生物信息学

序列分析与比对

吴凌云
中国科学院数学与系统科学研究院

目录

- 序列的组成
- 单一序列分析
- 序列比对
- 序列搜索
- 多序列比对

碱基分类

- 嘌呤（purine）
  - A：adenine，腺嘌呤，氨基，弱键
  - G：guanine，鸟嘌呤，酮基，强健
- 嘧啶（pyrimidine）
  - C：cytosine，胞嘧啶，氨基，强健
  - T：thymine，胸腺嘧啶，酮基，弱键
- 氨基: A, C / 酮基: T, G
- 弱键: A, T / 强键: C, G

IUB/IUPAC代码

代码	碱基	说明
R	A或G	嘌呤
Y	T或C	嘧啶
W	A或T	弱键
S	C或G	强键
M	A或C	氨基
K	G或T	酮基
B	C, G或T	非A
D	A, G或T	非C
H	A, C或T	非G
V	A, C或G	非T
N	A, G, C或T	任意碱基

碱基分布

- 四个碱基的分布不同
- 相邻碱基的分布不独立
- 与遗传密码的关系

目录

- 序列的组成
- **单一序列分析**
- 序列比对
- 序列搜索
- 多序列比对

重复序列（repeats）

- 重复序列的查找
- 记分法（Karlin 1983）
- A, C, G, T: 0, 1, 2, 3

$$\text{score} = 1 + \sum_{i=1}^{k} \alpha_i 4^{k-i}$$

- 逐步查找长度为k的重复，k=2,3,...

$$\text{Example: TGACC} = 32011_4$$
$$1 + 3 \times 4^4 + 2 \times 4^3 + 0 \times 4^2 + 1 \times 4^1 + 1 \times 4^0 = 459$$

重复序列（repeats）

- 最长重复序列的统计学估计

$$\mu_L = \frac{0.6359 + 2\ln n + \ln(1-p)}{\ln(1/p)} - 1$$

$$\sigma_L^2 = \frac{1.645}{(\ln p)^2}$$

$$p = \sum_{i=1}^{4} p_i^2$$

- 用于搜索重复序列时k的初始值

几何学分析

- 碱基计数

$$A_n, C_n, G_n, T_n, \quad n = 1, 2, \cdots, L$$

- 第n个碱基对应的点：

$$x_n = 2(A_n + G_n) - n$$

$$y_n = 2(A_n + C_n) - n$$

$$z_n = 2(A_n + T_n) - n$$

$$x_n, y_n, z_n \in [-n, n], n = 1, 2, \cdots, L$$

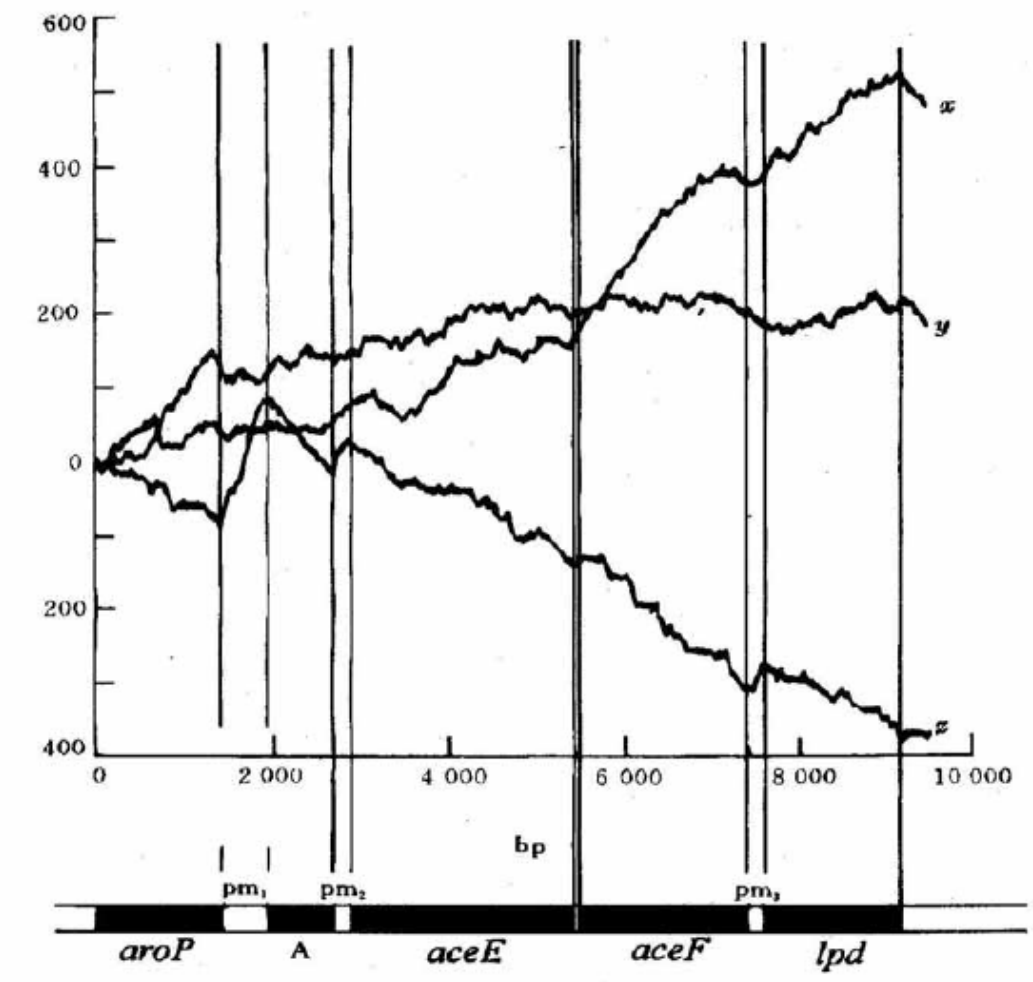
几何学分析

x_n – 嘌呤/嘧啶碱基沿序列的分布

y_n – 氨基/酮基碱基沿序列的分布

z_n – 弱键/强健碱基沿序列的分布

几何学分析



大肠杆菌ayoP基因族序列Z曲线的三个分量

目录

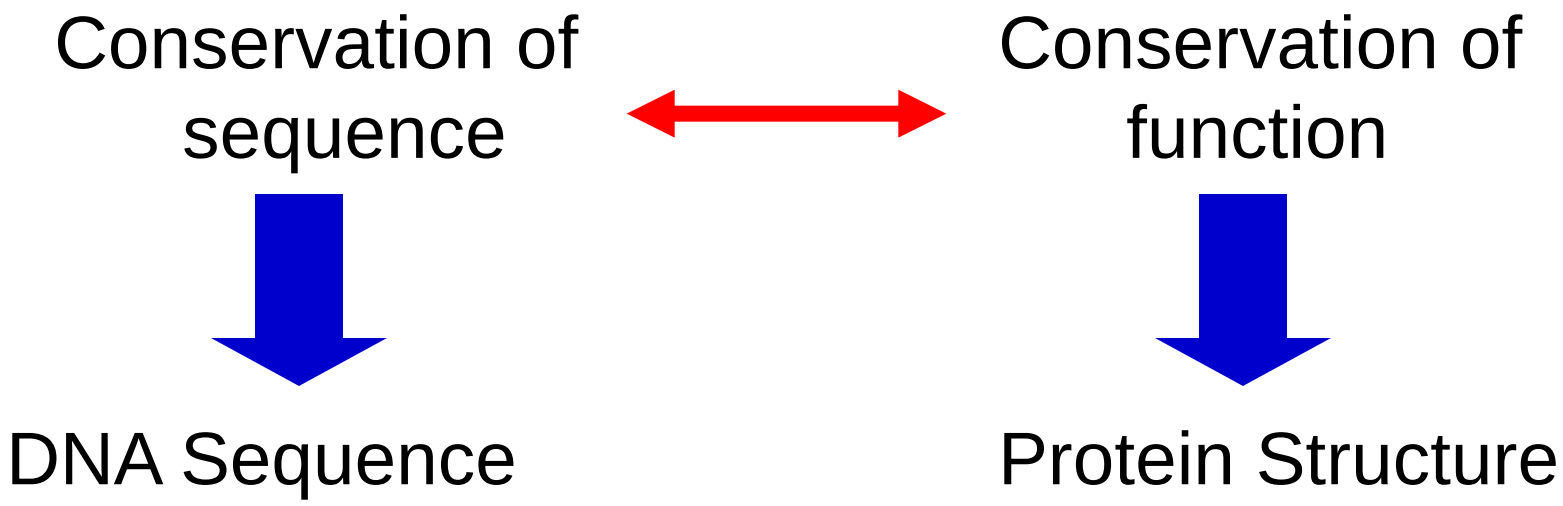
- 序列的组成
- 单一序列分析
- **序列比对**
- 序列搜索
- 多序列比对

什么是比对？

- 为了评估序列的相似程度和同源的可能性，将两个或者多个序列排列起来，以得到最高级别的一致性（在氨基酸序列中为保守性）。
  - 一致性：两个序列保持不变的区域。
  - 保守性：序列中某个位置的氨基酸发生了改变，但保留了原来残基的物理化学性质

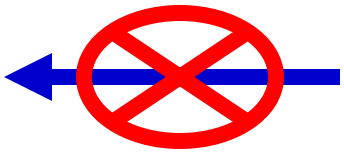
Why align sequences?

- Functional predictions based on identifying homologues
- Assumption:



Implicit Assumption

DNA Sequence  Protein Structure

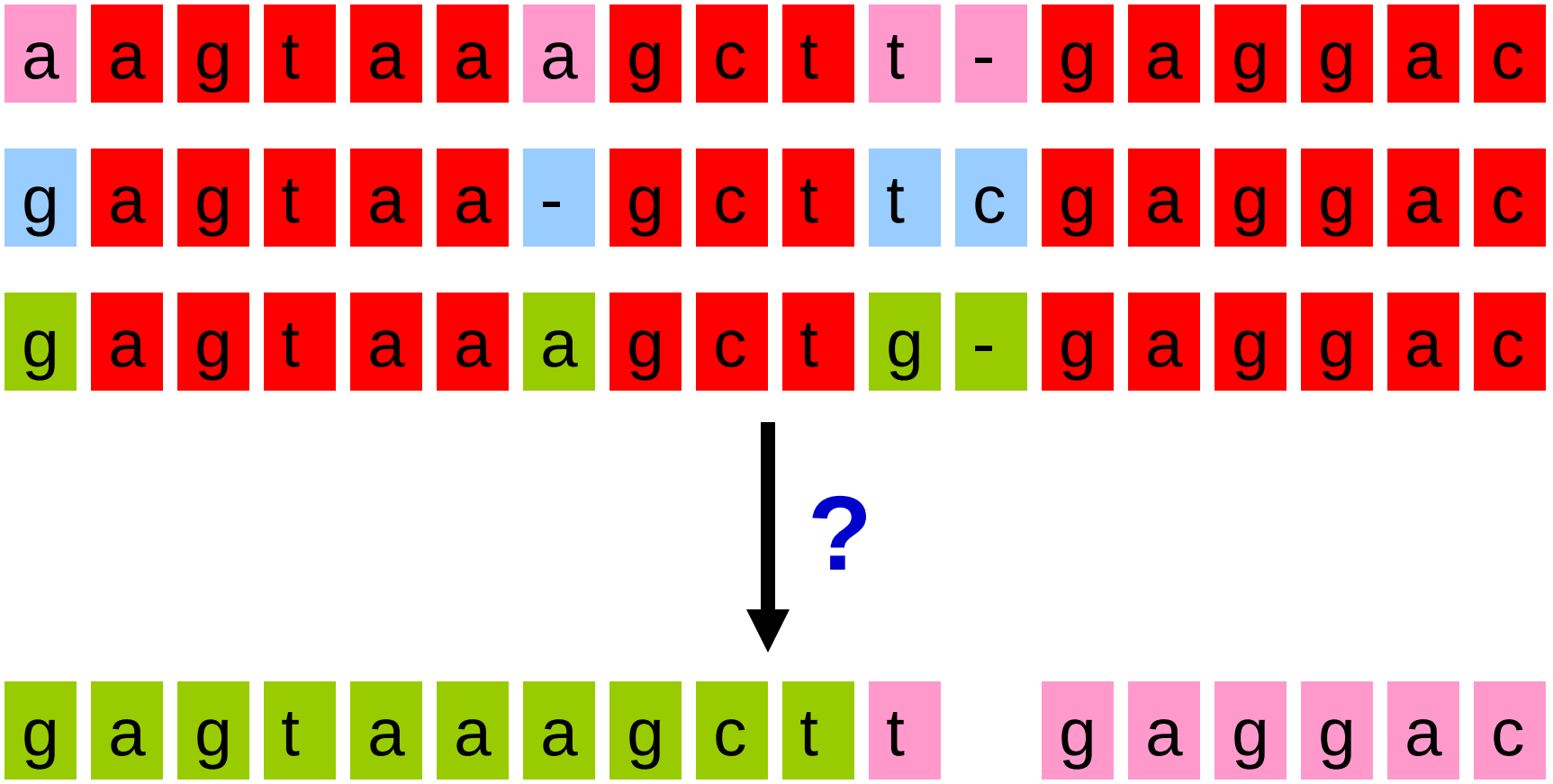
DNA Sequence  Protein Structure

Sequence Similarity

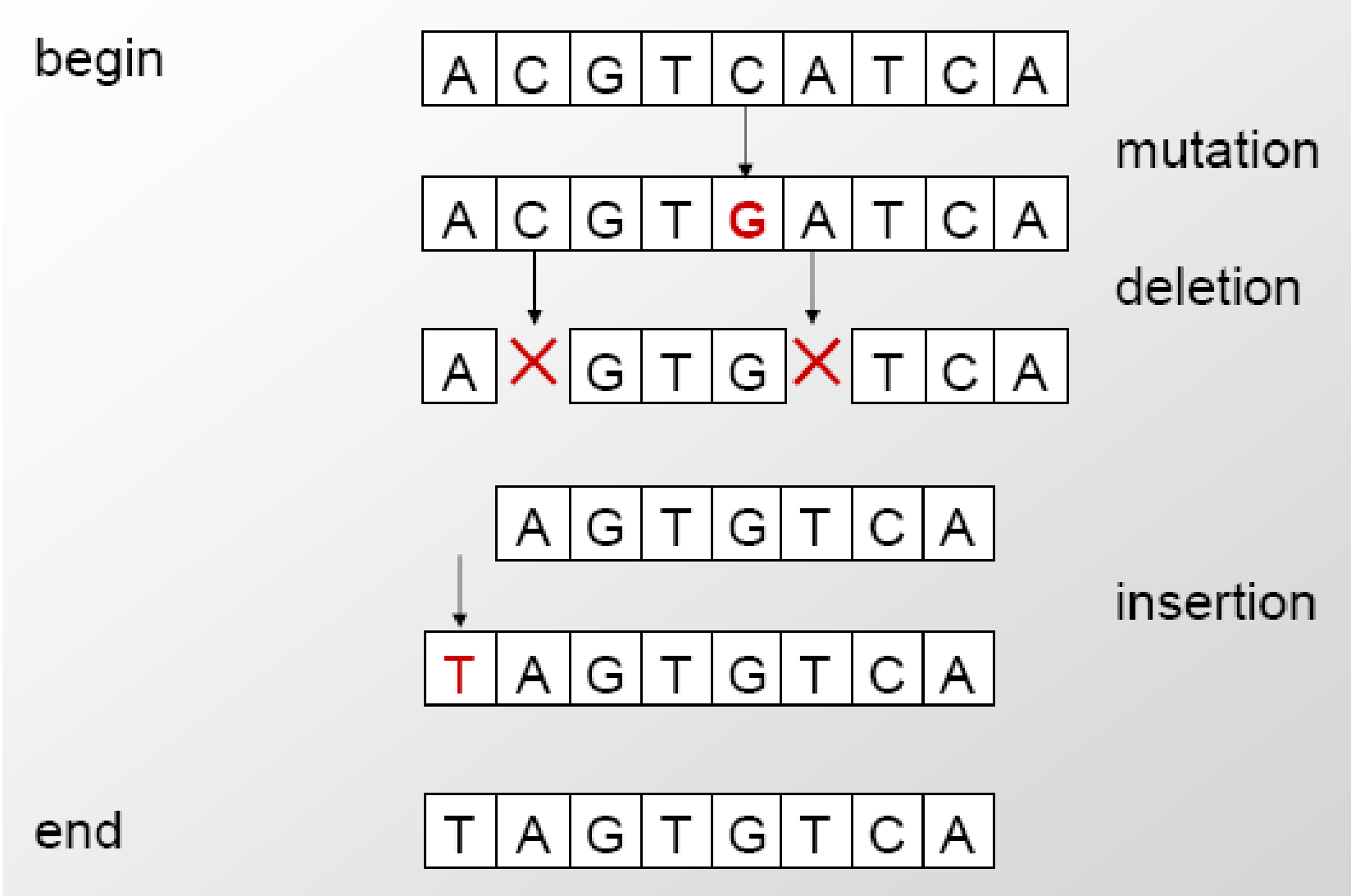
- How to measure sequence similarity?
- What's the last common ancestor?

a	a	g	t	a	a	a	g	c	t	t	g	a	g	g	a	c
g	a	g	t	a	a	g	c	t	t	c	g	a	g	g	a	c
g	a	g	t	a	a	a	g	c	t	g	g	a	g	g	a	c

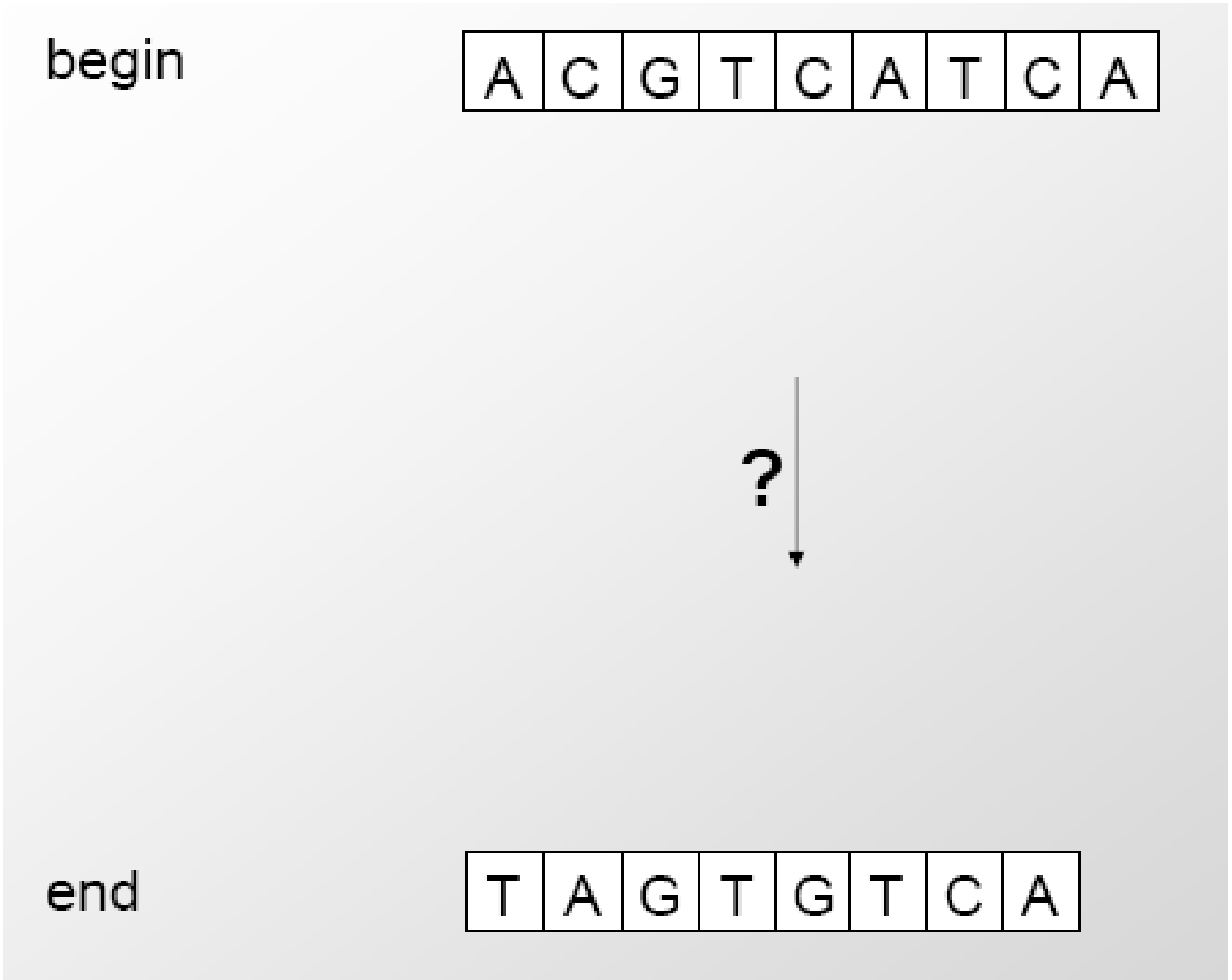
Last Common Ancestor



Sequence Evolution



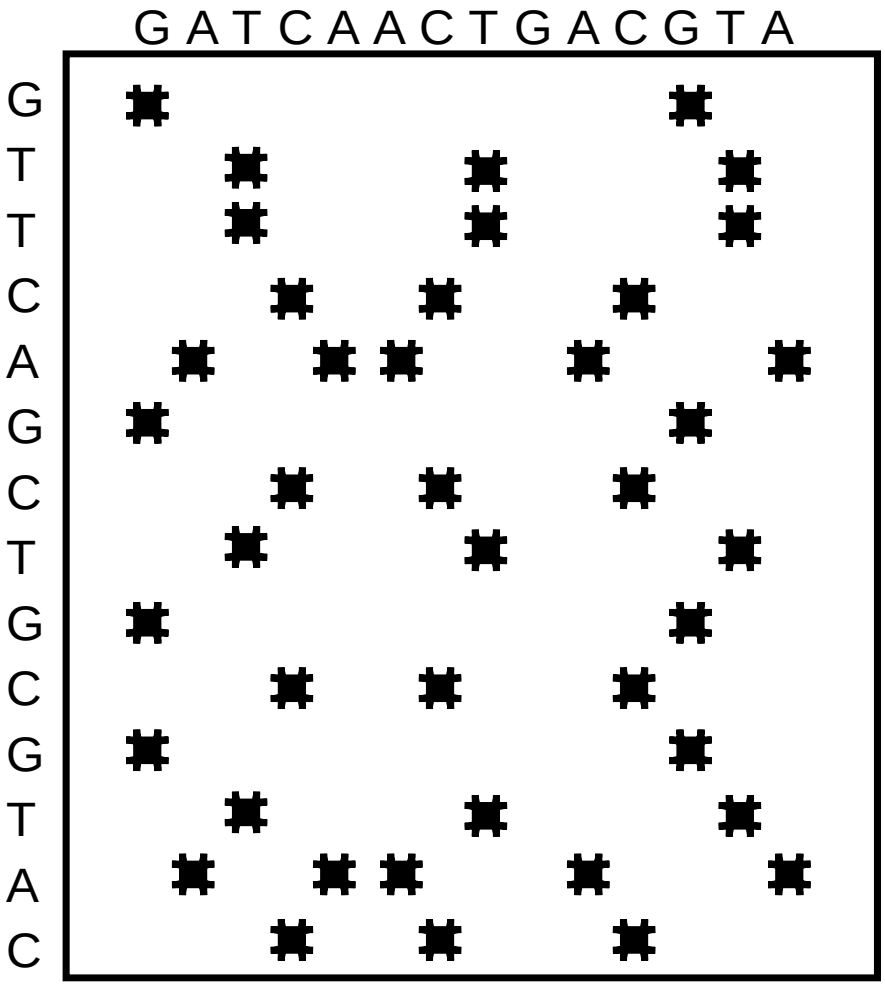
Infer Edit Operators



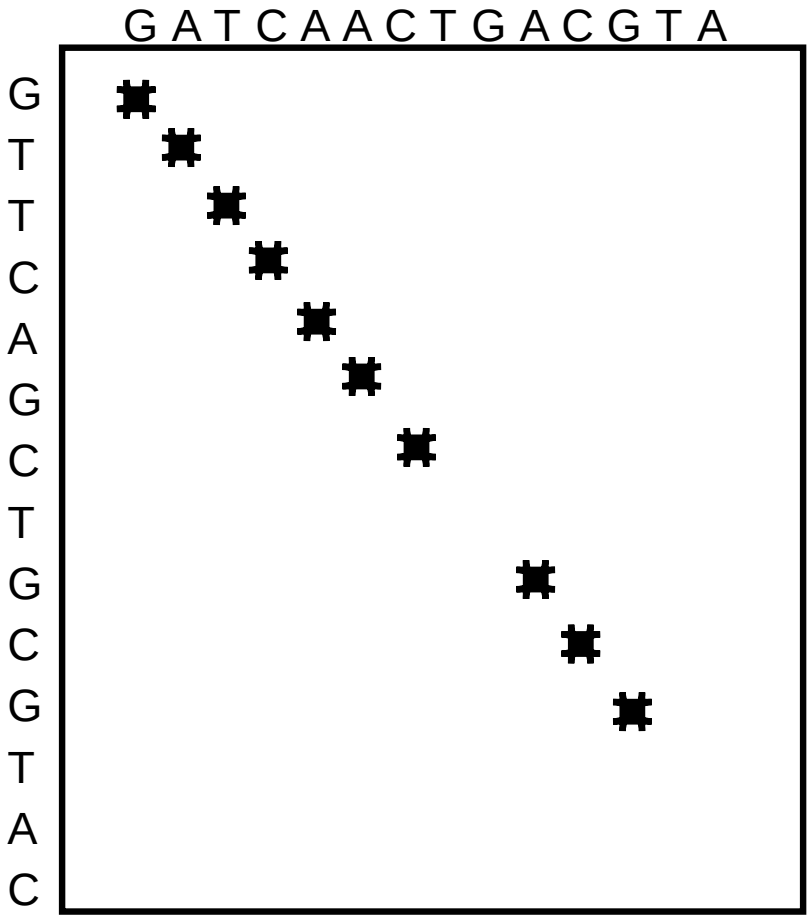
Methods

- Dot matrix analysis (intuitive)
- DP algorithm (exact)
- Word or k-tuple (FASTA, BLAST) (heuristic)

