Lecture 7:
SEQUENCE COMPARISON

✓PAIRWISE ALIGNMENT

Part II: Finding similarities

Warmup

Last week we learned :

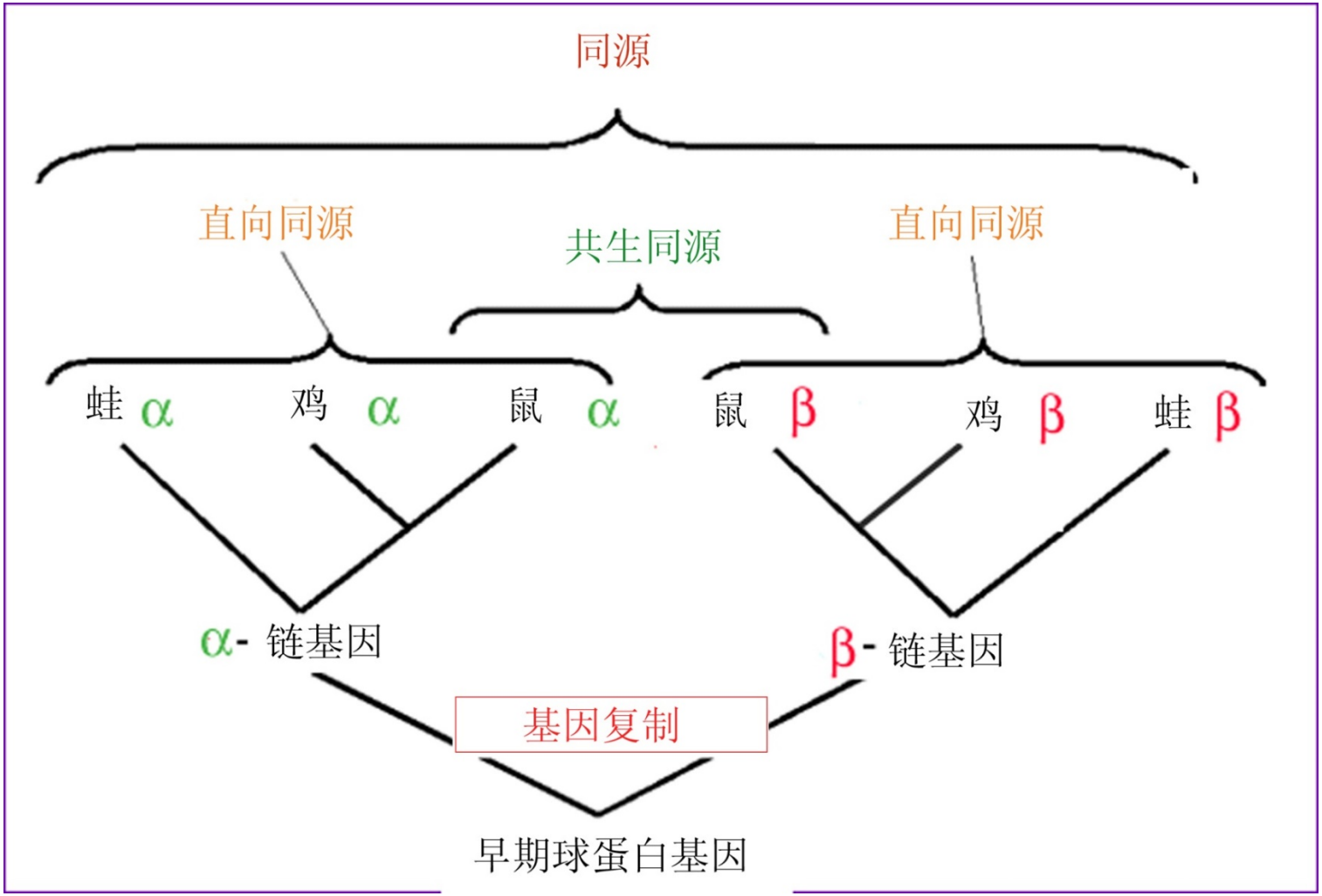
Concepts

- Identity & Similarity;
- mismatch & gap;
- homolog, ortholog, paralog

how to score: Scoring matrix

- DNA
- protein

homolog, ortholog, paralog



DNA Scoring Matrix

等价矩阵(unitary matrix)

	A	T	C	G
A	1	0	0	0
T	0	1	0	0
C	0	0	1	0
G	0	0	0	1

转换-颠换矩阵(transition-transversion matrix)

	A	T	C	G
A	1	5	5	1
T	5	-4	-4	-4
C	-5	1	-1	-5
G	-4	5	-4	-4
	-5	-1	1	-5
	-4	-4	5	-4
	-1	-5	-5	1
	-4	-4	-4	5

BLAST矩阵(Blast matrix)

protein Scoring Matrix

等价矩阵(unitary matrix)

遗传密码矩阵(genetic code matrix，GCM)

疏水性矩阵(hydrophobic matrix)

根据氨基酸侧链基团疏水性的不同，以及替换前后理化性质变化的大小而制定

适用于偏重蛋白质功能方面的序列比对

PAM矩阵

BLOSUM矩阵

	A	S	G	L	K	V	T	P	E	D	N	I	Q	R	F	Y	C	H	M	W	Z	B	X
A	2	1	1	2	2	1	1	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
S	1	0	1	1	2	1	1	1	1	1	1	1	2	1	1	1	1	2	2	1	2	2	2
G	1	1	0	2	2	1	2	2	1	1	2	2	2	1	2	2	1	2	2	1	2	2	2
L	2	2	2	0	2	1	2	2	2	2	2	1	2	2	1	1	2	2	1	2	2	2	2
K	2	2	2	2	0	2	2	2	1	1	1	2	1	1	2	2	2	2	2	2	1	1	2
V	1	2	1	2	2	0	2	2	2	2	2	1	2	2	1	2	2	2	1	2	2	2	2
T	1	1	2	2	2	2	0	2	2	2	1	1	2	1	2	2	2	2	1	2	2	2	2
P	1	1	2	1	2	2	1	0	2	2	2	2	1	1	2	2	2	1	2	2	2	2	2
E	2	1	1	2	1	2	2	2	0	1	2	2	1	2	2	2	2	2	2	1	2	2	2
D	2	1	1	2	1	2	2	2	1	0	2	2	2	2	2	1	2	1	2	2	2	1	2
N	2	1	2	2	1	2	1	2	2	1	0	1	2	2	2	1	2	1	2	2	2	1	2
I	2	1	2	1	1	1	1	2	2	2	1	0	2	1	1	2	2	2	1	2	2	2	2
Q	2	2	2	1	1	2	2	1	1	2	2	2	0	1	2	2	2	1	2	2	1	2	2
R	2	1	1	1	1	2	1	1	2	2	2	1	1	0	2	2	1	1	1	1	2	2	2
F	2	1	2	1	2	1	2	2	2	2	2	1	2	2	0	1	1	2	2	2	2	2	2
Y	2	1	2	2	2	2	2	2	2	1	1	2	2	2	1	0	1	1	3	2	2	1	2
C	2	1	1	2	2	2	2	2	2	2	2	2	2	1	1	1	0	2	2	1	2	2	2
H	2	2	2	1	2	2	2	1	2	1	1	2	1	1	2	1	2	0	2	2	2	1	2
M	2	2	2	1	1	1	1	2	2	2	2	1	2	1	2	3	2	2	0	2	2	2	2
W	2	1	1	1	2	2	2	2	2	2	2	2	2	1	2	2	1	2	2	0	2	2	2
Z	2	2	2	2	1	2	2	2	1	2	2	2	1	2	2	2	2	2	2	2	1	2	2
B	2	2	2	2	2	2	2	2	2	1	1	2	2	2	2	1	2	1	2	2	2	1	2
X	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2

Substitution - Replace a residue with another of similar physiochemical property.

Category	Amino Acid
Acids and Amides	Asp (D) Glu(E) Asn (N) Gln (Q)
Basic	His (H) Lys (K) Arg (R)
Aromatic	Phe (F) Tyr (Y) Trp (W)
Hydrophilic	Ala (A) Cys (C) Gly (G) Pro (P) Ser (S) Thr (T)
Hydrophobic	Ile (I) Leu (L) Met (M) Val (V)

PAM
point accepted matrix

- 基于氨基酸进化的点突变模型
- 如果两种氨基酸替换频繁，说明自然界容易接受这种替换，那么这对氨基酸替换的得分就高
- 从蛋白质序列全局比对结果统计而来
- PAM-1矩阵反映的是进化中1%氨基酸发生点突变的替换概率
- PAM-1矩阵自乘n次，得到PAM-n矩阵

BLOSUM
BLOck SUbstitution Matrix

基于蛋白质保守域模型

从蛋白质短序列局部比对结果统计而来

BLOSUM-62用来比较62%相似度的序列，

BLOSUM-80用来比较80%相似度的序列

Choice of Scoring Matrix

PAM-1

BLOSUM100

PAM-250

BLOSUM30

Small evolutionary distance
Strong similarity for short sequence

Large evolutionary distance
Weak similarity over stretched length

```
ATGCATGCTGCCAACGGATGTCCTG
|||  |  ||  |||  |||  |  ||||||
ATGAAAGCCGCCTACGAAAGTCCTG
```

We want computer programs which will compare sequences at all possible different alignments, looking for a degree of similarity greater than we would expect to find by chance.

But first we have to consider the implication of *gaps*…

Insertions and deletions are other possible forms of mutations and they can really mess up our simple alignments:

```
ATGCATGCTGGCCAACGGATGTCCTG
|||  |  ||  |||  |  |   |    |
ATGAAAGCCGCCTACGAAAGTCCTG
```

Gaps in Alignments

Consider these two obviously similar sequences:

```
 TTCCCAACTCTCCTCTTTCACCATGAAGCTCAAGGACAGATTCCACTCGCCCCAAAATCAAGCTCACCCCGTCCAAGAA
 | ||        |   || |||||||||||||||||||| ||||||||| ||| |||   |      |||   | |  |
TTCCCACCTCTCCTCTTTGCACCATGAAGCTCAAGGACAAATTCCACTCCCCCAAAATCAAGCGCACCCCGTCCCAGAA
```

In fact we realise that the most probable alignment (regarding biological origin) is with a small gap in each sequence:

```
TTCCCAACTCTCCTCTTT=CACCATGAAGCTCAAGGACAGATTCCACTCGCCCCAAAATCAAGCTCACCCCGTCCAAGAA
|||||| ||||||||||| |||||||||||||||||||| ||||||||| ||||||||||||||| |||||||||| ||||
TTCCCACCTCTCCTCTTTGCACCATGAAGCTCAAGGACAAATTCCACTC=CCCCAAAATCAAGCGCACCCCGTCCCAGAA
```

So in general we allow ourselves to insert gaps, until we find the optimal alignment.

But where should this process stop?

Cost is good… But…

The Downside of Gaps

Take two random sequences, with no 'real' similarity:

```
GACACTAGGTCGATGCGTGGTGGCGAGA

ACGCATCCGGATGTGCACCGTGGAACTG
```

And allow 'cost free' gaps:

```
GAC--ACT----AGGTCGATGC---GTGG---TGGCGAGA
 ||   | |     |   | | |||    ||||    ||
 ACGCA-TCCGGA--T-G-TGCACCGTGGAACTG
```

Clearly, although the alignment has *no mismatches*, it is obviously not biologically meaningful!

To prevent this we assign a cost to adding gaps which is offset against the benefit of finding matches – and this is the essence of 'finding gapped alignments'.

We want to find the 'alignment' between the two (or more) sequences which shows the greatest degree of similarity while introducing the fewest gaps …

Computers Can Detect Homology

```
ATGCATGCTGCCAACGGATGCCCTG
| | | | | | | | | | | | | | | | |
ATGAAAGCCGCCTACGACAGTCCTG
```

In fact computers are very good at this task – the two primary challenges are:

(a) performing the search fast enough to look through millions of sequence in a timescale compatible with a lab scientist's attention span

(b) at low levels of similarity, being able to distinguish between biologically related sequences and chance matches…

```
GCTGACTCGTAGCGCTTAGCTAGCT
  |       | |       |     |     |     |
CCAACATCTAGCCAGATTAGTTAGT
```

Ways to do Pairwise Alignment

Dot Plot (simplest method)

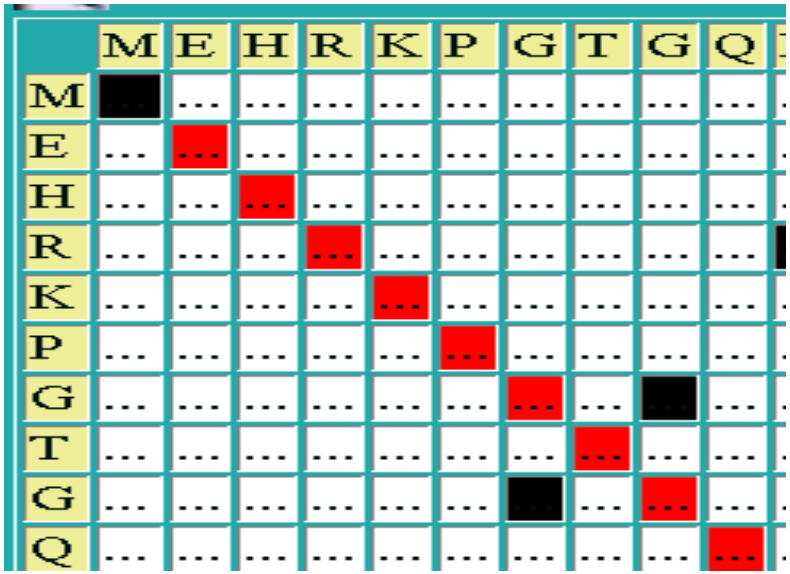
Statistical computation based

- **Local alignment** e.g. **BLAST**, **FASTA**
- **Global alignment**

STEPS IN DOT PLOT

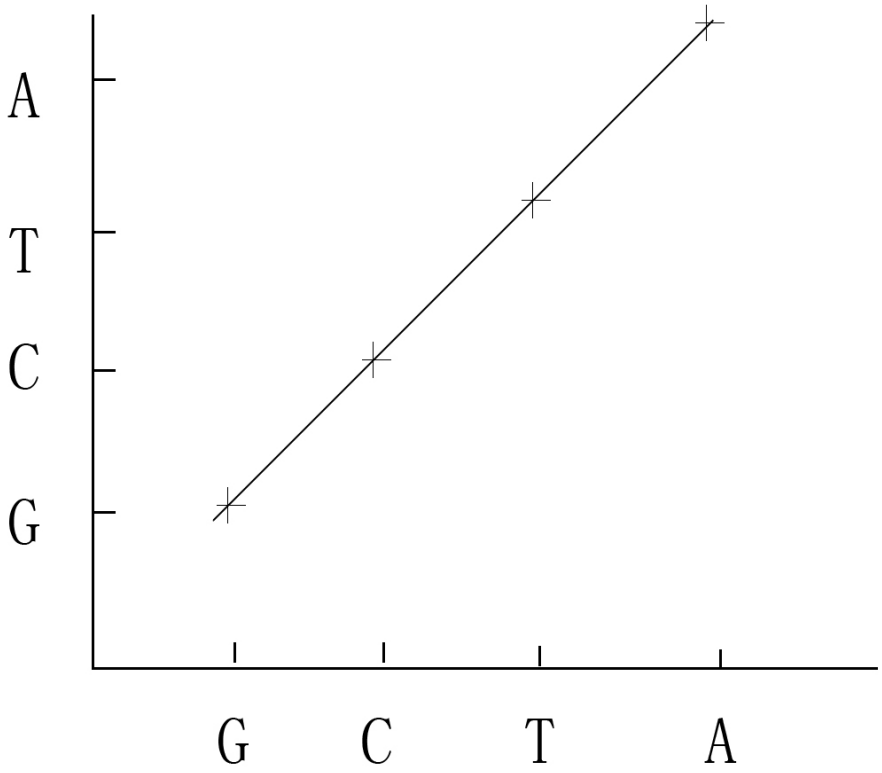
- Take two sequences to be compared
- Sequence A:MEHRKPGTGQ
- Sequence B:MEHRKPGTGQ
- Place sequence A in x-axis (Row). Place sequence B in y-axis (Column)

- Plot a dot everytime there is a match between an element of row sequence and an element of column sequence
- Do you see any diagonal line extending?
- If yes, then there is a match !



