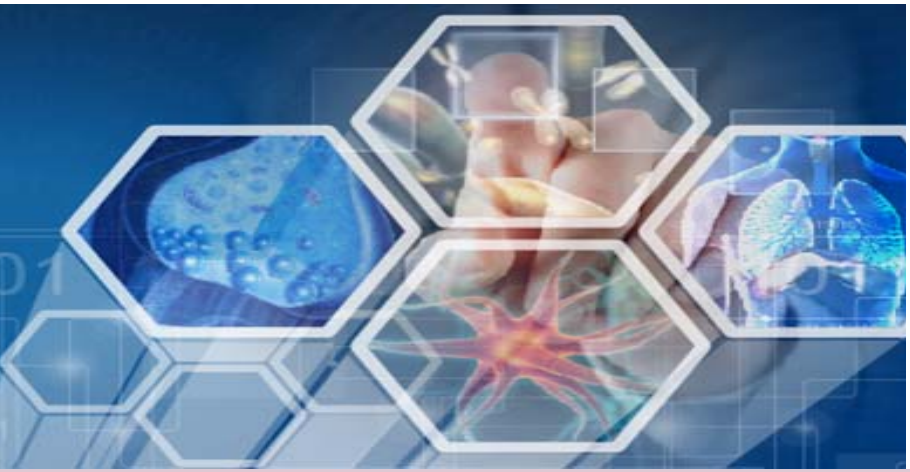
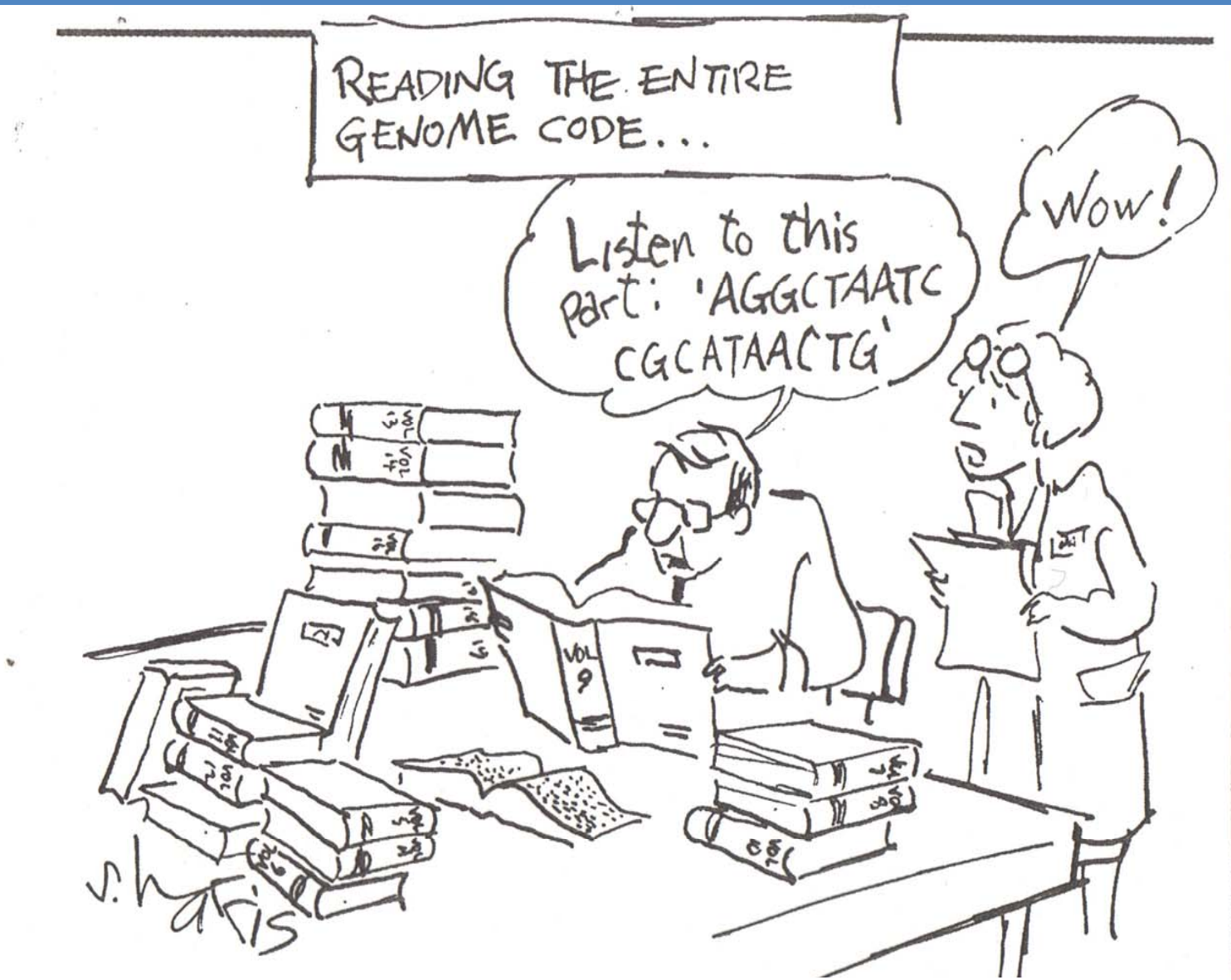


序列比对

第一节 引 言

Section 1 Introduction

人民卫生出版社
PEOPLE'S MEDICAL PUBLISHING HOUSE


READING THE ENTIRE GENOME CODE...
Listen to this part: 'AGGCTAATC CGCATAACTG'
Wow!
VOL 9
J. Harris

Best Sequence Matches

- **Depends on how you define "Best"**
- **Consider the two DNA sequences *v* and *w* :**

v : **A T A T A T A T**
w : **T A T A T A T A**

- **The Hamming distance: $d_H(v, w) = 8$ is large but the sequences are very similar**
- **What if we allowed insertions and deletions?**

Allowing Insertions and Deletions

By shifting one sequence over one position:

v : ATATATAT--
w : --TATATATA

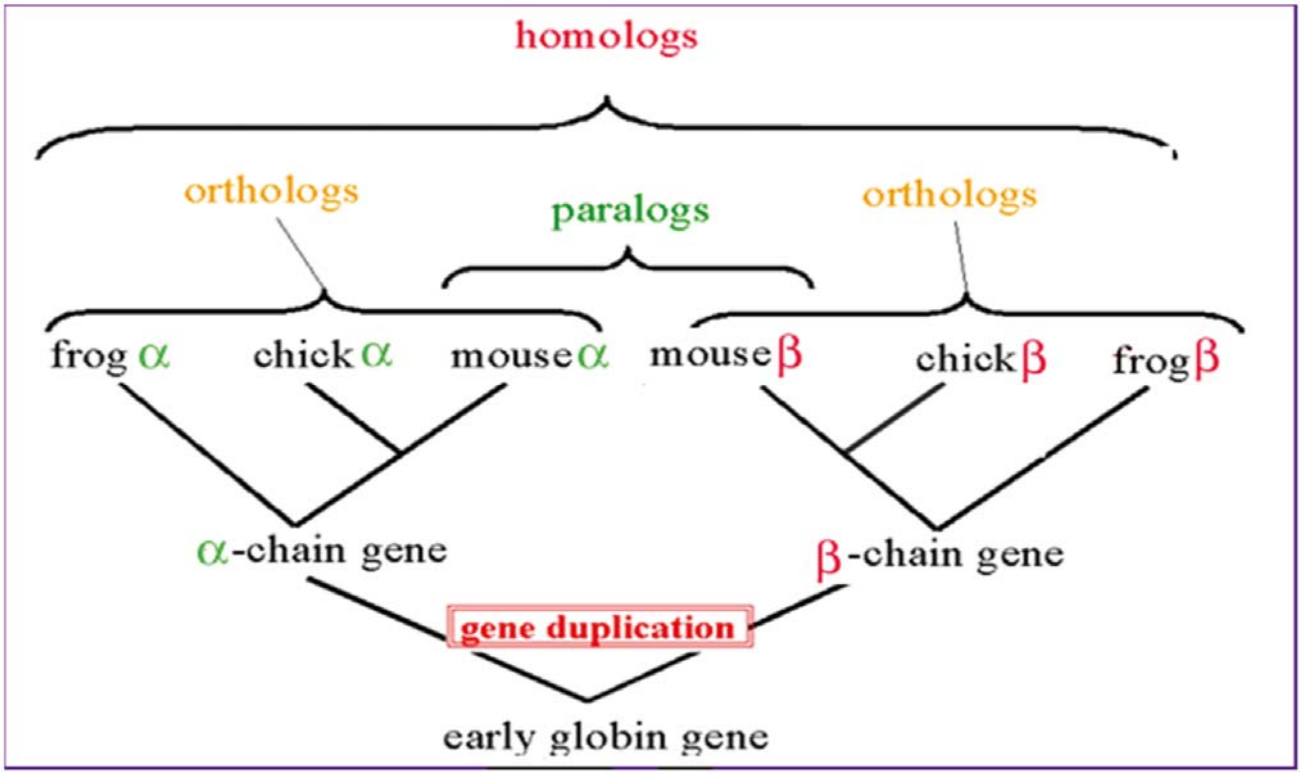
- The edit distance: $d_H(v, w) = 2$.
- Hamming distance neglects insertions and deletions in DNA

一 、同源、相似与距离

（一） 同源

- 两个序列享有一个共同的进化上的祖先，则这两个序列是同源的。
- 对于两个序列，他们或者同源或者不同源，不能说他们70%或80%同源。

➢ 同源可分为垂直同源（**ortholog**）和水平同源（**paralog**）



垂直同源与水平同源

（二）相似性与距离

- 相似性、距离：是两个定量描述多个序列相似度的度量。
- 相似性：被比对序列之间的相似程度。
- 距离：被比对序列间的差异程度。
- 相似性既可用于全局比对也可用于局部比对，而距离一般仅用于全局比对，因为它反映了把一个序列转换成另一个序列所需字符替换的耗费。

二、相似与距离的定量描述

- 相似性可定量地定义为两个序列的函数，即它可有多个值，值的大小取决于两个序列对应位置上相同字符的个数，值越大则表示两个序列越相似。
- 编辑距离（**edit distance**）也可定量地定义为两个序列的函数，其值取决于两个序列对应位置上差异字符的个数，值越小则表示两个序列越相似。

Levenshtein (1966) introduced the notion of an "edit distance" between two strings as the minimum number of elementary operations (insertions, deletions, and substitutions) to transform one string into the other.

