

生物信息学

天津医科大学
生物医学工程学院

2013-2014 学年上学期

- 只有正式上课前的请假有效。
- 提前 5 分钟到教室, 严禁迟到。
- 上课期间手机关机或调成震动。
- 上课期间离开教室先举手示意。
- 课上有疑问的话先举手后提问。
- 上课期间严禁交头接耳, 大声喧哗。
- 随机点名, 缺勤扣分如下: 1、3、6。
- 缺勤三次或三次以上者, 平时成绩为 0。



- 只有正式上课前的请假有效。
- 提前 5 分钟到教室, 严禁迟到。
- 上课期间手机关机或调成震动。
- 上课期间离开教室先举手示意。
- 课上有疑问的话先举手后提问。
- 上课期间严禁交头接耳, 大声喧哗。
- 随机点名, 缺勤扣分如下: 1、3、6。
- 缺勤三次或三次以上者, 平时成绩为 0。



- 只有正式上课前的请假有效。
- 提前 5 分钟到教室，严禁迟到。
- **上课期间手机关机或调成震动。**
- 上课期间离开教室先举手示意。
- 课上有疑问的话先举手后提问。
- 上课期间严禁交头接耳，大声喧哗。
- 随机点名，缺勤扣分如下：1、3、6。
- 缺勤三次或三次以上者，平时成绩为 0。



- 只有正式上课前的请假有效。
- 提前 5 分钟到教室, 严禁迟到。
- 上课期间手机关机或调成震动。
- **上课期间离开教室先举手示意。**
- 课上有疑问的话先举手后提问。
- 上课期间严禁交头接耳, 大声喧哗。
- 随机点名, 缺勤扣分如下 : 1、3、6。
- 缺勤三次或三次以上者, 平时成绩为 0。



- 只有正式上课前的请假有效。
- 提前 5 分钟到教室，严禁迟到。
- 上课期间手机关机或调成震动。
- 上课期间离开教室先举手示意。
- **课上有疑问的话先举手后提问。**
- 上课期间严禁交头接耳，大声喧哗。
- 随机点名，缺勤扣分如下：1、3、6。
- 缺勤三次或三次以上者，平时成绩为 0。



- 只有正式上课前的请假有效。
- 提前 5 分钟到教室，严禁迟到。
- 上课期间手机关机或调成震动。
- 上课期间离开教室先举手示意。
- 课上有疑问的话先举手后提问。
- **上课期间严禁交头接耳，大声喧哗。**
- 随机点名，缺勤扣分如下：1、3、6。
- 缺勤三次或三次以上者，平时成绩为 0。



- 只有正式上课前的请假有效。
- 提前 5 分钟到教室，严禁迟到。
- 上课期间手机关机或调成震动。
- 上课期间离开教室先举手示意。
- 课上有疑问的话先举手后提问。
- 上课期间严禁交头接耳，大声喧哗。
- 随机点名，缺勤扣分如下：1、3、6。
- 缺勤三次或三次以上者，平时成绩为 0。



- 只有正式上课前的请假有效。
- 提前 5 分钟到教室，严禁迟到。
- 上课期间手机关机或调成震动。
- 上课期间离开教室先举手示意。
- 课上有疑问的话先举手后提问。
- 上课期间严禁交头接耳，大声喧哗。
- 随机点名，缺勤扣分如下：1、3、6。
- 缺勤三次或三次以上者，平时成绩为 0。



自我介绍

姓 名 伊现富 (Yi Xianfu)

本 科 山东大学

硕 博 中国科学院

工作邮箱 yixfbio@gmail.com

生活邮箱 yixf1986@gmail.com

手 机 15620610763

个人博客 <http://yixf.name>

网络昵称 yixf



① 126 邮箱

- 账号 : bioinfo_TIJMU@126.com
- 密码 : C&563f&nzx!s

② 百度云网盘

- 账号 : bioinfo_TIJMU@126.com
- 密码 : 566&Us3Rp6#C



核酸序列分析

伊现富 (Yi Xianfu)

天津医科大学 (TJMU)
生物医学工程学院

2013 年 8 月



教学提纲

- 1 引言
- 2 DNA 序列转换与组份分析
- 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 重复序列分析

- 8 总结与答疑
- 9 引言
- 10 基因识别
- 11 mRNA 选择性剪接
- 12 miRNA 及其靶基因预测
- 13 lncRNA
- 14 查找数据库与工具
- 15 总结与答疑

