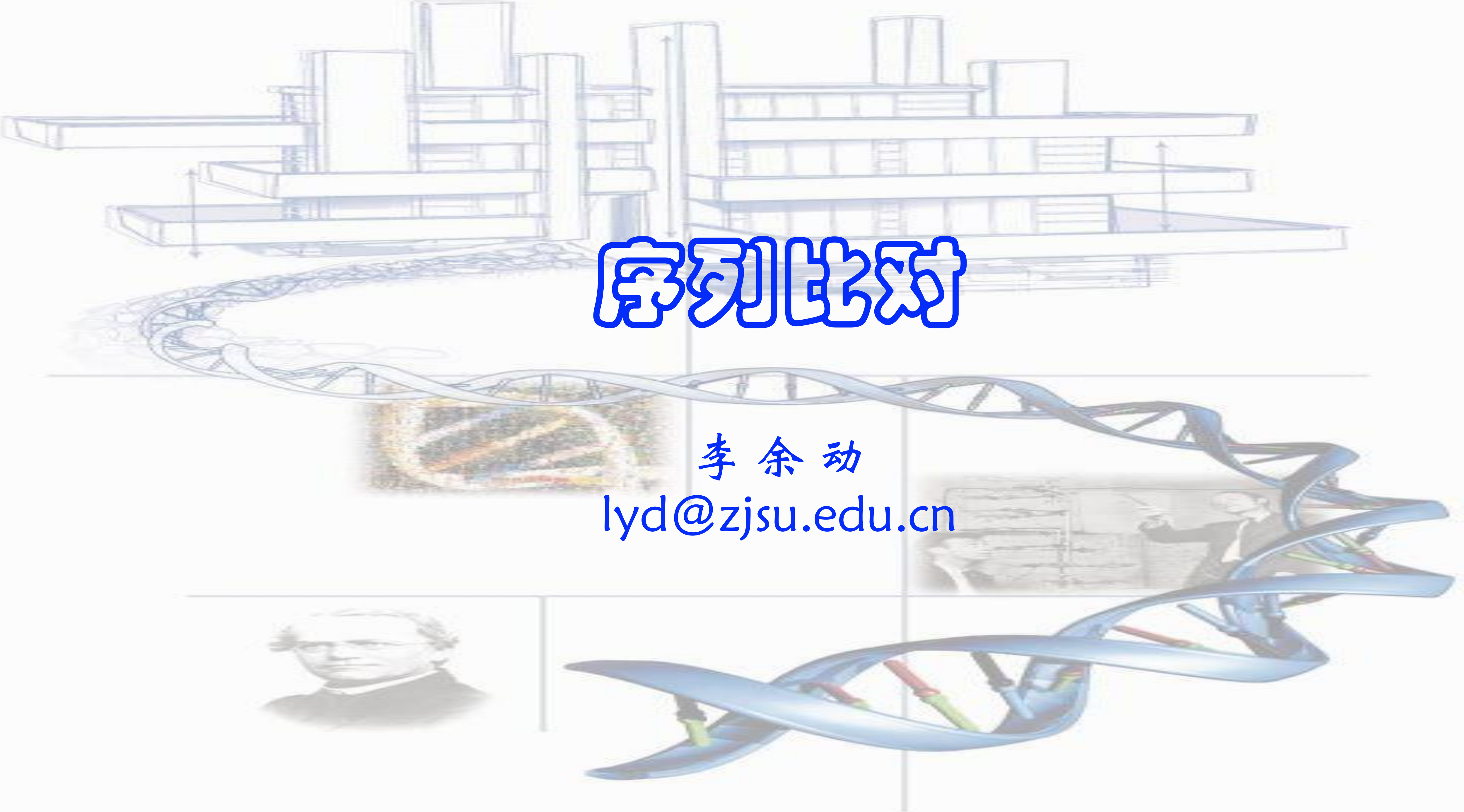


序列比对

李 余 动
lyd@zjsu.edu.cn

序列比对基础

- 比较是科学研究的常见方法之一，通过将研究对象进行相互比较，以寻找研究对象可能具备的某些特征和特性。
- 序列比对是生物信息学最基本的操作之一。

理论基础

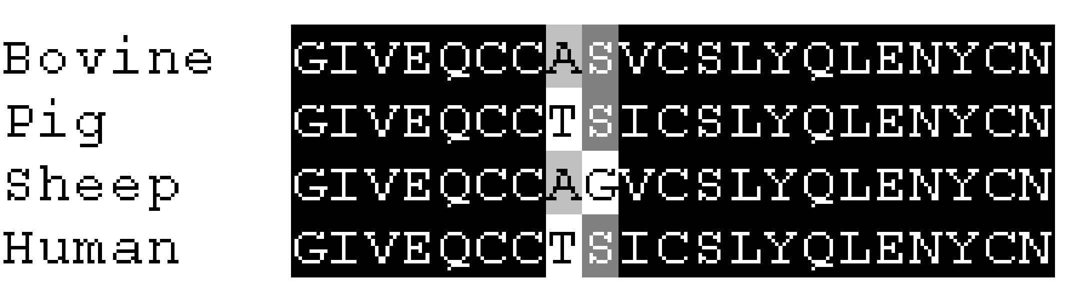
进化学说——不同序列不是随意产生，而是在进化上，不断发展演变而来

基本假设

生物学中序列决定结构，结构决定功能的普遍规律

序列比对(Sequence Alignment)

- 定义：运用某种特定的数学算法，找出两个或多个序列之间的最大匹配碱基或氨基酸残基数，判断序列之间的相似程度，从而推测它们的结构、功能及进化上的联系。



```
Bovine    GIVEQCCASVCSLYQLENYCN
Pig       GIVEQCCTSICSLYQLENYCN
Sheep     GIVEQCCAGVCSLYQLENYCN
Human     GIVEQCCTSICSLYQLENYCN
```

- 分类：

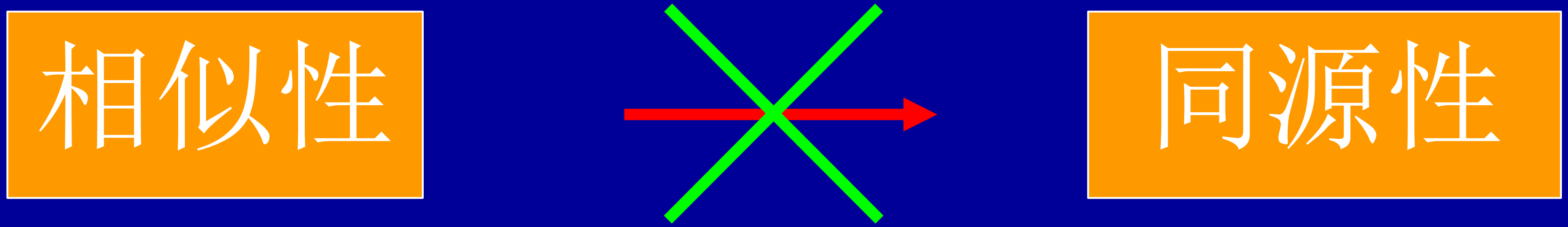
➢ 双序列比对 (Pairwise Alignment)：两条序列

➢ 多序列比对(Multiple Sequence Alignment, MSA)：三条或以上序列

➢ 全局比对 (Global Alignment)： 全长序列

➢ 局部比对(Local Alignment)：部分子序列

相似性(Similarity)和同源性(Homology)



相似性

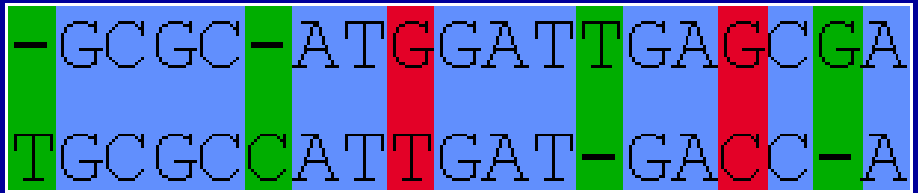
同源性

(一致性)

Similarity = an observable quantity often expressed as % identity.

Homology = ? (hint: there are no degrees of homology).

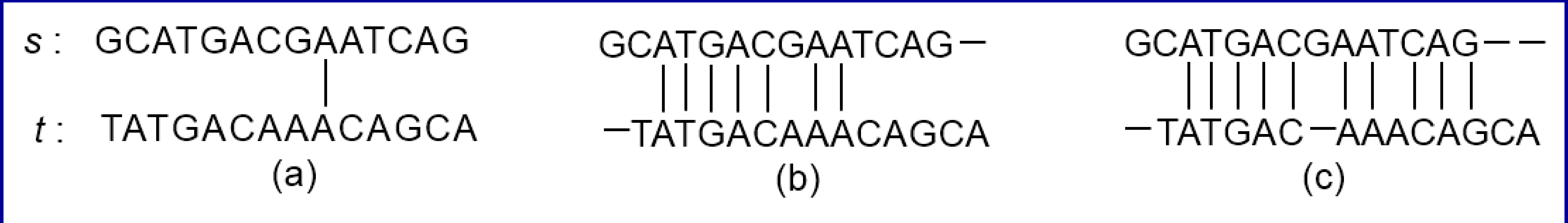
序列比对



```
-GCGC-ATGGATTGAGCGA
TGCGCCATTGAT-GACC-A
```

- 字符相同： match/identity
- 字符替代(mismatch/replace)： 氨基酸/碱基之间的替代和突变
- 插入和缺失(Insertion/Deletion, indel)
- 空位(gap): 由插入或删除事件引起的变化
  - ➢一般用横杠‘-’表示空位

序列s与序列t的三种比对结果



```
s : GCATGACGAATCAG        GCATGACGAATCAG-        GCATGACGAATCAG--
            |               |||||  ||              ||||| || |||
t : TATGACAAACAGCA        -TATGACAAACAGCA        -TATGAC-AAACAGCA
         (a)                    (b)                     (c)
```

- 序列比对根据序列的条数和每条序列长度不同往往有多种结果。
- 考虑由插入/删除事件引起的空位（空位罚分），导致比对的复杂性大大增加。

如何确定最优的比对结果？

Three key steps to answer if two sequences are related

(1) The scoring (打分) system used to rank alignments;

(2) The algorithm (算法) used to find optimal (or good) scoring alignments;

(3) The statistical (统计) methods used to evaluate the significance of an alignment score.

双序列比对打分

序列1：	V	D	S	–	C	Y
序列2：	V	E	S	L	C	Y
比对分数：	1	0	1	-1	1	1

打分模型

假设打分

- **匹配得分：1**
- **失配得分：0**
- **空位罚分：－1**

- 两序列比对的总分：
- Score=Σ(AA pair scores) – gap penalty = 3

常见的DNA替换记分矩阵

1、等价矩阵：相同核苷酸之间的匹配得分为1，不同核苷酸间的替换得分为0

	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、转换-颠换矩阵：核酸的碱基按照结构特征被划分为两类，嘌呤类（A、G）和嘧啶类（C、T）。碱基在类之间的替换称为转换，得分为-1，在类与类之间的替换成为颠换，得分为-5。

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

3、BLAST矩阵：被比对的两个核苷酸相同时得分为+5，反之为-4

	A	T	C	G
A	5	-4	-4	-4
T	-4	5	-4	-4
C	-4	-4	5	-4
G	-4	-4	-4	5

常见的蛋白质替换记分矩阵

PAM(Point Accepted Mutation)

- PAM—N矩阵是从蛋白质序列的全局比对结果中推导而来的。
- 基础的PAM-1矩阵反映进化产生的每一百个氨基酸平均发生一个突变的量值(统计方法得到)。
- PAM-1自乘n次，可以得到PAM-n，即发生了更多次突变，如PAM-250。
- PAM矩阵用于寻找蛋白质的进化起源。

BLOSUM(Blocks Substitution Matrix)

- BLOSUM—N矩阵则基于蛋白质序列的局部比对块，包含较远的相关序列。
- BLOSUM矩阵的相似性是根据真实数据产生的
- BLOSUM矩阵的编号，比如BLOSUM-80中的80，代表该矩阵是由一致度80%的序列计算而来的。
- BLOSUM矩阵用于发现蛋白质的保守区域。

PAM 矩阵与BLOSUM矩阵

氨基酸差异%	PAM	BLOSUM
1	PAM-1	BLOSUM-99
10	PAM-11	BLOSUM-90
20	PAM-23	BLOSUM-80
30	PAM-38	BLOSUM-70
40	PAM-56	BLOSUM-60
50	PAM-80	BLOSUM-50
60	PAM-112	BLOSUM-40
70	PAM-159	BLOSUM-30
80	PAM-246	BLOSUM-20