(But, he gave no solution)

- 对于一个比对，不论使用什么计分函数进行计分，相似性被定义为总等值于最大的计分：

$$similarity(s_1, s_2, ...s_k) = \max \sum_{i=1}^{|s_1'|} score(s_1'(i), s_2'(i), ...s_k'(i))$$

- 对于**k**个序列，如果用一个函数***cost***（）对每一列的所有替换操作进行计分，则多个序列之间的距离等值于最小的计分：

$$distance(s_1, s_2, ...s_k) = \min \sum_{i=1}^{|s_1'|} cost(s_1'(i), s_2'(i), ...s_k'(i))$$

		seq1 =	ATC	AGGCT	GCTAGCTA
		seq2 =	TAC	ACCTT	CGTGAGCA
打分规则1	p(a, a)=1				
	p(a, b)=0 (a≠b)	相似性得分=	1	2	2
打分规则2	p(a, a)=0.8				
	p(a, b)=0.2 (a≠b)	相似性得分=	1.2	2.2	2.8
打分规则BLAST		相似性得分=	-3	-2	-6

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

对相似性的计分

seq1 =	ATC	AGGCT	GCTAGCTA
seq2 =	TAC	ACCTT	CGTGAGCA
Hamming Distance(seq1,seq2)=	2	3	6

Edit Distance vs Hamming Distance

Hamming distance always compares ***i***-th letter of **v** with ***i***-th letter of **w**

v = ATATATAT

w = TATATATA

Just one shift

Lines them up

Edit distance may compare ***i***-th letter of **v** with ***j***-th letter of **w**

v = - ATATATAT

w = TATATATA

Hamming distance:

$d(\mathbf{v}, \mathbf{w})=8$

Computing Hamming distance is a trivial task

Edit distance:

$d(\mathbf{v}, \mathbf{w})=2$

Computing edit distance is a non-trivial task

Edit Distance: Example

TGCATAT → ATCCGAT in 5 steps

TGCATAT → (DELETE last T)
TGCATA → (DELETE last A)
TGCAT → (INSERT A at front)
ATGCAT → (SUBSTITUTE C for 3rd G)
ATCCAT → (INSERT G before last A)
ATCCGAT (Done)

What is the edit distance? 5?

Edit Distance: Example (cont'd)

TGCATAT → ATCCGAT in 4 steps

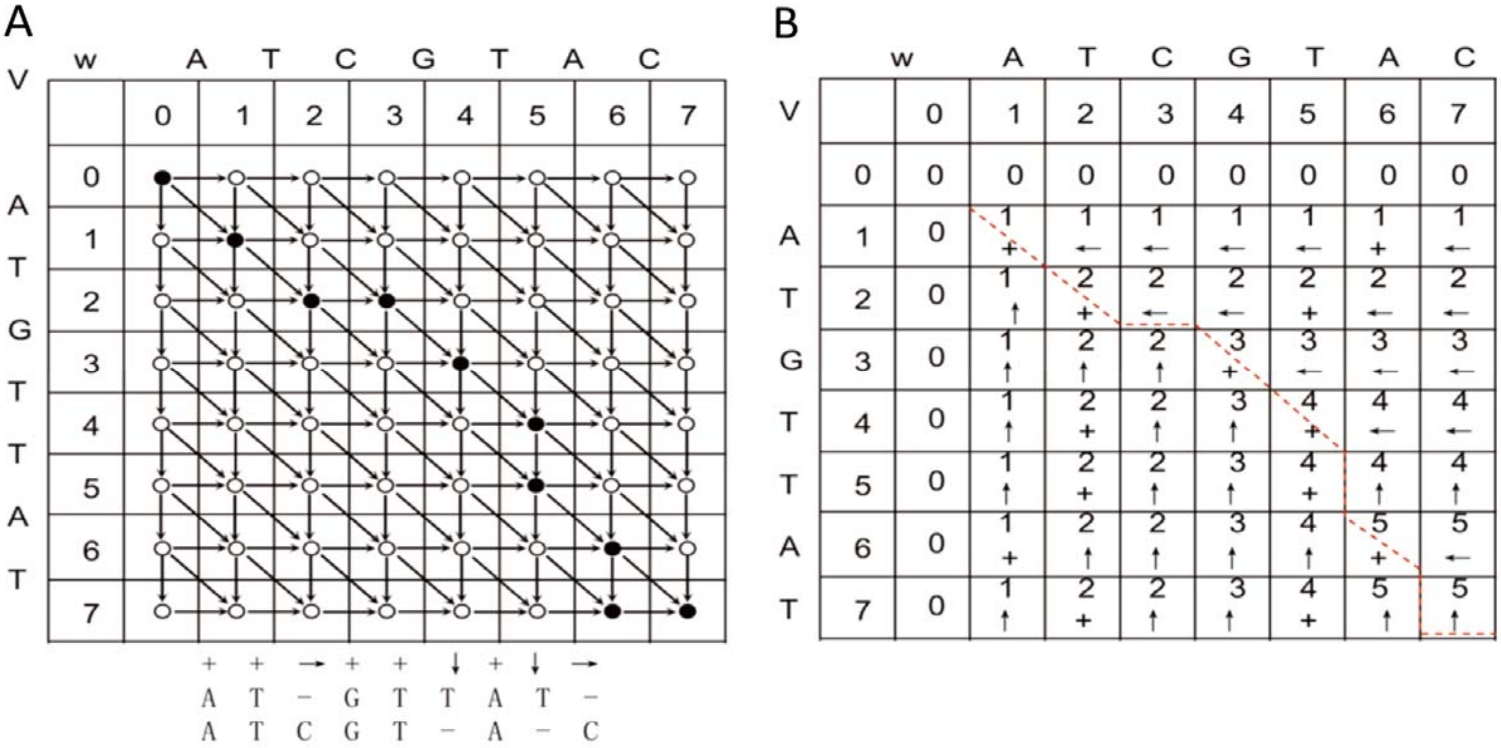
TGCATAT → (INSERT A at front)
ATGCATAT → (DELETE 6th T)
ATGCAAT → (SUBSTITUTE G for 5th A)
ATGCGAT → (SUBSTITUTE C for 3rd G)
ATCCGAT (Done)

Is 4 the minimum edit distance? 3?

A little jargon: *Since the effect of insertion in one string can be accomplished via a deletion in the other string these two operations are quite similar. Often algorithms will consider them together as a single operation called* ***INDEL***

三、算法实现的比对

- 用计算机科学的术语来说，比对两个序列就是找出两个序列的最长公共子序列（**longest common subsequence，LCS**），它反映了两个序列的最高相似度。