Dot Matrix



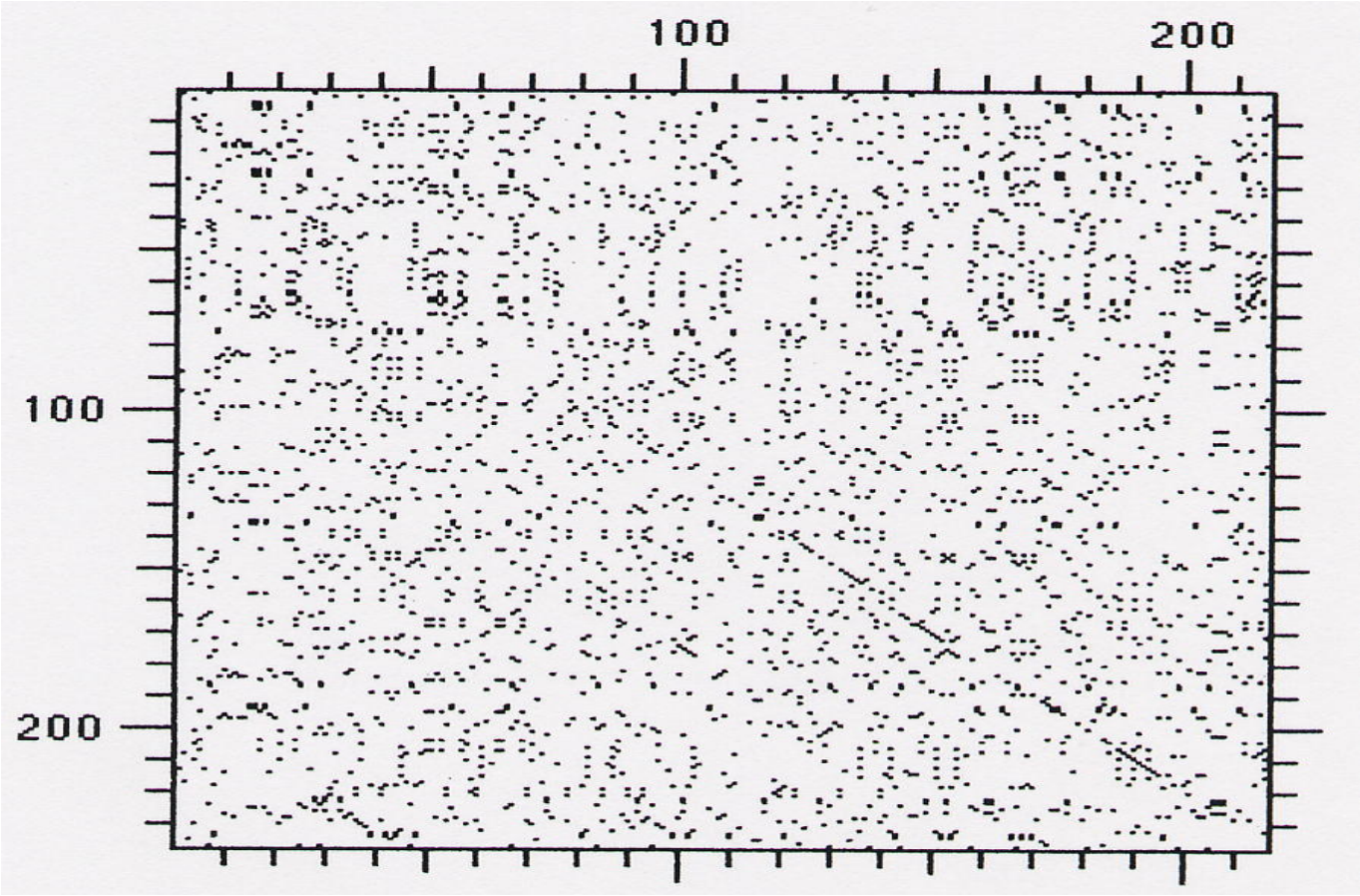
Filtered Dot Matrix



4 bases window

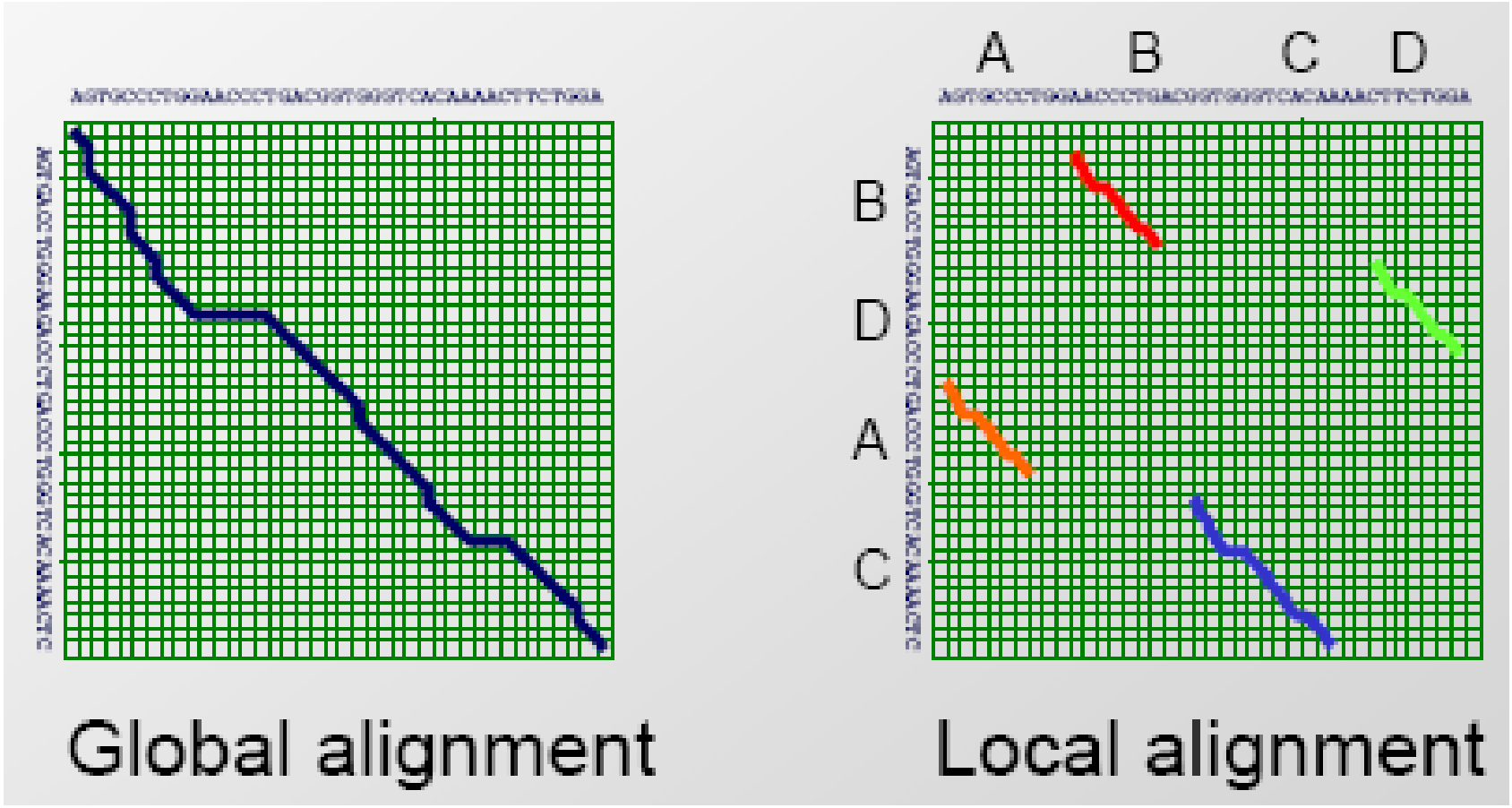
3 stringency

Dot Matrix with Real Data



The amino acid sequences of the phage λ*cI* (horizontal sequence) and phage P22 *c2* (*vertical sequence)* repressors. The window size and stringency are both 1.

Global vs. Local



Needleman-Wunsch算法

$$H_{ij} = \max \begin{cases} H_{i-1,j-1} + S(a_i, b_j) \\ \max\limits_{1 \le k \le i} (H_{i-k,j-1} - w_k + S(a_i, b_j)) \\ \max\limits_{1 \le k \le j} (H_{i-1,j-k} - w_k + S(a_i, b_j)) \end{cases}$$

$$H_{i0} = H_{0j} = 0$$

Needleman-Wunsch算法例子

	M	P	R	C	L	C	Q	R	J	N	C	B	A
P	0	1	0	0	0	0	0	0	0	0	0	0	0
B	0	0	1	1	1	1	1	1	1	1	1	2	1
R	0	0	2	1	1	1	1	2	1	1	1	1	2
C	0	0	1	3	2	3	2	2	2	2	3	2	2
K	0	0	1	2	3	3	3	3	3	3	3	3	3
C	0	0	1	3	3	4	3	3	3	3	4	3	3
R	0	0	2	2	3	3	4	?					
N													
J									1				
C				1		1					1		
J									1				
A													1

Needleman-Wunsch算法例子

	M	P	R	C	L	C	Q	R	J	N	C	B	A
P	0	1	0	0	0	0	0	0	0	0	0	0	0
B	0	0	1	1	1	1	1	1	1	1	1	2	1
R	0	0	2	1	1	1	1	2	1	1	1	1	2
C	0	0	1	3	2	3	2	2	2	2	3	2	2
K	0	0	1	2	3	3	3	3	3	3	3	3	3
C	0	0	1	3	3	4	3	3	3	3	4	3	3
R	0	0	2	2	3	3	4	5	4	4	4	4	4
N	0	0	1	2	3	3	4	4	5	6	5	5	5
J	0	0	1	2	3	3	4	4	6	5	6	6	6
C	0	0	1	3	3	4	4	4	5	6	7	6	6
J	0	0	1	2	3	3	4	4	6	6	6	7	7
A	0	0	1	2	3	3	4	4	5	6	6	7	8

```
MP-RCLCQR-JNCBA
  |  ||  |  |  |  |  |
-PBRCKC-RNJ-CJA
```

Smith-Waterman算法

$$H_{ij} = \max \begin{cases} H_{i-1,j-1} + S(a_i, b_j) \\ \max\limits_{1 \le k \le i} (H_{i-k,j} - w_k) \\ \max\limits_{1 \le k \le j} (H_{i,j-k} - w_k) \\ 0 \end{cases}$$

$$H_{i0} = H_{0j} = 0$$

Smith-Waterman算法例子

	C	A	G	C	C	U	C	G	C	U	U	A	G
A	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
A	0.0	1.0	0.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.7
U	0.0	0.0	0.8	0.3	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	0.7
G	0.0	0.0	1.0	0.3	0.0	0.0	0.7	1.0	0.0	0.0	0.7	0.7	1.0
C	1.0	0.0	0.0	2.0	1.3	0.3	1.0	0.3	2.0	0.7	0.3	0.3	0.3
C	1.0	0.7	0.0	1.0	3.0	1.7	?						
A													
U													
U													
G													
A													
C													
G													
G													

Smith-Waterman算法例子

	C	A	G	C	C	U	C	G	C	U	U	A	G
A	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
A	0.0	1.0	0.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.7
U	0.0	0.0	0.8	0.3	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	0.7
G	0.0	0.0	1.0	0.3	0.0	0.0	0.7	1.0	0.0	0.0	0.7	0.7	1.0
C	1.0	0.0	0.0	2.0	1.3	0.3	1.0	0.3	2.0	0.7	0.3	0.3	0.3
C	1.0	0.7	0.0	1.0	3.0	1.7	1.3	1.0	1.3	1.7	0.3	0.0	0.0
A	0.0	2.0	0.7	0.3	1.7	2.7	1.3	1.0	0.7	1.0	1.3	1.3	0.0
U	0.0	0.7	1.7	0.3	1.3	2.7	2.3	1.0	0.7	1.7	2.0	1.0	1.0
U	0.0	0.3	0.3	1.3	1.0	2.3	2.3	2.0	0.7	1.7	2.7	1.7	1.0
G	0.0	0.0	1.3	0.0	1.0	1.0	2.0	3.3	2.0	1.7	1.3	2.3	2.7
A	0.0	1.0	0.0	1.0	0.3	0.7	0.7	2.0	3.0	1.7	1.3	2.3	2.0
C	1.0	0.0	0.7	1.0	2.0	0.7	1.7	1.7	3.0	2.7	1.3	1.0	2.0
G	0.0	0.7	1.0	0.3	0.7	1.7	0.3	2.7	1.7	2.7	2.3	1.0	2.0
G	0.0	0.0	1.7	0.7	0.3	0.3	1.3	1.3	2.3	1.3	2.3	2.0	2.0

GCC-UCG

GCCAUUG

算法比较

Needleman-Wunsch

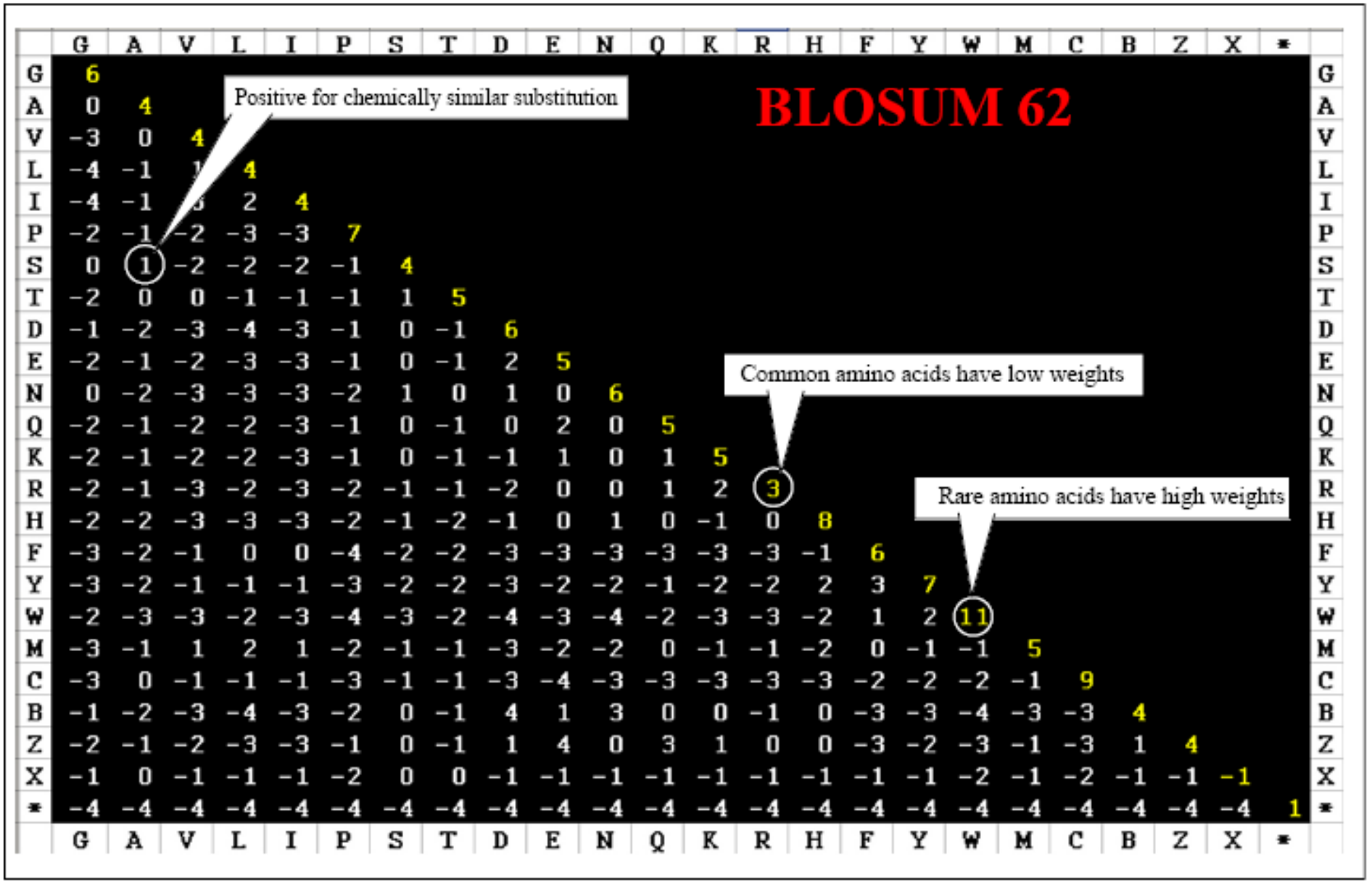
1. 全局比对
2. 距离分数：必须非负
3. 不需要gap惩罚
4. 在一条路径中分数不会减少

