAACDAC.DCBCDB
CCBADAB.DBBDCC
AAACAA...BBCCC
```

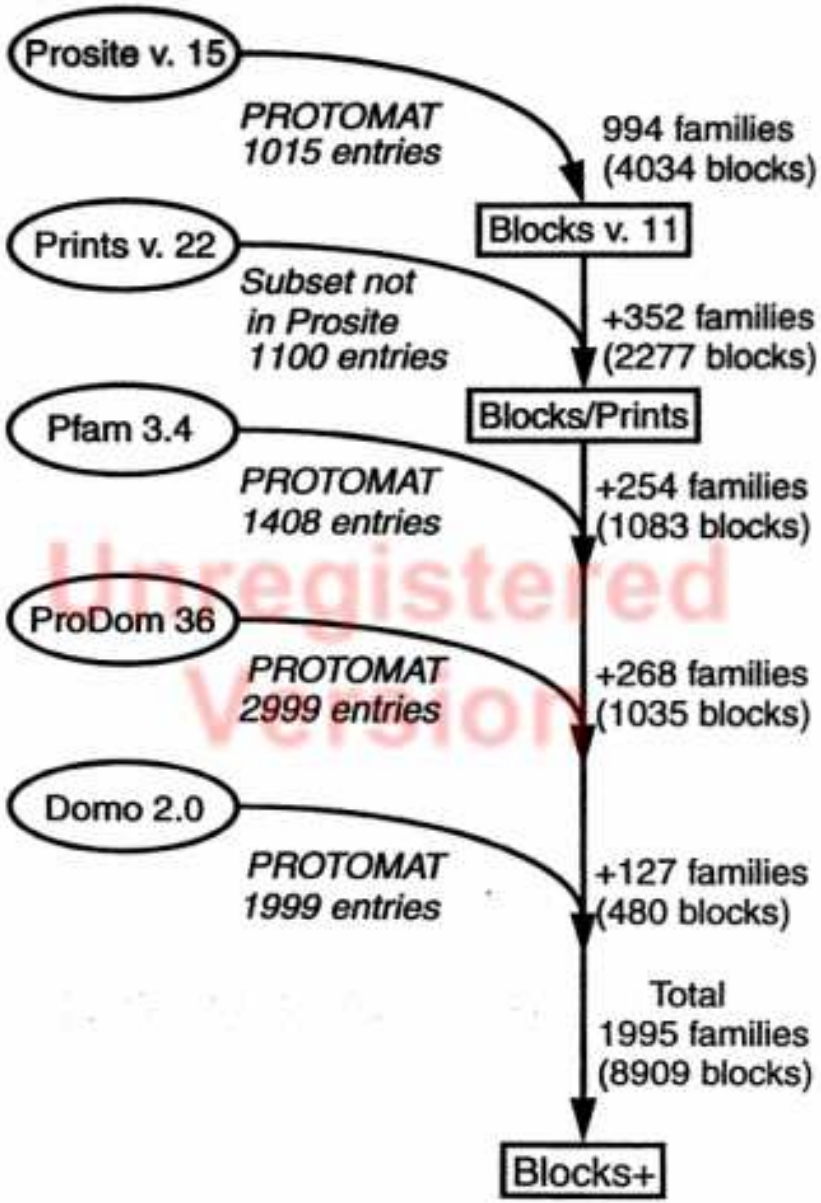


Fig. 1. Flow chart diagram of the hierarchical procedure used to build Blocks+ from constituent databases.

Some details in BLOSUM 62.

- Trypophan (W/W) pair score +11 why leucine (L/L) pair score +4.

Why shouldn't all identities get the same score?

The rarer the amino acid is, the more surprising it would be to set two of them align together by chance

Using the alignment data that BLOSUM was trained on, L/L pairs are more common than W/W pairs ($P_{LL} = 0.0371$; $P_{WW} = 0.0065$)

But W is a much rarer amino acid ($f_L = 0.099$, $f_W = 0.013$). Run those numbers, you get +3.8 for L/L and +10.5 for W/W, which were rounded to +4 and +11.

6. Choix de la Matrice?

→ Score total = f(matrice, gap)
→ Comparaison globale PAM et BLOSUM

Comparable Blosum and PAM Tables

Blosum Tables	(Entropy)	PAM Tables	(Entropy)	Percent Sequence Identity PAM Tables
Blosum 90	(1.18)	PAM 100	(1.18)	43
Blosum 80	(0.99)	PAM 120	(0.98)	38
Blosum 60	(0.66)	PAM 160	(0.70)	30
Blosum 52	(0.52)	PAM 200	(0.51)	25
Blosum 45	(0.38)	PAM 250	(0.36)	20

Shannon's Entropy = $\sum_{i=1}^{20} \sum_{j=1}^{i} q_{ij} s_{ij}$ ($\sum_{i=1}^{k} p_i \log_2 p_i$)

q_{ij} : observed probability that AA i and j are aligned by evolution

BUT differences e.g. BLOSUM is more tolerant than PAM to AA with hydrophobic character.

