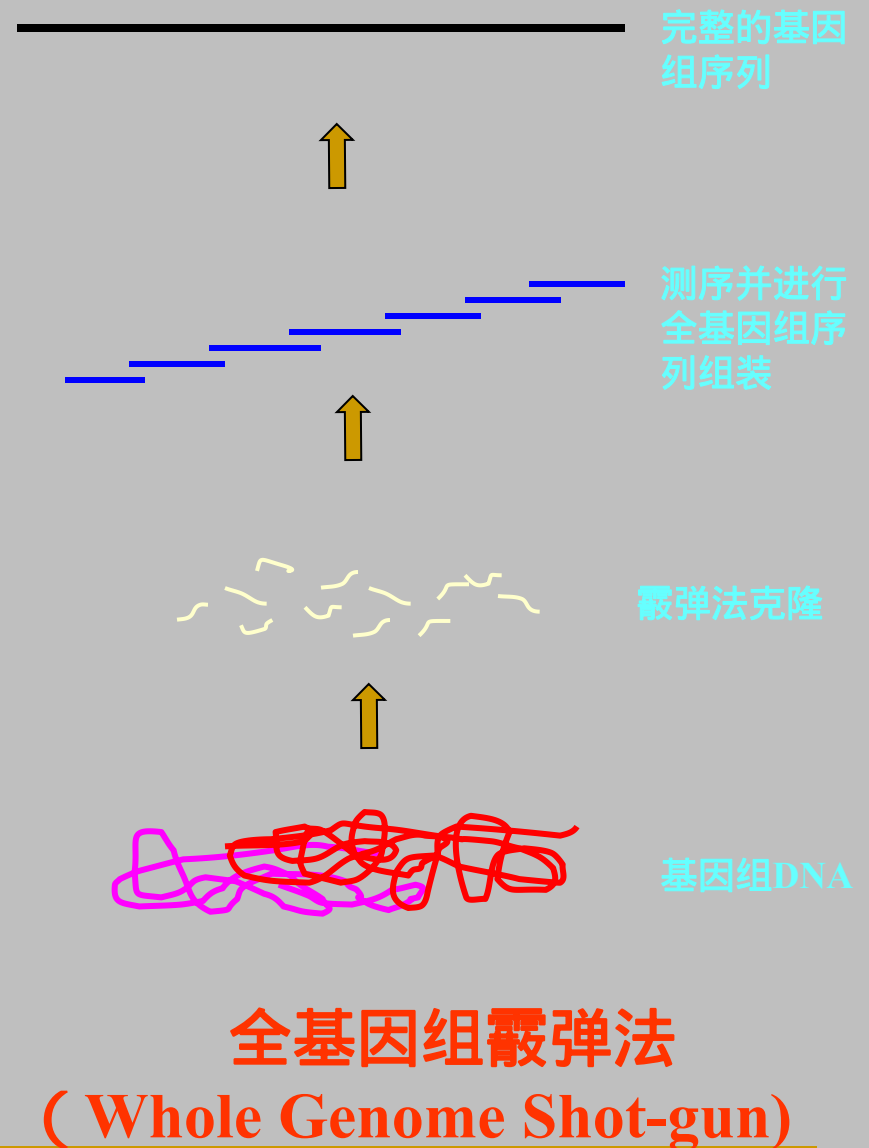
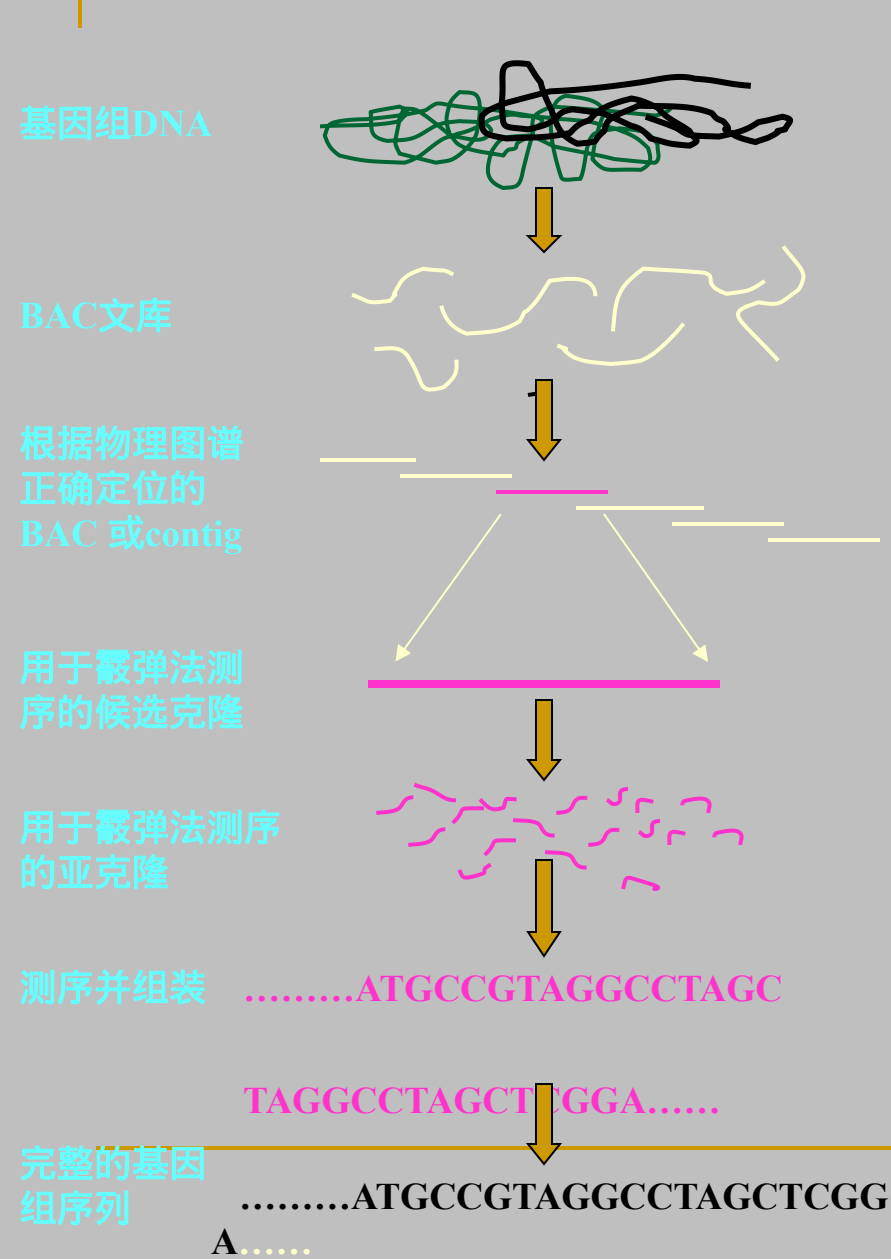