动态规划法示意

（A）使用动态规划法寻找两个序列的最长公共部分；

（B）动态规划表的填写。

四、序列比对的作用

- 获得共性序列
- 序列测序
- 突变分析
- 种系分析
- 保守区段分析
- 基因和蛋白质功能分析

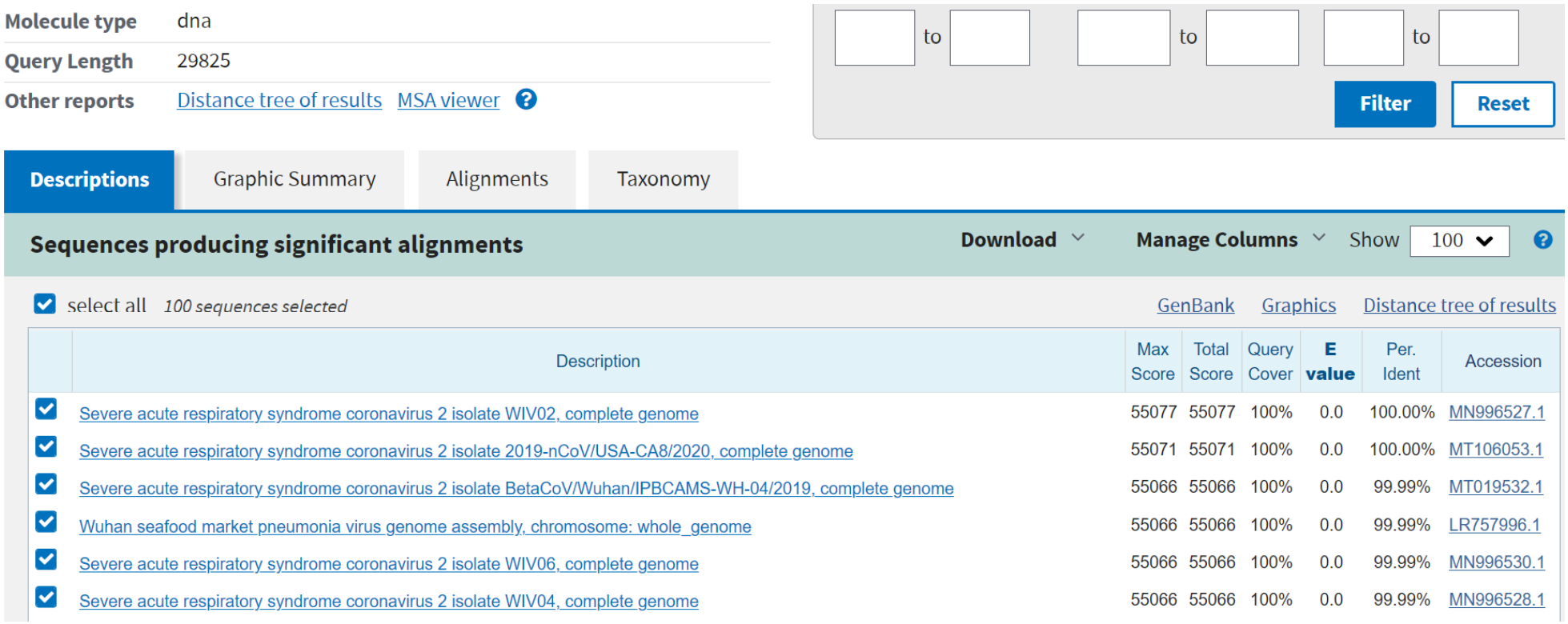
基本方法

- 序列比对

```
>MN996527.1 Severe acute respiratory syndrome coronavirus 2 isolate WIV02, complete
genome
AACCAACTTTCGATCTCTTGTAGATCTGTTCTCTAAACGAACTTTAAAATCTGTGTGGCTGTCACTCGGC
TGCATGCTTAGTGCACTCACGCAGTATAATTAATAACTAATTACTGTCGTTGACAGGACACGAGTAACTC
GTCTATCTTCTGCAGGCTGCTTACGGTTTCGTCCGTGTTGCAGCCGATCATCAGCACATCTAGGTTTCGT
CCGGGTGTGACCGAAAGGTAAGATGGAGAGCCTTGTCCCTGGTTTCAACGAGAAAACACACGTCCAACTC
AGTTTGCCTGTTTTACAGGTTCGCGACGTGCTCGTACGTGGCTTTGGAGACTCCGTGGAGGAGGTCTTAT
CAGAGGCACGTCAACATCTTAAAGATGGCACTTGTGGCTTAGTAGAAGTTGAAAAAGGCGTTTTGCCTCA
ACTTGAACAGCCCTATGTGTTCATCAAACGTTCGGATGCTCGAACTGCACCTCATGGTCATGTTATGGTT
GAGCTGGTAGCAGAACTCGAAGGCATTCAGTACGGTCGTAGTGGTGAGACACTTGGTGTCCTTGTCCCTC
ATGTGGGCGAAATACCAGTGGCTTACCGCAAGGTTCTTCTTCGTAAGAACGGTAATAAAGGAGCTGGTGG
CCATAGTTACGGCGCCGATCTAAAGTCATTTGACTTAGGCGACGAGCTTGGCACTGATCCTTATGAAGAT
TTTCAAGAAAACTGGAACACTAAACATAGCAGTGGTGTTACCCGTGAACTCATGCGTGAGCTTAACGGAG
GGGCATACACTCGCTATGTCGATAACAACTTCTGTGGCCCTGATGGCTACCCTCTTGAGTGCATTAAAGA
CCTTCTAGCACGTGCTGGTAAAGCTTCATGCACTTTGTCCGAACAACTGGACTTTATTGACACTAAGAGG
GGTGTATACTGCTGCCGTGAACATGAGCATGAAATTGCTTGGTACACGGAACGTTCTGAAAAGAGCTATG
AATTGCAGACACCTTTTGAAATTAAATTGGCAAAGAAATTTGACACCTTCAATGGGGAATGTCCAAATTT
TGTATTTCCCTTAAATTCCATAATCAAGACTATTCAACCAAGGGTTGAAAAGAAAAAGCTTGATGGCTTT
ATGGGTAGAATTCGATCTGTCTATCCAGTTGCGTCACCAAATGAATGCAACCAAATGTGCCTTTCAACTC
```

基本方法

- 序列比对



Molecule type	dna
Query Length	29825
Other reports	Distance tree of results MSA viewer

to to to Filter Reset

Descriptions | Graphic Summary | Alignments | Taxonomy

Sequences producing significant alignments Download Manage Columns Show 100

select all *100 sequences selected* GenBank Graphics Distance tree of results

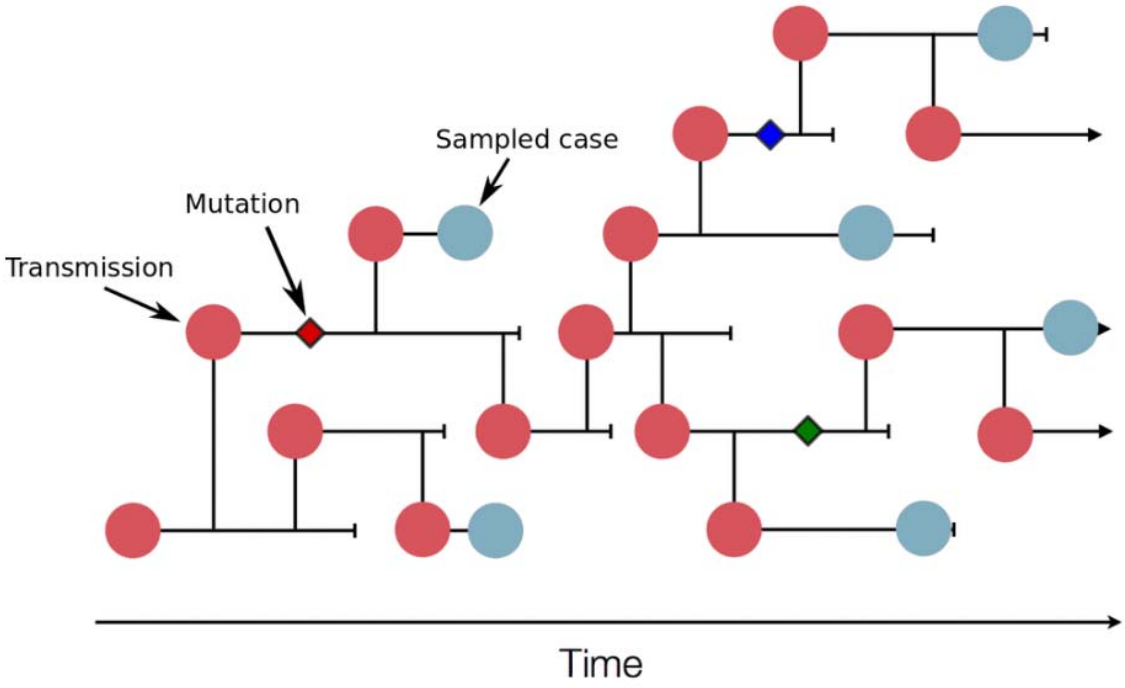
Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
Severe acute respiratory syndrome coronavirus 2 isolate WIV02, complete genome	55077	55077	100%	0.0	100.00%	MN996527.1
Severe acute respiratory syndrome coronavirus 2 isolate 2019-nCoV/USA-CA8/2020, complete genome	55071	55071	100%	0.0	100.00%	MT106053.1
Severe acute respiratory syndrome coronavirus 2 isolate BetaCoV/Wuhan/IPBCAMS-WH-04/2019, complete genome	55066	55066	100%	0.0	99.99%	MT019532.1
Wuhan seafood market pneumonia virus genome assembly, chromosome: whole_genome	55066	55066	100%	0.0	99.99%	LR757996.1
Severe acute respiratory syndrome coronavirus 2 isolate WIV06, complete genome	55066	55066	100%	0.0	99.99%	MN996530.1
Severe acute respiratory syndrome coronavirus 2 isolate WIV04, complete genome	55066	55066	100%	0.0	99.99%	MN996528.1

