Multiple sequence alignments

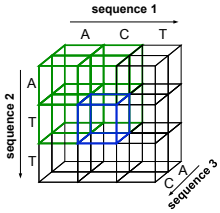
Jacques van Helden

Jacques.van-Helden@univ-amu.fr
Aix-Marseille Université (AMU), France
Lab. Technological Advances for Genomics and Clinics (TAGC, INSERM Unit U1090)
http://tagc.univ-mrs.fr/

FORMER ADDRESS (1999-2011)
Université Libre de Bruxelles, Belgique
Bioinformatique des Génomes et des Réseaux (BiGRe lab)
http://www.bigre.ulb.ac.be/

Dynamical programming - multiple alignment

- Dynamical programming can be extended to treat a set of 3 sequences
  - build a 3-dimensional matrix
  - the best score of each cell is calculated on the basis of the preceding cells in the 3 directions, and a scoring scheme (substitution matrix + gap cost)
- Can be extended to n sequencs by using a n-dimensional hyper-cube
- Problem: matrix size and execution time increase **exponentially** with the number of sequences
  - 2 sequences L1 x L2
  - 3 sequences L1 x L2 x L3
  - 4 sequences L1 x L2 x L3 x L4
  - n sequences L1 x L2 x ... x Ln
- Aligning n sequences with dynamical programming requires $O(L^n)$ operations, which becomes thus very rapidly impractical.
- The efficiency can be improved by only considering a subspace of the n-dimensional matrix. However, even with this kind of algorithmic improvement, the number of sequences that can be aligned is still restricted (~8 sequences maximum).

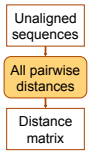


Progressive alignment

- Another approach to align multiple sequences is to perform a progressive alignment. The algorithm proceeds in several steps:
  - Calculate a ***distance matrix***, representing the distance between each pair of sequences.
  - From this matrix, build a ***guide tree*** regrouping the closest sequences first, and the more distant sequences later.
  - Use this tree as guide to progressively align the sequences.
- This is a heuristics
  - it is a practically tractable approach, but it cannot guarantee to return the optimal solution

Progressive alignment – step 1:compute distance matrix

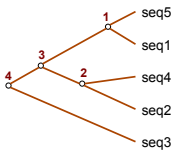
- Perform a pairwise alignment between each pair of sequences (dynamical programming or faster heuristic algorithm).
  - For n sequences: $n*(n-1)/2$ pairwise alignments.
- From each pairwise alignment, calculate the distance between the two sequences.
  - $d_{i,j}=s_{i,j}/L_{i,j}$
    - $d_{i,j}$ distance between sequences i and j
    - $L_{i,j}$ length of the alignment
    - $s_{i,j}$ number of substitutions
- Remarks
  - Gaps are not taken into account in the distance metric.
  - The matrix is symmetrical : $d_{i,j} = d_{j,i}$
  - Diagonal elements are null: $d_{i,i} = 0$

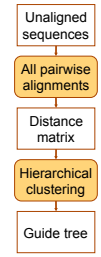


	seq 1	seq 2	...	seq n
seq 1	d1,1	d1,2	...	d1,n
seq 2	d2,1	d2,2	...	d2,n
...	...	...	...	...
seq n	dn,1	dn,2	...	dn,n

Progressive alignment – step 2: build guide tree

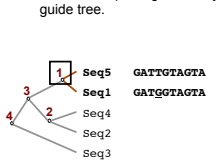
- A tree can be computed from the matrix distance by hierarchical clustering.
  - first regroup the two closest sequences (cluster **1**)
  - next, progressively regroup the closest remaining clusters
    - the two closest sequences (cluster **2**)
    - one cluster with another cluster (cluster **3**)
    - one sequence with a previous cluster (cluster **4**)
- This tree will then be used as ***guide*** to determine the order of incorporation of the sequences in the multiple alignment.
- Beware ! The guide tree should not be interpreted as a phylogenetic tree.
  - Its only purpose will be to identify the closest similarities between sequences in order to build a multiple alignment.

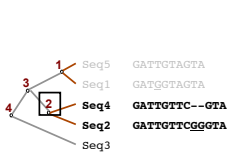


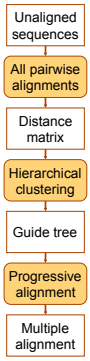


Progressive alignment – step 3: multiple alignment

- Build a multiple alignment, by progressively incorporating the sequences according to the guide tree.