PAM-250矩阵

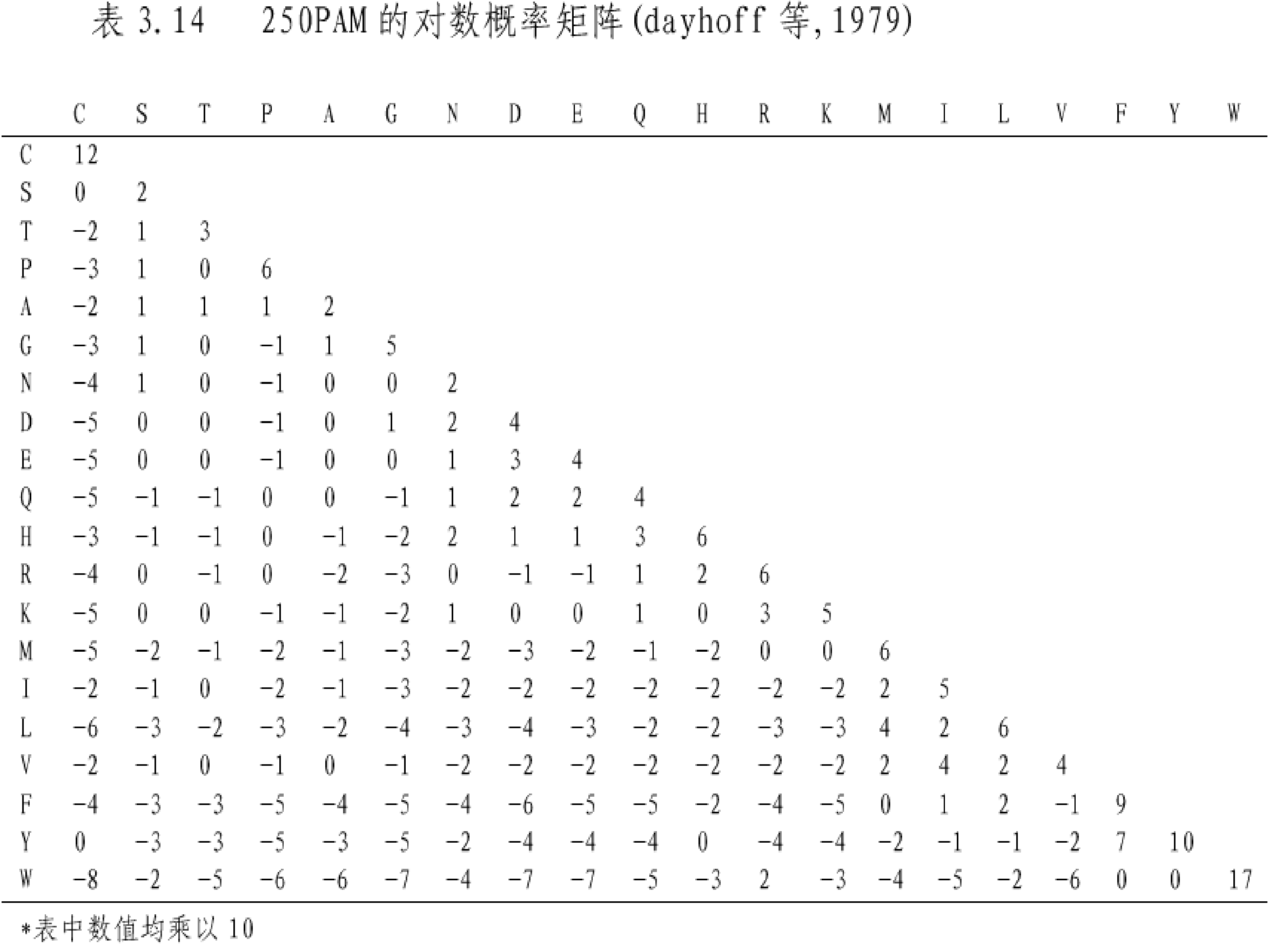


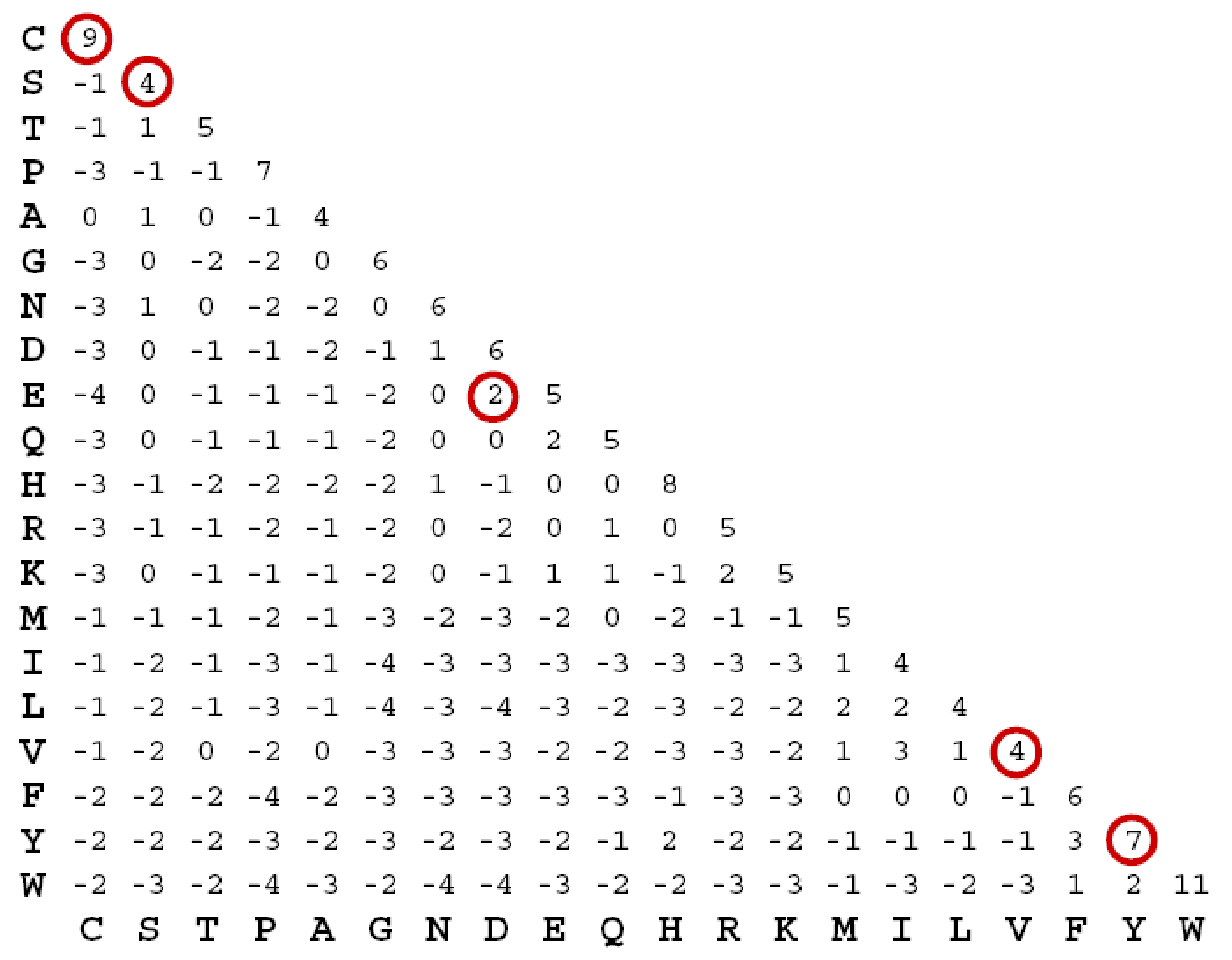
表 3.14　250PAM 的对数概率矩阵(dayhoff 等,1979)

	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	6								
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	-6	-3	-2	-3	-2	-4	-3	-4	-3	-2	-2	-3	-3	4	2	6				
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

*表中数值均乘以 10

- 对角线上的数值为匹配氨基酸的得分;
- 其他位置上，≥0的得分代表对应氨基酸对为相似氨基酸。

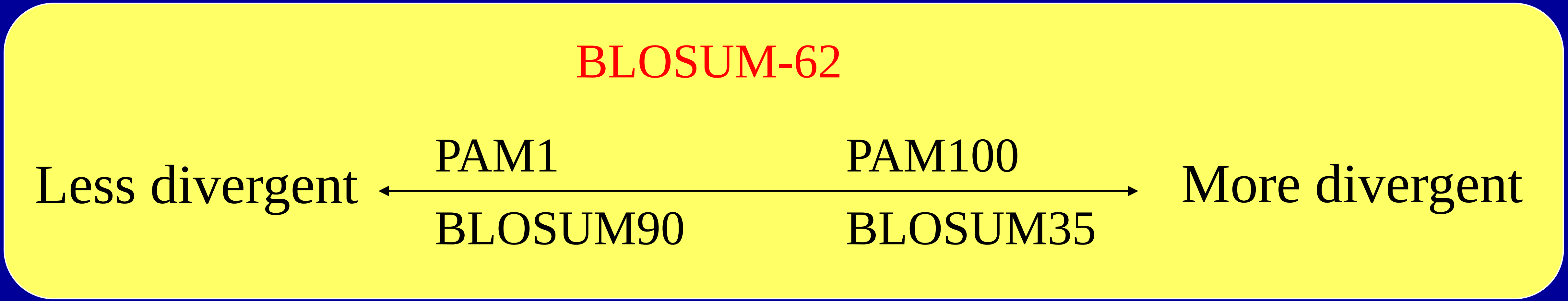
BLOSUM-62替换矩阵



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

- 在BLOSUM62矩阵中可以看到D->E替换分值为2，
- 分数越高，两氨基酸间越容易发生突变。

选 PAM-n? 还是BLOSUM-n?



- 对于关系较远的序列之间的比较，由于PAM-250是推算而来，所以其准确度受到一定限制，BLOSUM-45更具优势。
- 对于关系较近的序列之间的比较，用PAM或BLOSUM矩阵做出的比对结果，差别不大。
- 最常用的:BLOSUM-62

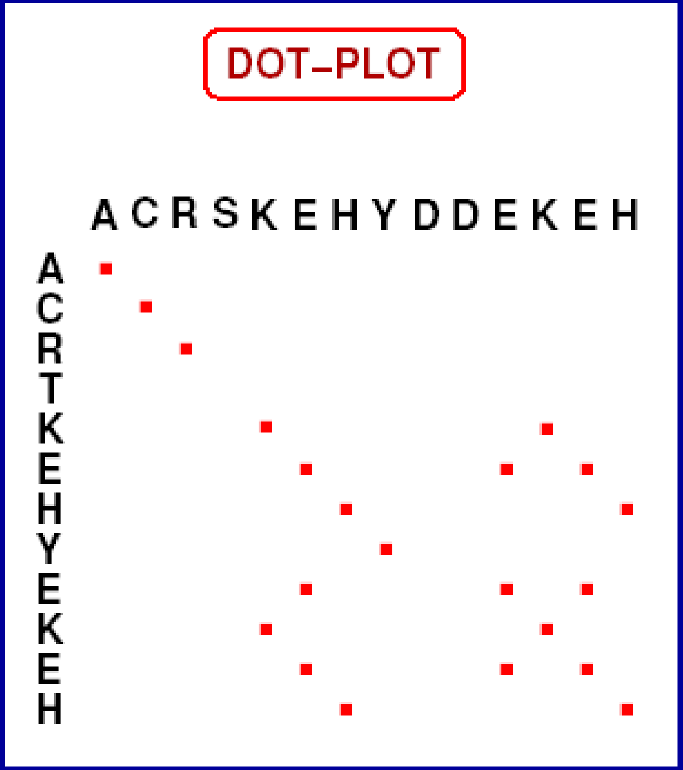
Three principle methods of pair-wise sequence alignment

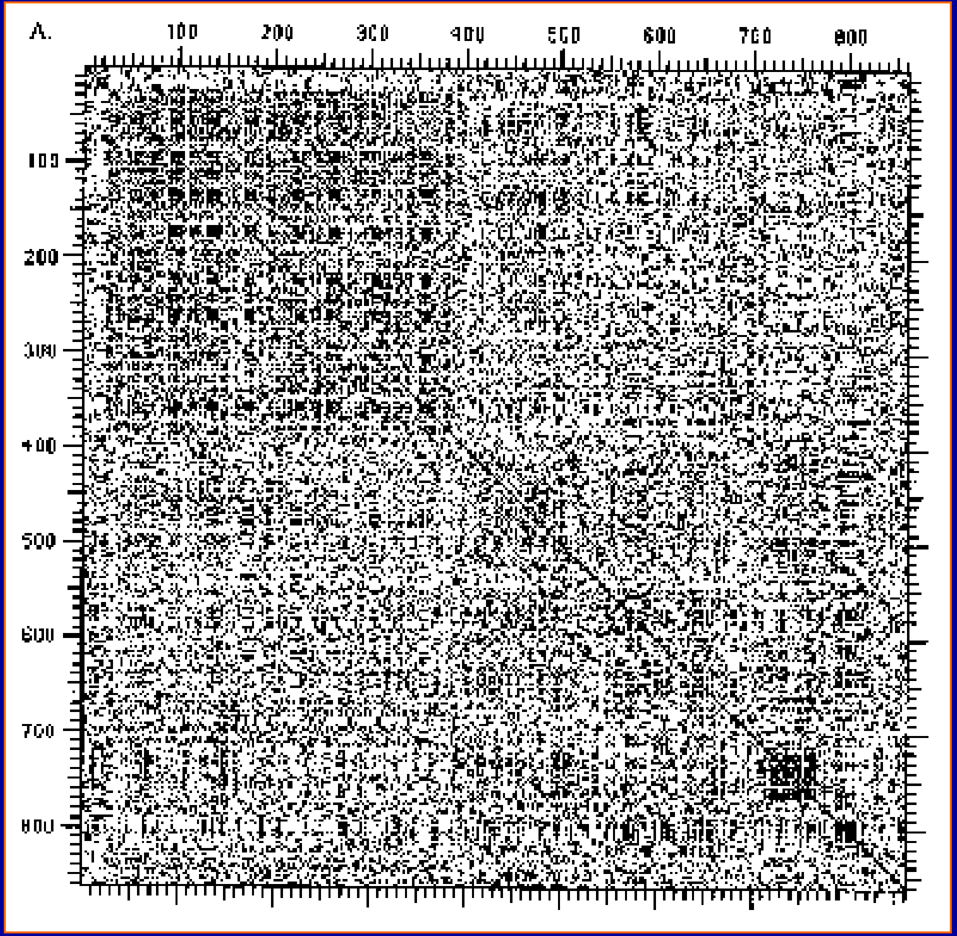
- **Dot matrix (点阵) pair-wise sequence comparison**