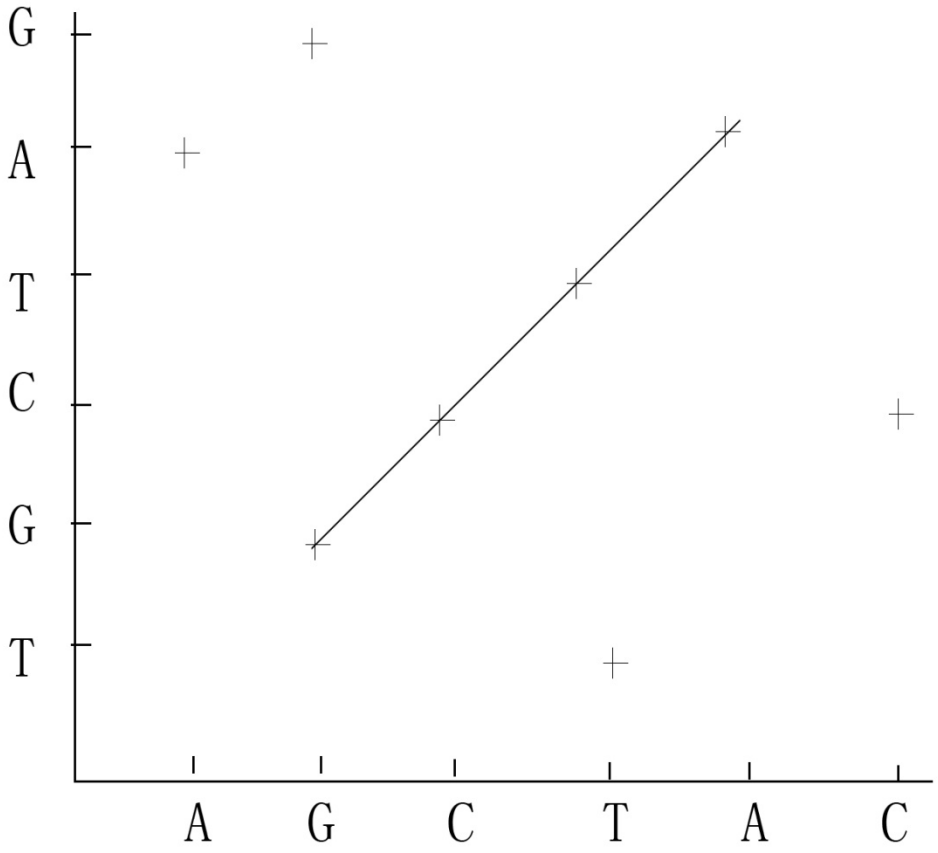
通过点矩阵进行序列比较

两条序列完全相同



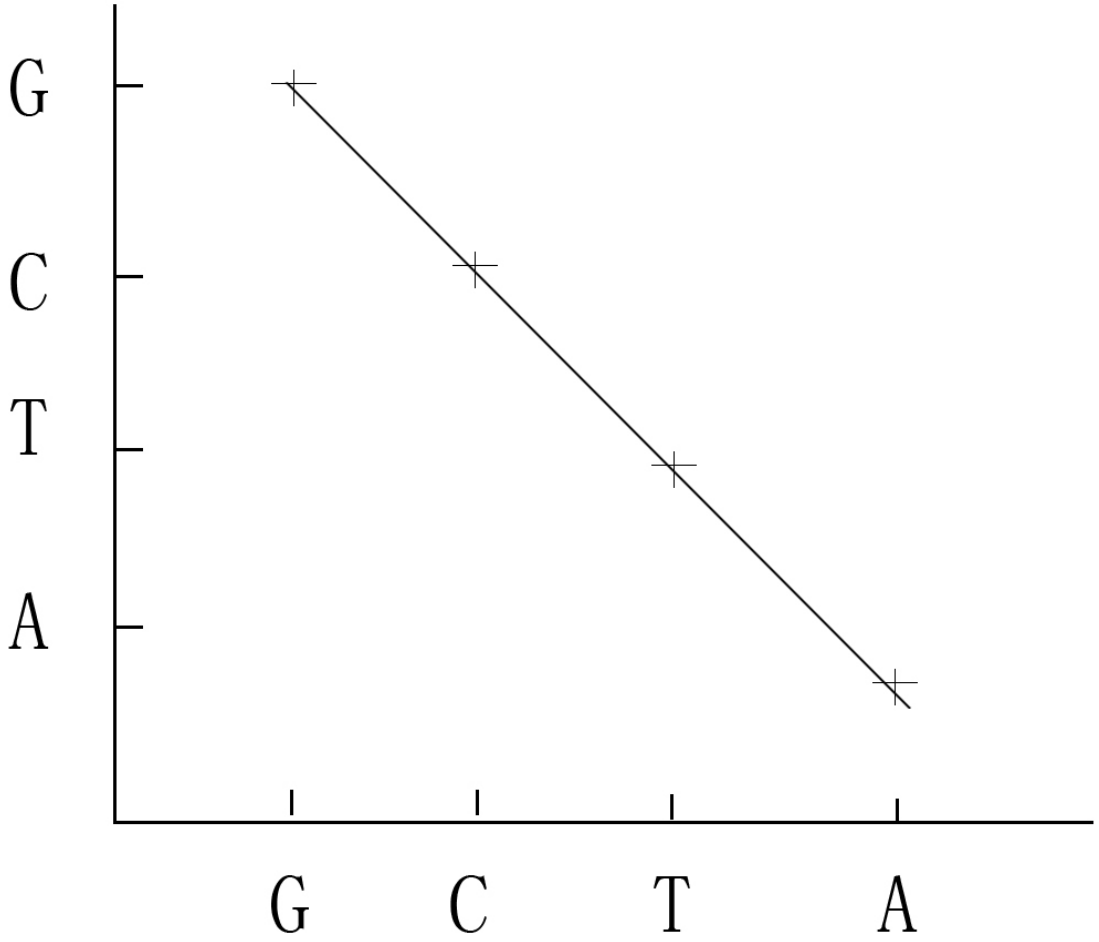

```
G C T A
| | | |
G C T A
```

两条序列有一条共同的子序列




```
A G C T A C
  | | | |
T G C T A G
```

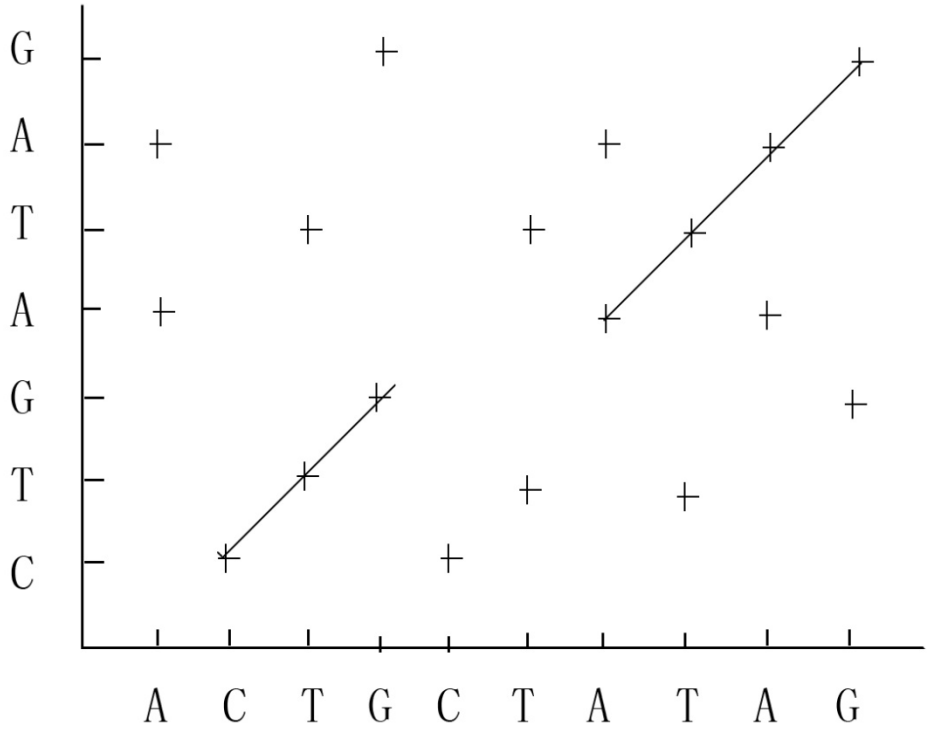
两条序列反向匹配



G C T A

A T C G

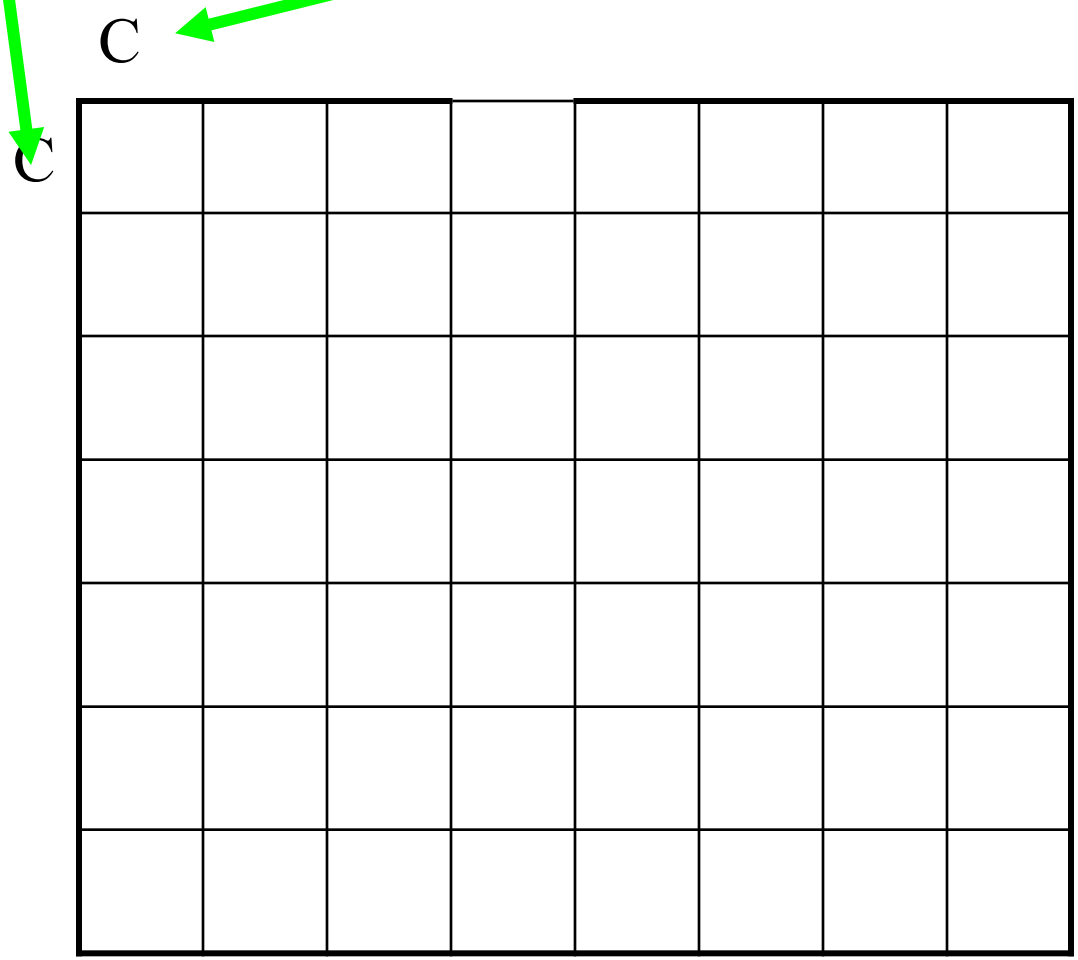
两条序列存在不连续的子序列




```
ACTGCTATAG
 |||   ||||
-CTG--ATAG
```

Graphic representation of an alignment

Sequence a: CTTAACT Sequence b: CGGATCAT

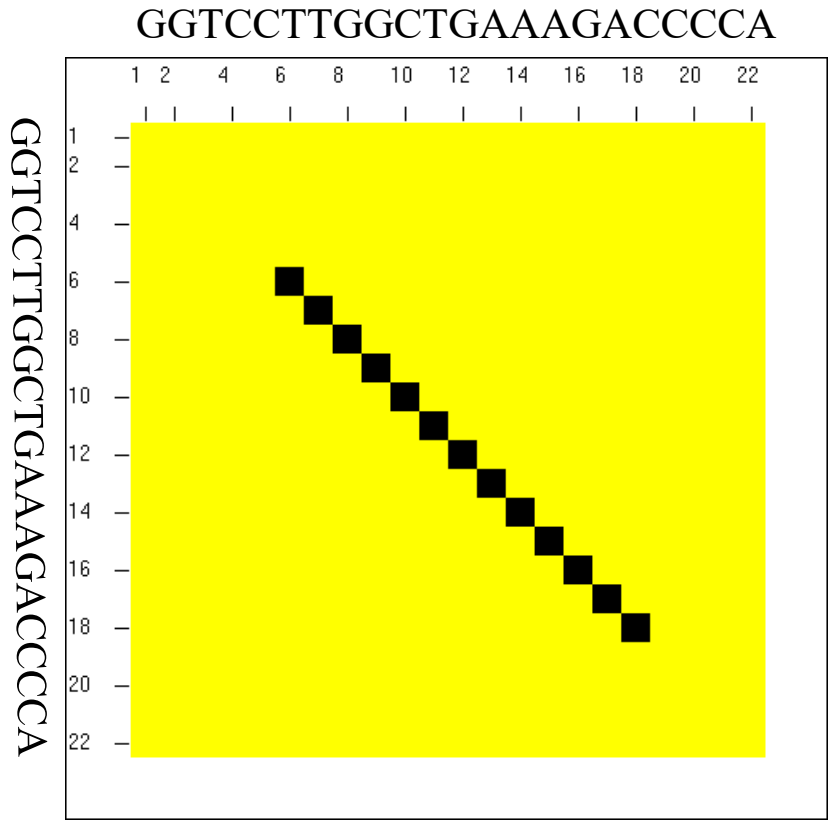


手动比一下
打点比一下

Patterns in Dot Plot

When two sequences are “identical”

Sequence : GGTCCTTGGCTGAAAG ACCCCA



Application of Dot Plot

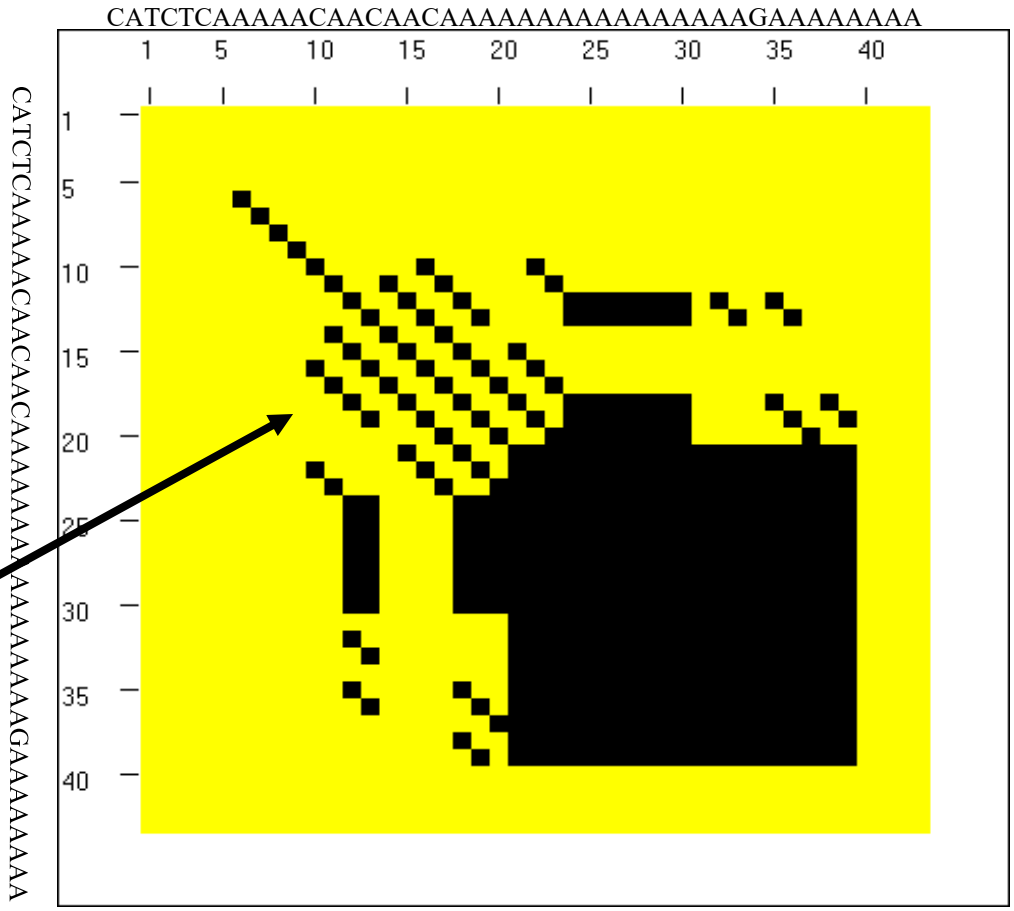
Using self comparison : Finding Repeats

Sequence used:

Human ALU sequence

CATCTCAAAAACAACAA CAAAAAAAAAAAAAAAAA GAAAAAAAA

- Omit main diagonal
- Clusters of diagonal lines show repeats in the sequence.



Notes:What are repeats?

Repeats:are stretches of repeated regions of residues in a sequence.