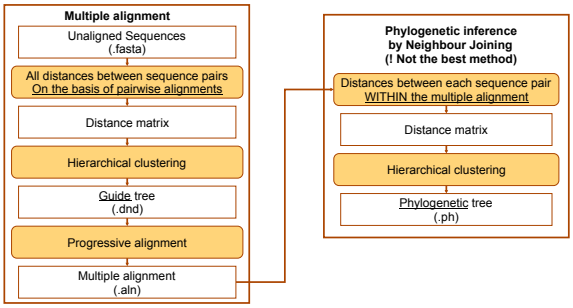




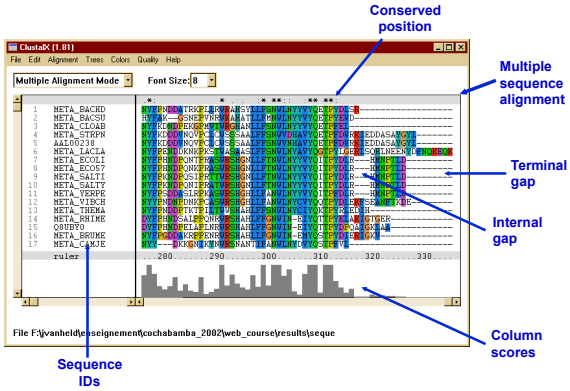


Progressive alignment and Neighbour-Joining (NJ) tree with clustalX

- Attention ! The guide tree is not a phylogenetic tree
  - Its only role is to propose an order of incorporation of the sequences for building the multiple alignment.
  - It does not aim at predicting the evolutionary history of the divergences between sequences.
- In a second time, it is possible to infer a phylogenetic tree from the multiple alignment, using the ***Neighbor Joining (NJ)*** method.
  - However, this method is sub-optimal for phylogenetic inference.

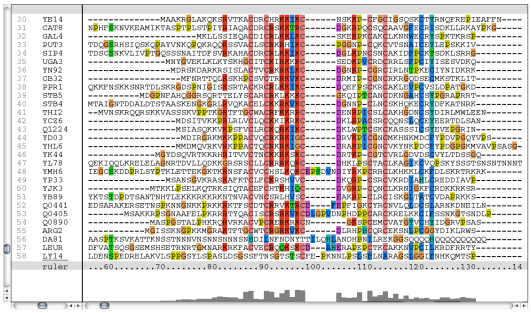


Global multiple alignment : Homoserine-O-dehydrogenase



Alignment of proteins containing a Zinc cluster domain

- The alignment of yeast Zn(2)Cys(6) binuclear cluster proteins is a difficult case.
  - The conserved region is restricted to the Zinc cluster domain.
  - This domain is not contiguous, it contains conserved and variable positions.
  - The alignment highlights 5 of the 6 characteristic cysteins.



Local multiple alignment

Progressive alignment - summary

- Processing time
  - Building the tree: proportional to n x n
  - Aligning sequences: linear with number of sequences
- Heuristic method
  - cannot guarantee to return the optimal alignment.

- clustalX is a window-based environment for clustalw, which provides additional functionalities
  - Mark low scoring segments
  - The alignment can be refined manually
    - Realign selected sequences
    - Realign selected positions

Sequence motifs

Profile matrices (=position-specific scoring matrices, PSSM)

- Starting from a multiple alignment, one can build a matrix which reflects the preferred residues at each position
  - Each column represents a position
  - Each row represents a residue (20 rows for proteins, 4 rows for DNA)
  - The cells indicate the frequency of each residue at each position of the multiple alignment.