- **The dynamic programming (DP，动态规划) algorithm**
  - ***Needleman and Wunsch (1970)***
  - ***Smith and Waterman (1981)***

- **Word or *k*-tuple methods (字符串或*k*-元法)**
  - **heuristic algorithms, used by the programs of FASTA and BLAST**

点阵法

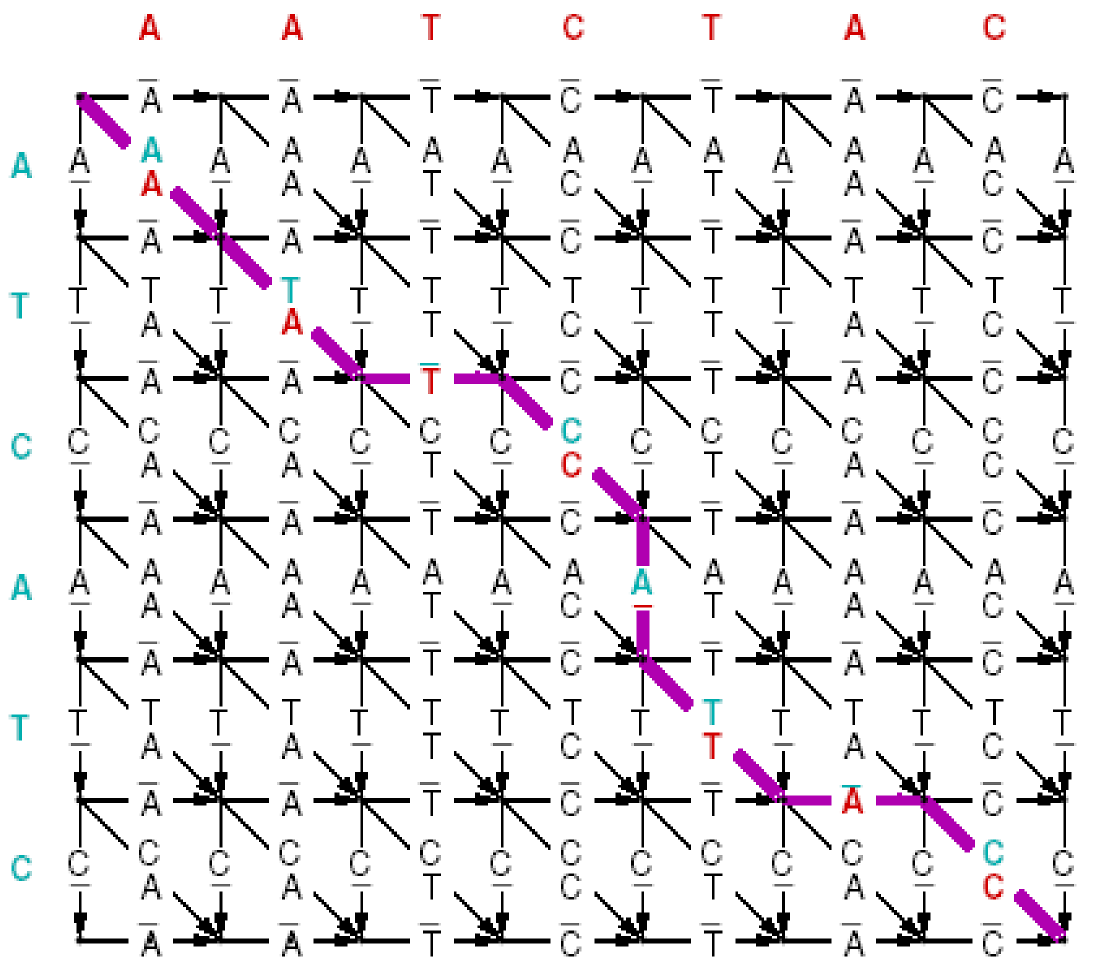




Dot matrix analysis of the human LDL receptor against itself

动态规划算法

- 这个方法把一个问题分解成计算合理的子问题，并使用这些子问题的结果来计算最终答案。
- 序列比对中某一位点有三种可能性：
  - 匹配、不匹配和空位。
- 动态规往往被用于一个复杂的空间中寻找一条最优路径。



Alignment corresponding to the colored path:

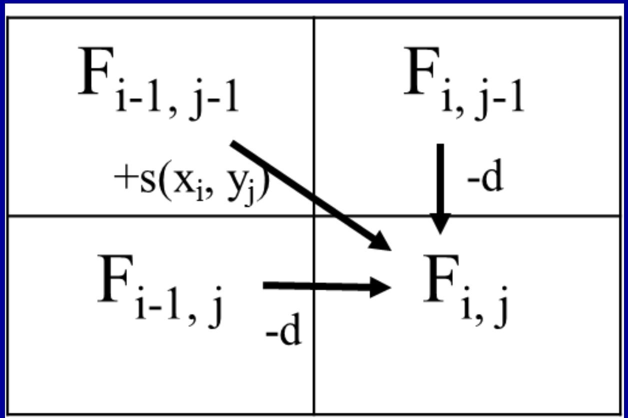
A T – C A T – C
A A T C – T A C

两条序列VESLCY和VDSCY比对第一位点的三种情况

(1)两条序列都不加空位

(2)给第一条序列加一个空位

(3)给第二条序列加一个空位



$F_{i-1,\, j-1}$

$F_{i,\, j-1}$

$+s(x_i,\, y_j)$

-d

$F_{i-1,\, j}$

-d

$F_{i,\, j}$

第一位点	得分	待比对的剩余序列
V	+1	ESLCY
V		DSCY
_	-1	VESLCY
V		DSCY
V	-1	ESLCY
_		VDSCY

全局比对(Global alignment)

	Gap	V	D	S	C	Y
Gap	0	1gap	2gap	…		
V	1gap					
E	2gap					
S	…					
L						
C						
Y						

本例：线性罚分

$$r(g) = -gd$$

全局比对 (2)

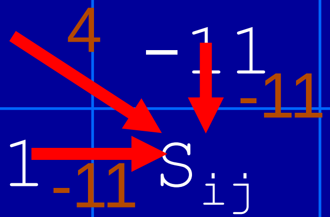
	Gap	V	D	S	C	Y
Gap	0	-11	-22	-33	-44	-55
V	-11	S_{ij}				
E	-22					
S	-33					
L	-44					
C	-55					
Y	-66					

要求解S_{ij}的分数，我们必须先知道$S_{i-1, j-1}$, $S_{i-1, j}$, 以及$S_{i, j-1}$的分数，这种方法叫做递归算法；采用这种方法，可以把大的问题分割成小的问题逐一解决，即动态规划算法；需要存储如何得到S_{ij}分数的过程。

$$\begin{tikzcd}
F_{i-1,\, j-1} \arrow[dr, "+s(x_i,\, y_j)"'] & F_{i,\, j-1} \arrow[d, "-d"] \\
F_{i-1,\, j} \arrow[r, "-d"'] & F_{i,\, j}
\end{tikzcd}$$

全局比对 (3)

	Gap	V	D	S	C	Y
Gap	0	-11	-22	-33	-44	-55
V	-11	S_{ij}				
E	-22					
S	-33					
L	-44					
C	-55					
Y	-66					



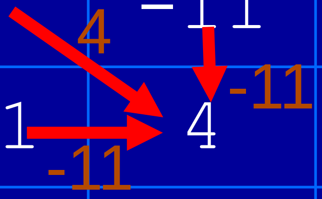
Needleman-Wunsch算法；

时间复杂度O(n^2)；

$$S_{ij} = \text{max of} \begin{cases} S_{i-1,\, j-1} + \sigma(x_i,\, y_j) \\ S_{i-1,\, j} - d \text{ (从左到右)} \\ S_{i,\, j-1} - d \text{ (从上到下)} \end{cases}$$

全局比对 (4)

	Gap	V	D	S	C	Y
Gap	0	-11	-22	-33	-44	-55
V	-11	4				
E	-22					
S	-33					
L	-44					
C	-55					
Y	-66					



全局比对 (5)

	Gap	V	D	S	C	Y
Gap	0	-11	-22 -3	-33	-44	-55
V	-11	4 -11	S_{ij} -11			
E	-22					
S	-33					
L	-44					
C	-55					
Y	-66					

全局比对 (6)

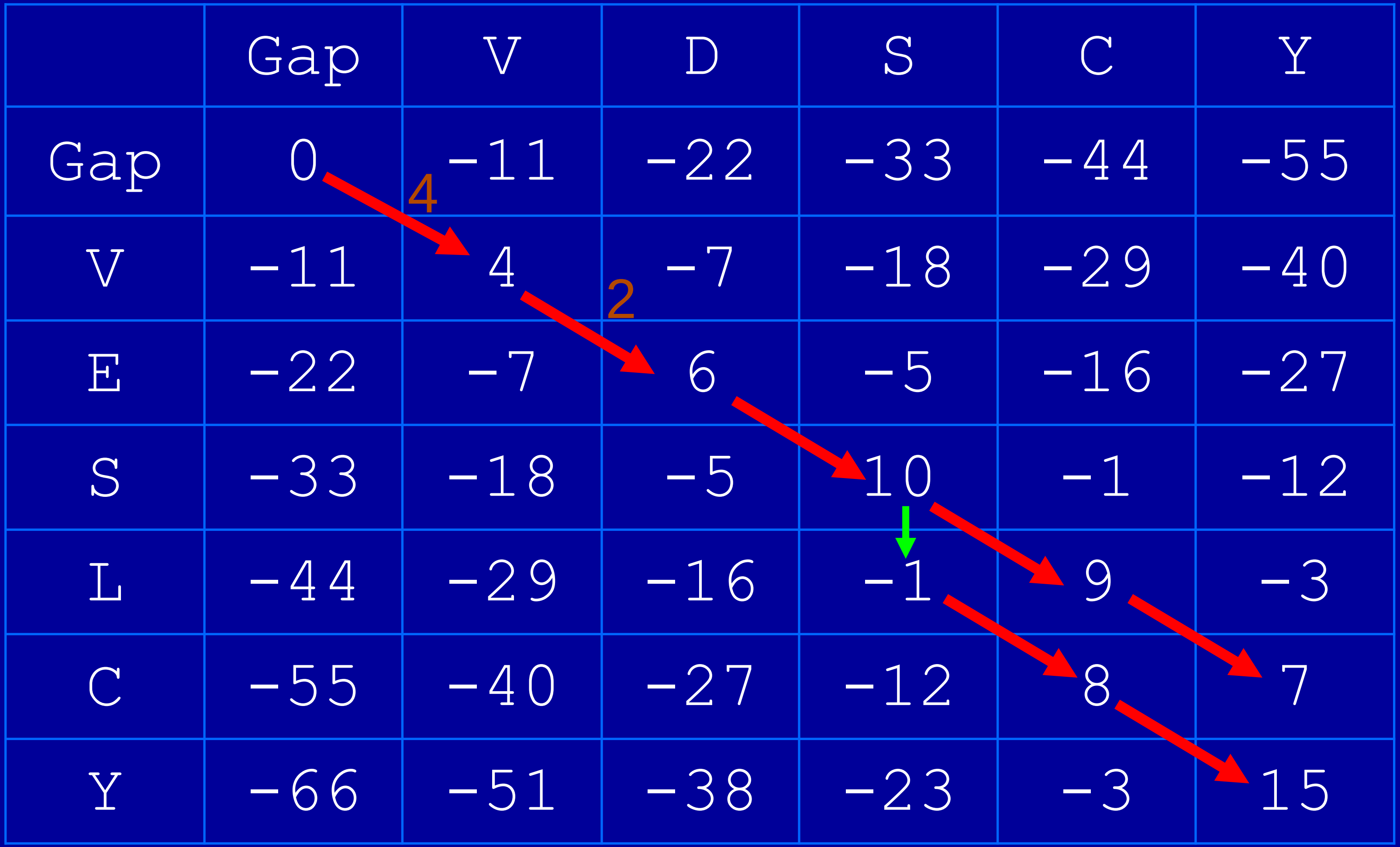
	Gap	V	D	S	C	Y
Gap	0	-11	-22	-33	-44	-55
V	-11	4	-7			
E	-22					
S	-33					
L	-44					
C	-55					
Y	-66					