Importance of repeats:

In protein:

- Regulatory regions
- Binding sites

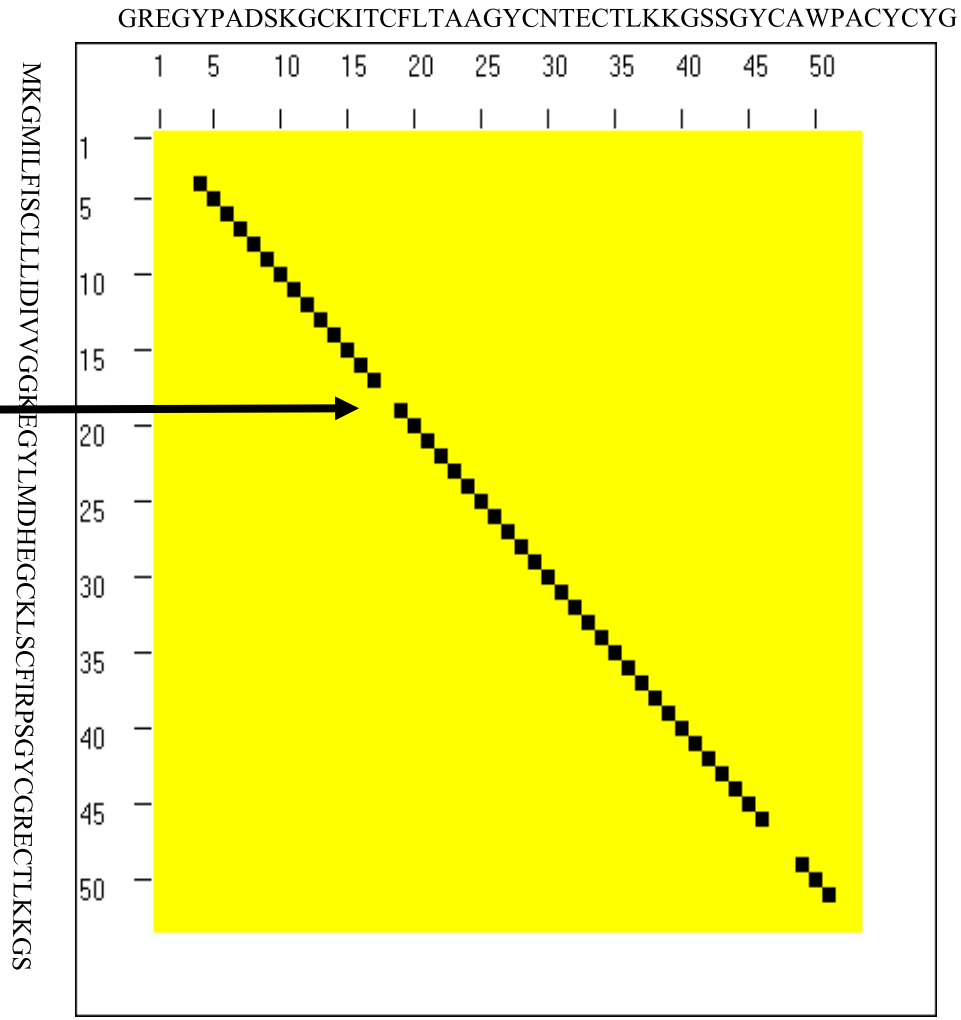
In DNA:

- Present in Transposons, chromosomal mutational hotspots, many genetic diseases related with repeats eg.Huntington.

Patterns in Dot Plot

When two sequences are similar :

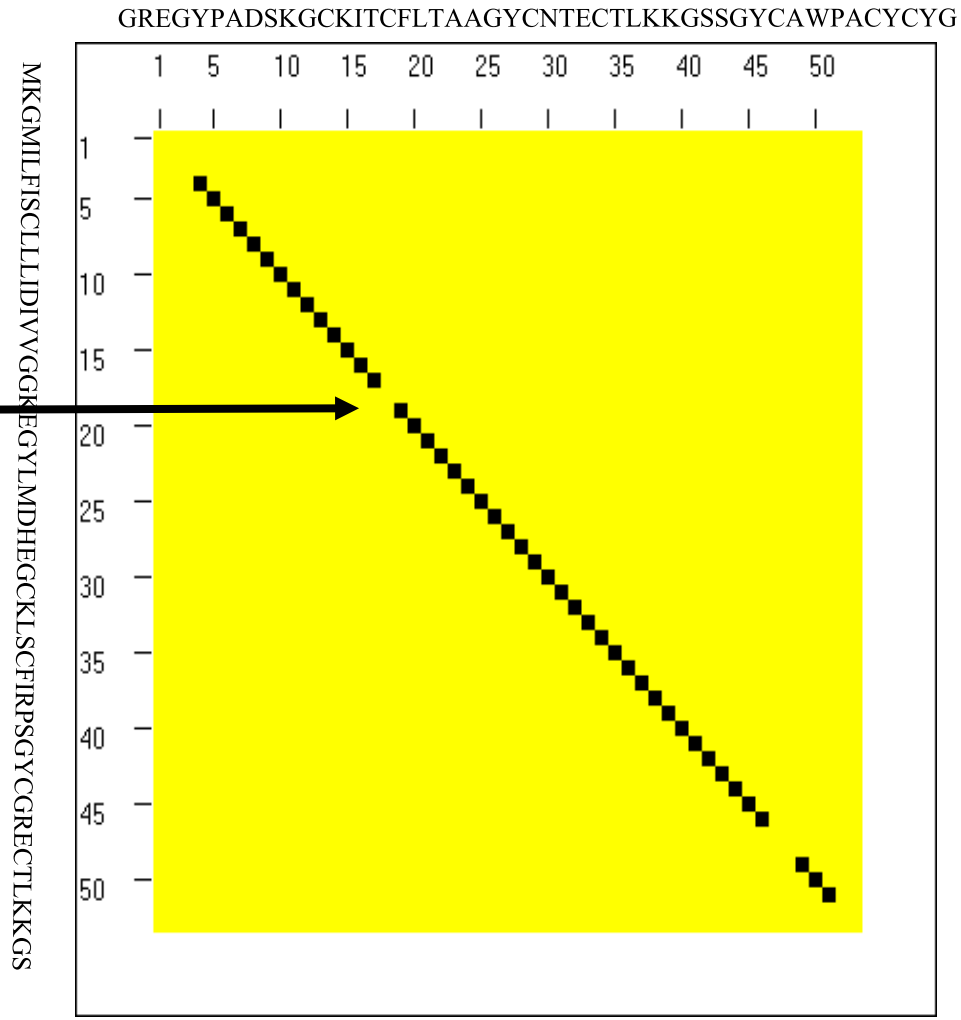
Broken ?



Patterns in Dot Plot

When two sequences are similar :

Broken diagonal, the interrupted region shows regions of mismatch

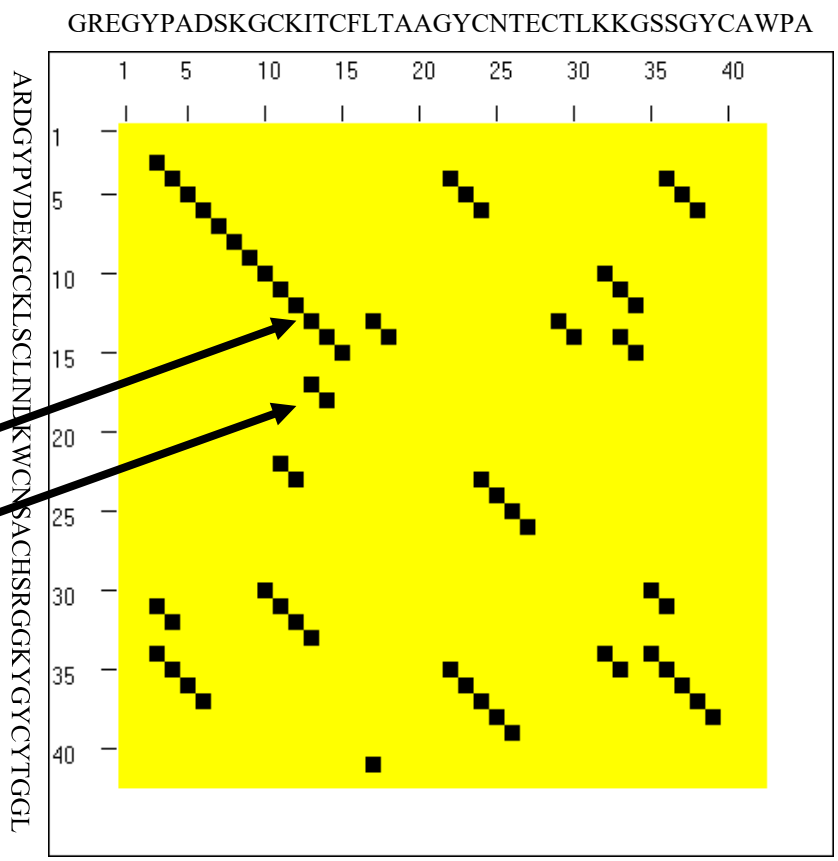


Patterns in Dot Plot

Two different, but related sequences

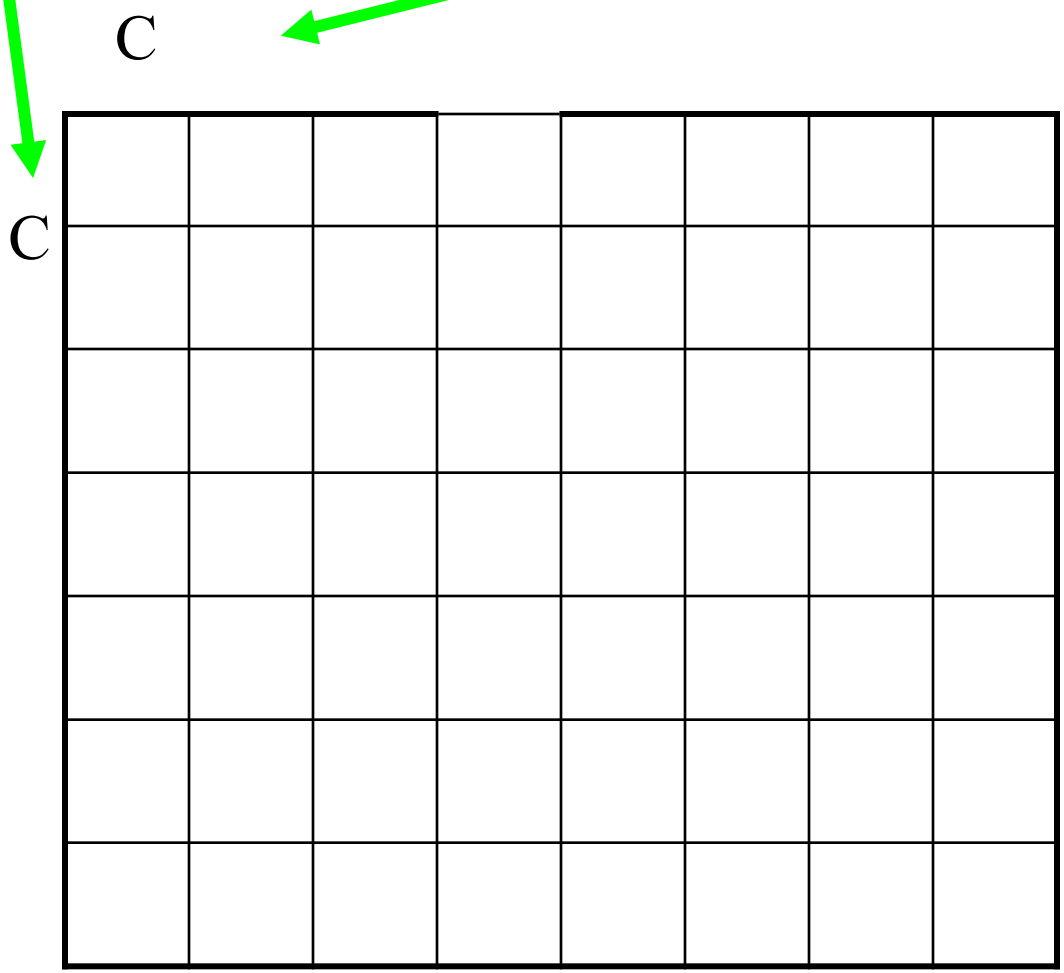
Broken diagonal clusters of dots parallel to the central diagonal.

Distance between the lines show no. of insertions done to get the alignment.



Graphic representation of an alignment

Sequence a: CTTAACT Sequence b: CGGATCAT

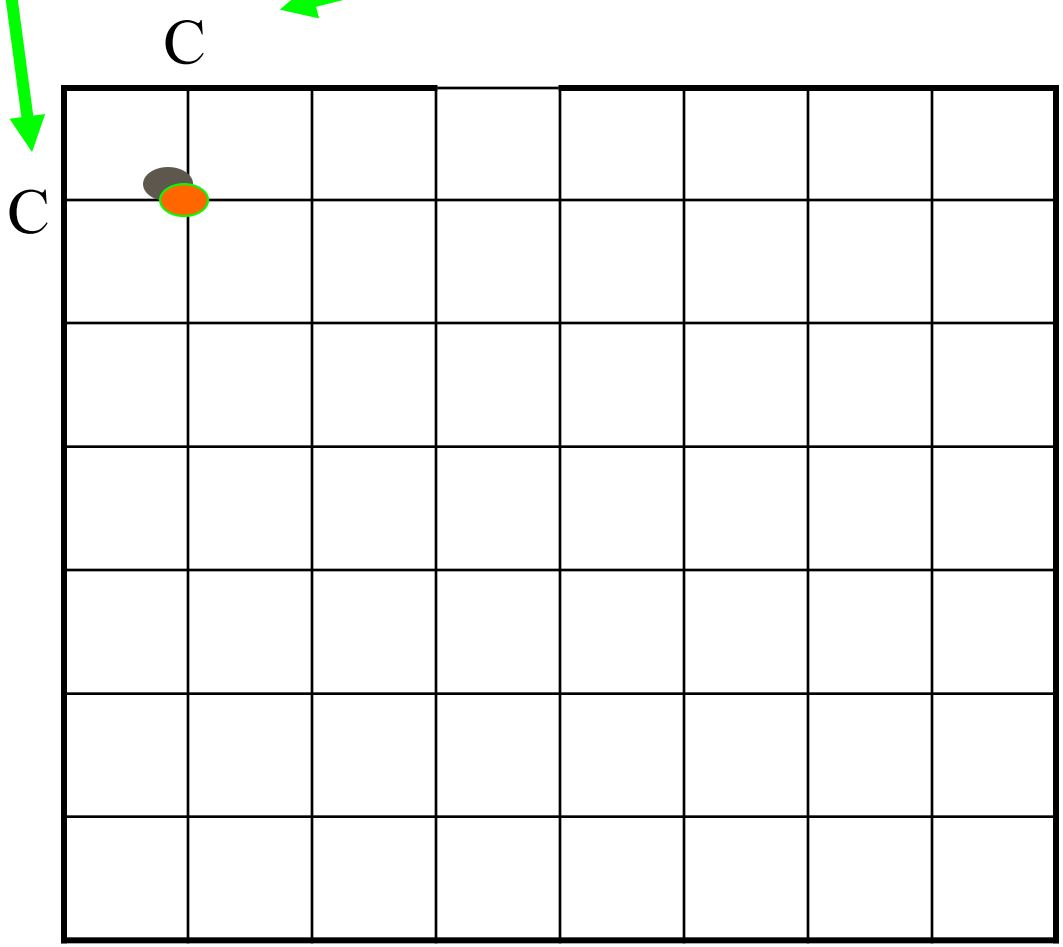


手动比一下
打点比一下

Answers?

Graphic representation of an alignment

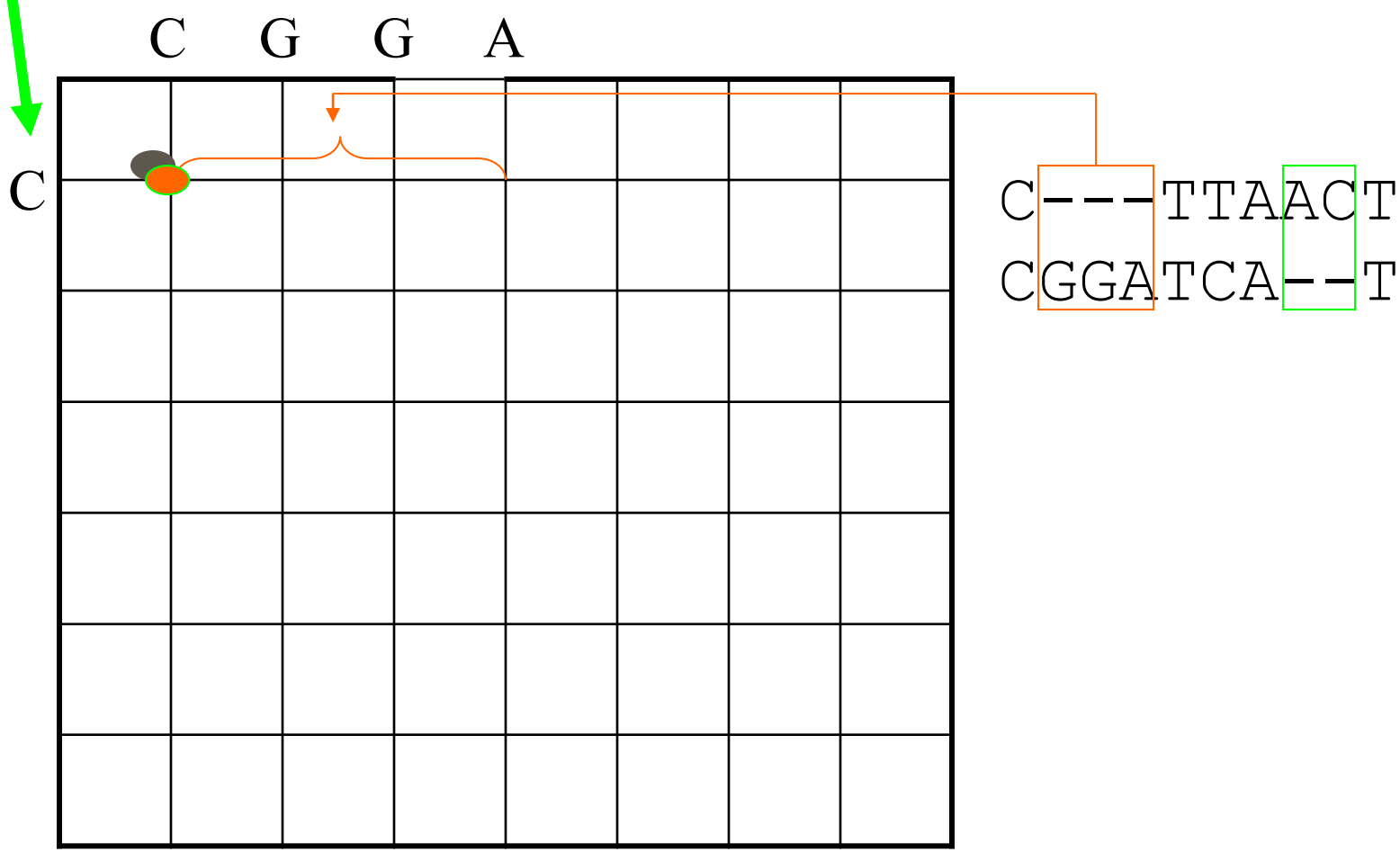
Sequence a: CTTAACT Sequence b: CGGATCAT




```
C---TTAACT
CGGATCA--T
```

