双序列比对

高建召

致谢

此PPT内容参考了
杨建益老师的PPT。
https://yanglab.nankai.edu.cn/

双序列比对的算法

- Dot Matrix，点阵法
- 动态规划算法：
  - Global: Needleman-Wunsch
  - Local：Smith-Waterman
- Word or *k*-tuple算法：FASTA，BLAST

为什么要做序列比对

>Protein a

MVLSEGEWQLVLHVWAKVEADVAGHGQD
ILIRLFKSHPETLEKFDRVKHLKTEAEMKAS
EDLKKHGVTVLTALGAILKKKGHHEAELKP
LAQSHATKHKIPIKY

>Protein b

MNIFEMLRIDEGLRLKIYKDTEGYYTIGIGHLLTKSPS
LNAAAKSELDKAIGRNTNGVITKDEAEKLFNQDVDA
AVRGILRNAKLKPVYDSLDAVRRAALINMVFQMGET
GVAGFTNSLRMLQ

Do they have similar structure and function?

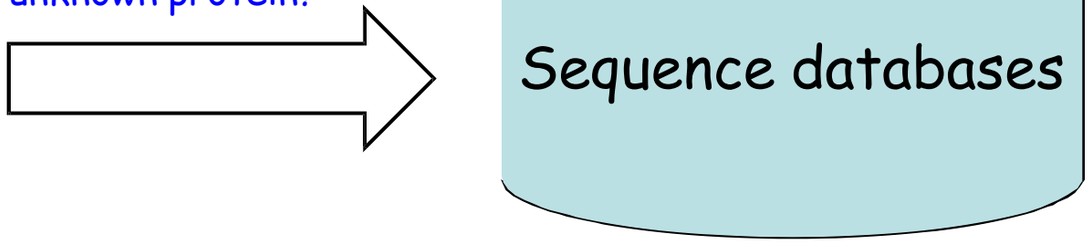
```
Length of sequence 1:  104 ->a.fasta
Length of sequence 2:  123 ->b.fasta
Aligned length:   93
Identical length:   22
Sequence identity:    0.179 (=  22/ 123)

------MVLSEGEWQLVLHVWAKVEADVA---GHGQDILIRLFKSHPETLEKFDRVKHLKTEAEMKASEDLKKHGVTVLTALGAILKKKGHHEAELKPLAQS-HATK-------------------HKIPIKY
          ::   : :      :       ::    :             :      :       : :     : :   ::       : :::   :  :                        
MNIFEMLRIDEG---LRLKIYKDTEGYYTIGIGH---LLTKSPSLNAAAKSELDKAIGRNTNGVITKDEAEKLFNQDVDAAVRGILRN-----AKLKPVYDSLDAVRRAALINMVFQMGETGVAGFTNSLRMLQ
12345678901234567890123456789012345678901234567890123456789012345678901234567890123456789012345678901234567890123456789012345678901234
```

I. Sequence alignment can help establish relationship of two proteins (roughly speaking, sequences having higher sequence identity usually come from the same ancestor and therefore have similar structure and function).
These proteins are called homology.

为什么做序列比对

>Query sequence
MVLSEGEWQLVLHVWAKVEADVAGHGQD
ILIRLFKSHPETLEKFDRVKHLKTEAEMKAS
EDLKKHGVTVLTALGAILKKKGHHEAELKP
LAQSHATKHKIPIKY

Can I find proteins in the databases, which are homologous to my unknown protein?



(GeneBank for DNA sequences)
(UniProt for protein sequences)
(PDB for protein structures)

II. Sequence alignment can help identify homologies from known databases, to generate structure and function predictions for the unknown proteins.

目前已有的核酸、蛋白质数据库

1. **GeneBank**: contains ~950M DNA sequences
2. **UniProt Swiss-Prot/trEMBL**: ~100M protein sequences (~550K with known function)
3. **Protein Data Bank (PDB)**: contains ~140k protein structures

小结

Purposes:

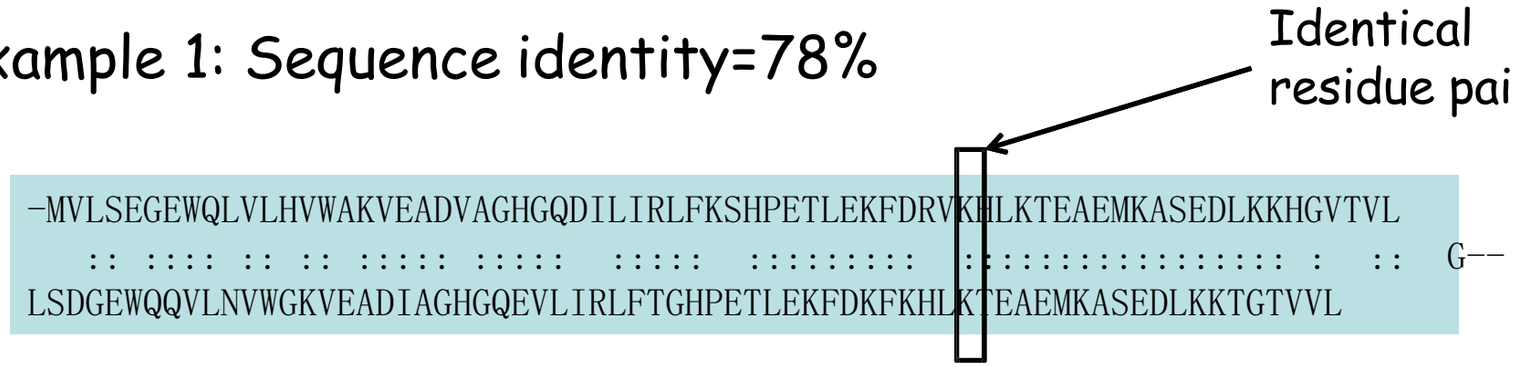
- Study the relationship between two proteins
- Scan a database with a query sequence and identify possible structure and function of the query protein

If two sequences are simiar,the following may be true

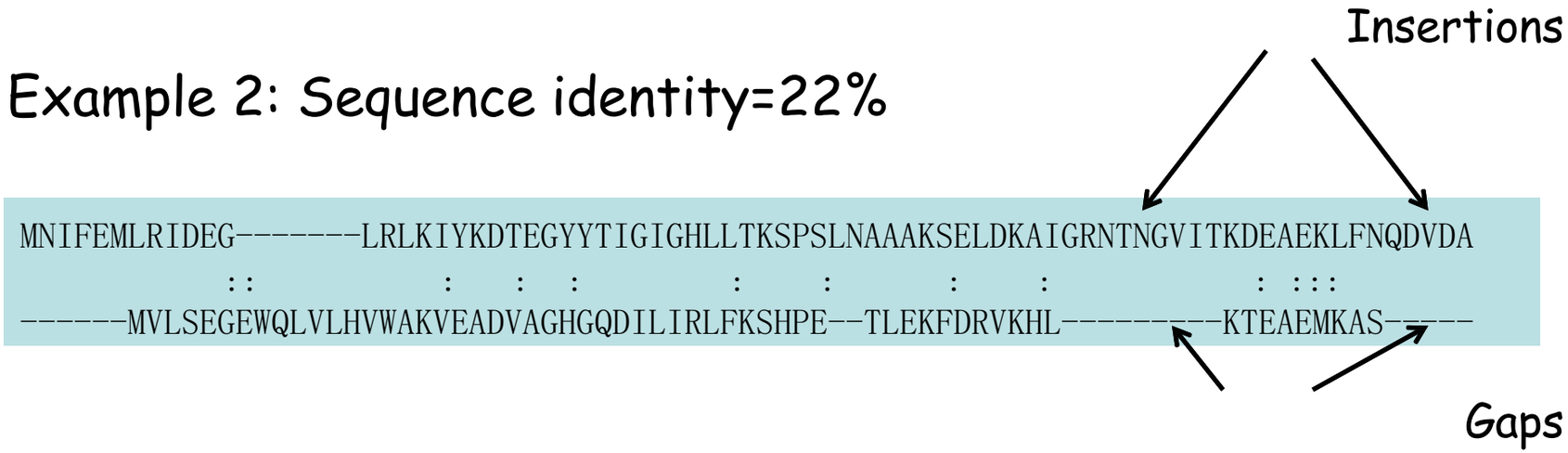
- The proteins may share a common evolutionary origin
- The proteins may have a similar 3-dimensional structure
- The proteins may have the same or related function

序列比对是什么

Example 1: Sequence identity=78%



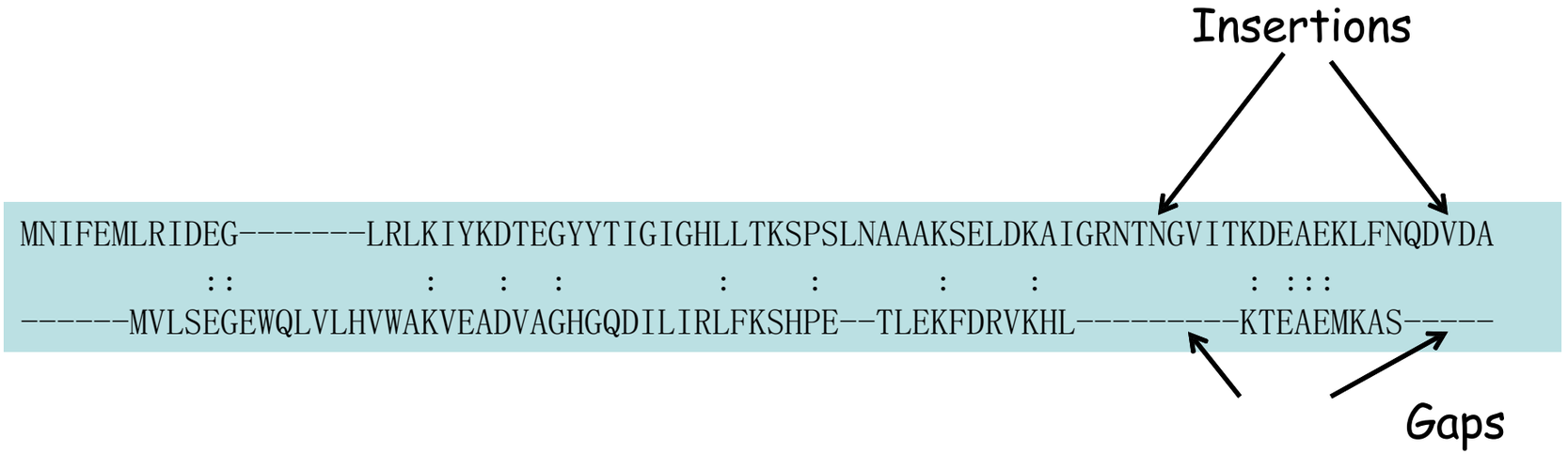
Example 2: Sequence identity=22%



Sequence identity = Number of identical residue pairs/Length of query sequence

序列比对的原则

- We want to align as many as possible THE SAME or THE SIMILAR residues
- We do not want gaps/insertions