基本方法

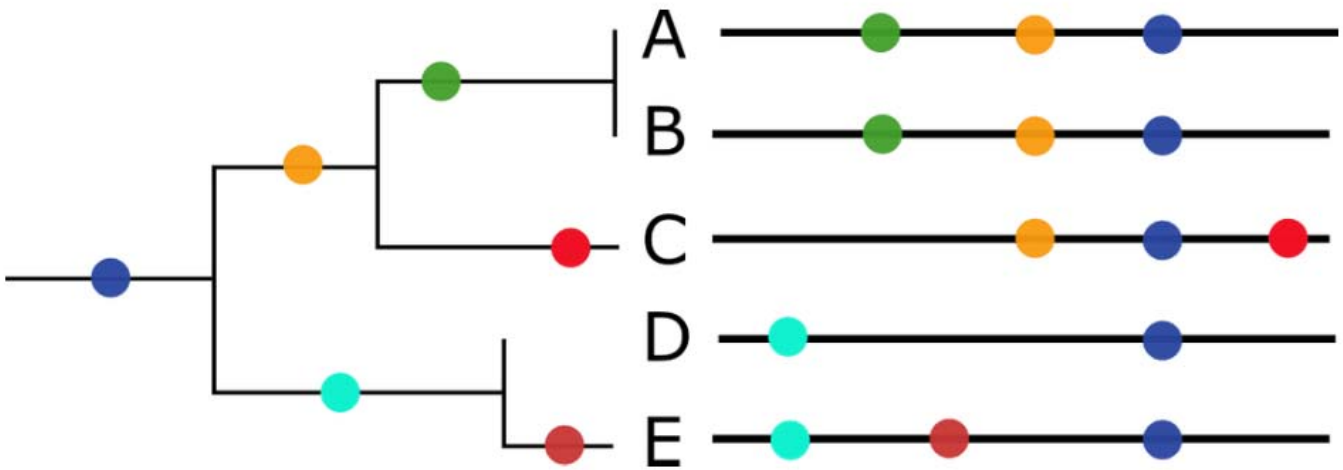
- 序列比对

☑	Severe acute respiratory syndrome coronavirus 2 isolate 2019-nCoV/USA-CA1/2020, complete genome	55027	55027	100%	0.0	99.97%	MN994467.1
☑	Severe acute respiratory syndrome coronavirus 2 isolate HZ-1, complete genome	55025	55025	99%	0.0	99.99%	MT039873.1
☑	Severe acute respiratory syndrome coronavirus 2 isolate SNU01, complete genome	55016	55016	100%	0.0	99.96%	MT039890.1
☑	Severe acute respiratory syndrome coronavirus 2 isolate SARS0CoV-2/61-TW/human/2020/ NPL, complete genome	55001	55001	99%	0.0	99.99%	MT072688.1
☑	Severe acute respiratory syndrome coronavirus 2 isolate Australia/VIC01/2020, complete genome	54985	54985	100%	0.0	99.95%	MT007544.1
☑	Severe acute respiratory syndrome coronavirus 2 isolate 2019-nCoV/USA-CA6/2020, complete genome	54911	54911	100%	0.0	99.91%	MT044258.1
☑	Bat coronavirus RaTG13, complete genome	48627	48627	100%	0.0	96.11%	MN996532.1
☑	Severe acute respiratory syndrome coronavirus 2 isolate nCoV-FIN-29-Jan-2020, partial genome	39748	54706	99%	0.0	99.84%	MT020781.1
☑	Bat SARS-like coronavirus isolate bat-SL-CoVZC45, complete genome	26892	35252	95%	0.0	89.10%	MG772933.1
☑	Bat SARS-like coronavirus isolate bat-SL-CoVZXC21, complete genome	22218	35193	94%	0.0	88.65%	MG772934.1
☑	SARS coronavirus ZS-C, complete genome	15208	22487	88%	0.0	82.34%	AY395003.1
☑	SARS coronavirus ZS-B, complete genome	15208	22513	88%	0.0	82.34%	AY394996.1
☑	SARS coronavirus SZ16, complete genome	15197	22507	88%	0.0	82.33%	AY304488.1
☑	SARS coronavirus SZ3, complete genome	15197	22492	88%	0.0	82.33%	AY304486.1
☑	SARS coronavirus BJ182-12, complete genome	15180	22420	88%	0.0	82.31%	EU371564.1
☑	Coronavirus BtRs-BetaCoV/YN2018B, complete genome	15171	22585	91%	0.0	82.31%	MK211376.1

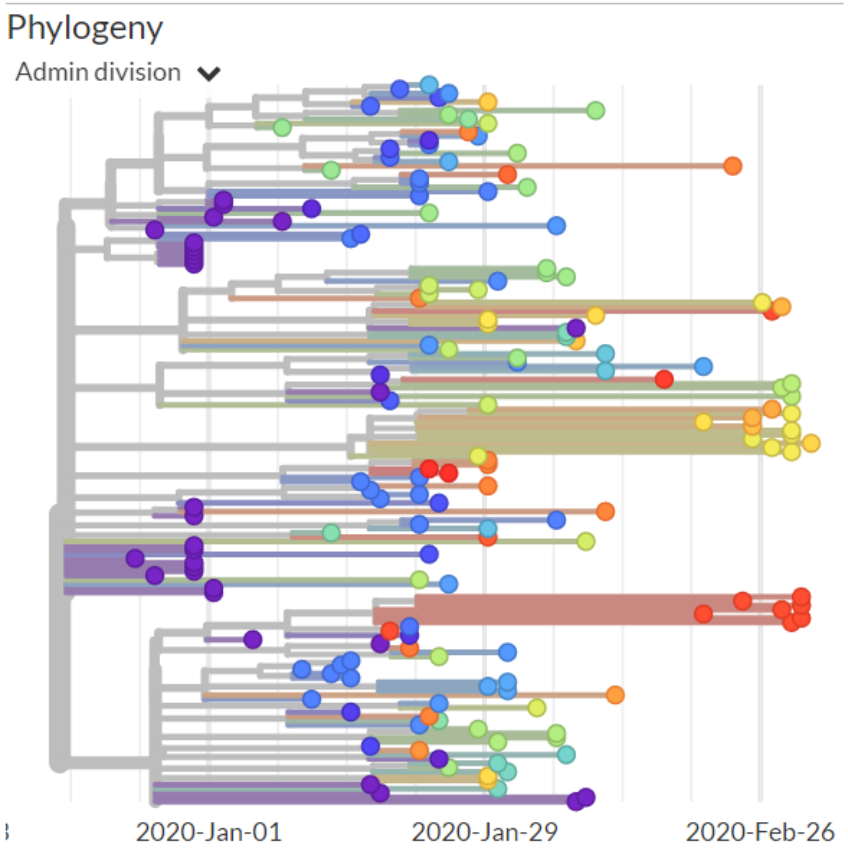
病毒溯源

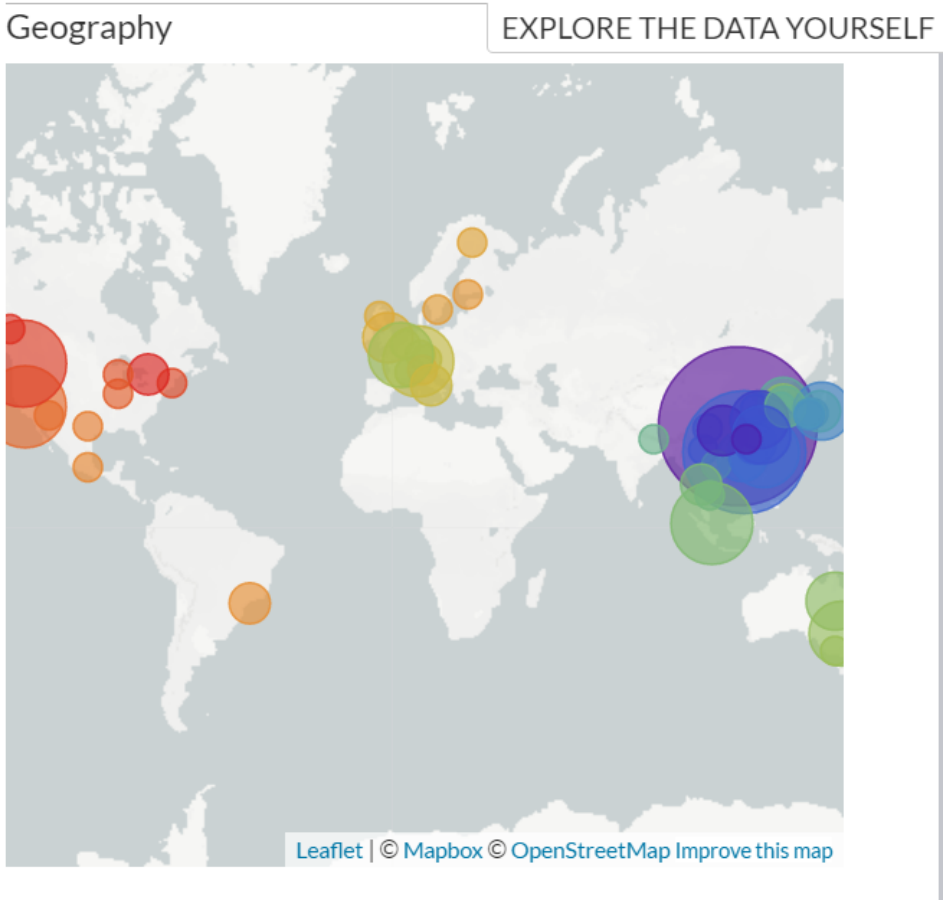


病毒溯源



病毒溯源


Phylogeny
Admin division
2020-Jan-01
2020-Jan-29
2020-Feb-26


Geography
EXPLORE THE DATA YOURSELF
Leaflet | © Mapbox © OpenStreetMap Improve this map

Narrative: 新型冠状病毒（COVID-19）传...

