

生物信息学

天津医科大学
生物医学工程学院

2013-2014 学年下学期 (春)
2011 级生信班

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

2014 年 3 月



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 EMBOSS
- 8 总结与答疑
- 9 引言
- 10 重复序列分析

- 11 基因识别
- 12 查找数据库与分析工具
- 13 总结与答疑
- 14 引言
- 15 mRNA 选择性剪接
- 16 miRNA 及其靶基因预测
- 17 lncRNA
- 18 学习数据库与分析工具的使用
- 19 总结与答疑
- 20 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

EMBOSS

8

总结与答疑

9

引言

10

重复序列分析

11

基因识别

12

查找数据库与分析工具

13

总结与答疑

14

引言

15

mRNA 选择性剪接

16

miRNA 及其靶基因预测

17

lncRNA

18

学习数据库与分析工具的使用

19

总结与答疑

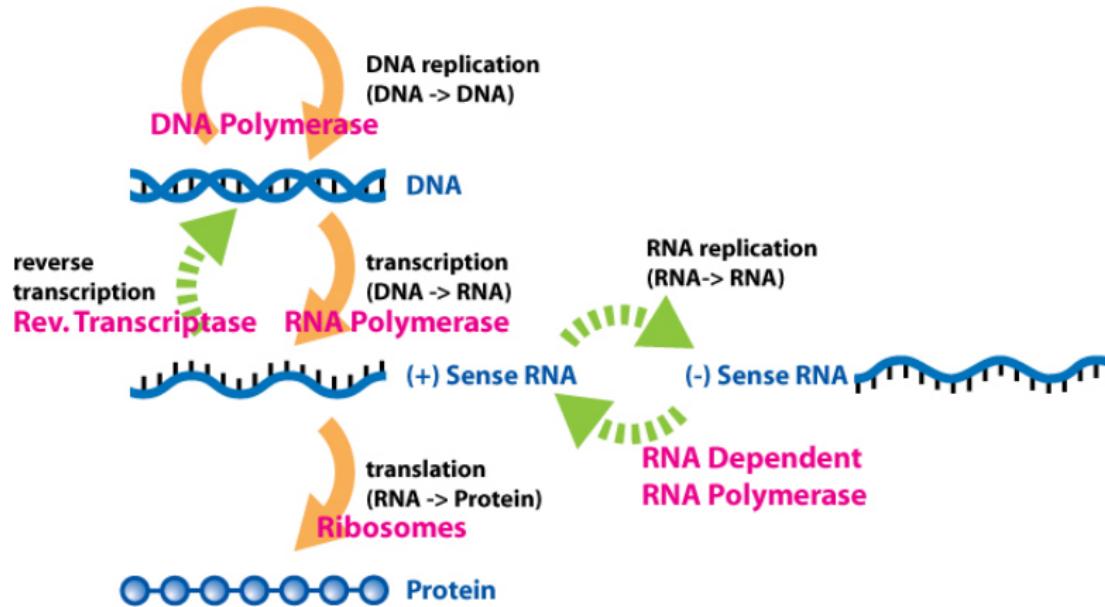
20

复习思考题

引言 | 大千世界



引言 | 中心法则



A COMPLETE GENOME IN TIME

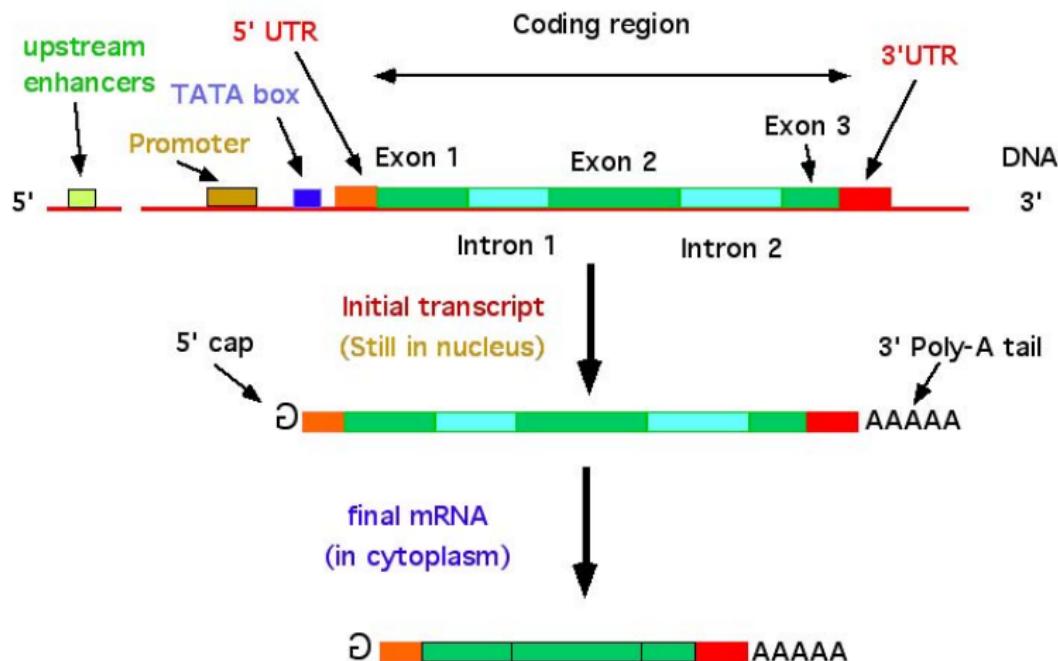
by Yonder Biology



引言 | ACGT⇒ 生信



引言 | 遗传信息



教学提纲

- 1 引言
- 2 DNA 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 EMBOSS
- 8 总结与答疑
- 9 引言
- 10 重复序列分析

- 11 基因识别
- 12 查找数据库与分析工具
- 13 总结与答疑
- 14 引言
- 15 mRNA 选择性剪接
- 16 miRNA 及其靶基因预测
- 17 lncRNA
- 18 学习数据库与分析工具的使用
- 19 总结与答疑
- 20 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

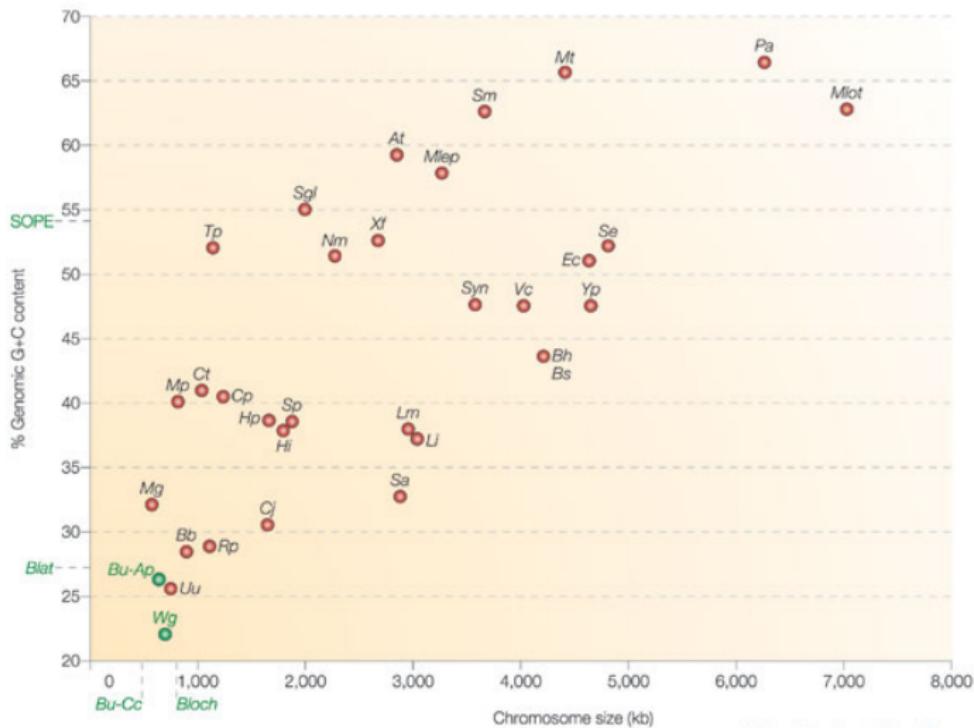


GC 含量 (GC content)