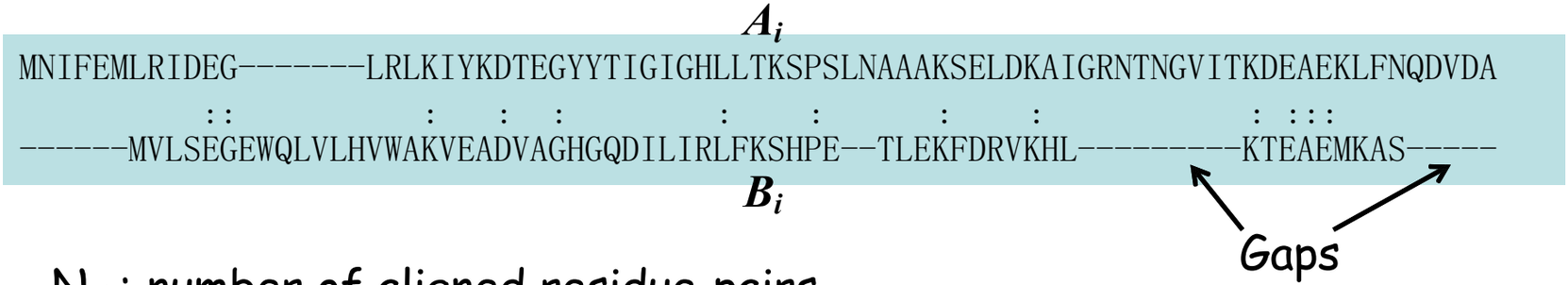
序列比对的原则

Mathematically, the goal is to maximize the following score:

$$Score = \sum_{i=1}^{N_{ali}} M(A_i, B_i) - GapPenalty$$

Residues of similar property should match together

Score for adding gap is always negative



N_{ali}: number of aligned residue pairs

A_i: amino acid identity of the i-th aligned resideu at the first sequence

B_i: amino acid identity of the i-th aligned resideu at the second sequence

$M(A_i, B_i)$: preference score of matching between amino acids A_i and B_i

打分矩阵

$$Score = \sum_{i=1}^{N_{ali}} M(A_i, B_i) - GapPenalty$$

The simplest scoring matrix is the unit matrix:

$$M = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}_{20\times 20}$$

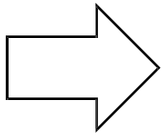
Question: What will be the problem if we use this simple solution?

Answer: All the similarity due to the evolutionary mutation has been neglected.

如何构造打分矩阵

The most often-used scoring matrices

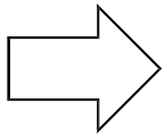
PAM250



DAYHOFF et al, 1978

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	13	6	9	9	5	8	9	12	6	8	6	7	7	4	11	11	11	2	4	9
R	3	17	4	3	2	5	3	2	6	3	2	9	4	1	4	4	3	7	2	2
N	4	4	6	7	2	5	6	4	6	3	2	5	3	2	4	5	4	2	3	3
D	5	4	8	11	1	7	10	5	6	3	2	5	3	1	4	5	5	1	2	3
C	2	1	1	1	52	1	1	2	2	2	1	1	1	1	2	3	2	1	4	2
Q	3	5	5	6	1	10	7	3	7	2	3	5	3	1	4	3	3	1	2	3
E	5	4	7	11	1	9	12	5	6	3	2	5	3	1	4	5	5	1	2	3
G	12	5	10	10	4	7	9	27	5	5	4	6	5	3	8	11	9	2	3	7
H	2	5	5	4	2	7	4	2	15	2	2	3	2	2	3	3	2	2	3	2
I	3	2	2	2	2	2	2	2	2	10	6	2	6	5	2	3	4	1	3	9
L	6	4	4	3	2	6	4	3	5	15	34	4	20	13	5	4	6	6	7	13
K	6	18	10	8	2	10	8	5	8	5	4	24	9	2	6	8	8	4	3	5
M	1	1	1	1	0	1	1	1	1	2	3	2	6	2	1	1	1	1	1	2
F	2	1	2	1	1	1	1	1	3	5	6	1	4	32	1	2	2	4	20	3
P	7	5	5	4	3	5	4	5	5	3	3	4	3	2	20	6	5	1	2	4
S	9	6	8	7	7	6	7	9	6	5	4	7	5	3	9	10	9	4	4	6
T	8	5	6	6	4	5	5	6	4	6	4	6	5	3	6	8	11	2	3	6
W	0	2	0	0	0	0	0	0	1	0	1	0	0	1	0	1	0	55	1	0
Y	1	1	2	1	3	1	1	1	3	2	2	1	2	15	1	2	2	3	31	2
V	7	4	4	4	4	4	4	4	5	4	15	10	4	10	5	5	5	72	4	17

BLOSUM62



Henikoff and Henikoff, PNAS, 1992

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

Questions:

1. How these matrices are obtained?
2. What are the differences between PAM and BLOSUM?

Margaret Dayhoff (1925 - 1983, US)

1945 - BA in Mathematics at NYU
1948 - PhD in Quantum Chemistry

1965 - Protein Atlas (65 proteins) (PIR)

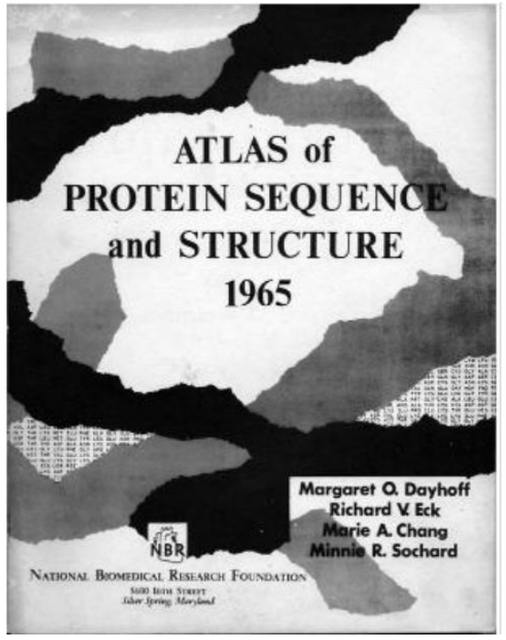
the first public comprehensive, computerised and publicly available database of protein sequences. It is the model for GenBank and many other molecular databases.

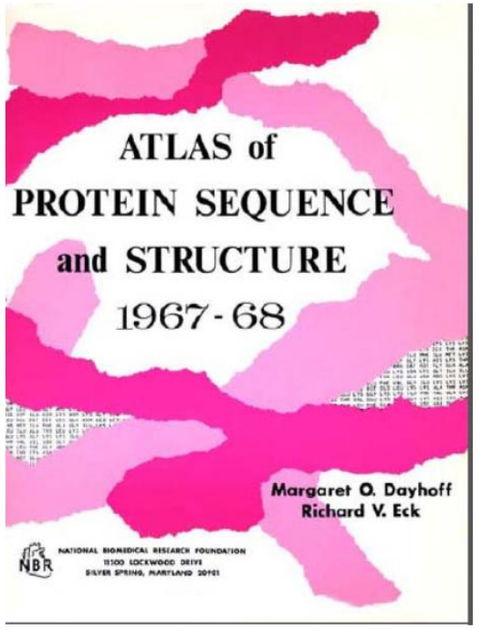
1980 - President of Biophysical Society

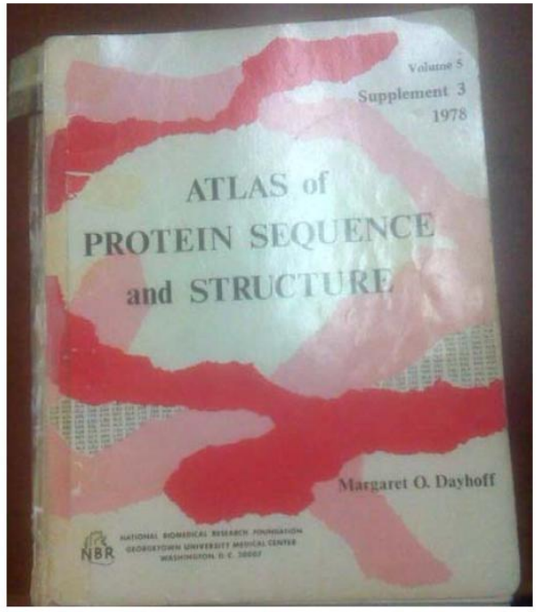
one of the founders in the field of Bioinformatics

https://www.whatisbiotechnology.org/index.php/people/summary/Dayhoff

Altas of protein sequence and structure 蛋白质序列和结构图谱







打分矩阵 PAM（Percent Accepted Mutation）

PAM (Percent Accepted Mutation) Matrix (by Dayhoff et al 1978):

- **Reference：** DAYHOFF, M., R. SCHWARTZ, AND B. ORCUTT. 1978. A model of evolutionary change in proteins. Pages 345--352 in Atlas of protein sequence and structure, Volume 5 (M. Dayhoff, ed.). National Biomedical Research Foundation, Washington, D.C.
- **Database**: 1,572 mutations, 75 homologous sequence groups, minimum sequence identity is 85%
- **Purpose：** to derive the mutation probability between amino acids

打分矩阵 PAM

Three steps for building the PAM matrix:

Step 1: Counting the number of mutations

Step 2: Relative mutability of amino acid

Step 3: Probability of mutations between amino acids (M_{ij})

Scoring matrix PAM

Step 1: Counting the number of mutations

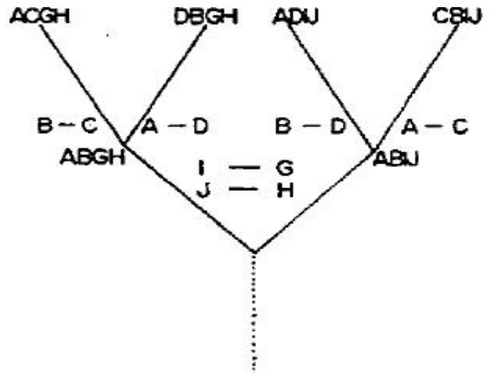
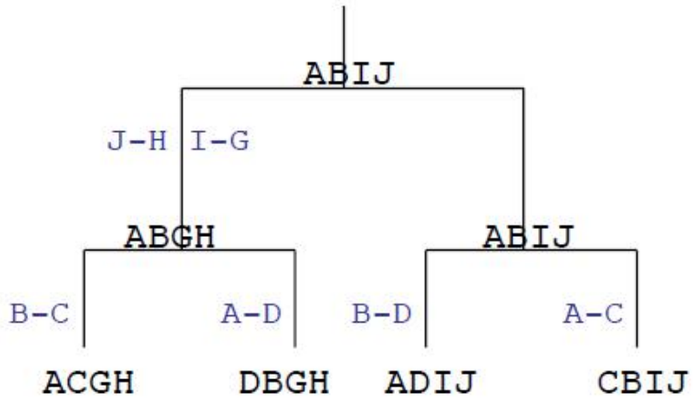


Figure 78. Simplified phylogenetic tree. Four "observed" proteins are shown at the top. Inferred ancestors are shown at the nodes. Amino acid exchanges are indicated along the branches.



	A	B	C	D	G	H	I	J
A			1	1				
B			1	1				
C	1	1						
D	1	1						
G							1	
H								1
I					1			
J						1		

Figure 79. Matrix of accepted point mutations derived from the tree of Figure 78.

Scoring matrix PAM

Glu-Asp 谷氨酸-天冬氨酸；

Mutation matrix from 1572 changes in 71 groups (id>85%)

Gly-Trp 甘氨酸-色氨酸

$A = (A_{ij})$原始矩阵

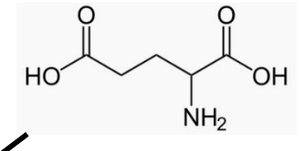
		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	Ala																				
R	Arg	30																			
N	Asn	109	17																		
D	Asp	154	0	532																	
C	Cys	33	10	0	0																
Q	Gln	93	120	50	76	0															
E	Glu	266	0	94	831	0	422														
G	Gly	579	10	156	162	10	30	112													
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	50	23	43	39									
M	Met	29	17	0	0	0	20	7	7	0	57	207	20								
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

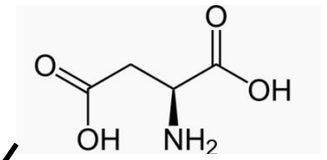
Figure 80. Numbers of accepted point mutations (X10) accumulated from closely related sequences. Fifteen hundred and seventy-two exchanges are shown. Fractional exchanges result when ancestral sequences are ambiguous.

注：总观测到1572次替换，表中次数均乘以10，祖先序列不明时，次数以平分处理。

Scoring matrix PAM

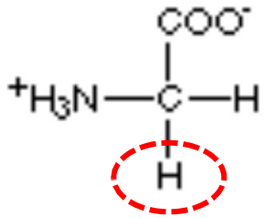
Two factors may influence the mutation numbers:



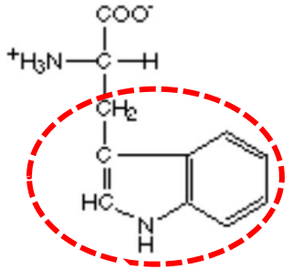


- Codon reason: mutation between Glu (=*GAA*, *GAG*) and Asp (=*GAC*, *GAU*) is the most frequent

- Physical reason: due to the volume difference, mutation between Gly (=*GGG*) and Trp (=*UGG*) never happens



Glycine
(gly)



Tryptophan
(trp)

Scoring matrix PAM

Step 2: Relative mutability of amino acid

$$m_i = \frac{N_{mut}(i)}{N_{comp}(i)}, i = 1,2,\cdots,20$$

Example:

Aligned	A D A		
sequences	A D B		
Amino acids	A	B	D
Changes	1	1	0
Frequency of occurrence (total composition)	3	1	2
Relative mutability	.33	1	0

Figure 81. Sample computation of relative mutability. The two aligned sequences may be two experimentally observed sequences or an observed sequence and its inferred ancestor.

Scoring matrix PAM

Table 21

Relative Mutabilities of the Amino Acids[a]

Asn	134	His	66
Ser	120	Arg	65
Asp	106	Lys	56
Glu	102	Pro	56
Ala	100	Gly	49
Thr	97	Tyr	41
Ile	96	Phe	41
Met	94	Leu	40
Gln	93	Cys	20
Val	74	Trp	18

[a] The value for Ala has been arbitrarily set at 100.

Scoring matrix PAM

Step 3: Probability of mutations between amino acids (M_{ij})

$$M_{ij} = \begin{cases} \lambda \dfrac{m_j A_{ij}}{\sum_{k=1}^{20} A_{kj}}, & 1 \le i, j \le 20; \quad i \ne j \\ 1 - \lambda m_j, & i = j \end{cases}$$

A_{ij}: Observed number of mutations between a_i and a_j

m_j: Relative mutate probability of a_j to all other amino acids

λ: A constant to decide the evolution distance

j

PAM1

ORIGINAL AMINO ACID

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

i

$$M_{ij}\ (j \rightarrow i)$$

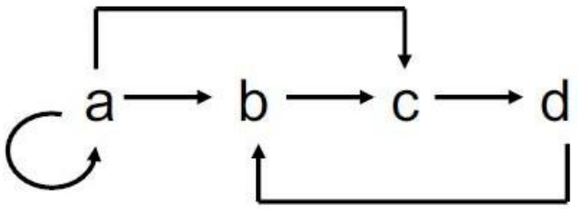
突变概率矩阵

Asymmetric?

For clarity, the values have been multiplied by 10000

From PAM1 to PAM2, PAM100, PAM250, etc…

Remark (from graph theory)



	a	b	c	d
a	1	1	1	0
b	0	0	1	0
c	0	0	0	1
d	0	1	0	0

Matrix **Q** indicates the number of paths going from one node to another in 1 step

	a	b	c	d
a	1	1	2	1
b	0	0	0	1
c	0	1	0	1
d	0	1	1	1

Matrix $\mathbf{Q}^2$ indicates the number of paths going from one node to another in 2 steps

	a	b	c	d
a	…	…	…	…
b	…	…	…	…
c	…	…	…	…
d	…	…	…	…

Matrix $\mathbf{Q}^n$ indicates the number of paths going from one node to another in *n* steps

Source: J. van Helden

From PAM1 to PAM2, PAM100, PAM250, etc…

$PAM2 = PAM1^{2}$

$PAM100 = PAM1^{100}$

$PAM250 = PAM1^{250}$

PAM1 相当于所有氨基酸平均有1%发生了变化。
PAM250 表示一种进化距离，数字越大，进化距离越远。
注意：PAM与进化时间之间没有大致对应关系，因为不同蛋白质家族的进化速率不同。
当两序列进行相似性比较时，不知道进化时间是恰当的。

PAM250：应用最广的替换矩阵

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	13	6	9	9	5	8	9	12	6	8	6	7	7	4	11	11	11	2	4	9
R	3	17	4	3	2	5	3	2	6	3	2	9	4	1	4	4	3	7	2	2
N	4	4	6	7	2	5	6	4	6	3	2	5	3	2	4	5	4	2	3	3
D	5	4	8	11	1	7	10	5	6	3	2	5	3	1	4	5	5	1	2	3
C	2	1	1	1	52	1	1	2	2	2	1	1	1	1	2	3	2	1	4	2
Q	3	5	5	6	1	10	7	3	7	2	3	5	3	1	4	3	3	1	2	3
E	5	4	7	11	1	9	12	5	6	3	2	5	3	1	4	5	5	1	2	3
G	12	5	10	10	4	7	9	27	5	5	4	6	5	3	8	11	9	2	3	7
H	2	5	5	4	2	7	4	2	15	2	2	3	2	2	3	3	2	2	3	2
I	3	2	2	2	2	2	2	2	2	10	6	2	6	5	2	3	4	1	3	9
L	6	4	4	3	2	6	4	3	5	15	34	4	20	13	5	4	6	6	7	13
K	6	18	10	8	2	10	8	5	8	5	4	24	9	2	6	8	8	4	3	5
M	1	1	1	1	0	1	1	1	1	2	3	2	6	2	1	1	1	1	1	2
F	2	1	2	1	1	1	1	1	3	5	6	1	4	32	1	2	2	4	20	3
P	7	5	5	4	3	5	4	5	5	3	3	4	3	2	20	6	5	1	2	4
S	9	6	8	7	7	6	7	9	6	5	4	7	5	3	9	10	9	4	4	6
T	8	5	6	6	4	5	5	6	4	6	4	6	5	3	6	8	11	2	3	6
W	0	2	0	0	0	0	0	0	1	0	1	0	0	1	0	1	0	55	1	0
Y	1	1	2	1	3	1	1	1	3	2	2	1	2	15	1	2	2	3	31	2
V	7	4	4	4	4	4	4	4	5	4	15	10	4	10	5	5	5	72	4	17

For clarity, the values have been multiplied by 100

Interpretation of the PAM250 matrix

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

In comparing 2 sequences at this evolutionary distance (250 PAM), there is:

* * * * **A** * * * * *

↓ **250 PAM**

* * * * **A** * * * * *	**probability of 13%**
* * * * **R** * * * * *	**probability of 3%**
* * * * **N** * * * * *	**probability of 4%**
* * * * **W** * * * * *	**probability of 0%**

...