-3

-11

-11

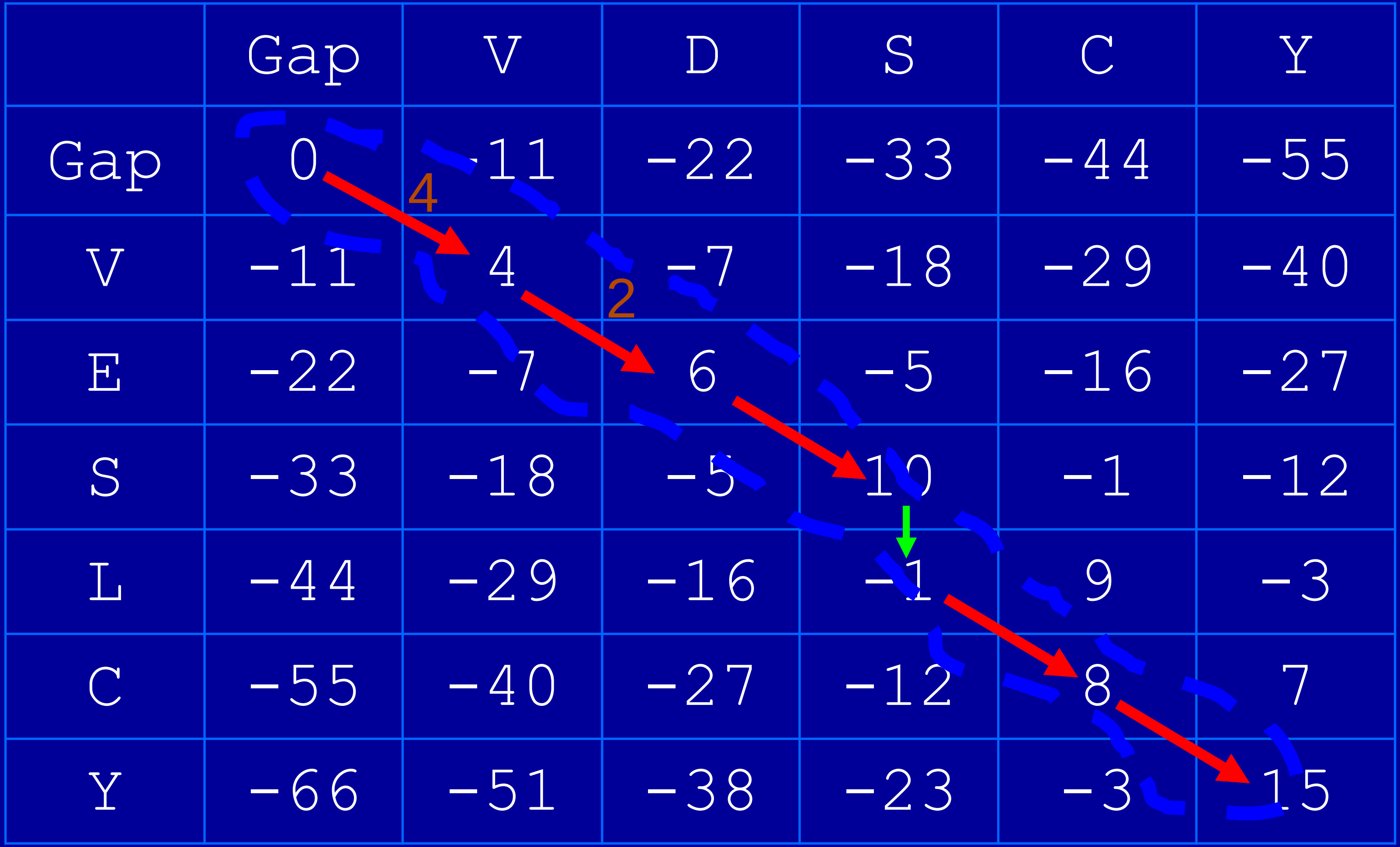
全局比对 (7)



	Gap	V	D	S	C	Y
Gap	0	-11	-22	-33	-44	-55
V	-11	4	-7	-18	-29	-40
E	-22	-7	6	-5	-16	-27
S	-33	-18	-5	10	-1	-12
L	-44	-29	-16	-1	9	-3
C	-55	-40	-27	-12	8	7
Y	-66	-51	-38	-23	-3	15

比对结果

```
V D S - C Y
V E S L C Y
```

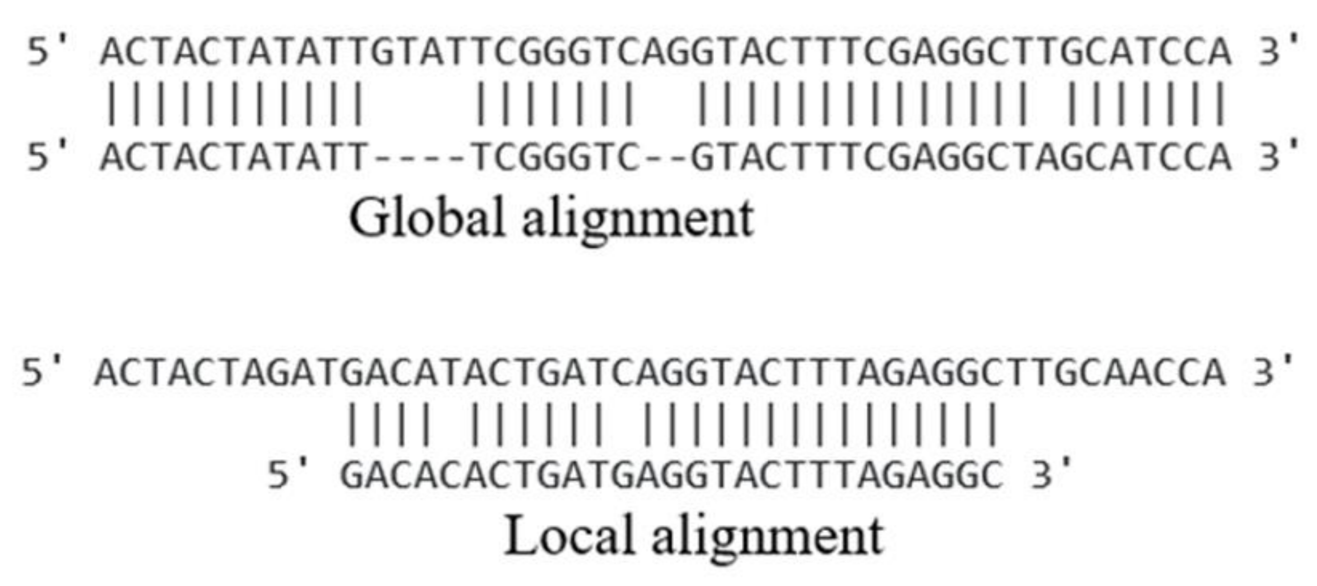


	Gap	V	D	S	C	Y
Gap	0	-11	-22	-33	-44	-55
V	-11	4	-7	-18	-29	-40
E	-22	-7	6	-5	-16	-27
S	-33	-18	-5	10	-1	-12
L	-44	-29	-16	-1	9	-3
C	-55	-40	-27	-12	8	7
Y	-66	-51	-38	-23	-3	15

全局比对与局部比对

- **全局比对(Global alignment)从全长序列出发，整体比较**

- **局部比对(Local alignment)则比较子序列的相似性**



```
5' ACTACTATATTGTATTCGGGTCAGGTACTTTCGAGGCTTGCATCCA 3'
   |||||||||||    |||||||  |||||||||||||| |||||||
5' ACTACTATATT----TCGGGTC--GTACTTTCGAGGCTAGCATCCA 3'
```

Global alignment

```
5' ACTACTAGATGACATACTGATCAGGTACTTTAGAGGCTTGCAACCA 3'
             |||| |||||| |||||||||||||||
         5' GACACACTGATGAGGTACTTTAGAGGC 3'
```

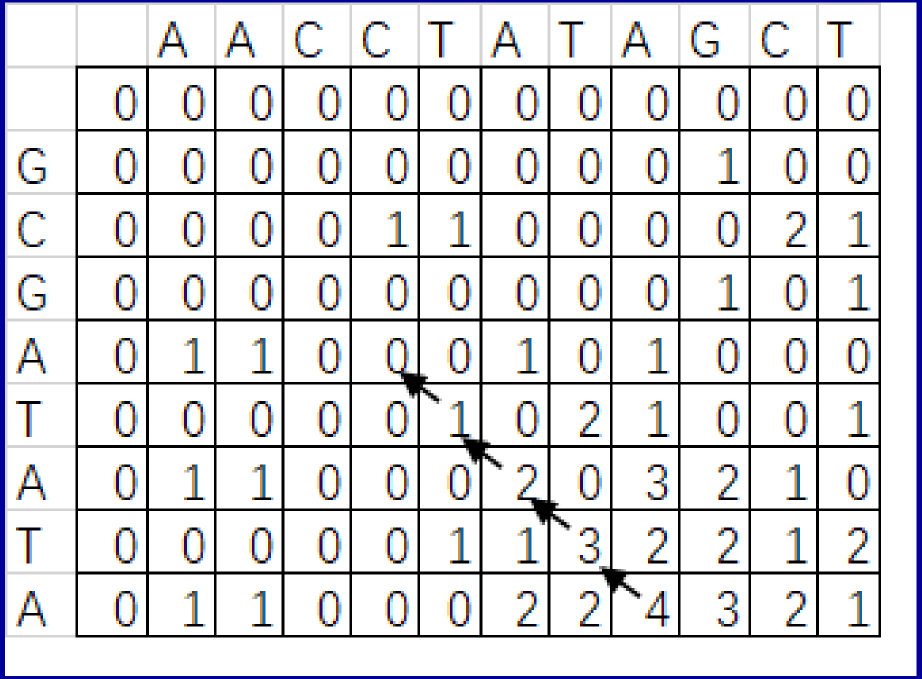
Local alignment

局部比对

Smith-Waterman算法;

时间复杂度O(n^2);

$$F_{ij} = \text{max of} \begin{cases} F_{i-1,\, j-1} + s\ (x_i,\, y_j) \\ F_{i-1,\, j} + d \\ F_{i,\, j-1} + d \\ 0 \end{cases}$$

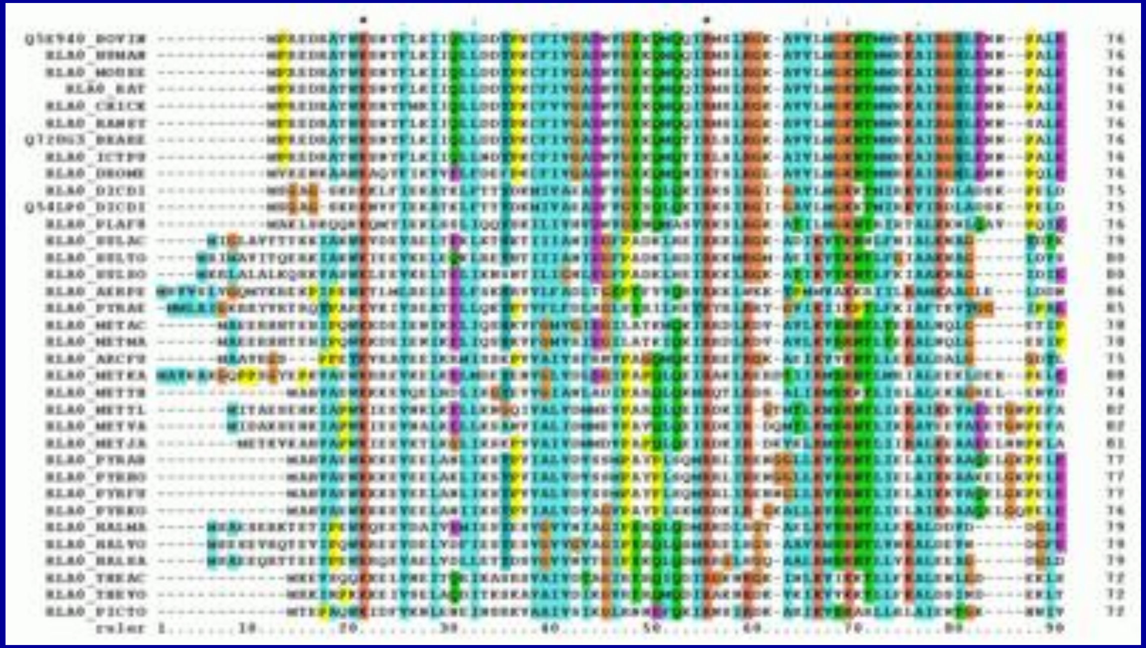


		A	A	C	C	T	A	T	A	G	C	T
	0	0	0	0	0	0	0	0	0	0	0	0
G	0	0	0	0	0	0	0	0	0	1	0	0
C	0	0	0	0	1	1	0	0	0	0	2	1
G	0	0	0	0	0	0	0	0	0	1	0	1
A	0	1	1	0	0	0	1	0	1	0	0	0
T	0	0	0	0	0	1	0	2	1	0	0	1
A	0	1	1	0	0	0	2	0	3	2	1	0
T	0	0	0	0	0	1	1	3	2	2	1	2
A	0	1	1	0	0	0	2	2	4	3	2	1