中文

至少两次的输入感染导致在意大利的爆发，这两次都可能存在社区传播

我们目前有3个来自意大利的序列，其中两个来自罗马地区，一个来自意大利北部的伦巴第.

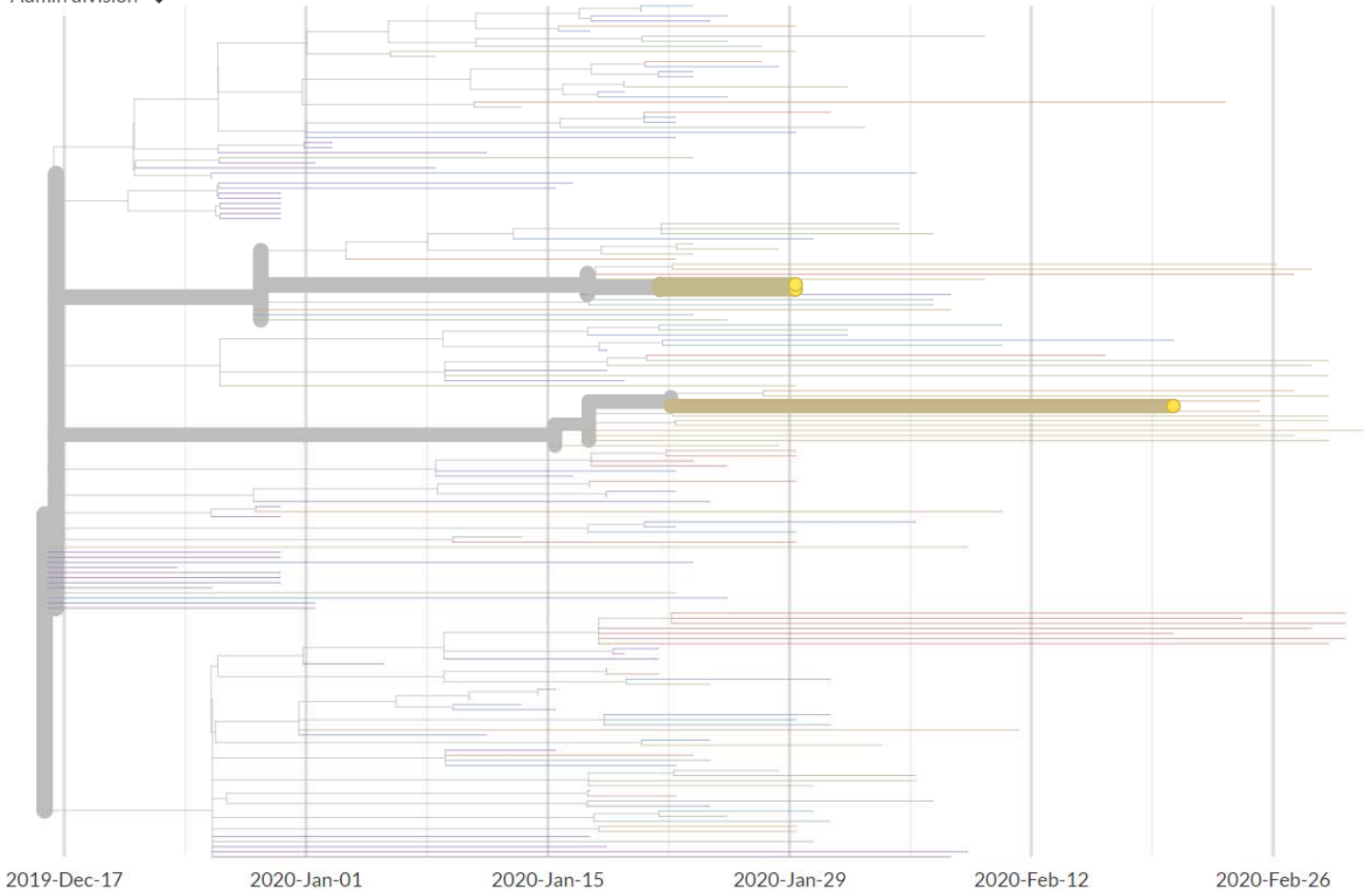
这3个序列在疫情早期共享一个共同的祖先(在树基附近，左边)，这强烈地表明至少有两个病例输入到意大利，并通过社区传播遍了该地区.

努诺·法里亚博士团队很好地细分了巴西和其他全球序列如何表明"意大利北部的疫情很可能是多个人传入该地区的结果，而不是来自一个单一的来源 链接 (英文).

Phylogeny

EXPLORE THE DATA YOURSE

Admin division



Narrative: 新型冠状病毒（COVID-19）传...

中文

SARS-CoV-2可能在西雅图地区蔓延

现在在大西雅图地区和整个美国报告了许多例COVID-19病例. 新分离和测序的病例在基因上与1月中旬在同一地区分离的一例密切相关.

对此，我们有两种可能的解释. 一种解释是，该病毒可能至少两次从中国的共同源头传入大西雅图地区. 而另一种解释是，病毒在该地区传播了一段时间，只是没有被发现.

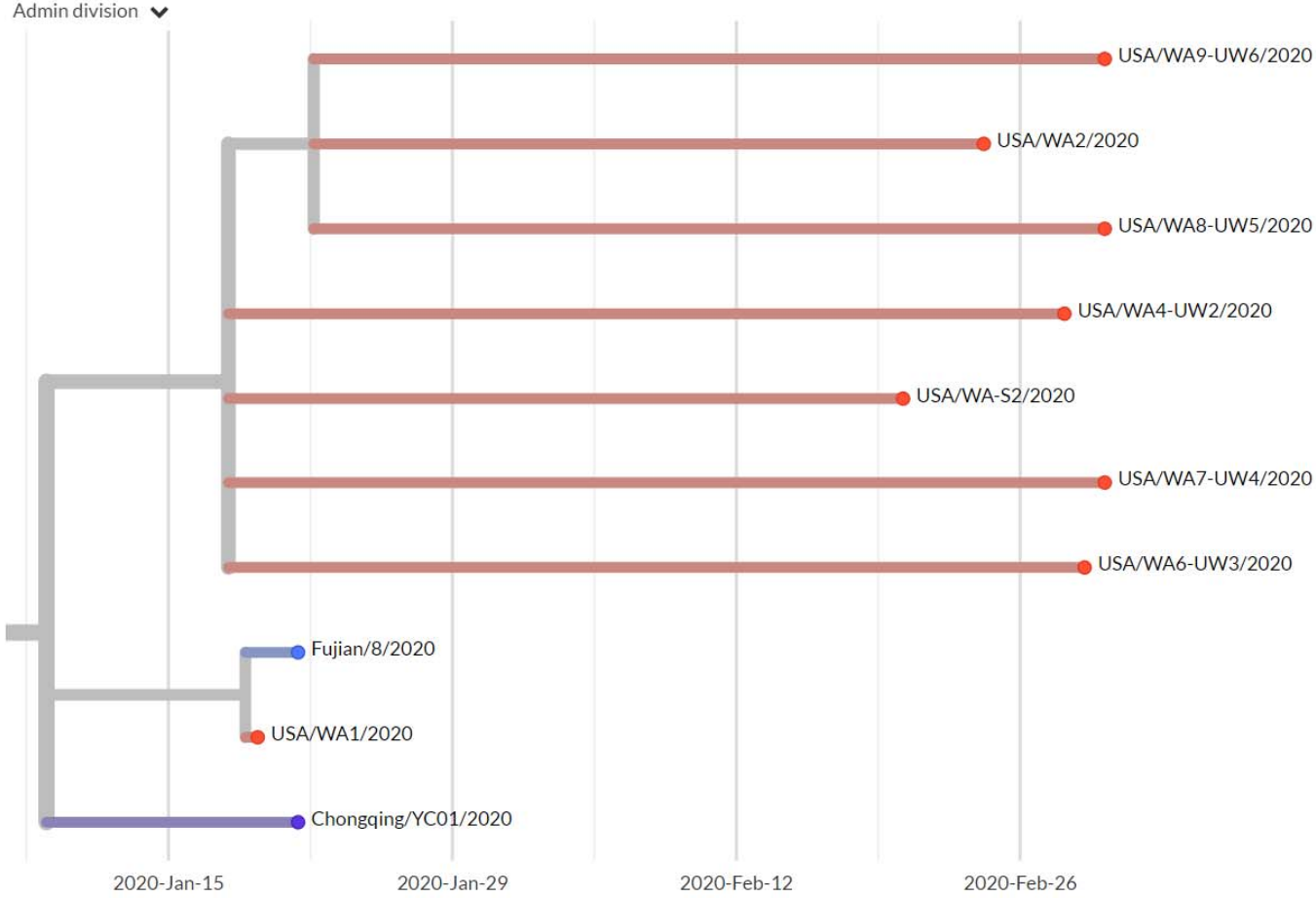
特雷弗·贝德福德(Nextstrain的联合创始人)就这些可能性写了一篇很棒的博客文章，你可以在这里读到 链接 (英文).