- 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) : $\frac{G+C}{A+T}$



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					



任务分析

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



任务分析

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



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]



- 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 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 EMBOSS
- 8 总结与答疑
- 9 引言
- 10 重复序列分析

- 11 基因识别
- 12 查找数据库与分析工具
- 13 总结与答疑
- 14 引言
- 15 mRNA 选择性剪接
- 16 miRNA 及其靶基因预测
- 17 lncRNA
- 18 学习数据库与分析工具的使用
- 19 总结与答疑
- 20 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
- 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：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

EMBOSS

8

总结与答疑

9

引言

10

重复序列分析

11

基因识别

12

查找数据库与分析工具

13

总结与答疑

14

引言

15

mRNA 选择性剪接

16

miRNA 及其靶基因预测

17

lncRNA

18

学习数据库与分析工具的使用

19

总结与答疑

20

复习思考题

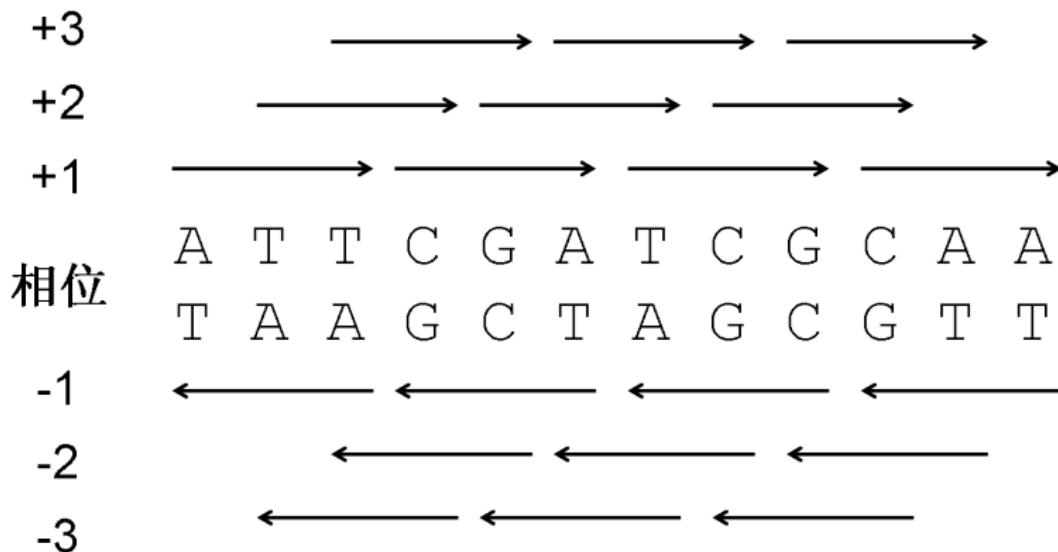


开放阅读框 (Open Reading Frame, ORF)

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



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



- 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 EMBOSS
- 8 总结与答疑
- 9 引言
- 10 重复序列分析

- 11 基因识别
- 12 查找数据库与分析工具
- 13 总结与答疑
- 14 引言
- 15 mRNA 选择性剪接
- 16 miRNA 及其靶基因预测
- 17 lncRNA
- 18 学习数据库与分析工具的使用
- 19 总结与答疑
- 20 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 (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。



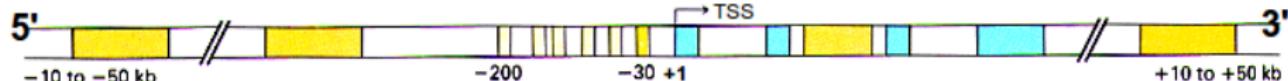
启动子 | 定义

启动子 (promoter)