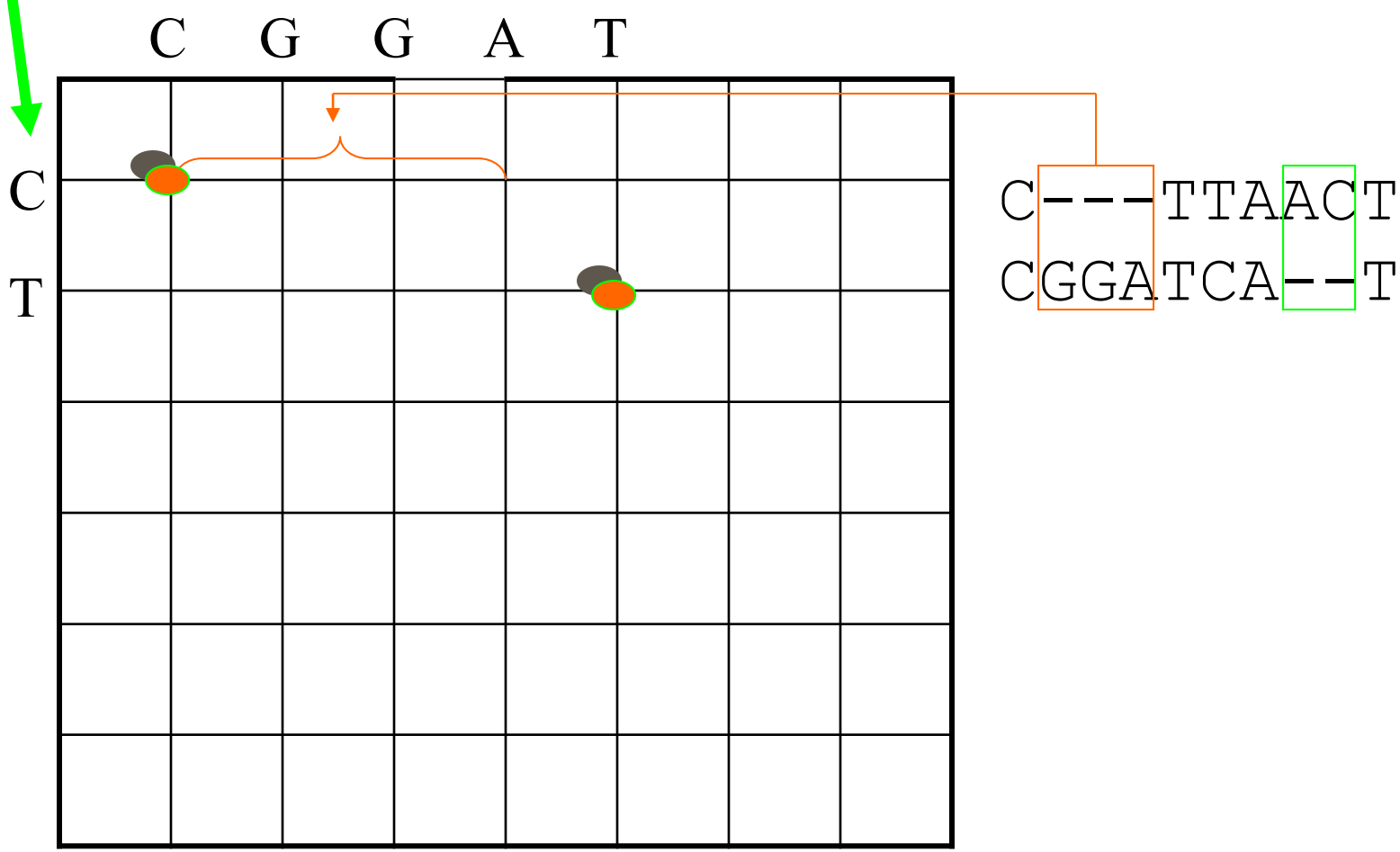
Graphic representation of an alignment

Sequence a: CTTAACT Sequence b: CGGATCAT



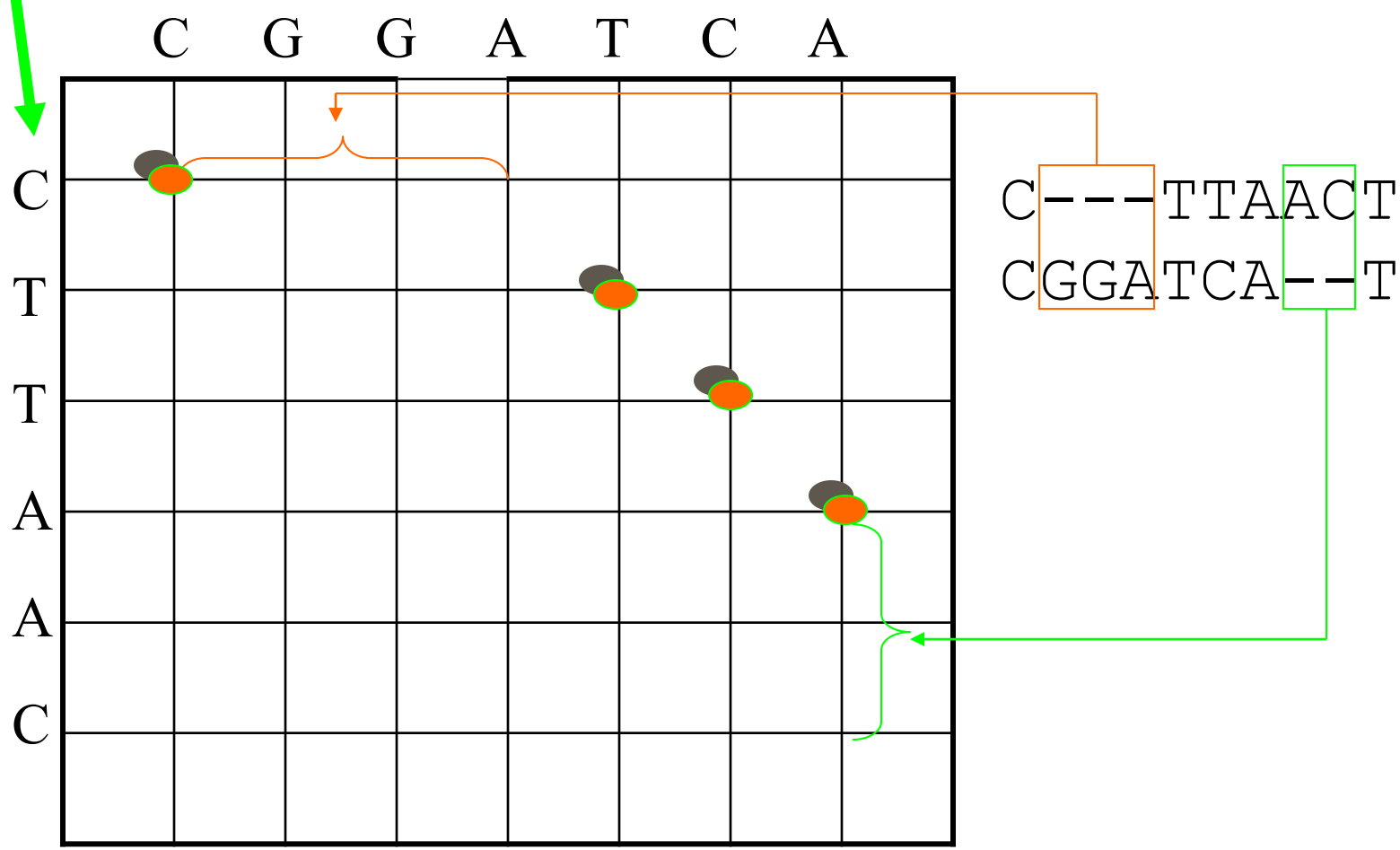
Graphic representation of an alignment

Sequence a: CTTAACT Sequence b: CGGATCAT



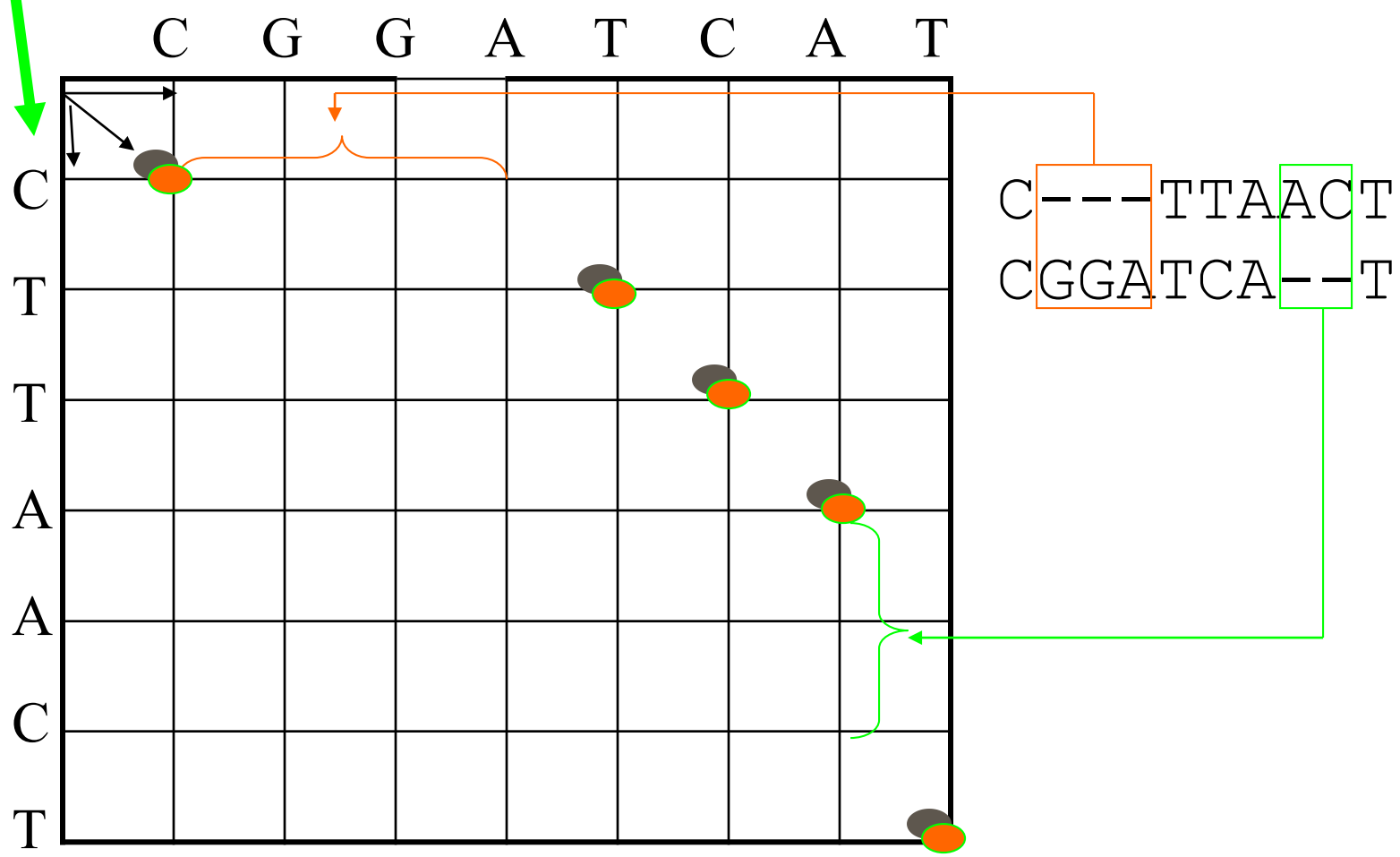
Graphic representation of an alignment

Sequence a: CTTAACT Sequence b: CGGATCAT



Graphic representation of an alignment

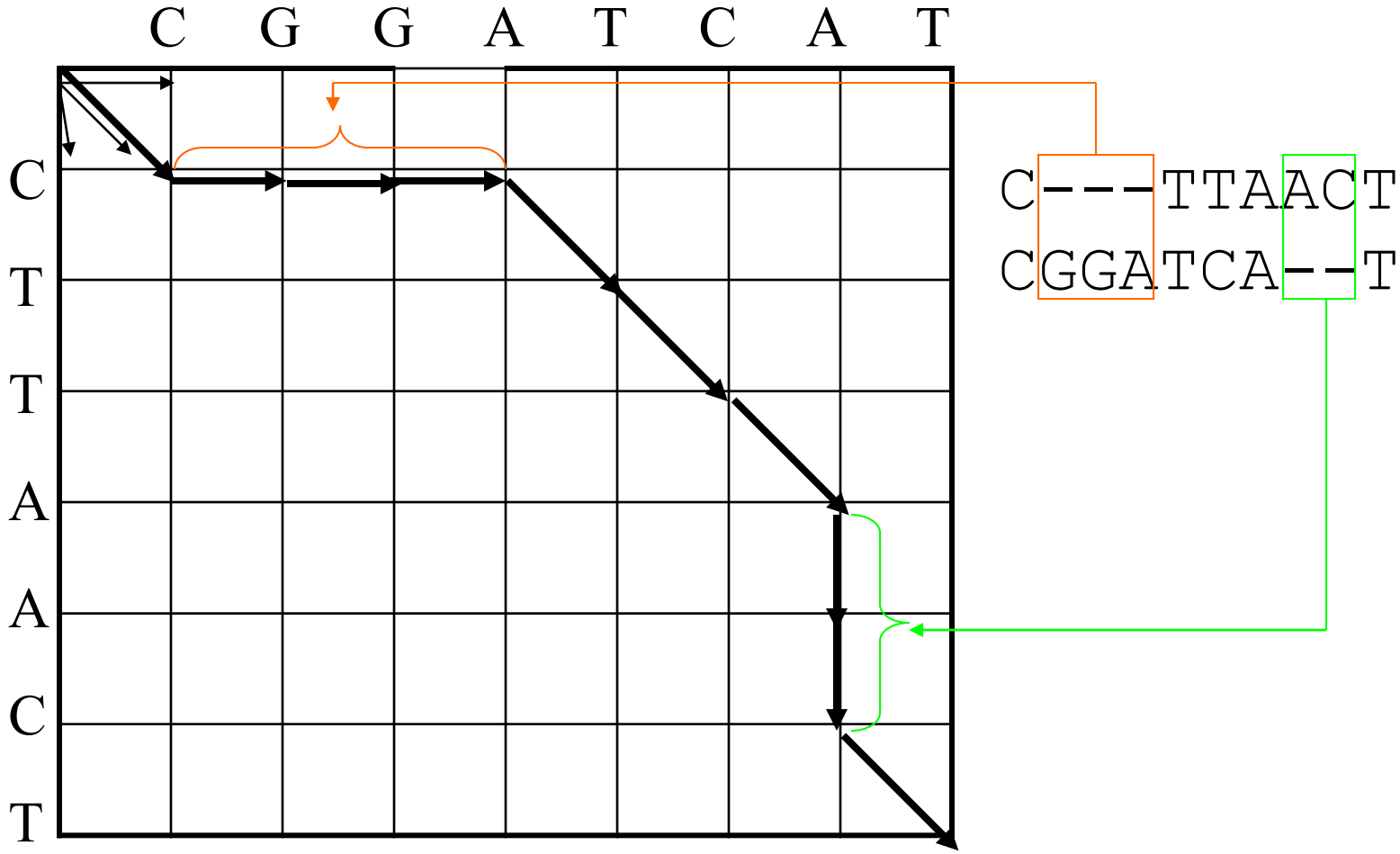
Sequence a: CTTAACT Sequence b: CGGATCAT