# 基因结构分析

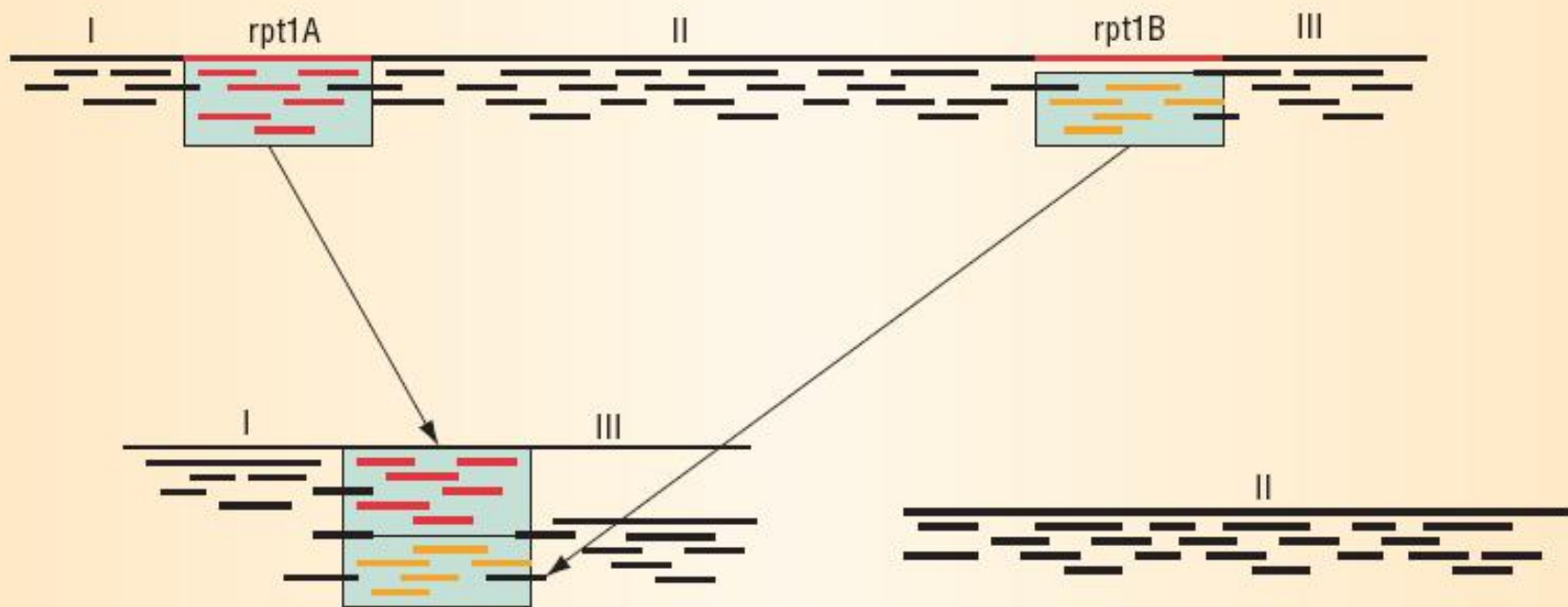
# 基因组从头测序数据分析

- ◆ 基因组拼装、统计
- ◆ 基因组注释
- ◆ 基因功能分类
- ◆ 比较基因组学及进化分析
- ◆ 建立数据库

# 逐步克隆法 (Clone by Clone)



# 重复序列带来干扰



# 基因组重测序数据分析

- ◆ 测序基本结果分析
- ◆ 基因组定位
- ◆ 单核苷酸多态性位点(SNPs)检测和功能分析
- ◆ 插入和缺失位点(Indels)检测和功能分析
- ◆ 拷贝数变异(Copy Number Variations, CNVs)检测和相关分析
- ◆ 其他基因组结构变异(Structural Variations, CVs)检测和相关分析
- ◆ 重新组装个体基因组

# 转录组测序数据分析

## ◆ 有参考基因组的转录组分析（Reference）

- 测序原始数据处理（质量剪切、污染序列去除等）
- 短序列mapping至参考基因组序列
- 基因功能注释
- 基因结构分析
- 可变剪切分析
- 基因表达差异分析
- 预测新的基因

## ◆ 无参考基因组转录组分析（De novo）

- 测序原始数据处理（质量剪切、污染序列去除等）
- Contig和Unigene统计分析
- 基因功能注释
- SNPs分析等
- 基因表达差异分析

# RNA 测序数据分析

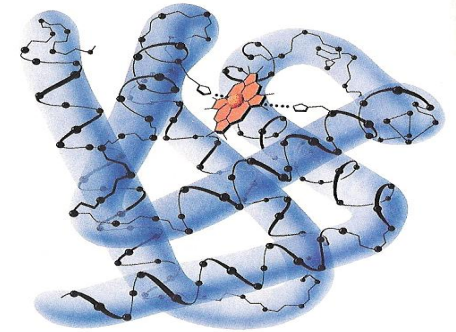
- ◆ 基因组定位
- ◆ 测序结果分类(microRNA, piRNA, tRNA, snRNA, rRNA, snoRNA 等)
- ◆ 计算已知 microRNA 的表达量
- ◆ MicroRNA的家族、基因座分类
- ◆ 预测新的 microRNA 基因
- ◆ 进化分析
- ◆ microRNA 基因簇分析

# 基因组功能分析

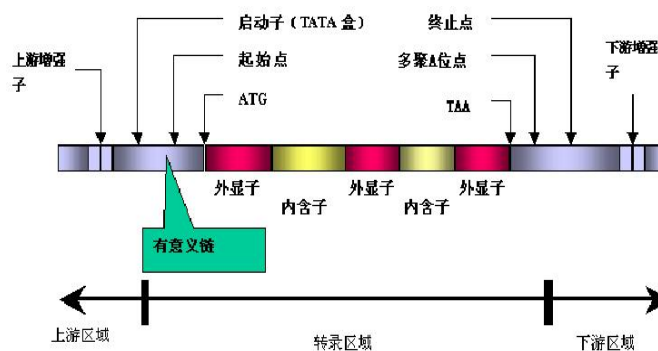
基因组序列  
cDNA序列

编码区预测 → 翻译 → 蛋白质序列 →

蛋白质理化性质  
二级结构预测  
结构域分析  
重要信号位点分析  
三级结构预测



基因结构分析



序列比对  
功能注释

KEGG

GO

系统发育树

Codon bias  
GC Content  
限制性酶切位点

选择性剪切  
转录调控因子



AAAGCCTTATTCACTTCCTTTTGTTCGCCAAATTCTCGAAATCGAAAAAGACGAGACTTTTAA  
CACTCAGACCTAAACCAACACTCTTCCTCTTTCTATCTCTCTCTCGCCGATCGGAAGCAGATT  
TTTCTCAAAAGGATGGATTCAACGAAGCTTAGTGAGCTAAAGGTCTTCATCGATCAATGCAAG  
TCTGACCCTTCCCTTCTCACTACTCCTTCACTCTCCTTCTTCCGTGACTATCTCGAGAGGTATA  
TATTTTTTTGTTTGGTTCAATTGAATTCGTTTGTGTGTTTTGCAGAAGAGTTTGTTTTTGATTG  
ATTGTTTCTGTTTAACAGAGGAGTTTCGTAGTGGAAGAGAGTGATGATGATATGGATGAACT  
GAAGAAGTAAAACCGAAAGTGGAGGAAGAAGAAGAGAGGATGAGATTGTTGAATCTGATGT  
AGAGCTTGAAGGAGACACTGTTGAGCCTGATAATGATCCTCCTCAGAAGGTACCATGATGACT  
AAGCAAAACCTTAGCTTTTGATTTTACTAATTTCTGTCTGATTGTGTGTTTCTGACTCTCTTGGT  
TTCGATACAGCTAGTGTCTACATTAAGTTGAAGAAGCCAAACGCTGCTATTTCGAGATGCAAAC  
GCAGCATTGGAGGTACAGTGGTCTACAATTTTCCTTGAGAATTTTGTTTTATGGTGATGATATT  
CAATCGTTTTGTGTTTATGATCACAGATTAACCTGATTCTGCCAAGGGATACAAGTCACGAGG  
TATGGCTCGTGCCATGCTTGAGAAATCGGCAGAGGCTGCAAAAGACCTTCACCTTGCATCTA  
CGATAGACTATGATGAGGAAATAGTGTTGTTCTCAAAAGGATTCATGTGTTTGTCTATATGT  
TTAATCACTTTTGATTCATTAACAATGCATAAGCGTGATTCTTATACAAACATTTTTTTGAAATTTG  
TAGGTTGAACCTAATGCACATAAGCTTGAGGAGCACCGTAGAAAGTATGACAGATTACGTAAG  
GAAAGAGAGGACAAAAAGGCTGAACGGGATAGATTACGTCGCCGTGCTGAAGCACAGGTAG  
CTTCATAGACCATGAAATCAATCAGAATAATTTTTGATAACATTAGTCAGTTTTTGATTGACCTT  
ATCACTTAGCTATAGACATGGGTTTTGTTTATATAATGTTGGGCTTATCCATATCTCTTTTTCTGT  
AGATCCTGAGCTAATGACGGCATTTAGCGACCCTGAAGTCATGGCTGCTCTTCAAGATGGTAT  
GATGACTCAGTTTAAAGTTGCAGTTTATTAATGTTTTTTGAATATTTTTTCGAGTCTTGAATGTTGT  
TTGAATGCTATAACAGTGATGAAGAACCCTGCGAATCTAGCGAAGCATCAGGCGAATCCGAAG  
GTGGCTCCCGTGATTGCAAAGATGATGGGCAAATTTGCAGGACCTCAGTAAACAAAACAAGA  
AGCTTGCTTTTTCTTTGCCAATTTCTGTGTTTAATTGCGTGAGATAAGAGATATGTTGGAGAACT  
TTTGTTTTCTTTTATGTTGTCGTTGCAGAGGAACTTTAACAGGAACAAAACCTTTTTCTCTTCG  
TTAGTAATCTACCCTCTTCTCGTTTTTCACTTCCTGAGTTAGAAGATTTATATTGAAAGATTCGTA  
TAAGTATAACACTTCCAACATTGTTTTTATGCGTCGTGAGA