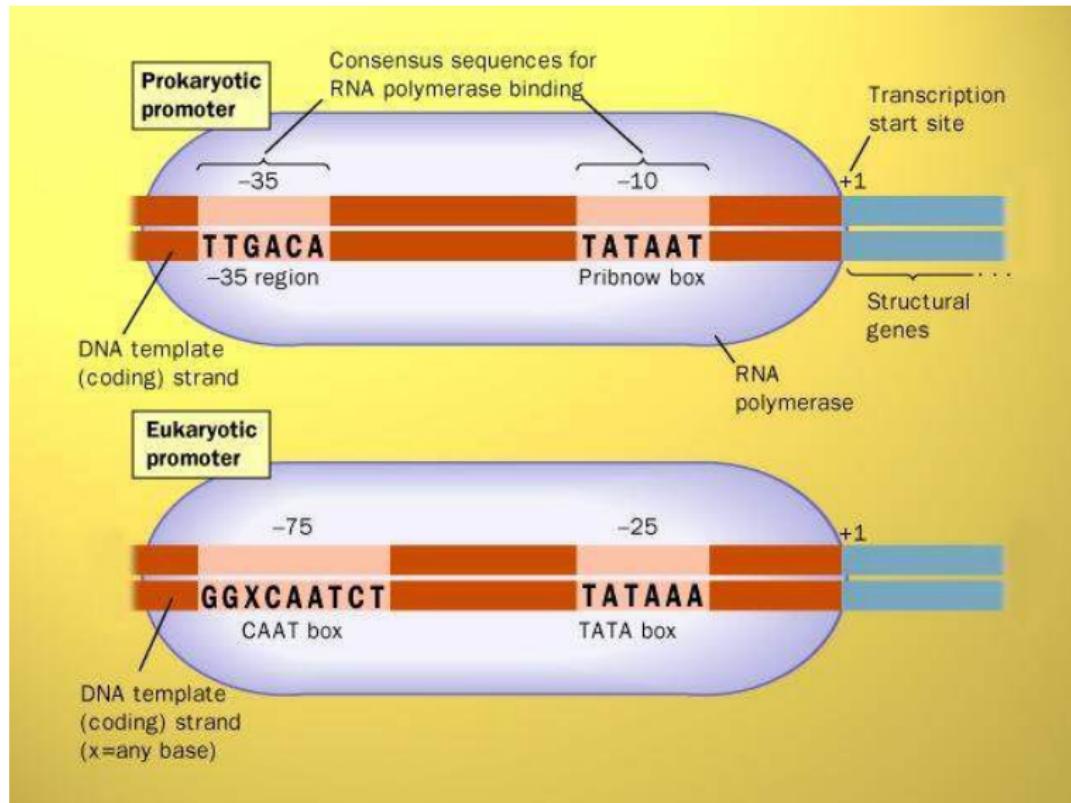
一段位于转录起始位点 5' 端上游区的 DNA 序列，能活化 RNA 聚合酶，使之与模板 DNA 准确地结合并具有转录起始的特异性。

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 | 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：有注释、非冗余的真核生物 RNA 聚合酶 II 启动子数据集
- Promoter Scan, Promoter 2.0

- 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)



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 EMBOSS
- 8 总结与答疑
- 9 引言
- 10 重复序列分析

- 11 基因识别
- 12 查找数据库与分析工具
- 13 总结与答疑
- 14 引言
- 15 mRNA 选择性剪接
- 16 miRNA 及其靶基因预测
- 17 lncRNA
- 18 学习数据库与分析工具的使用
- 19 总结与答疑
- 20 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
- 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 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 EMBOSS
- 8 总结与答疑
- 9 引言
- 10 重复序列分析

- 11 基因识别
- 12 查找数据库与分析工具
- 13 总结与答疑
- 14 引言
- 15 mRNA 选择性剪接
- 16 miRNA 及其靶基因预测
- 17 lncRNA
- 18 学习数据库与分析工具的使用
- 19 总结与答疑
- 20 复习思考题