Smith-Waterman

1. 局部比对
2. 相似性分数：可以是正数或者负数
3. 必须有gap惩罚
4. 在一条路径中分数可能增加、减少或者不变

Substitution Matrix

- PAM (Point Accepted Mutations)
  - Based on substitution data from alignment between similar proteins
  - 1% expected substitutions = 1PAM
  - PAMn = $(1PAM)^n$
- BLOSUM (BLOck Scoring Matrix)
  - Multiple alignment of distantly related proteins
  - BLOSUMn = sequences with n% or more of identical residues were clustered to compute log-odds ratio



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

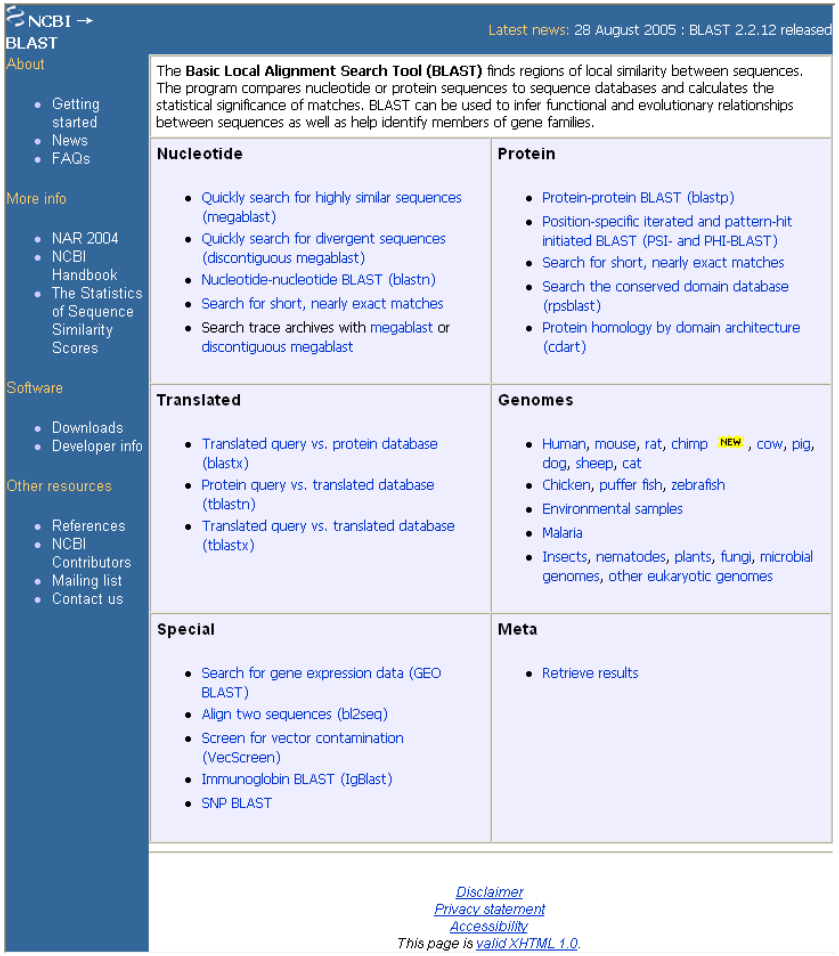
目录

- 序列的组成
- 单一序列分析
- 序列比对
- **序列搜索**
- 多序列比对

序列搜索

- 将查询序列和数据库中的所有记录进行比对
- 将每个比对打分
- 按照分值对数据库中的所有记录排序
- 报告具有高分值的那些数据库记录

BLAST主页



NCBI → BLAST

Latest news: 28 August 2005 : BLAST 2.2.12 released

About
- Getting started
- News
- FAQs

More info
- NAR 2004
- NCBI Handbook
- The Statistics of Sequence Similarity Scores

Software
- Downloads
- Developer info

Other resources
- References
- NCBI Contributors
- Mailing list
- Contact us

The **Basic Local Alignment Search Tool (BLAST)** finds regions of local similarity between sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance of matches. BLAST can be used to infer functional and evolutionary relationships between sequences as well as help identify members of gene families.

Nucleotide
- Quickly search for highly similar sequences (megablast)
- Quickly search for divergent sequences (discontiguous megablast)
- Nucleotide-nucleotide BLAST (blastn)
- Search for short, nearly exact matches
- Search trace archives with megablast or discontiguous megablast

Protein
- Protein-protein BLAST (blastp)
- Position-specific iterated and pattern-hit initiated BLAST (PSI- and PHI-BLAST)
- Search for short, nearly exact matches
- Search the conserved domain database (rpsblast)
- Protein homology by domain architecture (cdart)

Translated
- Translated query vs. protein database (blastx)
- Protein query vs. translated database (tblastn)
- Translated query vs. translated database (tblastx)

Genomes
- Human, mouse, rat, chimp NEW, cow, pig, dog, sheep, cat
- Chicken, puffer fish, zebrafish
- Environmental samples
- Malaria
- Insects, nematodes, plants, fungi, microbial genomes, other eukaryotic genomes

Special
- Search for gene expression data (GEO BLAST)
- Align two sequences (bl2seq)
- Screen for vector contamination (VecScreen)
- Immunoglobin BLAST (IgBlast)
- SNP BLAST

Meta
- Retrieve results

Disclaimer
Privacy statement
Accessibility
This page is valid XHTML 1.0.

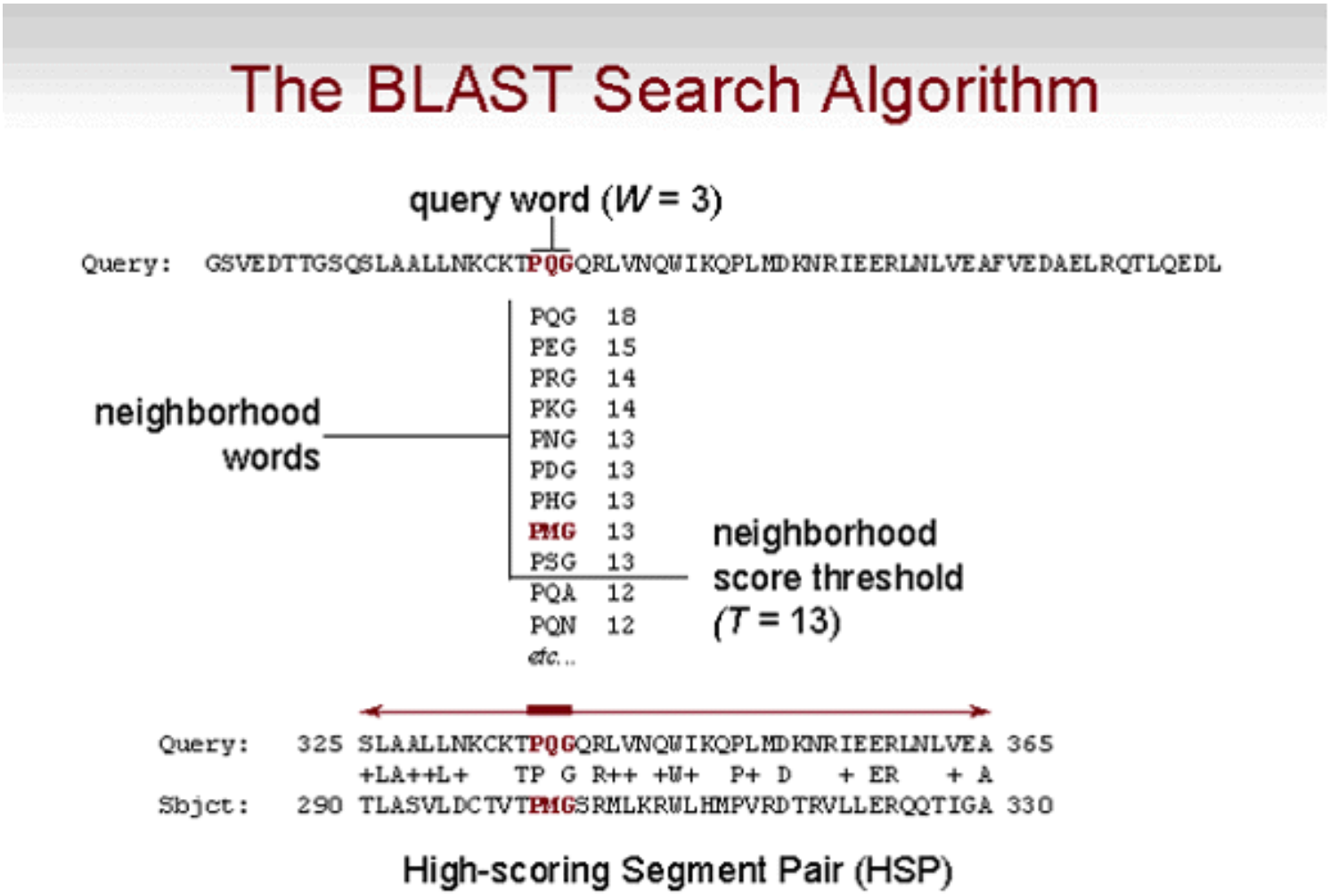
BLAST算法

- BLAST使用预先索引好的数据库
  - 每个“单词”在数据库记录中的位置都已经被记录下来
- 将待查询的序列分解成“单词”
- 在数据库中搜索相似的单词，用替换矩阵（substitution matrix）进行打分，相似的要求为分值不低于阈值T

BLAST算法

- 查询序列中两个不重叠的“单词”，而且相互距离在一定范围内，如果都与数据库中的记录匹配上，而且之间没有间隙，则该区域称为segment pair
- 将segment pair向两个方向扩展，直到其分值比其最高值下降了X
- 报告所有E value低于指定值的结果

High-scoring Segment Pair



BLAST算法的参数

- “单词”长度W
  - DNA序列：W=11
  - 蛋白质序列：W=3
- Segment pair的最低分值T
  - 较低的T可以查找更远的相似关系
- Segment pair扩展截止X
  - 较高的X可以增加HSP的长度
- E value过滤阈值
  - 对分值没有影响，只影响报告的结果数量

E value

- 在数据库中随机进行搜索时，找到分值高于或等于S的不同比对的期望次数。
- E value越小，则该分值的统计显著性越大。
- 和查询序列的长度有关

极值分布（EVD）

- the sum of a large number of independent identically distributed (i.i.d) random variables tends to a normal distribution
- the maximum of a large number of i.i.d. random variables tends to an extreme value distribution

$$E = Kmn\, e^{-\lambda S}$$

Bit scores

- “标准化”后的HSP分数

$$S' = \frac{\lambda S - \ln K}{\ln 2}$$

- 等价于如下的极值分布

$$E = mn\,2^{-S'}$$

P-value

- 在数据库搜索中，发现至少一个相同分值或者更高分值的比对发生的概率。

$$P = 1 - e^{-E}$$

- E-value比P-value直观
- 当E<0.01时，两者很接近

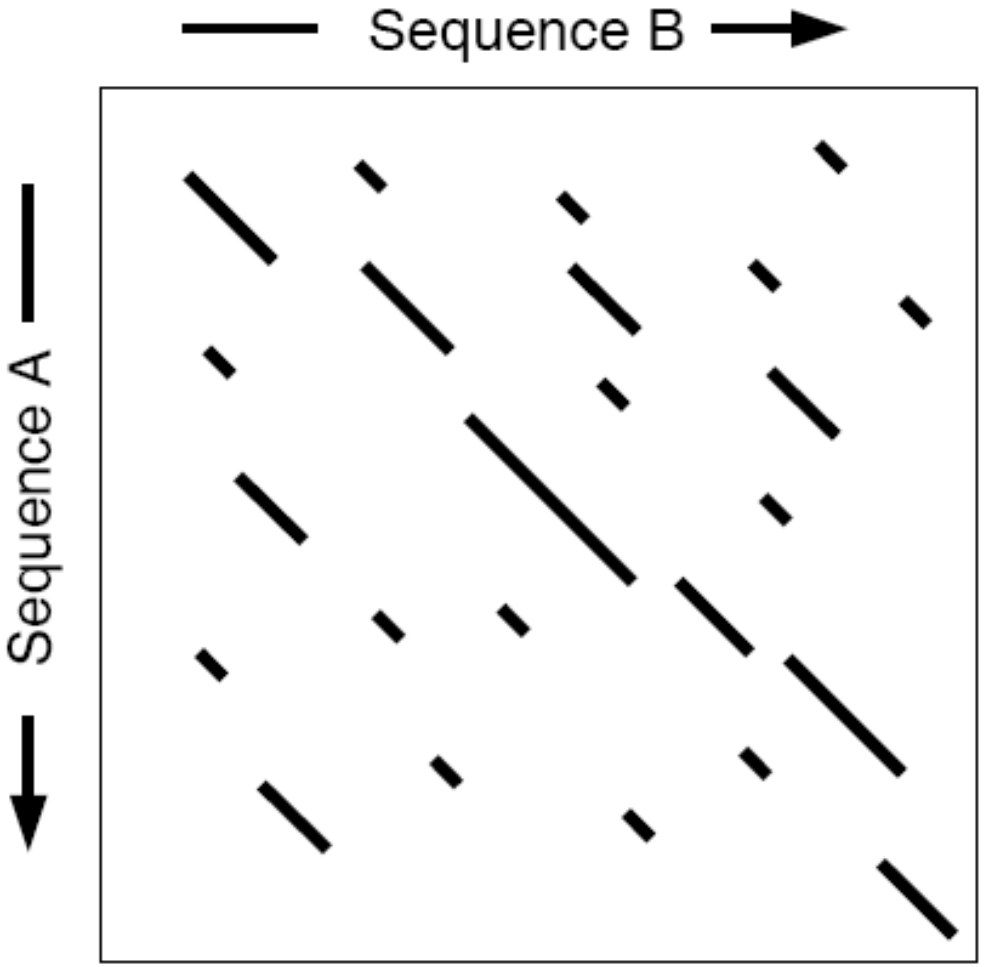
BLAST程序

	Sequence	Database
blastn	nucleotide	nucleotide
blastp	protein	protein
blastx	nucleotide	protein
tblastn	protein	nucleotide
tlbastx	nucleotide	nucleotide
PSI-blast	protein	protein
PHI-blast	protein	protein