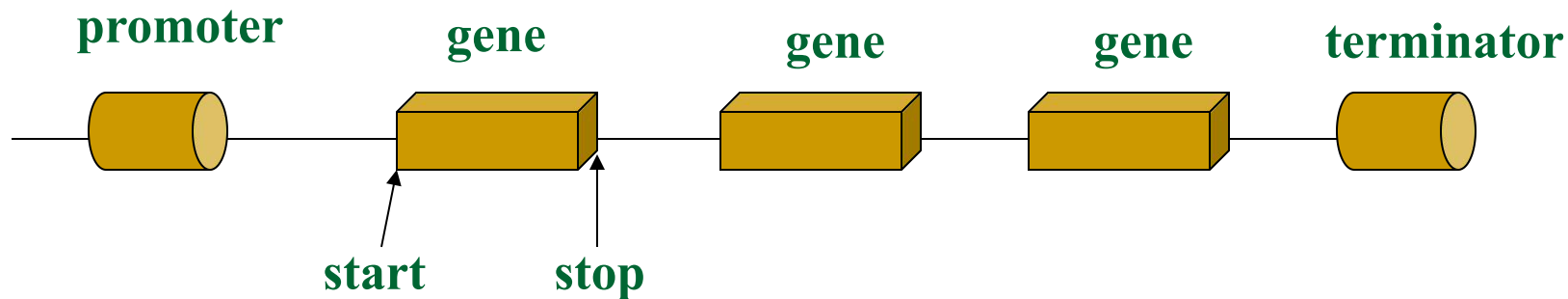
AAAGCCTTATTCACTTCCTTTTGTTCGCCAAATTCTCGAAATCGAAAAAGACGAGACTTTTAA  
CACTCAGACCTAAACCAACACTCTTCCTCTTTCTATCTCTCTCGCCGATCGGAAGCAGATT  
TTTCTCAAAAGGATGGATTCAACGAAGCTTAGTGAGCTAAAGGTCTTCATCGATCAATGCAAG  
TCTGACCCTTCCCTTCTCACTACTCCTTCACTCTCCTTCTTCCGTGACTATCTCGAGAGGTATA  
TATTTTTTTGTTTGGTTCAATTGAATTCGTTTGTGTGTTTGCAGAAGAGTTTGTTTTTGATTG  
ATTGTTTCTGTTTAACAGAGGAGTTTCGTAGTGGAAGAGAGTGATGATGATATGGATGAACT  
GAAGAAGTAAAACCGAAAGTGGAGGAAGAAGAAGAAGAGGATGAGATTGTTGAATCTGATGT  
AGAGCTTGAAGGAGACACTGTTGAGCCTGATAATGATCCTCCTCAGAAGGTACCATGATGACT  
AAGCAAACCTTAGCTTTTGATTTTACTAATTTCTGTCTGATTGTGTGTTTCTGACTCTCTTGGT  
TTCGATACAGCTAGTGTCTACATTAAGTTGAAGAAGCCAAACGCTGCTATTCGAGATGCAAAC  
GCAGCATTGGAGGTACAGTGGTCTACAATTTCCTTGAGAATTTTGTTTTATGGTGATGATATT  
CAATCGTTTTGTGTTTATGATCACAGATTAAACCCTGATTCTGCCAAGGGATACAAGTCACGAGG  
TATGGCTCGTGCCATGCTTGGAGAATGGGCAGAGGCTGCAAAGACCTTCACCTTGCATCTA  
CGATAGACTATGATGAGGAAATTAGTGCTGTTCTCAAAAAGGTATGCATGTGTTTGTCTATATGT  
TTAATCACTTTTGATTCATTAACAATGCATAAGCGTGATTCTTATACAAACATTTTTTTGAAATTTG  
TAGGTTGAACCTAATGCACATAAGCTTGAGGAGCACCGTAGAAAGTATGACAGATTACGTAAG  
GAAAGAGAGGACAAAAAGGCTGAACGGGATAGATTACGTCGCCGTGCTGAAGCACAGGTAG  
CTTCATAGACCATGAAATCAATCAGAATAATTTTTGATAACATTAGTCAGTTTTGTATTGACCTT  
ATCACTTAGCTATAGACATGGGTTTTGTTTATATAATGTTGGGCTTATCCATATCTCTTTTTCTGT  
AGATCCTGAGCTAATGACGGCATTTAGCGACCCTGAAGTCATGGCTGCTCTTCAAGATGGTAT  
GATGACTCAGTTTAAAGTTGCAGTTTATTAATGTTTTTTGAATATTTTTTCGAGTCTTGAATGTTGT  
TTGAATGCTATAACAGTGATGAAGAACCCTGCGAATCTAGCGAAGCATCAGGCGAATCCGAAG  
GTGGCTCCCGTGATTGCAAAGATGATGGGCAAATTTGCAGGACCTCAGTAAACAAAACAAGA  
AGCTTGCTTTTCTTTGCCAATTTCTGTGTTTAATTGCGTGAGATAAGAGATATGTTGGAGAACT  
TTTGTTTTCTTTTATGTTGTCGTTGCAGAGGAACTTAACAGGAACAAACTCTTTTCTCTTCG  
TTAGTAATCTACCCTCTTCTCGTTTTTCACTTCCTGAGTTAGAAGATTTATATTGAAAGATTCGTA  
TAAGTATAACACTTCCAACATTGTTTTTATGCGTCGTGAGA

# 基因识别

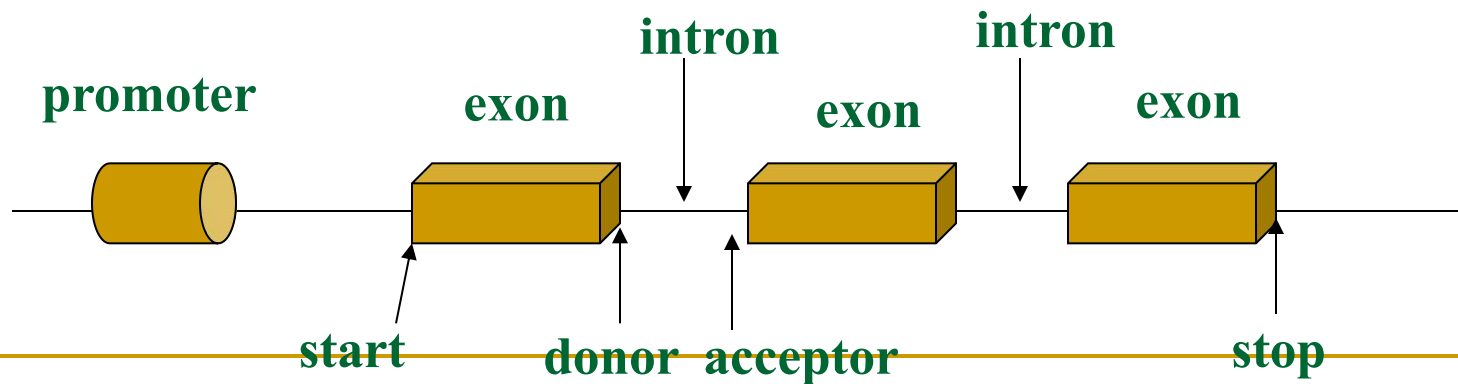
基因识别——使用计算机手段识别**DNA**序列上的具有生物学特征的片段，其对象主要是蛋白质编码基因，也包括其他具有一定生物学功能的因子，如**RNA**、**MicroRNA**基因等一些非编码基因，基因识别是生物信息学领域里的一个重要研究内容。