Pathway of an alignment

Sequence a: CTTAACT

Sequence b: CGGATCAT



Graphic representation of an alignment

Sequence a: CTTAACT Sequence b: CGGATCAT

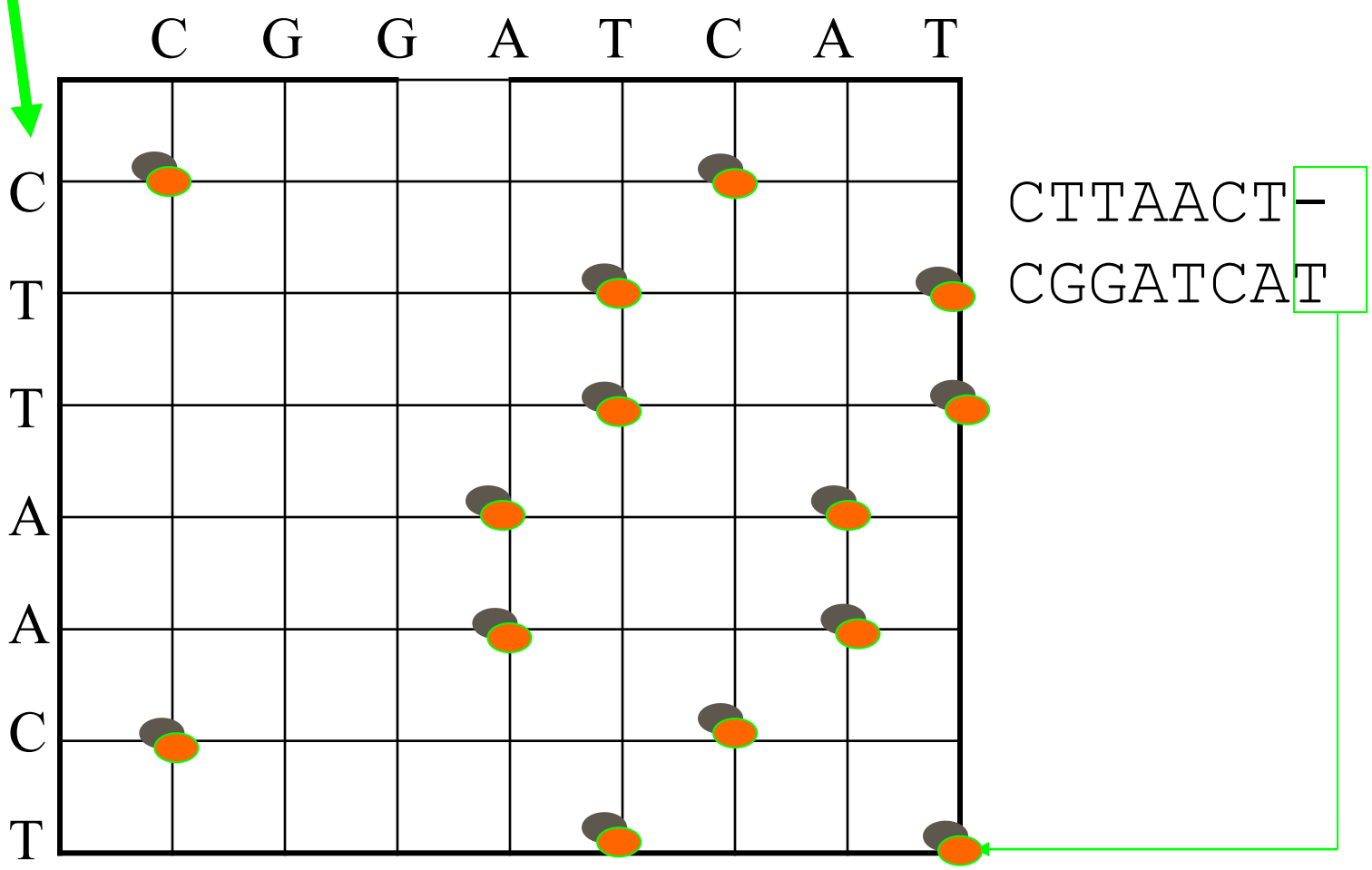
	C	G	G	A	T	C	A	T
C								
T								
T								
A								
A								
C								
T								

每人画两条路径

Answers?

Graphic representation of an alignment

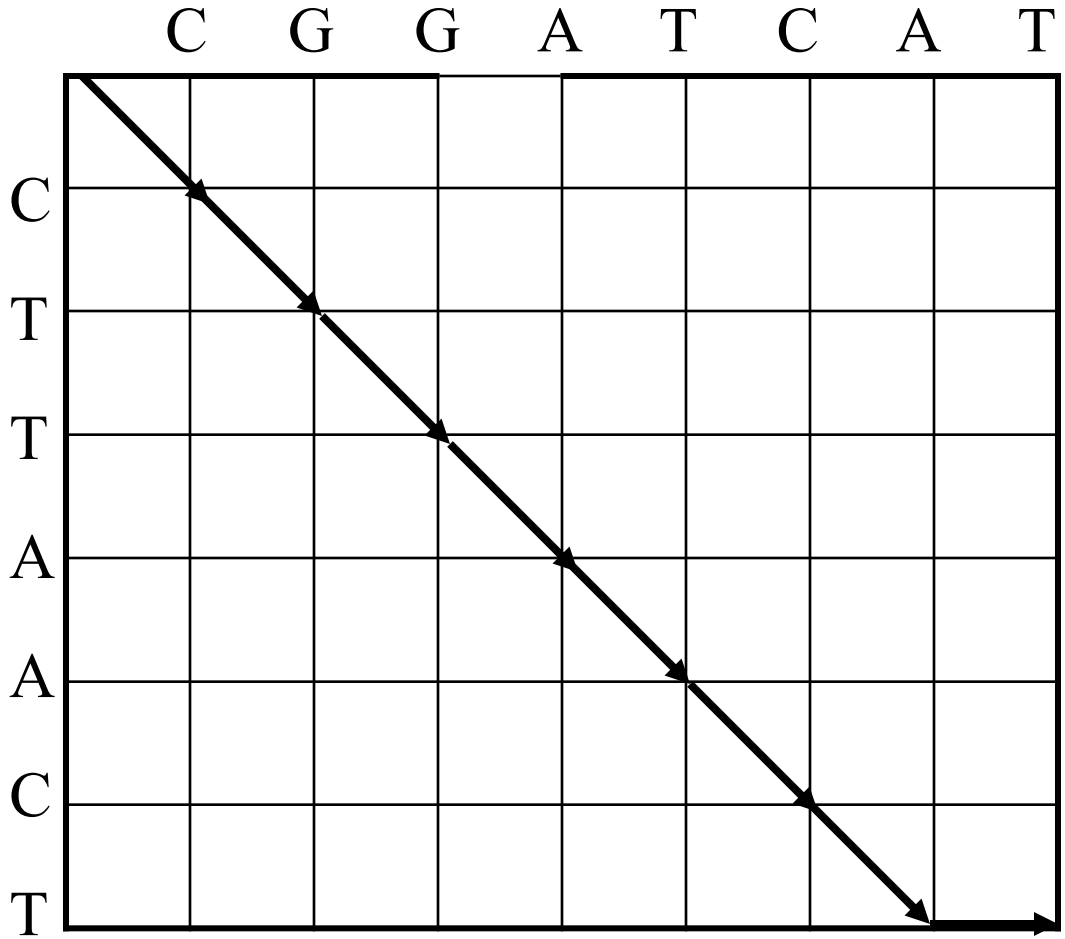
Sequence a: CTTAACT Sequence b: CGGATCAT



