Pairwise Sequence Alignment

Pengyu Hong
CS178A

Sequence Comparison





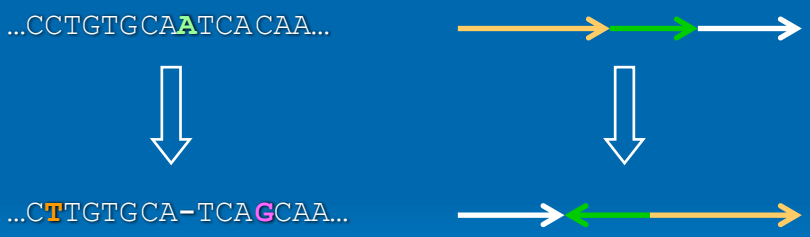



- Similarity → Homology
- Reveal evolutionary relationships
- Infer properties of genes/proteins
- Search for genes/proteins with same or similar functions






Genome Level Evolution



Sequence Alignment

```
AAB24882        TYHMCQFHCRYVNNHSGEKLYECNERSKAFSCPSHLQCHKRRQIGEKTHEHNQCGKAFPT 60
AAB24881        --------------------YECNQCGKAFAQHSSLKCHYRTHIGEKPYECNQCGKAFSK 40
                                     ****: .***:  * *:** * :****.:* *******..

AAB24882        PSHLQYHERTHTGEKPYECHQCGQAFKKCSLLQRHKRTHTGEKPYE-CNQCGKAFAQ- 116
AAB24881        HSHLQCHKRTHTGEKPYECNQCGKAFSQHGLLQRHKRTHTGEKPYMNVINMVKPLHNS 98
                 **** *:***********:***:**.: .***************    :  *.: :
```

clustalw

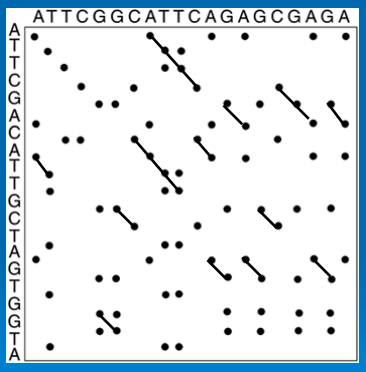
```
Global  FTFTALILLAVAV
        F--TAL-LLA-AV

Local   FTFTALILL-AVAV
        --FTAL-LLAAV--
```

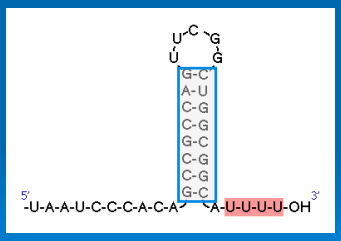
- Dot matrix
- Dynamic programming
- Word (*k*-tuple)