Quelle matrice de substitution choisir ?

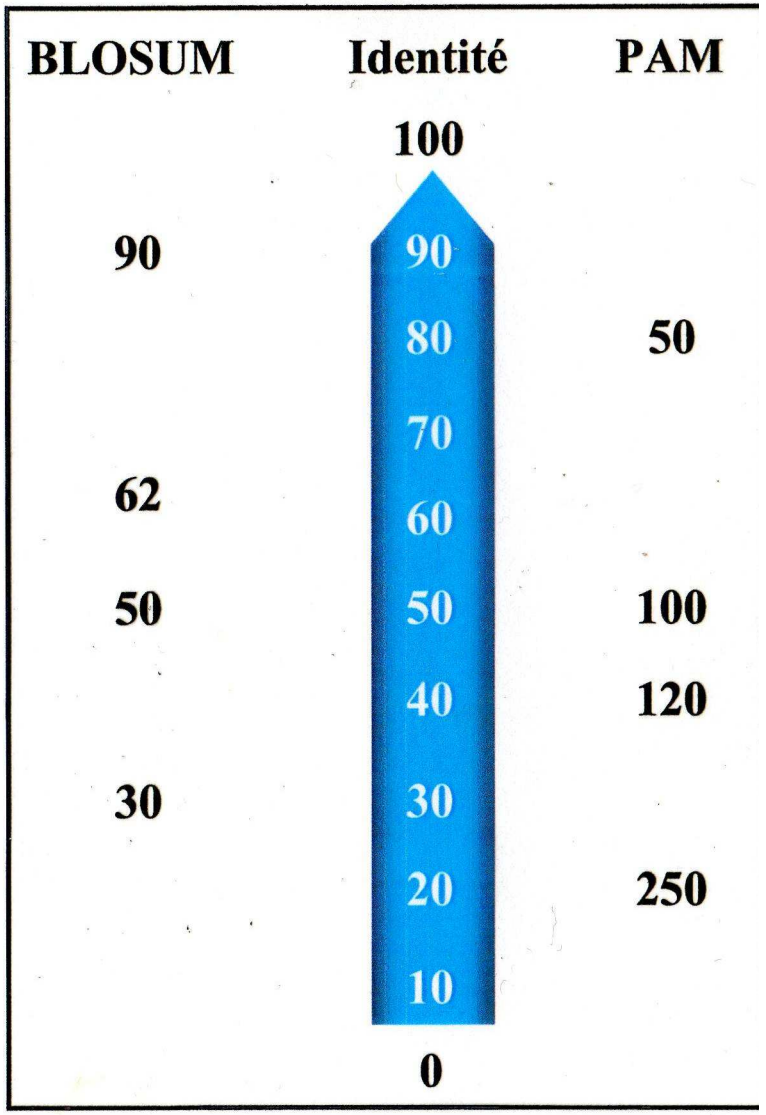
• **Pas** de matrice **idéale (mécanismes d'évolution)**;

• Les matrices dérivées des **mutations observées** donnent, pour les protéines, de **meilleurs résultats** que les matrices basées sur l'identité, le code génétique ou les propriétés physico-chimiques.

• **<u>Matrices PAM établies par M. Dayhoff</u>** (1978) :
  • absence de paires → Matrice Gonnet

• **<u>Matrices BLOSUM</u>** (1992) :
  • construites à partir de plus de données ;
  • BLOSUM62 : matrice par défaut sur serveurs

→Profils, Incorporations de données dérivées des structures quand possible



7. MATRICES PROTEIQUES FONCTIONS DES POSITIONS OU DE L'ENVIRONNEMENT.

% identity ≤ 30% (case of close) 2 AA cons. / 150 AA

- Matrice fonction de l'environnement des AA (accessibilité au solvent, structure secondaire, FISCHER 1996)

$g_{ij} = S_{xy}(\text{PAM, BLOSUM}) + w\, h_{ij}$.

w = cte.

$h_{ij} = v * \text{rel SS}_{ij}$ { v=1 résidus i (query) et j (homolog) ont m structures secondaires.