---TATA---
---TATA---

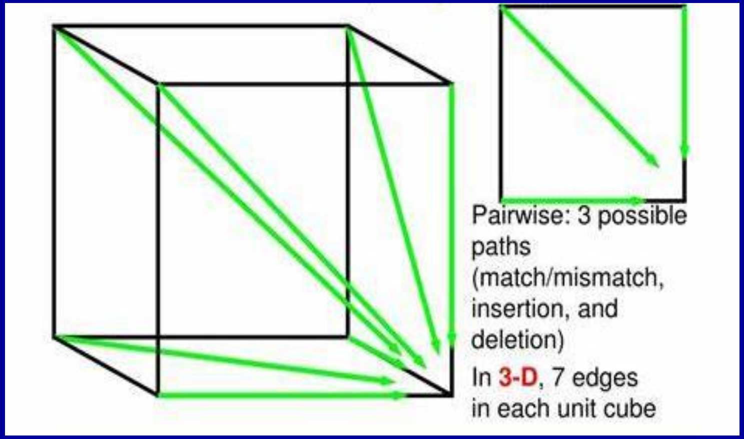
多重序列比对

- 多序列比对是研究基因或蛋白质功能的常用方法，可以发现同源序列中的保守结构域，预测基因结构及构建进化树等。

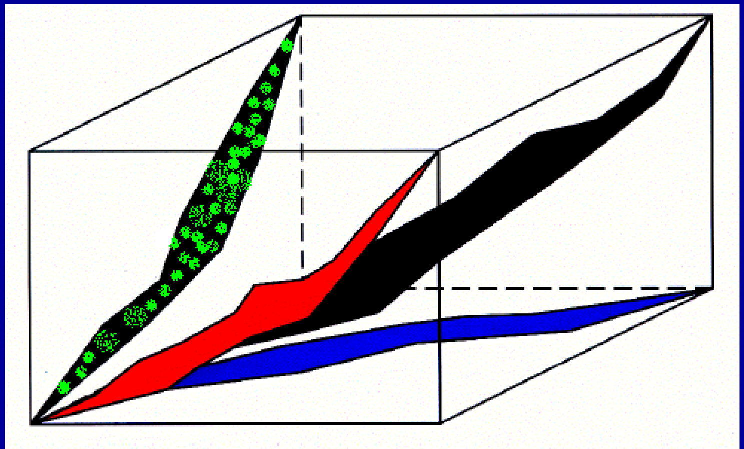


多重序列比对

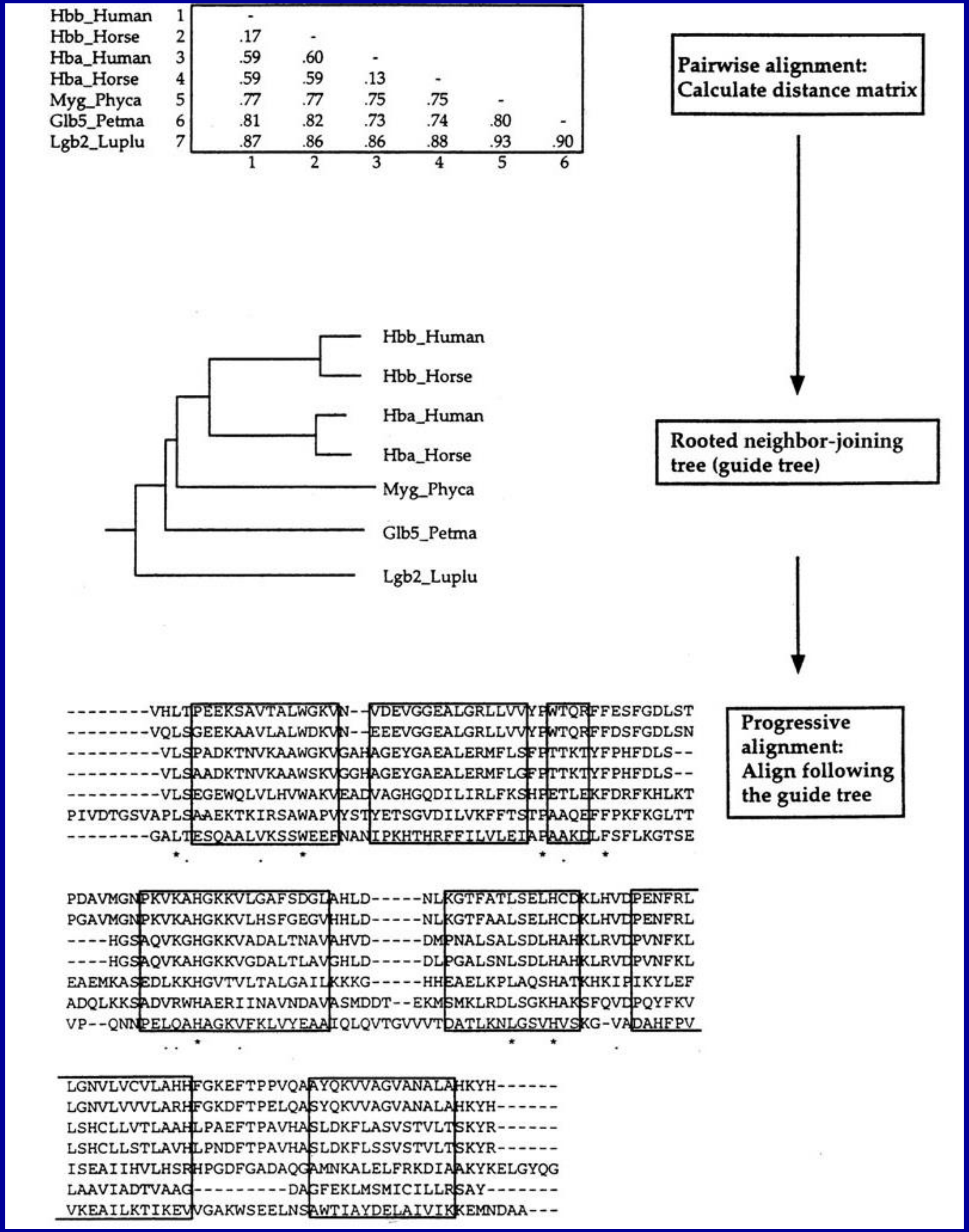
- 多重序列比对可直接应用动态规划算法，双序列比对得分矩阵相当于二维平面；而三条序列比对得分会形成一个三维晶格，每一维对应于一条序列，每一种可能的比对可用三维晶格中的一条路径表示。
- 随着序列数量增多，计算复杂度迅速增大，MSA计算时间复杂度是O(L^N)，L代表序列长度，N代表序列数量
- 目前，多序列比对算法大多是基于渐进比对(progressive alignment)的思想，在两两序列比对的基础上，逐步优化多序列比对的结果。



Pairwise: 3 possible paths (match/mismatch, insertion, and deletion)
In 3-D, 7 edges in each unit cube



CLUSTAL算法



		1	2	3	4	5	6
Hbb_Human	1	-					
Hbb_Horse	2	.17	-				
Hba_Human	3	.59	.60	-			
Hba_Horse	4	.59	.59	.13	-		
Myg_Phyca	5	.77	.77	.75	.75	-	
Glb5_Petma	6	.81	.82	.73	.74	.80	-
Lgb2_Luplu	7	.87	.86	.86	.88	.93	.90

Pairwise alignment:
Calculate distance matrix

Rooted neighbor-joining
tree (guide tree)

```
--------VHLTPEEKSAVTALWGKVN--VDEVGGEALGRLLVVYPWTQRFFESFGDLST
--------VQLSGEEKAAVLALWDKVN--EEEVGGEALGRLLVVYPWTQRFFDSFGDLSN
---------VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLS--
---------VLSAADKTNVKAAWSKVGGHAGEYGAEALERMFLGFPTTKTYFPHFDLS--
---------VLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLEKFDRFKHLKT
PIVDTGSVAPLSAAEKTKIRSAWAPVYSTYETSGVDILVKFFTSTPAAQEFFPKFKGLTT
--------GALTESQAALVKSSWEEFNANIPKHTHRFFILVLEIAPAAKDLFSFLKGTSE
          *.      .   *                      * .   *

PDAVMGNPKVKAHGKKVLGAFSDGLAHLD-----NLKGTFATLSELHCDKLHVDPENFRL
PGAVMGNPKVKAHGKKVLHSFGEGVHHLD-----NLKGTFAALSELHCDKLHVDPENFRL
----HGSAQVKGHGKKVADALTNAVAHVD-----DMPNALSALSDLHAHKLRVDPVNFKL
----HGSAQVKAHGKKVGDALTLAVGHLD-----DLPGALSNLSDLHAHKLRVDPVNFKL
EAEMKASEDLKKHGVTVLTALGAILKKKG-----HHEAELKPLAQSHATKHKIPIKYLEF
ADQLKKSADVRWHAERIINAVNDAVASMDDT--EKMSMKLRDLSGKHAKSFQVDPQYFKV
VP--QNNPELQAHAGKVFKLVYEAAIQLQVTGVVVTDATLKNLGSVHVSKG-VADAHFPV
         .. *   .                         *   *     .

LGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH------
LGNVLVVVLARHFGKDFTPELQASYQKVVAGVANALAHKYH------
LSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR------
LSHCLLSTLAVHLPNDFTPAVHASLDKFLSSVSTVLTSKYR------
ISEAIIHVLHSRHPGDFGADAQGAMNKALELFRKDIAAKYKELGYQG
LAAVIADTVAAG----------DAGFEKLMSMICILLRSAY-------
VKEAILKTIKEVVGAKWSEELNSAWTIAYDELAIVIKKEMNDAA---
```