Img src: http://en.wikipedia.org/wiki/Sequence_alignment

Dot Plot

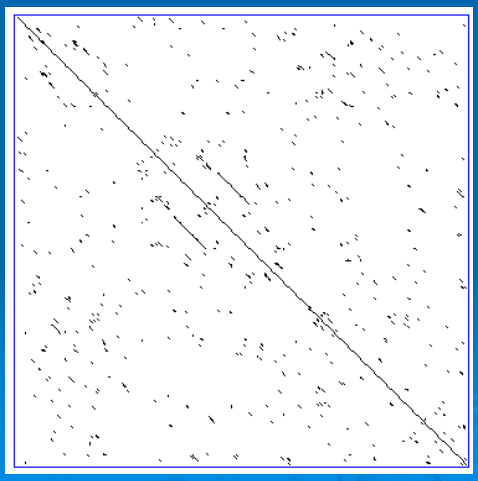


Identify repeats and inverted repeats



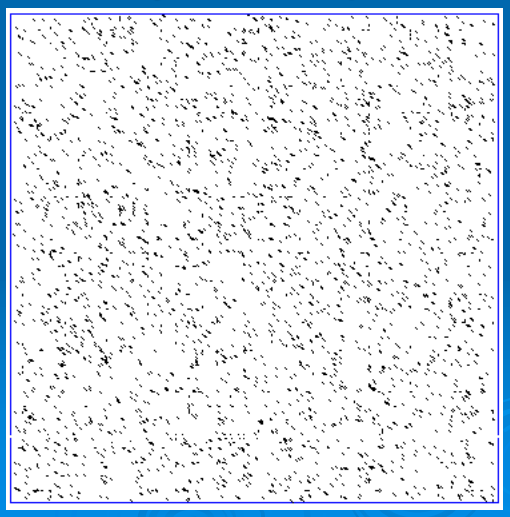
Dot Plot

Two identical sequences



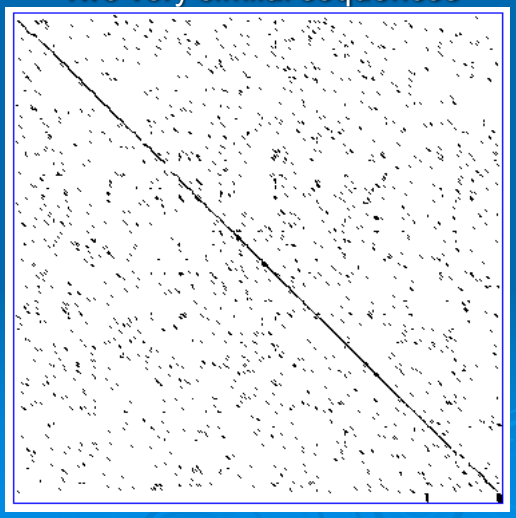
Dot Plot

Two very different sequences



Dot Plot

Two very similar sequences

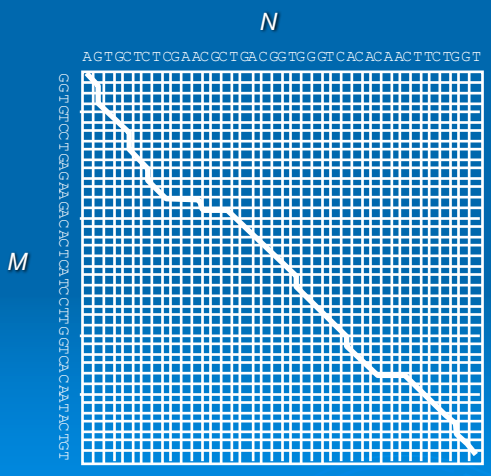


Dot Plot

- Dotmatcher
  - http://bioweb.pasteur.fr/seqanal/interfaces/dotmatcher.html
- MatrixPlot
  - http://www.cbs.dtu.dk/services/MatrixPlot

Dynamic Programming

– Compute the Best Alignment



$O(2^{M+N})$ possible alignments

Introduce scoring functions

Best alignment = Best score

Alignment → Optimization

Scoring Function

- Sequence edits:

  AGGCCTC
  - Mutations AGGACTC
  - Insertions AGGGCCTC
  - Deletions AGG–CTC

Scoring Function for Alignment:

Match: $+m$

Mismatch: $-u$

Gap: $-g$

Score = (# matches)×m - (# mismatches)×u – (#gaps)×g

Alignment Is Additive

$$x_1 \ldots x_i x_{i+1} \ldots x_M$$
$$y_1 \ldots y_j y_{j+1} \ldots y_N$$

$$\text{score}\begin{pmatrix} x_1 \ldots x_i x_{i+1} \ldots x_M \\ y_1 \ldots y_j y_{j+1} \ldots y_N \end{pmatrix}$$

$$= \text{score}\begin{pmatrix} x_1 \ldots x_i \\ y_1 \ldots y_j \end{pmatrix} + \text{score}\begin{pmatrix} x_{i+1} \ldots x_M \\ y_{j+1} \ldots y_N \end{pmatrix}$$

Divide and Conquer

Dynamic Programming

Let $F(i, j)$ = optimal score of aligning $\begin{matrix} x_1 \ldots x_{i-1} x_i \\ y_1 \ldots y_{j-1} y_j \end{matrix}$

We can define $F(i, j)$ recursively so that it can be calculated efficiently.