Multiple alignment

W	S	K	T	N	V	T	S	T	L	H	I	C	W	G	A	Q	A	G	L
W	S	K	T	N	V	T	S	T	L	N	I	C	W	G	A	Q	A	G	L
W	T	Q	S	H	V	T	H	S	R	T	L	N	I	C	W	A	A	Q	A	A	V
F	L	K	Q	N	V	T	T	S	S	S	M	Y	I	C	W	G	A	M	A	A	L
W	S	V	T	N	V	T	S	S	T	T	I	H	I	C	W	G	A	A	A	G	L
W	S	K	D	H	V	T	S	S	T	L	F	V	C	W	A	V	Q	A	A	G	L
W	S	K	S	H	V	Y	S	S	T	L	F	V	C	W	A	V	Q	A	A	A	L
W	S	K	S	H	V	Y	S	S	T	L	H	I	C	W	G	A	Q	A	A	A	L
W	T	T	S	N	V	H	S	S	T	L	N	V	C	W	G	G	M	A	A	A	V
W	T	K	D	H	V	T	S	S	T	L	F	V	C	W	A	V	Q	A	A	A	L
W	A	K	D	H	V	T	T	S	S	T	L	F	H	I	C	W	A	V	Q	A	A	L
W	S	K	T	H	V	Y	S	S	T	L	H	I	C	W	G	A	Q	A	G	G	L
W	S	R	H	N	V	Y	S	S	T	M	F	I	C	W	A	A	Q	A	G	L	L
W	A	K	A	H	V	T	S	S	T	L	Y	I	C	W	A	A	Q	A	A	G	L
W	A	K	E	H	V	T	S	S	T	L	F	V	C	W	A	V	Q	A	A	A	L
W	T	Q	T	N	V	H	S	S	T	L	N	V	C	W	G	A	M	A	A	A	I
W	S	K	T	H	V	Y	S	S	T	L	H	I	C	W	G	A	Q	A	A	G	L

Position-Specific Scoring Matrix (counts)

Position	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Residue																				
A	0	4	0	1	0	0	0	0	0	0	0	0	0	0	8	11	0	17	10	0
C	0	0	0	0	0	0	0	0	0	0	0	0	17	0	0	0	0	0	0	0
D	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
E	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
F	1	0	0	0	0	0	0	0	0	0	0	6	0	0	0	0	0	0	0	0
G	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9	1	0	0	7	0
H	0	0	0	1	10	0	3	0	0	0	0	6	0	0	0	0	0	0	0	0
I	0	0	0	0	0	0	0	0	0	1	0	0	10	0	0	0	0	0	0	1
K	0	0	12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
L	0	1	0	0	0	0	0	0	0	0	0	14	0	0	0	0	0	0	0	14
M	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	3	0	0	0
N	0	0	0	0	7	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0
P	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	14	0	0
Q	0	2	1	0	0	0	0	0	0	0	0	0	0	0	0	14	0	0	0	0
R	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
S	0	9	0	2	0	0	0	16	2	0	0	0	0	0	0	0	0	0	0	0
T	0	3	1	7	0	0	10	0	15	0	0	0	0	0	0	0	0	0	0	0
V	0	0	1	0	0	17	0	0	0	0	0	7	0	0	5	0	0	0	0	2
W	16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	17	0	0	0	0
Y	0	0	0	0	0	0	0	4	0	0	0	2	0	0	0	0	0	0	0	0
sum	17	17	17	17	17	17	17	17	17	17	17	17	17	17	17	17	17	17	17	17

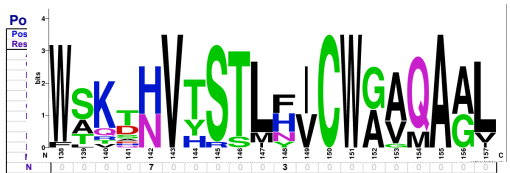
Multiple alignment

W	S	K	T	N	V	T	S	T	L	H	I	C	W	G	A	Q	A	G	L
W	S	K	T	N	V	T	S	T	L	N	I	C	W	G	A	Q	A	G	L
W	T	Q	S	H	V	T	H	R	T	L	N	I	C	W	A	A	Q	A	A	A	V
F	L	K	Q	N	V	T	T	S	S	M	Y	I	C	W	G	A	M	A	A	L
W	S	V	T	N	V	T	S	S	T	I	H	I	C	W	G	A	A	A	G	L
W	S	K	D	H	V	T	S	S	T	L	F	V	C	W	A	V	Q	A	A	G	L
W	S	K	S	H	V	Y	S	S	T	L	F	V	C	W	A	V	Q	A	A	A	L
W	S	K	S	H	V	Y	S	S	T	L	H	I	C	W	G	A	Q	A	A	A	L
W	T	T	T	N	V	H	S	S	T	L	N	V	C	W	G	G	M	A	A	A	V
W	T	K	D	H	V	T	S	S	T	L	F	V	C	W	A	V	Q	A	A	A	L
W	A	K	D	H	V	T	T	S	S	T	L	F	H	I	C	W	A	V	Q	A	A	L
W	S	K	T	H	V	Y	S	S	T	L	H	I	C	W	G	A	Q	A	G	G	L
W	S	R	H	N	V	Y	S	S	T	M	F	I	C	W	A	A	Q	A	G	L	L
W	A	K	A	H	V	T	S	S	T	L	Y	I	C	W	A	A	Q	A	A	G	L
W	A	K	E	H	V	T	S	S	T	L	F	V	C	W	A	V	Q	A	A	A	L
W	T	Q	T	N	V	H	S	S	T	L	N	V	C	W	G	A	M	A	A	A	I
W	S	K	T	H	V	Y	S	S	T	L	H	I	C	W	G	A	Q	A	A	G	L