华盛顿最近的其他序列告诉我们另一件事：这些序列和来自大西雅图地区的序列，在分析中被聚集在了一起. 这强烈表明已存在社区传播，且SARS-CoV-2病毒已经在该地区传播了一段时间.

Phylogeny

Admin division

EXPLORE THE DATA YOURSELF

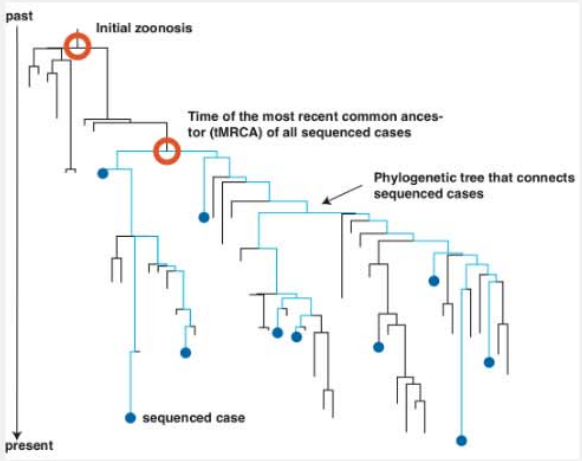


Narrative: 新型冠状病毒（COVID-19）传...

中文

测定最近共同祖先的时间

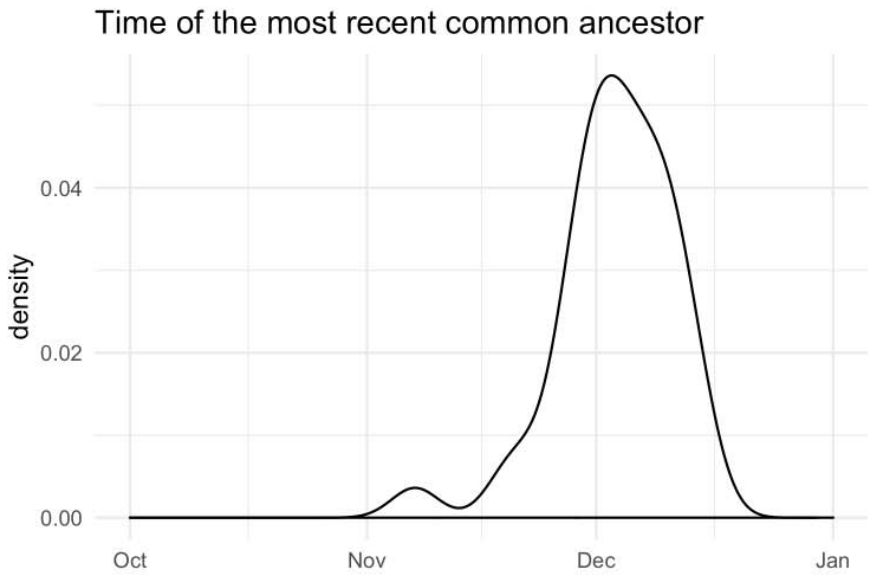
一组测序案例的最近共同祖先(或tMRCA)的时间表示这些测序案例最后共享共同祖先的时间. 这一时间可以早于病毒首次进入人类群体的时间，但也可以大大晚于此时间，如下图所示.



爆发病毒共同祖先的日期

数个研究小组估算了最近共同祖先的时间 - 查看 A Rambaut的文章 (英文) 或 T Stadler的文章 (英文).

所有序列的共同祖先最有可能出现在11月中旬至12月中旬之间. 这与目前所有测序是从武汉华南海鲜市场获得的最初病例群衍生下来的说法一致.

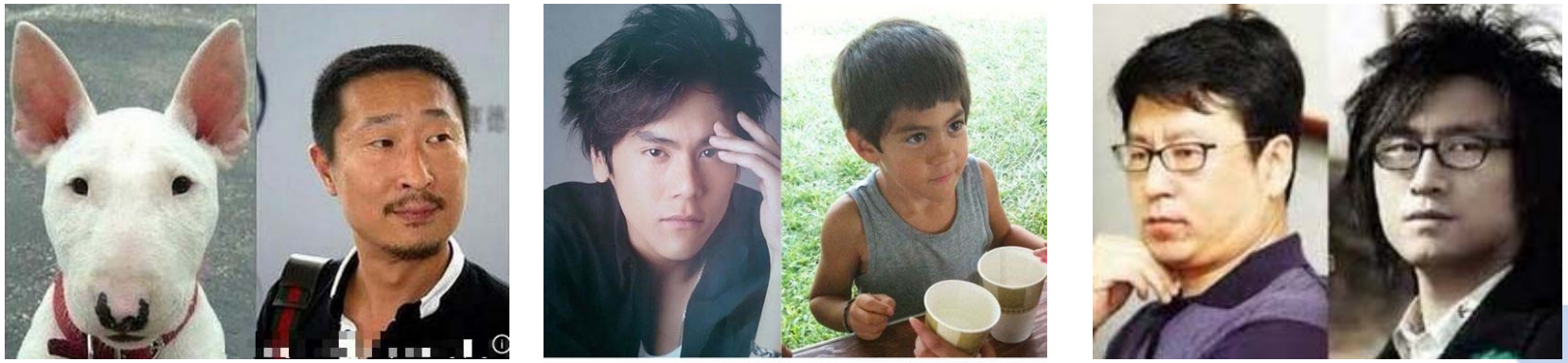


序列相似性

- 序列相似性的重要性

相似的序列往往起源于一个共同的祖先序列。它们很可能有相似的空间结构和生物学功能，因此对于一个已知序列但未知结构和功能的蛋白质，如果与它序列相似的某些蛋白质的结构和功能已知，则可以推测这个未知结构和功能的蛋白质的结构和功能。

结构相似？功能相似？



序列相似性

- 序列一致度（identity）与相似度（similarity）

一致度：又称同一度，如果两个序列（蛋白质或核酸）长度相同，那么它们的一致度定义为他们对应位置上相同的残基（一个字母，氨基酸或碱基）的数目占总长度的百分数。

相似度：如果两个序列（蛋白质或核酸）长度相同，那么它们的相似度定义为他们对应位置上相似的残基与相同的残基的数目和占总长度的百分数。

问题：哪个残基与哪个残基算作相似

答：残基两两相似的量化关系被替换记分矩阵所定义。

	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala	4																			
Arg	-1	5																		
Asn	-2	0	6																	
Asp	-2	-2	1	6																
Cys	0	-3	-3	-3	9															
Gln	-1	1	0	0	-3	5														
Glu	-1	0	0	2	-4	2	5													
Gly	0	-2	0	-1	-3	-2	-2	6												
His	-2	0	1	-1	-3	0	0	-2	8											
Ile	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
Leu	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
Lys	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5								
Met	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
Phe	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
Pro	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
Ser	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
Thr	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5			
Trp	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Tyr	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
Val	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

序列 1 ： CLHK

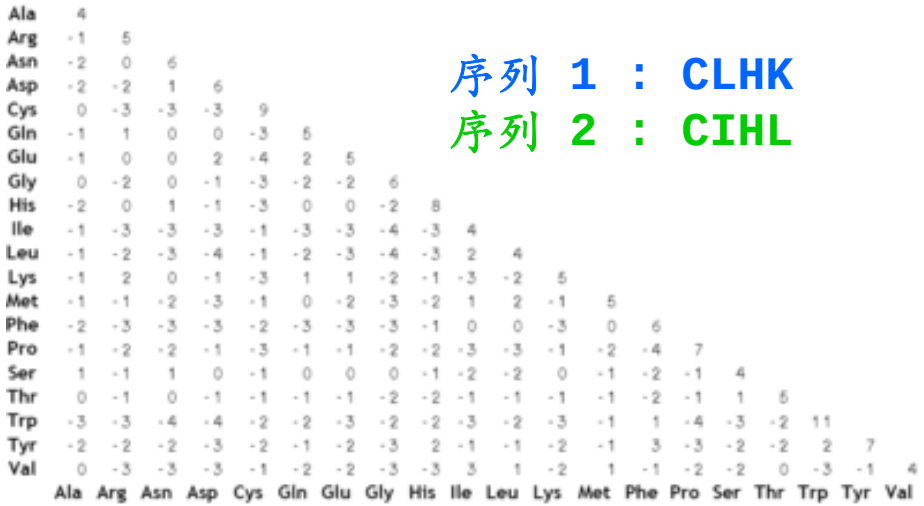
序列 2 ： CIHL

替换记分矩阵

替换记分矩阵（Substitution Matrix）：反映残基之间相互替换率的矩阵，它描述了残基两两相似的量化关系。分为DNA替换记分矩阵和蛋白质替换记分矩阵。

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

DNA替换记分矩阵



序列 1 : CLHK
序列 2 : CIHL

蛋白质替换记分矩阵

DNA序列的替换记分矩阵

- 3种常见的DNA序列的替换记分矩阵

1. 等价矩阵（unitary matrix）：最简单的替换记分矩阵，其中，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。由于不含有碱基的理化信息和不区别对待不同的替换，在实际的序列比较中较少使用。

