

生物信息学

天津医科大学
生物医学工程与技术学院

2018-2019 学年上学期 (秋)
2016 级生信班

第四章 核酸序列分析

伊现富 (Yi Xianfu)

天津医科大学 (TJMU)
生物医学工程与技术学院

2018 年 11 月



章节内容概览

4.1 DNA 序列信息分析

- ① DNA 序列的基本信息：组份，序列转换，限制酶位点
- ② DNA 序列的特征信息：开放阅读框，启动子，CpG 岛

4.2 基因组结构注释分析

- ① 重复序列
- ② 基因识别

4.3 RNA 序列分析

- ① mRNA 选择性剪接
- ② miRNA 及其靶基因



教学提纲

- 1 引言
- 2 DNA 组份分析与序列转换
- 3 限制酶位点分析
- 4 开放阅读框分析
- 5 启动子分析
- 6 CpG 岛识别
- 7 总结与答疑
- 8 引言

- 9 重复序列分析
- 10 基因识别
- 11 总结与答疑
- 12 引言
- 13 mRNA 选择性剪接
- 14 miRNA 及其靶基因预测
- 15 总结与答疑
- 16 复习思考题

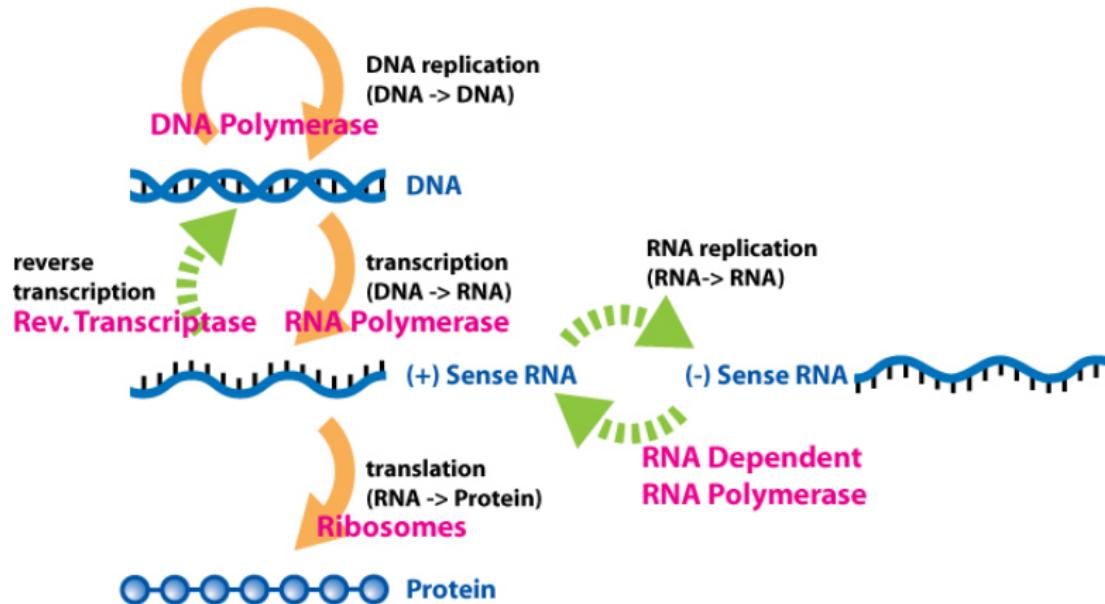
教学提纲

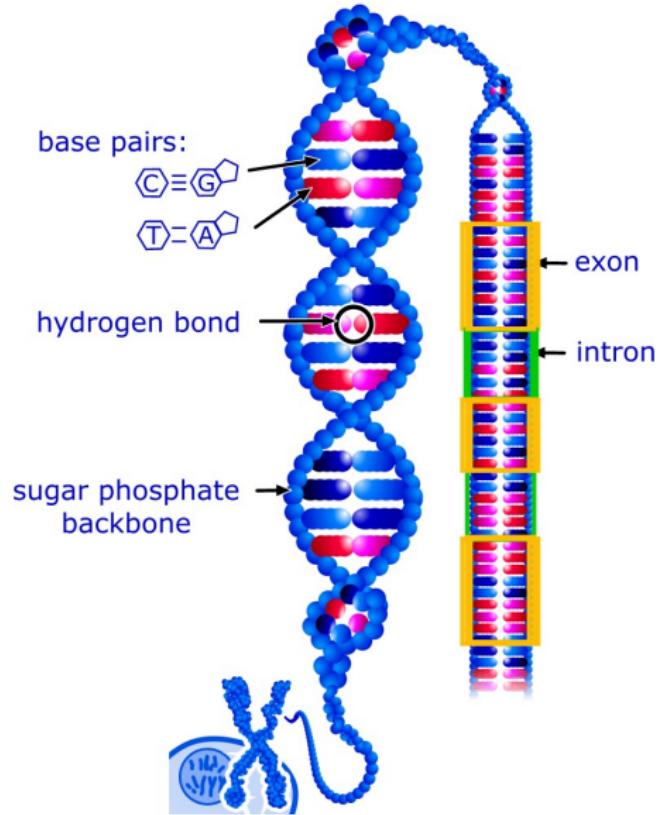
- 1 引言
- 2 DNA 组份分析与序列转换
- 3 限制酶位点分析
- 4 开放阅读框分析
- 5 启动子分析
- 6 CpG 岛识别
- 7 总结与答疑
- 8 引言

- 9 重复序列分析
- 10 基因识别
- 11 总结与答疑
- 12 引言
- 13 mRNA 选择性剪接
- 14 miRNA 及其靶基因预测
- 15 总结与答疑
- 16 复习思考题

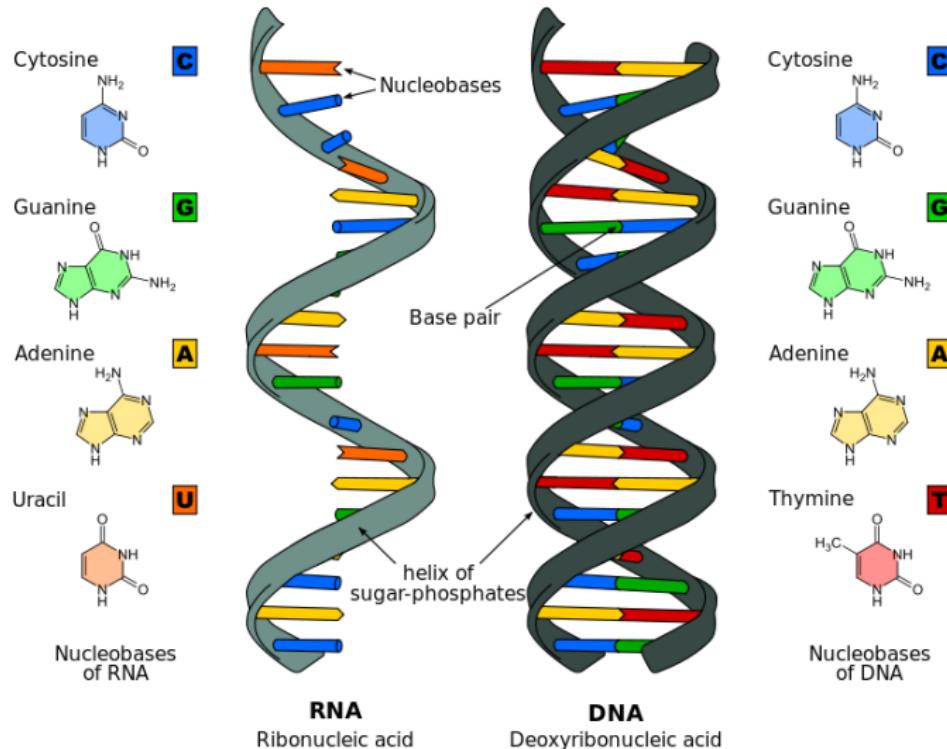


引言 | 中心法则

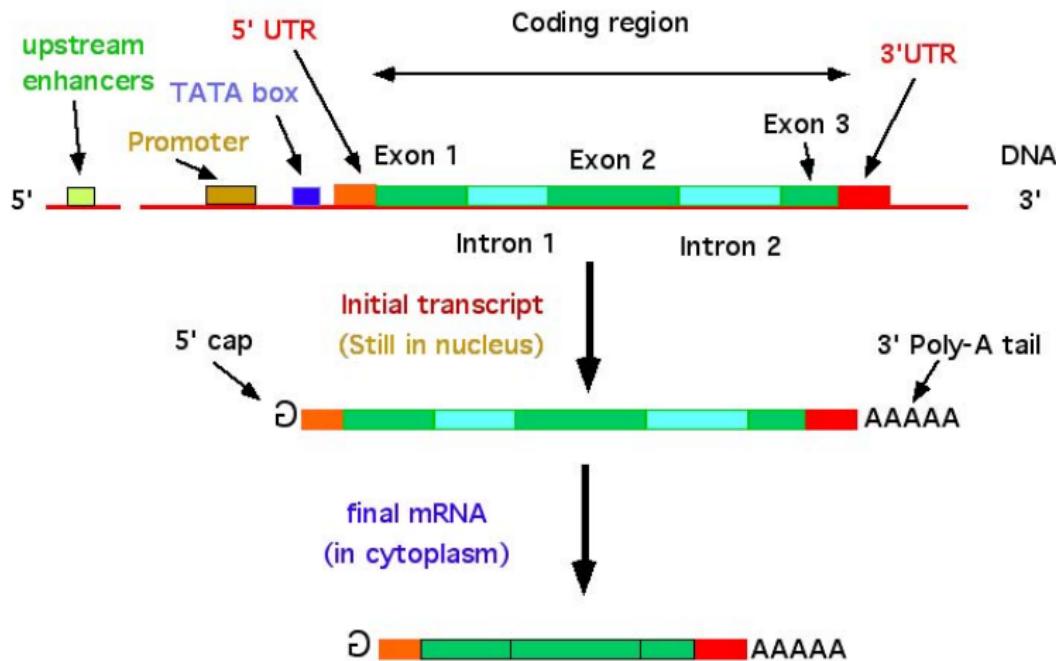




引言 | DNA & RNA



引言 | 遗传信息



基本问题

- 总的 GC 含量或者其他核苷酸成分是多少？
- 有哪些重复的 DNA 序列，在什么地方？
- 一共有多少个基因（编码蛋白质的序列）？

深层问题

- 为什么会有各种特征序列？（物理、化学性质？进化压力？）
- 需要从哪些方面分析序列特征？
- 怎样描述这些序列特征？

序列分析

通过实验或计算等方式，确定核苷酸或氨基酸序列中可能与特定功能、结构或生化过程相关联的**具有生物学意义的序列特征**，或者**序列自身的规律**。

基本问题

- 总的 GC 含量或者其他核苷酸成分是多少？
- 有哪些重复的 DNA 序列，在什么地方？
- 一共有多少个基因（编码蛋白质的序列）？

深层问题

- 为什么会有各种特征序列？（物理、化学性质？进化压力？）
- 需要从哪些方面分析序列特征？
- 怎样描述这些序列特征？

序列分析

通过实验或计算等方式，确定核苷酸或氨基酸序列中可能与特定功能、结构或生化过程相关联的**具有生物学意义的序列特征**，或者**序列自身的规律**。

引言 | 序列分析

基本问题

- 总的 GC 含量或者其他核苷酸成分是多少？
- 有哪些重复的 DNA 序列，在什么地方？
- 一共有多少个基因（编码蛋白质的序列）？

深层问题

- 为什么会有各种特征序列？（物理、化学性质？进化压力？）
- 需要从哪些方面分析序列特征？
- 怎样描述这些序列特征？

序列分析

通过实验或计算等方式，确定核苷酸或氨基酸序列中可能与特定功能、结构或生化过程相关联的**具有生物学意义的序列特征**，或者**序列自身的规律**。

教学提纲

1 引言

2 DNA 组份分析与序列转换

3 限制酶位点分析

4 开放阅读框分析

5 启动子分析

6 CpG 岛识别

7 总结与答疑

8 引言

9 重复序列分析

10 基因识别

11 总结与答疑

12 引言

13 mRNA 选择性剪接

14 miRNA 及其靶基因预测

15 总结与答疑

16 复习思考题



查戈夫法则

第一法则 $A = T, G = C \implies A + C = T + G, A + G = C + T$

第二法则 AT/GC 的比值因生物种类不同而异



序列长度

序列长度是具有独立生物学功能的序列片段（如基因、启动子等）的基本性质。物种的基因组长度也是重要参数之一。

蛋白质编码基因的序列长度

- 原核： ~ 1000 个核苷酸
- 脊椎动物： ~ 30000 个核苷酸
- 人： $20000\sim 50000$ 个核苷酸



序列长度

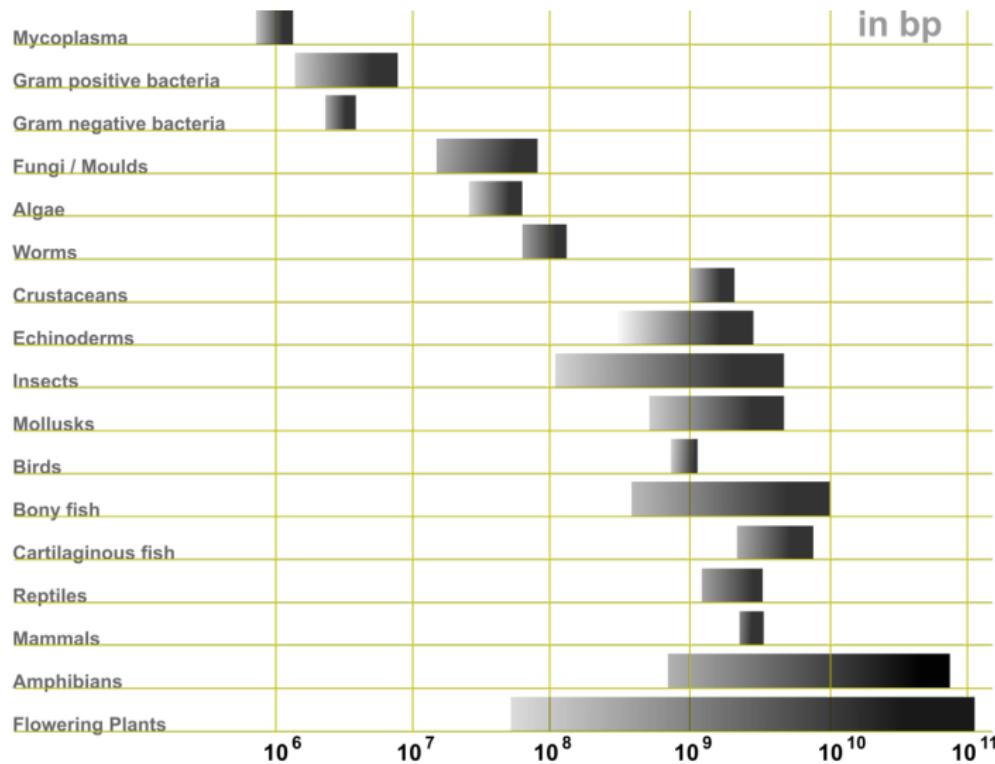
序列长度是具有独立生物学功能的序列片段（如基因、启动子等）的基本性质。物种的基因组长度也是重要参数之一。

蛋白质编码基因的序列长度

- 原核： ~ 1000 个核苷酸
- 脊椎动物： ~ 30000 个核苷酸
- 人： $20000\sim 50000$ 个核苷酸



DNA 序列 | 序列长度 | 基因组



碱基组成

- 核酸序列由 ACGT 四种碱基组成
- 不同物种的 DNA 碱基组成存在差异
- 同一基因组内不同区段（基因、基因间）的碱基组成有差异
- 同一基因内部不同片段（外显子、内含子）的碱基组成也有差异

碱基频率

- 对于随机分布的 DNA 序列，每种核苷酸的出现是均匀分布的（出现频率各为 0.25）；真实基因组的核苷酸分布则是非均匀的（酵母： $A/T=0.325$, $G/C=0.175$ ）
- 如果同时计算 DNA 的正反两条链，A 和 T、G 和 C 的出现频率相同（碱基配对原则）；如果仅统计一条链，则虽然 A 和 T、G 和 C 的出现频率不同，但是数值接近（酵母： $A=0.344$, $T=0.343$, $G=0.157$, $C=0.155$ ）