FASTA算法

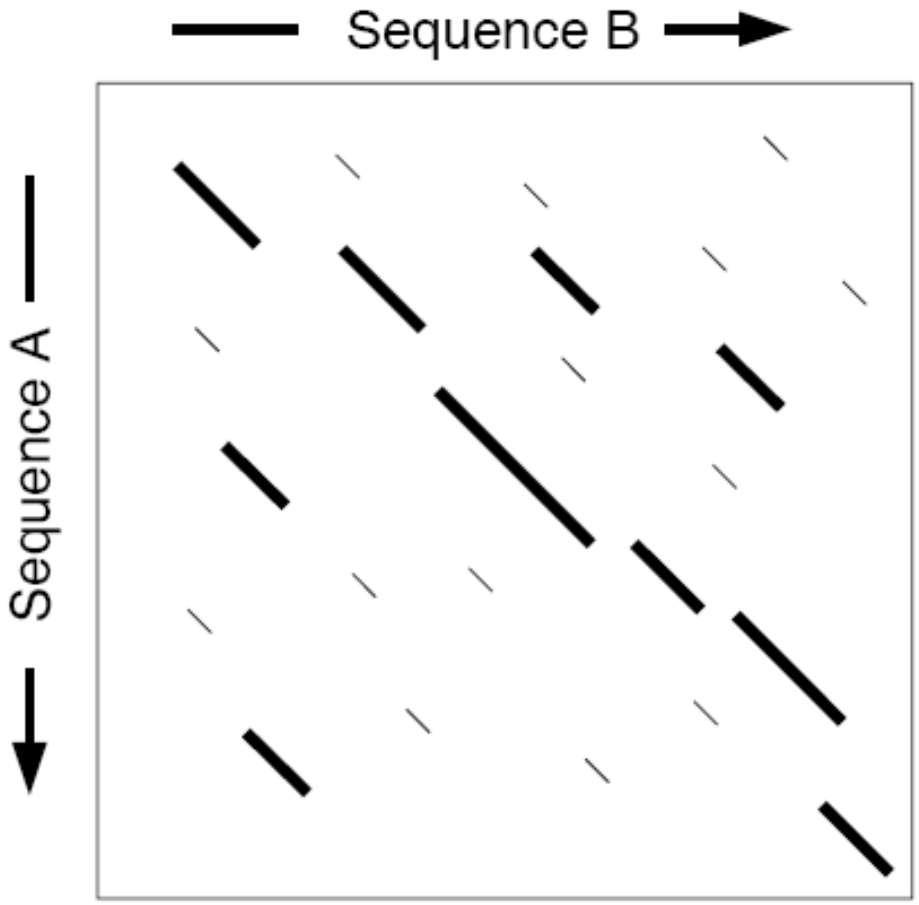
- 速度介于Smith-Waterman算法和BLAST之间
- 敏感度（sensitivity）也介于Smith-Waterman算法和BLAST之间
- 是Smith-Waterman算法的快速近似
- FASTA算法可以进行全局比对，而BLAST只是局部比对

FASTA算法（step 1）



Find runs of identities

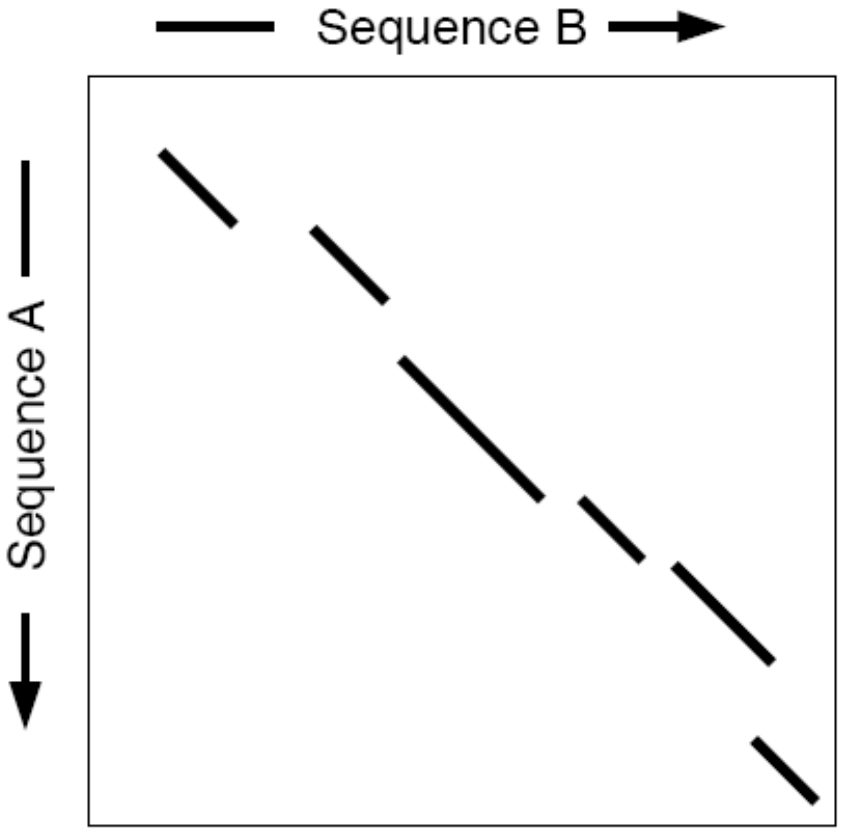
FASTA算法（step 2）



Re-score using PAM matrix

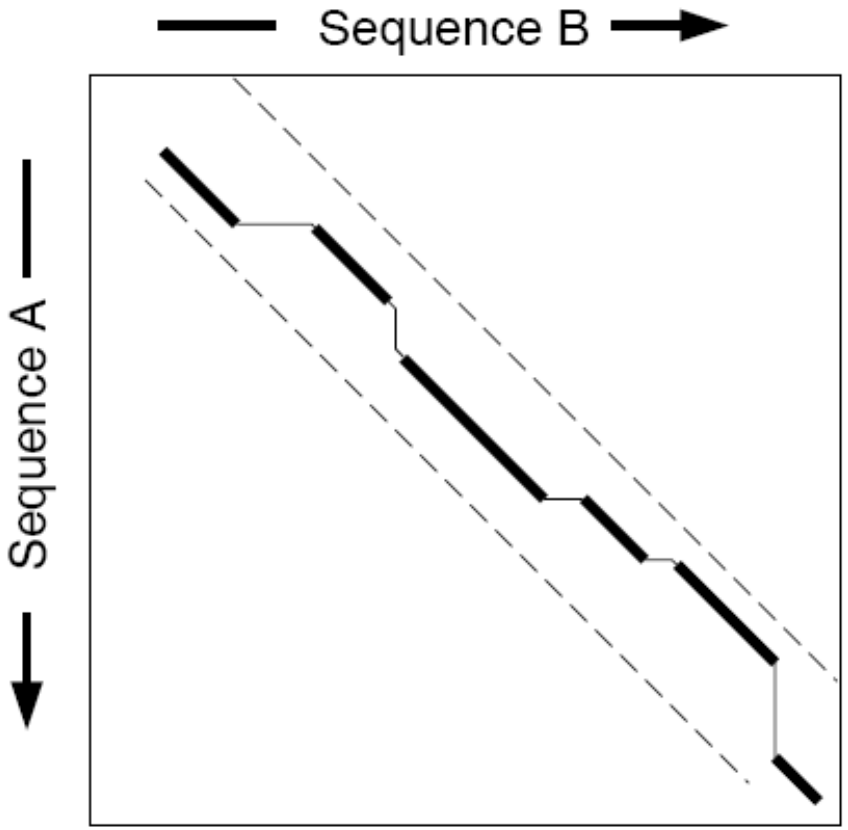
Keep top scoring segments.

FASTA算法（step 3）



Apply "joining threshold"
to eliminate segments that
are unlikely to be part of the alignment
that includes highest scoring segment.

FASTA算法（step 4）



Use dynamic programming to optimise the alignment in a narrow band that encompasses the top scoring segments.

目录

- 序列的组成
- 单一序列分析
- 序列比对
- 序列搜索
- **多序列比对**

多序列比对（MSA）

- Needleman-Wunsch算法和Smith-Waterman算法可以直接推广到K条序列比较
- DP算法的计算复杂性
  - 从$O(n^2)$增长为$O(n^k)$
  - k>6基本上就无法求解
- MSA问题是NP-hard（对于k来说）

解决方法

- 启发式算法（heuristic）
- 固定参数复杂性（fixed parameter complexity）
- 多项式时间近似算法（polynomial time approximation algorithm）

启发式算法（1）

- Bains (1986): consensus alignment
  - 每个序列都和consensus序列做比对
  - 产生新的consensus
  - 直到consensus不再变化
  - 依赖于每个序列比对的顺序

启发式算法（2）

- Higgins and Sharp (1989): Clustal
  - 计算所有两两比对的分值
  - 建立相似性矩阵，对序列进行聚类
  - 用consensus方法对类进行比对
  - 对类中的序列依次进行比对
- ClustalW：局部比对
- ClustalV：全局比对

MSA和进化树

- MSA的结果会影响进化树的推断
- 不同的进化树会导致不同的MSA
- 鸡和蛋的问题
- 同时考虑MSA和构建进化树
  - Jotun Hein: TREEALIGN

MSA数据格式

.....ACATGTCGAT.....AGGTG

TGCAC.....TCGATACATAAGGTG

TGCACACATGT...

ACATG.....TCGAT.....AGGTG

.....TGCACTCGATACATAAGGTG

TGCACACATGT...

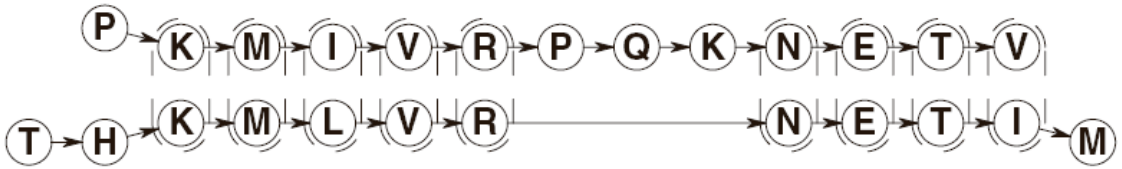
Partial Order Alignment

(a)

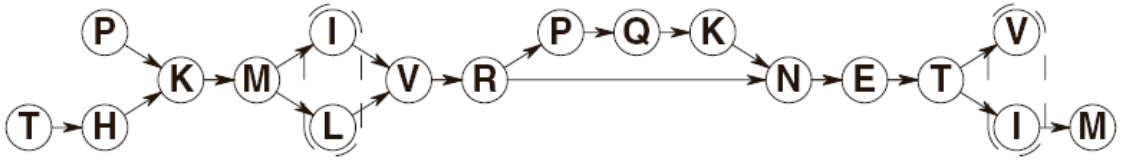
.	.	P	K	M	I	V	R	P	Q	K	N	E	T	V	.
T	H	.	K	M	L	V	R	.	.	.	N	E	T	I	M

(b)

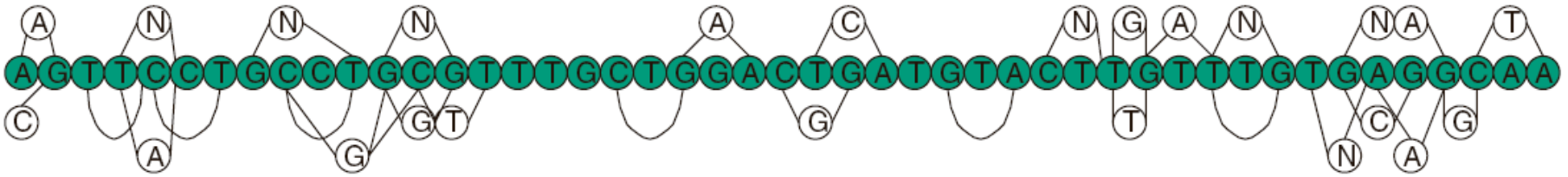
(c)