Po

Pos

Res



Weight matrix

Count matrix

Position	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	Sum	Freq
Residue																						
A	0	4	0	1	0	0	0	0	0	0	0	0	0	0	8	11	0	17	10	0	51	0.150
C	0	0	0	0	0	0	0	0	0	0	0	0	17	0	0	0	0	0	0	0	17	0.050
D	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0.012
E	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0.003
F	1	0	0	0	0	0	0	0	0	0	0	6	0	0	0	0	0	0	0	0	7	0.021
G	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9	1	0	0	7	0	17	0.050
H	0	0	0	1	10	0	3	0	0	0	0	6	0	0	0	0	0	0	0	0	20	0.059
I	0	0	0	0	0	0	0	0	0	1	0	0	10	0	0	0	0	0	0	1	12	0.035
K	0	0	12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	12	0.035
L	0	1	0	0	0	0	0	0	0	0	0	14	0	0	0	0	0	0	0	14	29	0.085
M	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	3	0	0	0	5	0.015
N	0	0	0	0	7	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	10	0.029
P	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	14	0	0	2	0.006
Q	0	2	1	0	0	0	0	0	0	0	0	0	0	0	0	14	0	0	0	0	17	0.050
R	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0.006
S	0	9	0	2	0	0	0	16	2	0	0	0	0	0	0	0	0	0	0	0	29	0.085
T	0	3	1	7	0	0	10	0	15	0	0	0	0	0	0	0	0	0	0	0	36	0.106
V	0	0	1	0	0	17	0	0	0	0	0	7	0	0	5	0	0	0	0	2	32	0.094
W	16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	17	0	0	0	0	33	0.097
Y	0	0	0	0	0	0	0	4	0	0	0	2	0	0	0	0	0	0	0	0	6	0.018
sum	17	17	17	17	17	17	17	17	17	17	17	17	17	17	17	17	17	17	17	17	340	1.000