简介

EMBOSS (The European Molecular Biology Open Software Suite) 是一个开源、免费的序列分析软件包，整合了目前可以获得的大部分序列分析软件。

使用 EMBOSS，可以将系列分析工作进行无缝整合，弥补了许多软件功能分散、分析效率低下的缺陷。

使用

- 操作系统：Linux, Mac, Windows
- 界面：JEMBOSS (java) , EMBOSS Explorer (web)



- 最重要的程序。Wossname：根据关键字查找程序；Showdb：显示所有整合的数据库。
- 序列编辑。Revseq：将序列反转并互补；Seqret：序列格式转换。
- 两个序列相似性图形表达。Dottup：精确匹配；Dotmatcher：近似匹配。
- 双序列比对。Needle：全局比对；Water：局部比对。
- 多序列比对。Emma：clustalW。
- 寻找 SNP。Deffseq：仅限于双序列比对中。
- 其他。Plotorf, Getorf：翻译；Iep：等电点预测；Tmap：跨膜区预测；Pepinfo：蛋白质性质；Patmatmotifs：Motif 搜索。



组份分析

- compseq: Calculate the composition of unique words in sequences
- geecee: Calculate fractional GC content of nucleic acid sequences
- revseq: Reverse and complement a nucleotide sequence

CpG 岛分析

- extractseq: Extract regions from a sequence
- cpgplot: Identify and plot CpG islands in nucleotide sequence(s)
- cpgreport: Identify and report CpG-rich regions in nucleotide sequence(s)
- isochore: Plot isochores in DNA sequences

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 EMBOSS
- 8 总结与答疑
- 9 引言
- 10 重复序列分析
- 11 基因识别
- 12 查找数据库与分析工具
- 13 总结与答疑
- 14 引言
- 15 mRNA 选择性剪接
- 16 miRNA 及其靶基因预测
- 17 lncRNA
- 18 学习数据库与分析工具的使用
- 19 总结与答疑
- 20 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

技能——解决问题的思路

- 首先分析任务的属性
- 寻找可能的解决方案
- 确定最合适的方法
- 先易后难, 由浅入深

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 EMBOSS
- 8 总结与答疑
- 9 引言
- 10 重复序列分析

- 11 基因识别
- 12 查找数据库与分析工具
- 13 总结与答疑
- 14 引言
- 15 mRNA 选择性剪接
- 16 miRNA 及其靶基因预测
- 17 lncRNA
- 18 学习数据库与分析工具的使用
- 19 总结与答疑
- 20 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

→ 非重叠序列

→ 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



- 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 EMBOSS
- 8 总结与答疑
- 9 引言
- 10 重复序列分析

- 11 基因识别
- 12 查找数据库与分析工具
- 13 总结与答疑
- 14 引言
- 15 mRNA 选择性剪接
- 16 miRNA 及其靶基因预测
- 17 lncRNA
- 18 学习数据库与分析工具的使用
- 19 总结与答疑
- 20 复习思考题



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
 - 卫星 DNA (satellite DNA) : 5 ~ 200bp, 几百万个拷贝, 着丝粒部位
 - 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：分散于染色体的各位点上
 - 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 ²	Functions, features of family members
Satellites — large tandem arrays				
Microsatellite	2–5	0.2–0.5	3×10^5	Repeat expansion causes cancer
Minisatellite	~15	0.5–3	10^5	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				
Retrotransposons				
<i>LTR-containing elements</i>				
<i>copia², gypsy²</i>	~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
SINES				
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



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



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 EMBOSS
- 8 总结与答疑
- 9 引言
- 10 重复序列分析

- 11 基因识别
- 12 查找数据库与分析工具
- 13 总结与答疑
- 14 引言
- 15 mRNA 选择性剪接
- 16 miRNA 及其靶基因预测
- 17 lncRNA
- 18 学习数据库与分析工具的使用
- 19 总结与答疑
- 20 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。