1 引言

2 DNA 序列转换与组份分析

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 重复序列分析

8 总结与答疑

9 引言

10 基因识别

11 mRNA 选择性剪接

12 miRNA 及其靶基因预测

13 lncRNA

14 查找数据库与工具

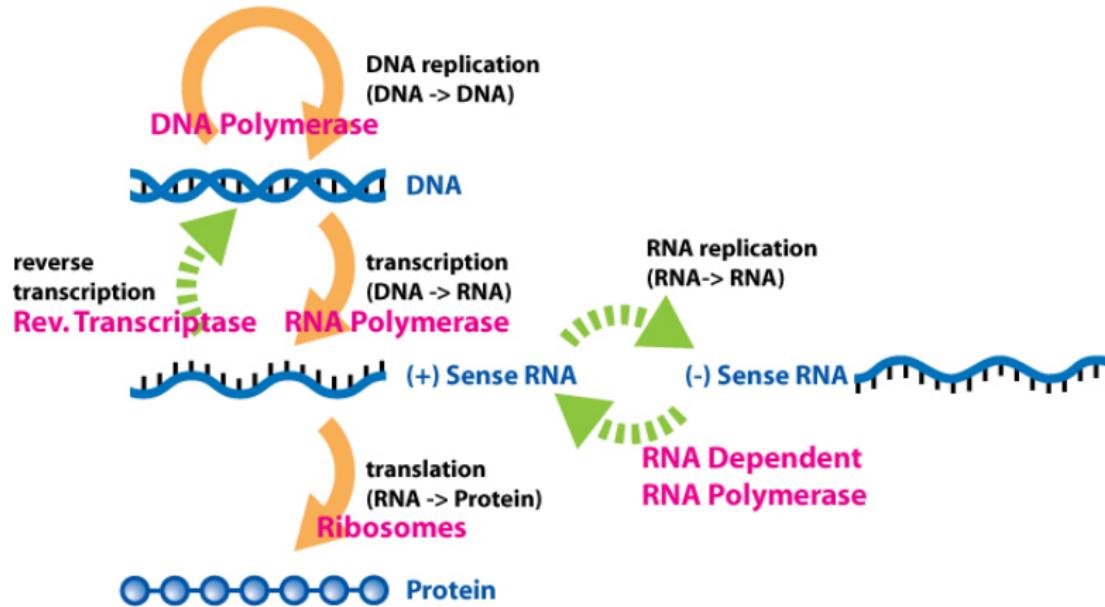
15 总结与答疑



引言 | 大千世界



引言 | 中心法则



A COMPLETE GENOME IN TIME

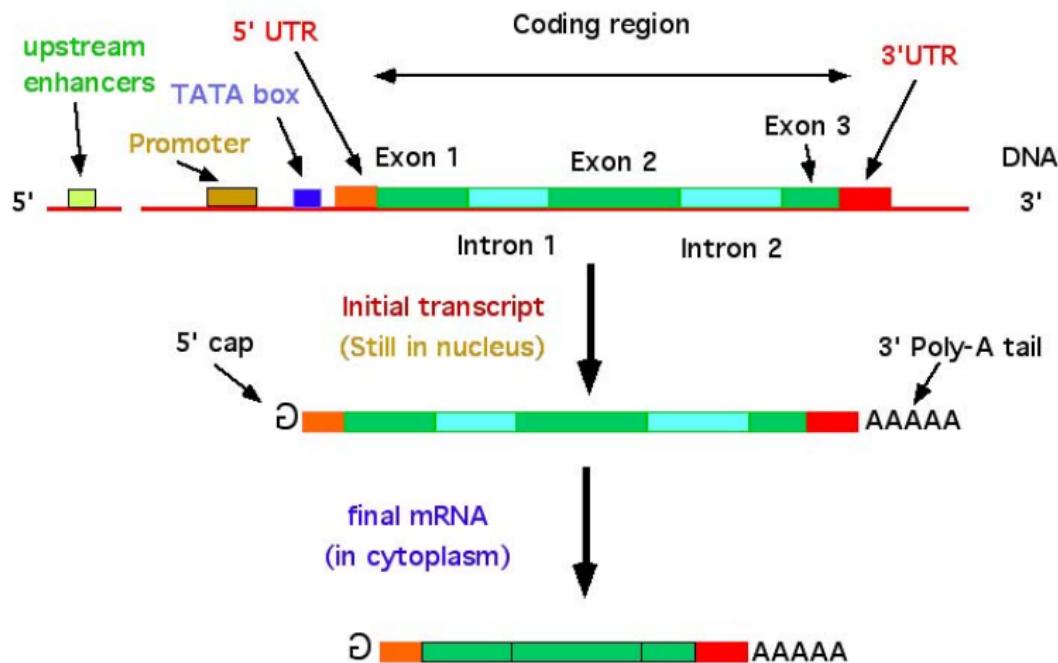
by Yonder Biology



引言 | ACGT⇒ 生信



引言 | 遗传信息



教学提纲

1 引言

2 DNA 序列转换与组份分析

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 重复序列分析

8 总结与答疑

9 引言

10 基因识别

11 mRNA 选择性剪接

12 miRNA 及其靶基因预测

13 lncRNA

14 查找数据库与工具

15 总结与答疑



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

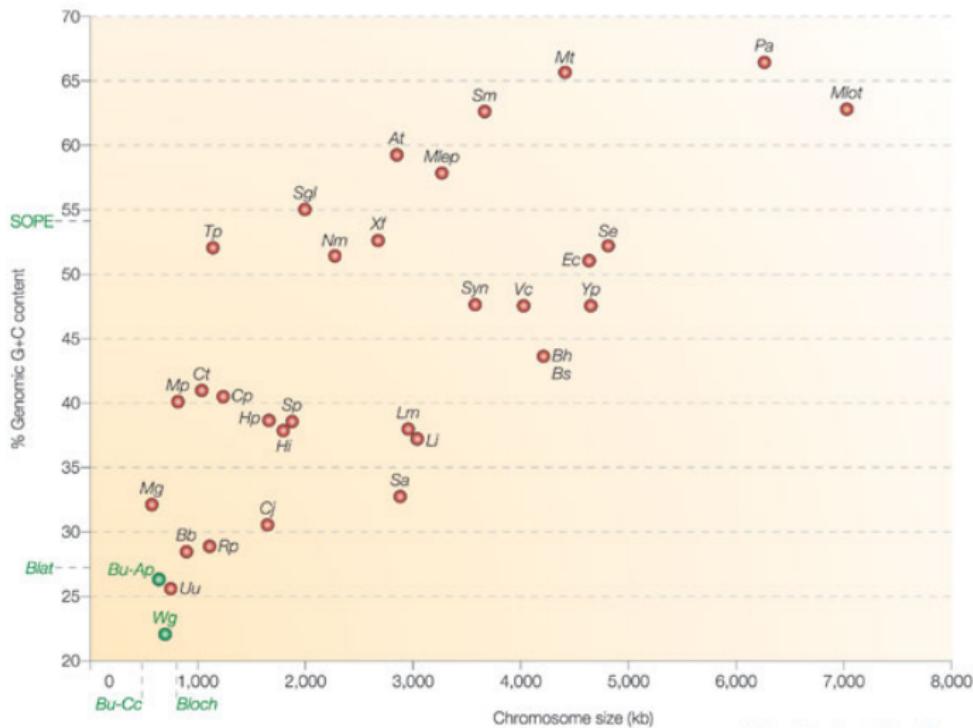


GC 含量 (GC content)

- 鸟嘌呤 (G) 和胞嘧啶 (C) 所占的比例
- GC 含量高的 DNA 更加稳定
- GC 含量随 DNA 不同而异
- 计算公式： $\frac{G+C}{A+T+G+C} \times 100$
- GC 比 (GC-ratio) : $\frac{G+C}{A+T}$



DNA 序列 | GC 含量 | 基因组



Nature Reviews | Genetics



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>Y. 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 ICEhin1056	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 pIP1202	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]



任务分析

- 序列长短
- 序列数目
- 任务数量
- 任务频率
- 工作时间
- ...