Progressive
alignment:
Align following
the guide tree

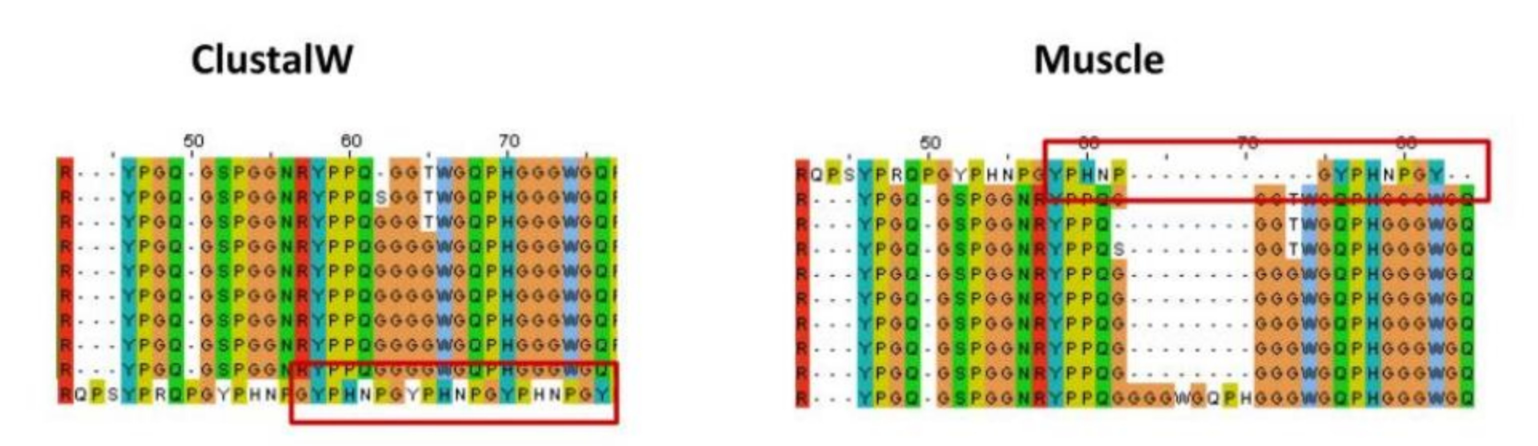
两两比对，构建距离矩阵

指导树的构建

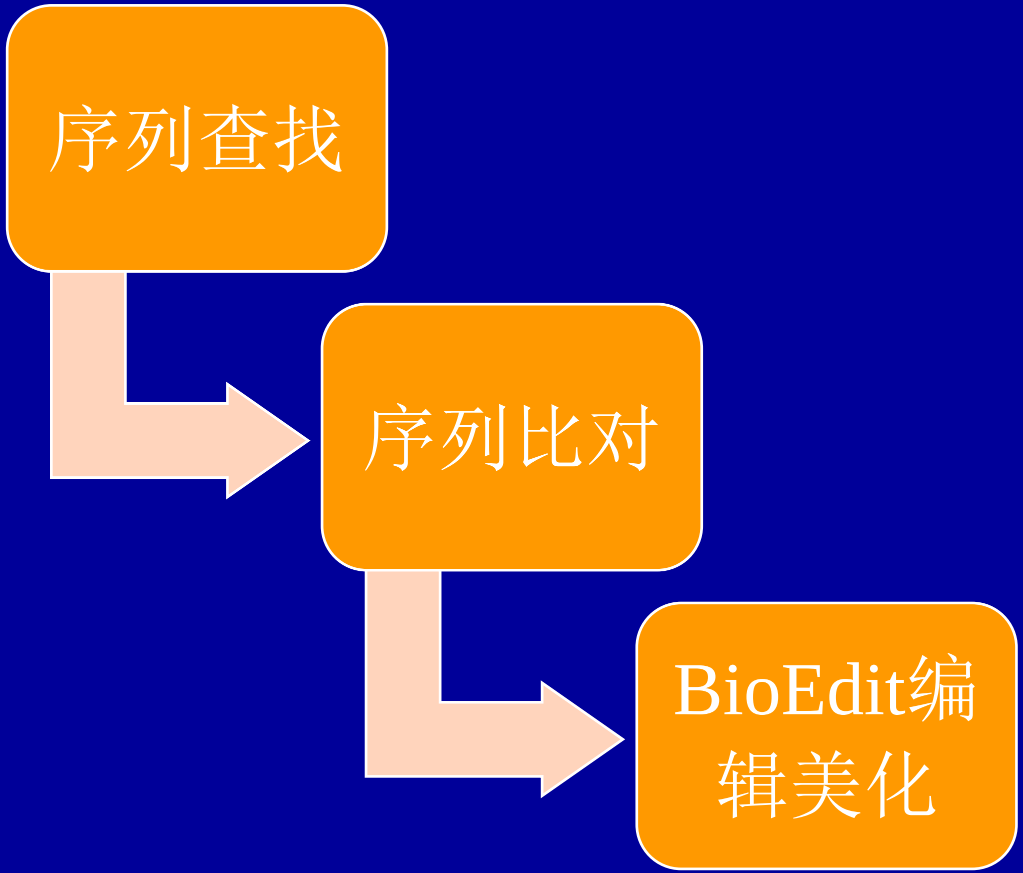
渐进比对

多重序列比对软件

- 常用的多序列比对软件有Clustal、Muscle等。
- Clustal系列有Clustal Omega、ClustalW和ClustalX三个版本



双序列比对实践



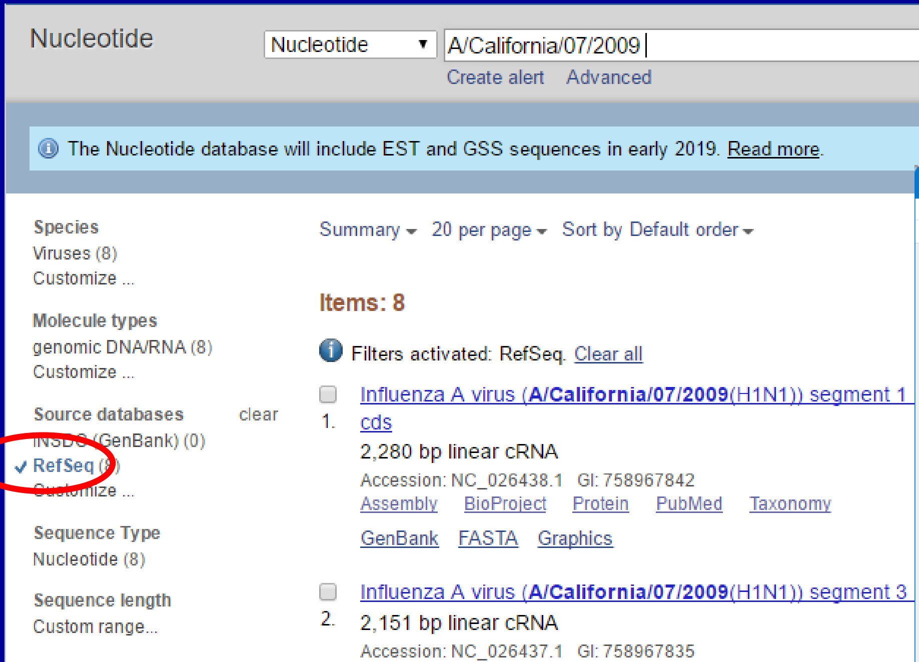
序列查找

序列比对

BioEdit编辑美化

基因序列查找

NCBI网站：https://www.ncbi.nlm.nih.gov



Nucleotide

Nucleotide ▾ A/California/07/2009

Create alert Advanced

The Nucleotide database will include EST and GSS sequences in early 2019. Read more.

Species
Viruses (8)
Customize ...

Molecule types
genomic DNA/RNA (8)
Customize ...

Source databases clear
INSDC (GenBank) (0)
✓ RefSeq (8)
Customize ...

Sequence Type
Nucleotide (8)

Sequence length
Custom range...

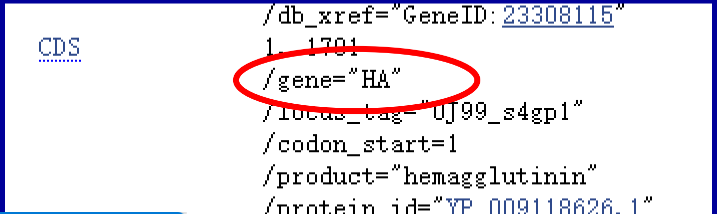
Summary ▾ 20 per page ▾ Sort by Default order ▾

Items: 8

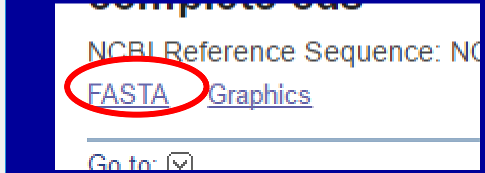
Filters activated: RefSeq. Clear all

□ 1. Influenza A virus (A/California/07/2009(H1N1)) segment 1 cds
2,280 bp linear cRNA
Accession: NC_026438.1 GI: 758967842
Assembly BioProject Protein PubMed Taxonomy
GenBank FASTA Graphics

□ 2. Influenza A virus (A/California/07/2009(H1N1)) segment 3
2,151 bp linear cRNA
Accession: NC_026437.1 GI: 758967835



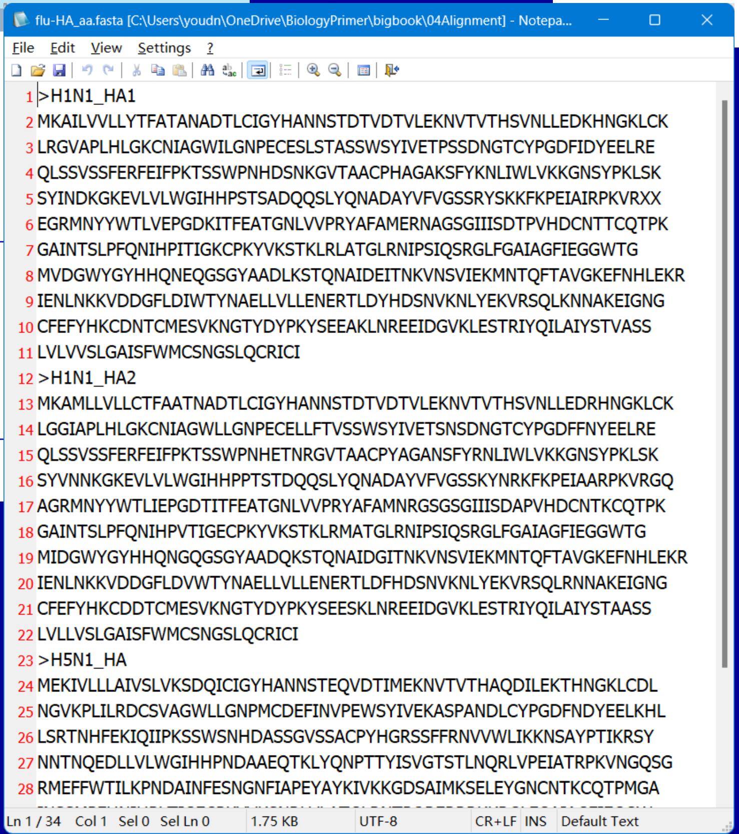
```
                     /db_xref="GeneID:23308115"
     CDS             1..1701
                     /gene="HA"
                     /locus_tag="UJ99_s4gp1"
                     /codon_start=1
                     /product="hemagglutinin"
                     /protein_id="YP_009118626.1"
                     /db_xref="GeneID:23308115"
```



NCBI Reference Sequence: NC

FASTA Graphics

Go to: ☑



flu-HA_aa.fasta [C:\Users\youdn\OneDrive\BiologyPrimer\bigbook\04Alignment] - Notepa...

```
>H1N1_HA1
MKAILVVLLYTFATANADTLCIGYHANNSTDTVDTVLEKNVTVTHSVNLLEDKHNGKLCK
LRGVAPLHLGKCNIAGWILGNPECESLSTASSWSYIVETPSSDNGTCYPGDFIDYEELRE
QLSSVSSFERFEIFPKTSSWPNHDSNKGVTAACPHAGAKSFYKNLIWLVKKGNSYPKLSK
SYINDKGKEVLVLWGIHHPSTSADQQSLYQNADAYVFVGSSRYSKKFKPEIAIRPKVRXX
EGRMNYYWTLVEPGDKITFEATGNLVVPRYAFAMERNAGSGIIISDTPVHDCNTTCQTPK
GAINTSLPFQNIHPITIGKCPKYVKSTKLRLATGLRNIPSIQSRGLFGAIAGFIEGGWTG
MVDGWYGYHHQNEQGSGYAADLKSTQNAIDEITNKVNSVIEKMNTQFTAVGKEFNHLEKR
IENLNKKVDDGFLDIWTYNAELLVLLENERTLDYHDSNVKNLYEKVRSQLKNNAKEIGNG
CFEFYHKCDNTCMESVKNGTYDYPKYSEEAKLNREEIDGVKLESTRIYQILAIYSTVASS
LVLVVSLGAISFWMCSNGSLQCRICI
>H1N1_HA2
MKAMLLVLLCTFAATNADTLCIGYHANNSTDTVDTVLEKNVTVTHSVNLLEDRHNGKLCK
LGGIAPLHLGKCNIAGWLLGNPECELLFTVSSWSYIVETSNSDNGTCYPGDFFNYEELRE
QLSSVSSFERFEIFPKTSSWPNHETNRGVTAACPYAGANSFYRNLIWLVKKGNSYPKLSK
SYVNNKGKEVLVLWGIHHPPTSTDQQSLYQNADAYVFVGSSKYNRKFKPEIAARPKVRGQ
AGRMNYYWTLIEPGDTITFEATGNLVVPRYAFAMNRGSGSGIIISDAPVHDCNTKCQTPK
GAINTSLPFQNIHPVTIGECPKYVKSTKLRMATGLRNIPSIQSRGLFGAIAGFIEGGWTG
MIDGWYGYHHQNGQGSGYAADQKSTQNAIDGITNKVNSVIEKMNTQFTAVGKEFNHLEKR
IENLNKKVDDGFLDVWTYNAELLVLLENERTLDFHDSNVKNLYEKVRSQLRNNAKEIGNG
CFEFYHKCDDTCMESVKNGTYDYPKYSEESKLNREEIDGVKLESTRIYQILAIYSTAASS
LVLLVSLGAISFWMCSNGSLQCRICI
>H5N1_HA
MEKIVLLLAIVSLVKSDQICIGYHANNSTEQVDTIMEKNVTVTHAQDILEKTHNGKLCDL
NGVKPLILRDCSVAGWLLGNPMCDEFINVPEWSYIVEKASPANDLCYPGDFNDYEELKHL
LSRTNHFEKIQIIPKSSWSNHDASSGVSSACPYHGRSSFFRNVVWLIKKNSAYPTIKRSY
NNTNQEDLLVLWGIHHPNDAAEQTKLYQNPTTYISVGTSTLNQRLVPEIATRPKVNGQSG
RMEFFWTILKPNDAINFESNGNFIAPEYAYKIVKKGDSAIMKSELEYGNCNTKCQTPMGA
```

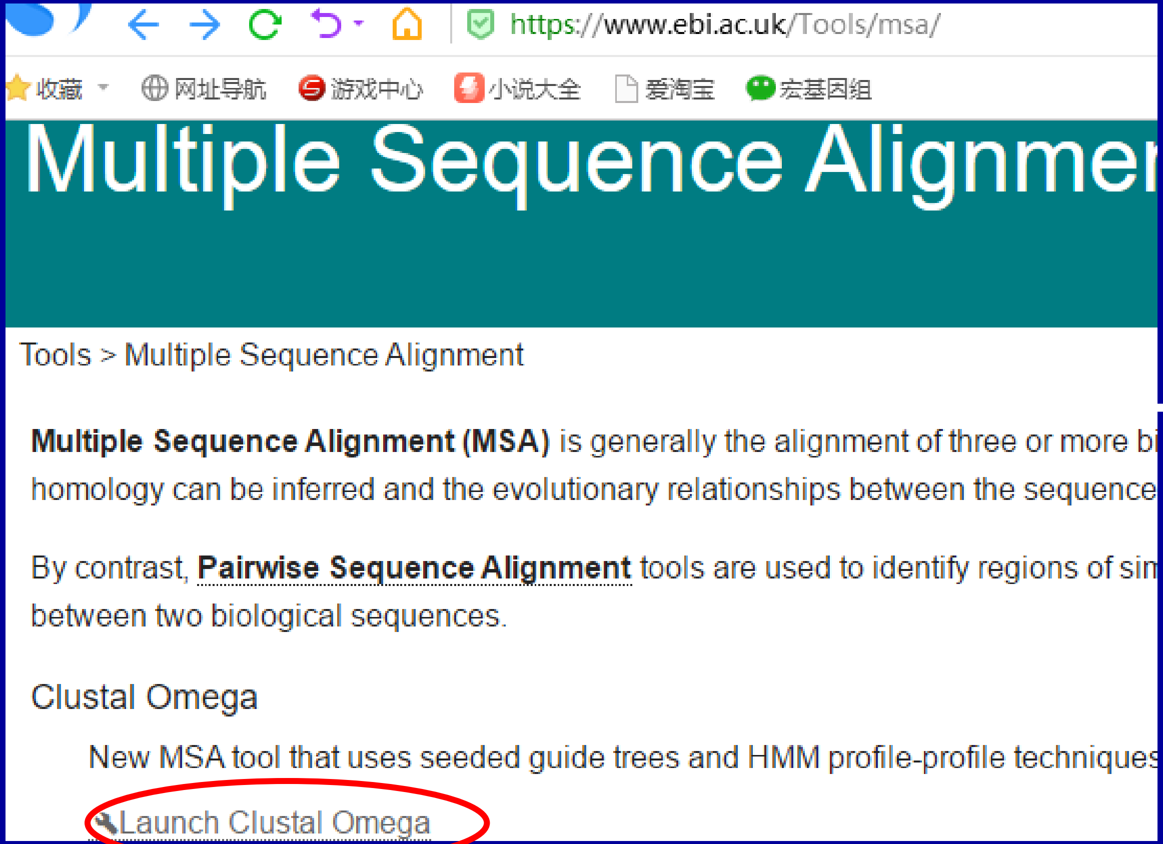
Ln 1 / 34 Col 1 Sel 0 Sel Ln 0 1.75 KB UTF-8 CR+LF INS Default Text