PAM250 对数概率矩阵Log-odds of PAM250

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

$$S_{ij} = 10 \log_{10} \frac{M_{ij}}{P_i}$$

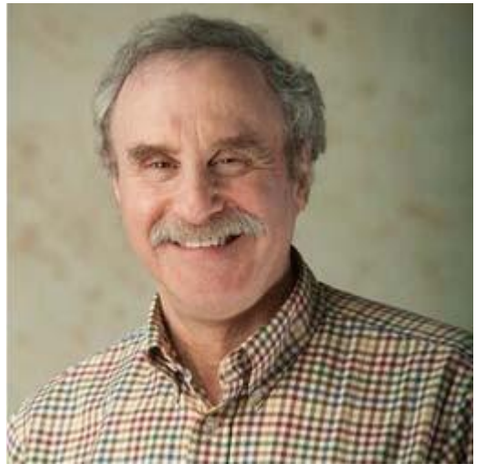
P_i: Probability of a_i in sequences

Log-odds matrix backs to symmetric

BLOSUM:
BLOcks
SUbstitution
Matrix

Scoring matrix BLOSUM

Henikoff S, **Henikoff JG**. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci U S A. 1992 Nov 15;89(22):10915-9



Steve Henikoff

HHMI Investigator NAS member

Jorja G. Henikoff

Henikoff

Steven Henikoff

Member in Basic Sciences, Fred Hutchinson Cancer Research Center
在 fhcrc.org 的电子邮件经过验证 - 首页

Genetics

引用次数		查看全部
	总计	2013 年至今
引用	71761	26214
h 指数	125	76
i10 指数	291	218

标题	引用次数	年份
Amino acid substitution matrices from protein blocks S Henikoff, JG Henikoff Proceedings of the National Academy of Sciences 89 (22), 10915-10919	5740	1992
Unidirectional digestion with exonuclease III creates targeted breakpoints for DNA sequencing S Henikoff Gene 28 (3), 351-359	4110	1984
Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm P Kumar, S Henikoff, PC Ng Nature protocols 4 (7), 1073	3763	2009
SIFT: Predicting amino acid changes that affect protein function PC Ng, S Henikoff Nucleic acids research 31 (13), 3812-3814	3062	2003

Scoring matrix BLOSUM

Dataset: >2000 blocks（蛋白质序列中的高度保守区，称为block）

Four steps for building the BLOSUM matrix:

Step 1: Count frequency table f_{ij}

Step 2: Calculate the observed occurrence probability q_{ij}

Step 3: Calculate the expected occurrence probability e_{ij}

Step 4: Calculate the log-odds matrix S_{ij}

直接利用多序列比对分析亲缘关系较远的蛋白质，而不是用近源蛋白序列。

优点：符合实际观测结果；

缺点：不能与进化挂钩。

总体上来说，BLOSUM矩阵比PAM矩阵更适合生物学关系的分析和局部相似性搜索。

Scoring matrix BLOSUM

Step 1: Count frequency table f_{ij}

A block of known conserved sequences (gapless):

```
LVLHVWAKVEADVAGHGQDILIRLFKSHPETLE
LVLWDWAKVEADVAGHGQDILIRLFKSHPETLE
LDLHVWAKVGGDVAGHGQAALIRLFKSHPETLE
LCLHVWAKVEADVAGGGQGGLIRLFKSHPETLE
DVLHVWAKVEADVAGHGQDILIRLFKSHPETLE
LVLHVWAKVEADVAGHGQDILIRLFKSHPETLE
```

DD pairs: 6
DA pairs: 4
DG pairs: 4
AG pairs: 1
Total pairs at this column: 6x5/2=15

Total pairs in all columns:

$$w \times s(s-1)/2$$

s: number of sequences,
w: number of columns

Scoring matrix BLOSUM

Step 2: Calculate the observed occurrence probability q_{ij}

Probability of occurrence of each i-j pairs:

$$q_{ij} = \frac{f_{ij}}{\sum_{i=1}^{20}\sum_{j=1}^{i} f_{ij}}, \quad 1 \le j \le i \le 20$$

Comparison with PAM

$$M_{ij} = \begin{cases} \lambda \dfrac{m_j A_{ij}}{\sum_{k=1}^{20} A_{kj}}, & 1 \le i, j \le 20; \quad i \ne j \\ 1 - \lambda m_j, & i = j \end{cases}$$

Scoring matrix BLOSUM

Step 3: Calculate the expected occurrence probability e_{ij}

1. Probability of occurrence of the i-th amino acid:

$$p_i = q_{ii} + \frac{1}{2}\sum_{j \neq i} q_{ij}, \quad 1 \leq i \leq 20$$

2. Expected probability of i-j pairs（在完全独立的情况下）:

$$e_{ij} = \begin{cases} p_i^2, & \text{if } i = j \\ 2 p_i p_j, & \text{otherwise} \end{cases}$$

Scoring matrix BLOSUM

Step 4: Calculate the log-odds matrix S_{ij}

$$S_{ij} = 2\log_2 \frac{q_{ij}}{e_{ij}}, \quad 1 \le j \le i \le 20$$

Comparison with PAM

$$S_{ij} = 10\log_{10} \frac{M_{ij}}{P_i}$$

Scoring matrix BLOSUM62

Sequence identity of the blocks is at least 62%

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

S_{ij}<0, probability is less than expected

S_{ij}>0, probability is more than expecte

BLOSUM vs. PAM

- BLOSUM系列比PAM系列好

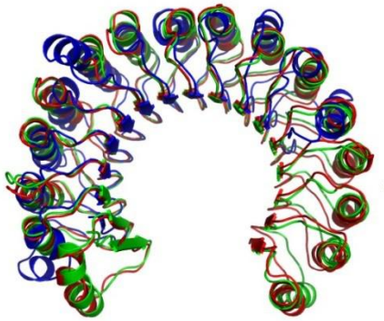
Matrix aligned	Program	Residue positions missed*	
		All positions	Side chains
	MSA	12	6
PAM 120	MULTALIN	31	22
PAM 160	MULTALIN	30	22
PAM 250	MULTALIN	30	22
+6/−1	MULTALIN	34	26
BLOSUM 45	MULTALIN	9	5
BLOSUM 62	MULTALIN	6	4
BLOSUM 80	MULTALIN	9	6

A potential research project

One of the major difficulty in the field is to detect remote-homology proteins. （远程同源蛋白）

How can we derive a matrix that is more suitable for aligning remote-homology proteins?

One way is probably to use structure alignment to construct blocks for the mutation matrix construction.



mTM-align

for

multiple protein structure alignment

空位罚分

Gap penalty

- What is alignment gap?

```
MVLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLEKFDRVKHLKTEAEMKASEDLK
SLEWMVNWAMVNWAAVV-------DDFYQELFKAHPEYQNKFGFFKAHPEYQNKFGFKGVALG
```

Gap opening　　Gap extension

- Gap penalty:

$$w(k) = a + b(k-1)$$

  - a: gap-opening penalty
  - b: gap-entension penalty (usually b≤ a)
  - k: length of the gaps

空位罚分a	空位扩展b	比对影响	适用
大	大	极少插入和缺失	非常相关蛋白质间的比对
大	小	少量大块插入	整个功能域可能插入的情况
小	大	大量小块插入	亲缘关系较远的蛋白质同源性分析

Gap penalty

$$Score = \sum_{i=1}^{N_{ali}} M(A_i, B_i) - GapPenalty$$

Question:

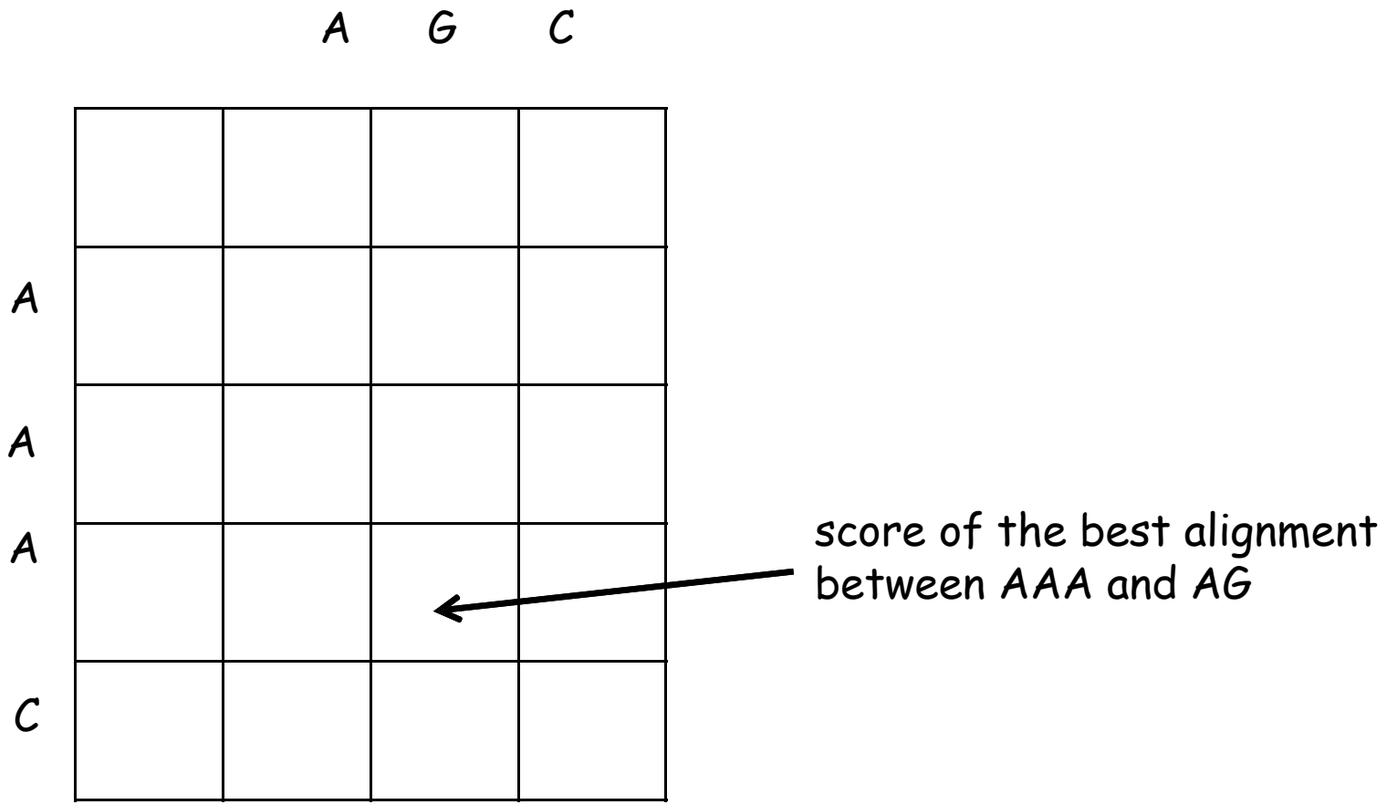
For a given score matrix and gap penalty protocol, how to find the best alignment of two protein sequences?