- SeqTools.pl
- EMBOSS
- bioinfx(Free Online Tools for Bioinformatics)
- Complementary Sequence Conversion Tool
- DNA Sequence Reverse and Complement Online Tool
- DNA/RNA GC Content Calculator
- Oligo Calculator
- ...

教学提纲

1 引言

2 DNA 序列转换与组份分析

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 重复序列分析

8 总结与答疑

9 引言

10 基因识别

11 mRNA 选择性剪接

12 miRNA 及其靶基因预测

13 lncRNA

14 查找数据库与工具

15 总结与答疑



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
- 平滑末端



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
- 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



- 识别、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

《题金山寺》，北宋·苏轼

潮随暗浪雪山倾，远浦渔舟钓月明。轻鸥数点千峰雪，水接云边四望遥。
桥对寺门松径小，槛当泉眼石波清。晴日晚霞红霭霭，晓天江树绿迢迢。
迢迢绿树江天晓，霭霭红霞晚日晴。清波石眼泉当槛，小径松门寺对桥。
遥望四边云接水，雪峰千点数鸥轻。明月钓舟渔浦远，倾山雪浪暗随潮。

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。



限制酶 | II 型 | 回文

《题金山寺》，北宋·苏轼

潮随暗浪雪山倾，远浦渔舟钓月明。轻鸥数点千峰雪，水接云边四望遥。
桥对寺门松径小，槛当泉眼石波清。晴日晚霞红霭霭，晓天江树绿迢迢。
迢迢绿树江天晓，霭霭红霞晚日晴。清波石眼泉当槛，小径松门寺对桥。
遥望四边云接水，雪峰千点数鸥轻。明月钓舟渔浦远，倾山雪浪暗随潮。

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。



限制酶 | II 型 | 回文

《题金山寺》，北宋·苏轼

潮随暗浪雪山倾，远浦渔舟钓月明。轻鸥数点千峰雪，水接云边四望遥。
桥对寺门松径小，槛当泉眼石波清。晴日晚霞红霭霭，晓天江树绿迢迢。
迢迢绿树江天晓，霭霭红霞晚日晴。清波石眼泉当槛，小径松门寺对桥。
遥望四边云接水，雪峰千点数鸥轻。明月钓舟渔浦远，倾山雪浪暗随潮。

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。



限制酶 | II 型 | 黏性末端

酵素名称	来源	辨识序列	切法
EcoRI	<i>Escherichia coli</i>	5'GAATTC 3'CTTAAG	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 | 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'



限制酶 | 数据库与分析工具

- REBASE
- NEBCutter V2.0



教学提纲

- 1 引言
- 2 DNA 序列转换与组份分析
- 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 重复序列分析

- 8 总结与答疑
- 9 引言
- 10 基因识别
- 11 mRNA 选择性剪接
- 12 miRNA 及其靶基因预测
- 13 lncRNA
- 14 查找数据库与工具
- 15 总结与答疑



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 (ATG) 到终止密码子 (TAA、TAG、TGA) 之间的一段编码蛋白质的碱基序列。



开放阅读框 | 相位

Frame +3 +2 +1 -1 -2 -3

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



开放阅读框 | ORF VS. CDS

- 一个 ORF 对应一个候选的 CDS (编码序列, Coding Sequence)
- ORF : 理论预测
- CDS : 实验证实



- 一个 ORF 对应一个候选的 CDS (编码序列, Coding Sequence)
- ORF : 理论预测
- CDS : 实验证实



- 确定第一个 ATG 和终止密码子
- 最长 ORF 法（原核生物）
- ORF Finder



教学提纲

1 引言

2 DNA 序列转换与组份分析

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 重复序列分析

8 总结与答疑

9 引言

10 基因识别

11 mRNA 选择性剪接

12 miRNA 及其靶基因预测

13 lncRNA

14 查找数据库与工具

15 总结与答疑



- 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
- 两者相互作用参与基因表达调控



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 | 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。

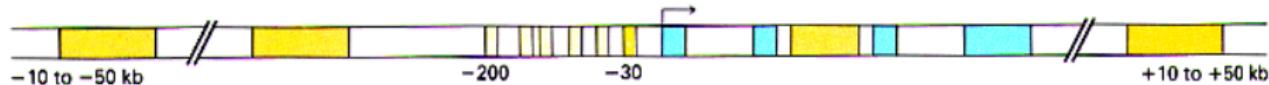


启动子 (promoter)