HA: 流感病毒凝集素基因
(segnment 4 hemagglutinin gene)

序列比对 - Clustal

CLUSTAL是欧洲生物信息研究所（European Bioinformatics Institute）开发的一套比对序列工具。

网址：https://www.ebi.ac.uk/Tools/msa



Multiple Sequence Alignmen

Tools > Multiple Sequence Alignment

Multiple Sequence Alignment (MSA) is generally the alignment of three or more bi
homology can be inferred and the evolutionary relationships between the sequence

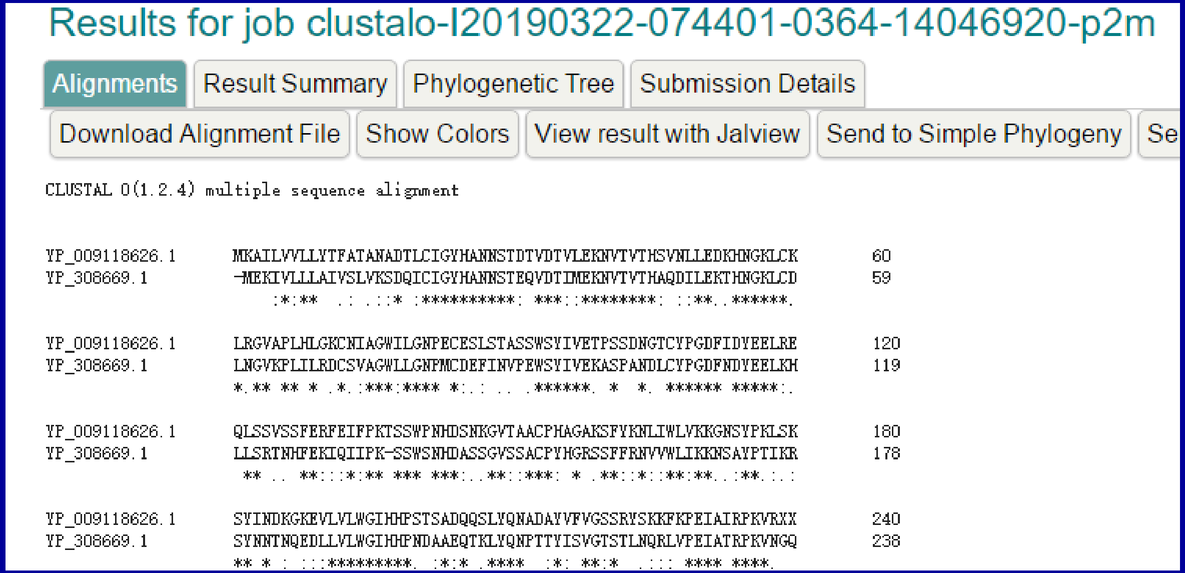
By contrast, **Pairwise Sequence Alignment** tools are used to identify regions of sim
between two biological sequences.

Clustal Omega

New MSA tool that uses seeded guide trees and HMM profile-profile techniques

Launch Clustal Omega

→

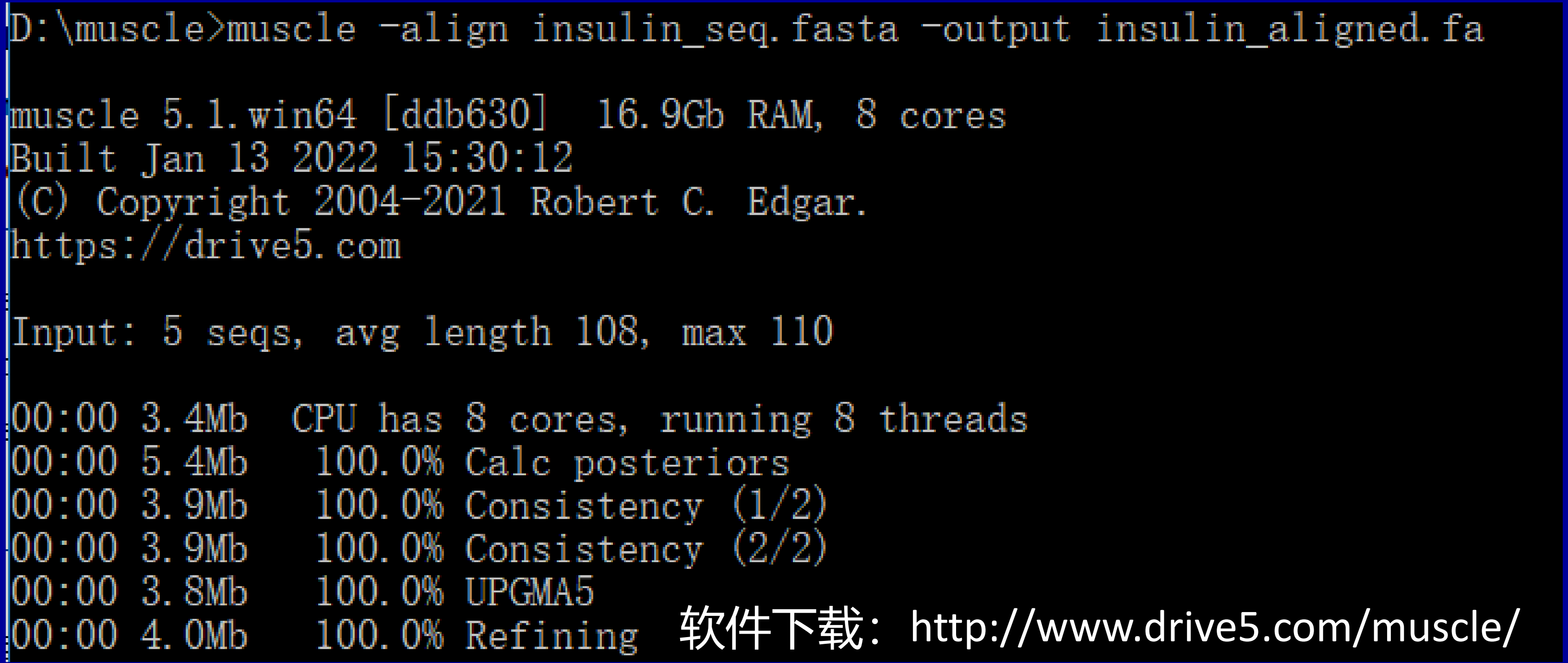


Results for job clustalo-I20190322-074401-0364-14046920-p2m

```
CLUSTAL O(1.2.4) multiple sequence alignment


YP_009118626.1      MKAILVVLLYTFATANADTLCIGYHANNSTDTVDTVLEKNVTVTHSVNLLEDKHNGKLCK      60
YP_308669.1         -MEKIVLLLAIVSLVKSDQICIGYHANNSTEQVDTIMEKNVTVTHAQDILEKTHNGKLCD      59
                        :*:**  .: .::* :**********: ***::********: ::**..******.

YP_009118626.1      LRGVAPLHLGKCNIAGWILGNPECESLSTASSWSYIVETPSSDNGTCYPGDFIDYEELRE      120
YP_308669.1         LNGVKPLILRDCSVAGWLLGNPMCDEFINVPEWSYIVEKASPANDLCYPGDFNDYEELKH      119
                    *.** ** * .*.:***:**** *:.: .. .******. *  *. ****** *****:.

YP_009118626.1      QLSSVSSFERFEIFPKTSSWPNHDSNKGVTAACPHAGAKSFYKNLIWLVKKGNSYPKLSK      180
YP_308669.1         LLSRTNHFEKIQIIPK-SSWSNHDASSGVSSACPYHGRSSFFRNVVWLIKKNSAYPTIKR      178
                     ** .. **:::*:** *** ***:..**::***: * .**::*::**:**..:**.:.:

YP_009118626.1      SYINDKGKEVLVLWGIHHPSTSADQQSLYQNADAYVFVGSSRYSKKFKPEIAIRPKVRXX      240
YP_308669.1         SYNNTNQEDLLVLWGIHHPNDAAEQTKLYQNPTTYISVGTSTLNQRLVPEIATRPKVNGQ      238
                    ** * : :::*********. :*:* .****  :*: **:*  .::: **** ****.
```

序列比对 - MUSCLE

- MUSCLE (Multiple Protein Sequence Alignment)是一款简单好用的多序列比对软件，相比ClustalW，在不损失精度的情况下速度提升了数倍。
- 它使用十分方便，大多数情况下用户只需要指定输入/输出文件即可，输入/输出文件默认为fasta格式。

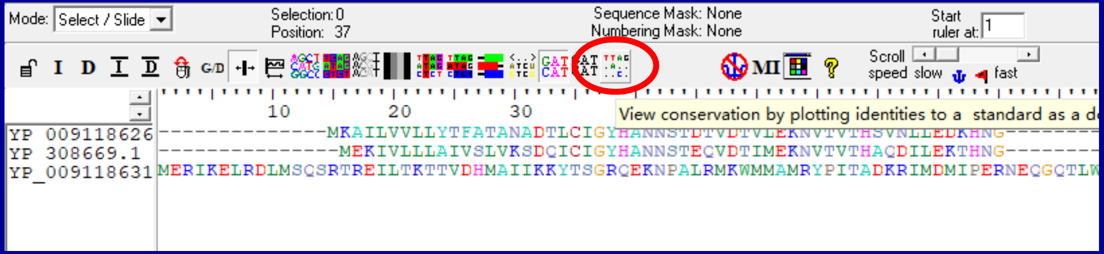


```
D:\muscle>muscle -align insulin_seq.fasta -output insulin_aligned.fa

muscle 5.1.win64 [ddb630]  16.9Gb RAM, 8 cores
Built Jan 13 2022 15:30:12
(C) Copyright 2004-2021 Robert C. Edgar.
https://drive5.com

Input: 5 seqs, avg length 108, max 110

00:00 3.4Mb  CPU has 8 cores, running 8 threads
00:00 5.4Mb   100.0% Calc posteriors
00:00 3.9Mb   100.0% Consistency (1/2)
00:00 3.9Mb   100.0% Consistency (2/2)
00:00 3.8Mb   100.0% UPGMA5
00:00 4.0Mb   100.0% Refining
```

软件下载：http://www.drive5.com/muscle/

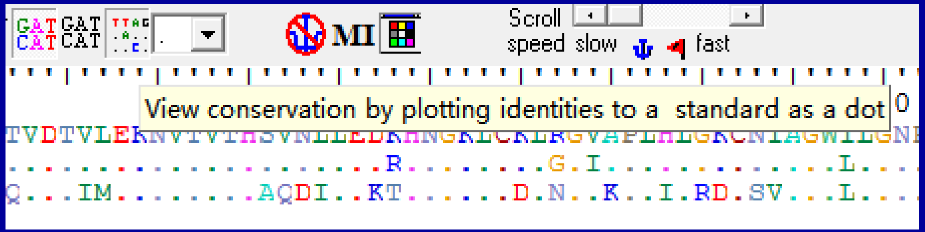
BioEdit

1. 查看比对结果，可选中工具栏中的“View conservation by plotting identities to a standard as a dot”，以点显示与第一行相同的字母。
2. 如果观察到有些位置的序列比对不合理，可进行序列编辑，或调整空位(gap)位置等。



View conservation by plotting identities to a standard as a d

```
YP 009118626 -------------MKAILVVLLYTFATANADTLCIGYHANNSTDTVDTVLEKNVTVTHSVNLLEDKHNG--------
YP 308669.1  ---------------MEKIVLLLAIVSLVKSDQICIGYHANNSTEQVDTIMEKNVTVTHAQDILEKTHNG--------
YP_009118631 MERIKELRDLMSQSRTREILTKTTVDHMAIIKKYTSGRQEKNPALRMKWMMAMRYPITADKRIMDMIPERNEQGQTLW
```