Pathway of an alignment

Sequence a: CTTAACT

Sequence b: CGGATCAT

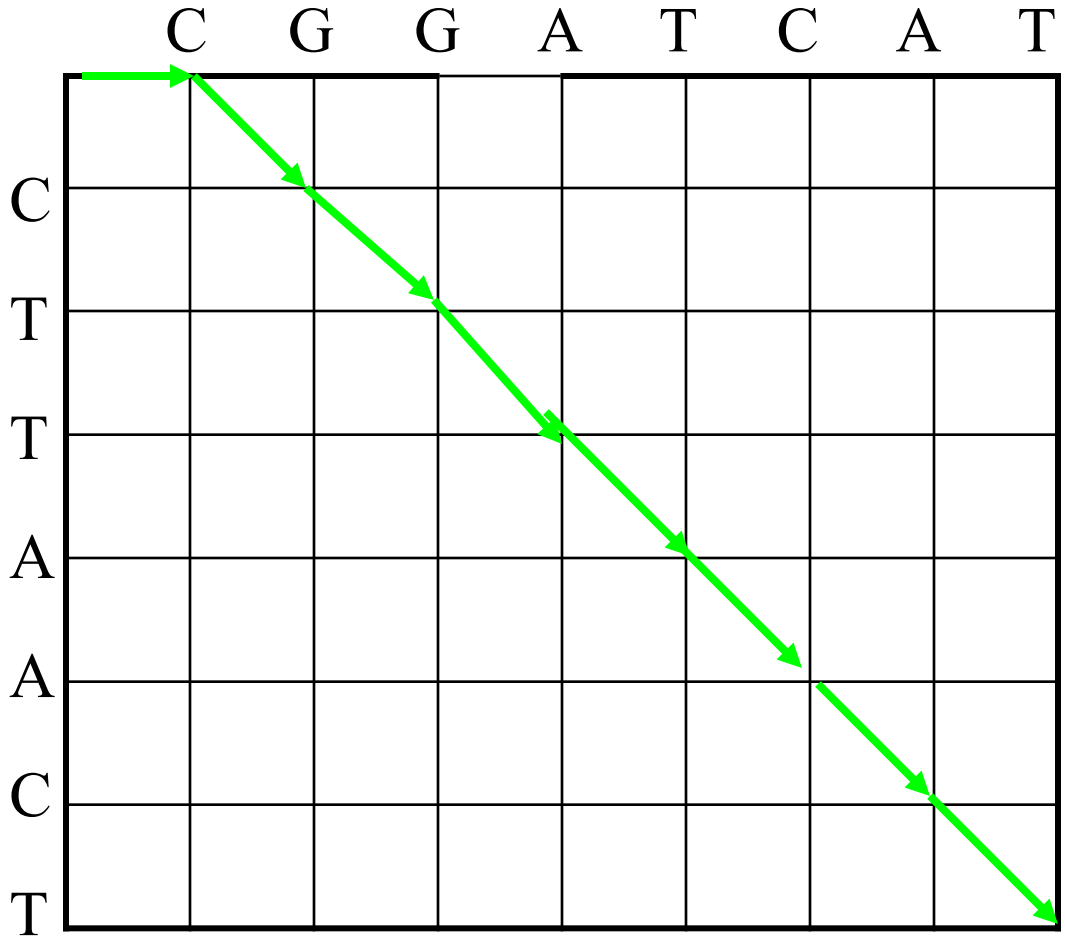



```
CTTAACT-
CGGATCAT
```

Use of graph to generate alignments

Sequence a: CTTAACT

Sequence b: CGGATCAT

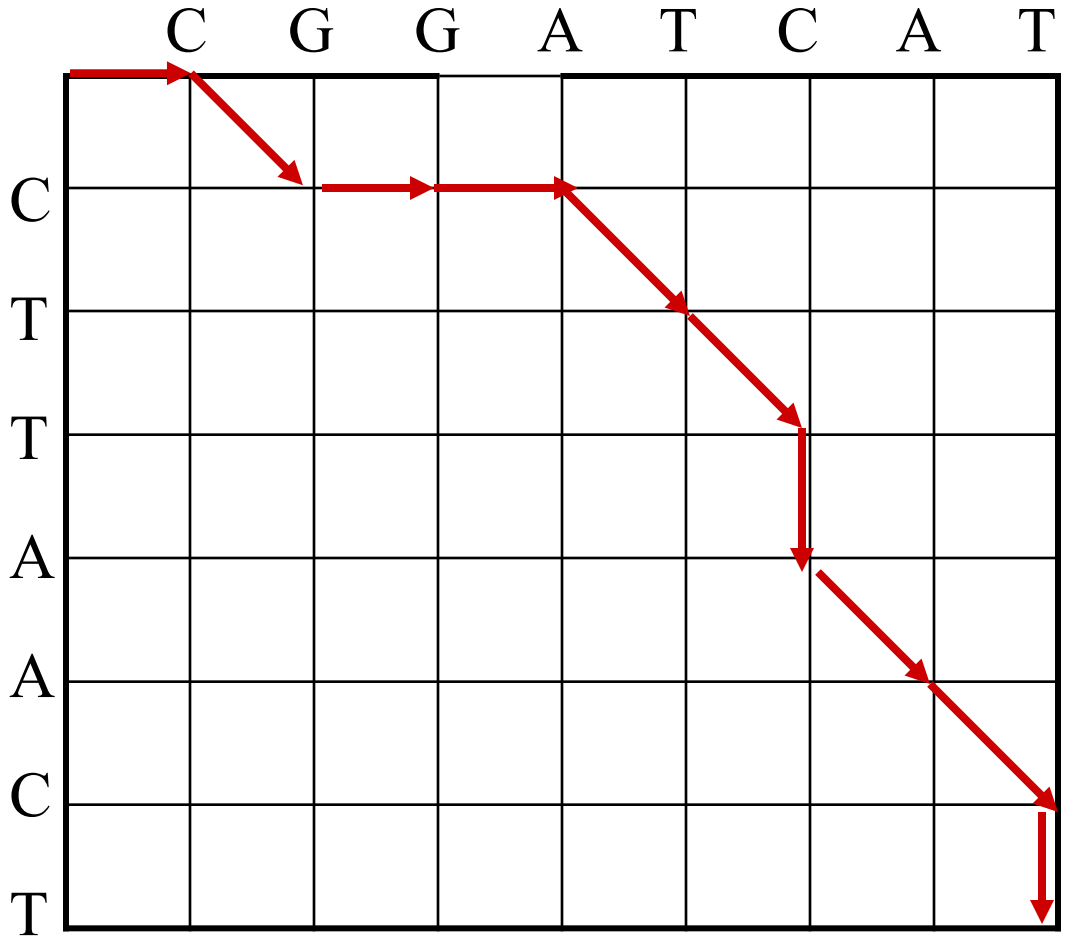


- CTTAACT
CGGATCAT

Use of graph to generate alignments

Sequence a: CTTAACT

Sequence b: CGGATCAT

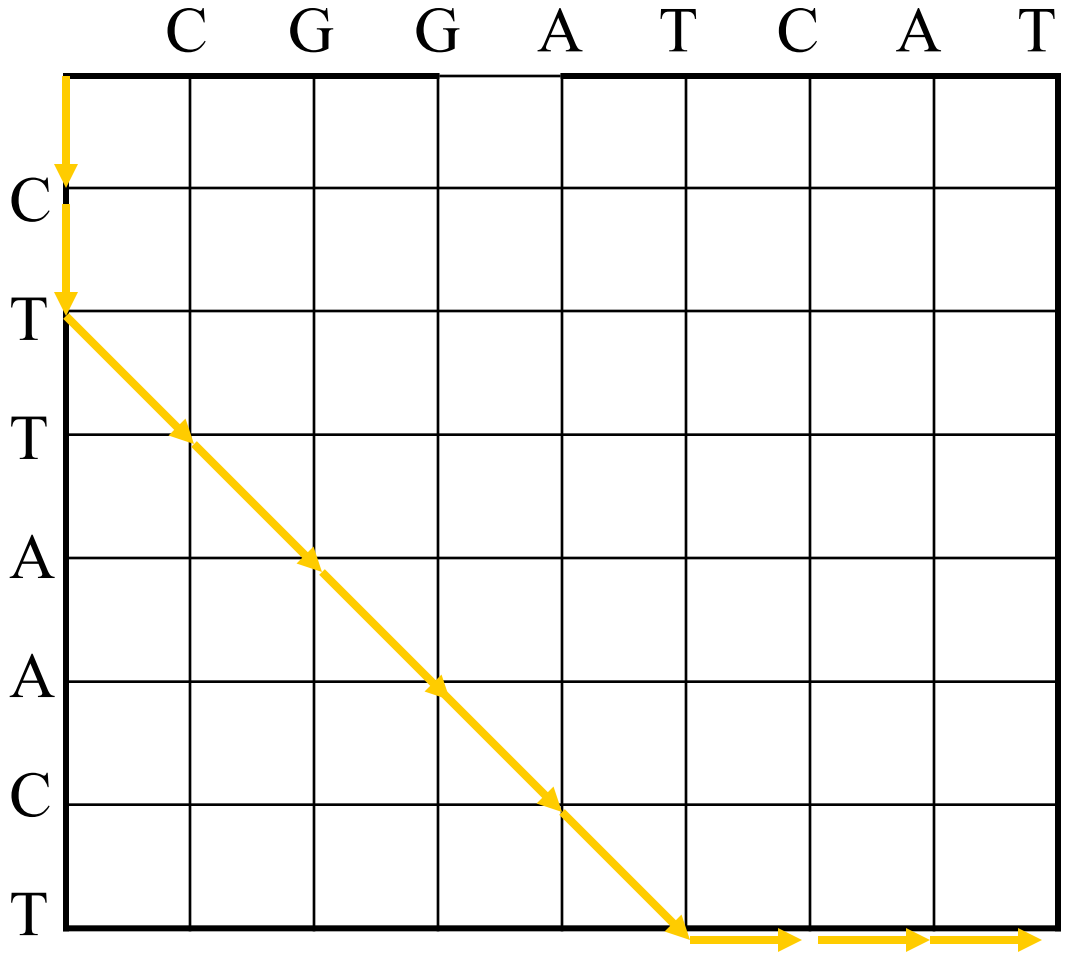


- C - - TTAACT
CGGATC - AT -

Use of graph to generate alignments

Sequence a: CTTAACT

Sequence b: CGGATCAT



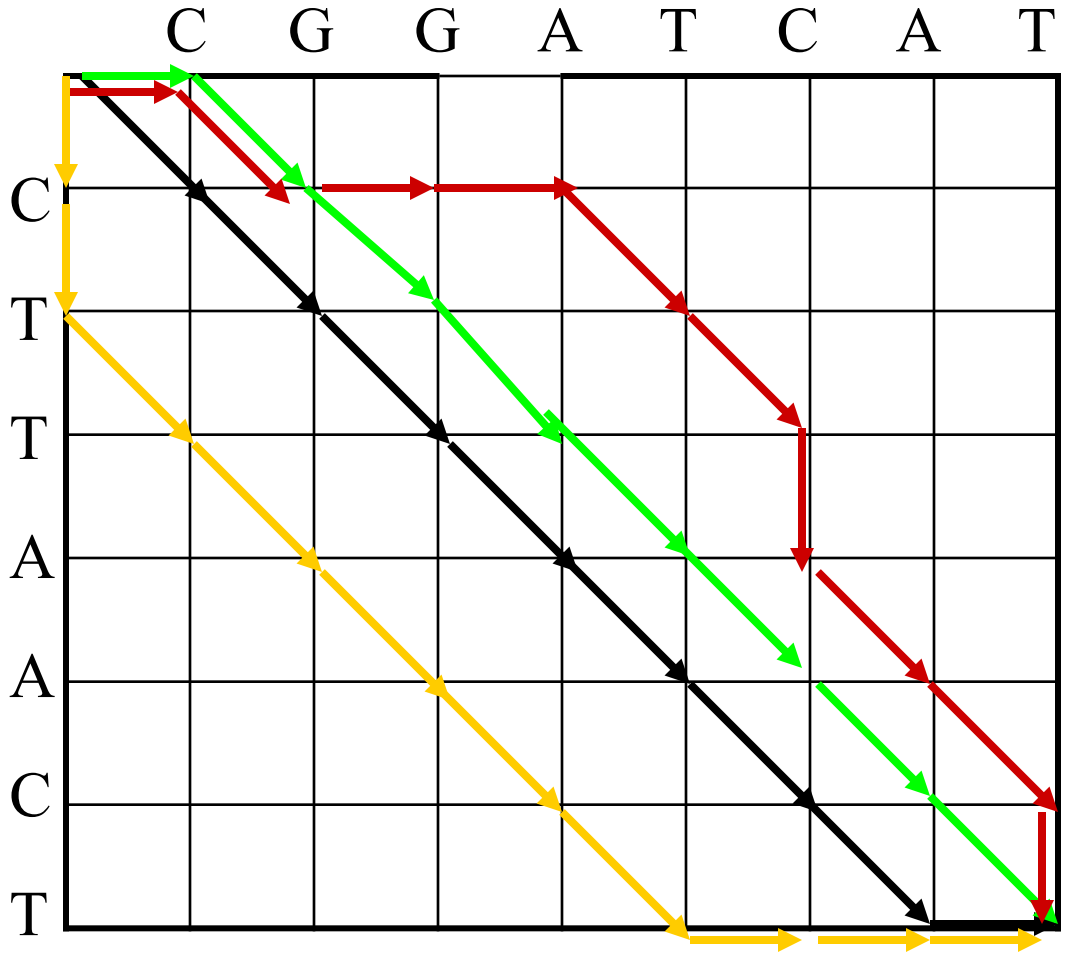
CTTAACT - - -

- - CGGATCAT

Which pathway is better?

Sequence a: CTTAACT

Sequence b: CGGATCAT



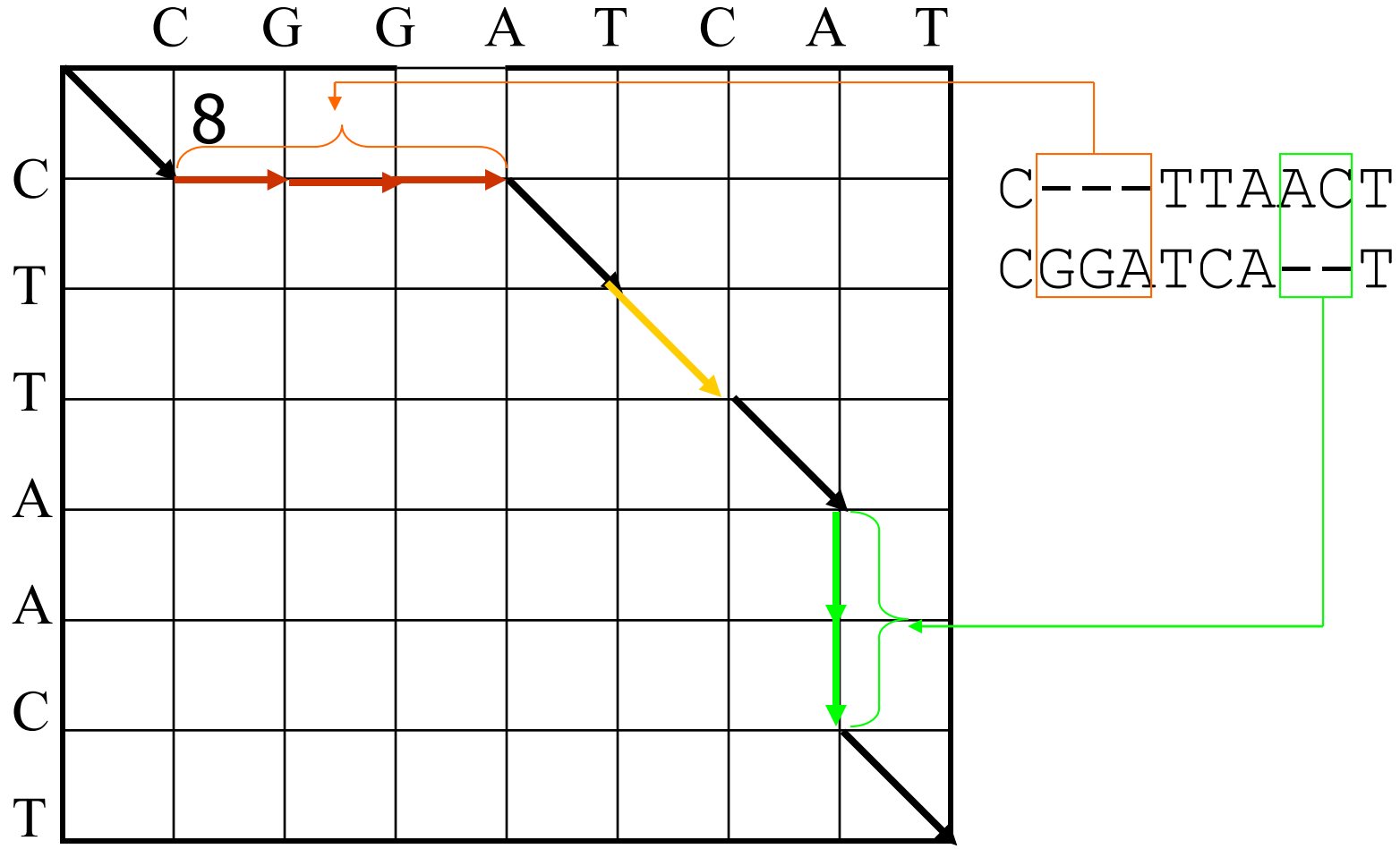
Multiple pathways

Each with a unique scoring function

Alignment Score

Sequence a: CTTAACT

Sequence b: CGGATCAT

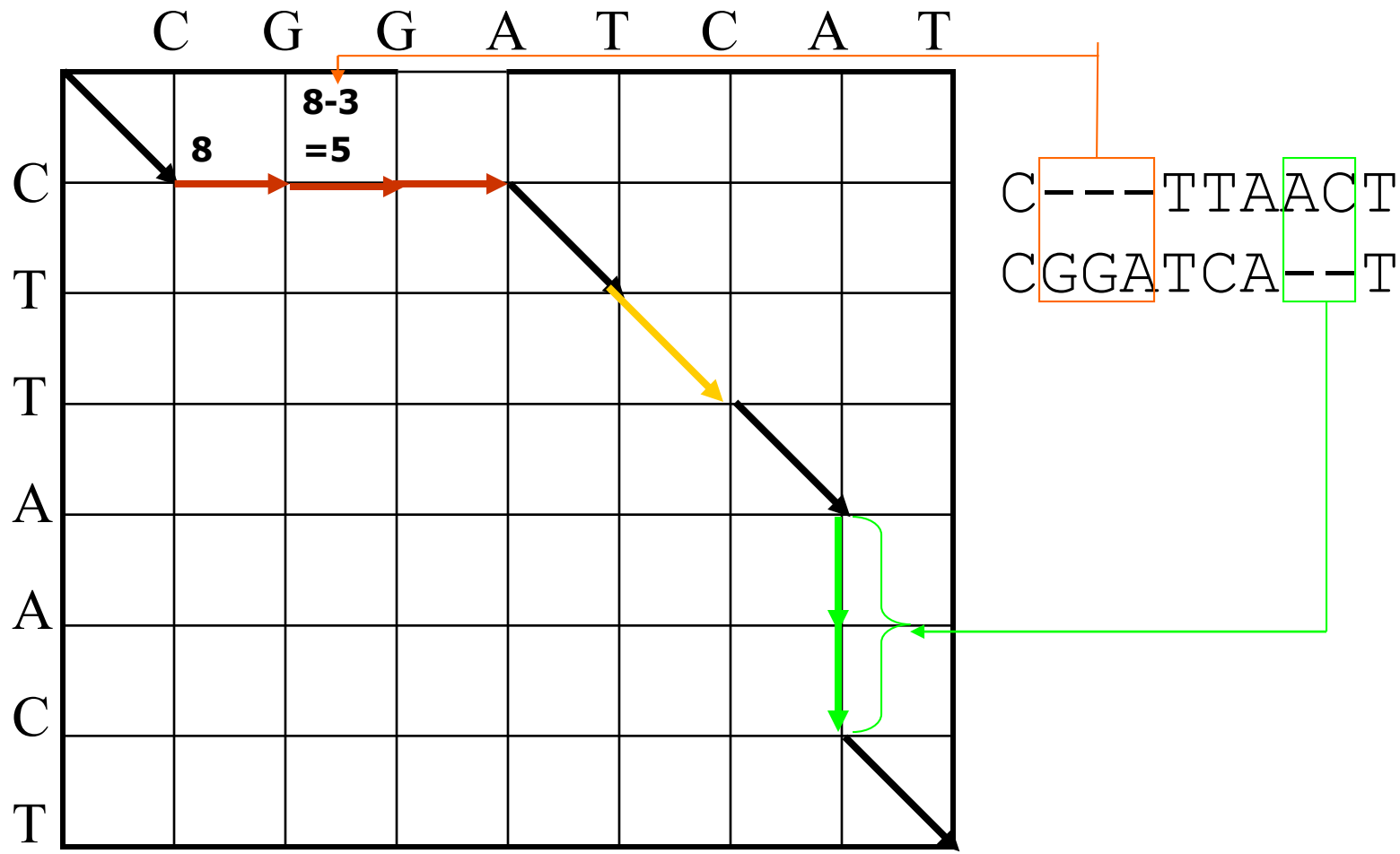


Alignment Score

Match: 8
Gap open:-3
Gap ext: -3
Mismatch: -3

Sequence a: CTTAACT

Sequence b: CGGATCAT

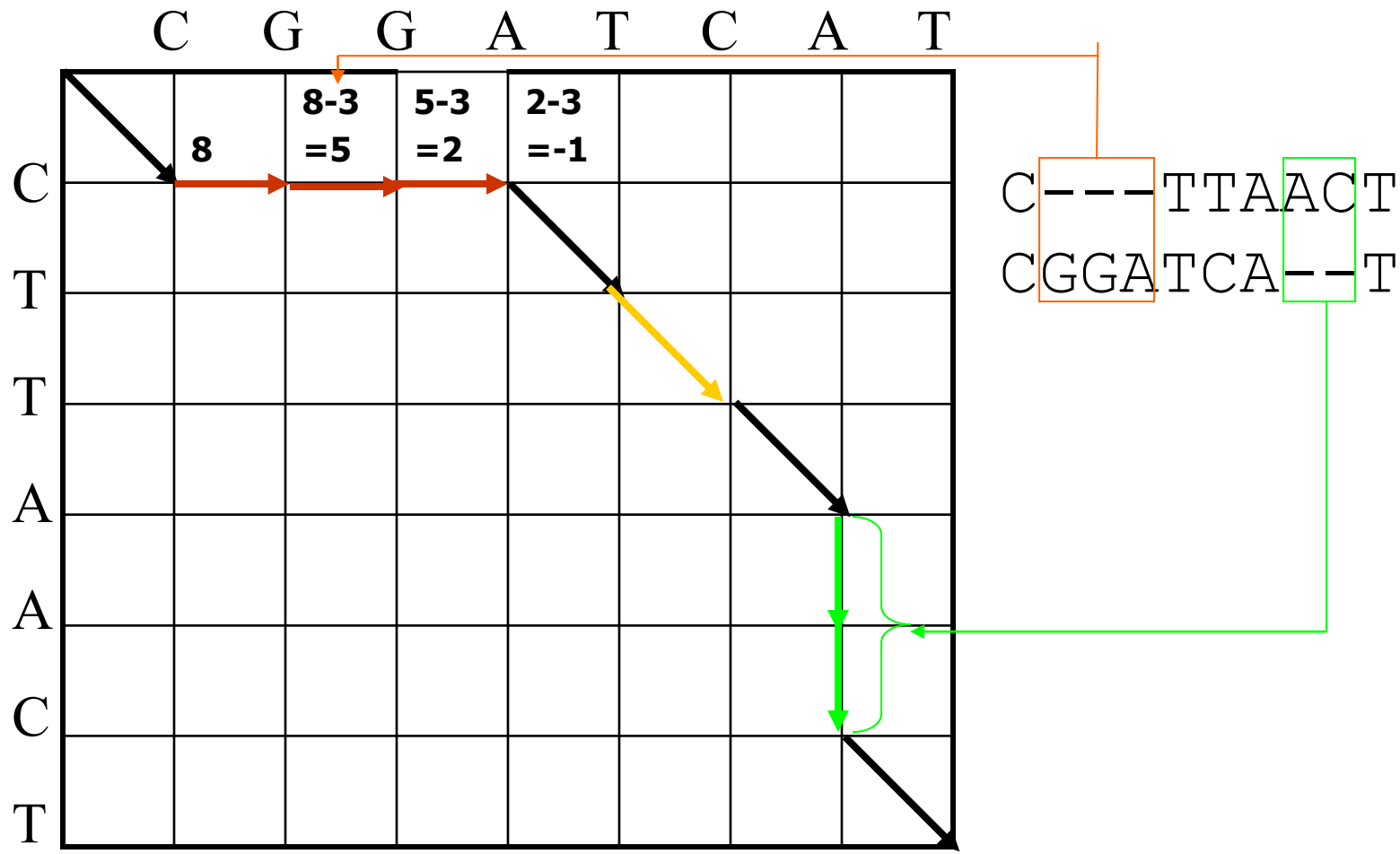


Alignment Score

Match: 8
Gap open:-3
Gap ext: -3
Mismatch: -3

Sequence a: CTTAACT

Sequence b: CGGATCAT

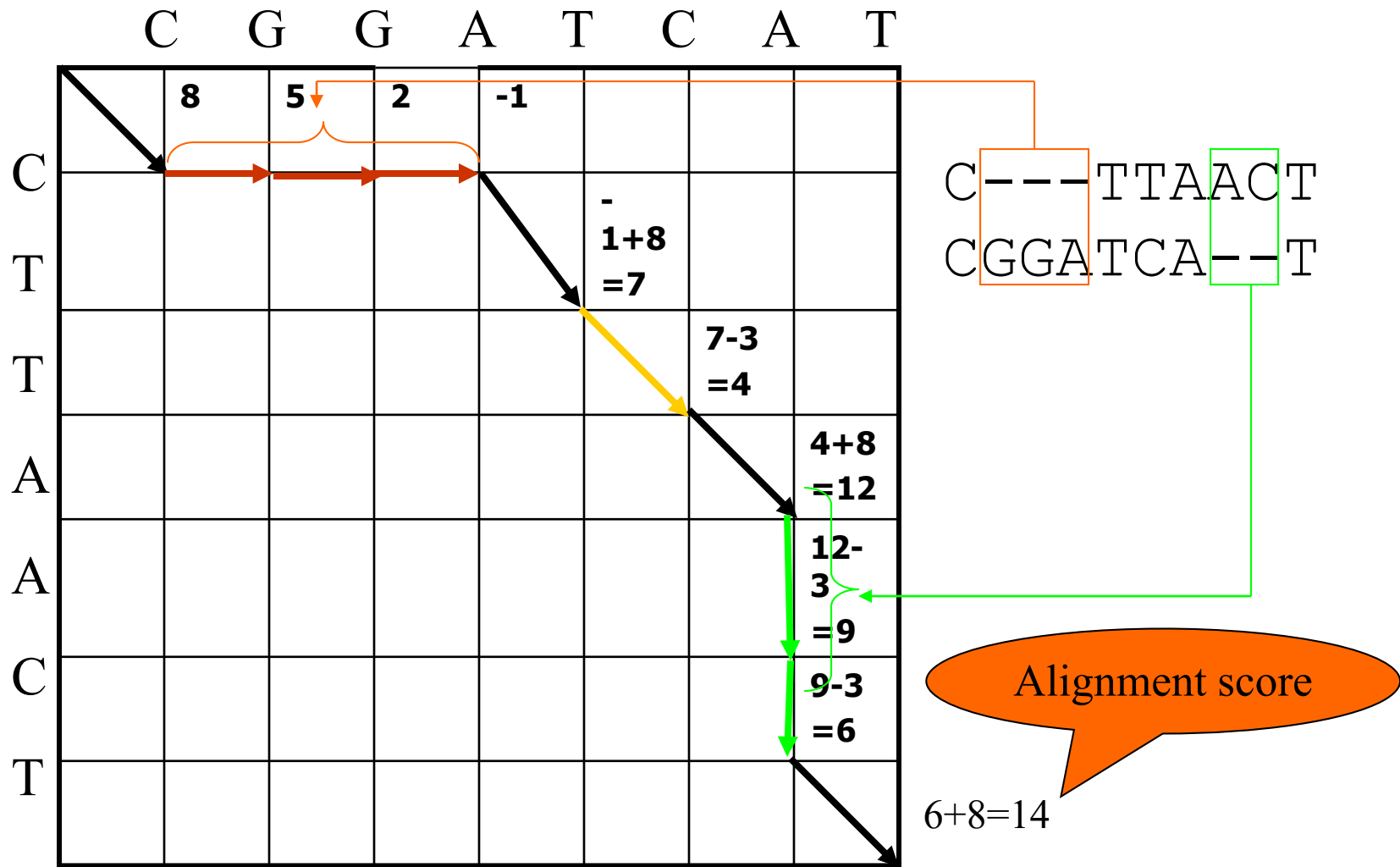


Alignment Score

Match: 8
Gap open:-3
Gap ext: -3
Mismatch: -3

Sequence a: CTTAACT

Sequence b: CGGATCAT



全局比对 vs. 局部比对

全局比对：在整个序列上达到尽可能多的字符匹配

- 序列在全长上有比较高的相似度
- 比对的序列长度基本接近
- 比对中允许插入空格

```
ACTCGGCCCC GCGCTCACTG C
| | | | | |  | | | | | | |  | |  |
ACTCGGAC --- GCGCTCAGTG C
```