# 基因结构回顾

## 原核基因



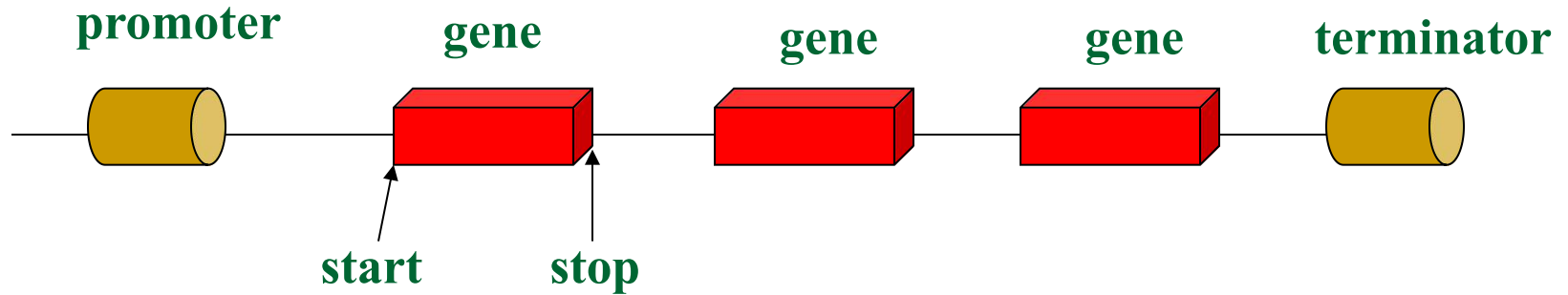
## 真核基因



# 原核基因识别

# 原核基因基本结构识别

原核基因是DNA分子的一个片段，具有连续编码的结构特征。



原核基因识别——重点在于识别编码区域。

- ◆ 对于核酸序列（不考虑互补链），根据密码子的起始位置，可以按照三种方式进行解释。

例如，序列ATTCTGATCGCAA

(1) ATT CGA TCG CAA

(2) TTC GAT CGC AAN

(3) TCG ATC GCA ANN

- ◆ 这三种阅读顺序称为阅读框（reading frames）

- ◆ 开放阅读框（open reading frame, ORF）是一段可编码蛋白质的核苷酸序列，其间不存在任何终止编码的密码子。



原核基因识别任务的重点是识别开放阅读框。



# 基于基因密码子特性的识别方法

- ◆ 辨别编码区与非编码区的一种方法
  - 检查终止密码子的出现频率

如一条核酸序列是均匀随机分布的，那么终止密码子出现的期望次数为：

每21个 ( $64/3$ ) 密码子出现一次终止密码子

绝大部分原核生物蛋白质的长度大于60个氨基酸  
(*E. coli*: 蛋白质编码区域平均长度为316.8个密码子，  
长度小于60个密码子的基因不到1.8%)

## ◆ 基本思想：

- 如果能够找到一个比较长的序列，其相应的密码子序列不含终止密码子，则这段序列可能就是编码区域。

## ◆ 基本方法：

- 扫描**DNA**序列，在三个不同的阅读框中寻找较长的**ORF**。遇到终止密码子以后，回头寻找起始密码子。

# SOFTWARE: NCBI ORF Finder

## ORF Finder (Open Reading Frame Finder)

Enter GI or ACCESSION

or sequence in FASTA format

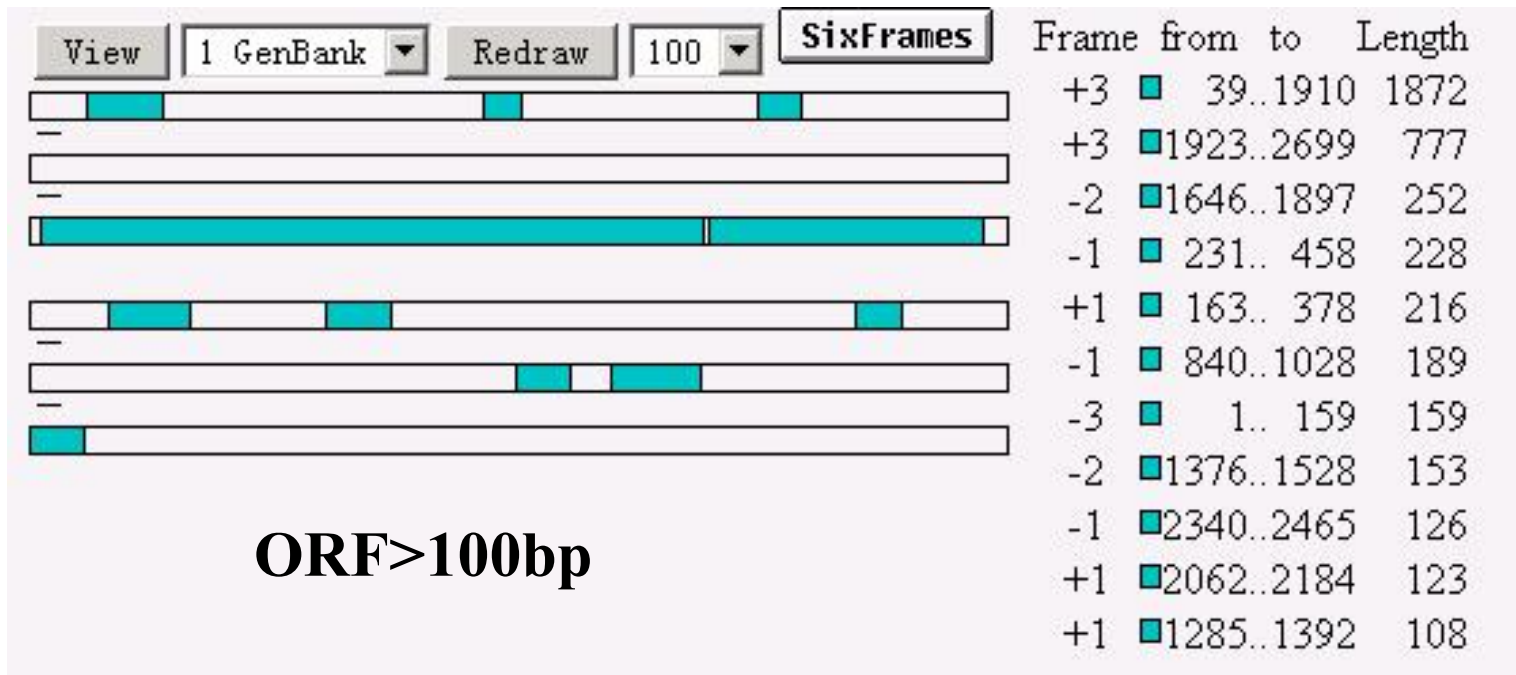


FROM:  TO:

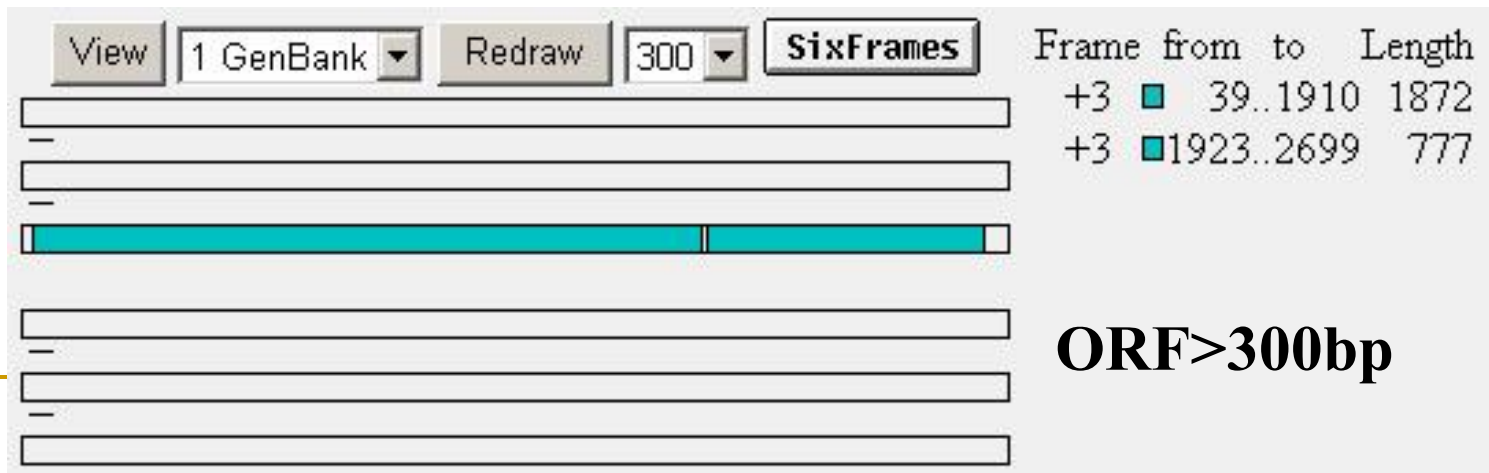
[Genetic codes](#)  

<http://www.ncbi.nlm.nih.gov/gorf/gorf.html>

# NCBI ORF Finder 分析实例



分析序列  
GenBank:  
L03845



# 改善原核基因预测方法

- ❖ 简单的**ORF Finder**只能正确地找到原核基因组大约**85%**的蛋白质编码区域，该算法过于简单，不适合于处理短的**ORF**。
- ❖ 在一些情况下不得不用更复杂的方法，比如**GeneMark**和**Glimmer**。这些情况包括：寻找很短的蛋白质；分析在不同的阅读框下有重合的**ORF**。
- ❖ 这些方法不是简单地找不间断的阅读框，同时考虑序列的统计特性计算每种可能**ORF**的概率。

# 当前著名的原核基因预测软件

## 1、GeneMark系列软件（包括最新版本GeneMarkS）

翻译起始位置预测准确率为83.2%-94.4%

——Besemer, J., Lomsadze, A. and Borodovsky, M. (2001) GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic Acids Res.*, 29: 2607-2618.