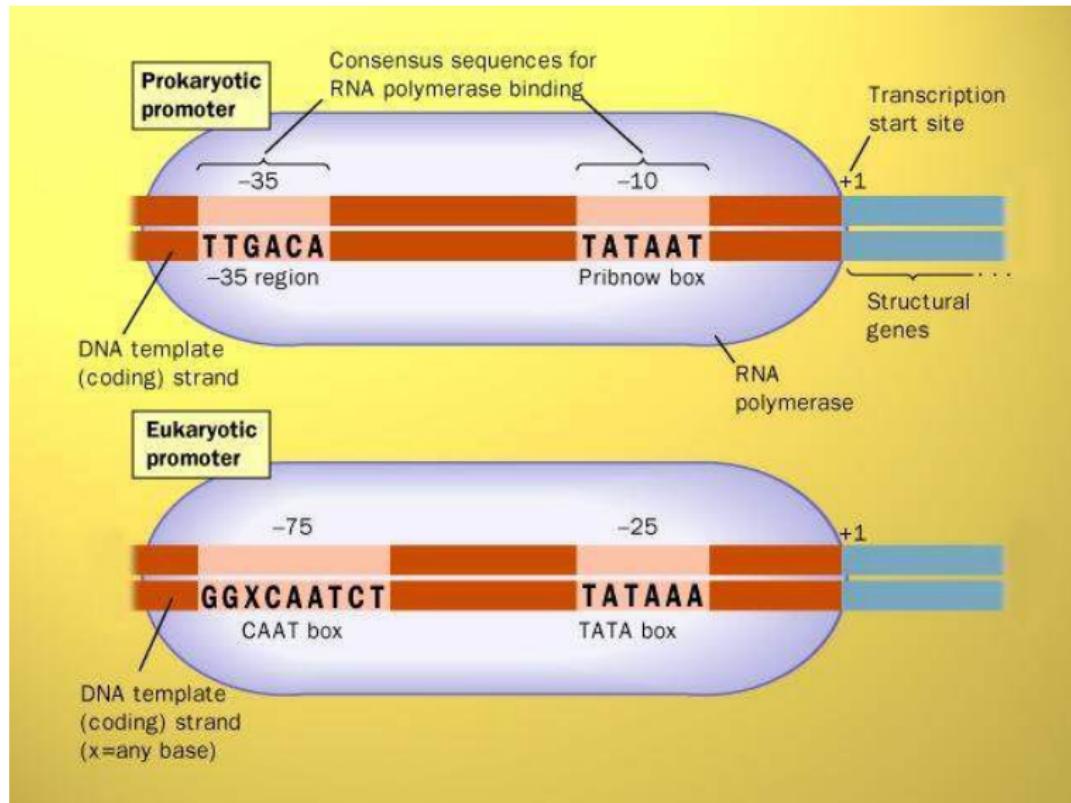
一段位于转录起始位点 5' 端上游区的 DNA 序列，能活化 RNA 聚合酶，使之与模板 DNA 准确地结合并具有转录起始的特异性。

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 | 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。

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。



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。

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。



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
 - Promoter Scan, Promoter 2.0
- 转录因子
 - TRANSFAC
 - Tfblast (TRANSFAC BLAST)



教学提纲

- 1 引言
- 2 DNA 序列转换与组份分析
- 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 重复序列分析

- 8 总结与答疑
- 9 引言
- 10 基因识别
- 11 mRNA 选择性剪接
- 12 miRNA 及其靶基因预测
- 13 lncRNA
- 14 查找数据库与工具
- 15 总结与答疑



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
- 一般位于基因的 5' 端区域（转录起始位点附近），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
- 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
- 一般位于基因的 5' 端区域（转录起始位点附近），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
- 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

- 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%

$$\bullet \frac{\text{Num of CpG}}{\text{Num of C} \times \text{Num of G}} \times \text{Total number of nucleotides in the sequence}$$

- ④ 500bp, 55%, 65%



- 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%

$$\bullet \frac{\text{Num of CpG}}{\text{Num of C} \times \text{Num of G}} \times \text{Total number of nucleotides in the sequence}$$

- ④ 500bp, 55%, 65%



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



教学提纲

- 1 引言
- 2 DNA 序列转换与组份分析
- 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 重复序列分析

- 8 总结与答疑
- 9 引言
- 10 基因识别
- 11 mRNA 选择性剪接
- 12 miRNA 及其靶基因预测
- 13 lncRNA
- 14 查找数据库与工具
- 15 总结与答疑



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) : 在整个基因组中只含有 2 ~ 10 个拷贝
- 中度重复序列 (moderately repetitive sequence) : 重复次数为几十次到几千次，重复单元的平均长度约 300bp
- 高度重复序列 (highly repetitive sequence) : 重复几百万次，一般是少于 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) : 在整个基因组中只含有 2 ~ 10 个拷贝
- 中度重复序列 (moderately repetitive sequence) : 重复次数为几十次到几千次，重复单元的平均长度约 300bp
- 高度重复序列 (highly repetitive sequence) : 重复几百万次，一般是少于 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
 - 卫星 DNA (satellite DNA) : 一类高度重复序列
 - 小卫星 (minisatellite, VNTR) : 由 10 ~ 100bp 的基本单位串联而成，总长通常不超过 20kb，重复次数在群体中是高度变异的
 - 微卫星 (microsatellite, SSR, STR) : 两个或多个核苷酸重复排列，只有 2 ~ 10bp，串联成簇，长度 50 ~ 100bp，STR 遗传多态性
- 散在重复序列：分散于染色体的各位点上
 - 短散在重复序列 (Short Interspersed Nuclear Element, SINE) : 长度在 500bp 以下，在人基因组中的重复拷贝数达 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
 - 长散在重复序列 (Long Interspersed Nuclear Element, LINE) : 长度在 1000bp 以上，在人基因组中有上万份拷贝；可以自主转座的一类反转录转座子，来源于 RNA 聚合酶 II 的转录产物；L1

- Repbase
- L1Base
- STRBase
- RepeatMasker：四个搜索引擎
 - Cross_match
 - ABBLast
 - RMBLast
 - HMMER



教学提纲

1 引言

2 DNA 序列转换与组份分析

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 重复序列分析

总结与答疑

8 引言

9 基因识别

10 mRNA 选择性剪接

11 miRNA 及其靶基因预测

12 lncRNA

13 查找数据库与工具

14 总结与答疑



知识点

- DNA 序列基本信息分析——查戈夫法则，序列转换，GC 含量
- 限制酶位点分析——命名，II 型
- 开放阅读框分析——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——判别依据及标准
- 重复序列分析——分类

技能

- 解决问题的思路
- 搜索、学习软件

教学提纲

- 1 引言
- 2 DNA 序列转换与组份分析
- 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 重复序列分析

- 8 总结与答疑
- 9 引言
- 10 基因识别
- 11 mRNA 选择性剪接
- 12 miRNA 及其靶基因预测
- 13 lncRNA
- 14 查找数据库与工具
- 15 总结与答疑