Weight matrix

Position	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Residue																				
A	-1.72	0.19	-1.72	-0.39	-1.72	-1.72	-1.72	-1.72	-1.72	-1.72	-1.72	0.49	0.63	-1.72	0.82	0.59	-1.72			
C	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	1.28	-1.26	-1.26	-1.26	-1.26	-1.26			
D	-0.70	-0.70	-0.70	1.21	-0.70	-0.70	-0.70	-0.70	-0.70	-0.70	-0.70	-0.70	-0.70	-0.70	-0.70	-0.70	-0.70			
E	-0.30	-0.30	-0.30	1.02	-0.30	-0.30	-0.30	-0.30	-0.30	-0.30	-0.30	-0.30	-0.30	-0.30	-0.30	-0.30	-0.30			
F	0.42	-0.90	-0.90	-0.90	-0.90	-0.90	-0.90	-0.90	-0.90	-0.90	1.18	-0.90	-0.90	-0.90	-0.90	-0.90	-0.90			
G	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	1.00	0.07	-1.26	-1.26			
H	-1.32	-1.32	-1.32	0.00	0.98	-1.32	0.46	-1.32	-1.32	-1.32	0.76	-1.32	-1.32	-1.32	-1.32	-1.32	-1.32			
I	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11	0.21	-1.11	1.19	-1.11	-1.11	-1.11	-1.11	-1.11			
K	-1.11	-1.11	1.27	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11			
L	-1.48	-0.15	-1.48	-1.48	-1.48	-1.48	-1.48	-1.48	0.97	-1.48	-1.48	-1.48	-1.48	-1.48	-1.48	-1.48	-1.48			
M	-0.78	-0.78	-0.78	-0.78	-0.78	-0.78	-0.78	0.83	-0.78	-0.78	-0.78	-0.78	-0.78	-0.78	1.01	-0.78	-0.78			
N	-1.04	-1.04	-1.04	-1.04	1.11	-1.04	-1.04	-1.04	-1.04	0.74	-1.04	-1.04	-1.04	-1.04	-1.04	-1.04	-1.04			
P	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00			
Q	-1.26	-1.26	0.36	0.07	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	1.19	-1.26	-1.26			
R	-0.48	-0.48	0.85	-0.48	-0.48	-0.48	-0.48	0.85	-0.48	-0.48	-0.48	-0.48	-0.48	-0.48	-0.48	-0.48	-0.48			
S	-1.48	0.78	-1.48	0.14	-1.48	-1.48	-1.48	1.03	0.14	-1.48	-1.48	-1.48	-1.48	1.03	0.14	-1.48	-1.48			
T	-1.57	0.22	-0.25	0.58	-1.57	-1.57	0.73	-1.57	0.91	-1.57	-1.57	-1.57	-1.57	-1.57	-1.57	-1.57	-1.57			
V	-1.52	-1.52	-0.20	-1.52	-1.52	1.01	-1.52	-1.52	-1.52	-1.52	0.63	-1.52	-1.52	-1.52	0.49	-1.52	-1.52			
W	0.98	-1.53	-1.53	-1.53	-1.53	-1.53	-1.53	-1.53	-1.53	1.00	-1.53	-1.53	-1.53	-1.53	-1.53	1.00	-1.53			
Y	-0.85	-0.85	-0.85	-0.85	-0.85	-0.85	1.06	-0.85	-0.85	-0.85	0.77	-0.85	-0.85	-0.85	-0.85	-0.85	-0.85			
sum	-17.8	-14.4	-13.7	-10.7	-17.2	-19.1	-15.7	-17.8	-17.6	-16.3	-14.1	-17.2	-19.1	-19.1	-17.2	-16	-17.4	-19.1	-17.2	-16.3

$$f_{i,j} = \frac{n_{i,j} + p_i k}{\sum_{r=1} n_{r,j} + k}$$

$$W_{i,j} = \ln\left(\frac{f_{i,j}}{p_i}\right)$$

Scoring a sequence with a profile matrix

Weight matrix

Position	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Residue																				
A	-1.72	0.19	-1.72	-0.39	-1.72	-1.72	-1.72	-1.72	-1.72	-1.72	-1.72	0.49	0.63	-1.72	0.82	0.59	-1.72			
C	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	1.28	-1.26	-1.26	-1.26	-1.26	-1.26			
D	-0.70	-0.70	-0.70	1.21	-0.70	-0.70	-0.70	-0.70	-0.70	-0.70	-0.70	-0.70	-0.70	-0.70	-0.70	-0.70	-0.70			
E	-0.30	-0.30	-0.30	1.02	-0.30	-0.30	-0.30	-0.30	-0.30	-0.30	-0.30	-0.30	-0.30	-0.30	-0.30	-0.30	-0.30			
F	0.42	-0.90	-0.90	-0.90	-0.90	-0.90	-0.90	-0.90	-0.90	-0.90	1.18	-0.90	-0.90	-0.90	-0.90	-0.90	-0.90			
G	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	1.00	0.07	-1.26	-1.26			
H	-1.32	-1.32	-1.32	0.00	0.98	-1.32	0.46	-1.32	-1.32	-1.32	0.76	-1.32	-1.32	-1.32	-1.32	-1.32	-1.32			
I	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11	0.21	-1.11	1.19	-1.11	-1.11	-1.11	-1.11	-1.11			
K	-1.11	-1.11	1.27	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11			
L	-1.48	-0.15	-1.48	-1.48	-1.48	-1.48	-1.48	-1.48	0.97	-1.48	-1.48	-1.48	-1.48	-1.48	-1.48	-1.48	-1.48			
M	-0.78	-0.78	-0.78	-0.78	-0.78	-0.78	-0.78	0.83	-0.78	-0.78	-0.78	-0.78	-0.78	-0.78	1.01	-0.78	-0.78			
N	-1.04	-1.04	-1.04	-1.04	1.11	-1.04	-1.04	-1.04	-1.04	0.74	-1.04	-1.04	-1.04	-1.04	-1.04	-1.04	-1.04			
P	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00			
Q	-1.26	-1.26	0.36	0.07	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	1.19	-1.26	-1.26			
R	-0.48	-0.48	0.85	-0.48	-0.48	-0.48	-0.48	0.85	-0.48	-0.48	-0.48	-0.48	-0.48	-0.48	-0.48	-0.48	-0.48			
S	-1.48	0.78	-1.48	0.14	-1.48	-1.48	-1.48	1.03	0.14	-1.48	-1.48	-1.48	-1.48	1.03	0.14	-1.48	-1.48			
T	-1.57	0.22	-0.25	0.58	-1.57	-1.57	0.73	-1.57	0.91	-1.57	-1.57	-1.57	-1.57	-1.57	-1.57	-1.57	-1.57			
V	-1.52	-1.52	-0.20	-1.52	-1.52	1.01	-1.52	-1.52	-1.52	-1.52	0.63	-1.52	-1.52	-1.52	0.49	-1.52	-1.52			
W	0.98	-1.53	-1.53	-1.53	-1.53	-1.53	-1.53	-1.53	-1.53	1.00	-1.53	-1.53	-1.53	-1.53	-1.53	1.00	-1.53			
Y	-0.85	-0.85	-0.85	-0.85	-0.85	-0.85	1.06	-0.85	-0.85	-0.85	0.77	-0.85	-0.85	-0.85	-0.85	-0.85	-0.85			
sum	-17.8	-14.4	-13.7	-10.7	-17.2	-19.1	-15.7	-17.8	-17.6	-16.3	-14.1	-17.2	-19.1	-19.1	-17.2	-16	-17.4	-19.1	-17.2	-16.3