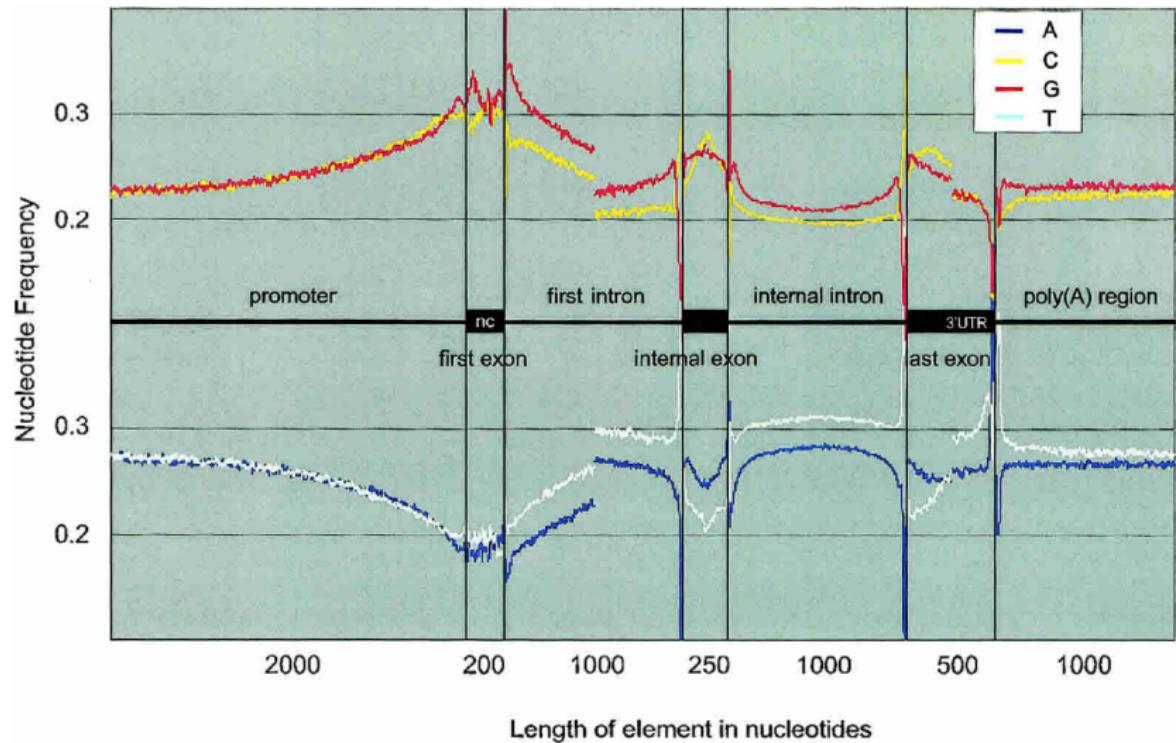
碱基组成

- 核酸序列由 ACGT 四种碱基组成
- 不同物种的 DNA 碱基组成存在差异
- 同一基因组内不同区段（基因、基因间）的碱基组成有差异
- 同一基因内部不同片段（外显子、内含子）的碱基组成也有差异

碱基频率

- 对于随机分布的 DNA 序列，每种核苷酸的出现是均匀分布的（出现频率各为 0.25）；真实基因组的核苷酸分布则是非均匀的（酵母： $A/T=0.325$, $G/C=0.175$ ）
- 如果同时计算 DNA 的正反两条链，A 和 T、G 和 C 的出现频率相同（碱基配对原则）；如果仅统计一条链，则虽然 A 和 T、G 和 C 的出现频率不同，但是数值接近（酵母： $A=0.344$, $T=0.343$, $G=0.157$, $C=0.155$ ）

DNA 序列 | 碱基组成 | 实例



GC 含量 (GC content)

- 对象：核酸片段、基因、基因组、……
- 鸟嘌呤 (G) 和胞嘧啶 (C) 所占的比例
- GC 含量随 DNA 不同而异
- GC 含量高的 DNA 更加稳定
- 计算公式： $\frac{G+C}{A+T+G+C} \times 100$
- GC 比 (GC-ratio) : $\frac{G+C}{A+T}$
- 结合滑动窗口进行计算



特点

- 不同物种基因组中 GC 含量不同。 (15%~75%，两头少中间多。疟原虫为 20%，啤酒酵母为 38%，人约为 40%，天蓝色链霉菌 A3 为 72%。)
- 同一基因组内，GC 含量不均匀。
- GC 含量与多种生物学特征相关，比如基因密度、内含子、外显子等。

应用

- 根据 GC 含量差异识别细菌种类
- 真核基因组具有 GC 含量较高或较低的近似均匀片段
- 不同物种的密码子使用与其 GC 含量有关
- GC 含量与 DNA 双链的熔解温度有关，是进行核酸杂交的重要参数

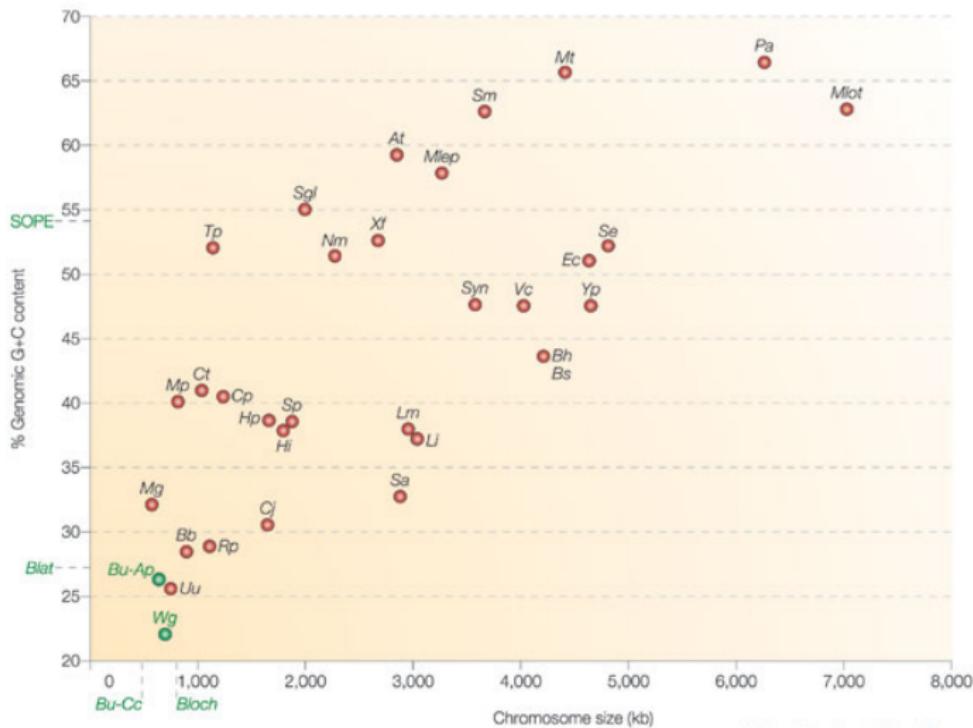
特点

- 不同物种基因组中 GC 含量不同。 (15%~75%，两头少中间多。疟原虫为 20%，啤酒酵母为 38%，人约为 40%，天蓝色链霉菌 A3 为 72%。)
- 同一基因组内，GC 含量不均匀。
- GC 含量与多种生物学特征相关，比如基因密度、内含子、外显子等。

应用

- 根据 GC 含量差异识别细菌种类
- 真核基因组具有 GC 含量较高或较低的近似均匀片段
- 不同物种的密码子使用与其 GC 含量有关
- GC 含量与 DNA 双链的熔解温度有关，是进行核酸杂交的重要参数

DNA 序列 | GC 含量 | 基因组



DNA 序列 | GC 含量 | 基因区

Introns tend to be slightly richer in AT residues compared to their neighbouring exons.

Code	Yeast	Coding "o" regions ^(a)		Intron regions ^{(b)(c)}			Difference Intron-exon ^(b)
		GC	AT	GC	AT	n	
	<i>S. cerevisiae</i>	39.6	60.4	33.4	66.6	260	6.2
AT	<i>S. servazzii</i>	34.7	65.3	27.6	72.4	22	7.1
AU	<i>S. kluyveri</i>	41.5	58.5	36.8	63.2	27	4.7
AZ	<i>K. marxianus</i>	42.3	57.7	34.5	65.5	13	7.8
BD	<i>C. tropicalis</i>	34.5	65.5	26.9	73.1	7	7.6
BC	<i>D. hansenii</i>	36.5	63.5	33.6	66.4	12	2.9
BB	<i>P. angusta</i>	48.5	51.5	41.8	58.2	29	6.7
AW	<i>V. lipolytica</i>	53.0	47.0	48.5	51.5	15	4.5

(a) Génolevures, 2000, *FEBS Lett.*, 487, 1-149.

(b) Bon et al., 2003.

(c) Only entire introns



DNA 序列 | GC 含量 | 基因 vs. 基因组

Gene	Gene ID	Bacterium	RefSeq	Gene GC %	Genome GC %
tetA	2716475	<i>Escherichia coli</i> plasmid pC15-1a	NC_005327.1	63.66	52.6
	8877592	<i>Klebsiella pneumoniae</i> plasmid pKF3-140	NC_013951.1	63.21	52.5
	7886608	<i>Salmonella enterica</i> plasmid pAM04528	NC_012693.1	62.43	51.9
	7003405	<i>Haemophilus influenzae</i> plasmid lCEhin1056	NC_011409.1	43.36	39.1
	2653967	<i>Serratia marcescens</i> plasmid R478	NC_004989.1	43.28	36.9
	4927413	<i>Yersinia pestis</i> biovar Orientalis str. IP275 plP1202	NC_009141.1	57.63	52.9
	6002612	<i>Acinetobacter baumannii</i> AYE	NC_010410.1	63.21	39.3
	1794537	<i>Rhodopirellula baltica</i> SH 1	NC_005027.1	57.55	55.4
	3433250	<i>Corynebacterium jeikeium</i> K411	NC_007164.1	68.50	61.40
	2797858	<i>Listeria monocytogenes</i> serotype 4b str. F2365	NC_002973.6	42.30	38.00
<i>p</i> = 0.02					
transferase	1238790	<i>Klebsiella pneumoniae</i> plasmid pJHCMW1	NC_003486.1	52.97	49
	13919580	<i>Providencia stuartii</i> plasmid pTC2	NC_019375.1	53.03	52.5
	9487131	<i>Klebsiella pneumoniae</i> plasmid pNL194	NC_014368.1	53.03	53.1
	9487121	<i>Klebsiella pneumoniae</i> plasmid pNL194	NC_014368.1	52.9	53.1
	1055588	<i>Citrobacter freundii</i> plasmid pCTX-M3	NC_004464.2	51.89	51
	7156160	<i>Escherichia coli</i> UMN026	NC_011751.1	58.37	50.6
	6810778	<i>Salmonella enterica</i> subsp. <i>enterica</i> serovar Paratyphi A str. AKU_12601	NC_011147.1	54.11	52.2
	6455499	<i>Cupriavidus taiwanensis</i> LMG 19424, Chr 2	NC_010530.1	70.94	67
	6928720	<i>Burkholderia cenocepacia</i> J2315, Chr 2	NC_011001.1	71.33	66.9
	2662391	<i>Bordetella bronchiseptica</i> RB50	NC_002927.3	73.45	68.1
<i>p</i> = 0.2					



序列转换

- 反向序列
- 互补序列
- 反向互补序列
- DNA 双链
- RNA 序列

书写惯例

- DNA/RNA : [左] 5' \rightarrow 3' [右]
- 多肽/蛋白质 : [左] N 端 (氨基端) \rightarrow C 端 (羧基端) [右]



序列转换

- 反向序列
- 互补序列
- 反向互补序列
- DNA 双链
- RNA 序列

书写惯例

- DNA/RNA : [左] 5' \rightarrow 3' [右]
- 多肽/蛋白质 : [左] N 端 (氨基端) \rightarrow C 端 (羧基端) [右]



教学提纲

- 1 引言
- 2 DNA 组份分析与序列转换
- 3 限制酶位点分析
- 4 开放阅读框分析
- 5 启动子分析
- 6 CpG 岛识别
- 7 总结与答疑
- 8 引言
- 9 重复序列分析
- 10 基因识别
- 11 总结与答疑
- 12 引言
- 13 mRNA 选择性剪接
- 14 miRNA 及其靶基因预测
- 15 总结与答疑
- 16 复习思考题



限制酶 (restriction enzyme)

又称限制内切酶或限制性内切酶 (restriction endonuclease) , 全称限制性核酸内切酶, 是可以识别 DNA 的特异序列、并在识别位点或其周围切割双链 DNA 的一类内切酶。

切割末端

- 黏性末端 vs. 平滑末端



限制酶 | 定义

限制酶 (restriction enzyme)

又称限制内切酶或限制性内切酶 (restriction endonuclease) , 全称限制性核酸内切酶, 是可以识别 DNA 的特异序列、并在识别位点或其周围切割双链 DNA 的一类内切酶。

切割末端

- 黏性末端 vs. 平滑末端



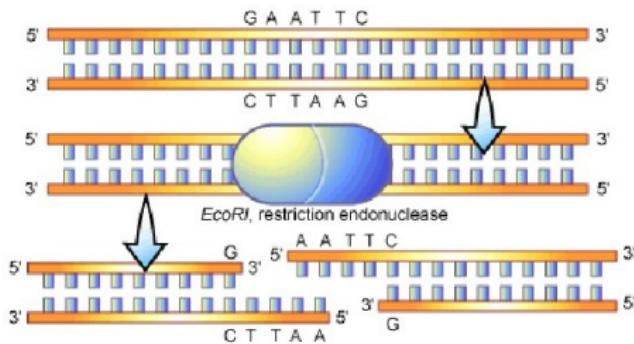
限制酶 | 定义

限制酶 (restriction enzyme)

又称限制内切酶或限制性内切酶 (restriction endonuclease) , 全称限制性核酸内切酶, 是可以识别 DNA 的特异序列、并在识别位点或其周围切割双链 DNA 的一类内切酶。

切割末端

- 黏性末端 vs. 平滑末端



Derivation of the EcoRI name

Abbreviation	Meaning	Description
E	<i>Escherichia</i>	genus
co	<i>coli</i>	species
R	RY13	strain
I	First identified	order of identification in the bacterium



II型限制酶

- 识别与切割位点：专一
 - 识别序列：4-8个碱基，回文对称结构
 - 切割序列：识别序列，切割位点对称
- 切割末端：黏性末端，平滑末端
 - 黏性末端：切割位点在回文序列的一侧
 - 平滑末端：切割位点在回文序列的中间

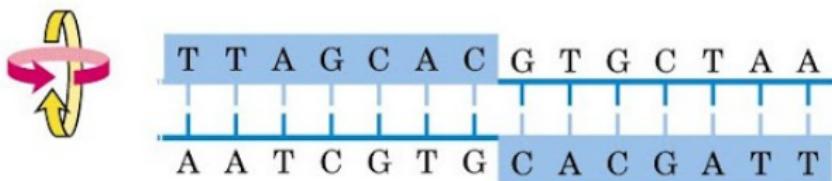


限制酶 | II 型 | 回文对称

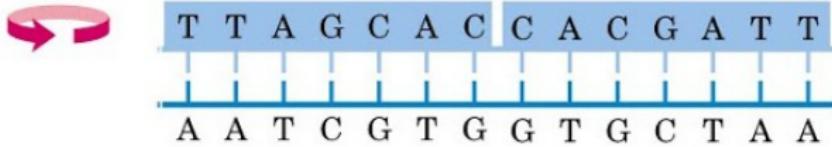
回文对称 (palindrome)

特指 DNA 的一种具有反向重复的结构。具有这种结构的 DNA，其一条链从左向右读和另一条链从右向左读的序列是相同的。

Palindrome

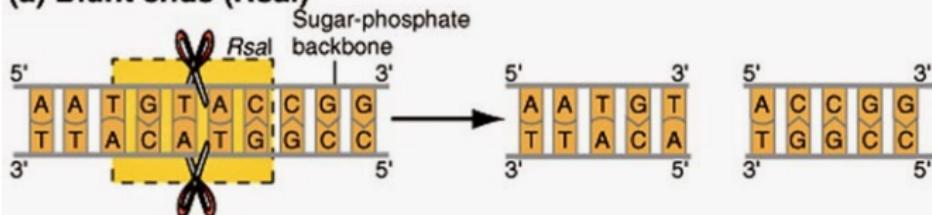


Mirror repeat

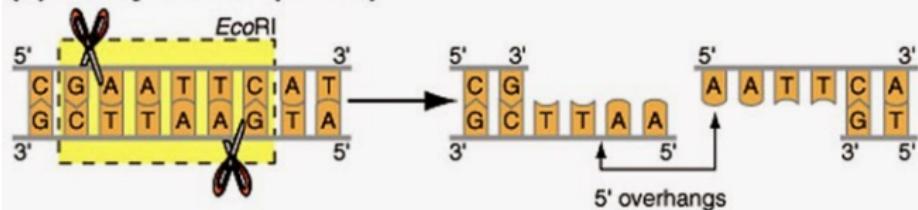


限制酶 | II 型 | 末端

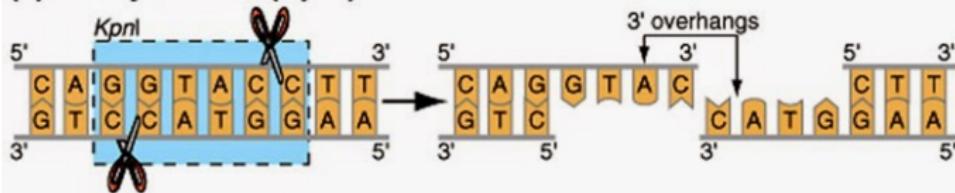
(a) Blunt ends (*Rsa*I)