2. 转换-颠换矩阵（transition-transversion matrix）：核酸的碱基按照环结构特征被划分为两类，一类是嘌呤（腺嘌呤A、鸟嘌呤G），它们有两个环；另一类是嘧啶（胞嘧啶C、胸腺嘧啶T），它们只有一个环。如果DNA碱基的替换保持环数不变，则称为转换，如A→G、C→T；如果环数发生变化，则称为颠换，如A→C、T→G等。在进化过程中，转换发生的频率远比颠换高。为了反映这一情况，通常该矩阵中转换的得分为-1，而颠换的得分为-5。

3. BLAST矩阵：经过大量实际比对发现，如果令被比对的两个核苷酸相同时得分为+5，反之为-4，则比对效果较好。这个矩阵广泛地被DNA序列比较所采用。

①

	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

	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

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

蛋白质序列的替换记分矩阵

- 3种常见的蛋白质序列的替换记分矩阵

1. **等价矩阵（unitary matrix）**：与DNA等价矩阵道理相同，相同氨基酸之间的匹配得分为1，不同氨基酸间的替换得分为0。在实际的序列比对中较少使用。

2. **PAM矩阵（Dayhoff突变数据矩阵）**：PAM矩阵基于进化原理。如果两种氨基酸替换频繁，说明自然界易接受这种替换，那么这对氨基酸替换得分就应该高。PAM矩阵是目前蛋白质序列比较中最广泛使用的记分方法之一，基础的PAM-1矩阵反应的是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。

3. **BLOSUM矩阵（blocks substitution matrix）**：BLOSUM矩阵都是通过对大量符合特定要求的序列计算而来的。PAM-1矩阵是基于相似度大于85%的序列计算产生的，那些进化距离较远的矩阵，如PAM-250，是通过PAM-1自乘得到的。即，BLOSUM矩阵的相似性是根据真实数据产生的，而PAM矩阵是通过矩阵自乘外推而来的。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，同理，BLOSUM-62是指该矩阵由一致度≥62%的序列计算而来的。

蛋白质序列的替换记分矩阵

- 3种常见的蛋白质序列的替换记分矩阵

BLOSUM-62

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。

	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala	4																			
Arg	-1	5																		
Asn	-2	0	6																	
Asp	-2	-2	1	6																
Cys	0	-3	-3	-3	9															
Gln	-1	1	0	0	-3	5														
Glu	-1	0	0	2	-4	2	5													
Gly	0	-2	0	-1	-3	-2	-2	6												
His	-2	0	1	-1	-3	0	0	-2	8											
Ile	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
Leu	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
Lys	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5								
Met	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
Phe	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
Pro	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
Ser	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
Thr	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5			
Trp	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Tyr	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
Val	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

蛋白质序列的替换记分矩阵

- 选 **PAM-1** 还是 **PAM-250**?

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

- 选 PAM-？　还是　BLOSUM-？

BLOSUM-80	BLOSUM-62	BLOSUM-45
PAM-1	PAM-120	PAM-250
亲缘关系较近的序列之间的比对	⟷	亲缘关系较远的序列之间的比对

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

一致度和相似度的计算

- 序列一致度（identity）与相似度（similarity）

一致度：如果两个序列（蛋白质或DNA）长度相同，那么它们的一致度定义为他们对应位置上相同的残基（一个字母，氨基酸或碱基）的数目占总长度的百分数。

相似度：如果两个序列（蛋白质或DNA）长度相同，那么它们的相似度定义为他们对应位置上相似的残基与相同的残基的数目和占总长度的百分数。

问题：哪个残基与哪个残基算作相似

答：残基两两相似的量化关系被替换记分矩阵所定义。

```
序列 1 : CLHK
序列 2 : CIHL
         ? ?

一致度 = 2/4 = 50%
相似度 = ?
```

一致度和相似度的计算

- 序列一致度（identity）与相似度（similarity）

	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala	4																			
Arg	-1	5																		
Asn	-2	0	6																	
Asp	-2	-2	1	6																
Cys	0	-3	-3	-3	9															
Gln	-1	1	0	0	-3	5														
Glu	-1	0	0	2	-4	2	5													
Gly	0	-2	0	-1	-3	-2	-2	6												
His	-2	0	1	-1	-3	0	0	-2	8											
Ile	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
Leu	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
Lys	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5								
Met	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
Phe	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
Pro	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
Ser	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
Thr	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5			
Trp	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Tyr	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
Val	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