Sequence	L	W	A	K	D	H	V	T	S	T	M	F	V	C	W	A	V	M	A	A	SUM
Score	-1.48	-1.53	-1.72	-1.11	-0.7	-1.32	-1.52	-1.57	0.136	-1.57	-0.78	-0.9	-1.52	-1.26	-1.53	0.628	-1.52	-0.78	0.587	-1.72	-21.1626

Scoring a sequence with a profile matrix

Weight matrix

Position	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Residue																				
A	-1.72	0.19	-1.72	-0.39	-1.72	-1.72	-1.72	-1.72	-1.72	-1.72	-1.72	0.49	0.63	-1.72	0.82	0.59	-1.72			
C	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	1.28	-1.26	-1.26	-1.26	-1.26	-1.26			
D	-0.70	-0.70	-0.70	1.21	-0.70	-0.70	-0.70	-0.70	-0.70	-0.70	-0.70	-0.70	-0.70	-0.70	-0.70	-0.70	-0.70			
E	-0.30	-0.30	-0.30	1.02	-0.30	-0.30	-0.30	-0.30	-0.30	-0.30	-0.30	-0.30	-0.30	-0.30	-0.30	-0.30	-0.30			
F	0.42	-0.90	-0.90	-0.90	-0.90	-0.90	-0.90	-0.90	-0.90	-0.90	1.18	-0.90	-0.90	-0.90	-0.90	-0.90	-0.90			
G	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	1.00	0.07	-1.26	-1.26			
H	-1.32	-1.32	-1.32	0.00	0.98	-1.32	0.46	-1.32	-1.32	-1.32	0.76	-1.32	-1.32	-1.32	-1.32	-1.32	-1.32			
I	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11	0.21	-1.11	1.19	-1.11	-1.11	-1.11	-1.11	-1.11			
K	-1.11	-1.11	1.27	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11			
L	-1.48	-0.15	-1.48	-1.48	-1.48	-1.48	-1.48	-1.48	0.97	-1.48	-1.48	-1.48	-1.48	-1.48	-1.48	-1.48	-1.48			
M	-0.78	-0.78	-0.78	-0.78	-0.78	-0.78	-0.78	0.83	-0.78	-0.78	-0.78	-0.78	-0.78	-0.78	1.01	-0.78	-0.78			
N	-1.04	-1.04	-1.04	-1.04	1.11	-1.04	-1.04	-1.04	-1.04	0.74	-1.04	-1.04	-1.04	-1.04	-1.04	-1.04	-1.04			
P	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00			
Q	-1.26	-1.26	0.36	0.07	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	1.19	-1.26	-1.26			
R	-0.48	-0.48	0.85	-0.48	-0.48	-0.48	-0.48	0.85	-0.48	-0.48	-0.48	-0.48	-0.48	-0.48	-0.48	-0.48	-0.48			
S	-1.48	0.78	-1.48	0.14	-1.48	-1.48	-1.48	1.03	0.14	-1.48	-1.48	-1.48	-1.48	1.03	0.14	-1.48	-1.48			
T	-1.57	0.22	-0.25	0.58	-1.57	-1.57	0.73	-1.57	0.91	-1.57	-1.57	-1.57	-1.57	-1.57	-1.57	-1.57	-1.57			
V	-1.52	-1.52	-0.20	-1.52	-1.52	1.01	-1.52	-1.52	-1.52	-1.52	0.63	-1.52	-1.52	-1.52	0.49	-1.52	-1.52			
W	0.98	-1.53	-1.53	-1.53	-1.53	-1.53	-1.53	-1.53	-1.53	1.00	-1.53	-1.53	-1.53	-1.53	-1.53	1.00	-1.53			
Y	-0.85	-0.85	-0.85	-0.85	-0.85	-0.85	1.06	-0.85	-0.85	-0.85	0.77	-0.85	-0.85	-0.85	-0.85	-0.85	-0.85			
sum	-17.8	-14.4	-13.7	-10.7	-17.2	-19.1	-15.7	-17.8	-17.6	-16.3	-14.1	-17.2	-19.1	-19.1	-17.2	-16	-17.4	-19.1	-17.2	-16.3