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。

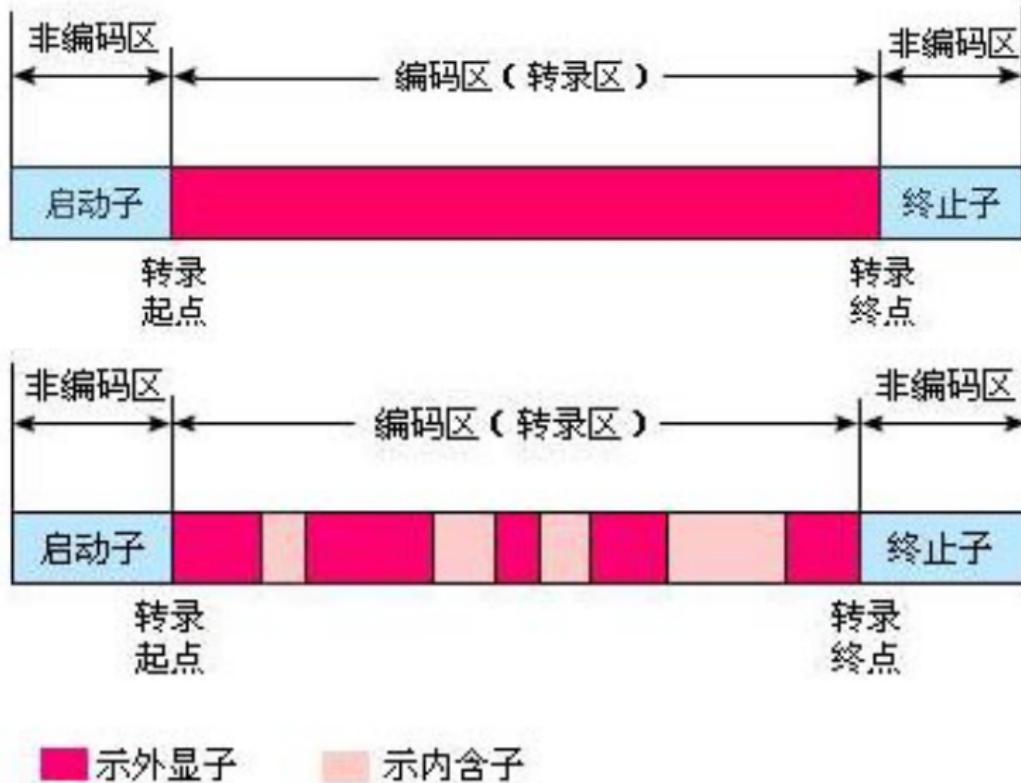
一个完整的基因，不仅包括编码区，还包括 5' 末端和 3' 末端长度不等的特异性序列。