BLOSUM-62

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。

一致度和相似度的计算

- 序列一致度（identity）与相似度（similarity）

一致度：如果两个序列（蛋白质或DNA）长度相同，那么它们的一致度定义为他们对应位置上相同的残基（一个字母，氨基酸或碱基）的数目占总长度的百分数。

相似度：如果两个序列（蛋白质或DNA）长度相同，那么它们的相似度定义为他们对应位置上相似的残基与相同的残基的数目和占总长度的百分数。

问题：哪个残基与哪个残基算作相似

答：残基两两相似的量化关系被替换记分矩阵所定义。

序列 1 : CLHK
序列 2 : CIHL

一致度 = 2/4 = 50%
相似度 = (2+1)/4 = 75%

第二节

比对算法概要

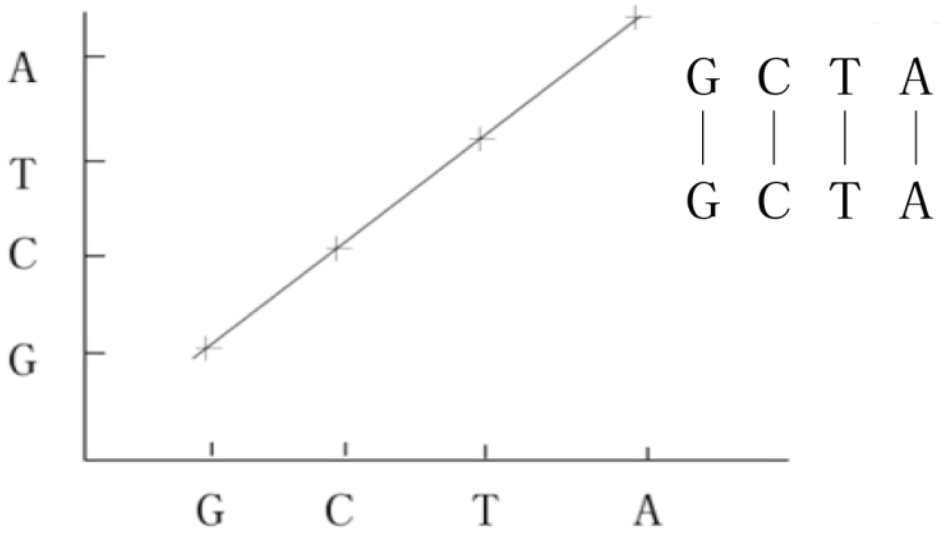
Section 2 Alignment Algorithms

人民卫生出版社
PEOPLE'S MEDICAL PUBLISHING HOUSE

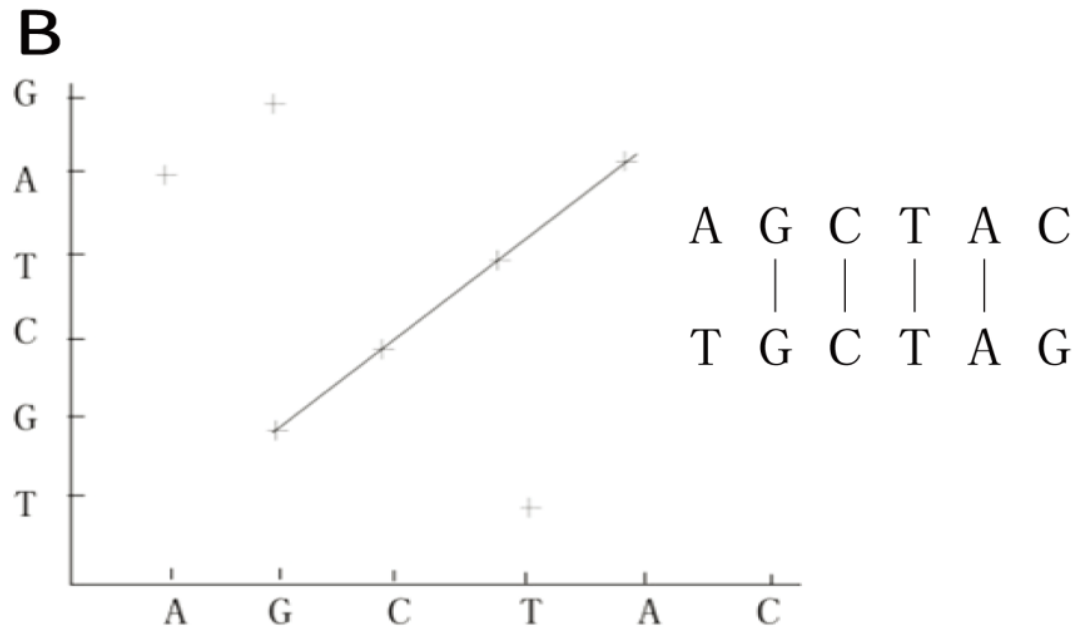
一、替换计分矩阵

（一）通过点矩阵对序列比较进行计分

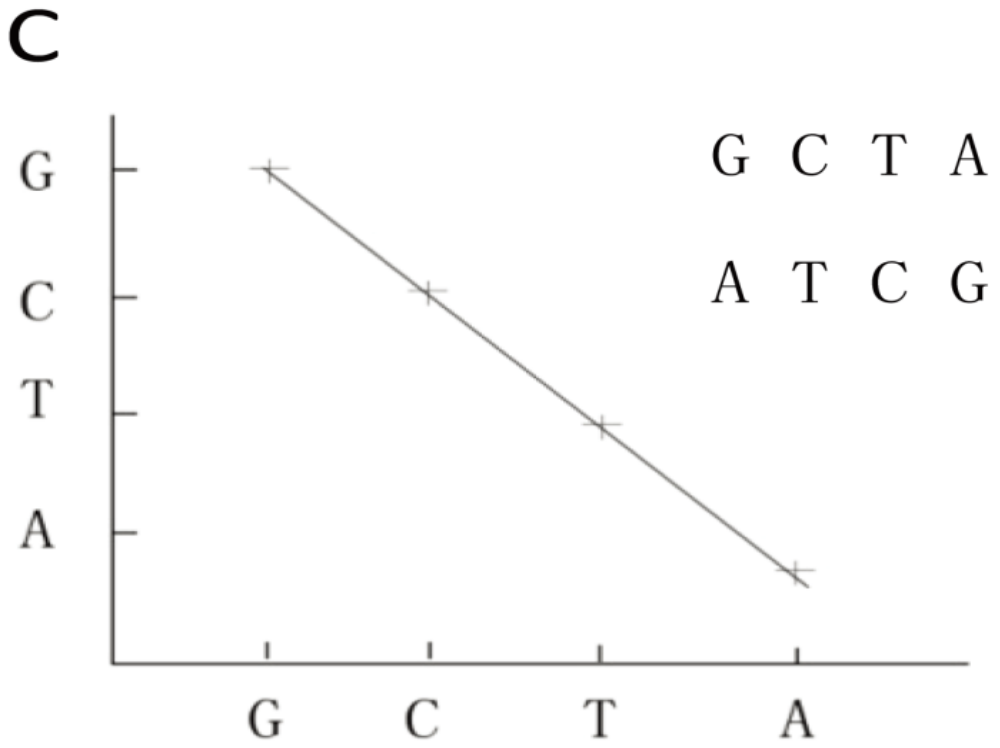
A



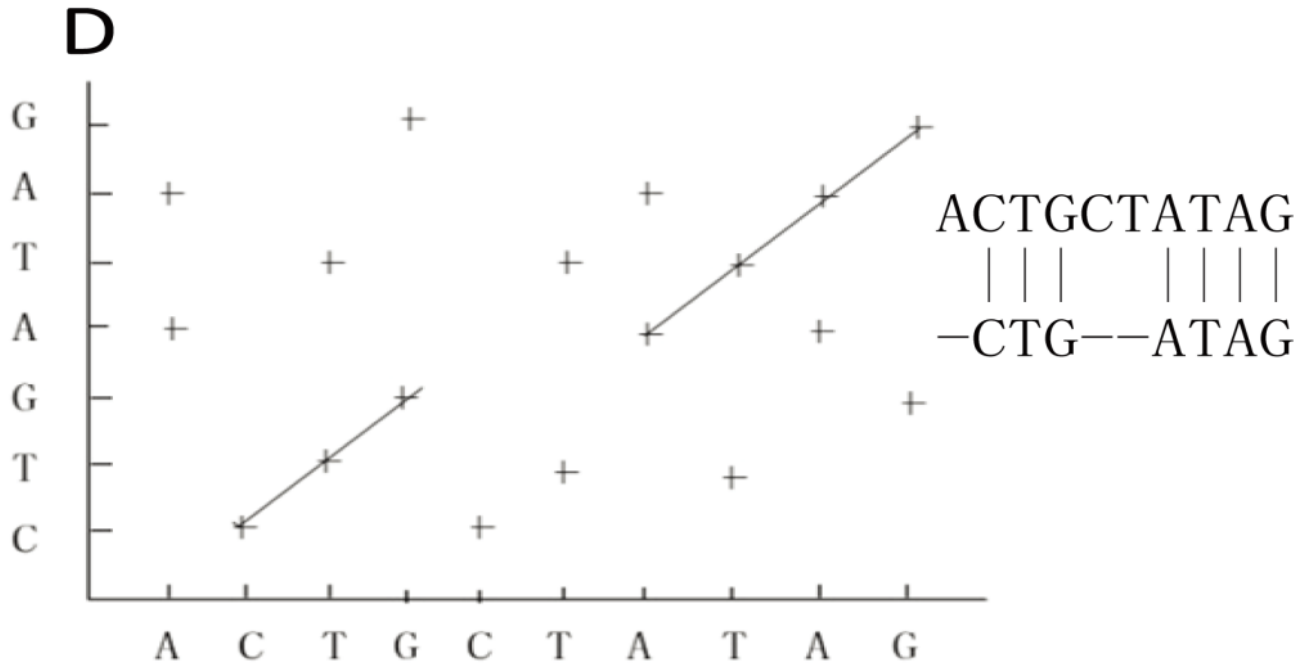
A.两条序列完全相同



B.两条序列有一个共同的子序列



C.两条序列反向匹配



D.两条序列存在不连续的两条子序列

最简单的比较两个序列的方法，理论上可以用  来完成。

Seq1：
THEFASTCAT

Seq2：
THEFATCAT

最简单的比较两个序列的方法，理论上可以用  来完成。

Seq1:
THEFASTCAT

Seq2:
THEFATCAT

Seq 1

Seq 2	T	H	E	F	A	S	T	C	A	T
T	X						X			X
H		X								
E			X							
F				X						
A					X				X	
T	X						X			X
C								X		
A					X				X	
T	X						X			X

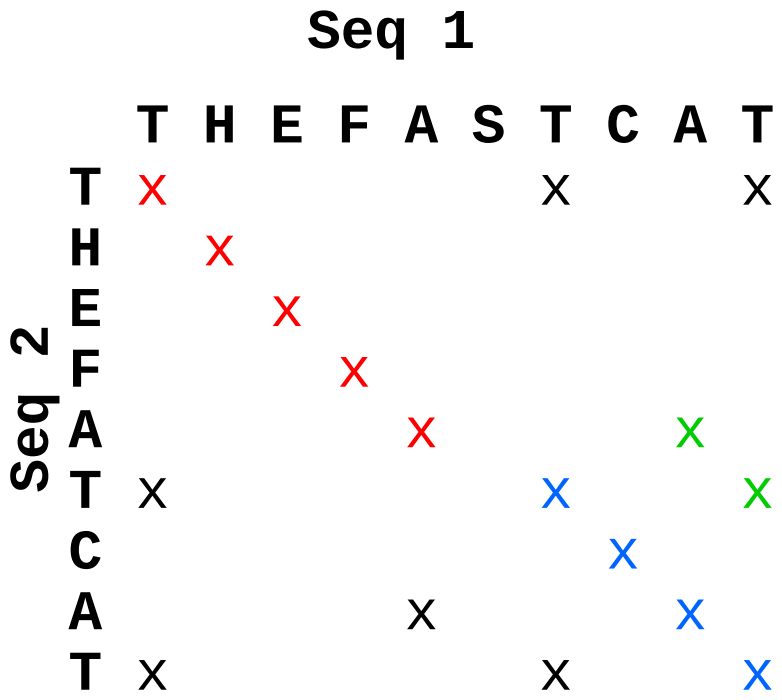
打点法的用途

1. THEFA
2. TCAT
3. AT

连续的对角线及对角线的平行线代表两条序列中相同的区域

Seq1:
THEFASTCAT

Seq2:
THEFATCAT



打点法的用途

可以用一条序列自己对自己打点，从而可以发现序列中重复的片段。这样的打点矩阵必然是对称的，并且有一条主对角线。在横向或纵向上，与主对角线平行的短平行线所对应的序列片段就是重复的部分。

Seq1：
THEFASTHESTHE

	T	H	E	F	A	S	T	H	E	Y	T	H	E
T	X						X				X		
H		X						X				X	
E			X						X				X
F				X									
A					X								
S						X							
T	X						X						
H		X						X					
E			X						X				
Y										X			
T	X										X		
H		X										X	
E			X										X

打点法的用途

发现串联重复序列（tandem repeat）

Seq1：

FASABCABCABCTHE

短串联重复序列（short tandem repeat, STR）也叫做微卫星DNA，是一类广泛存在于真核生物基因组中的DNA串联重复序列。它由2-6bp的核心序列组成，重复次数通常在15-30次。STR具有高度多态性，即存在重复次数的个体间差异，而且这种差异在基因遗传过程中一般遵循孟德尔共显性遗传规律，所以它被广泛研究。

	F	S	A	B	C	A	B	C	A	B	C	T	H
F	X												
S		X											
A			X			X			X				
B				X			X			X			
C					X			X			X		
A			X			X			X				
B				X			X			X			
C					X			X			X		
A			X			X			X				
B				X			X			X			
C					X			X			X		
T												X	
H													X

MOOC内容

5月5日学习

- 第三章 序列比较 3.4 3.5 3.6