Three possible cases: $\begin{matrix} x_1 \ldots x_{i-1} x_i \\ y_1 \ldots y_{j-1} y_j \end{matrix}$

(a) x_i aligns to y_j

$x_1 \ldots\ldots x_{i-1} \quad x_i$

$y_1 \ldots\ldots y_{j-1} \quad y_j$

$$F(i,j) = F(i\text{-}1, j\text{-}1) + \begin{cases} m, \text{ if } x_i = y_j \\ -u, \text{ otherwise} \end{cases}$$

$s(x_i, y_j)$

(b) x_i aligns to a gap

$x_1 \ldots\ldots x_{i-1} \quad x_i$

$y_1 \ldots\ldots y_j \quad -$

$$F(i,j) = F(i\text{-}1, j) - g$$

(c) y_j aligns to a gap

$x_1 \ldots\ldots x_i \quad -$

$y_1 \ldots\ldots y_{j-1} \quad y_j$

$$F(i,j) = F(i, j\text{-}1) - g$$

$$F(i,j) = \max(F(i\text{-}1, j\text{-}1) + s(x_i, y_j),\ F(i, j\text{-}1) - g,\ F(i\text{-}1, j) - g)$$

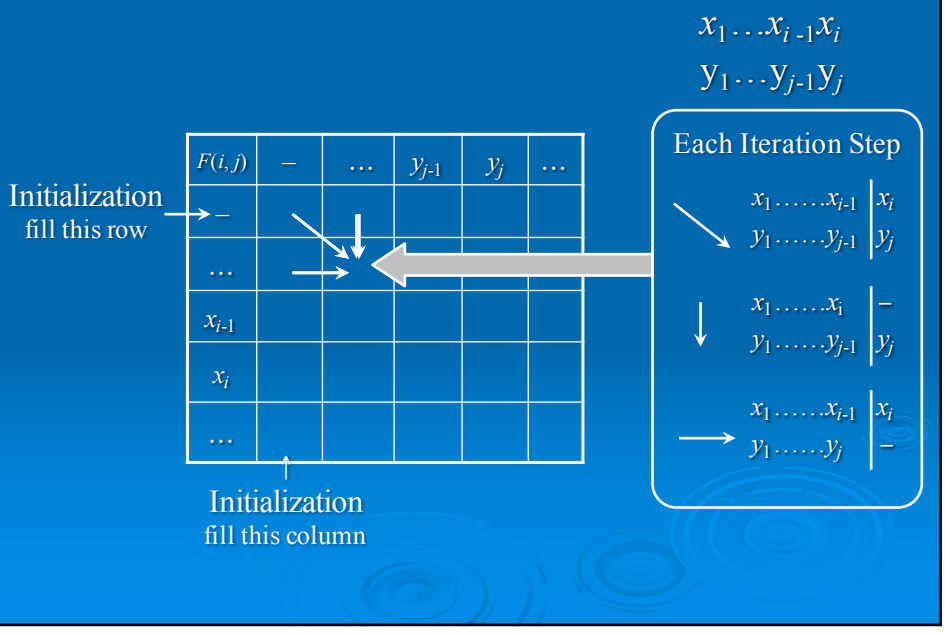
The Needleman-Wunsch Algorithm (Global)

1. Initialization.
   a) $F(0, 0) = 0$
   b) $F(0, j) = -j \times g$
   c) $F(i, 0) = -i \times g$

2. Iteration. Filling-in partial alignments
   for $i = 1$ to M
   for $j = 1$ to N
   $F(i, j) = \max(F(i\text{-}1, j\text{-}1) + s(x_i, y_j), F(i, j\text{-}1) - g, F(i\text{-}1, j) - g)$
   $p(i,j) =$ Diag, Up, or Left
   end
   end

3. Termination. $F(M, N)$ is the optimal score, and use $p(M, N)$ to trace back the optimal alignment

- Time: $O(NM)$. Space: $O(NM)$



Example

x = AGTA, y = ATA

$m = 1, u = 1, g = 1$

		0	1	2	3	4
	$F(i,j)$	-	A	G	T	A
0	-					
1	A					
2	T					
3	A					

		0	1	2	3	4
	$p(i,j)$	-	A	G	T	A
0	-					
1	A					
2	T					
3	A					

Example

x = AGTA, y = ATA

$m = 1, u = 1, g = 1$

		0	1	2	3	4
	$F(i,j)$	-	A	G	T	A
0	-	0	-1	-2	-3	-4
1	A	-1	1	0	-1	-2
2	T	-2	0	0	1	0
3	A	-3	-1	-1	0	2