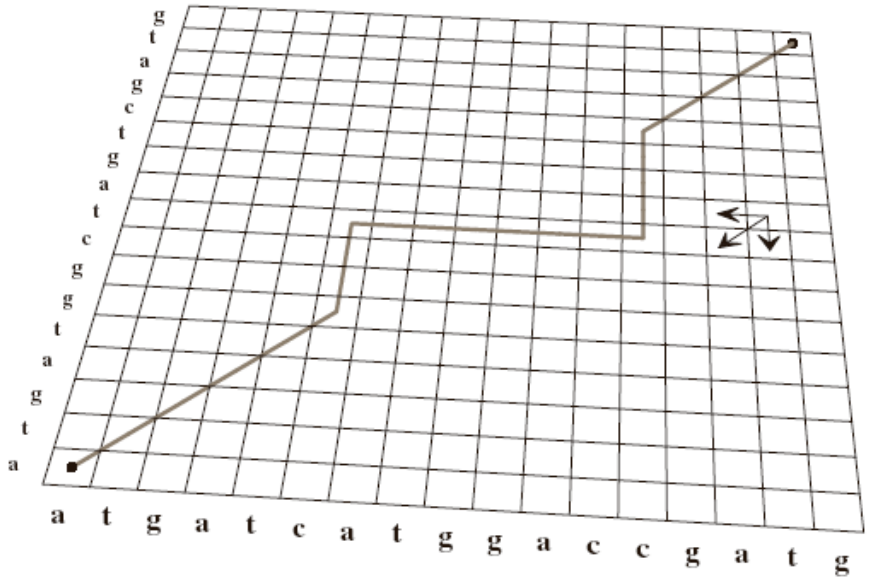
(d)



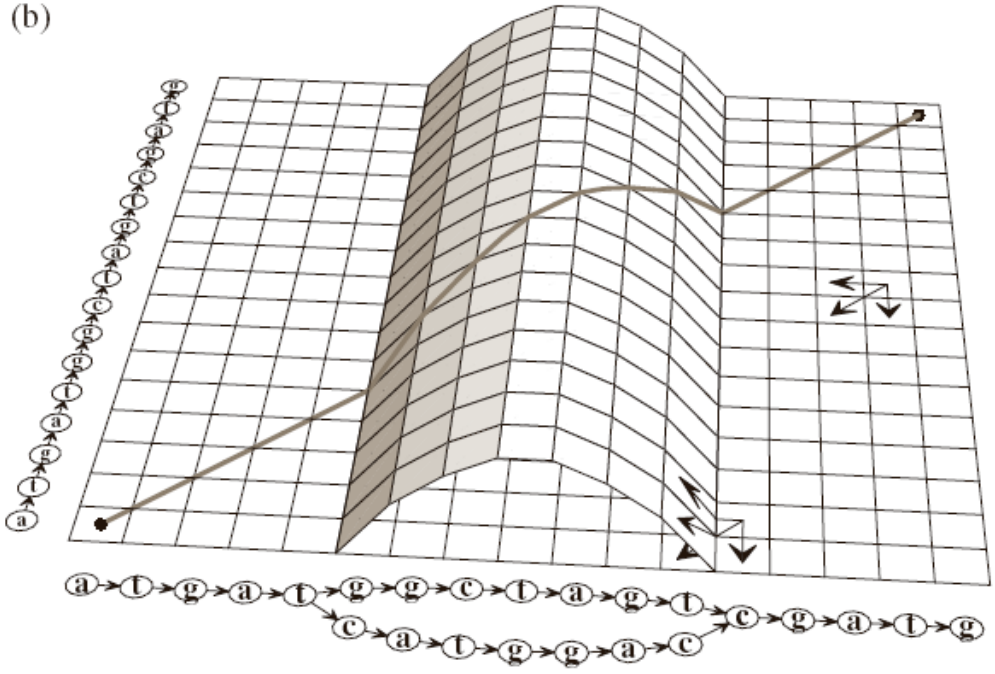
```
CONSENS1          ..........................TGTACNT.GTTTGTGAGG.CTA
CONSENS0          A.GTTCCTGC.TGCGTTTGCTGGACTGATGTACTT.GTTTGTGAGG.CAA
Hs#S663801        A.GTTCCTGC.TGCGTTTGCTGGACTTATGTACTT.GTTTGTGAGG.CAA
Hs#S337687        AAGTTCCTGC.TGCGTTTGCTGGACTGATGTACTTGGTTTGTGNAGGCAA
Hs#S629177        A.GTTCCTGC.TGCGTTTGCTGGACTGATGTACTT.GTTTGTNAGG.CAA
Hs#S672957        A.GTTCCTGC.TGCGTTTGCT......................
Hs#S672182        A.GTTCCTGC.TGCGTTTGCTGGACTGATGTACTT.GTTT.......
Hs#S674099        A.GTTCCTGC.TGCGTTTGCTGGACTGATGTACTT.GTTTGTGAGG.CAA
Hs#S196113        A.GTTNCTGN TGNGTTTGCTGGACTGATGTACTT.GTTTGTGAGG.CAA
Hs#S994400        ..........................GTACNT.GTTTGTGAGG.CTA
Hs#S550772        A.GTTCCTGC.TGCGTTTGCTGGACTGATGTACTT.GTTTGTGAGG.CAA
Hs#S80460         A.GTTCCTGC.TGCGTTTGCTGGACTGATGTACTT.GTTTGTGAGG.CAA
Hs#S39701         A.GTTCCTGC.TGCGTTTGCTGGACTGATGTACTT.GTTTGTGAGG.CAA
Hs#S1988018       A.GTTCCTGC.TGCTTTTGCTGGACTGATGTACTT.GATTGTGAGG.CAA
Hs#S341915        A.GTTCCTGC.TGCGTTTGCTGGACTGATGTACTT.GTTTGTGAGG.CAA
Hs#S1794113       A.GTTCCTGC.TGCGCTTGCTGGACTGATGTACTT.GTTTGTGAGG.CAA
Hs#S4698          A.GTTCCTGC.TGCGTTTGCTGGACTGATGTACTT.GTTTGTGCGG.CAA
Hs#S813765        A.GT.CCTGC.G.CGTTTGC.GGACGGATGTACTT.GTT.GTGAGG.CAA
Hs#S1184845       ..........................................G.CAA
Hs#S1577463       ..........................................GG.CAA
Hs#S914987        ......................CTGATGTACTT.GTT.GTGAGGGCAA
Hs#S1985364       A.GTTCCTGC.TGCGTTTGCTGGACTGATGTACTT.GTTTGTGAGG.CAA
Hs#S1465644       ..GTTC.TGCCTGCGTTTGCTGAACTGATGTACTT.GTTAGT.AAG.CAA
Hs#S1850471       C.GTTACTGC.GGGGTTTGCTGGACTCATG.ACTTTGTTNGT.AGG.CAA
```



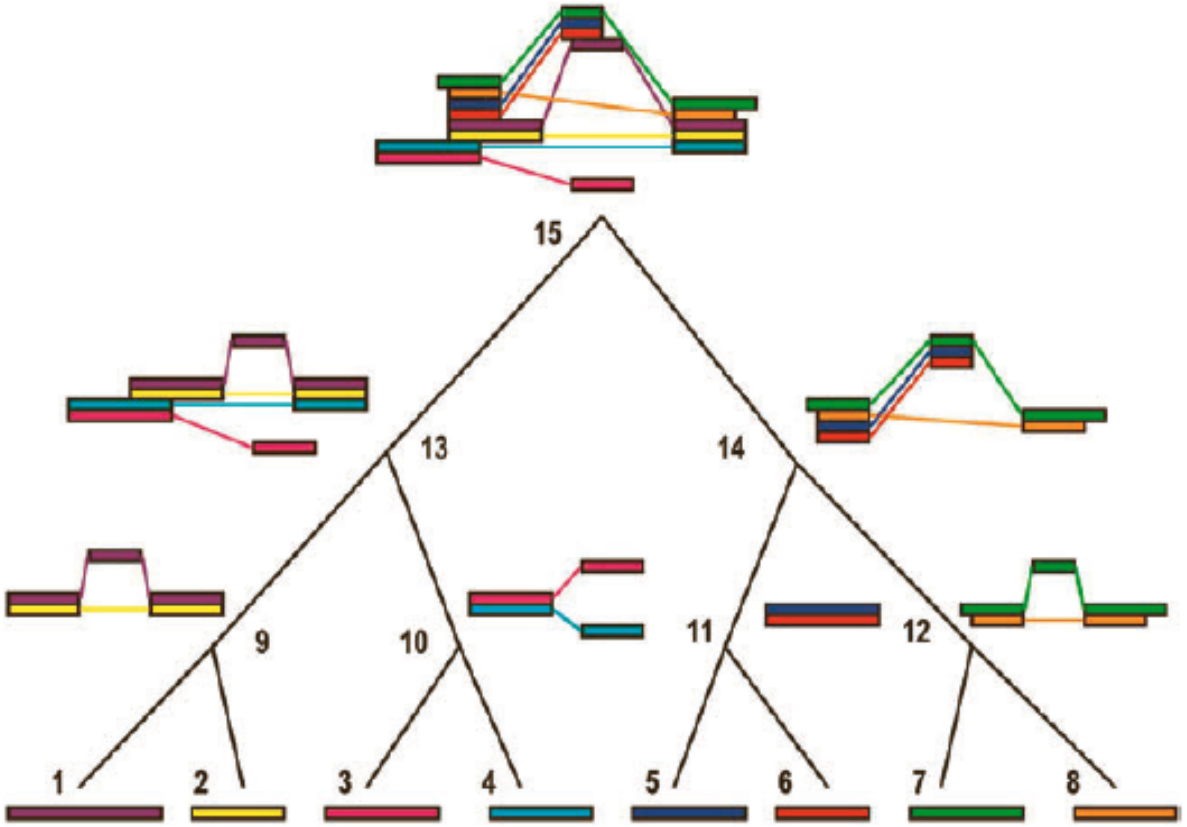
(a)


g t a g c t g a t c g g t a a g t a
a t g a t c a t g g a c c g a t g

(b)


a t g a t g g c t a g t c g a t g
c a t g g a c

Progressive Aligment



一些网址

- BLAST
  - http://www.ncbi.nlm.nih.gov/BLAST/
- EMBOSS (Pairwise Alignment):
  - http://www.ebi.ac.uk/emboss/align/
  - http://emboss.sourceforge.net/
- ClustalW (Multiple Alignment):
  - http://www.ebi.ac.uk/clustalw/

生物信息学

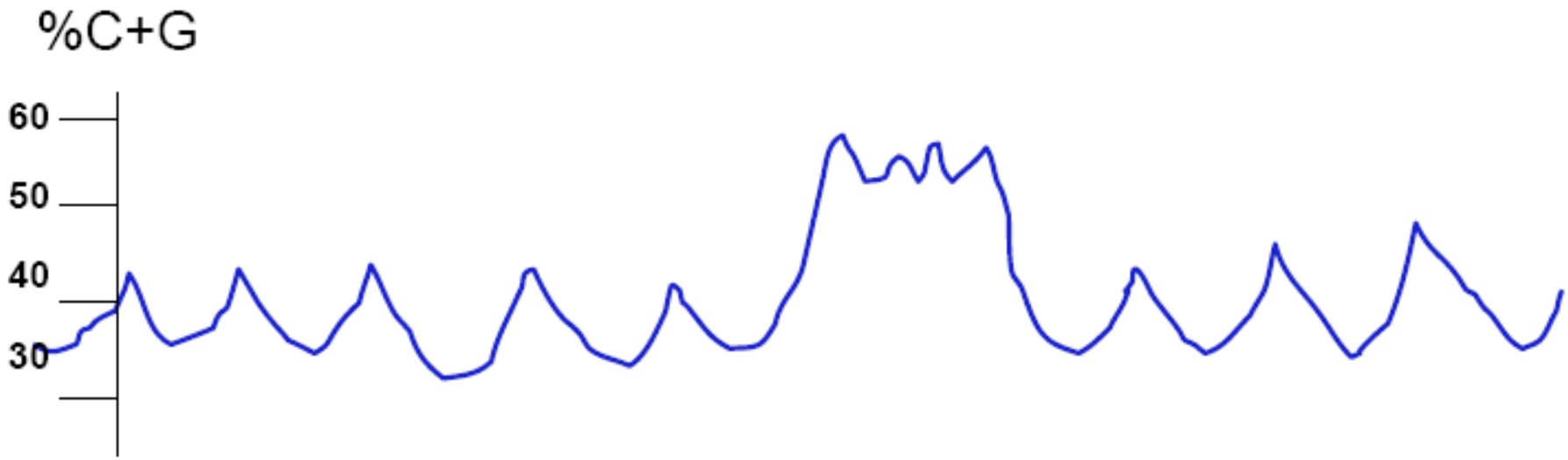
基因识别

吴凌云
中国科学院数学与系统科学研究院

Gene Finding

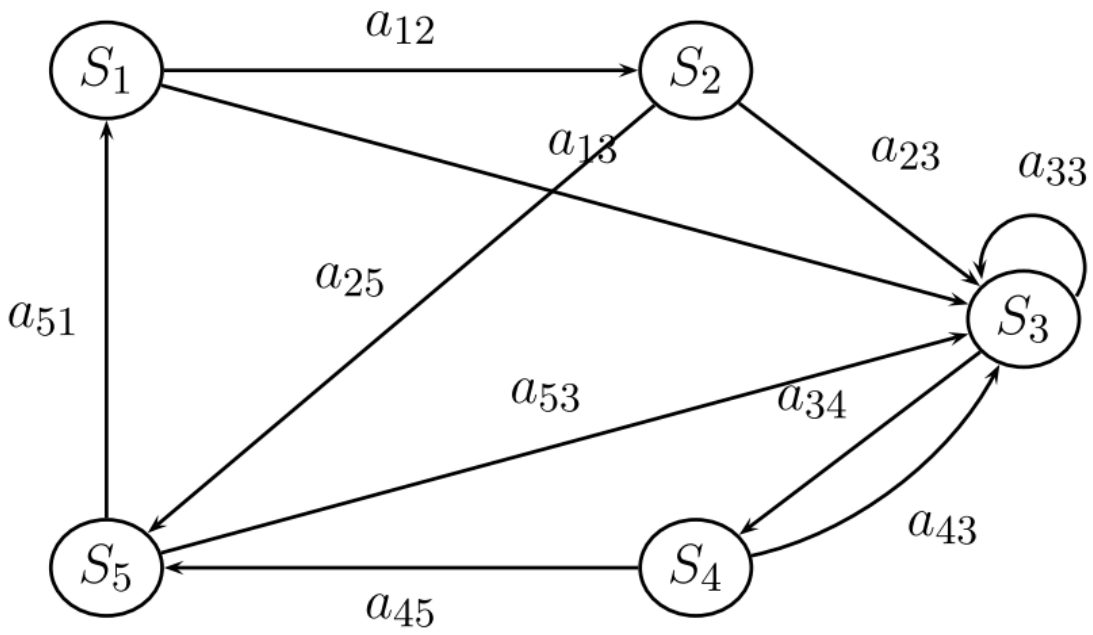
- Homology sequences (BLAST)
- Statistical difference of sequences
- Gene Prediction (HMM)

CpG Island



大约70%的人类启动子区域有很高的CpG含量.