动态规划算法–全局比对

Global alignment: Needleman-Wunsch

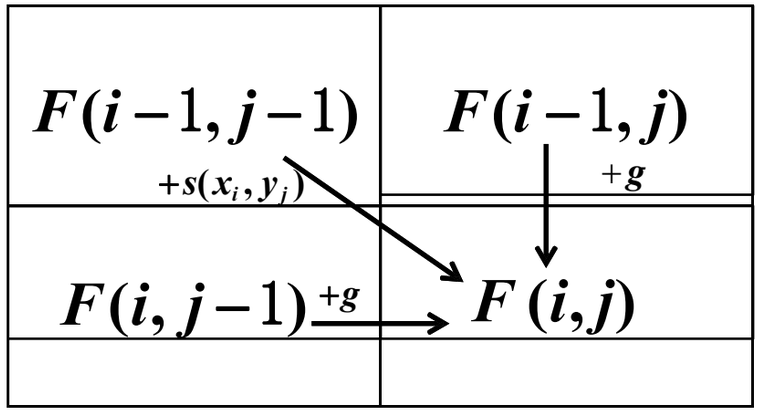
Needleman-Wunsch's dynamic programming (DP) idea

- Given an n-character sequence x, and an m-character sequence y
- Construct an *(n+1)x(m+1)* matrix *F(0…n, 0…m)*
- F(i,j)=score of the best alignment between x[1…i] and y[1…j]



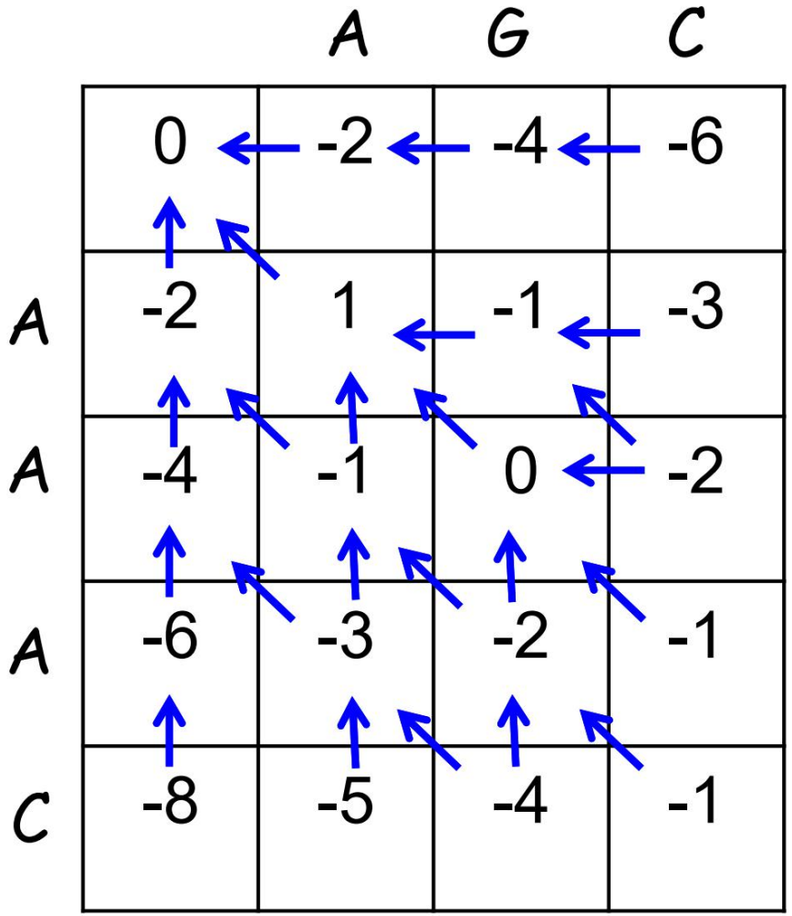
Iteration formula

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



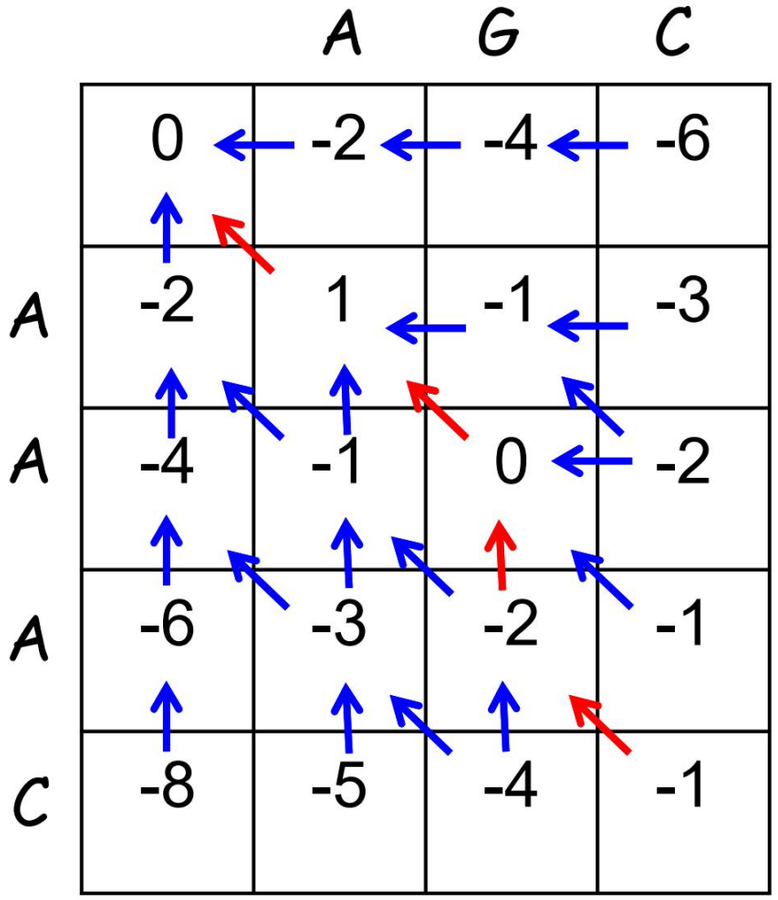
Two steps in Needleman-Wunsch algorithm

Step 1: Fill in the matrix F iteratively



Two steps in Needleman-Wunsch algorithm

Step 2: Traceback to find the optimal alignment



x: AAAC
y: AG-C

Two steps in Needleman-Wunsch algorithm

$$s(x_i, y_j) = \begin{cases} 1, & \text{if } x_i = y_j \\ -1, & \text{otherwise} \end{cases}$$

Gap penalty: g=-2

extension = opening

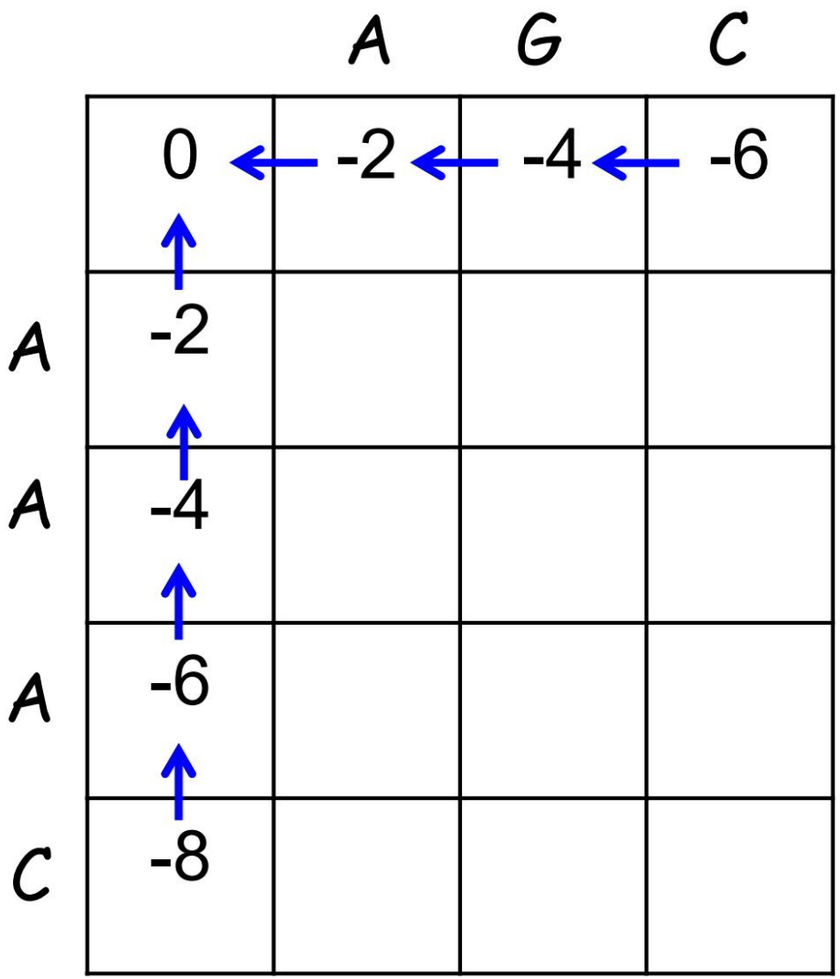
Step 1: Fill in the matrix F iteratively