● 基本信息分析

- 序列转换
- 碱基比例
- GC 含量
- 寻找限制酶切位点

● 序列特征分析

● 基因识别



● 基本信息分析

● 序列转换

- 碱基比例
- GC 含量
- 寻找限制酶切位点

● 序列特征分析

● 基因识别



- 基本信息分析

- 序列转换
- 碱基比例
- GC 含量
- 寻找限制酶切位点

- 序列特征分析

- 基因识别



- 基本信息分析

- 序列转换
- 碱基比例
- GC 含量
- 寻找限制酶切位点

- 序列特征分析

- 基因识别



● 基本信息分析

- 序列转换
- 碱基比例
- GC 含量
- **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

- 序列转换
- 碱基比例
- GC 含量
- 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

- 序列转换
- 碱基比例
- GC 含量
- 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

- 序列转换
- 碱基比例
- GC 含量
- 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

- 序列转换
- 碱基比例
- GC 含量
- 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

→ 非重叠序列

→ 基因识别



● 基本信息分析

- 序列转换
- 碱基比例
- GC 含量
- 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

- 序列转换
- 碱基比例
- GC 含量
- 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

- 序列转换
- 碱基比例
- GC 含量
- 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 序列转换与组份分析

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 重复序列分析

8 总结与答疑

9 引言

10 基因识别

11 mRNA 选择性剪接

12 miRNA 及其靶基因预测

13 lncRNA

14 查找数据库与工具

15 总结与答疑



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。



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。

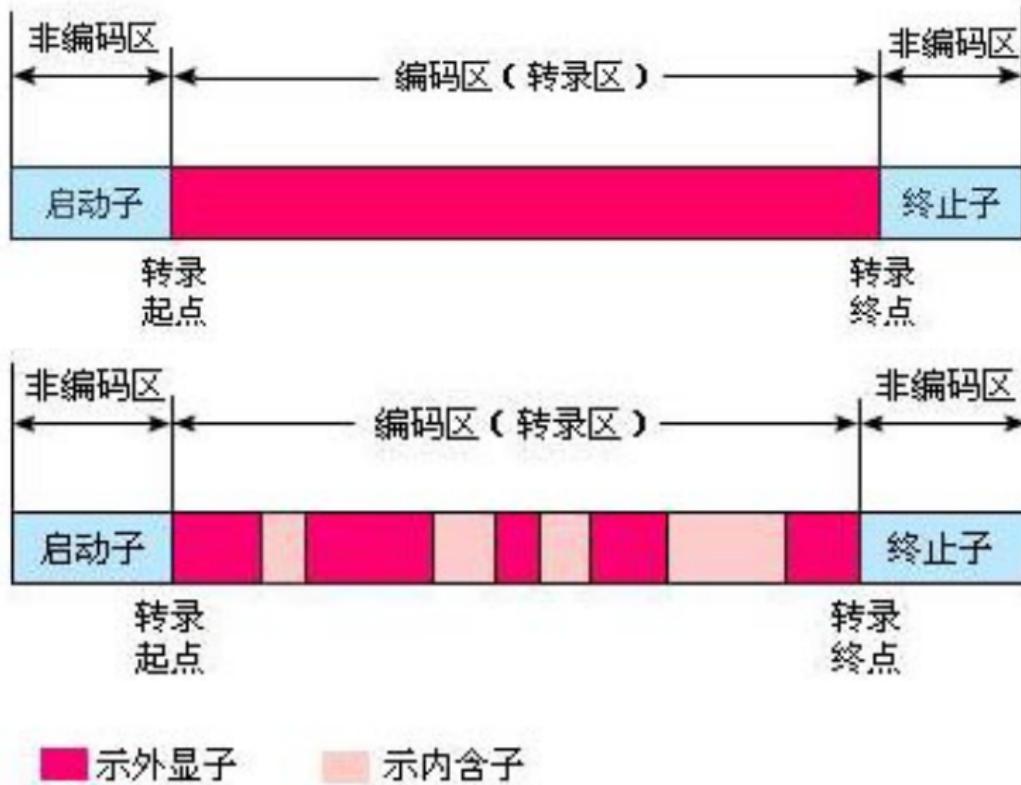
一个完整的基因，不仅包括编码区，还包括 5' 末端和 3' 末端长度不等的特异性序列。

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。



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, 基于基因的两种类型的特征：
 - “信号” : 由一些特殊的序列构成, 通常预示着其周围存在着一个基因 ; TATA box、CAAT box、供体位点与受体位点、起始密码子、终止密码子、polyA 信号序列、...
 - “内容” : 蛋白质编码基因所具有的某些统计学特征 ; 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)、...
- 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



- GeneMarkS
- Glimmer
- GENSCAN
- GRAIL
- List of gene prediction software(Wikipedia)