(b) Sticky 5' ends (*Eco*RI)



(c) Sticky 3' ends (*Kpn*I)



限制酶 | II 型 | 末端 | 黏性末端

酵素名称	来源	辨识序列	切法
EcoRI	<i>Escherichia coli</i>	5'GAATTC 3'CTTAAAG	5'---G AATTC---3' 3'---CTTAA G---5'
BamHI	<i>Bacillus amyloliquefaciens</i>	5'GGATCC 3'CCTAGG	5'---G GATCC---3' 3'---CCTAG G---5'
HindIII	<i>Haemophilus influenzae</i>	5'AAGCTT 3'TTCGAA	5'---A AGCTT---3' 3'---TTCGA A---5'
TaqI	<i>Thermus aquaticus</i>	5'TCGA 3'AGCT	5'---T CGA---3' 3'---AGC T---5'
NotI	<i>Nocardia otitidis</i>	5'GCGGCCGC 3'CGCCGGCG	5'---GC GGCGC---3' 3'---CGCCGG CG---5'



限制酶 | II 型 | 末端 | 平滑末端

PovII*	<i>Proteus vulgaris</i>	5' CAGCTG 3' GTCGAC	5' ---CAG CTG---3' 3' ---GTC GAC---5'
SmaI*	<i>Serratia marcescens</i>	5' CCCGGG 3' GGGCCC	5' ---CCC GGG---3' 3' ---GGG CCC---5'
HaeIII*	<i>Haemophilus egytius</i>	5' GGCC 3' CCGG	5' ---GG CC---3' 3' ---CC GG---5'
AluI*	<i>Arthrobacter luteus</i>	5' AGCT 3' TCGA	5' ---AG CT---3' 3' ---TC GA---5'
EcoRV*	<i>Escherichia coli</i>	5' GATATC 3' CTATAG	5' ---GAT ATC---3' 3' ---CTA TAG---5'



资源

- REBASE：收录了限制酶的所有信息
- NEBCutter V2.0：产生 DNA 序列的酶切位点分析结果



教学提纲

1

引言

2

DNA 组份分析与序列转换

3

限制酶位点分析

4

开放阅读框分析

5

启动子分析

6

CpG 岛识别

7

总结与答疑

8

引言

9

重复序列分析

10

基因识别

11

总结与答疑

12

引言

13

mRNA 选择性剪接

14

miRNA 及其靶基因预测

15

总结与答疑

16

复习思考题

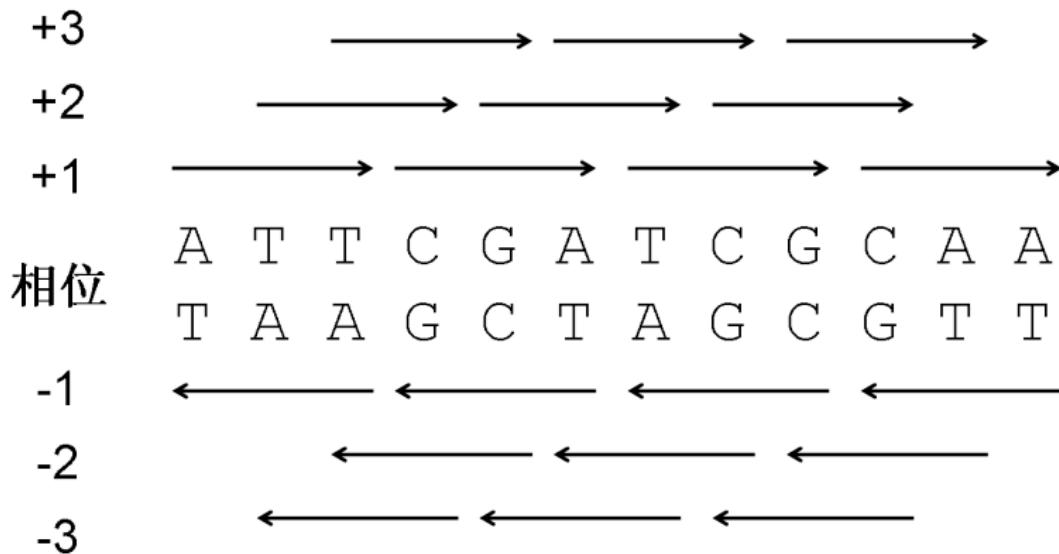


开放阅读框 (Open Reading Frame, ORF)

在给定的阅读框架中，不包含终止密码子的一串序列，是生物个体的基因组中可能作为蛋白质编码序列的部分，包含从 5' 端翻译起始密码子 (AUG) 到终止密码子 (UAA、UAG、UGA) 之间的一段编码蛋白质的碱基序列。



开放阅读框 | 相位 (frame)



开放阅读框 | ORF VS. CDS

- 一个 ORF 对应一个候选的 CDS (编码序列, Coding DNA Sequence)
- ORF : 理论预测
- CDS : 实验证实
- 分析 DNA 序列中的 ORF 是对该序列是否为 CDS 的初步判断



开放阅读框 | ORF VS. CDS

- 一个 ORF 对应一个候选的 CDS (编码序列, Coding DNA Sequence)
- ORF : 理论预测
- CDS : 实验证实
- 分析 DNA 序列中的 ORF 是对该序列是否为 CDS 的初步判断



- 确定第一个 AUG 和终止密码子
- 原核生物：最长 ORF 法
- 真核生物：特征统计、模式识别、同源比对
- ORF Finder：NCBI 的在线分析工具



教学提纲

- 1 引言
- 2 DNA 组份分析与序列转换
- 3 限制酶位点分析
- 4 开放阅读框分析
- 5 启动子分析
- 6 CpG 岛识别
- 7 总结与答疑
- 8 引言

- 9 重复序列分析
- 10 基因识别
- 11 总结与答疑
- 12 引言
- 13 mRNA 选择性剪接
- 14 miRNA 及其靶基因预测
- 15 总结与答疑
- 16 复习思考题



- 顺式作用元件 (cis-acting element) : 核酸序列
 - 启动子 (promoter)
 - 增强子 (enhancer)
 - ...
- 反式作用因子 (trans-acting factor) : 蛋白质
- 两者相互作用实现转录调控



启动子 | 定义

启动子 (promoter)

一段位于转录起始位点 5' 端上游区的 DNA 序列，能活化 RNA 聚合酶，使之与模板 DNA 准确地结合并具有转录起始的特异性。

转录起始位点 (Transcription Start Site, TSS)

与新生 RNA 链第一个核苷酸相对应 DNA 链上的碱基，研究证实通常为一个嘌呤。



启动子 (promoter)

一段位于转录起始位点 5' 端上游区的 DNA 序列，能活化 RNA 聚合酶，使之与模板 DNA 准确地结合并具有转录起始的特异性。

转录起始位点 (Transcription Start Site, TSS)

与新生 RNA 链第一个核苷酸相对应 DNA 链上的碱基，研究证实通常为一个嘌呤。

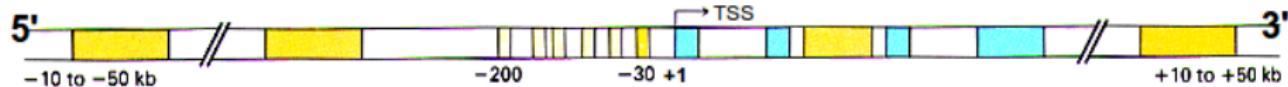


启动子 (promoter)

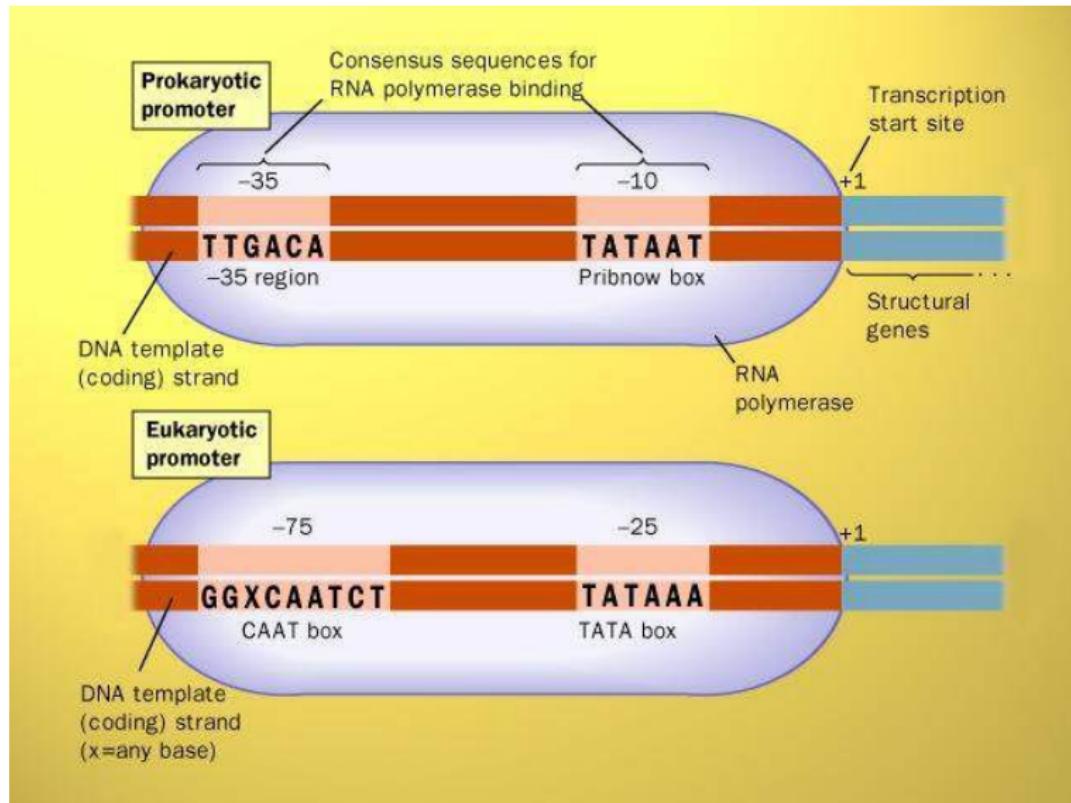
一段位于转录起始位点 5' 端上游区的 DNA 序列，能活化 RNA 聚合酶，使之与模板 DNA 准确地结合并具有转录起始的特异性。

转录起始位点 (Transcription Start Site, TSS)

与新生 RNA 链第一个核苷酸相对应 DNA 链上的碱基，研究证实通常为一个嘌呤。



启动子 | 结构



转录因子 (transcription factor)

能够结合在某基因上游特异核苷酸序列上的蛋白质，这些蛋白质能调控其基因的转录。

- 功能区域：DNA 结合结构域、效应结构域
- 作用方式：与启动子区域结合、与其他 TF 形成复合体
- 调控模式：同时调控多个基因

转录因子结合位点 (Transcription Factor Binding Site, TFBS)

与转录因子结合的 DNA 序列，长度约为 5~20bp，它们与转录因子相互作用进行基因的转录调控。

- 保守性 vs. 可变性



转录因子 (transcription factor)

能够结合在某基因上游特异核苷酸序列上的蛋白质，这些蛋白质能调控其基因的转录。

- 功能区域：DNA 结合结构域、效应结构域
- 作用方式：与启动子区域结合、与其他 TF 形成复合体
- 调控模式：同时调控多个基因

转录因子结合位点 (Transcription Factor Binding Site, TFBS)

与转录因子结合的 DNA 序列，长度约为 5~20bp，它们与转录因子相互作用进行基因的转录调控。

- 保守性 vs. 可变性



启动子 | TFBS

M00671 TCF-4

	A	C	G	T
01	1	3	2	0
02	0	6	0	0
03	0	1	0	5
04	0	0	0	6
05	0	0	0	6
06	0	0	5	1
07	6	0	0	0
08	3	0	1	2



M00761 TP53

	A	C	G	T
01	25	3	16	2
02	14	0	32	0
03	25	0	21	0
04	2	39	4	1
05	32	2	4	8
06	23	2	2	19
07	3	0	43	0
08	9	15	5	17
09	2	28	9	7
10	5	22	5	14



M00789 GATA

	A	C	G	T
01	50	8	8	39
02	1	0	103	1
03	104	0	1	0
04	0	0	0	105
05	89	1	3	12
06	58	3	39	5
07	28	18	48	11



- 启动子

- EPD：有注释、非冗余的真核生物 RNA 聚合酶 II 启动子数据集
- Promoter Scan (同源性分析) , Promoter 2.0 (人工神经网络技术)

- 转录因子

- TRANSFAC : 真核生物顺式作用元件和反式作用因子数据库
- Tfblast (TRANSFAC BLAST)
- JASPAR: The high-quality transcription factor binding profile database
- CIS-BP Database: Catalog of Inferred Sequence Binding Preferences
- footprintDB
- HOCOMOCO: expansion and enhancement of the collection of transcription factor binding sites models
- MotifMap: genome-wide maps of regulatory elements.
- UniPROBE (Universal PBM Resource for Oligonucleotide Binding Evaluation) database
- ENCODE TF ChIP-seq datasets
- Human Protein-DNA Interactome (hPDI)



教学提纲

- 1 引言
- 2 DNA 组份分析与序列转换
- 3 限制酶位点分析
- 4 开放阅读框分析
- 5 启动子分析
- 6 CpG 岛识别
- 7 总结与答疑
- 8 引言

- 9 重复序列分析
- 10 基因识别
- 11 总结与答疑
- 12 引言
- 13 mRNA 选择性剪接
- 14 miRNA 及其靶基因预测
- 15 总结与答疑
- 16 复习思考题



CpG 岛

在基因组的某些区段，CpG 保持或高于正常概率，这些区段被称作 CpG 岛 (CpG island)。

特征

- 几乎看家基因都含有 CpG 岛（人类和小鼠分别有 55.9% 和 46.9% 的基因与 CpG 岛有密切关联）
- 一般位于基因的 5' 端区域（转录起始位点附近，有助于基因的识别），长度约 300~3000bp
- 大多数 CpG 岛是未甲基化的，未甲基化 CpG 岛说明基因可能具有潜在活性（表观遗传学中重要的作用区域，甲基化异常常常伴随着疾病的发生）
- CpG 岛中的核小体中 H1 含量低，其他组蛋白被广泛乙酰化，并具有超敏感位点

CpG 岛

在基因组的某些区段，CpG 保持或高于正常概率，这些区段被称作 CpG 岛 (CpG island)。

特征

- 几乎看家基因都含有 CpG 岛（人类和小鼠分别有 55.9% 和 46.9% 的基因与 CpG 岛有密切关联）
- 一般位于基因的 5' 端区域（转录起始位点附近，有助于基因的识别），长度约 300~3000bp
- 大多数 CpG 岛是未甲基化的，未甲基化 CpG 岛说明基因可能具有潜在活性（表观遗传学中重要的作用区域，甲基化异常常常伴随着疾病的发生）
- CpG 岛中的核小体中 H1 含量低，其他组蛋白被广泛乙酰化，并具有超敏感位点

CpG 岛 | 识别依据与判别标准

- ① CpG 岛长度：至少 200bp
- ② GC 含量：超过 50%
- ③ CpG 的观察值与预测值的比率：高于 60%
 - 观测值： $Num\ of\ CpG$
 - 预测值： $\frac{Num\ of\ C \times Num\ of\ G}{Length\ of\ sequence}$
 - 比率： $\frac{Num\ of\ CpG}{Num\ of\ C \times Num\ of\ G} \times Length\ of\ sequence$
 - 实例： $ACGCGACGCG$; $\frac{4}{4 \times 4} = \frac{4}{16} \times 10 = 2.5$
- ④ 500bp, 55%, 65%



CpG 岛 | 识别依据与判别标准

- ① CpG 岛长度：至少 200bp
- ② GC 含量：超过 50%
- ③ CpG 的观察值与预测值的比率：高于 60%
 - 观测值： $Num\ of\ CpG$
 - 预测值： $\frac{Num\ of\ C \times Num\ of\ G}{Length\ of\ sequence}$
 - 比率： $\frac{Num\ of\ CpG}{Num\ of\ C \times Num\ of\ G} \times Length\ of\ sequence$
 - 实例： $ACGCGACGCG$; $\frac{4}{4 \times 4} = \frac{4}{16} \times 10 = 2.5$
- ④ 500bp, 55%, 65%