Sequence	W	A	K	D	H	V	T	S	T	M	F	V	C	W	A	V	M	A	A	L	SUM
Score	0.975	0.192	1.268	1.21	0.981	1.014	0.735	1.029	0.91	0.835	1.18	0.631	1.277	1.061	0.491	0.486	1.007	0.817	0.587	0.972	17.59818

Scoring a sequence with a profile matrix

Weight matrix

Position	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Residue																				
A	-1.72	**0.19**	-1.72	-0.39	-1.72	-1.72	-1.72	-1.72	-1.72	-1.72	-1.72	-1.72	-1.72	-1.72	**0.49**	**0.63**	-1.72	**0.82**	**0.59**	-1.72
C	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	**1.28**	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26
D	-0.70	-0.70	-0.70	**1.21**	-0.70	-0.70	-0.70	-0.70	-0.70	-0.70	-0.70	-0.70	-0.70	-0.70	-0.70	-0.70	-0.70	-0.70	-0.70	-0.70
E	-0.30	-0.30	-0.30	**1.02**	-0.30	-0.30	-0.30	-0.30	-0.30	-0.30	-0.30	-0.30	-0.30	-0.30	-0.30	-0.30	-0.30	-0.30	-0.30	-0.30
F	**0.42**	-0.90	-0.90	-0.90	-0.90	-0.90	-0.90	-0.90	-0.90	-0.90	**1.18**	-0.90	-0.90	-0.90	-0.90	-0.90	-0.90	-0.90	-0.90	-0.90
G	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	**1.00**	**0.07**	-1.26	-1.26	**0.89**	-1.26	
H	-1.32	-1.32	-1.32	**0.00**	**0.98**	-1.32	**0.46**	-1.32	-1.32	-1.32	**0.76**	-1.32	-1.32	-1.32	-1.32	-1.32	-1.32	-1.32	-1.32	-1.32
I	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11	**0.21**	-1.11	**1.19**	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11	**0.21**
K	-1.11	-1.11	**1.27**	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11
L	-1.48	-0.15	-1.48	-1.48	-1.48	-1.48	-1.48	-1.48	-1.48	**0.97**	-1.48	-1.48	-1.48	-1.48	-1.48	-1.48	-1.48	-1.48	-1.48	**0.97**
M	-0.78	-0.78	-0.78	-0.78	-0.78	-0.78	-0.78	-0.78	-0.78	**0.83**	-0.78	-0.78	-0.78	-0.78	-0.78	-0.78	**1.01**	-0.78	-0.78	-0.78
N	-1.04	-1.04	-1.04	-1.04	**1.11**	-1.04	-1.04	-1.04	-1.04	-1.04	**0.74**	-1.04	-1.04	-1.04	-1.04	-1.04	-1.04	-1.04	-1.04	-1.04
P	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
Q	-1.26	-1.26	**0.36**	**0.07**	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	**1.19**	-1.26	-1.26	-1.26
R	-0.48	-0.48	**0.85**	-0.48	-0.48	-0.48	-0.48	**0.85**	-0.48	-0.48	-0.48	-0.48	-0.48	-0.48	-0.48	-0.48	-0.48	-0.48	-0.48	-0.48
S	-1.48	**0.78**	-1.48	**0.14**	-1.48	-1.48	-1.48	**1.03**	**0.14**	-1.48	-1.48	-1.48	-1.48	-1.48	-1.48	-1.48	-1.48	-1.48	-1.48	-1.48
T	-1.57	**0.22**	-0.25	**0.58**	-1.57	-1.57	**0.73**	-1.57	**0.91**	-1.57	-1.57	-1.57	-1.57	-1.57	-1.57	-1.57	-1.57	-1.57	-1.57	-1.57
V	-1.52	-1.52	-0.20	-1.52	-1.52	**1.01**	-1.52	-1.52	-1.52	-1.52	-1.52	**0.63**	-1.52	-1.52	-1.52	**0.49**	-1.52	-1.52	-1.52	**0.09**
W	**0.98**	-1.53	-1.53	-1.53	-1.53	-1.53	-1.53	-1.53	-1.53	-1.53	-1.53	-1.53	-1.53	**1.00**	-1.53	-1.53	-1.53	-1.53	-1.53	-1.53
Y	-0.85	-0.85	-0.85	-0.85	-0.85	-0.85	**1.06**	-0.85	-0.85	-0.85	**0.77**	-0.85	-0.85	-0.85	-0.85	-0.85	-0.85	-0.85	-0.85	-0.85
sum	**-17.8**	**-14.4**	**-13.7**	**-10.7**	**-17.2**	**-19.1**	**-15.7**	**-17.8**	**-17.6**	**-16.3**	**-14.1**	**-17.2**	**-19.1**	**-19.1**	**-17.2**	**-16**	**-17.4**	**-19.1**	**-17.2**	**-16.3**