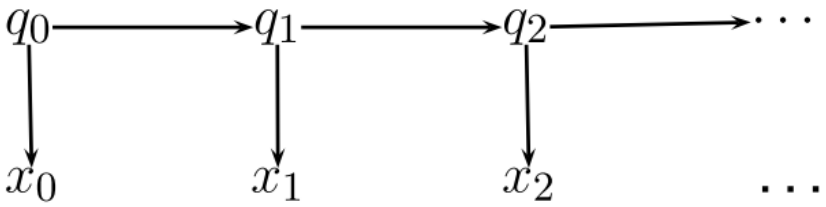
离散Markov过程



- States $S = \{S_1, \ldots, S_N\}$
- Initial state distribution $\pi_i = P(q_1 = S_i)$
- Transition coefficients $a_{ij} = P(q_t = S_j | q_{t-1} = S_i) \geq 0$, such that $\sum_{j=1}^{N} a_{ij} = 1$

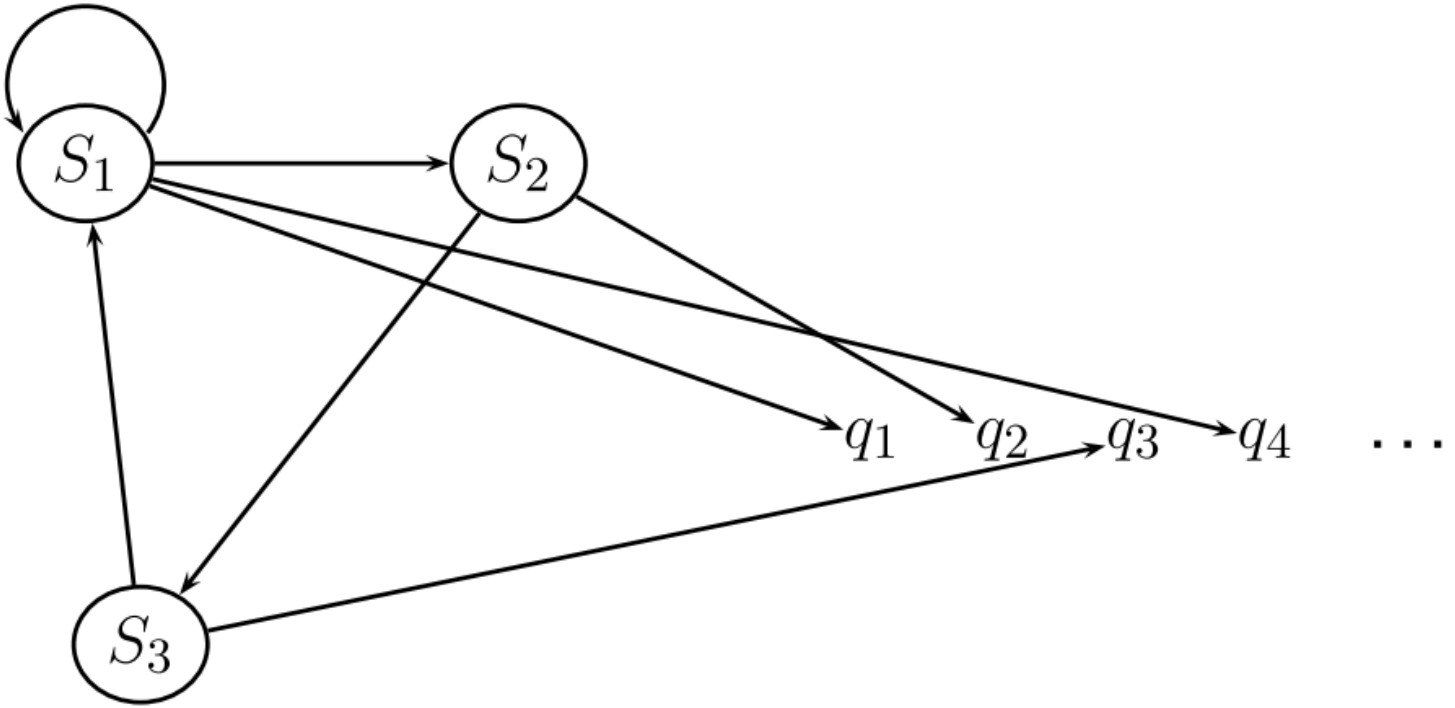
Hidden Markov Model

Every node generates an output symbol x_i; we observe x_i but not q_i.



- States $S = \{S_1, \ldots, S_N\}$
- Initial state distribution $\pi = \{\pi_i = P(q_1 = S_i)\}$
- Transition coefficients $a_{ij} = P(q_t = S_j | q_{t-1} = S_i) \geq 0$, such that $\sum_{j=1}^{N} a_{ij} = 1$
- Alphabet $V = \{v_1, \ldots, v_M\}$
- Observation probabilities $b_j(k) = P(x_t = v_k | q_t = S_j)$

HMM



HMM的三大问题

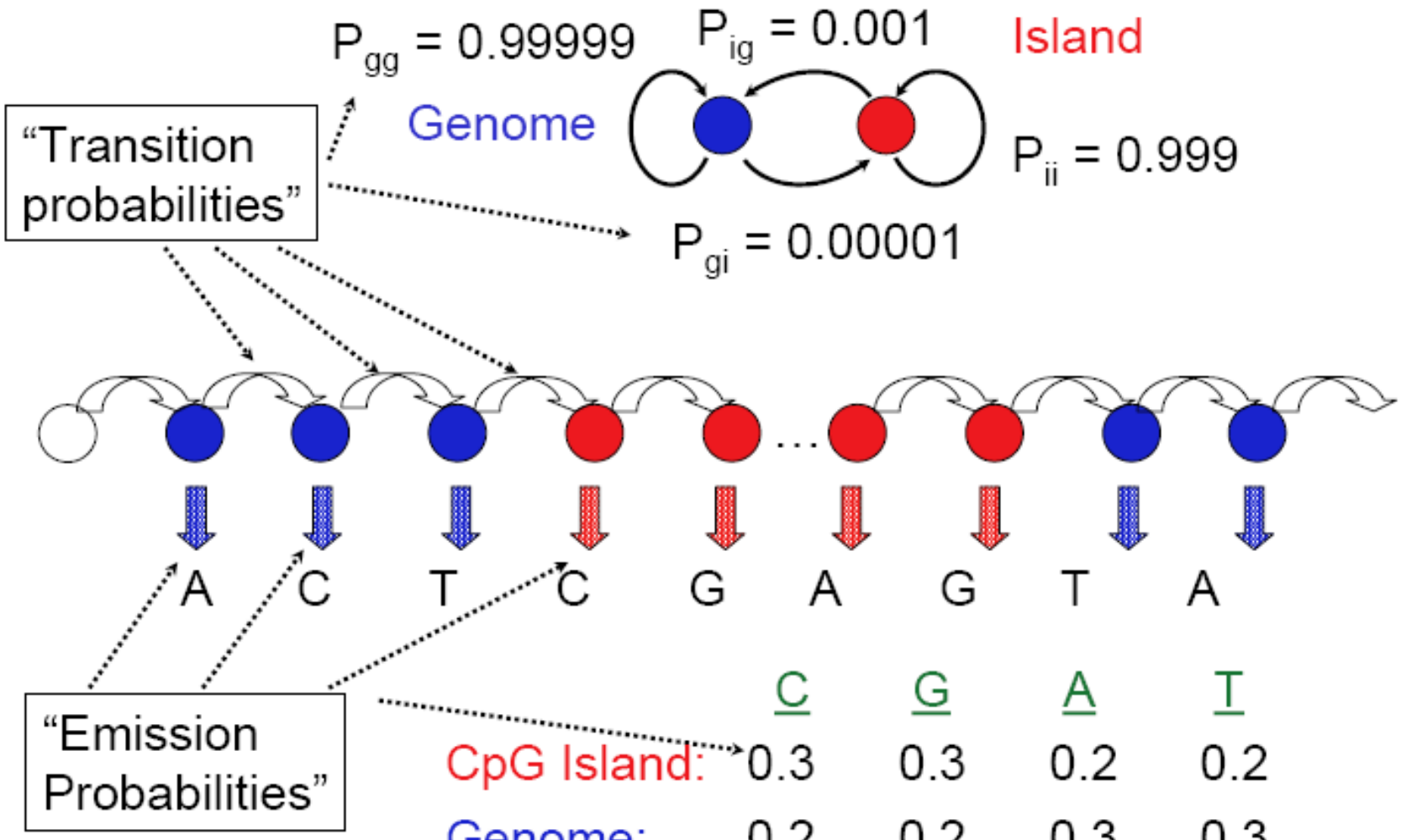
Consider

- a Hidden Markov Model (HMM) $\lambda = (A, B, \pi)$
- observation sequence $X = x_0, \ldots, x_T$

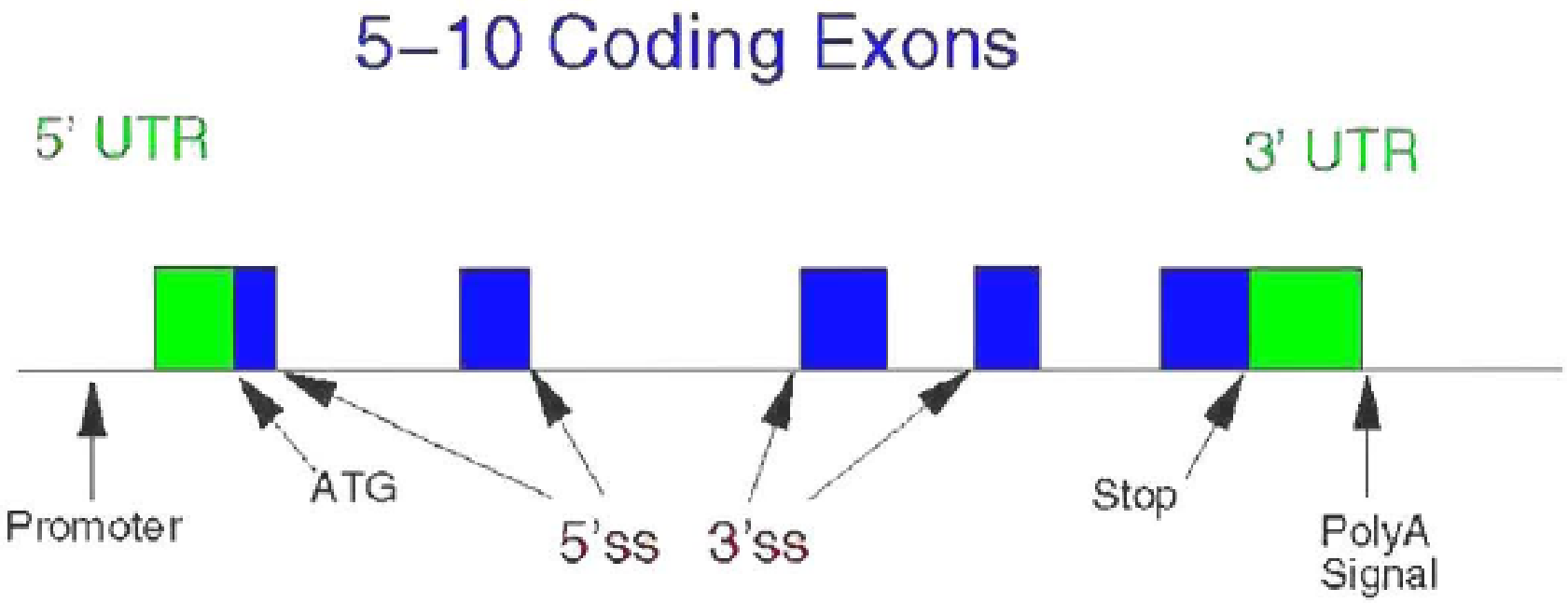
The following problems will arise.