## 2、Glimmer 2.02

基因识别的准确率约为97-98%

——Delcher, A. L., Harmon, D., Kasif, S., White, O., and Salzberg, S. L. (1999) Improved microbial gene identification with GLIMMER. *Nucleic Acids Res.*, 27, 4636-4641

# 真核基因识别

# 真核生物的基因结构——断裂基因结构

真核基因远比原核基因复杂：

- 真核生物的基因一般为断裂基因（interrupted gene），基因的编码序列在DNA上不是连续的，而是被不编码的序列隔开；
- 外显子（exon）：基因中编码的序列，与成熟mRNA的序列相对应。
- 内含子（Intron）：基因中不编码的序列。
- 真核基因具有丰富的调控信息。



# 真核基因识别的主要方法

两大类识别方法：

## ◆ 基于同源序列比较的方法

- 利用已知的基因编码信息（如mRNA或蛋白质序列），通过同源比较，在DNA序列中发现潜在的编码序列。

## ◆ 基于序列特征的方法

- 根据DNA本身所具有的序列特征预测编码基因

理想的方法：综合两大类方法优点，开发混合算法。

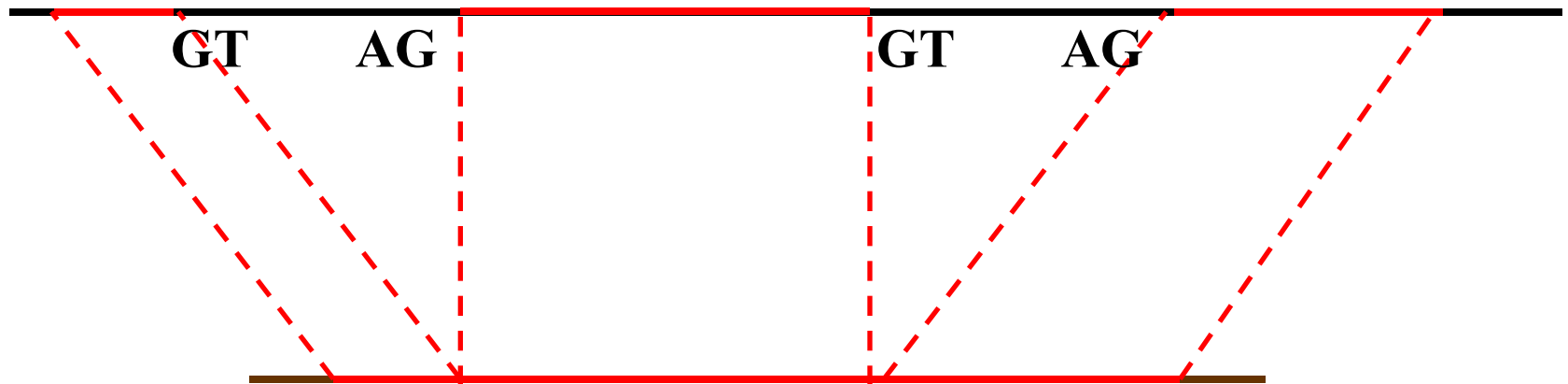
# 基于同源序列比较的方法

通过相似性比对发现编码序列

**DNA序列**

外显子

内含子



**EST, cDNA或蛋白质序列**

**DNA序列**的某一片段与**mRNA**或蛋白质序列具有高度相似性，这说明该**DNA**片段极有可能是编码序列。

# 基于同源序列比较的方法

- ◆ **BLAST**(最原始的方法，一般选用其子程序**BLASTn**和**BLASTx**进行分析，只能确定编码区的大致位置)

---

# 分析举例（BLAST）

- BLAST（<http://blast.ncbi.nlm.nih.gov/Blast.cgi>）
    - 2) 提供需要分析的DNA序列
    - 3) 针对数据库选择适合的BLAST子程序
      - a) 对EST数据库中同一生物的cDNA序列进行比较分析（如Blastn）
      - b) 在6个阅读框中进行翻译并与蛋白质数据库中的序列进行比较分析（如Blastx）
    - 4) 确定基因数目和对应的ORF
-

## 基于序列特征的方法

一般意义上基因具有两种类型的特征

- ◆ 一类特征是“信号”，由一些特殊的序列构成，通常预示着其周围存在着基因。
- ◆ 另一类特征是“内容”，即蛋白质编码基因所具有的某些统计学特征。

# 基于序列特征的方法

## ◆ 序列特征信号

- 转录启动信号
- 起始密码子
- 外显子 剪接位点（GT-AG法则）
- 终止密码子
- 转录终止信号（包含多聚A序列）
- 整个基因编码区长度一定是3的倍数

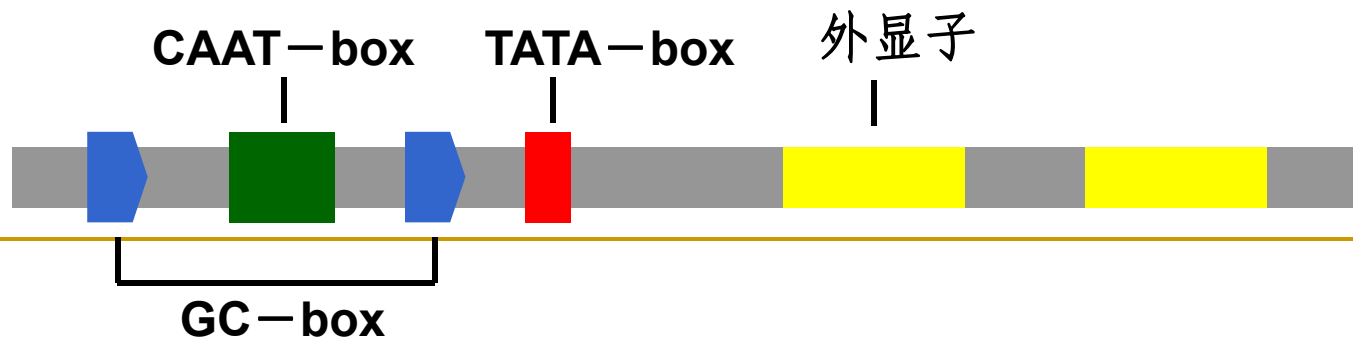
# 真核基因特征信号——转录启动区

启动子是一段特定的直接与RNA聚合酶及其转录因子相结合、决定基因转录起始与否的DNA序列。

**TATA - box:** 其一致顺序为TATAATAAT。约在基因转录起始点上游约-30-50bp处，基本上由A-T碱基对组成，是决定基因转录始的选择。

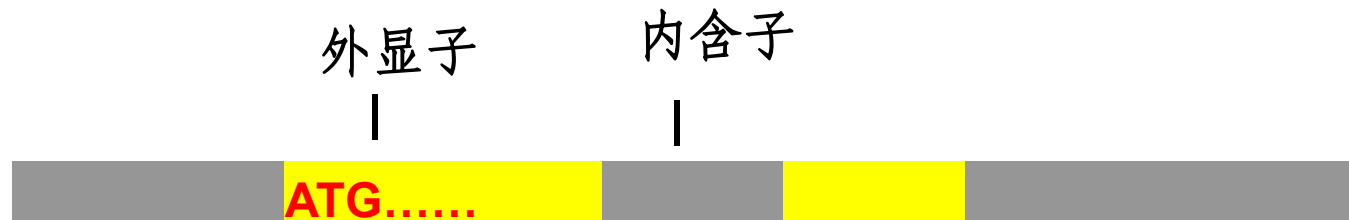
**CAAT - box:** 其一致顺序为GGGTCAATCT，是真核生物基因常有的调节区，位于转录起始点上游约-80-100bp处。

**GC - box:** 有两个拷贝，位于CAAT框的两侧，由GGCGGG组成。



# 真核基因特征信号——起始密码子

翻译起始位置：ATG



GCC**ATG**GCGA .....