- EMBOSS 中的 CpGPlot/CpGReport/Isochore
- CpG Island Searcher
- CpGcluster2



教学提纲

- 1 引言
- 2 DNA 组份分析与序列转换
- 3 限制酶位点分析
- 4 开放阅读框分析
- 5 启动子分析
- 6 CpG 岛识别
- 7 总结与答疑
- 8 引言

- 9 重复序列分析
- 10 基因识别
- 11 总结与答疑
- 12 引言
- 13 mRNA 选择性剪接
- 14 miRNA 及其靶基因预测
- 15 总结与答疑
- 16 复习思考题



知识点——DNA 序列的基本信息与特征信息分析

- DNA 序列基本信息分析——查戈夫法则, GC 含量, 序列转换
- 限制酶位点分析——命名, II 型的特点
- 开放阅读框分析——相位, ORF 与 CDS
- 启动子与转录因子结合位点分析——启动子结构
- CpG 岛识别——概念、判别依据及标准



教学提纲

1 引言

2 DNA 组份分析与序列转换

3 限制酶位点分析

4 开放阅读框分析

5 启动子分析

6 CpG 岛识别

7 总结与答疑

8 引言

9 重复序列分析

10 基因识别

11 总结与答疑

12 引言

13 mRNA 选择性剪接

14 miRNA 及其靶基因预测

15 总结与答疑

16 复习思考题



● 基本信息分析

- 碱基比例
- GC 含量
- 序列转换
- 寻找限制酶切位点

● 序列特征分析

● 基因识别



- 基本信息分析

- 碱基比例

- GC 含量

- 序列转换

- 寻找限制酶切位点

- 序列特征分析

- 基因识别



- 基本信息分析

- 碱基比例
- GC 含量
- 序列转换
- 寻找限制酶切位点

- 序列特征分析

- 基因识别



- 基本信息分析

- 碱基比例
- GC 含量
- 序列转换
- 寻找限制酶切位点

- 序列特征分析

- 基因识别



- 基本信息分析

- 碱基比例
- GC 含量
- 序列转换
- 寻找限制酶切位点

- 序列特征分析

- 开放阅读框的预测
- 启动子和终止子结合位点的分析

- 基因识别



● 基本信息分析

- 碱基比例
- GC 含量
- 序列转换
- 寻找限制酶切位点

● 序列特征分析

- 开放阅读框的预测
- 启动子和转录因子结合位点的分析
- CpG 岛的识别

● 基因识别



- 基本信息分析

- 碱基比例
- GC 含量
- 序列转换
- 寻找限制酶切位点

- 序列特征分析

- **开放阅读框的预测**

- 启动子和转录因子结合位点的分析
 - CpG 岛的识别

- 基因识别



- 基本信息分析

- 碱基比例
- GC 含量
- 序列转换
- 寻找限制酶切位点

- 序列特征分析

- 开放阅读框的预测
- 启动子和转录因子结合位点的分析
- CpG 岛的识别

- 基因识别



- 基本信息分析

- 碱基比例
- GC 含量
- 序列转换
- 寻找限制酶切位点

- 序列特征分析

- 开放阅读框的预测
- 启动子和转录因子结合位点的分析
- CpG 岛的识别

- 基因识别

- ◆ 简单重复序列

- ◆ 复杂识别



- 基本信息分析

- 碱基比例
- GC 含量
- 序列转换
- 寻找限制酶切位点

- 序列特征分析

- 开放阅读框的预测
- 启动子和转录因子结合位点的分析
- CpG 岛的识别

- 基因识别

- 屏蔽重复序列
- 基因识别



- 基本信息分析

- 碱基比例
- GC 含量
- 序列转换
- 寻找限制酶切位点

- 序列特征分析

- 开放阅读框的预测
- 启动子和转录因子结合位点的分析
- CpG 岛的识别

- 基因识别

- **屏蔽重复序列**
- 基因识别



- 基本信息分析

- 碱基比例
- GC 含量
- 序列转换
- 寻找限制酶切位点

- 序列特征分析

- 开放阅读框的预测
- 启动子和转录因子结合位点的分析
- CpG 岛的识别

- 基因识别

- 屏蔽重复序列
- 基因识别



教学提纲

1

引言

2

DNA 组份分析与序列转换

3

限制酶位点分析

4

开放阅读框分析

5

启动子分析

6

CpG 岛识别

7

总结与答疑

8

引言

9

重复序列分析

10

基因识别

11

总结与答疑

12

引言

13

mRNA 选择性剪接

14

miRNA 及其靶基因预测

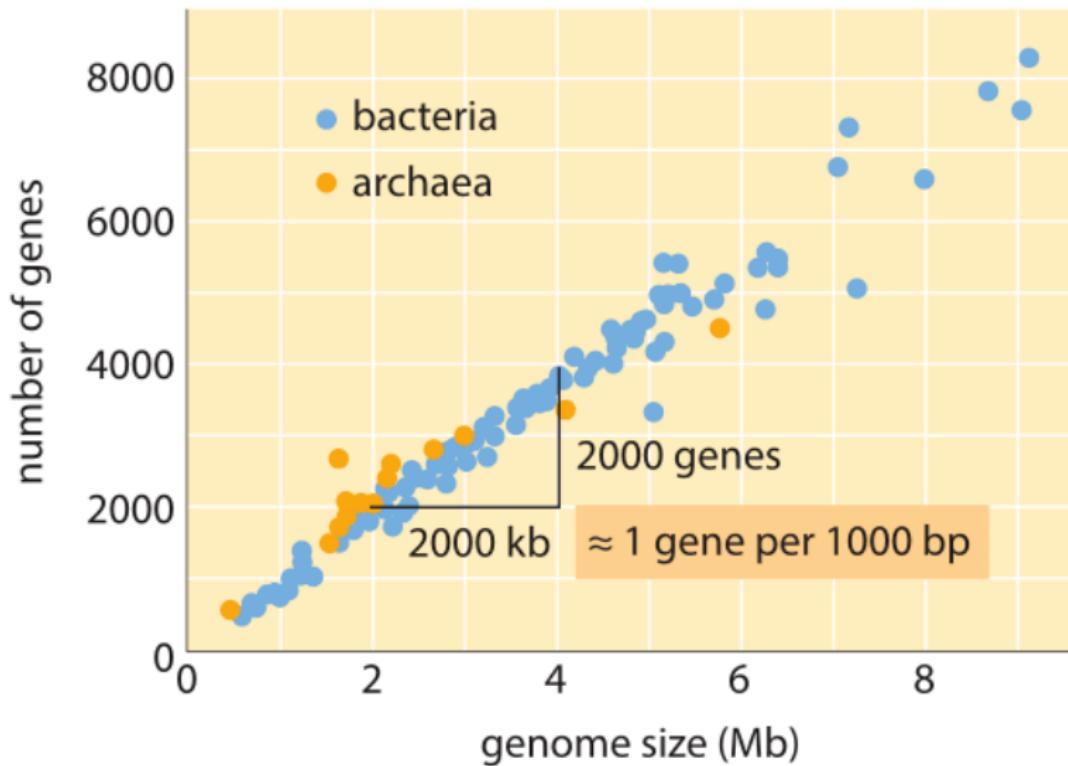
15

总结与答疑

16

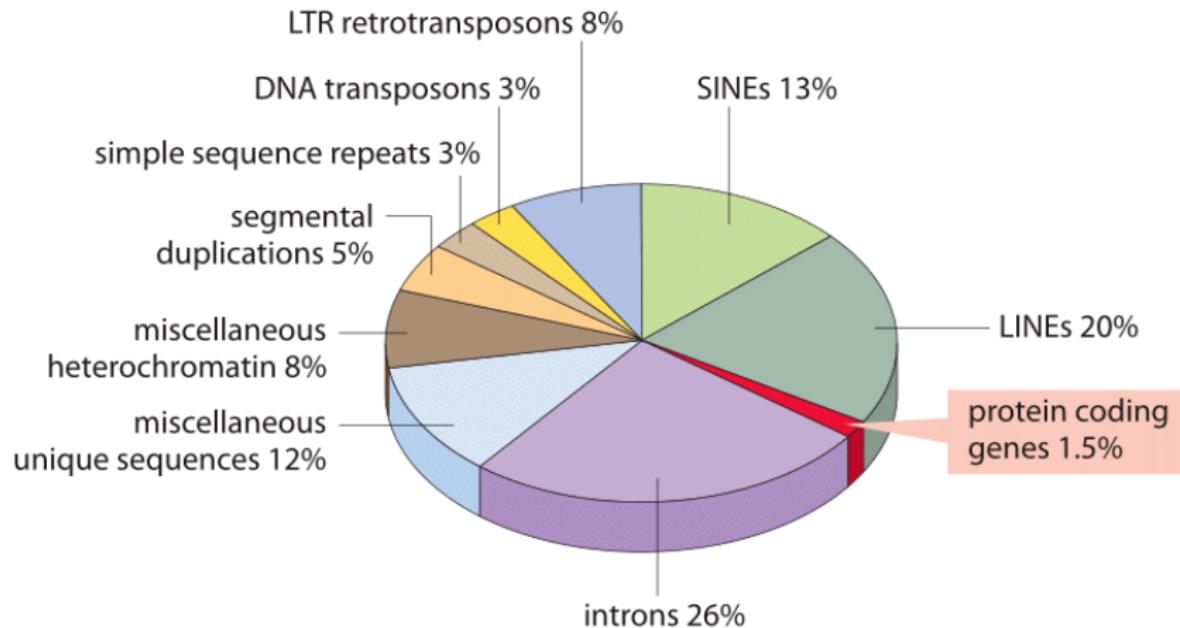
复习思考题





重复序列 | 基因组构成 | 真核

main components of the human genome



重复序列 (repetitive sequence, repeated sequence)

真核生物基因组中重复出现的核苷酸序列，一般不编码多肽，在基因组内可成簇排布，也可散布于基因组。

重复次数

- 低度重复序列 (lowly repetitive sequence) : 2~10 个拷贝
- 中度重复序列 (moderately repetitive sequence) : 重复几十次到几千次，平均长 300bp
- 高度重复序列 (highly repetitive sequence) : 重复几百万次，少于 10 个核苷酸残基组成的短片段



重复序列 (repetitive sequence, repeated sequence)

真核生物基因组中重复出现的核苷酸序列，一般不编码多肽，在基因组内可成簇排布，也可散布于基因组。

重复次数

- 低度重复序列 (lowly repetitive sequence) : 2~10 个拷贝
- 中度重复序列 (moderately repetitive sequence) : 重复几十次到几千次，平均长 300bp
- 高度重复序列 (highly repetitive sequence) : 重复几百万次，少于 10 个核苷酸残基组成的短片段



组织形式

- 串联重复序列：成簇存在于染色体的特定区域，依重复单位的长度分类
 - 卫星 DNA (satellite DNA) : 5~200bp, 几百万个拷贝, 着丝粒部位, 高度重复序列
 - 小卫星 (minisatellite, VNTR) : 10~100bp 的基本单位, 总长不超过 20kb, 重复次数高度变异, 靠近端粒的位置
 - 微卫星 (microsatellite, SSR, STR) : 2~10bp, 长度 50~100bp, STR 遗传多态性, 内含子
- 散在重复序列：比较均匀地分散于染色体的各位点上，中度重复序列
 - 短散在重复序列 (SINE) : 500bp 以下, 重复拷贝数达 10 万以上；非自主转座的反转录转座子；来源于 RNA 聚合酶 III 的转录产物；Alu (300bp, 100 万个拷贝)
 - 长散在重复序列 (LINE) : 1000bp 以上, 上万份拷贝；可以自主转座的反转录转座子；来源于 RNA 聚合酶 II 的转录产物；L1 (6100bp, 3500 个拷贝)

重复序列 | 分类

Sequence types	Repeat size(bp)	Array size (kb)	Copy number ^a	Functions, features of family members
Satellites — large tandem arrays		10–25% of total DNA		
Microsatellite	2–5	0.2–0.5	3×10^3	Repeat expansion causes cancer
Minisatellite	~15	0.5–3	10^3	Changes in sequence cause cancer
Satellite	5–100	100,000	10^7	Centromere and telomere function
Megasatellite	4–10 kb	30–100	30–100	?
Interpersed elements		35–40% of total DNA		
Retrotransposons				
<i>LTR-containing elements</i>				
<i>copia</i> ² , <i>gypsy</i> ²	~5 kb	NA	20–60	Can be found as free circular DNA Horizontal transfer of genes; can infect germline cells
Yeast Ty	6.3 kb	NA	40	Ty1 and Ty3 transpose specifically to genes transcribed by RNA polymerase III; Repair of chromosomal breaks
<i>Poly-A elements</i>				
LINE1 (L1)	1–7 kb	NA	$\sim 10^5$	Mutant sequences can promote cancer Some provide polyadenylation signals Some copies mobile
HeT-A, TART ²	6–10 kb	5–10	$\sim 10^4$	Maintenance of telomeres
<i>SINEs</i>				
Alu	300	NA	$\sim 10^6$	Retinoic acid receptor-binding site Enhancer of gene activity Silencer of gene activity Negative calcium response element Alters protein synthesis Insertion can cause disease



数据库

- Repbase Update (RU) : 真核生物 DNA 重复序列数据库
- L1Base : L1 数据库
- STRBase : STR 数据库

分析工具

- RepeatMasker : 识别、分类和屏蔽重复序列
 - Cross_match : 速度慢、精度高
 - ABBLast : 速度快、精度略低
 - RMBlast : NCBI Blast 的兼容版
 - HMMER : 只适用于人类基因组序列



数据库

- Repbase Update (RU) : 真核生物 DNA 重复序列数据库
- L1Base : L1 数据库
- STRBase : STR 数据库

分析工具

- RepeatMasker : 识别、分类和屏蔽重复序列
 - Cross_match : 速度慢、精度高
 - ABBLast : 速度快、精度略低
 - RMBlast : NCBI Blast 的兼容版
 - HMMER : 只适用于人类基因组序列



重复序列 | RepeatMasker

```
=====
file name: sequence.fasta
sequences:          1
total length:      50830 bp (50830 bp excl N/X-run)
GC level:          36.75 %
bases masked:      4990 bp ( 9.82 %)
=====
                                number of           length   percentage
                                elements*        occupied   of sequence
-----
SINEs:                  1                 32 bp    0.06 %
  ALUs:                  0                 0 bp     0.00 %
  MIRs:                  0                 0 bp     0.00 %
LINEs:                  3                 142 bp   0.28 %
  LINE1:                2                 93 bp    0.18 %
  LINE2:                1                 49 bp    0.10 %
  L3/CR1:               0                 0 bp     0.00 %
LTR elements:           0                 0 bp     0.00 %
  ERVL:                 0                 0 bp     0.00 %
  ERVL-MaLRs:           0                 0 bp     0.00 %
  ERV_classI:           0                 0 bp     0.00 %
  ERV_classII:          0                 0 bp     0.00 %
DNA elements:           7                 1516 bp   2.98 %
  hAT-Charlie:          0                 0 bp     0.00 %
  TcMar-Tigger:          7                 1516 bp   2.98 %
Unclassified:           0                 0 bp     0.00 %
Total interspersed repeats: 1690 bp  3.32 %
```



教学提纲

1

引言

2

DNA 组份分析与序列转换

3

限制酶位点分析

4

开放阅读框分析

5

启动子分析

6

CpG 岛识别

7

总结与答疑

8

引言

9

重复序列分析

10

基因识别

11

总结与答疑

12

引言

13

mRNA 选择性剪接

14

miRNA 及其靶基因预测

15

总结与答疑

16

复习思考题



基因 (gene)

产生一条多肽链或功能 RNA 所需的全部核苷酸序列。一段具有特定功能和结构的连续的 DNA 片段，携带着遗传信息，是编码蛋白质或 RNA 分子遗传信息、控制性状的基本遗传单位。

一个完整的基因，不仅包括编码区，还包括 5' 末端和 3' 末端长度不等的特异性序列。

基因识别 (gene prediction, gene finding)

使用生物学实验或计算机等手段识别 DNA 序列上的具有生物学特征的片段，是基因组研究的基础。

对象主要是蛋白质编码基因（还有 RNA 基因和调控因子等）。



基因 (gene)

产生一条多肽链或功能 RNA 所需的全部核苷酸序列。一段具有特定功能和结构的连续的 DNA 片段，携带着遗传信息，是编码蛋白质或 RNA 分子遗传信息、控制性状的基本遗传单位。