教学提纲

1 引言

2 DNA 序列转换与组份分析

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 重复序列分析

8 总结与答疑

9 引言

10 基因识别

11 mRNA 选择性剪接

12 miRNA 及其靶基因预测

13 lncRNA

14 查找数据库与工具

15 总结与答疑



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。



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 | 实例

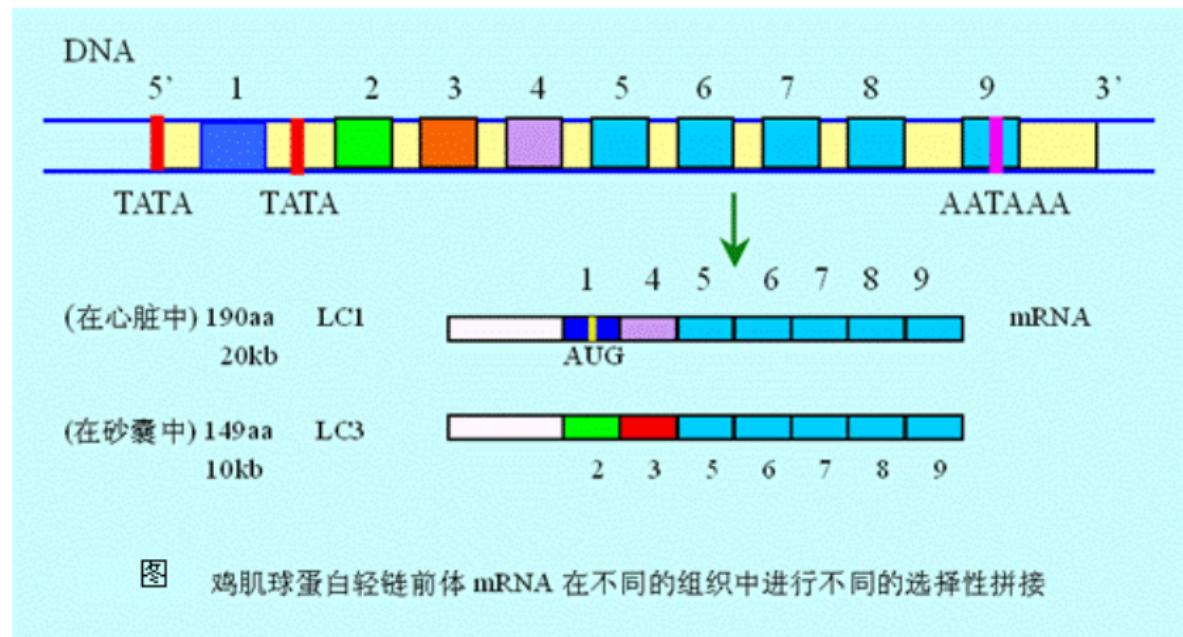
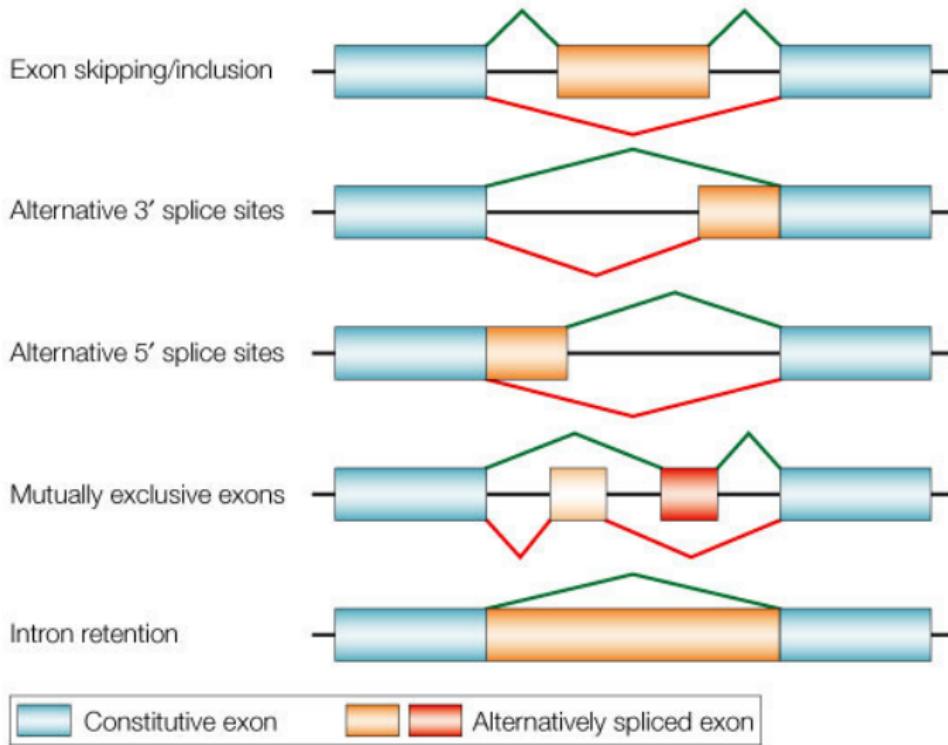