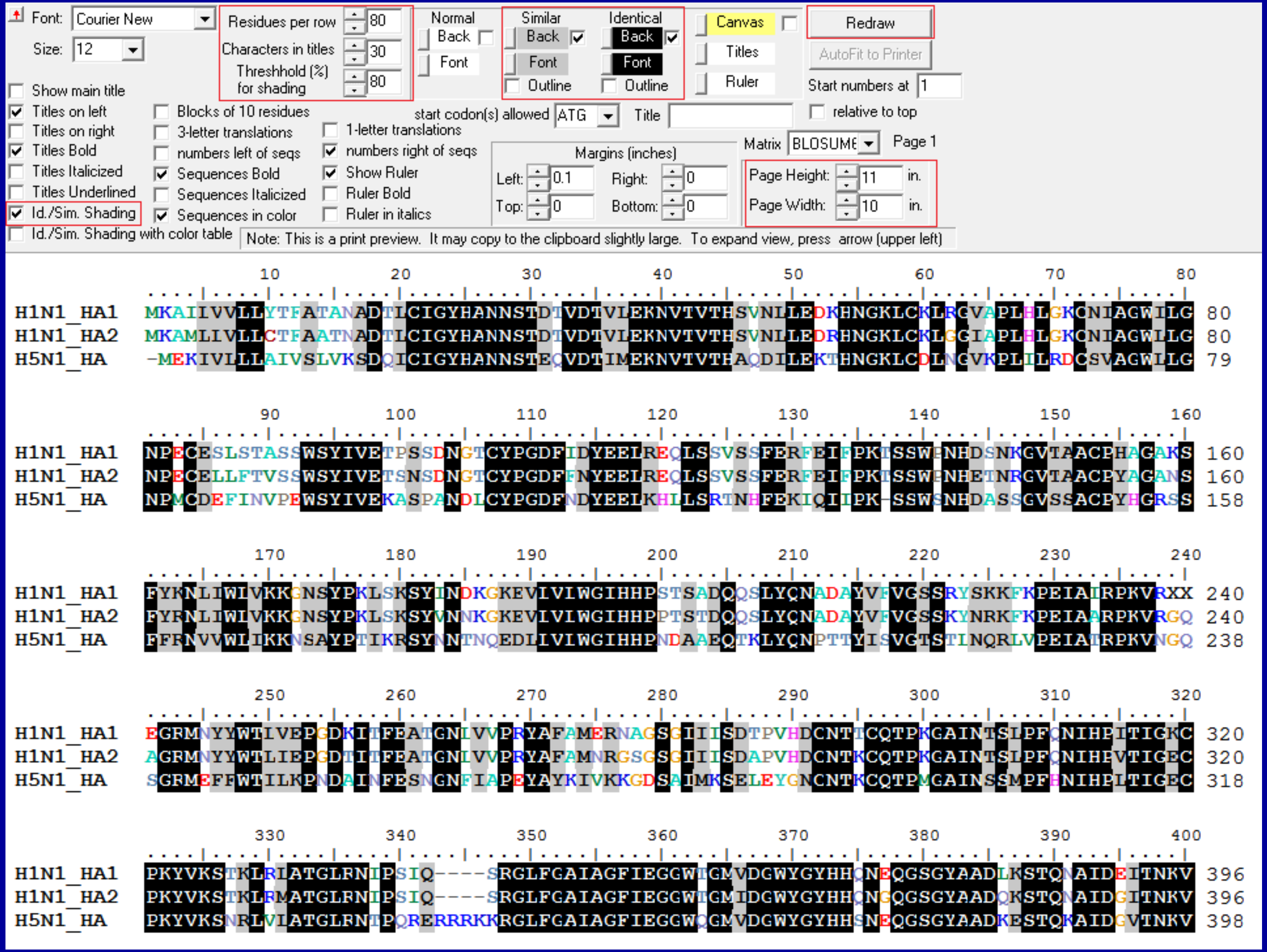
View conservation by plotting identities to a standard as a dot

```
TVDTVLEKNVTVTHSVNLLEDKHNGKICKIRGVAPLHLGKCNIAGWILGN
.....................R........G.I.............L...
Q...IM........AQDI..KT......D.N..K..I.RD.SV...L...
```

注意：修改前需要把工具栏中Mode的状态改成“Edit/Insert”，才能进行删除或修改操作。

序列美化

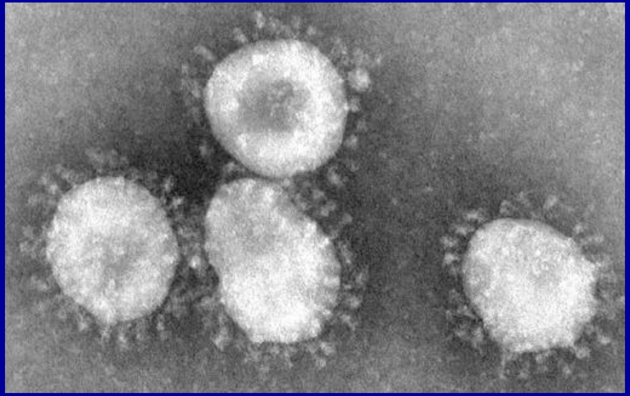
- 显示序列比对：File菜单→Graphic view,可以对比对显示进行美化。
- 有许多显示参数可修改，但有些参数修改后需要按右上角的Redraw按钮察看结果。



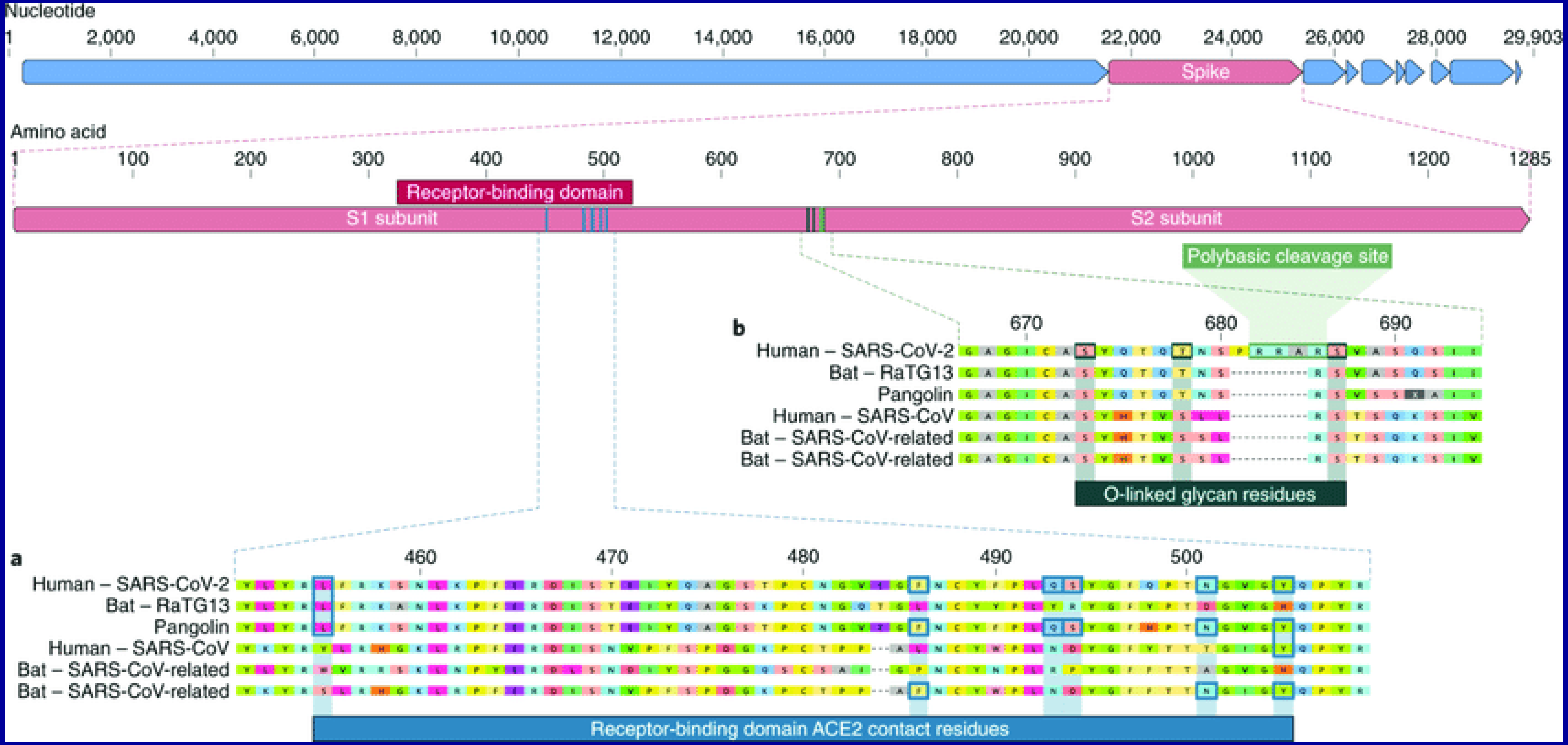
Font: Courier New
Size: 12
Residues per row 80
Characters in titles 30
Threshhold (%) for shading 80
Normal: Back □ Font
Similar: Back ☑ Font □ Outline
Identical: Back ☑ Font □ Outline
Canvas □
Titles
Ruler
Redraw
AutoFit to Printer
Start numbers at 1
□ relative to top
□ Show main title
☑ Titles on left
□ Titles on right
☑ Titles Bold
□ Titles Italicized
□ Titles Underlined
☑ Id./Sim. Shading
□ Id./Sim. Shading with color table
□ Blocks of 10 residues
□ 3-letter translations
□ numbers left of seqs
☑ Sequences Bold
□ Sequences Italicized
☑ Sequences in color
start codon(s) allowed ATG
Title
□ 1-letter translations
☑ numbers right of seqs
☑ Show Ruler
□ Ruler Bold
□ Ruler in italics
Margins (inches)
Left: 0.1 Right: 0
Top: 0 Bottom: 0
Matrix BLOSUM6
Page 1
Page Height: 11 in.
Page Width: 10 in.
Note: This is a print preview. It may copy to the clipboard slightly large. To expand view, press arrow (upper left)

```
                      10        20        30        40        50        60        70        80
              ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
H1N1_HA1      MKAIIVVLLYTFATANADTLCIGYHANNSTDTVDTVLEKNVTVTHSVNLLEDKHNGKLCKLRGVAPLHLGKCNIAGWILG  80
H1N1_HA2      MKAMLIVLLCTFAATNADTLCIGYHANNSTDTVDTVLEKNVTVTHSVNLLEDRHNGKLCKLGGIAPLHLGKCNIAGWLLG  80
H5N1_HA       -MEKIVLLLAIVSLVKSDQICIGYHANNSTEQVDTIMEKNVTVTHAQDILEKTHNGKLCDLNGVKPLILRDCSVAGWLLG  79

                      90       100       110       120       130       140       150       160
              ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
H1N1_HA1      NPECESLSTASSWSYIVETPSSDNGTCYPGDFIDYEELREQLSSVSSFERFEIFPKTSSWPNHDSNKGVTAACPHAGAKS  160
H1N1_HA2      NPECELLFTVSSWSYIVETSNSDNGTCYPGDFFNYEELREQLSSVSSFERFEIFPKTSSWPNHETNRGVTAACPYAGANS  160
H5N1_HA       NPMCDEFINVPEWSYIVEKASPANDLCYPGDFNDYEELKHLLSRTNHFEKIQIIPK-SSWSNHDASSGVSSACPYHGRSS  158

                     170       180       190       200       210       220       230       240
              ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
H1N1_HA1      FYKNLIWLVKKGNSYPKLSKSYINDKGKEVLVLWGIHHPSTSADQQSLYQNADAYVFVGSSRYSKKFKPEIAIRPKVRXX  240
H1N1_HA2      FYRNLIWLVKKGNSYPKLSKSYVNNKGKEVLVLWGIHHPPTSTDQQSLYQNADAYVFVGSSKYNRKFKPEIAARPKVRGQ  240
H5N1_HA       FFRNVVWLIKKNSAYPTIKRSYNNTNQEDLLVLWGIHHPNDAAEQTKLYQNPTTYISVGTSTLNQRLVPEIATRPKVNGQ  238

                     250       260       270       280       290       300       310       320
              ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
H1N1_HA1      EGRMNYYWTLVEPGDKITFEATGNLVVPRYAFAMERNAGSGIIISDTPVHDCNTTCQTPKGAINTSLPFQNIHPITIGKC  320
H1N1_HA2      AGRMNYYWTLLEPGDTITFEATGNLVVPRYAFAMNRGSGSGIIISDAPVHDCNTKCQTPKGAINTSLPFQNIHPVTIGEC  320
H5N1_HA       SGRMEFFWTILKPNDAINFESNGNFIAPEYAYKIVKKGDSAIMKSELEYGNCNTKCQTPMGAINSSMPFHNIHPLTIGEC  318

                     330       340       350       360       370       380       390       400
              ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
H1N1_HA1      PKYVKSTKLRLATGLRNIPSIQ----SRGLFGAIAGFIEGGWTGMVDGWYGYHHQNEQGSGYAADLKSTQNAIDEITNKV  396
H1N1_HA2      PKYVKSTKLRMATGLRNIPSIQ----SRGLFGAIAGFIEGGWTGMIDGWYGYHHQNGQGSGYAADQKSTQNAIDGITNKV  396
H5N1_HA       PKYVKSNRLVLATGLRNTPQRERRRKKRGLFGAIAGFIEGGWQGMVDGWYGYHHSNEQGSGYAADKESTQKAIDGVTNKV  398
```

新冠病毒(SARS-CoV-2)



- 外壳上是刺突状的Spike蛋白(S)。



Andersen, K.G., Rambaut, A., Lipkin, W.I. et al. The proximal origin of SARS-CoV-2. *Nat Med* 26, 450–452 (2020).

作业

- 从NCBI核酸数据库下载SARS-CoV-2、SARS-CoV (2003)及RaTG13病毒株的S(Spike)蛋白质序列，进行多序列比对后观察S蛋白的氨基酸序列差异？
- 可使用下面任一种方法：
  - 使用EBI在线工具Clustal Omega
  - 使用BioEdit软件的ClustalW程序
  - 在Linux命令行中的MUSCLE程序 (`$sudo apt install muscle`)