Draw an (n+1)x(m+1) matrix F(o…n, o…m) first

		A	G	C
A				
A				
A				
C				

Initialize the 1st column and 1st row

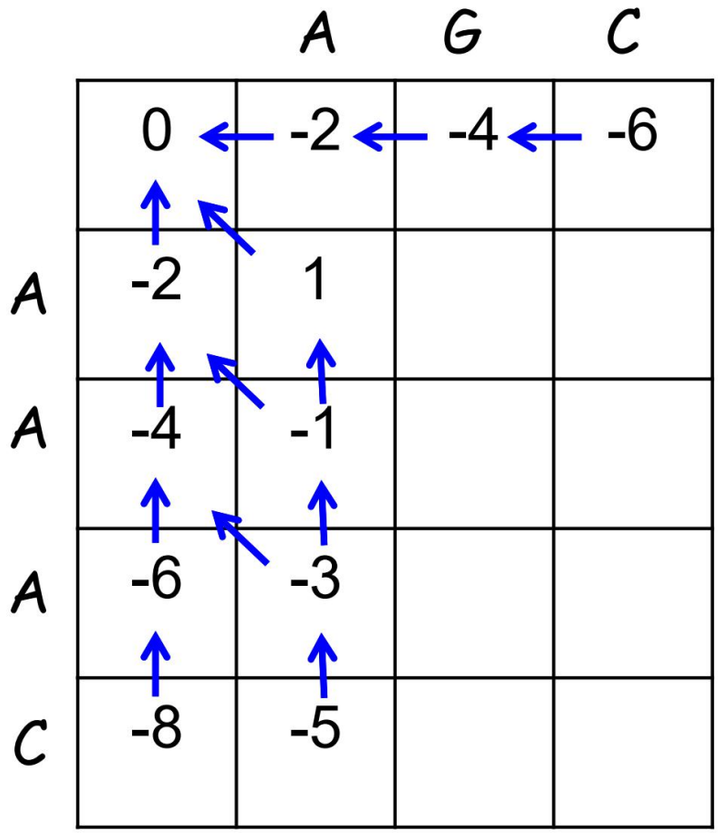
Gap penalty: g=-2 extensic opening



Begin filling in column-wise or row-wise order

$$s(x_i, y_j) = \begin{cases} 1, & \text{if } x_i = y_j \\ -1, & \text{otherwise} \end{cases}$$

Gap penalty: w(k)=-2k

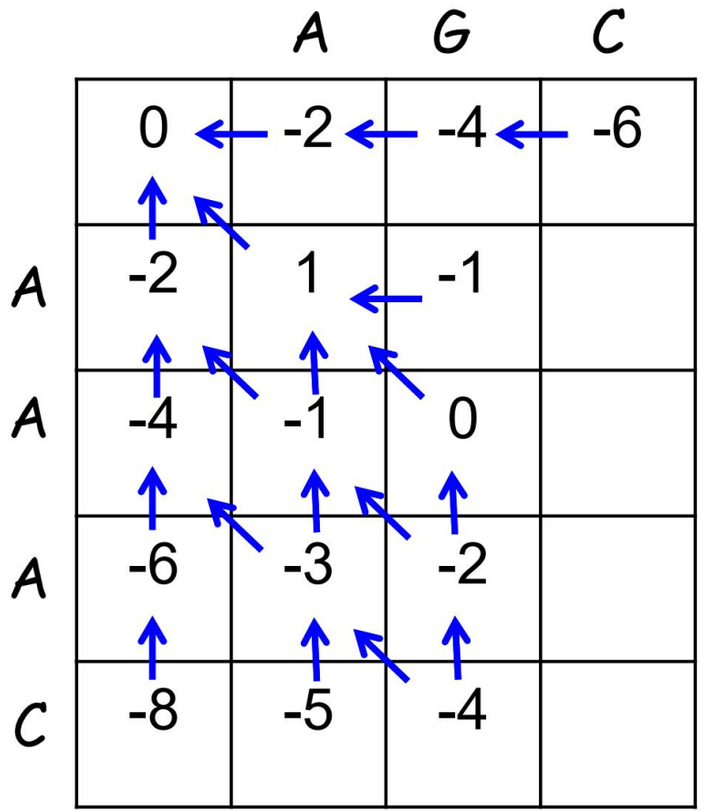


filling…

$$s(x_i, y_j) = \begin{cases} 1, & \text{if } x_i = y_j \\ -1, & \text{otherwise} \end{cases}$$

Gap penalty: -2

extension = opening

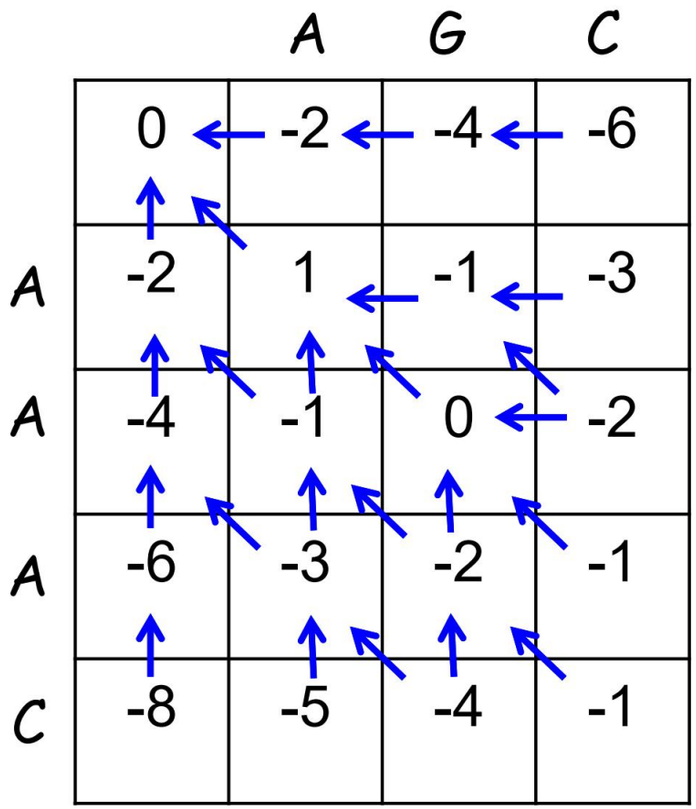


finally

$$s(x_i, y_j) = \begin{cases} 1, & \text{if } x_i = y_j \\ -1, & \text{otherwise} \end{cases}$$

Gap penalty: -2

extension = opening

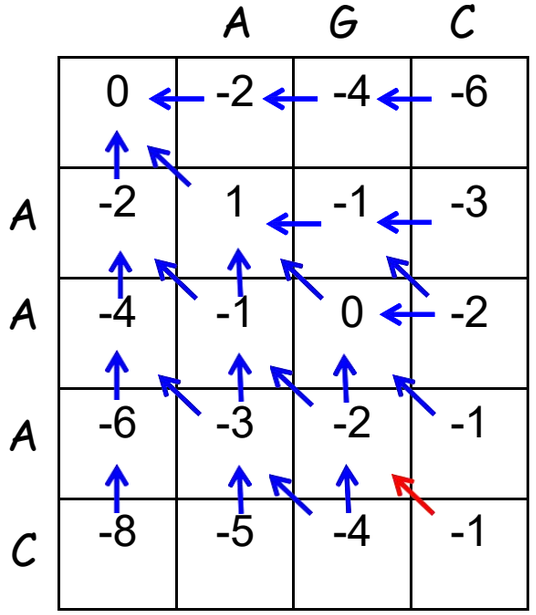


Traceback

Step 2: Traceback to find the optimal alignment

Starting from F(n,m) to F(0,0)