一个完整的基因，不仅包括编码区，还包括 5' 末端和 3' 末端长度不等的特异性序列。

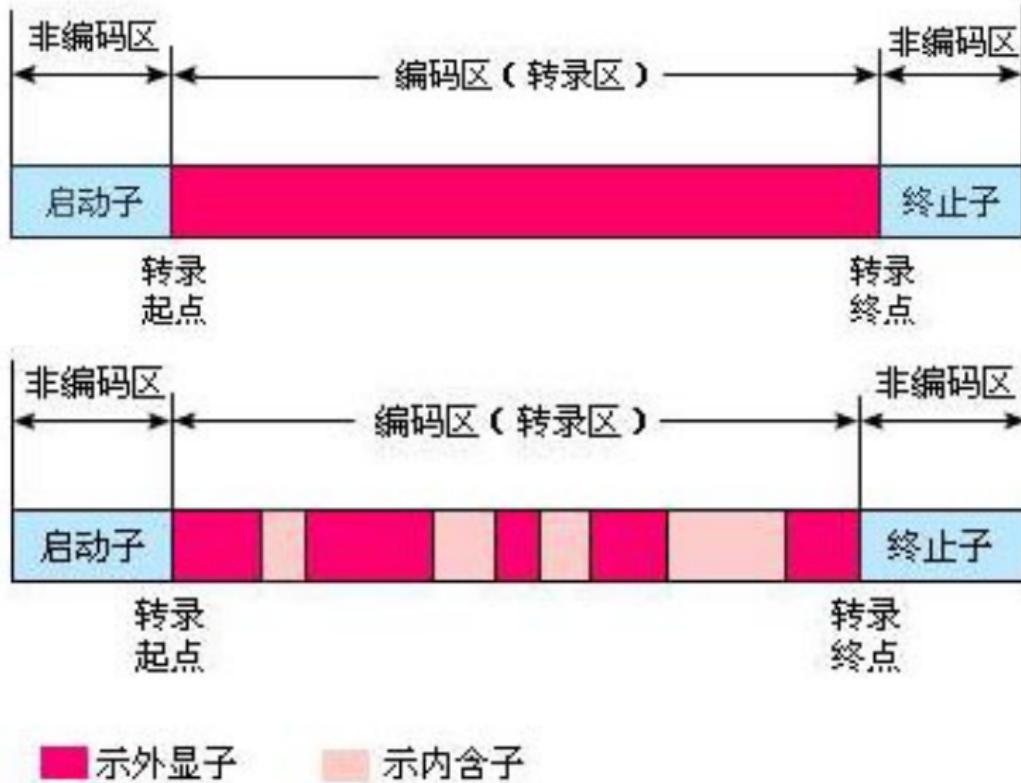
基因识别 (gene prediction, gene finding)

使用生物学实验或计算机等手段识别 DNA 序列上的具有生物学特征的片段，是基因组研究的基础。

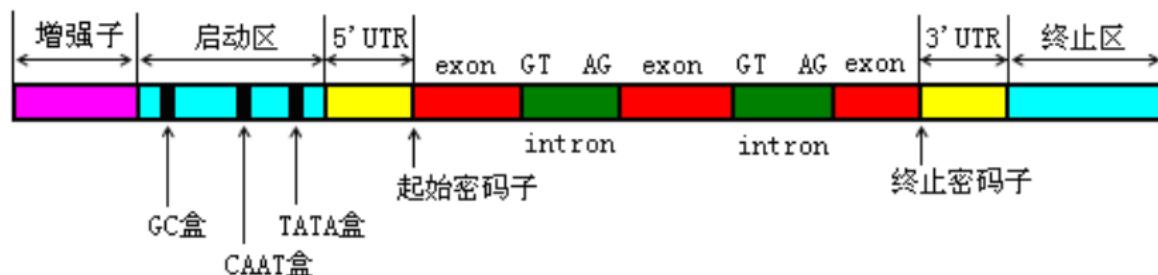
对象主要是蛋白质编码基因（还有 RNA 基因和调控因子等）。



基因识别 | 基因结构 | 连续 vs. 不连续



基因识别 | 基因结构



- ① 间接识别法 (Extrinsic Approach) : 利用已知的 mRNA 或蛋白质序列为线索在 DNA 序列中搜寻所对应的片段
- ② 从头计算法 (*Ab Initio Approach*) : 基因预测 (通常仍需实验证实)。基于基因的两种类型的特征：
 - “信号” : 由一些特殊的序列构成, 通常预示着周围存在着一个基因
 - “内容” : 蛋白质编码基因所具有的某些统计学特征
- ③ 比较基因组学的方法 : 自然选择的力量使得基因和 DNA 序列上具有生物学功能的片段较其他部分有较慢的变异速率, 在前者的变异更有可能对生物体的生存产生负面影响, 因而难以得到保存



信号

- 不连续的局部序列模体，一般都有一致性序列（consensus sequence）
- 启动子，剪接供体和受体位点，起始和终止密码子，polyA 位点

内容

- 不同长度的扩展序列，没有一致性序列，但具有把自己与周围 DNA 区分开来的保守特征
- 密码子使用偏好性（codon usage bias），双联密码子出现频率，基因组等值区（isochore）



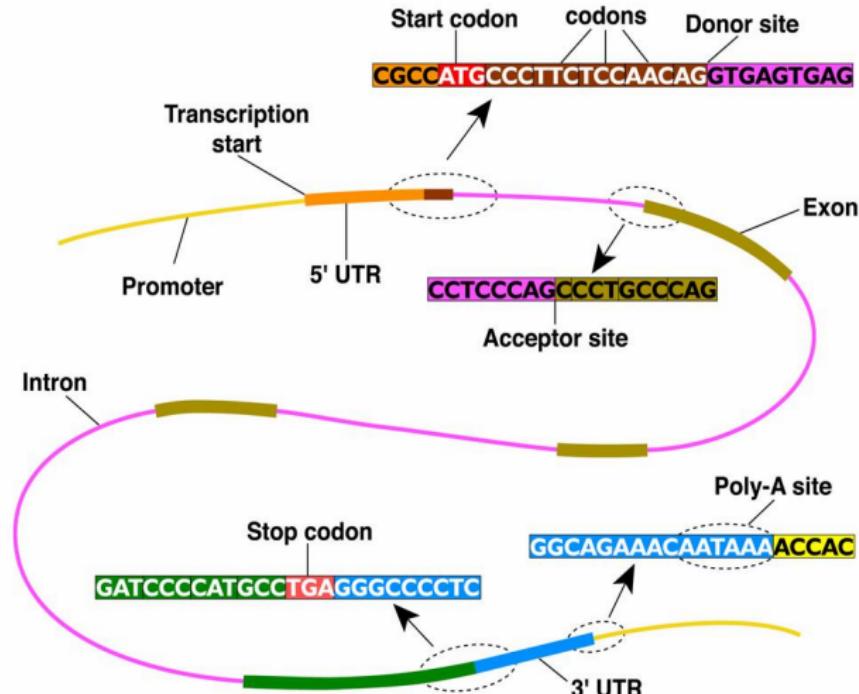
信号

- 不连续的局部序列模体，一般都有一致性序列（consensus sequence）
- 启动子，剪接供体和受体位点，起始和终止密码子，polyA 位点

内容

- 不同长度的扩展序列，没有一致性序列，但具有把自己与周围 DNA 区分开来的保守特征
- 密码子使用偏好性（codon usage bias），双联密码子出现频率，基因组等值区（isochore）





基因识别 | 基因预测 | 内容 | 密码子使用偏好性

CODON USAGE IN *E. COLI* GENES¹

	Codon	Amino acid ²	% ³	Ratio ⁴	Codon	Amino acid	%	Ratio	Codon	Amino acid	%	Ratio	Codon	Amino acid	%	Ratio	
U	UUU	Phe (F)	1.9	0.51	UCU	Ser (S)	1.1	0.19	UAU	Tyr (Y)	1.6	0.53	UGU	Cys (C)	0.4	0.43	U
	UUC	Phe (F)	1.8	0.49	UCC	Ser (S)	1.0	0.17	UAC	Tyr (Y)	1.4	0.47	UGC	Cys(C)	0.6	0.57	C
	UUA	Leu (L)	1.0	0.11	UCA	Ser (S)	0.7	0.12	UAA	STOP	0.2	0.62	UGA	STOP	0.1	0.30	A
	UUG	Leu (L)	1.1	0.11	UCG	Ser (S)	0.8	0.13	UAG	STOP	0.03	0.09	UGG	Trp (W)	1.4	1.00	G
C	CUU	Leu (L)	1.0	0.10	CCU	Pro (P)	0.7	0.16	CAU	His (H)	1.2	0.52	CGU	Arg (R)	2.4	0.42	U
	CUC	Leu (L)	0.9	0.10	CCC	Pro (P)	0.4	0.10	CAC	His (H)	1.1	0.48	CGC	Arg (R)	2.2	0.37	C
	CUA	Leu (L)	0.3	0.03	CCA	Pro (P)	0.8	0.20	CAA	Gln (Q)	1.3	0.31	CGA	Arg (R)	0.3	0.05	A
	CUG	Leu (L)	5.2	0.55	CCG	Pro (P)	2.4	0.55	CAG	Gln (Q)	2.9	0.69	CGG	Arg (R)	0.5	0.08	G
A	AUU	Ile (I)	2.7	0.47	ACU	Thr (T)	1.2	0.21	AAU	Asn (N)	1.6	0.39	AGU	Ser (S)	0.7	0.13	U
	AUC	Ile (I)	2.7	0.46	ACC	Thr (T)	2.4	0.43	AAC	Asn (N)	2.6	0.61	AGC	Ser (S)	1.5	0.27	C
	AUA	Ile (I)	0.4	0.07	ACA	Thr (T)	0.1	0.30	AAA	Lys (K)	3.8	0.76	AGA	Arg (R)	0.2	0.04	A
	AUG	Met (M)	2.6	1.00	ACG	Thr (T)	1.3	0.23	AAG	Lys (K)	1.2	0.24	AGG	Arg (R)	0.2	0.03	G
G	GUU	Val (V)	2.0	0.29	GCU	Ala (A)	1.8	0.19	GAU	Asp (D)	3.3	0.59	GGU	Gly (G)	2.8	0.38	U
	GUC	Val (V)	1.4	0.20	GCC	Ala (A)	2.3	0.25	GAC	Asp (D)	2.3	0.41	GGC	Gly (G)	3.0	0.40	C
	GUА	Val (V)	1.2	0.17	GCA	Ala (A)	2.1	0.22	GAA	Glu (E)	4.4	0.70	GGA	Gly (G)	0.7	0.09	A
	GUG	Val (V)	2.4	0.34	GCG	Ala (A)	3.2	0.34	GAG	Glu (E)	1.9	0.30	GGG	Gly (G)	0.9	0.13	G
	U				C				A				G				



Codon	Human	Drosophila	E. coli
Arginine:			
AGA	22 %	10 %	1 %
AGG	23 %	6 %	1 %
CGA	10 %	8 %	4 %
CGC	22 %	49 %	39 %
CGG	14 %	9 %	4 %
CGU	9 %	18 %	49 %
Total number of arginine codons	2403	506	149
Total number of genes	195	46	149



信号

启动子序列（Pribnow 盒），转录因子结合位点

内容

连续的开放阅读框，统计学特征

总结

信号容易识别，内容容易判别，预测能达到相对较高的精度



信号

启动子序列（Pribnow 盒），转录因子结合位点

内容

连续的开放阅读框，统计学特征

总结

信号容易识别，内容容易判别，预测能达到相对较高的精度



信号

启动子（TATA box, CAAT box, GC box），供体和受体位点，起始和终止密码子，polyA 信号序列，CpG 岛

内容

密码子使用偏好性，双联密码子出现频率，基因组等值区，核苷酸周期性规律

总结

- 综合信号信息确定外显子的边界，识别编码区域
- 通过内容统计值区分外显子、内含子和基因间区域
- 信号复杂，内容难判别，预测相当有挑战性
- 联合信号和内容检测以及同源性搜索，提高识别效率

信号

启动子（TATA box, CAAT box, GC box），供体和受体位点，起始和终止密码子，polyA 信号序列，CpG 岛

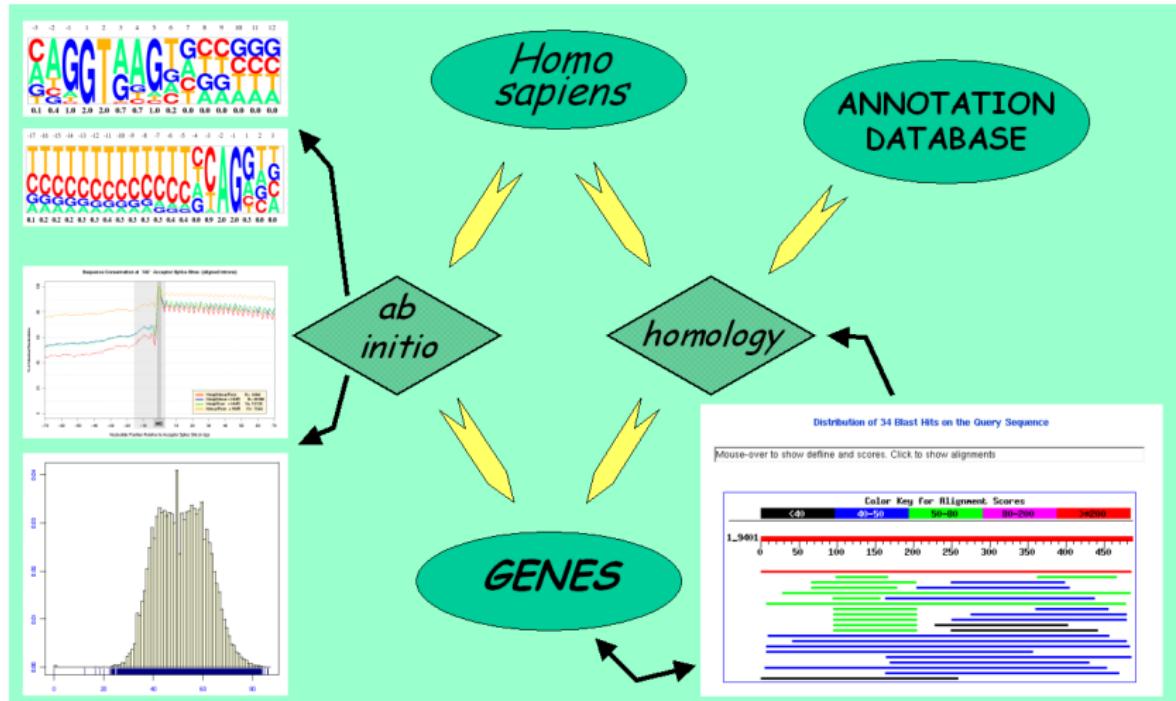
内容

密码子使用偏好性，双联密码子出现频率，基因组等值区，核苷酸周期性规律

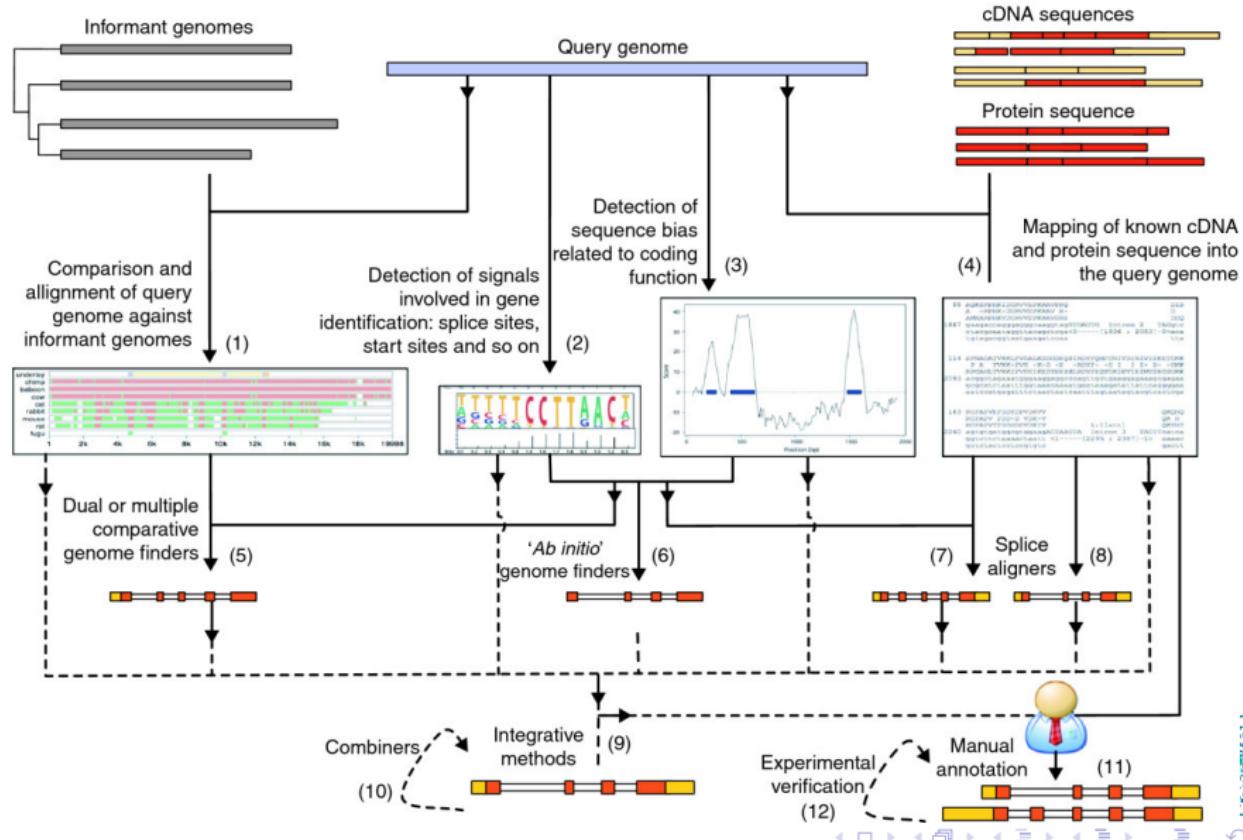
总结

- 综合信号信息确定外显子的边界，识别编码区域
- 通过内容统计值区分外显子、内含子和基因间区域
- 信号复杂，内容难判别，预测相当有挑战性
- 联合信号和内容检测以及同源性搜索，提高识别效率