局部比对：仅保留最高的得分区域以达到最佳的匹配

- 序列在全长上不一定相似，但是在某些区域有很高的相似度
- 允许序列长度差别较大
- 比对中尽可能少插入空格

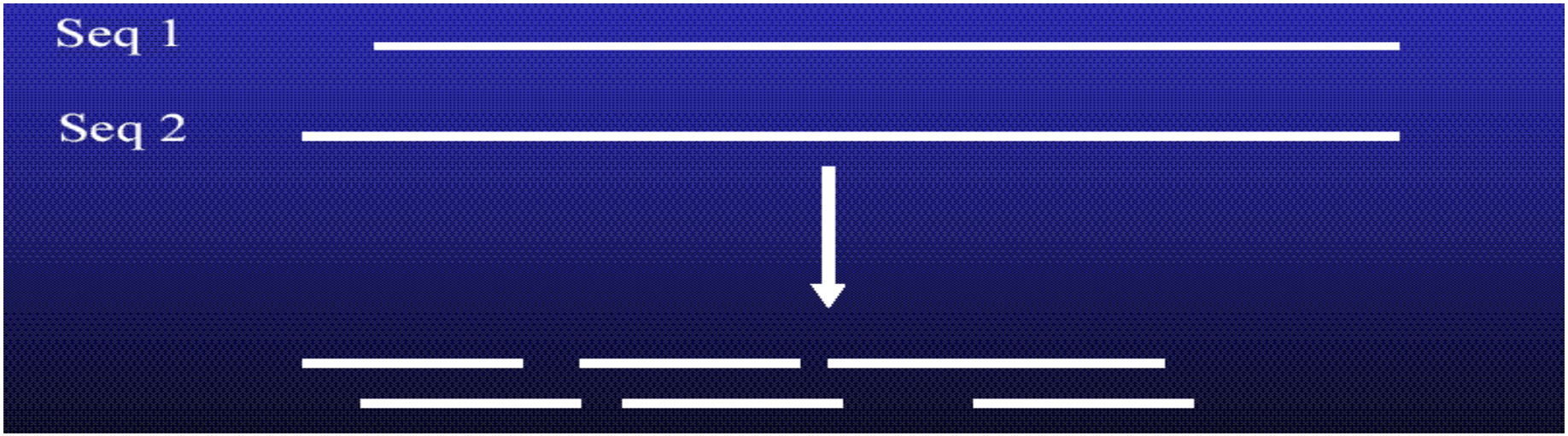
```
- - - -AGCT- - - -
ATGCAGCTGTCT
```

Two models of alignment: Local and Global alignments

Global alignment:

Looks for similarity across full extent of sequences

Needleman-Wunsch algorithm based on this model.



全局 Needleman-Wunsch

a、b是两条DNA或者蛋白质序列，长度分别是m和n

S(i, j)是a[1, i]和b[1, j]的最大相似性得分

$w(a_i, b_j)$为a_i和b_j按照替换记分矩阵计算的得分

gap为插入删除的罚分

初始化S(i, 0)=0 S(0, j)=0

$$S(i, j)=\max\begin{cases} S(i-1, j-1)+w(a_i, b_j) & \text{匹配或错配} \\ S(i-1, j)+gap & \text{插入} \\ S(i, j-1)+gap & \text{缺失} \end{cases}$$

	gap	a1	a2	a3	a4
gap	0	1 gap	2 gaps	3 gaps	4 gaps
b1	1 gap	s11	s21		
b2	2 gaps	s12	s22		
b3	3 gaps				
b4	4 gaps				

$$S = \begin{array}{c|ccccccccc} & - & A & C & A & C & A & C & T & A \\ \hline - & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ A & 0 & 2 & 1 & 2 & 1 & 2 & 1 & 0 & 2 \\ G & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 1 \\ C & 0 & 0 & 3 & 2 & 3 & 2 & 3 & 2 & 1 \\ A & 0 & 2 & 2 & 5 & 4 & 5 & 4 & 3 & 4 \\ C & 0 & 1 & 4 & 4 & 7 & 6 & 7 & 6 & 5 \\ A & 0 & 2 & 3 & 6 & 6 & 9 & 8 & 7 & 8 \\ C & 0 & 1 & 4 & 5 & 8 & 8 & 11 & 10 & 9 \\ A & 0 & 2 & 3 & 6 & 7 & 10 & 10 & 10 & 12 \end{array}$$

A- CACACTA
AGCACAC- A

$$S = \begin{array}{c|ccccccccc}
 & - & A & C & A & C & A & C & T & A \\
\hline
- & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
A & 0 & 2 & 1 & 2 & 1 & 2 & 1 & 0 & 2 \\
G & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 1 \\
C & 0 & 0 & 3 & 2 & 3 & 2 & 3 & 2 & 1 \\
A & 0 & 2 & 2 & 5 & 4 & 5 & 4 & 3 & 4 \\
C & 0 & 1 & 4 & 4 & 7 & 6 & 7 & 6 & 5 \\
A & 0 & 2 & 3 & 6 & 6 & 9 & 8 & 7 & 8 \\
C & 0 & 1 & 4 & 5 & 8 & 8 & 11 & 10 & 9 \\
A & 0 & 2 & 3 & 6 & 7 & 10 & 10 & 10 & 12
\end{array}$$

全局比对的统计学显著性

典型方法：

- 将两条待比较的序列分别随机打乱
- 使用相同的程序与打分函数（或打分矩阵）进行比对
- 计算这些随机序列的相似性得分
- 重复这一过程（50～100次）用μ和δ分别表示其平均值与标准差。
- 设原来两条序列的比对得分为x，利用下式计算大于或等于x的比对得分概率： $z = (x - \mu)/\delta$
- 根据z值判断两个序列相似得分的显著性，当z值是3.1、4.3、5.2时，x出现的概率为10^{-3}、10^{-5}、10^{-7}
- Z > 5，同源；
- Z < 3，不同源；
- Z = 3~5, 可能同源

经验法则（针对蛋白质序列）：

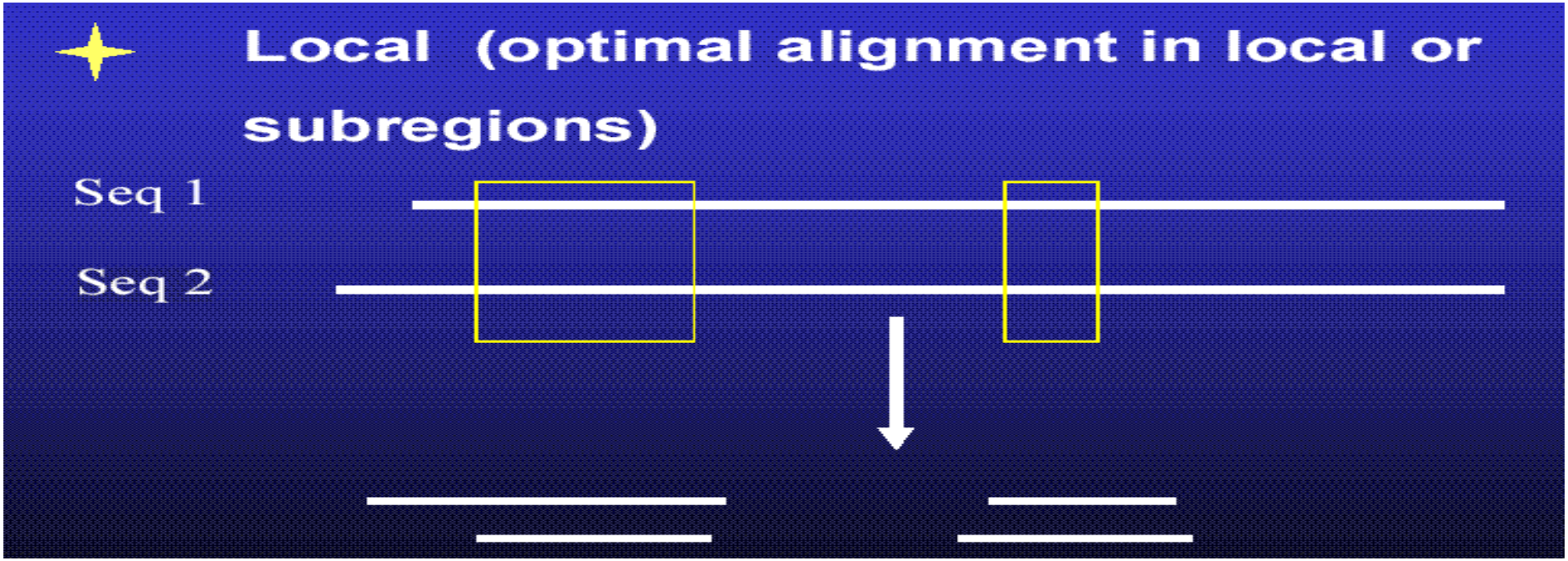
- 如果两个序列的长度都大于100，在适当地加入空位之后，它们配对的相同率达到25%以上，则两个序列相关；
- 如果配对的相同率小于15%，则不管两个序列的长度如何，它们都不可能相关；
- 如果两个序列的相同率在15%~25%之间，它们可能是相关的。

Local alignment

Looks for regions of similarity in parts of the sequences only.

Smith-Waterman algorithm based on this model.

Softwares : BLAST, FASTA.



局部 Smith-Waterman

如果当前比对分数小于0，赋值为0，比对从目前位置重新开始。

回溯的时候不是从最后开始，而是从最大的分数开始

$$S(i, j)=\max \begin{cases} 0 \\ S(i-1, j-1)+w(a_i, b_j) & \text{匹配或错配} \\ S(i-1, j)+gap & \text{插入} \\ S(i, j-1) & \text{缺失不罚分} \end{cases}$$

----AGCT----

ATGCAGCTGCTT

Why two different models?

- Global alignment

  High degree of Homology

  Good for modelling

- Local Alignment

  Localised Similarity (conserved regions with structural , functional importance, Repeats, Domains)

不同比对算法

算法	准确度 （敏感度，特异度）	速度
详尽的（exhaustive）： 动态规划	高	非常慢
启发式的（heuristic）： FASTA BLAST	Not best Good enough	较快

What is BLAST?

Basic Local Alignment Search Tool (BLAST)

Method for Pairwise Alignment.

Is used to search for homologous sequences from a database (of nucleotide/protein sequence) for a given query sequence.

Modified version of FASTA

- Faster in generating output.

Sites for doing BLAST:

http://www.ncbi.nlm.nih.gov

How does it work?

The main task of any sequence comparison program is to test all possible mutual alignments of two sequence and see how good the match is:

query CCGAGCTTCTCATTGCTCTTCCTAACAGTG

1st database sequence CTCTTCTCATTGCTCTTCCTAACAGTGCTAATAGGCTAACGCTAATGGGCTTCT

This would actually be a very slow search process if implemented like this…

How does BLAST work?

BLAST achieves its speed through two strategies:

- **it takes a WORD based approach**

- **it pre-INDEXES database sequences**

BLAST: WORDS and INDEXING

Database of sequences

```
1 GACAAATCCAAACCCCTGAAGTTCTCCACCAGCAAAGCCA
2 TAAGCAAATTTAATTTTGTTTACATTTTC
3 GTTAAGACCTTCCCTGACATTTGCAGCAGTTTCAAATGTA
```

Numbered list of *all possible* 'words'

```
AAAAAAAA 00001
AAAAAAAC 00002
AAAAAAAG 00003
:
ACAAATCC 07967
ACAAATCC 07968
ACAAATCC 07979
:
GACAAATC 33568
GACAAATG 33569
:
TCCAAACC 64321
TCCAAACC 64322
:
```

Build a position index of all words in the database

sequence	position	word
1	1	33658
1	2	07967
1	3	16210
:		
3	15	33568
3	16	07967
:		

Analyse the Query Sequence

QUERY SEQUENCE

```
>query
AGACAAATCCAAACCCCTGAAGTTCTCCACCAGCAAAGCCA
```

Numbered list of *all possible* 'words'

```
AAAAAAAA 00001
AAAAAAAC 00002
AAAAAAAG 00003
:
ACAAATCC 07967
ACAAATCC 07968
ACAAATCC 07979
:
GACAAATC 33568
GACAAATG 33569
:
TCCAAACC 64321
TCCAAACC 64322
:
```

Analyse QUERY SEQUENCE

position	word
1	14236
2	33658
3	07967
:	

Index of database

sequence	position	word
1	1	33658
1	2	07967
1	3	16210
:		
3	15	33568
3	16	07967
:		

Expand from Word Based Matches

We ‘instantly’ know which sequences in the database have at least a word length match with our query sequence, and at what relative position.

Next, the potential alignments are expanded, adding up a score for (total matches - mismatches - gap penalties), to make the best possible alignment. But this is usually for a *tiny* proportion of the sequences in the database – so overall it is *much* quicker.

The highest scoring alignments are reported.

But we can potentially miss alignments with no word-size bits in common, consider BLASTn with a default word-size of 11:

```
TCGGAAGTGGAAGCTGAACCTGATTGTAGAGTTGGAGGCCAGTGTTCTGGCTGAGC|||||||||
||||| |||||||||| |||||||||| |||| ||||| |||||||
TCGGAAGTGTAAGCTCAACCTGATTGCAGAGTTGGAGTCCAGAGTTCTAGCTGAGC
```

Care is sometimes needed…

BLAST output for a nucleotide query sequence from a spider.

Sequences producing significant alignments:

		Score (bits)	E Value
emb|X16893.1|ECHEMSUA	Tarantula mRNA for hemocyanin subunit a	4183	0.0
emb|AJ290430.1|ECA290430	Eurypelma californicum mRNA for he...	111	1e-21
emb|AJ290429.1|ECA290429	Eurypelma californicum mRNA for he...	105	7e-20
emb|AJ277492.1|ECA277492	Eurypelma californicum mRNA for he...	100	4e-18
emb|AJ277491.1|ECA277491	Eurypelma californicum mRNA for he...	88	2e-14
gb|AF003253.1|AF003253	Manduca sexta pro-phenol oxidase sub...	72	1e-09
emb|AJ277489.1|ECA277489	Eurypelma californicum mRNA for he...	70	4e-09
emb|X16894.1|ECHEMSUE	Tarantula mRNA for hemocyanin subunit e	68	1e-08
emb|X04291.1|ECHEMER1	Tarantula hemocyanin chain e mRNA fra...	68	1e-08
emb|X16654.1|ECHEMED5	Tarantula exon 5 for hemocyanin subun...	68	1e-08
gb|AE003801.1|AE003801	Drosophila melanogaster genomic scaf...	48	0.014
gb|AF161261.1|AF161261	Sarcophaga bullata prophenoloxidase ...	48	0.014
gb|AF161260.1|AF161260	Sarcophaga bullata prophenoloxidase ...	48	0.014
gb|AC004640.1|AC004640	Drosophila melanogaster DNA sequence...	48	0.014
emb|X16652.1|ECHEMED3	Tarantula exon 3 for hemocyanin subun...	48	0.014
dbj|D45835.1|DROORA	Drosophila melanogaster pro-phenol oxid...	48	0.014
gb|AC007357.2|F3F19	Arabidopsis thaliana chromosome 1 BAC F...	44	0.22
emb|X16653.1|ECHEMED4	Tarantula exon 4 for hemocyanin subun...	44	0.22
gb|AF155223.1|AF155223	Porphyromonas gingivalis tonB-linked...	42	0.86
emb|Z93929.1|HS272E8	Human DNA sequence from clone 272E8 on...	42	0.86
emb|Y07618.1|PGTLAGEN	P.gingivalis tla gene	42	0.86
dbj|D49370.1|BMOPS1A	Bombyx mori mRNA for prophenoloxidase ...	42	0.86
gb|AC008080.1|AC008080	Homo sapiens clone RP11-89N17 from 7...	40	3.4
gb|AC024848.1|AC024848	Caenorhabditis elegans clone Y67D8A,...	40	3.4
ref|NM_009347.1|	Mus musculus tectorin alpha (Tecta), mRNA	40	3.4
gb|AF179375.1|AF179375	Mycoplasma fermentans orfD1 gene, In...	40	3.4
gb|AC007450.1|AC007450	Homo sapiens 12p BAC RPCI11-434C1 (R...	40	3.4
gb|AC005255.1|AC005255	Homo sapiens chromosome 19, CIT-HSP-...	40	3.4
gb|AC002988.1|AC002988	Human DNA from chromosome 19-specifi...	40	3.4
gb|AC002113.1|HSAC002113	Human Cosmid g1862x083 from 7q31.3...	40	3.4
emb|X99805.1|MMALPHTEC	Mus musculus mRNA for alpha tectorin	40	3.4
emb|AL158111.2|CNS01RGL	Human chromosome 14 DNA sequence **...	40	3.4
emb|AL136132.15|AL136132	Human DNA sequence from clone RP11...	40	3.4

Score (bits)

is the score given letter by letter during alignment based on the Subtitution matrices.