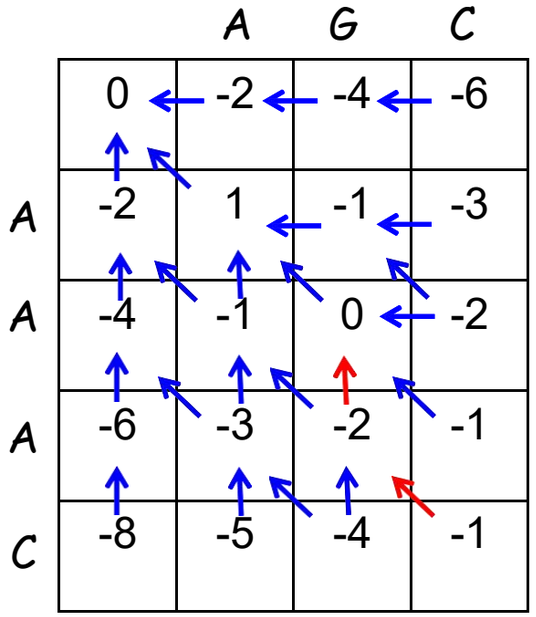
x: C
y: C



Traceback

Step 2: Traceback to find the optimal alignment

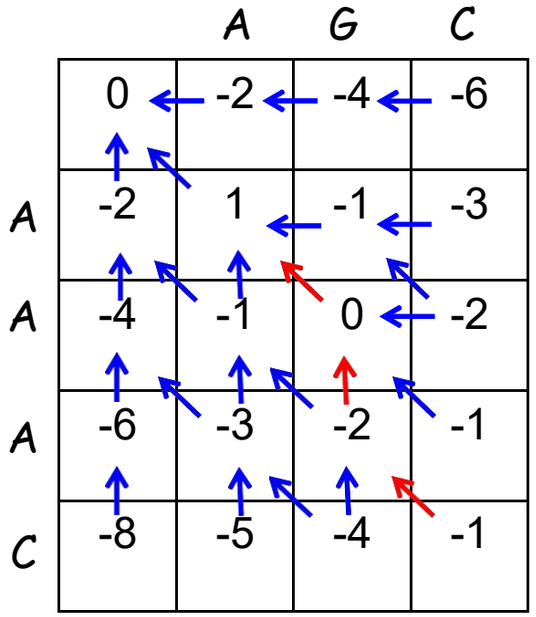
x: AC
y: −C



Traceback

Step 2: Traceback to find the optimal alignment

x: AAC
y: G–C

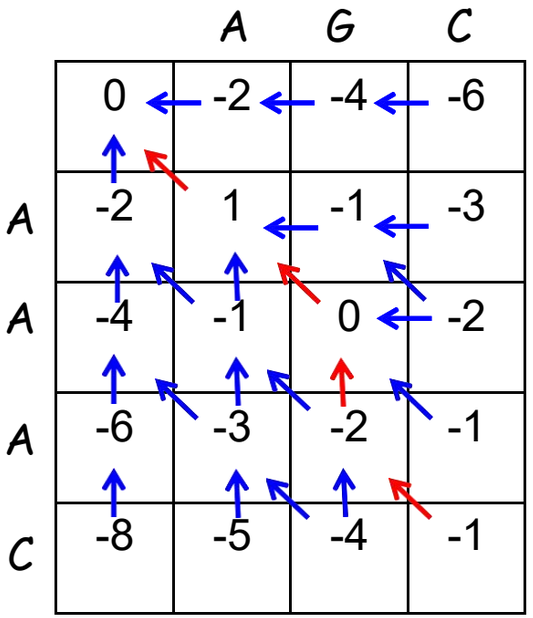


Traceback

Step 2: Traceback to find the optimal alignment

one optimal alignment

x: AAAC
y: AG-C

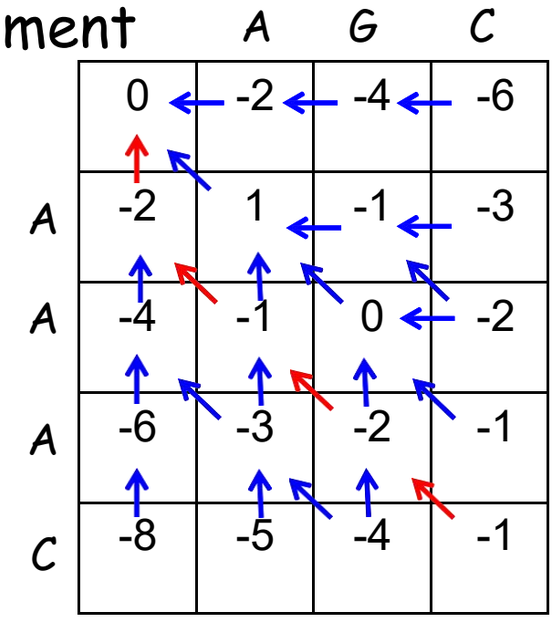


Traceback

Step 2: Traceback to find the optimal alignment

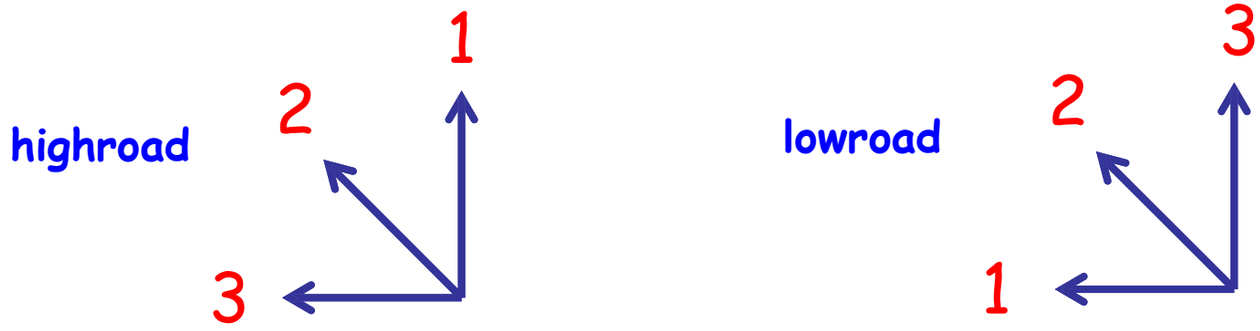
another optimal alignment

x: AAAC
y: −AGC



Equally optimal alignments

can use preference ordering over paths when doing traceback



Another example

$$s(x_i, y_j) = \begin{cases} 2, & \text{if } x_i = y_j \\ -3, & \text{otherwise} \end{cases}$$

Gap penalty: g=-2

extension = opening

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

Another example

$$s(x_i, y_j) = \begin{cases} 2, & \text{if } x_i = y_j \\ -3, & \text{otherwise} \end{cases}$$

Gap penalty: g=-2

extension = opening

		A	C	T	G	A	T	T	C	A
	0	-2	-4	-6	-8	-10	-12	-14	-16	-18
A	-2	2								
C	-4	0								
G	-6	-2								
C	-8	-4								
A	-10	-6								
T	-12	-8								
C	-14	-10								
A	-16	-12								

Another example

$$s(x_i, y_j) = \begin{cases} 2, & \text{if } x_i = y_j \\ -3, & \text{otherwise} \end{cases}$$

Gap penalty: g=-2

extension = opening

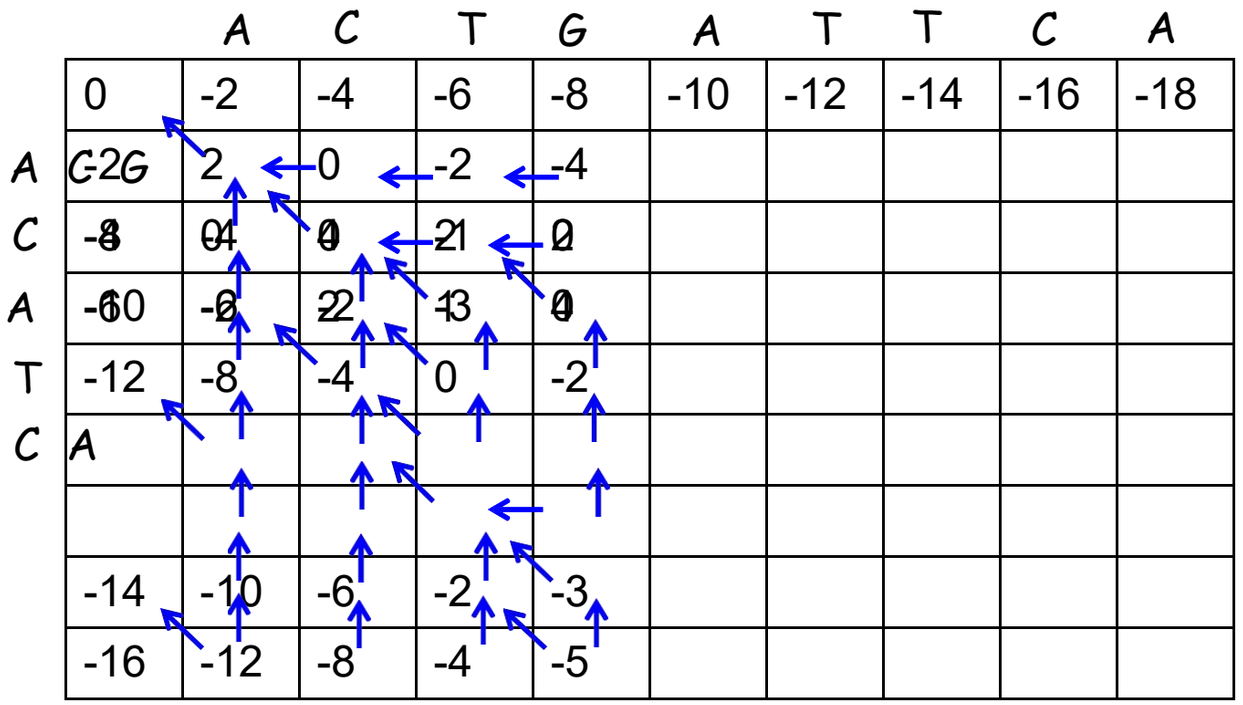
		A	C	T	G	A	T	T	C	A
	0	-2	-4	-6	-8	-10	-12	-14	-16	-18
A	-2	2	0							
C	-4	0	4							
G	-6	-2	2							
C	-8	-4	0							
A	-10	-6	-2							
T	-12	-8	-4							
C	-14	-10	-6							
A	-16	-12	-8							

Another example

$$s(x_i, y_j) = \begin{cases} 2, & \text{if } x_i = y_j \\ -3, & \text{otherwise} \end{cases}$$

Gap penalty: g=-2

extension = opening



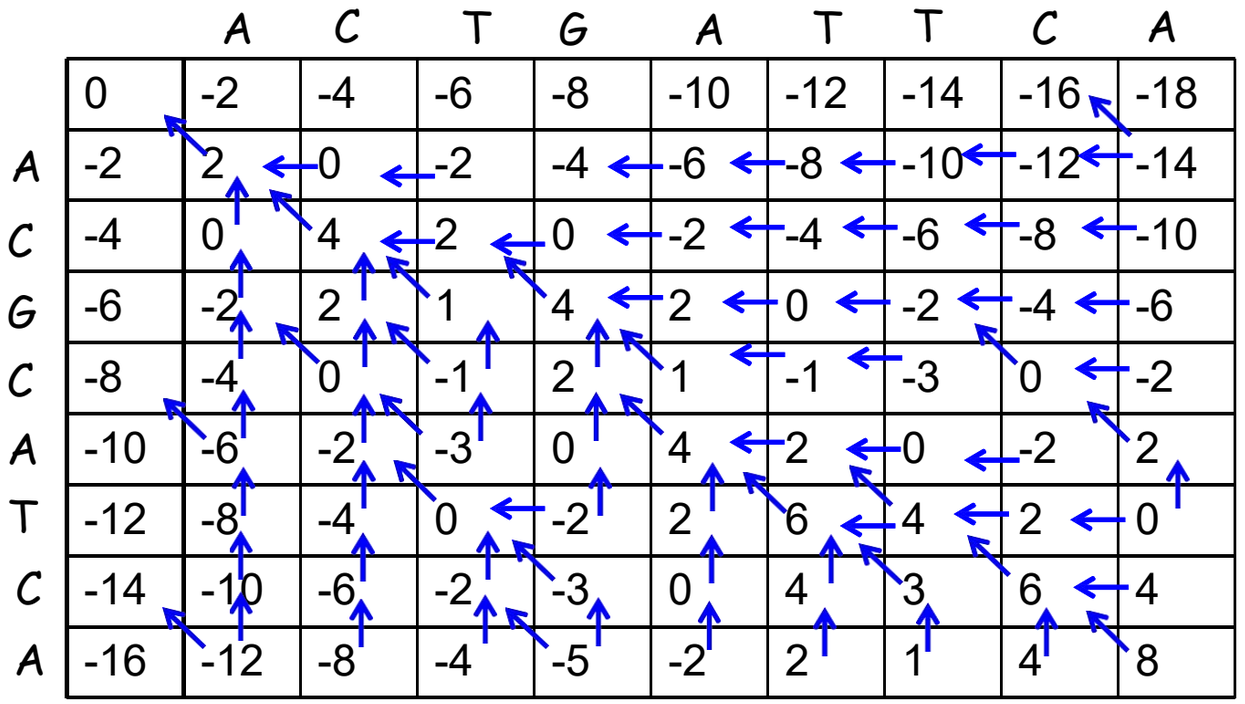
		A	C	T	G	A	T	T	C	A
	0	-2	-4	-6	-8	-10	-12	-14	-16	-18
A	-2	2	0	-2	-4					
C	-4	0	4	2	0					
A	-6	-2	2	3	1					
T	-8	-4	0	5	3					
C	-10	-6	-2	3	4					
	-12	-8	-4	1	2					
	-14	-10	-6	-1	0					
	-16	-12	-8	-3	-2					