		0	1	2	3	4
	$p(i,j)$	-	A	G	T	A
0	-		←	←	←	←
1	A	↑	↖	←	←	↖←
2	T	↑	↑	↖	↖	←
3	A	↑	↖↑	↖↑	↑	↖

Example

x = AGTA, y = ATA

$m = 1, u = 1, g = 1$

		0	1	2	3	4
	$F(i,j)$		A	G	T	A
0		0	-1	-2	-3	-4
1	A	-1	1	0	-1	-2
2	T	-2	0	0	1	0
3	A	-3	-1	-1	0	2

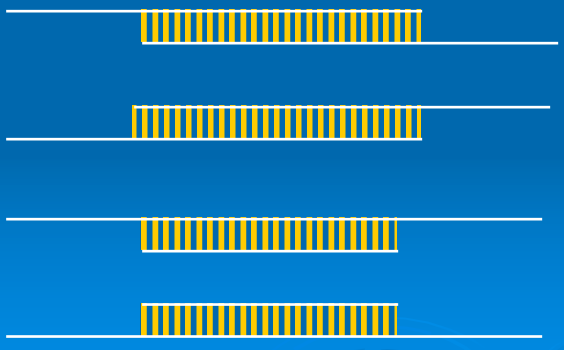
Optimal Alignment:

$F(4,3) = 2$

AGTA
A–TA

A Variation for Overlap Detection

- OK to have an unlimited number of gaps in the beginning and end.



Changes:

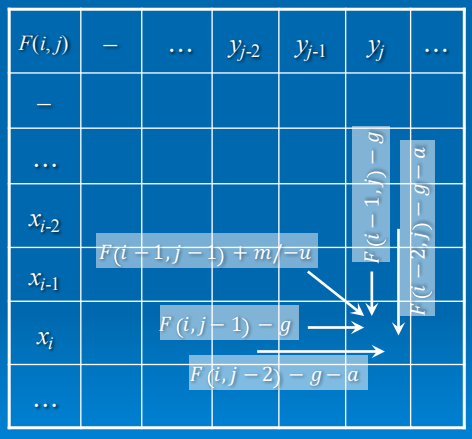
(a) Initialization:

(b) Iteration:

(c) Termination:

Gap Penalty Functions

- Gaps usually occur in bunches (an evolution event can insert/delete a segment)
- Affine gap penalty function: $g + a * (k - 1)$
  - Gap open penalty
    - Nucleotides 5
    - Proteins 11
  - Gap extension penalty
    - Nucleotides 2
    - Proteins 1



How about longer gaps?

Pairwise Sequence Alignment @ EMBL

http://www.ebi.ac.uk/Tools/psa/

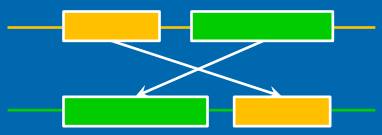
>seq1
CATCGAA

>seq2
CACGA

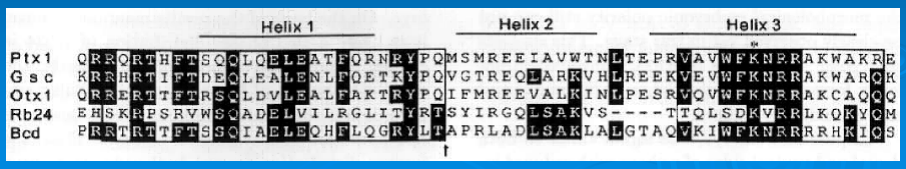
Try different gap penalty parameters

Why Local Alignment

- Genes are shuffled between genomes



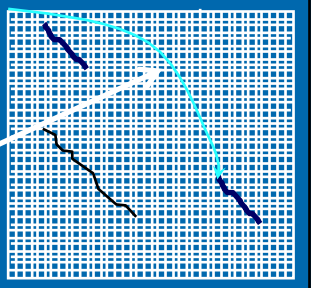
- Portions of proteins (e.g., domains), but not all, are conserved