基因识别 | 真核基因



基因识别 | 策略



基因识别 | 工具列表

Program	Class*	URL
BLAST [61]	4	http://blast.ncbi.nlm.nih.gov/Blast.cgi
Twinscan [62]	5	http://mblab.wustl.edu/
Sgp2 [63]	5	http://genome.imim.es/software/sgp2/
SLAM [64]	5	http://bio.math.berkeley.edu/slam/mouse/
DoubleScan [65]	5	http://www.sanger.ac.uk/Software/analysis/doublescan/
Augustus [66]	6	http://augustus.gobics.de/
GeneID [67]	6	http://genome.imim.es/software/geneid/
Genscan [68]	6	http://genes.mit.edu/GENSCANinfo.html
GlimmerHMM [69]	6	http://www.ccb.umd.edu/software/GlimmerHMM/
GeneMark [70]	6	http://exon.gatech.edu/GeneMark/
GenomeScan [71]	7	http://genes.mit.edu/genomescan.html
N-SCAN(_EST) [72]	7, 5	http://mblab.wustl.edu/



基因识别 | 工具列表

Name	Description	Species
ATGpr	identifying translational initiation sites in cDNA sequences	
AUGUSTUS	Eukaryote gene predictor	Eukaryotes
BGF	hidden Markov model (HMM) and dynamic programming based <i>ab initio</i> gene prediction program	
DIOGENES	a system for fast detection of coding regions in short genomic sequences	
Dragon Promoter Finder	software for recognition of vertebrate RNA Polymerase II promoters	
EUGENE	gene finding for <i>Arabidopsis thaliana</i>	<i>Arabidopsis thaliana</i>
FGENESH	HMM-based gene structure prediction (multiple genes, both chains)	Eukaryotes
FRAMED	find genes and frameshift in G+C rich prokaryotic sequences	Prokaryotes
GENIUS	linking ORFs in complete genomes to protein 3D structures	
gened	program to predict genes, exons, splice sites and other signals along a DNA sequence	Eukaryotes
GENEPARSER	Parse a DNA sequence into introns and exons	
GeneMark	family of gene prediction programs	Prokaryotes+Eukaryotes
GeneTack	prediction of genes with frameshifts in prokaryotic genomes	Prokaryotes
GENOMESCAN	predicts locations and exon-intron structures of genes in genomic sequences from a variety of organisms.	
GENSCAN	finding genes using Fourier transform	
GLIMMER	finding genes in microbial DNA	Prokaryotes
GLIMMERHMM	Eukaryotic gene-finding System	Eukaryotes
GraalEXP	predicts exons, genes, promoters, polyAs, CpG Islands, EST similarities, and	



- GeneMarkS : 迭代隐马尔科夫模型
- Glimmer : 插入式马尔科夫模型
- GENSCAN : 广义隐马尔科夫模型
- GRAIL : 人工神经网路
- FGENESH : HMM-based gene structure prediction
- [List of gene prediction software\(Wikipedia\)](#)
- Computational prediction of eukaryotic protein-coding genes, Box 2, Useful internet resources



基因识别 | GENSCAN | 结果

Gn.	Ex	Type	S	.Begin	...End	.Len	Fr	Ph	I/Ac	Do/T	CodRg	P....	Tscr..
1.00		Prom	+	1653	1692	40							-1.16
1.01		Init	+	5215	5266	52	0	1	83	75	151	0.925	12.64
1.02		Intr	+	5395	5562	168	2	0	89	75	163	0.895	15.02
1.03		Intr	+	11738	11899	162	0	0	74	113	101	0.990	11.15
1.04		Intr	+	12188	12424	237	0	0	71	86	197	0.662	15.39
1.05		Intr	+	14288	14623	336	0	0	82	98	263	0.986	22.19
1.06		Intr	+	17003	17203	201	0	0	116	86	102	0.976	12.06
1.07		Intr	+	17741	17859	119	0	2	78	109	51	0.984	6.38
1.08		Intr	+	18197	18264	68	1	2	103	72	81	0.541	5.70



Type

- Init: initial exon; Intr: internal exon; Term: terminal exon
- Sngl: single-exon gene; Prom: promoter region; PlyA: polyA signal

P

- 可能性极高的外显子 ($P>0.99$) : 预测结果几乎完全与真实注释的外显子相吻合, 准确度高达 97.7%
- 中等或高可能性的外显子 ($0.50 < P < 0.99$) : 预测结果在大多数情况下与实际相吻合, 准确度比 P 值略小 ($P>0.90$ 的准确度为 88%)
- 低可能性的外显子 ($P<0.50$) : 不可靠, 使用时要小心, 甚至可以将其忽略

Type

- Init: initial exon; Intr: internal exon; Term: terminal exon
- Sngl: single-exon gene; Prom: promoter region; PolyA: polyA signal

P

- 可能性极高的外显子 ($P>0.99$) : 预测结果几乎完全与真实注释的外显子相吻合, 准确度高达 97.7%
- 中等或高可能性的外显子 ($0.50 < P < 0.99$) : 预测结果在大多数情况下与实际相吻合, 准确度比 P 值略小 ($P>0.90$ 的准确度为 88%)
- 低可能性的外显子 ($P<0.50$) : 不可靠, 使用时要小心, 甚至可以直接将其忽略

教学提纲

1

引言

2

DNA 组份分析与序列转换

3

限制酶位点分析

4

开放阅读框分析

5

启动子分析

6

CpG 岛识别

7

总结与答疑

8

引言

9

重复序列分析

10

基因识别

11

总结与答疑

12

引言

13

mRNA 选择性剪接

14

miRNA 及其靶基因预测

15

总结与答疑

16

复习思考题



知识点——重复序列和基因识别

- 重复序列——分类
- 基因识别——原核和真核的基因结构，基因识别方法



教学提纲

1

引言

2

DNA 组份分析与序列转换

3

限制酶位点分析

4

开放阅读框分析

5

启动子分析

6

CpG 岛识别

7

总结与答疑

8

引言

9

重复序列分析

10

基因识别

11

总结与答疑

12

引言

13

mRNA 选择性剪接

14

miRNA 及其靶基因预测

15

总结与答疑

16

复习思考题



● DNA 序列分析

- 基本信息
- 序列特征
- 基因识别

● RNA 序列分析



- DNA 序列分析

- 基本信息
- 序列特征
- 基因识别

- RNA 序列分析



- DNA 序列分析

- 基本信息
- **序列特征**
- 基因识别

- RNA 序列分析



- DNA 序列分析

- 基本信息
- 序列特征
- 基因识别

- RNA 序列分析

- * rRNA 选择性剪接

- * tRNA 识别密码子



- DNA 序列分析

- 基本信息
- 序列特征
- 基因识别

- RNA 序列分析

- mRNA 选择性剪接
- miRNA 与靶基因



- DNA 序列分析

- 基本信息
- 序列特征
- 基因识别

- RNA 序列分析

- mRNA 选择性剪接
- miRNA 与靶基因



- DNA 序列分析

- 基本信息
- 序列特征
- 基因识别

- RNA 序列分析

- mRNA 选择性剪接
- miRNA 与靶基因

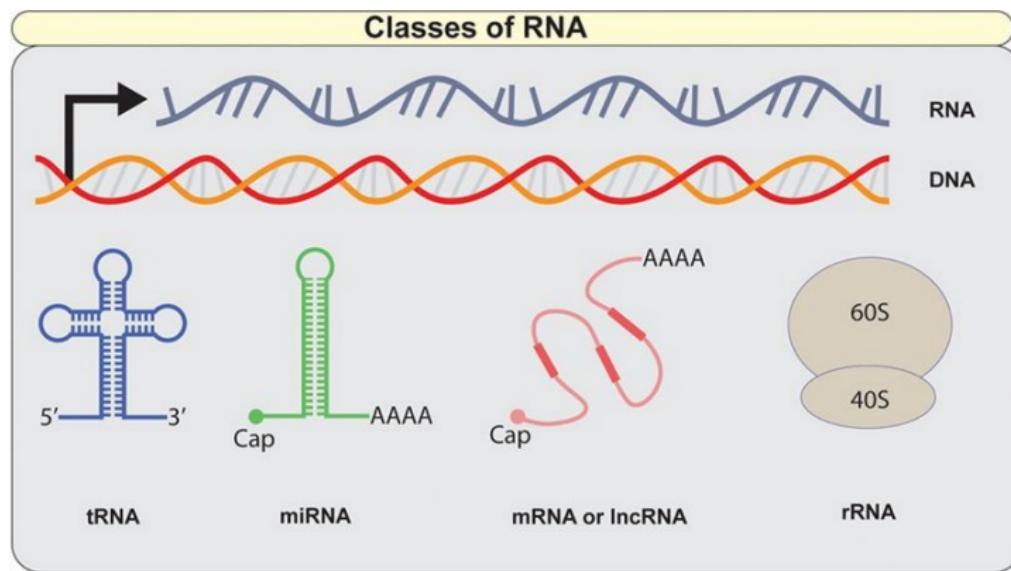


① 编码 RNA

- mRNA

② 非编码 RNA

- tRNA、rRNA
- miRNA、siRNA、lncRNA



Location and functions of different classes of RNA molecules

Class of RNA	Cell type	Location of function in eukaryotic cells ^a	Function
Ribosomal RNA (rRNA)	Bacterial and eukaryotic	Cytoplasm	Structural and functional components of the ribosome
Messenger RNA (mRNA)	Bacterial and eukaryotic	Nucleus and cytoplasm	Carries genetic code for proteins
Transfer RNA (tRNA)	Bacterial and eukaryotic	Cytoplasm	Helps incorporate amino acids into polypeptide chain
Small nuclear RNA (snRNA)	Eukaryotic	Nucleus	Processing of pre-mRNA
Small nucleolar RNA (snoRNA)	Eukaryotic	Nucleus	Processing and assembly of rRNA
Small cytoplasmic RNA (scRNA)	Eukaryotic	Cytoplasm	Variable
MicroRNA (miRNA)	Eukaryotic	Cytoplasm	Inhibits translation of mRNA
Small interfering RNA (siRNA)	Eukaryotic	Cytoplasm	Triggers degradation of other RNA molecules

^aAll eukaryotic RNAs are transcribed in the nucleus.

引言 | RNA | ncRNA

Non-coding RNA	Length (nt)	Species	Function
Ribosomal RNA (rRNA)	120~4700	All	Translation
Transfer RNA (tRNA)	70~100	All	Translation
Small nuclear RNA (snRNA)	70~350	Eukaryote	Splicing, mRNA processing
Small nucleolar RNA (snoRNA)	70~300	Eukaryote, archaea	RNA modification, rRNA processing
miRNA	21~25	Eukaryote	Translational regulation
siRNA	21~25	Eukaryote	Protection against viral infection
piRNA	24~30	Eukaryote	Genome stabilization
Long ncRNA	several hundreds~ several hundred thousands	Eukaryote	Transcription, splicing, transport regulation



教学提纲

1

引言

2

DNA 组份分析与序列转换

3

限制酶位点分析

4

开放阅读框分析

5

启动子分析

6

CpG 岛识别

7

总结与答疑

8

引言

9

重复序列分析

10

基因识别

11

总结与答疑

12

引言

13

mRNA 选择性剪接

14

miRNA 及其靶基因预测

15

总结与答疑

16

复习思考题



剪接 (splicing)

又称拼接，指基因信息在转录后的一种修饰，即将内含子移除及合并外显子，是真核生物的信使 RNA 前体 (precursor messenger RNA) 变成成熟 mRNA 的过程之一。

选择性剪接 (alternative splicing)

又称可变剪接，指用不同的剪接方式（选择不同的剪接位点组合）从一个 mRNA 前体产生不同的 mRNA 剪接异构体的过程。



选择性剪接 | 剪接与选择性剪接

剪接 (splicing)

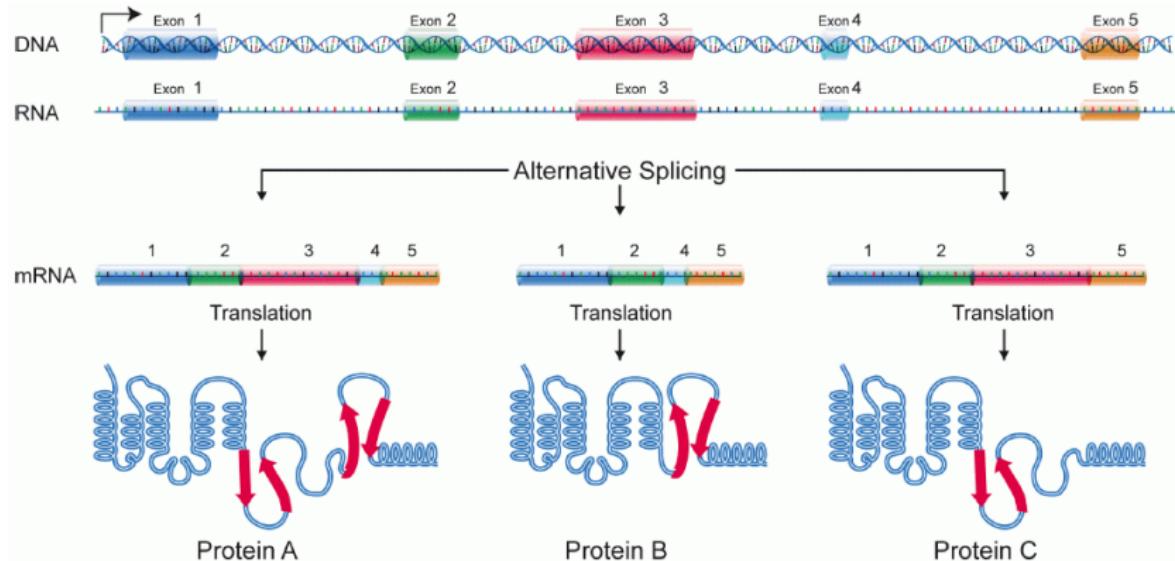
又称拼接，指基因信息在转录后的一种修饰，即将内含子移除及合并外显子，是真核生物的信使 RNA 前体 (precursor messenger RNA) 变成成熟 mRNA 的过程之一。

选择性剪接 (alternative splicing)

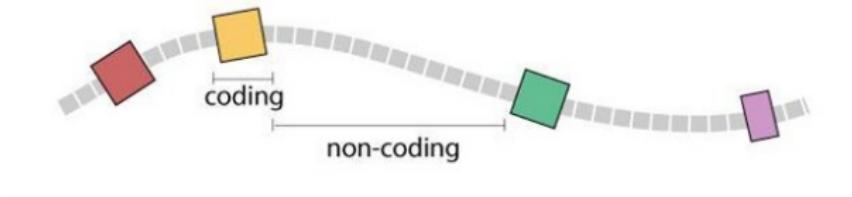
又称可变剪接，指用不同的剪接方式（选择不同的剪接位点组合）从一个 mRNA 前体产生不同的 mRNA 剪接异构体的过程。