Another example

$$s(x_i, y_j) = \begin{cases} 2, & \text{if } x_i = y_j \\ -3, & \text{otherwise} \end{cases}$$

Gap penalty: g=-2

extension = opening

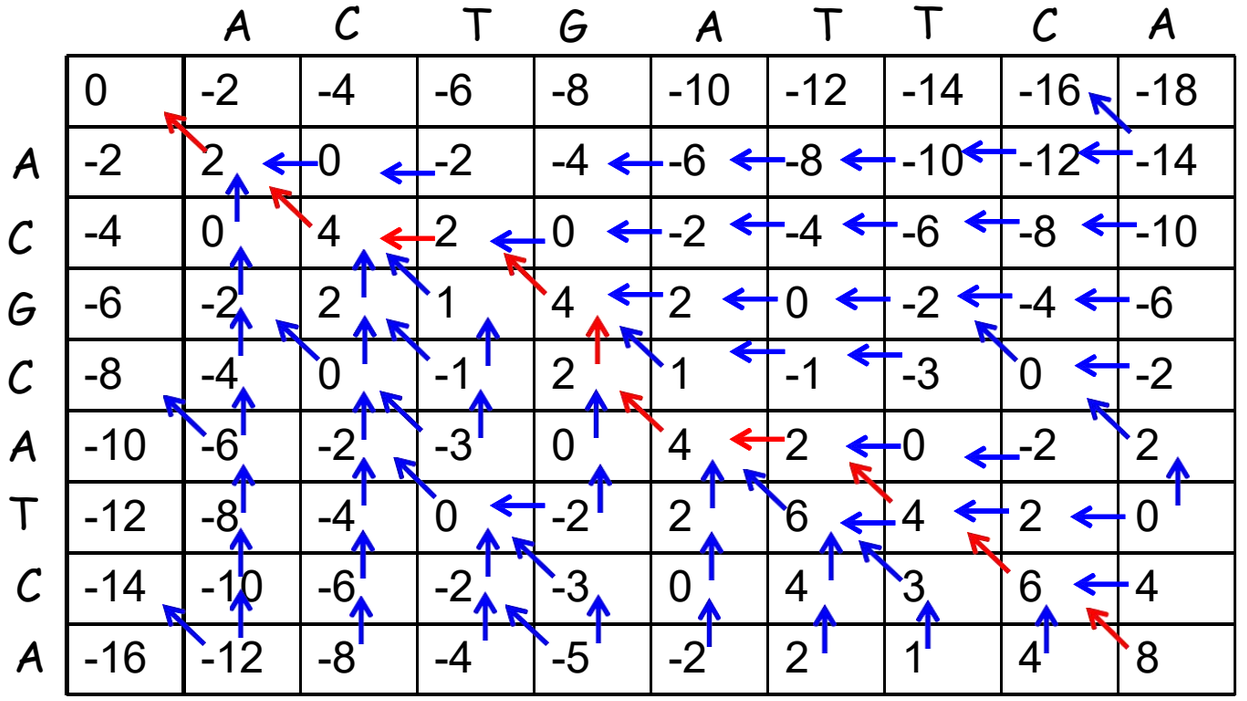


		A	C	T	G	A	T	T	C	A
	0	-2	-4	-6	-8	-10	-12	-14	-16	-18
A	-2	2	0	-2	-4	-6	-8	-10	-12	-14
C	-4	0	4	2	0	-2	-4	-6	-8	-10
G	-6	-2	2	1	4	2	0	-2	-4	-6
C	-8	-4	0	-1	2	1	-1	-3	0	-2
A	-10	-6	-2	-3	0	4	2	0	-2	2
T	-12	-8	-4	0	-2	2	6	4	2	0
C	-14	-10	-6	-2	-3	0	4	3	6	4
A	-16	-12	-8	-4	-5	-2	2	1	4	8

Another example

one optimal alignment

x: AC-GCA-TCA y:
ACTG-ATTCA

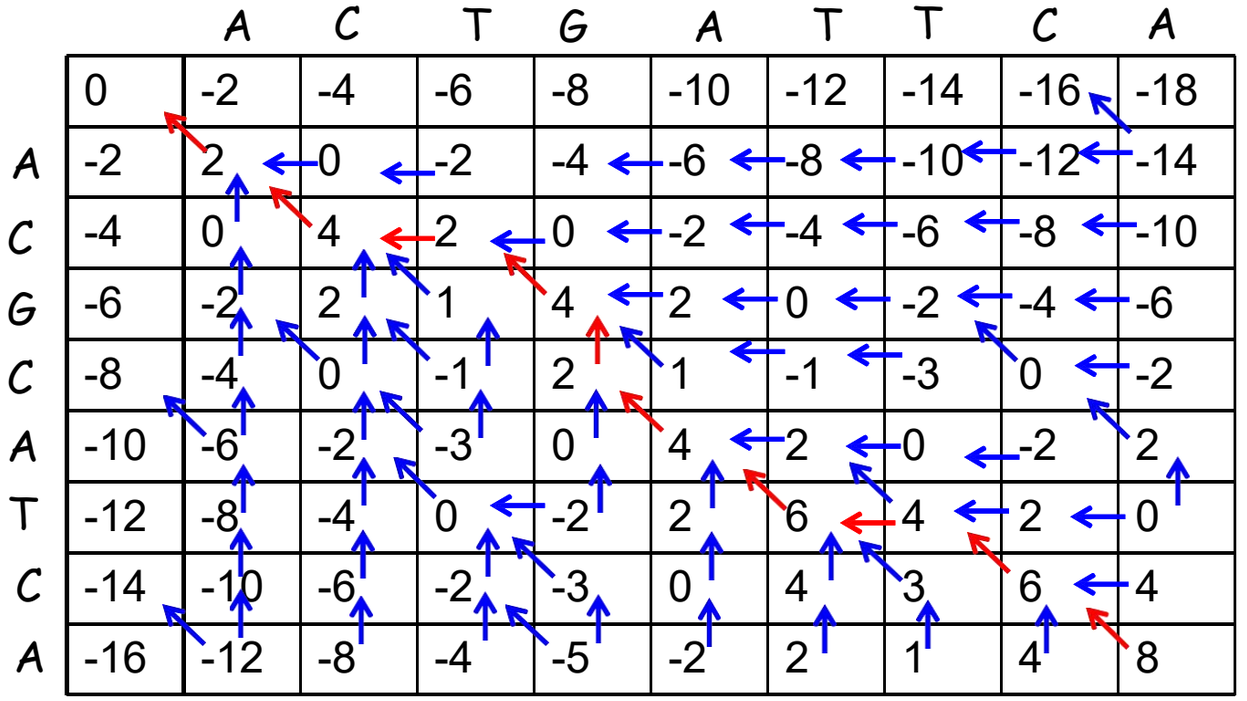


		A	C	T	G	A	T	T	C	A
	0	-2	-4	-6	-8	-10	-12	-14	-16	-18
A	-2	2	0	-2	-4	-6	-8	-10	-12	-14
C	-4	0	4	2	0	-2	-4	-6	-8	-10
G	-6	-2	2	1	4	2	0	-2	-4	-6
C	-8	-4	0	-1	2	1	-1	-3	0	-2
A	-10	-6	-2	-3	0	4	2	0	-2	2
T	-12	-8	-4	0	-2	2	6	4	2	0
C	-14	-10	-6	-2	-3	0	4	3	6	4
A	-16	-12	-8	-4	-5	-2	2	1	4	8

Another example

another optimal alignment

x: AC-GCAT-CA y: ACTG-ATTCA



		A	C	T	G	A	T	T	C	A
	0	-2	-4	-6	-8	-10	-12	-14	-16	-18
A	-2	2	0	-2	-4	-6	-8	-10	-12	-14
C	-4	0	4	2	0	-2	-4	-6	-8	-10
G	-6	-2	2	1	4	2	0	-2	-4	-6
C	-8	-4	0	-1	2	1	-1	-3	0	-2
A	-10	-6	-2	-3	0	4	2	0	-2	2
T	-12	-8	-4	0	-2	2	6	4	2	0
C	-14	-10	-6	-2	-3	0	4	3	6	4
A	-16	-12	-8	-4	-5	-2	2	1	4	8

作业

全局比对作业HomeWork

sequence x: GAATTCAGTTA

sequence y: GGATCGA

Score matrix: $s(x_i, y_j) = \begin{cases} 2, & \text{if } x_i = y_j \\ -1, & \text{otherwise} \end{cases}$

Gap penalty: g=-2

extension = opening

请按照全局比对算法，列出打分矩阵，回溯路径，以及比对后的结果。

Questions to think about

- How about local alignment?
- What happens if gap opening penalty and gap extension penalties are not equal?

Paper to read :

- T F Smith & M S Waterman, Identification of common molecular subsequences. J Mol Biol (1981) 147, 195-197.
- O. Gotoh. An improved algorithm for matching biological sequences. Journal of Molecular Biology 162 705-708 1982.