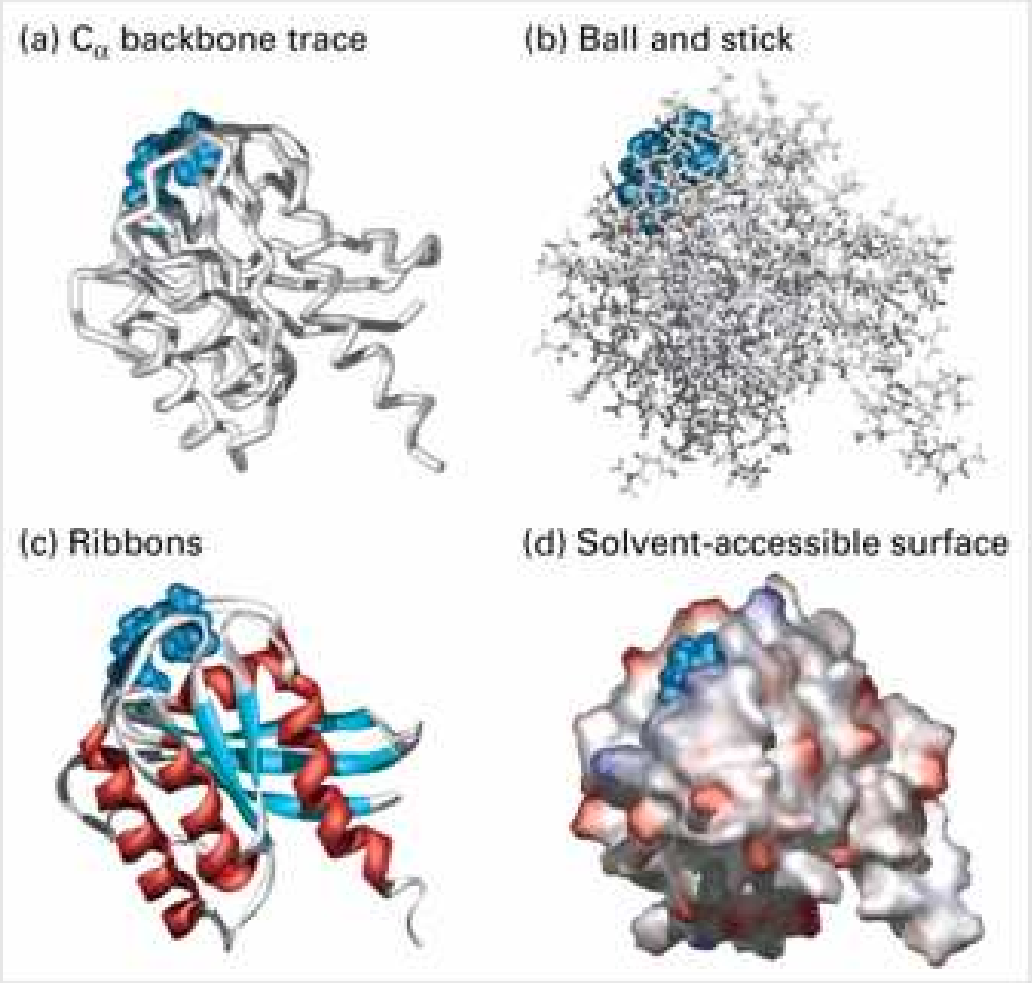
autre matrice H3P2

Matrices utilisées pour reconnaissance de folds. (structure must be known)

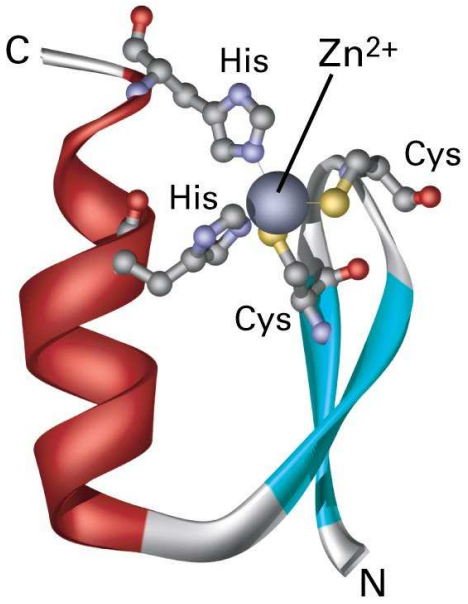
- Matrice fonction de la position des résidus dans la séquence (Gribskov, 1987) = profil.

↳ PSI-BLAST

Matrices dérivées de la structure



(b) Zinc-finger motif



Consensus sequence:

F/Y - C - - C - - - - F/Y - - - - - - - - H - - - H -

Position Specific Scoring Matrices

Sequence Information

- Ligands to Calcium
- Hydrophobic patches that stabilize structure

```
MCEG     FKEAFSLFDKDGDGTITTKELGTVMRSL
MCDO     FKEAFSLFDKDGDGSITTKELGTVMRSL
MCSP     FKEAFSLFDKDGDGCITTKELGTVMRSL
MCCHM    IREAFRVFDKDGNGYISAAELRHVMTNL
MCUR1C   IKAIIQKADANKDGKIDREEFMKLIKS.
MCUR2C   IDAIIKKADGNNDGKIRVQEFVKMIESS
KLBOB    FNKAFELYDQDGDGYIDENELDALLKDL
KLCHI    FNKAFEMYDQDGNGYIDENELDALLKDL
KLBOI    LDELFEELDKNGDGEVSFEEFQVLVKKI
KLPGI    LDDLFQELDKNGNGEVSFEEFQVLVKKI
```

Which matrix to be used?

- PAM results from global alignt
- BLOSUM results from local alignt.

but both matrices can be used for global and local alignment.

Is sequence similarity relevant biologically or due to random?

↳ E-value for score S.
and the various probability distributions