ACG**ATG**CTGT ....

GAC**ATG**GTAC ...

AGG**ATG**GGCT ...

GCG**ATG**TGGC ...

并非所有的ATG都是起始密码子



# 真核基因特征信号——起始密码子

起始密码子两侧位置碱基的出现具有偏爱性



下面哪个序列更像翻译起始位置？

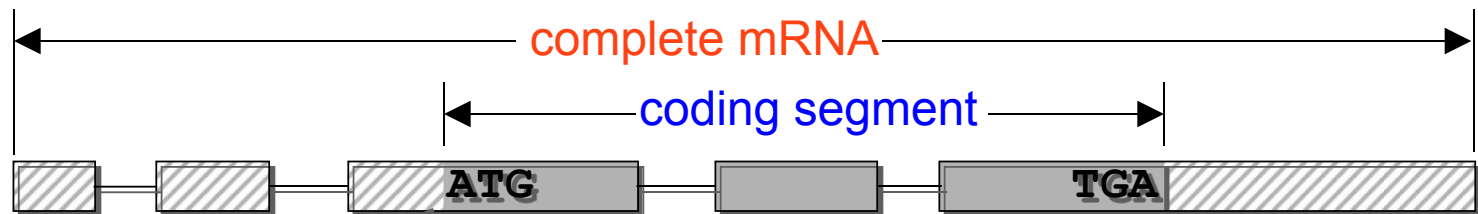
**CACC ATG GCG**

**TCGA ATG TTA**

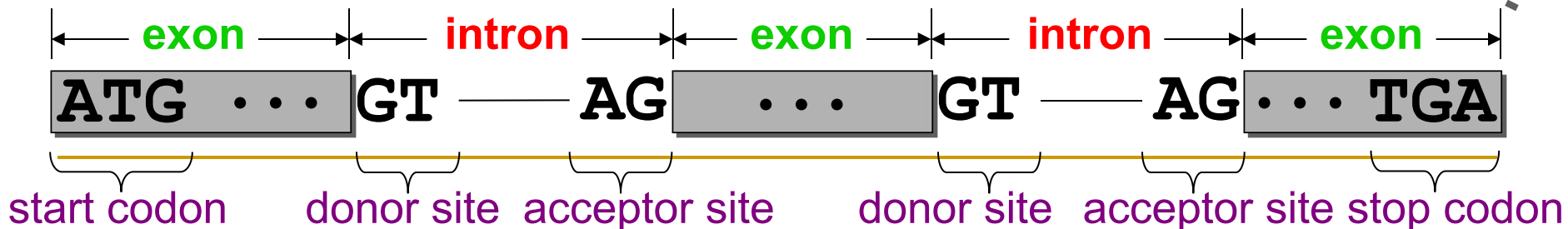
# 真核基因特征信号——剪接位点

## ❖ 外显子与内含子接头——“GT-AG法则”

外显子与内含子相连部位通常是一段高度保守的特定序列，即内含子5'端都是GT开始（供体位点），3'端都是AG结尾（受体位点），这种接头方式称为“GT-AG法则”。

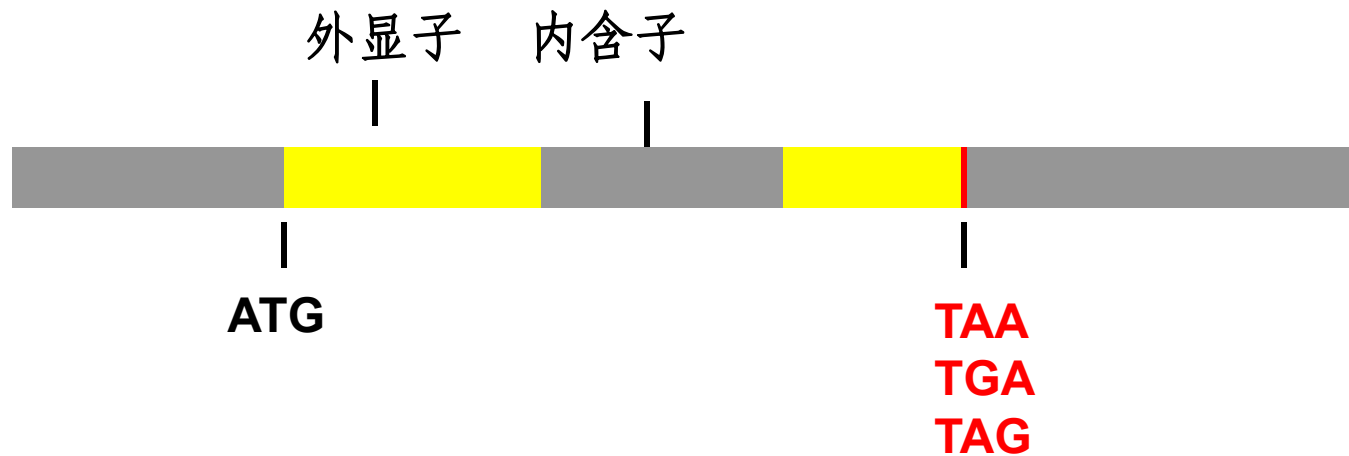


RNA中对应剪接的信号为GU-AG



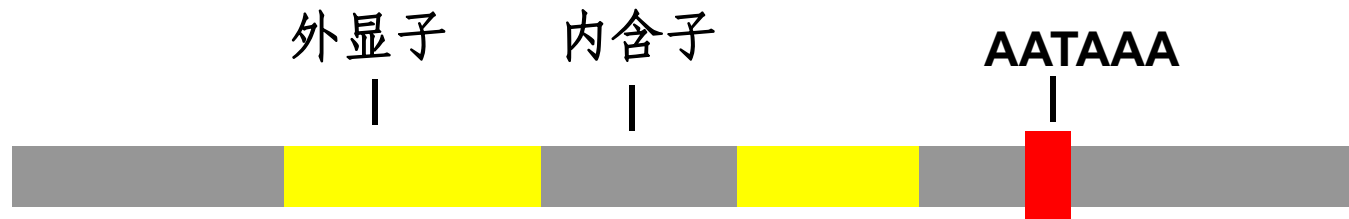
# 真核基因特征信号——终止密码子

终止密码子：TAA，TGA，TAG



# 真核基因特征信号——转录终止信号

转录终止子（Terminator）：由一段回文序列以及特定的序列AATAAA（或ATTAAA）组成。



# 基于序列特征的方法

- ◆ 编码序列统计学特征
  - 密码子使用偏爱性
  - 双联密码子出现频率

# 密码子使用偏爱性

密码子简并性：同一种氨基酸具有两个或更多个密码子编码的现象。

密码子偏爱性：兼并密码子的使用频率并不均等，各种生物编码某种氨基酸往往较多地采用其中一种。

遗 传 密 码 表

		第二 位					
		U	C	A	G		
第一 位 (5' 端)	U	UUU } phe UUC } UUA } leu UUG }	UCU } UCC } ser UCA } UCG }	UAU } tyr UAC } UAA 终止 UAG 终止	UGU } cys UGC } UGA 终止 UGG trp	U C A G	第三 位 (3' 端)
	C	CUU } CUC } leu CUA } CUG }	CCU } CCC } pro CCA } CCG }	CAU } his CAC } CAA } gln CAG }	CGU } CGC } arg CGA } CGG }	U C A G	
	A	AUU } AUC } ile AUA } AUG 起始	ACU } ACC } thr ACA } ACG }	AAU } asn AAC } AAA } lys AAG }	AGU } ser AGC } AGA } arg AGG }	U C A G	
	G	GUU } GUC } val GUA } GUG }	GCU } GCC } ala GCA } GCG }	GAU } asp GAC } GAA } glu GAG }	GGU } GGC } gly GGA } GGG }	U C A G	