The Smith-Waterman algorithm

Idea: Ignore badly aligning regions

Modifications to Needleman-Wunsch:



Initialization: $F(0, j) = F(i, 0) = 0$

Iteration: $$F(i, j) = \max \begin{cases} 0 \\ F(i-1, j) - g \\ F(i, j-1) - g \\ F(i-1, j-1) + s(x_i, y_j) \end{cases}$$

Termination:

- The best local alignment: $F_{OPT} = \max_{i,j} F(i, j)$
- All local alignments scoring $> t$: for all i, j that $F(i, j) > t$

Substitution Scoring Matrices

– Scoring Mismatches/Mutations

- Information accumulated in the gene/protein databases can be utilized to create scoring matrices.
- Matrices for DNA are relatively simple as there are only two options purine (A, G) & pyrimidine (C, T) and match & mismatch.
- Matrices for proteins are much more complex and there are many options.
  - PAM (point/percent accepted mutation)
  - BLOSUM (BLOck SUbstituion Matrix)

DNA Substitution Scoring Matrices

- Jukes-Cantor Model
  - All nucleotides are substituted with equal probability

	A	C	G	T
A		α	α	α
C	α		α	α
G	α	α		α
T	α	α	α	

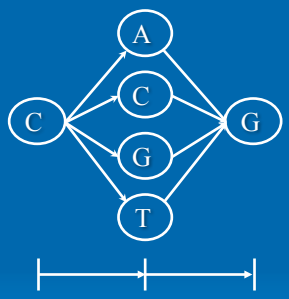
- Kimura Model
  - Mutation rates for transitions and transversi on are assumed to be different, which is more realistic

	A	T	G	C
A		β	α	β
T	β		β	α
G	α	β		β
C	β	α	β	

Transitions (α) occur more frequently than transversions (β)

Probability of C → G given time for 2 mutations

	A	C	G	T
A		α	α	α
C	α		α	α
G	α	α		α
T	α	α	α	



	A	T	G	C
A		β	α	β
T	β		β	α
G	α	β		β
C	β	α	β	

Some Concepts

- Sequence Similarity
- Sequence Identity
- Sequence Homology

In protein sequence alignment,
sequence identify = percentage of matches of the same amino acid residues between two aligned sequences.
sequence similarity = percentage of aligned amino acid residues that have similar physicochemical characteristics and can be substituted for each other.

Protein Substitution Scoring Matrices

- PAM (Point Accepted Mutation or Percent Accepted Mutation)
  - One PAM unit is an average of 1% change per 100 residues
  - PAM*n* is for sequences that have diverged *n* PAM units
    - PAM1 was constructed by *Margaret Dayhoff* from 71 protein sequence groups of at least 85% similarity.
    - PAMN = $(\text{PAM1})^N$
    - PAM250 represents a distance of 250 PAMs.
    - PAM matrices with lower series numbers are more suitable for aligning more closely related sequences.

PAM	Observed Mutation Rate (%)	Sequence Identity (%)
0	0	100
1	1	99
30	25	75
80	50	50
200	75	25
250	80	20

PAM1 × 10000

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

PAM1[*m*,*n*] = The probability of amino acid *m* being mutated to *n* after One PAM of mutation occurred.

Adapted from Figure 82. Atlas of Protein Sequence and Structure, Suppl 3, 1978, M.O. Dayhoff, ed. National Biomedical Research Foundation, 1979.

Mutation probabilities are converted into scoring matrix (log odds)