Sequence	A	K	D	H	V	T	S	T	M	F	V	C	W	A	V	M	A	A	L	V	SUM
Score	-1.72	-1.11	-0.7	1E-16	-1.52	-1.57	-1.48	-1.57	-0.78	-0.9	-1.52	-1.26	-1.53	-1.72	-1.52	-0.78	-1.72	**0.617**	-1.48	**0.094**	-21.9422

PSI-BLAST

- PSI-BLAST stands for Position-Specific Iterated BLAST (Altschul et al, 1997)
  - BLAST runs a first time in normal mode.
  - Resulting sequences are aligned together (Multiple sequence alignment) and a PSSM is calculated.
  - This PSSM is used to scan the database for new matches.
  - Steps 2-3 can be iterated several times.
- The PSSM increases the sensitivity of the search.

References

- Substitution matrices
  - PAM series
    - 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 substitution matrices
    - Henikoff, S. & Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci U S A 89, 10915-9.
  - Gonnet matrices, built by an iterative procedure
    - Gonnet, G. H., Cohen, M. A. & Benner, S. A. (1992). Exhaustive matching of the entire protein sequence database. Science 256, 1443-5. 1.
- Sequence alignment algorithms
  - Needleman-Wunsch (pairwise, global)
    - Needleman, S. B. & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 48, 443-53.
  - Smith-Waterman (pairwise, local)
    - Smith, T. F. & Waterman, M. S. (1981). Identification of common molecular subsequences. J Mol Biol 147, 195-7.
  - FastA (database searches, pairwise, local)
    - W. R. Pearson and D. J. Lipman. Improved tools for biological sequence comparison. Proc. Natl. Acad. Sci. USA, 85:2444–2448, 1988.
  - BLAST (database searches, pairwise, local)
    - S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. A basic local alignment search tool. J. Mol. Biol., 215:403–410, 1990.
    - S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs Nucleic Acids Res., 25:3389–3402, 1997.
  - Clustal (multiple, global)
    - Higgins, D. G. & Sharp, P. M. (1988). CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. Gene 73, 237-44.
    - Higgins, D. G., Thompson, J. D. & Gibson, T. J. (1996). Using CLUSTAL for multiple sequence alignments. Methods Enzymol 266, 383-402.
  - Dialign (multiple, local)
    - Morgenstern, B., Frech, K., Dress, A. & Werner, T. (1998). DIALIGN: finding local similarities by multiple sequence alignment. Bioinformatics 14, 290-4.
  - MUSCLE (multiple local)