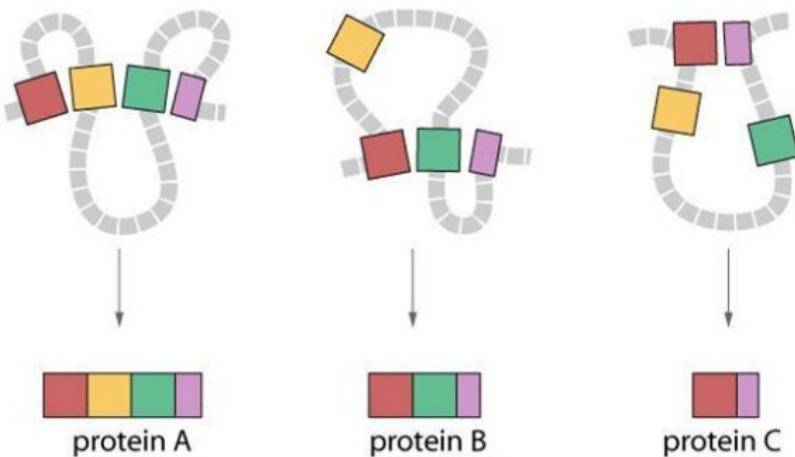
选择性剪接 | 剪接



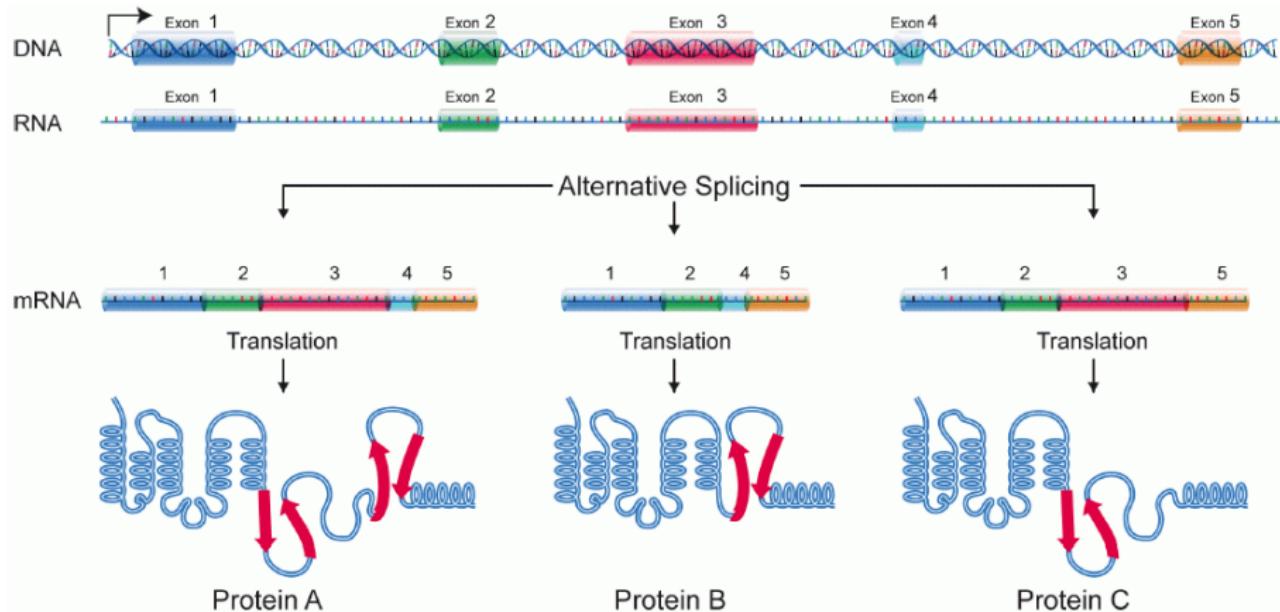
选择性剪接 | 模式图



splice variants lead to protein diversity



选择性剪接 | 模式图

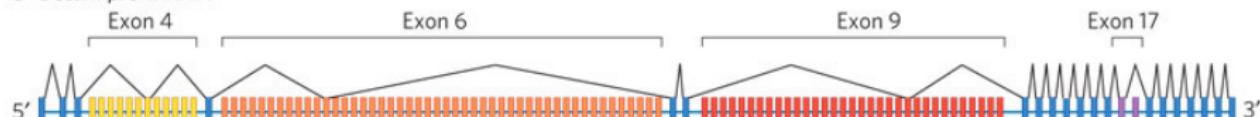


选择性剪接 | 实例

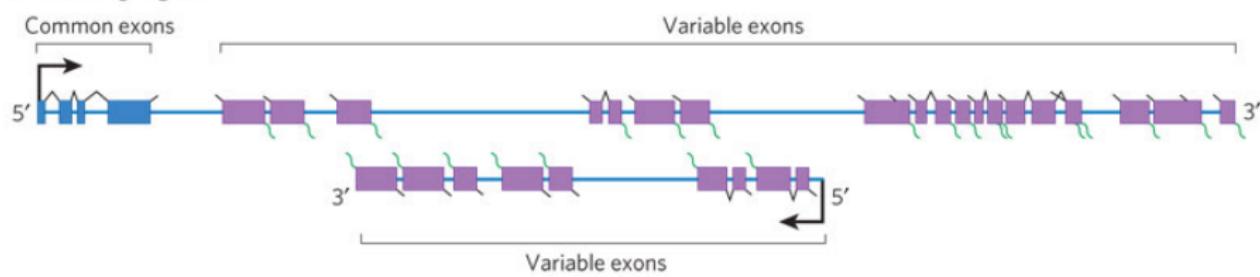
a KCNMA1 pre-mRNA



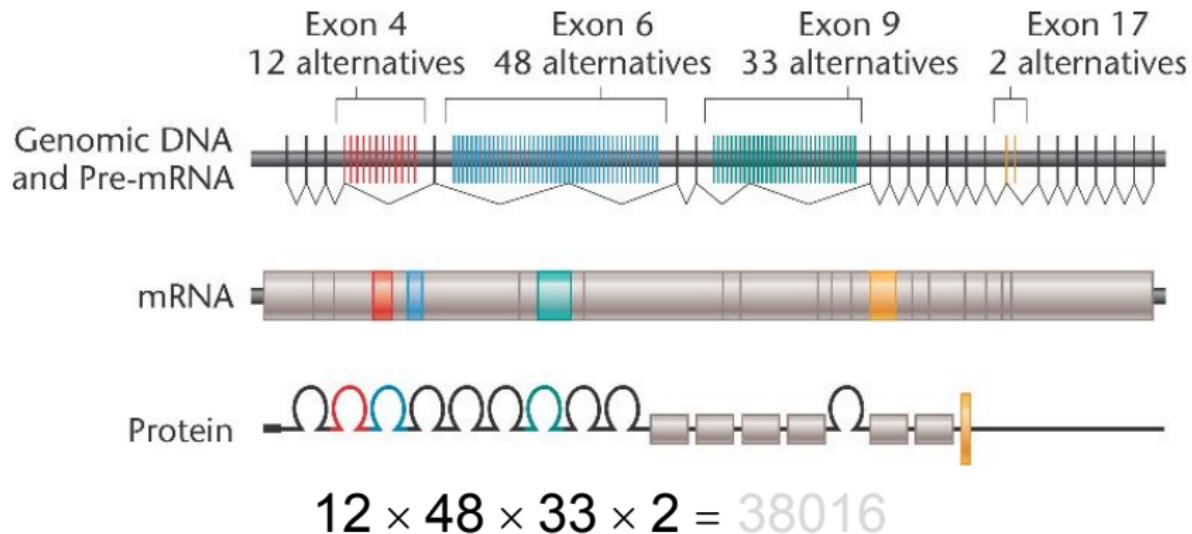
b Dscam pre-mRNA



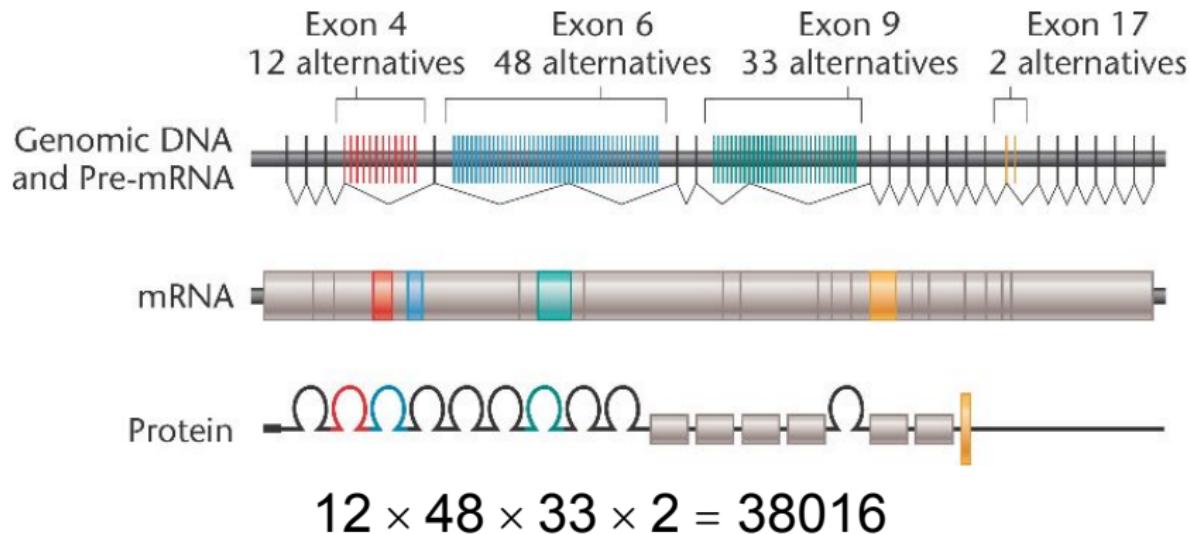
c mod(mdg4) gene



选择性剪接 | 实例 | *Dscam*



选择性剪接 | 实例 | *Dscam*

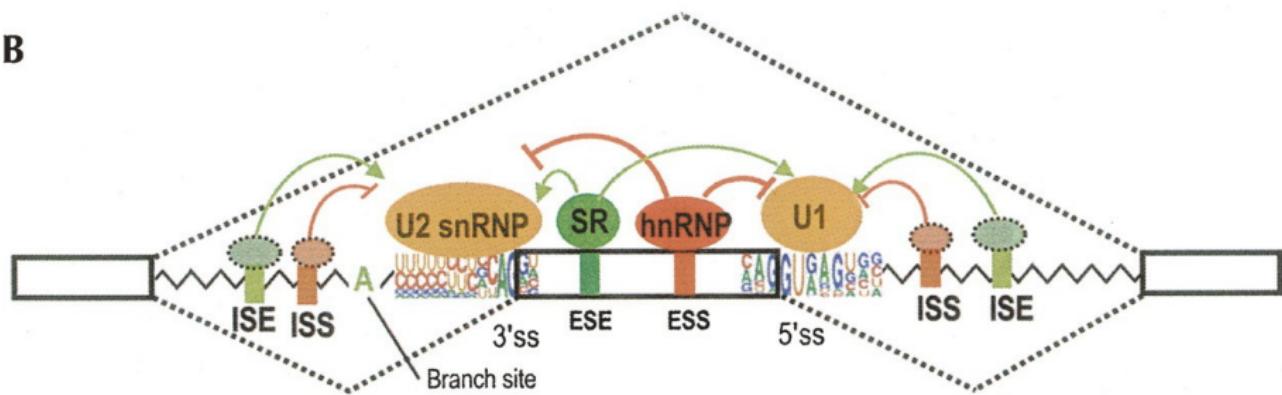


- 剪接因子与调节蛋白相互作用
- 剪接体的核心部分包括一组小核 RNA (snRNA) 以及与之结合的蛋白质，它们以严格的程序组装成剪接体
 - snRNA 成员分别为 U1、U2、U4、U5 和 U6，长度在 106(U6)~185(U2) 个核苷酸之间
 - snRNA 与蛋白质结合在一起形成小核核糖核蛋白 (snRNP)
- 剪接因子依据结合在 RNA 上的位置及作用方式，可以分为
 - 外显子增强子 (exonic splicing enhancer, ESE)
 - 外显子抑制子 (exonic splicing silencer, ESS)
 - 内含子增强子 (intronic splicing enhancer, ISE)
 - 内含子抑制子 (intronic splicing silencer, ISS)

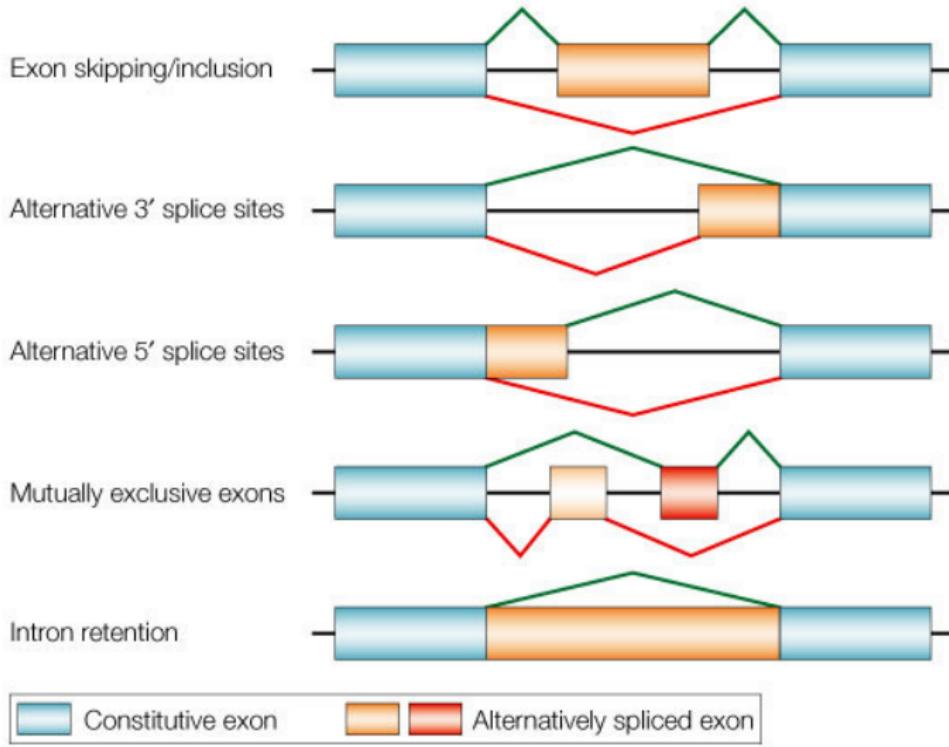


选择性剪接 | 调控

B

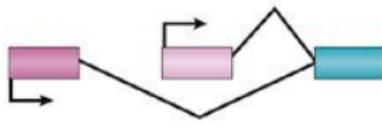


选择性剪接 | 机制 | 五种

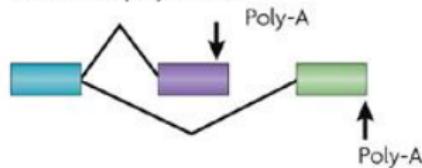


选择性剪接 | 机制 | 七种

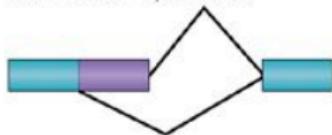
Alternative promoters



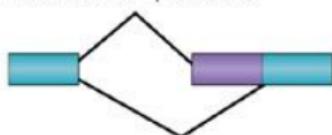
Alternative poly-A sites



Alternative 5' splice sites



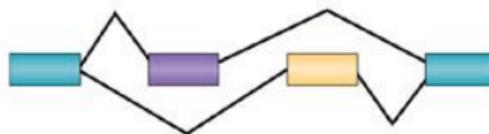
Alternative 3' splice sites



Cassette exon



Mutually exclusive exons

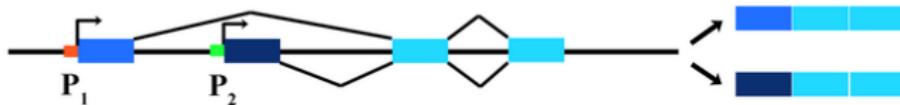


Retained intron

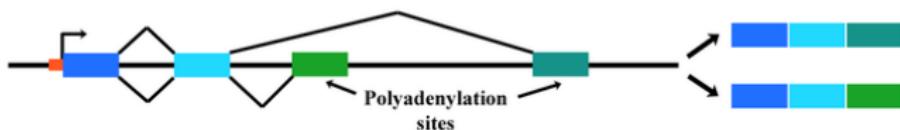


选择性剪接 | 机制 | 实例

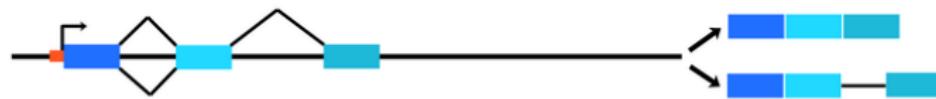
(a) Alternative selection of promoters (e.g., *myosin* primary transcript)



(b) Alternative selection of cleavage/polyadenylation sites (e.g., *tropomyosin* transcript)