# 密码子使用偏爱性

利用密码子使用频率对序列进行分析可以发现编码区的粗略位置。

	U		C		A		G	
U	UUU	Phe 57	UCU	Ser 16	UAU	Tyr 58	UGU	Cys 45
	UUC	Phe 43	UCC	Ser 15	UAC	Tyr 42	UGC	Cys 55
	UUA	Leu 13	UCA	Ser 13	UAA	Stp 62	UGA	Stp 30
	UUG	Leu 13	UCG	Ser 15	UAG	Stp 8	UGG	Trp 100
C	CUU	Leu 11	CCU	Pro 17	CAU	His 57	CGU	Arg 37
	CUC	Leu 10	CCC	Pro 17	CAC	His 43	CGC	Arg 38
	CUA	Leu 4	CCA	Pro 20	CAA	Gln 45	CGA	Arg 7
	CUG	Leu 49	CCG	Pro 51	CAG	Gln 66	CGG	Arg 10
A	AUU	Ile 50	ACU	Thr 18	AAU	Asn 46	AGU	Ser 15
	AUC	Ile 41	ACC	Thr 42	AAC	Asn 54	AGC	Ser 26
	AUA	Ile 9	ACA	Thr 15	AAA	Lys 75	AGA	Arg 5
	AUG	Met 100	ACG	Thr 26	AAG	Lys 25	AGG	Arg 3
G	GUU	Val 27	GCU	Ala 17	GAU	Asp 63	GGU	Gly 34
	GUC	Val 21	GCC	Ala 27	GAC	Asp 37	GGC	Gly 39
	GUA	Val 16	GCA	Ala 22	GAA	Glu 68	GGA	Gly 12
	GUG	Val 36	GCG	Ala 34	GAG	Glu 32	GGG	Gly 15

阅读框出现偏爱密码子，则该阅读框比较有可能是编码序列。

**CGC CTG ATT**

**CGA CTA ATA**

人类基因组密码子使用频率表



# 双联密码子出现频率

在实际应用中更多的是使用二联密码子出现频率

Name	ala	arg	asn	asp	cys	glu	gln	gly	his	ile	leu	lys	met	phe	pro	ser	thr	trp	tyr	val
ala	9.5	4.1	4.3	5.3	1.2	6	4.8	6.5	2	6.5	11.5	6	2.6	3.7	3.5	6.2	5	1.1	2.7	6.5
arg	7.9	5.5	3.9	5.3	1.1	6	5.5	5.9	2.6	6.5	11.4	5	2.2	4.7	3.6	5.5	4.4	1.4	4	6.6
asn	9.6	4.9	4.2	4.9	1	5.3	5.6	7.4	2.3	6	10	4.9	2	3.5	5.1	6.1	5.5	1.5	3.1	6.1
asp	9.3	4	4.7	5.1	1	6.7	2.9	7	1.8	7.1	9.6	6.3	2.3	4.3	3.9	5.9	5.1	1.6	3.6	6.6
cys	8.4	4.8	3.3	5.4	1.7	5.6	5.2	8.1	4.3	5.4	10.2	3.8	1.8	4.1	4.5	6.3	4.3	1.6	3.4	6.8
glu	9.4	5.8	3.6	4.5	0.8	4.9	7	5.8	2.6	5.9	12.7	5	2.4	4	3.5	5.4	5	1.1	2.8	6.8
gln	10.3	4.9	3	4.4	0.9	4.5	6.8	7	2.7	5.5	12.8	4.1	2	3.9	3.8	5.8	5.3	1.4	3	6.9
gly	8.1	4.8	3.9	5.1	1.2	6	4.6	6.4	2.4	6.8	10.5	5.8	2.7	4.8	2.4	5.8	5.1	1.4	3.7	7.5
his	7.3	4.7	4	4.8	1.5	4.9	5.6	6.9	3	6.2	10.8	4.8	1.6	5	5.2	6.8	4.9	1.7	4.2	5.1
ile	11	4.7	4.9	6.5	1.1	6.9	3.6	7.2	2.1	5.3	8.6	5.3	1.8	3.2	4.2	7	5.6	0.9	2.9	6.1
leu	10.4	4.2	4.3	5.2	1.1	5.2	3.7	6.8	2	5.6	10.6	5.3	2.3	3.8	4.5	7.4	6.2	1	2.6	6.6
lys	10.6	5.2	3.8	5.2	0.5	5.3	5.9	6.6	2.6	5.2	11.3	4.7	1.9	2.8	4.6	6	5.5	1.2	2.6	7.6
met	10.8	4.8	3.8	4.6	0.7	4.6	4.9	7	1.7	4.7	11.4	5.2	2.8	3.3	5.1	7.4	6.3	0.9	2	6.8
phe	9.6	3.7	5.2	6.5	1.2	6.4	2.7	7.9	1.9	6.7	7.4	5	2.5	3.9	3.6	8	5.8	1.3	3.3	6.3
pro	8.4	3.6	4.6	5.4	0.7	7.6	5.2	5.4	2.3	6.1	11.2	5.5	2.4	4.2	2.8	6.5	5.4	1.4	2.9	7.5
ser	9.1	4.6	3.7	5	1	5.4	5.2	7.2	2.6	6	11.6	4.5	2.2	4.1	4.1	6.5	5	1.2	3.2	6.8
thr	9.1	4.2	3.7	5.6	0.9	5.7	5.7	7.5	2.2	5.5	12	4.2	2	3.5	5.5	6.2	5.3	1.1	2.6	6.7
trp	7.1	6.3	3.2	4.8	1.3	3.9	8.5	6.6	3.6	5	14.2	3.2	2.4	4.6	3.9	5.8	4.3	1.3	3	6.1
tyr	7.9	6.5	3.6	4.9	1.2	4.5	7	7.1	2.6	5	11.7	4	1.6	4.7	4.9	6.4	4.6	1.5	3.4	5.7
val	9.6	4.1	4.4	5.9	1	6.2	3.4	6.4	1.8	6.5	10.2	5.2	2.5	3.7	3.8	7.2	6.1	1.1	2.7	7.1