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	
C	12																				C
S	0	2																			S
T	−2	1	3																		T
P	−3	1	0	6																	P
A	−2	1	1	1	2																A
G	−3	1	0	−1	1	5															G
N	−4	1	0	−1	0	0	2														N
D	−5	0	**0**	−1	0	1	2	4													D
E	−5	**0**	**0**	−1	0	0	1	3	4												E
Q	−5	−1	−1	**0**	0	−1	1	2	2	4											Q
H	−3	−1	−1	0	−1	−2	2	1	1	3	6										H
R	−4	0	−1	0	−2	−3	0	−1	−1	1	2	6									R
K	−5	0	0	−1	−1	−2	1	0	0	1	0	3	5								K
M	−5	−2	−1	−2	−1	−3	−2	−3	−2	−1	−2	0	0	6							M
I	−2	−1	0	−2	−1	−3	−2	−2	−2	−2	−2	−2	−2	2	5						I
L	−6	−3	−2	−3	−2	−4	−3	−4	−3	−2	−2	−3	−3	4	2	6					L
V	−2	−1	0	−1	0	−1	−2	−2	−2	−2	−2	−2	−2	2	4	2	4				V
F	−4	−3	−3	−5	−4	−5	−4	−6	−5	−5	−2	−4	−5	0	1	2	−1	9			F
Y	0	−3	−3	−5	−3	−5	−2	−4	−4	−4	0	−4	−4	−2	−1	−1	−2	7	10		Y
W	−8	−2	−5	−6	−6	−7	−4	−7	−7	−5	−3	2	−3	−4	−5	−2	−6	0	0	17	W
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	

$$\log_{10}\left(\frac{P_{ij}}{P_j}\right)$$

PAM250. Amino acids are grouped according to to the chemistry of the side group: (C) sulfhydryl, (STPAG)-small hydrophilic, (NDEQ) acid, acid amide and hydrophilic, (HRK) basic, (MILV) small hydrophobic, and (FYW) aromatic.

Positive values, 0, and negative values denote substitutions occurring more, equally, or less frequently than expected among evolutionarily conserved replacements respectively.

$$Total\ score = \sum_{k=1}^{n} (alignment\ score\ at\ each\ poistion) + Total\ gap\ penalty$$

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	
C	12																				C
S	0	2																			S
T	−2	1	3																		T
P	−3	1	0	6																	P
A	−2	1	1	1	2																A
G	−3	1	0	−1	1	5															G
N	−4	1	0	−1	0	0	2														N
D	−5	0	0	−1	0	1	2	4													D
E	−5	0	0	−1	0	0	1	3	4												E
Q	−5	−1	−1	0	0	−1	1	2	2	4											Q
H	−3	−1	−1	0	−1	−2	2	1	1	3	6										H
R	−4	0	−1	0	−2	−3	0	−1	−1	1	2	6									R
K	−5	0	0	−1	−1	−2	1	0	0	1	0	3	5								K
M	−5	−2	−1	−2	−1	−3	−2	−3	−2	−1	−2	0	0	6							M
I	−2	−1	0	−2	−1	−3	−2	−2	−2	−2	−2	−2	−2	2	5						I
L	−6	−3	−2	−3	−2	−4	−3	−4	−3	−2	−2	−3	−3	4	2	6					L
V	−2	−1	0	−1	0	−1	−2	−2	−2	−2	−2	−2	−2	2	4	2	4				V
F	−4	−3	−3	−5	−4	−5	−4	−6	−5	−5	−2	−4	−5	0	1	2	−1	9			F
Y	0	−3	−3	−5	−3	−5	−2	−4	−4	−4	0	−4	−4	−2	−1	−1	−2	7	10		Y
W	−8	−2	−5	−6	−6	−7	−4	−7	−7	−5	−3	2	−3	−4	−5	−2	−6	0	0	17	W
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	

YCNERSKA
YCN--SVA

BLOSUM

- Constructed based on more than 2,000 conserved amino acid patterns representing 500 groups of protein sequences
- BLOSUM matrices are actual percentage identity values of sequences selected for construction of the matrices.
  - BLOSUM62 indicates that the sequences selected for constructing the matrix share an average identity value of 62%.
  - Lower BLOSUM series number represent more divergency.

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

BLOSUM62

PAM *vs* BLOSUM

- PAM matrices
  - PAM1 derived from global alignment of full-length sequences
  - PAM*n* are derived from PAM1 using an evolutionary model
  - More often used for reconstructing phylogenetc trees.
  - Less realistic for divergent sequences
- BLOSUM matrices
  - Derived from direct observations
  - May have less evolutionary meaning
  - Empirically shown to outperform the PAM matrices in local alignment
  - May be more advantageous in searching databases and fining conserved domains in proteins.

Question

Any better way to generate a substitution matrix?