图 鸡肌球蛋白轻链前体 mRNA 在不同的组织中进行不同的选择性拼接

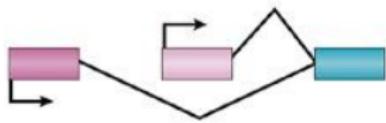


选择性剪接 | 机制 | 五种

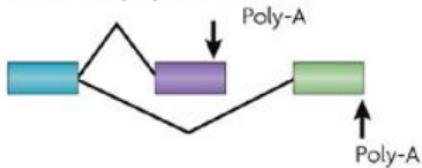


选择性剪接 | 机制 | 七种

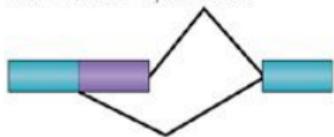
Alternative promoters



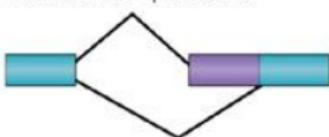
Alternative poly-A sites



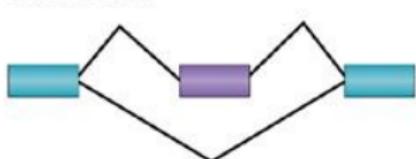
Alternative 5' splice sites



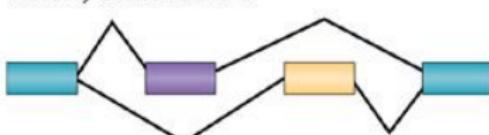
Alternative 3' splice sites



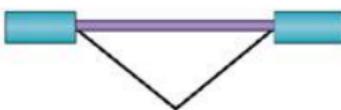
Cassette exon



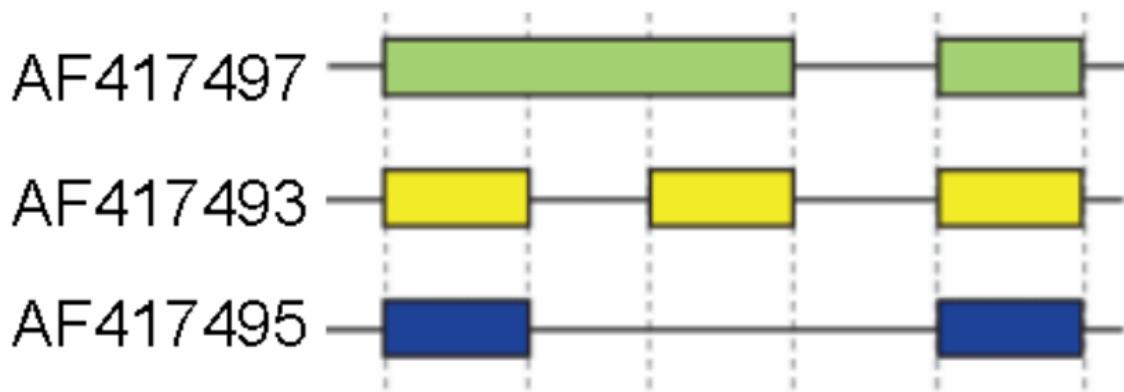
Mutually exclusive exons



Retained intron



选择性剪接 | 机制 | 复杂实例



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



教学提纲

1 引言

2 DNA 序列转换与组份分析

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 重复序列分析

8 总结与答疑

9 引言

10 基因识别

11 mRNA 选择性剪接

12 miRNA 及其靶基因预测

13 lncRNA

14 查找数据库与工具

15 总结与答疑



非编码 RNA (non-coding RNAs, ncRNA)

- 基础结构性 ncRNA (infrastructural non-coding RNAs) , 看家 ncRNA (housekeeping non-coding RNAs)
 - tRNA、rRNA、snRNA、snoRNA
- 调节性 ncRNA (regulatory non-coding RNAs)
 - 小 RNA (small RNAs, sRNA) : <200nt
 - miRNA、siRNA、piRNA
 - 长链非编码 RNA (long ncRNAs, lncRNA) : >200nt



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。

特点

- 20 ~ 24nt
- 不具有开放阅读框，不编码蛋白质
- 表达具有时序性和组织特异性
- 进化上具有高度的保守性



微 RNA (microRNAs, miRNA, 小分子 RNA)

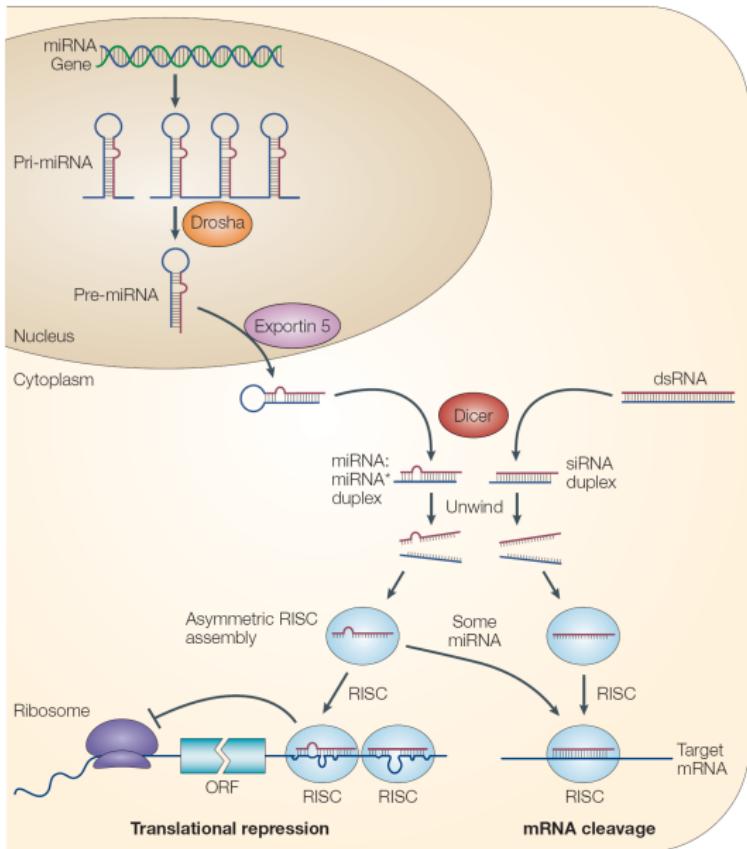
归属小 RNA 范畴，是真核生物中广泛存在的一种长约 20 到 24 个核苷酸的内源性非编码单链 RNA 分子。miRNA 通过 RNA 诱导沉默复合体 (RISC) 与靶基因的 3' 非翻译区 (3' UTR) 相结合，导致靶基因 mRNA 降解或者抑制其翻译，从而调节基因转录后的表达。