(c) Intron retaining mode (e.g., *transposase* primary transcript)

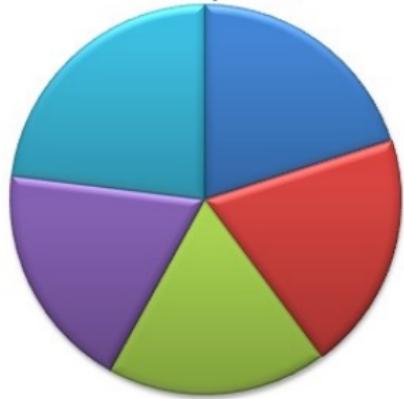


(d) Exon cassette mode (e.g., *troponin* primary transcript)



选择性剪接 | 机制 | 使用频率

Drosophila



Human



- skipped exon
- alt. donor
- alt. acceptor
- retained intron
- other



两大类数据库（依数据来源）

- 基于文献报道（收集整理实验数据和文献报道）
- 基于 EST 数据（EST 与基因组或 DNA、mRNA 比对）

数据库与工具

- ASTD = ASD (= AEDB + AltExtron + AltSplice) + ATD
- ASAP
- ESEfinder
- RESCUE-ESE
- ASPicDB



两大类数据库（依数据来源）

- 基于文献报道（收集整理实验数据和文献报道）
- 基于 EST 数据（EST 与基因组或 DNA、mRNA 比对）

数据库与工具

- ASTD = ASD (= AEDB + AltExtron + AltSplice) + ATD
- ASAP
- ESEfinder
- RESCUE-ESE
- ASPicDB



教学提纲

- 1 引言
- 2 DNA 组份分析与序列转换
- 3 限制酶位点分析
- 4 开放阅读框分析
- 5 启动子分析
- 6 CpG 岛识别
- 7 总结与答疑
- 8 引言

- 9 重复序列分析
- 10 基因识别
- 11 总结与答疑
- 12 引言
- 13 mRNA 选择性剪接
- 14 miRNA 及其靶基因预测
- 15 总结与答疑
- 16 复习思考题



微 RNA (microRNAs, miRNA, 小分子 RNA)

归属小 RNA 范畴，是真核生物中广泛存在的一种长约 20 到 24 个核苷酸的内源性非编码单链 RNA 分子。miRNA 通过 RNA 诱导沉默复合体 (RISC) 与靶基因的 3' 非翻译区 (3' UTR) 相结合，导致靶基因 mRNA 降解或者抑制其翻译，从而调节基因转录后的表达。



miRNA | 特征

序列

不具有开放阅读框，不编码蛋白质；成熟的 miRNA 5' 端为单一磷酸基团，3' 端为羟基

表达

具有时序性和组织特异性

调控

miRNA 与靶基因间呈多对多的关系

物理位置

倾向于成簇地出现在染色体上

进化的保守性

在物种间高度保守

miRNA | 特征

序列

不具有开放阅读框，不编码蛋白质；成熟的 miRNA 5' 端为单一磷酸基团，3' 端为羟基

表达

具有时序性和组织特异性

调控

miRNA 与靶基因间呈多对多的关系

物理位置

倾向于成簇地出现在染色体上

进化

在物种间高度保守

miRNA | 特征

序列

不具有开放阅读框，不编码蛋白质；成熟的 miRNA 5' 端为单一磷酸基团，3' 端为羟基

表达

具有时序性和组织特异性

调控

miRNA 与靶基因间呈多对多的关系

物理位置

倾向于成簇地出现在染色体上

进化

在物种间高度保守

miRNA | 特征

序列

不具有开放阅读框，不编码蛋白质；成熟的 miRNA 5' 端为单一磷酸基团，3' 端为羟基

表达

具有时序性和组织特异性

调控

miRNA 与靶基因间呈多对多的关系

物理位置

倾向于成簇地出现在染色体上

进化

在物种间高度保守

miRNA | 特征

序列

不具有开放阅读框，不编码蛋白质；成熟的 miRNA 5' 端为单一磷酸基团，3' 端为羟基

表达

具有时序性和组织特异性

调控

miRNA 与靶基因间呈多对多的关系

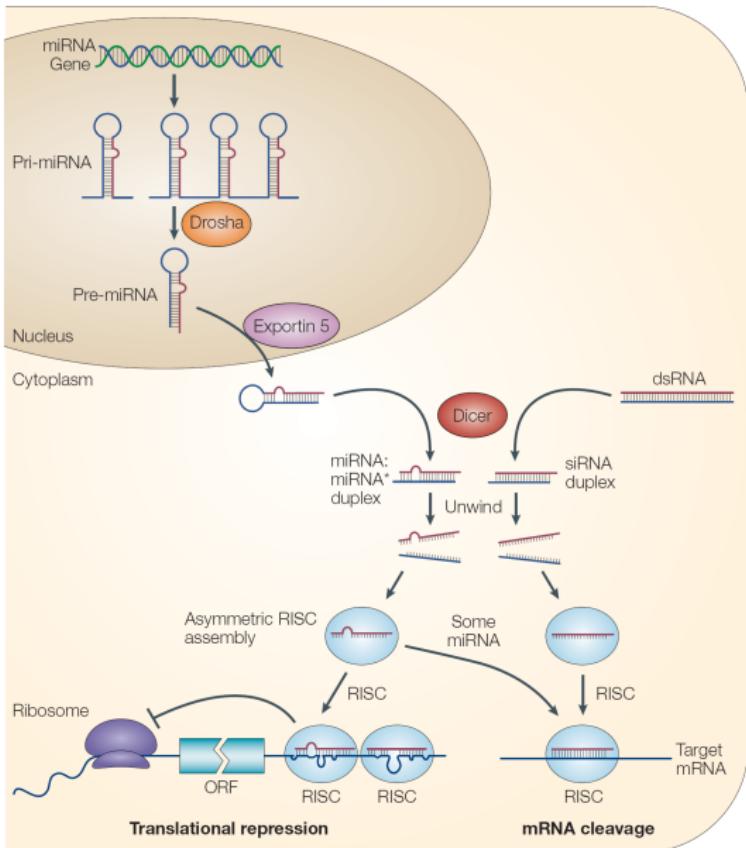
物理位置

倾向于成簇地出现在染色体上

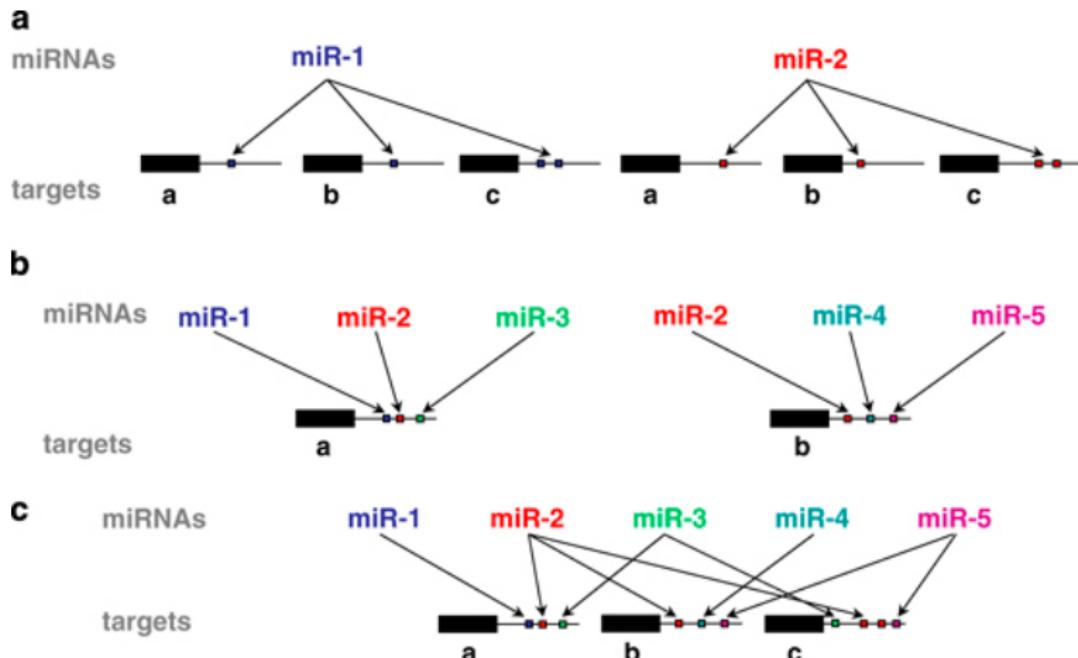
进化

在物种间高度保守

miRNA | 生成



miRNA | 作用网络



miRNA | 功能

pn-microRNA and removes the hairpin loop, leaving a double stranded microRNA duplex molecule.

3 In plant cells, the microRNA is usually perfectly complementary to its target mRNA molecule. The microRNA will bond with it, and cause the mRNA to break down.

acute proteins

4 In animal cells, the microRNA nucleotides typically don't pair up with the mRNA nucleotides as well. Their base pairing often follows a pattern though.

microRNA

Nucleotide 1
Has an A across from it

3' Poly-A tail

Deadenylation

5 The microRNA-protein complex's presence blocks translation as well as speeding up deadenylation (breakdown of the Poly-A tail), which causes the mRNA to be degraded sooner and translated less.

the formation and function of microRNAs



- 实验手段（cDNA 克隆测序、miRNA-seq），有局限性。
- 计算预测，有优势。



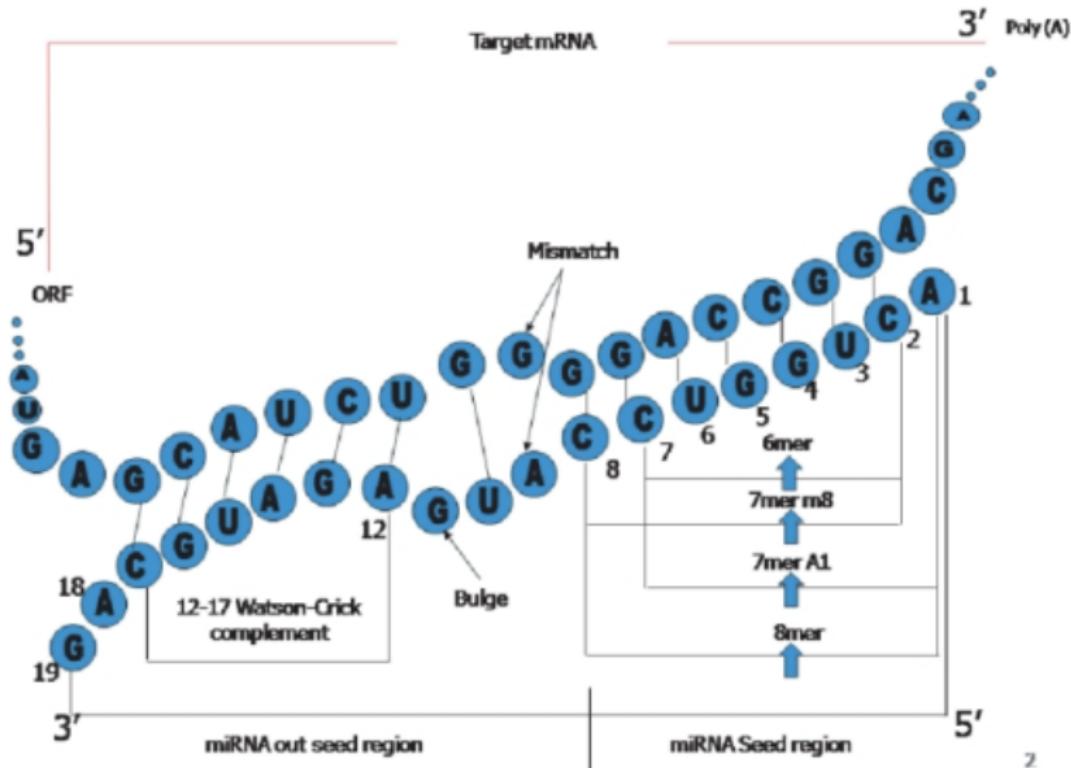
- ① 同源片段搜索方法。将已知 miRNA 或 pre-miRNA 序列在自身或其他相近基因组中用比对算法搜索同源序列，结合序列二级结构特征进行筛选。（局限于：与已知 miRNA/pre-miRNA 在序列上和结构上同源的 miRNA/pre-miRNA）
- ② 基于比较基因组学的预测方法。依据进化过程中的保守性在多物种中搜索潜在的 miRNA。（能够找到不与已知 miRNA 同源的新 miRNA；局限于：保守 miRNA）
 - 方法一：先在一个物种基因组中根据结构和序列特征找出可能的 pre-miRNA，而后与其他物种基因组比较，判断其序列和结构是否保守
 - 方法二：先通过比较两物种的基因组找出保守区域，而后在保守区域中根据结构和序列特征搜索可能的 miRNA



- ③ 基于序列和结构特征打分的预测方法。根据已知 miRNA 序列和结构的特征对全基因组范围内能形成茎环结构的片段进行筛选，是发现非同源、物种特异 miRNA 的方法。（为降低假阳性，用异常严格的标准筛选候选片段，可能遗漏大量的 miRNA）
- ④ 结合作用靶标的预测方法。依据 miRNA 与其靶基因序列间的碱基互补配对的保守性的特点预测 miRNA。
- ⑤ 基于机器学习的预测方法。通过对阳性 miRNA 和阴性 miRNA 数据集的训练来构建区分两者的分类器，根据所得分类器对未知序列进行预测。（支持向量机 SVM 是目前 miRNA 分类和预测最常用的机器学习方法）



miRNA | 种子区域



① 基于种子区域互补和保守性的规则预测

- miRanda
- TargetScan

② 基于机器学习方法训练参数进行靶基因预测

- PicTar
- miTarget



- 数据库：miRBase、miRTarBase、miRWalk2.0、TarBase、miRGen
- miRNA 预测：MiRscan、MiPred、miRFinder
- miRNA 靶基因预测：miRanda、TargetScan、PicTar、miTarget
- 微 RNA 与微 RNA 靶数据库（维基百科）

强调

- miRBase 是一个集 miRNA 序列、注释信息以及预测的靶基因数据为一体的数据库。
- miRTarBase: the experimentally validated microRNA-target interactions database
- miRWalk2.0: a comprehensive atlas of predicted and validated microRNA-target interactions
- TarBase 是一个存储已被实验证实的真实 miRNA 与靶基因间关系的数据库。

- 数据库：miRBase、miRTarBase、miRWalk2.0、TarBase、miRGen
- miRNA 预测：MiRscan、MiPred、miRFinder
- miRNA 靶基因预测：miRanda、TargetScan、PicTar、miTarget
- 微 RNA 与微 RNA 靶数据库（维基百科）

强调

- miRBase 是一个集 miRNA 序列、注释信息以及预测的靶基因数据为一休的数据库。
- miRTarBase: the experimentally validated microRNA-target interactions database
- miRWalk2.0: a comprehensive atlas of predicted and validated microRNA-target interactions
- TarBase 是一个存储已被实验证实的真实 miRNA 与靶基因间关系的数据休。

教学提纲

- 1 引言
- 2 DNA 组份分析与序列转换
- 3 限制酶位点分析
- 4 开放阅读框分析
- 5 启动子分析
- 6 CpG 岛识别
- 7 总结与答疑
- 8 引言

- 9 重复序列分析
- 10 基因识别
- 11 总结与答疑
- 12 引言
- 13 mRNA 选择性剪接
- 14 miRNA 及其靶基因预测
- 15 总结与答疑
- 16 复习思考题



知识点——mRNA 选择性剪接和 miRNA 分析

- mRNA 选择性剪接——选择性剪接的主要机制，数据资源
- miRNA——miRNA 的特点和作用机制，miRNA 预测方法与工具，miRNA 靶基因预测方法与工具



教学提纲

- 1 引言
- 2 DNA 组份分析与序列转换
- 3 限制酶位点分析
- 4 开放阅读框分析
- 5 启动子分析
- 6 CpG 岛识别
- 7 总结与答疑
- 8 引言

- 9 重复序列分析
- 10 基因识别
- 11 总结与答疑
- 12 引言
- 13 mRNA 选择性剪接
- 14 miRNA 及其靶基因预测
- 15 总结与答疑
- 16 复习思考题



知识点

- ① DNA 序列携带哪两类遗传信息？可以对 DNA 序列进行哪些分析？
- ② 简述限制性核酸内切酶的命名规则及 II 型限制酶的主要特点。
- ③ 简述 CpG 岛的概念及其识别依据和判别标准。
- ④ 简述重复序列依重复次数和组织形式的分类。
- ⑤ 简述基因识别的三大类方法及主要策略。
- ⑥ 简述选择性剪接的产生机制。
- ⑦ 简述 miRNA 预测和 miRNA 靶基因预测的方法。