有些二联密码子（亮氨酸与色氨酸）倾向于同时出现，另一些（脯氨酸与半胱氨酸）则不是



# 编码序列统计学特征

如果一段**DNA**序列的阅读框出现许多偏爱密码子（双联密码子），则该区域很有可能就是基因的编码区域。

# 构建真核基因模型

tccatgcagaccatggcggtacaggacatgccggtgcagctctgactt

找到可能起始密码和终止密码子

tccATGcagaccATGgcggtacaggacATGccggtgcagctctgactt

剪接位点形成外显子和内含子边界

tccATGagaccATGgcggtacaggacATGccggtgcagctctgactt

tcc**ATG**cagacc**ATG**gcg**gt**ac**agg**ac**ATG**ccg**gt**gc**ag**ctc**tga**ctt

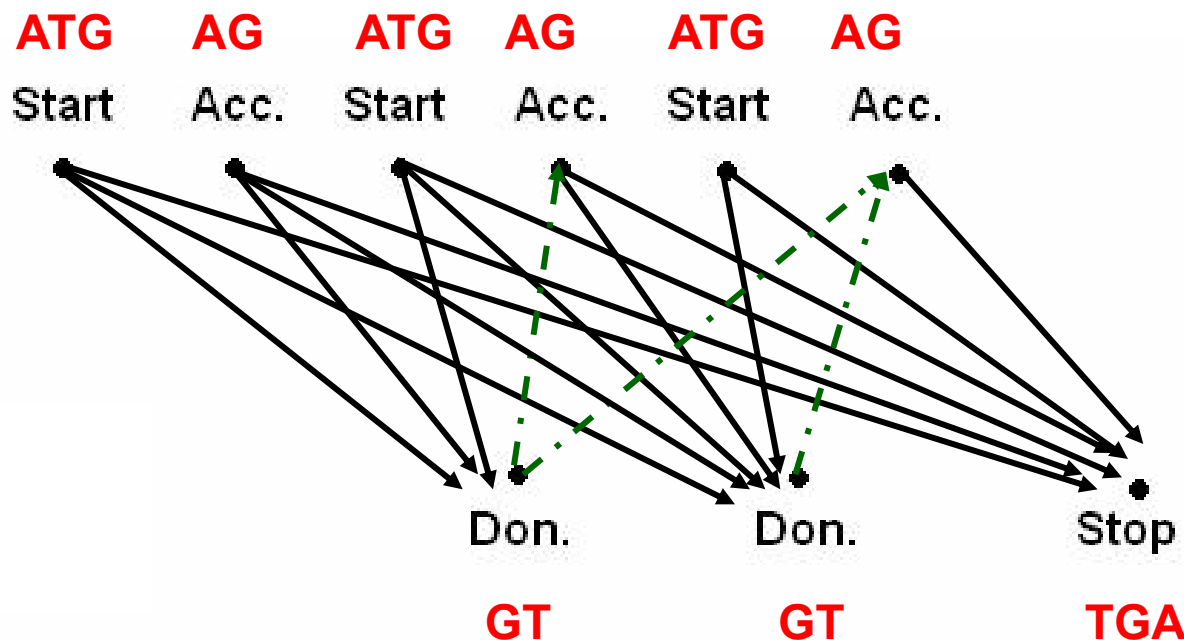


可能的外显子

- 1 **ATG**cagacc**ATG**gcg
- 2 **ATG**cagacc**ATG**gcg**gt**ac**agg**ac**ATG**ccg
- 3 **ATG**cagacc**ATG**gcg**gt**ac**agg**ac**ATG**ccg**gt**gcagctc**tga**
- 4 acc**ATG**gcg
- 5 acc**ATG**gcg**gt**ac**agg**ac**ATG**ccg
- 6 acc**ATG**gcg**gt**ac**agg**ac**ATG**ccg**gt**gcagctc**tga**
- 7 **ATG**gcg
- 8 **ATG**gcg**gt**ac**agg**ac**ATG**ccg
- 9 **ATG**gcg**gt**ac**agg**ac**ATG**ccg**gt**gcagctc**tga**
- 10 gac**ATG**ccg
- 11 gac**ATG**ccg**gt**gcagctc**tga**
- 12 **ATG**ccg
- 13 **ATG**ccg**gt**gcagctc**tga**
- 14 ctc**tga**

# 构建真核基因模型

tcc**ATG**c**ag**acc**ATG**gcg**gt**ac**ag**gac**ATG**ccg**gt**gc**ag**ctc**tga**ctt



→  
可能的外显子

---→  
可能的内含子

从起点到终点的任何一条路径代表一个可能的基因结构

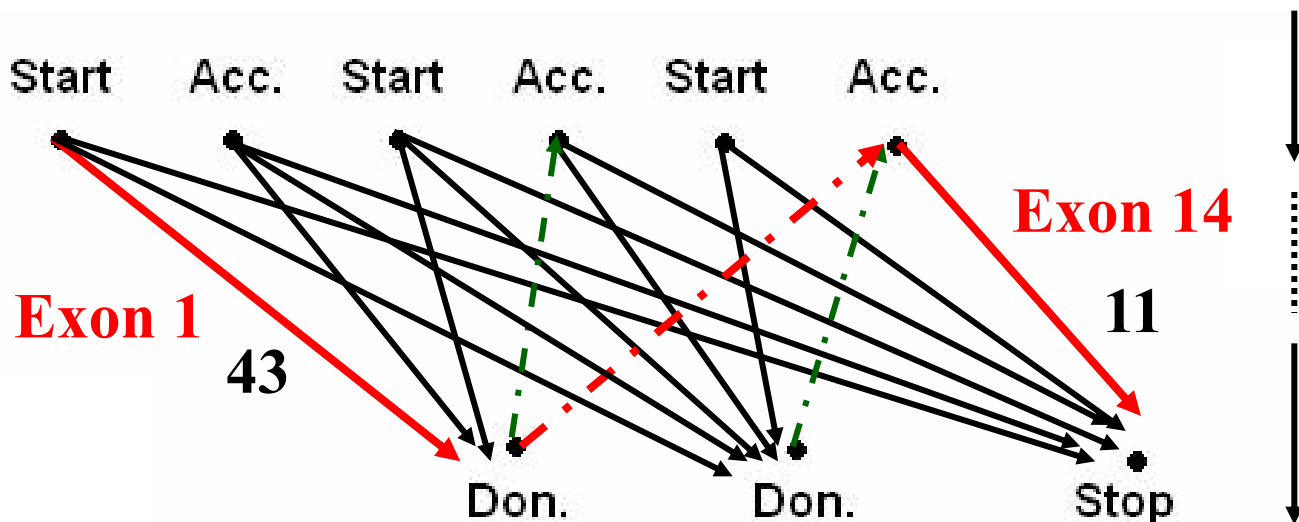
# 构建真核基因模型

1 根据编码区序列统计学特性确定可能外显子的概率大小

2 根据外显子的概率大小确定可能的基因结构

# 构建真核基因模型

tcc**ATG**cagacc**ATG**gcggtacaggac**ATG**ccggtgcagctctga**ctt**



tcc**ATG**cagaccatggcggtacaggacatgccggtgcagctctga**ctt**

外显子1

内含子

外显子2

**ATG**cagaccatggcgctctga

ORF

# 常用真核基因预测软件

## **(1)、FGENEH**

所用算法：线性判别分析（**Linear Discriminant Analysis**）方法

## **(2)、GeneID**

所用算法：法则系统（**Rule-based System**）算法

## **(3)、GeneParser**

所用算法：动态规划算法（**Dynamic Programming**）

## **(4)、GRAIL**

所用算法：人工神经网络方法

## **(5)、MZEF**

所用算法：决策树（**decision tree**）方法

## **(6)、GENSCAN**

所用算法：隐Markov模型（**Hidden Markov Model**）方法、动态规划算法

## **(7)、Genie**

所用算法：广义隐Markov模型（**Generalized Hidden Markov Model**）方法、动态规划算法

## **(8)、HMMgene**

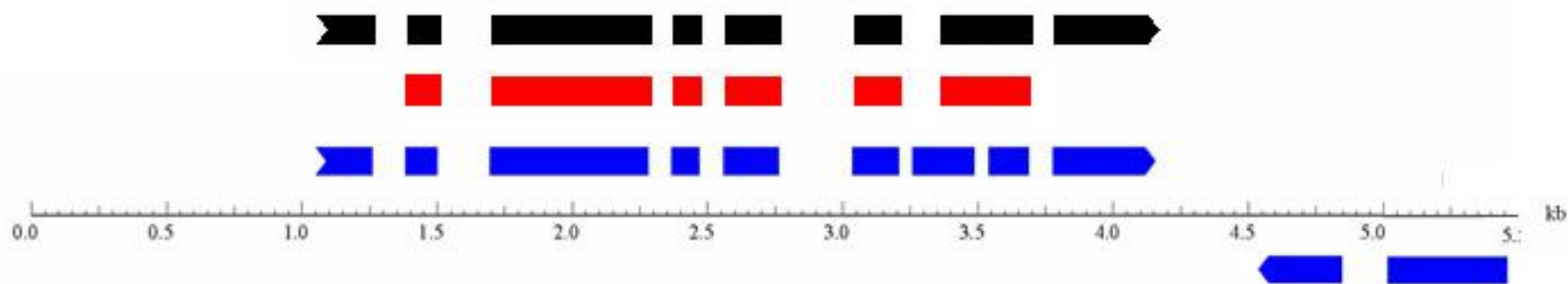
所用算法：隐Markov模型（**Hidden Markov Model**）方法





# 分析举例 ( Genscan )

- Genscan ( <http://genes.mit.edu/GENSCAN.html> )
- 2) 选择适合物种模式物种作为参照
    - a ) Vertebrate(脊椎动物)
    - b ) **Arabidopsis**(拟南芥)
    - c ) Maize (玉米)
  - 3) 提供需要预测分析的序列 ( GenBank:AB363664 ) , 并选择相关合适的参数 ( 默认参数 )
  - 4) 基因识别程序对提供的序列给出分析结果
-



# MZEF、Genscan基因预测与已注释结果的比较



Key:  Initial exon  Internal exon  Terminal exon  Single-exon gene

■ 已注释外显子  
■ MZEF预测的外显子  
■ Genscan预测的外显子

# 综合多个方法提高预测的准确性

- 目前还没有一个基因预测工具可以完全正确地预测一个基因组中的所有基因
- 目前最好的基因预测工具预测一个基因组中的所有外显子的准确率最多达到75%，预测基因结构的准确率< 50%
- 不同的基因预测软件分析结果有差异，尽量综合多个基因预测软件的分析结果