特点

- 20 ~ 24nt
- 不具有开放阅读框，不编码蛋白质
- 表达具有时序性和组织特异性
- 进化上具有高度的保守性



miRNA | 生成



miRNA | 作用网络

a

miRNAs

miR-1

targets

a

b

c

miR-2

a

b

c

b

miRNAs

miR-1

miR-2

miR-3

miR-2

miR-4

miR-5

targets

a

b

c

miRNAs

miR-1

miR-2

miR-3

miR-4

miR-5

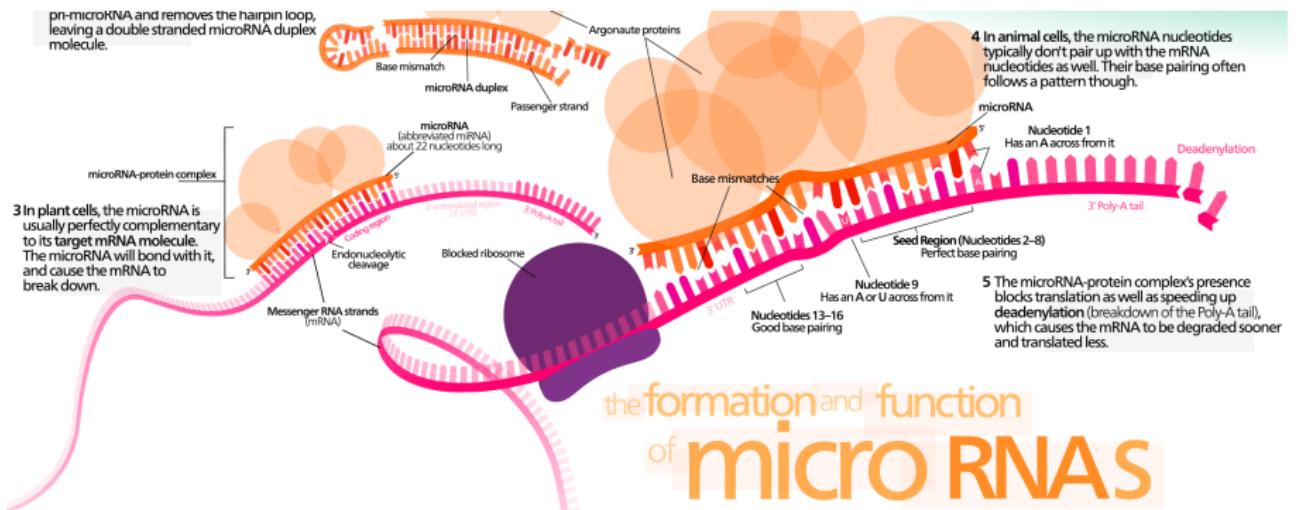
targets

a

b

c

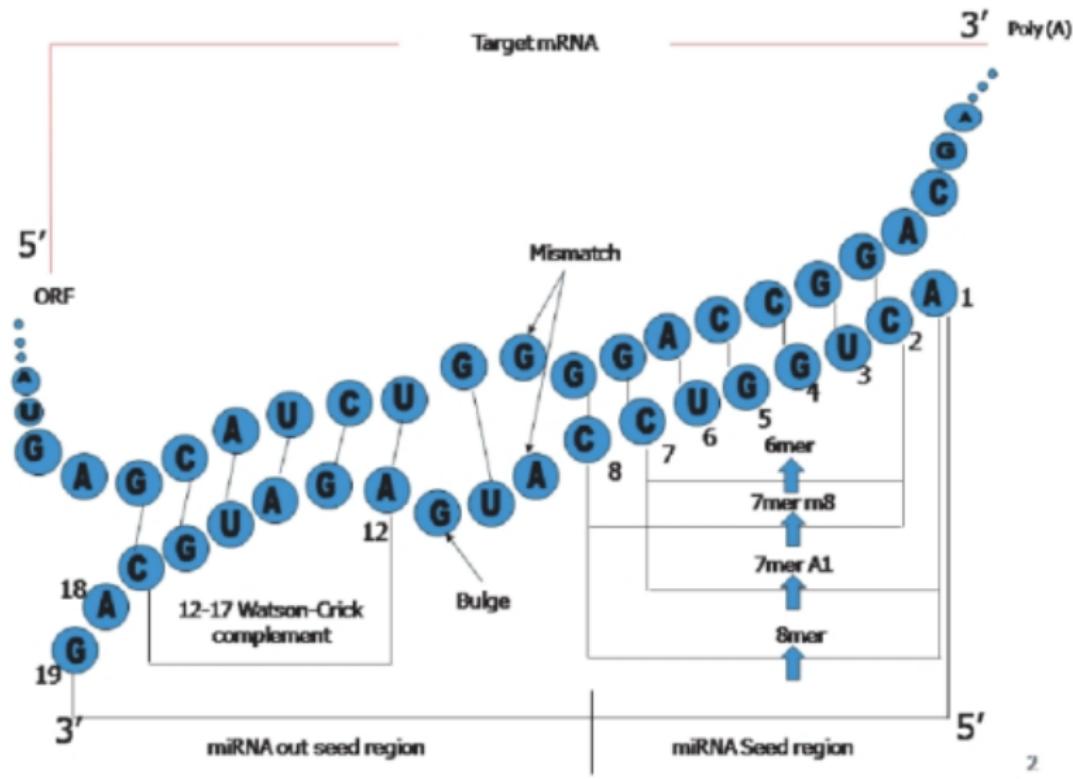
miRNA | 功能



- ① 同源片段搜索方法
- ② 基于比较基因组学的预测方法
- ③ 基于序列和结构特征打分的预测方法
- ④ 结合作用靶标的预测方法
- ⑤ 基于机器学习的预测方法



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、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）



教学提纲

1 引言

2 DNA 序列转换与组份分析

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 重复序列分析

8 总结与答疑

9 引言

10 基因识别

11 mRNA 选择性剪接

12 miRNA 及其靶基因预测

13 lncRNA

14 查找数据库与工具

15 总结与答疑



- 大多被 RNA 聚合酶 II 所转录
- 有 5' 帽子和 3' 端的 poly(A) 尾巴
- 主要富集在细胞核
- 长度偏短、外显子数目偏少
- 在不同物种间的保守性差
- 稳定性偏低
- 表达水平很低，且具有时空特异性
- 长链非编码 RNA 数据库（维基百科）



教学提纲

1 引言

2 DNA 序列转换与组份分析

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 重复序列分析

8 总结与答疑

9 引言

10 基因识别

11 mRNA 选择性剪接

12 miRNA 及其靶基因预测

13 lncRNA

14 查找数据库与工具

15 总结与答疑



- 借鉴相关文献中使用的数据库与工具
- 向特定领域的专家请教
- *Nucleic Acids Research* 每年的第一期为数据库专刊
- 维基百科等总结性网站
- [The Elements of Bioinformatics](#)
- 使用 Google 等搜索引擎搜索



教学提纲

1 引言

2 DNA 序列转换与组份分析

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 重复序列分析

8 总结与答疑

9 引言

10 基因识别

11 mRNA 选择性剪接

12 miRNA 及其靶基因预测

13 lncRNA

14 查找数据库与工具

15 总结与答疑



知识点

- 基因识别——原核和真核的基因结构，基因识别方法
- mRNA 可变剪接——选择性剪接的类型，数据资源
- miRNA——miRNA 的特点，miRNA 预测方法与工具，miRNA 靶基因预测方法与工具

技能

- 查找数据库——时效性
- 查找分析工具——适用范围



Powered by



T_EX L^AT_EX X_ET_EX Beamer