High score = less E value.

- E value: No. of chance alignments that one will get as hits.

- Lower the E value lesser no. of chance hits

- E value of zero or less than zero indicates very good hit (highly homologous sequence)

- E value is also known as P(N) in some BLAST programs

E Value
0.0
1e-21
7e-20
4e-18
2e-14
1e-09
4e-09
1e-08
1e-08
1e-08
0.014
0.014
0.014
0.014
0.014
0.014
0.22
0.22
0.86
0.86
0.86
0.86
3.4
3.4

In this example, the E value equals

1×10^{-21}

The letter "e" is used to show that -21 is the exponent. You would "expect" to find very few random sequences in this database that match the query sequence this well.

BLAST OUTPUT

Gives the identity

Gives the similarity

```
□>gi|223452|prf||0806225A    genome Y73
          Length = 812

 Score =  951 bits (2459), Expect = 0.0
 Identities = 455/458 (99%), Positives = 456/458 (99%)

Query: 1    VPSPYPSTLTGGGTVFVALYDYEARTTDDLSFKGGERFQIINNTEGDWWEARSIATGKTG 60
            VPSPYPSTLTGGGTVFVALYDYEARTTDDLSFK GERFQIINNTEGDWWEARSIATGKTG
Sbjct: 355 VPSPYPSTLTGGGTVFVALYDYEARTTDDLSFKKGERFQIINNTEGDWWEARSIATGKTG 414

Query: 61   YIPSNYVAPADSIEAEEWYFGKMGRKDAERLLLNPGNQRGIFLVRESETTKGAYSLSIRD 120
            YIPSNYVAPADSI+AEEWYFGKMGRKDAERLLLNPGNQRGIFLVRESETTKGAYSLSIRD
Sbjct: 415 YIPSNYVAPADSIQAEEWYFGKMGRKDAERLLLNPGNQRGIFLVRESETTKGAYSLSIRD 474

Query: 121 WDEVRGDNVKHYKIRKLDNGGYYITTRAQFESLQKLVKHSREHADGLCHKLTTVCPTVKP 180
            WDEVRGDNVKHYKIRKLDNGGYYITTRAQFESLQKLVKH REHADGLCHKLTTVCPTVKP
Sbjct: 475 WDEVRGDNVKHYKIRKLDNGGYYITTRAQFESLQKLVKHYREHADGLCHKLTTVCPTVKP 534
```

BLAST搜索的统计学显著性

对于两个随机序列s和t，随机观察到比对得分大于等于S的概率：

$$P(s \geq S) = 1 - \exp(-Kste^{-\lambda S})$$

BLAST返回比对得分大于阈值S的期望值为：

$$E = Kste^{-\lambda S}$$

随着S的增加，E值呈指数下降，比对随机发生的可能性就接近于0 （阈值越高，序列相似就越可信）

数据库的大小和探测序列的长度影响比对随机发生的可能性 （序列越长，序列相似就越可信）

K is a natural scale for the search space size

λ is a natural scale for the scoring system–

$P = 1 - \exp(-E)$

假阳性升高

E	*P*
10	0.99995
5	0.99326
2	0.86466
1	0.63212
0.1	0.09516
0.05	0.04877
0.001	0.0009995
0.0001	0.0001

E-values

The number of matches like the discovered match that I would expect to find by chance.

An E-value of 0.0 implies that I would expect no matches like this to arise by chance, therefore…

An E-value of 1 implies I would expect 1 match like this to arise by chance, so if I have a match with such an E-value…

Also “expect value“ or “expectation”

Example Calculation

For BLOSUM62, λ = 0.318 and K=0.14

Seq1: FMMLVKEEKVLMMF

Seq2: YMMLVQEDQVLMMY

Length(seq1)=250, Length(seq1)=470

Which scores 54 using BLOSUM 62

$E = Kste^{-\lambda S}$

E(x > 54) = 0.14 * 250 * 470 * e-(0.318*54)
= 5.734 x 10-4 = 0.0005734

E-values From First Principles

Some database statistics (23rd July 2005):

Database: NCBI RefSeq mRNA
272,619 sequences; 503,566,580 total letters (~5.0×10^8)

Database: NCBI nr
3,329,110 sequences; 14,601,814,750 total letters (~1.4×10^{10})

We will consider first searching a nucleotide sequence ('ACGTAGACGT') against a nucleotide database, e.g. the RefSeq mRNA above.

Then we will consider the more complex case of amino acid sequence (protein) searches. Which is of course what we mostly do.

Calculating an E-value

The RefSeq mRNA database has ~ 5.0×10^8 letters

There are 4 possible nucleotides - ACGT

How many matches do we expect to find by chance?

Query = 'A'

```
CCGCCAGCTACGGTCACCGAGCTTCTCATTGCTCTTCCTAACAGTGTGATAGGCTAACCGTAATGGCG
     A   A      A   A        A             AA A     A A    AA    AA
```

Expected number of matches = $(5.0 \times 10^8) / 4 = \sim 1.2 \times 10^8$

Query = 'AC'

```
CCGCCAGCTACGGTCACCGAGCTTCTCATTGCTCTTCCTAACAGTGTGATAGGCTAACCGTAATGGCG
          AC     AC                             AC                AC
```

Expected number of matches = $(5.0 \times 10^8) / (4 \times 4) = \sim 3.1 \times 10^7$

Query = 'ACG'

```
CCGCCAGCTACGGTCACCGAGCTTCTCATTGCTCTTCCTAACAGTGTGATAGGCTAACCGTAATGGCG
          ACG
```

Expected number of matches = $(5.0 \times 10^8) / (4 \times 4 \times 4) = \sim 8.1 \times 10^6$

Query = 'ACGTCGA…..CTGATTCG' - 60-mer

Expected number of matches = $(5.0 \times 10^8) / (4 \times 4 \times 4 \times 4 \ldots 60 \text{ times})$

$= (5.0 \times 10^8) / 10^{36}$

$= 5.0 \times 10^{-28}$

E-value $= 5.0 \times 10^{-28}$

E-values In Practice

So if I take a 60 nt sequence:

>sequence

ACAGCTCGTCCTCCTTCCGAGCCTACCGGGCCGCCCTCTCGGAGGTGGAACCGCCGTGCA

and actually BLAST it against the RefSeq mRNA database, I get:

```
BLAST OUTPUT:
>gi|27469838|gb|BC041710.1| ___ Homo sapiens Rap guanine nucleotide exchange factor
(GEF) 1, transcript variant 2, mRNA (cDNA clone MGC:49019 IMAGE:6051007), complete cds
          Length=6060
Score = 119 bits (60), Expect = 2e-26
Identities = 60/60 (100%), Gaps = 0/60 (0%) Strand=Plus/Plus

Query    1 ACAGCTCGTCCTCCTTCCGAGCCTACCGGGCCGCCCTCTCGGAGGTGGAACCGCCGTGCA 60
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct 2977 ACAGCTCGTCCTCCTTCCGAGCCTACCGGGCCGCCCTCTCGGAGGTGGAACCGCCGTGCA 3036
```

theoretical value was $5.0e^{-28}$ - !?

E-values: Effect of Database Size

The nr mRNA database has ~ 1.4×10^{10} letters (was RefSeq and 5.0×10^{8})
There are 4 possible nucleotides - ACGT
How many matches do we expect to find by chance?

Query = 'A'

```
CCGCCAGCTACGGTCACCGAGCTTCTCATTGCTCTTCCTAACAGTGTGATAGGCTAACCGTAATGGCG
     A   A      A   A        A              AA A      A A    AA    AA
```

Expected number of matches = (1.4×10^{10}) / 4 = ~3×10^{9}

Query = 'AC'

```
CCGCCAGCTACGGTCACCGAGCTTCTCATTGCTCTTCCTAACAGTGTGATAGGCTAACCGTAATGGCG
          AC     AC                              AC                AC
```

Expected number of matches = (1.4×10^{10}) / (4x 4) = ~1×10^{8}

Query = 'ACG'

```
CCGCCAGCTACGGTCACCGAGCTTCTCATTGCTCTTCCTAACAGTGTGATAGGCTAACCGTAATGGCG
          ACG
```

Expected number of matches = (1.4×10^{10}) / (4 x 4 x 4) = ~2×10^{7}

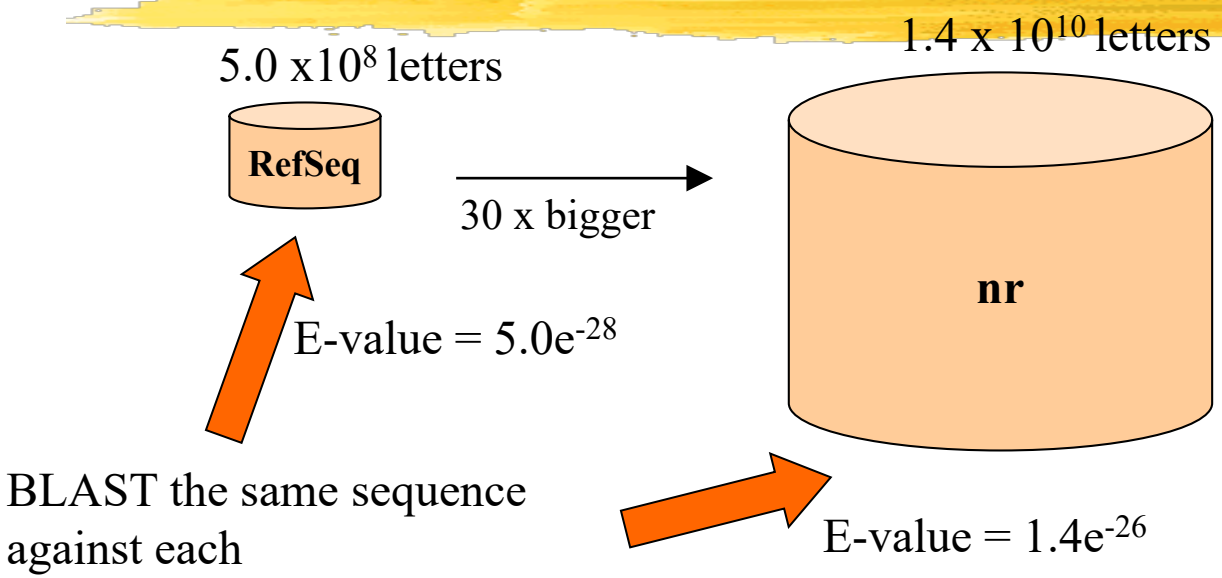
Query = 'ACGTCGA…..CTGATTCG' - 60-mer

Expected number of matches = (1.4×10^{10}) / (4 x 4 x 4 x 4 … 60 times)
= (1.4×10^{10}) / 10^{36}
= 1.4×10^{-26}

E-value = 1.4×10^{-26}

(was E-value = 5.0 x 10-28)

E-values: Effect of Database Size



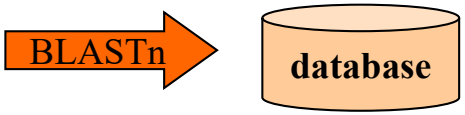
The database was ~30 times bigger and so the E-value was ~30 times bigger.

The E-value is simply *dependent on database size*.

E-values: Effect of Query Length

BLAST 500 nt sequence against a database

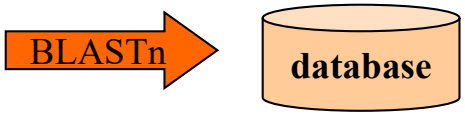
```
>sequence
ACTAGTCTAGCTAGACATCG
ATCGATGATGCTACACAGAT
AGACGATAGATAGTAAGTCG
ATCGATCGCGCATCGATCGT
CTAGATCGATCGCTCGCTGT
GTAGATAGATCGGCGATAGA
```



Get a full length match with sequence XYZ at an E-value = $5.0e^{-160}$

BLAST *half* of *the same* sequence against *the same* database

```
>sequence
ACTAGTCTAGCTAGACATCG
ATCGATGATGCTACACAGAT
AGACGATAGATAGTAAGTCG
```



Get a match with sequence XYZ again, but at an E-value = $5.0e^{-80}$

Biologically it's the same match! Does it mean we are any less sure that this match didn't occur by chance?
The E-value is simply *dependent on match length*.

Why not just use % identity?

At some levels this a good question.

But consider two very different searches, both of which give a 75% identity match

Query1 was 60 nt long:

```
CGGAGCTCAGGGCTTAACGACTGATATCTCCGCGCATGTCGAGAAACGATACAGCCAGCG
||||||||||| || | || | || || |||| | |  | ||||||  | ||||||||||
CGGAGCTCAGGCCTCACCGGCGGACATGTCCGGGAAAATAGAGAAAGCAGACAGCCAGCG
```

Which would have an E-value ~ 5.0×10^{-19}

And, Query2 only 16 nt long:

```
ACGTACGTACGTACGT
||| || | |||| ||
ACGCACCTTCGTAGGT
```

Which would have an E-value ~ 30

And intuitively we feel we would expect to see that sort of number of matches in the database just by chance…

So what's the real problem?

Basically you are usually trying to answer the question:

Can I find the ortholog of my gene in some other species, so that I can work out what it might be doing in my organism?

The difficulty is because:

BLAST → ~~**ORTHOLOGY**~~

biological knowledge

nature of query sequence

BLAST → **Similarity + Probability**

phylogenetic relationship

Query,match length, size of database…

Are there any useful guidelines though, at least for biological meaningfulness?

Rules of Thumb

How good does an E-value have to be before we might even think we have an ortholog?

← larger/worse smaller/better →

E-values 10^{-5} 10^{-10} 10^{-40} 10^{-100} 0.0

fantasy

borderline

encouraging

pretty good

can't get better

But note that in some gene families with closely related members you can get an E-value of 0.0 for several different matches, and then % identity may be more sensitive. Also bear in mind, in cases like this, that ideas of 'functional' orthology may break down, with more than one locus producing identical proteins which share the same function…

BLAST

BLAST query schemes:

- **Amino acid seq: against db?**
  - **Blastp (protein sequence db)**
  - **Tblastn (translated nucleotide sequence db)**

- **DNA seq: against db?**
  - **Blastn (nucleotide db)**
  - **Blastx (protein sequence db)**
  - **Tblastx (translated nucleotide sequence db)**

Flavours of BLAST

query sequence	other operation?	database sequences

BLASTn

ACGATAGATCCCATCCATAAAT

ATGACGATAGATCCCATCAT
CGATAGGACCACCACA
GATAGACCAGGATACATAGGATAATTA
AGCTCGCTTGGCTCGATGGCT

FAST

BLASTp

MQWCGYRWTYQGYRW

MKJLSPWERSYTRGHYTWER MGHTVNBZY
MKLPWRHGDBKJGMNDFD
MBKLRPIUHDFRTASGSLKWWRTVBN

FAST

BLASTx

6 frame translation

ACGATAGATCCCATCCATAAAT

MKJLSPWERSYTRGHYTWER MGHTVNBZY
MKLPWRHGDBKJGMNDFD
MBKLRPIUHDFRTASGSLKWWRTVBN

SLOW

tBLASTn

MQWCGYRWTYQGYRW

ATGACGATAGATCCCATCAT
CGATAGGACCACCACA
GATAGACCAGGATACATAGGATAATTA
AGCTCGCTTGGCTCGATGGCT

SLOWER

ACGATAGATCCCATCCATAAAT

tBLASTx

ATGACGATAGATCCCATCAT
CGATAGGACCACCACA
GATAGACCAGGATACATAGGATAATTA
AGCTCGCTTGGCTCGATGGCT

HORRIBLY SLOW!

Summary

Pair-wise:

- **Dot plot:** 图形化，有助于理解
- **Blast**
  - 如何快速： word + pre-indexing
  - E-value, similarity, 生物学意义

Summary

Scoring matrices
- **PAM**
- **BLOSUM**

Sequence alignment
- **Pair-wise: 2 seq.**
  - **Local pair-wise alignment**
    - **(Smith-Waterman Algorithm)**
    - **FASTA**
    - **BLAST ****
  - **Global pair-wise alignment**
    - **(Needleman-Wunsch Algorithm)**
- **multiple alignment: more than 3 seq.**