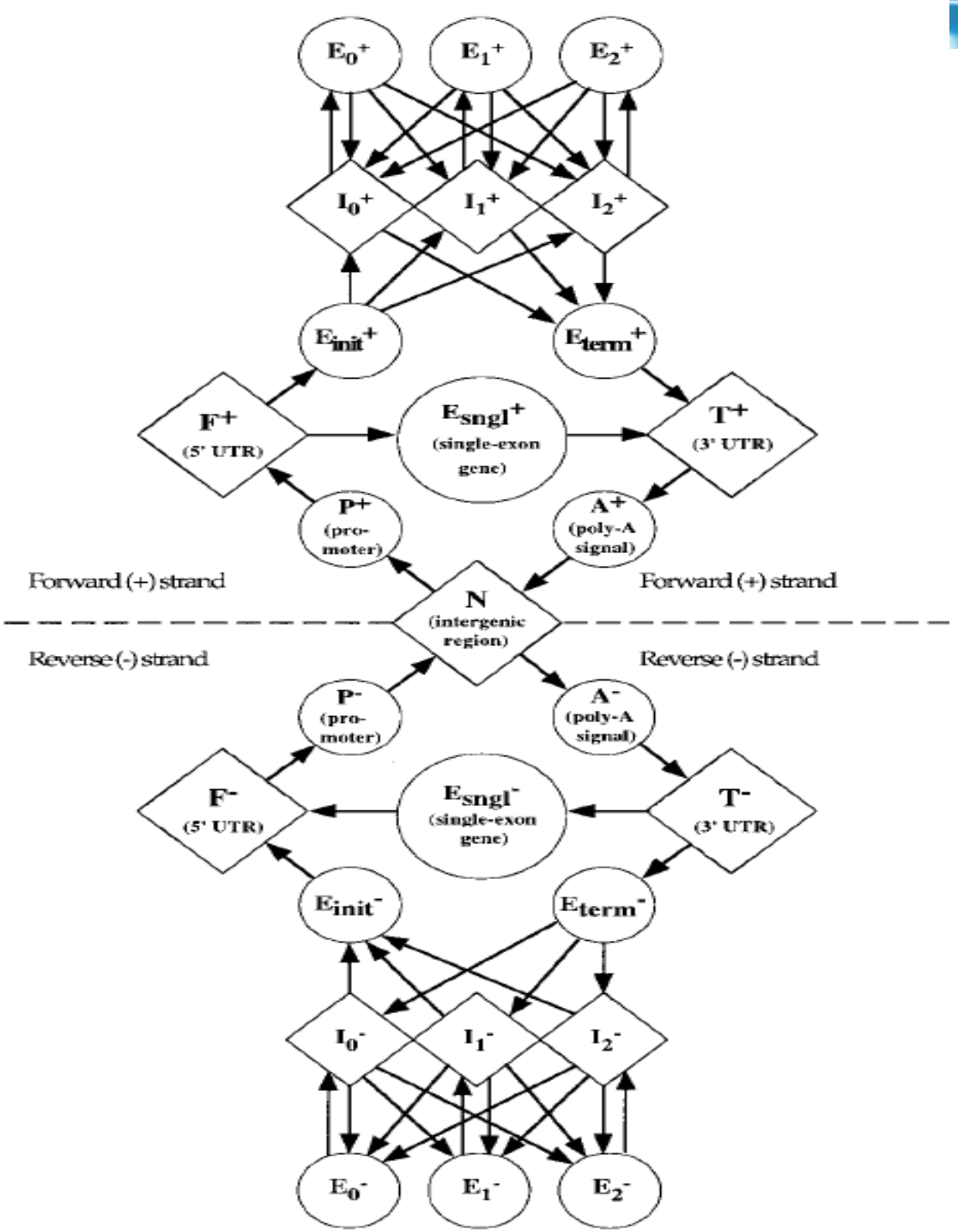
- **Evaluation Problem.** Given X and λ, compute $P(X|\lambda)$
- **Decoding Problem.** Given X and λ, choose an "optimal" state sequence corresponding to X
- **Learning Problem.** Given X, adjust λ to maximize $P(X|\lambda)$

HMM for CpG Island



Structure of Typical Gene




E0+
E1+
E2+
I0+
I1+
I2+
Einit+
Eterm+
F+
(5' UTR)
Esngl+
(single-exon gene)
T+
(3' UTR)
P+
(pro-moter)
A+
(poly-A signal)
Forward (+) strand
Forward (+) strand
N
(intergenic region)
Reverse (-) strand
Reverse (-) strand
P-
(pro-moter)
A-
(poly-A signal)
F-
(5' UTR)
Esngl-
(single-exon gene)
T-
(3' UTR)
Einit-
Eterm-
I0-
I1-
I2-
E0-
E1-
E2-

Genescan

Accuracy on Vertebrate Genes

Accuracy per nucleotide			Accuracy per exon					
Method	Sn	Sp	AC	Sn	Sp	(Sn+Sp)/2	ME	WE
GENSCAN	0.93	0.93	0.91	0.78	0.81	0.80	0.09	0.05
FGENEH	0.77	0.85	0.78	0.61	0.61	0.61	0.15	0.11
GeneID	0.63	0.81	0.67	0.44	0.45	0.45	0.28	0.24
GeneParser2	0.66	0.79	0.66	0.35	0.39	0.37	0.29	0.17
GenLang	0.72	0.75	0.69	0.50	0.49	0.50	0.21	0.21
GRAILII	0.72	0.84	0.75	0.36	0.41	0.38	0.25	0.10
SORFIND	0.71	0.85	0.73	0.42	0.47	0.45	0.24	0.14
Xpound	0.61	0.82	0.68	0.15	0.17	0.16	0.32	0.13

生物信息学

Motif预测

吴凌云
中国科学院数学与系统科学研究院

What are Motifs

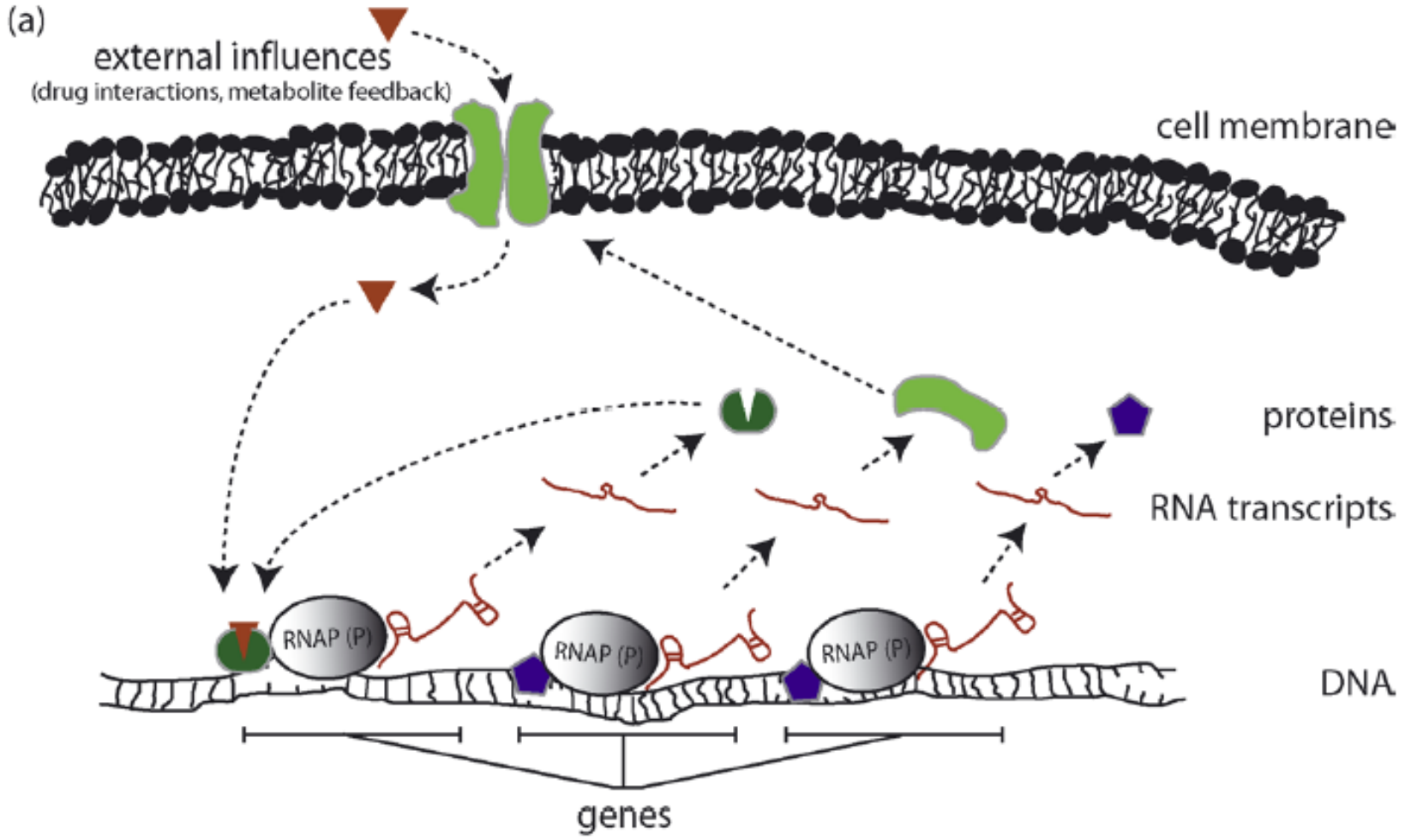
- Motifs: a recurring element in bio-sequence or structure
- A motif typically has functional implications since it is preserved (recurring) during evolution
- There are different types of motifs in biological data
  - sequence motifs
  - structure motifs
  - network motifs
  - …

Sequence Motif

- DNA sequence motifs: they generally function as **regulatory elements** in biological systems
  - transcription regulatory motifs

- Protein sequence motifs: they are generally **functional sites**

Transcription Regulation



Transcription Factors

- Different classes of transcription factors **regulate** the on and off of transcription as well as the efficiency
  - activators (general or specific)
  - enhancers (general or specific)
  - repressors (general or specific)
  - inhibitors (general or specific)
  - specificity regulators

Transcription Factors

- A transcription factor could regulate multiple genes

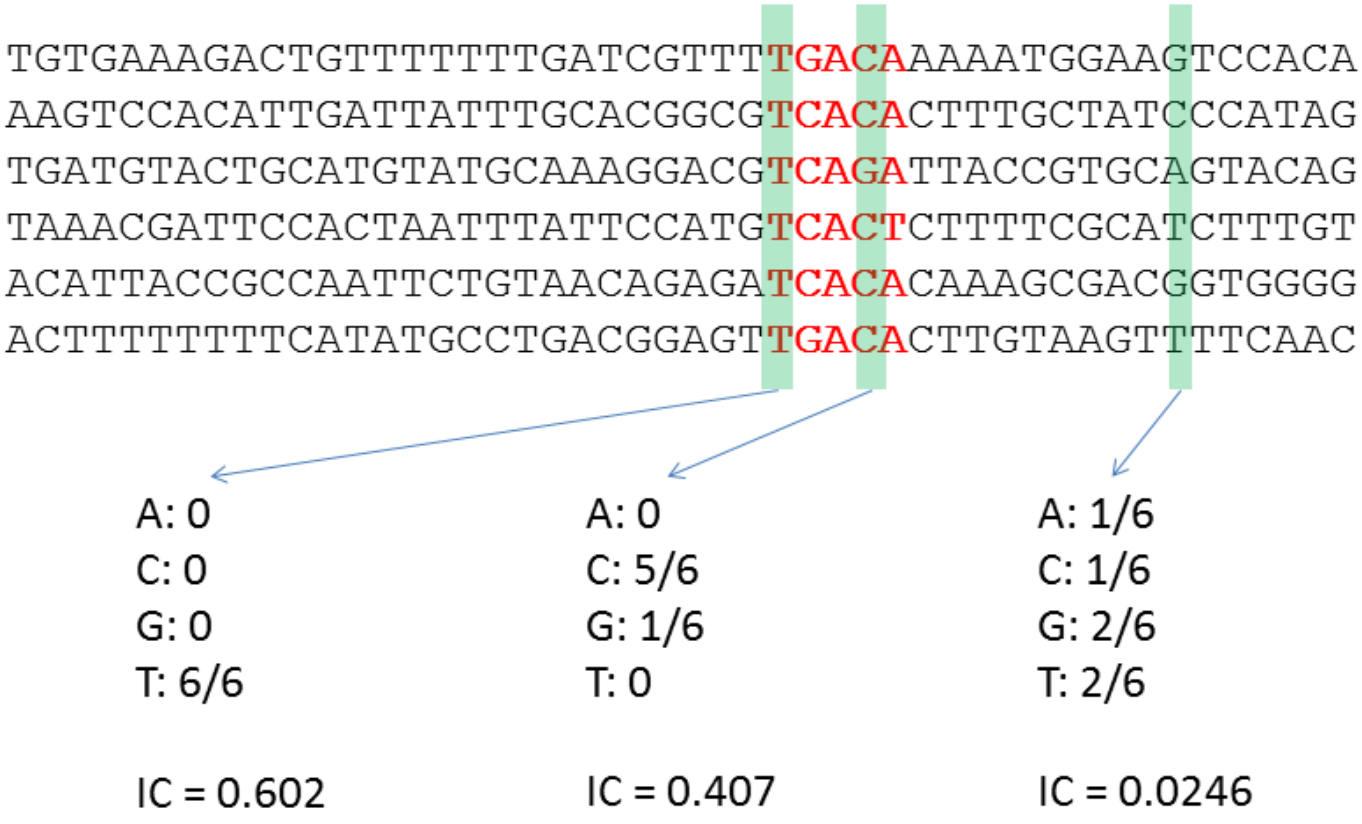
- Knowing the genes regulated by a transcription factor can help to elucidate a biological process responsible for a complex task – a systems biology view

Regulatory Binding Site

- In a genome, find genes that are transcriptionally (co-)regulated by the same transcription factor

- Identify genes sharing “common” regulatory binding sites
  - Binding sites of the same transcription factor does not have to be exactly the same in their sequences; rather they should be “conserved”

Identify Motif



Information content: Σ F(X) log (F(X)/0.25)

Sequence Logo

Problem

- A group of sequence motifs are considered as "conserved" if their aligned positions have "high" information content
- How to find blocks of DNA that have high information content?

- Simple if the sequences are already aligned! But ……

- How to find the "conserved" sites?

Brute-force Method

- Basic idea
  - assume that we know that the length of the “conserved” sites is K (K typically from 5 to 30)
  - go through all possible combinations of K-mers, one K-mer from each sequence, and calculate the information content
  - call a particular combination a “conserved” site if the IC is high

- Too many combinations to consider!

Approaches

- Heuristic methods
- Multiple sequence alignment

- MEME Suite
  - http://meme.sdsc.edu