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）



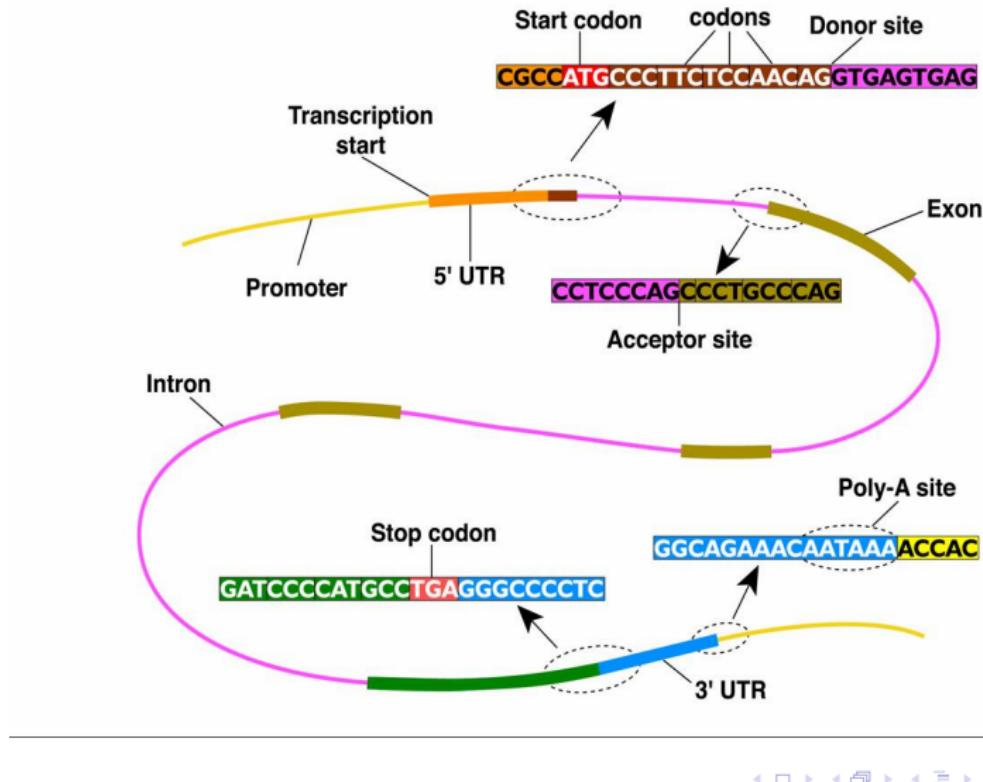
信号

- 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
	GUA	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

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

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

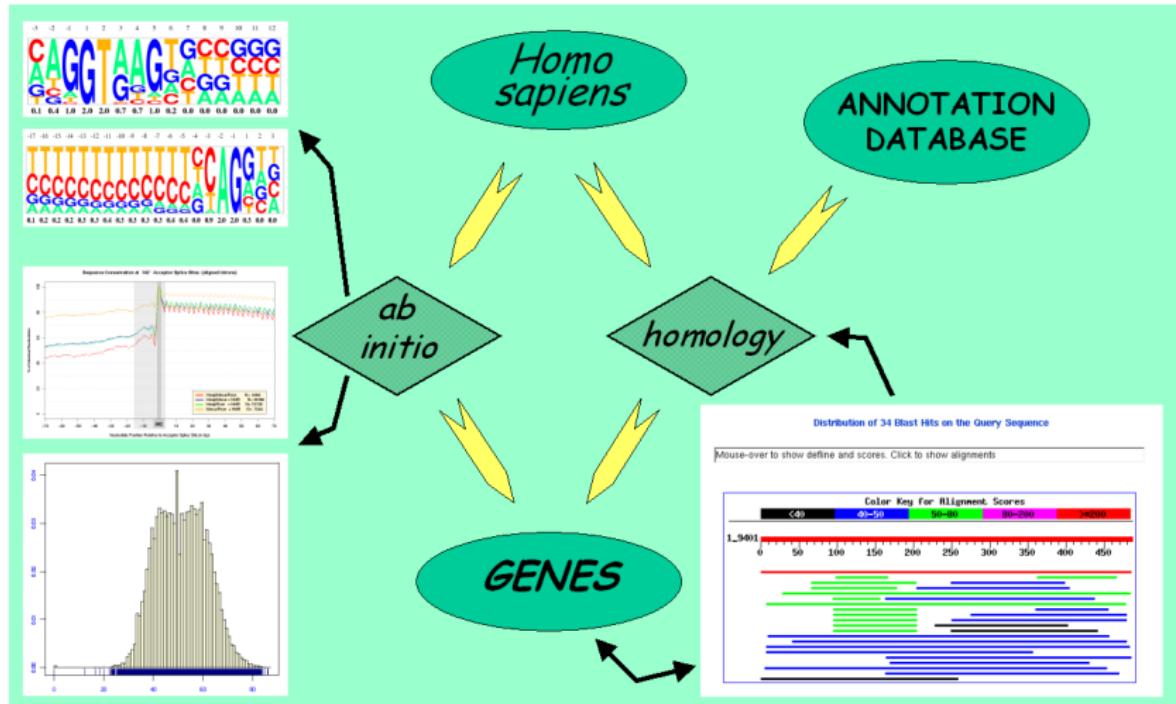
内容

密码子使用偏好性，双联密码子出现频率，基因组等值区

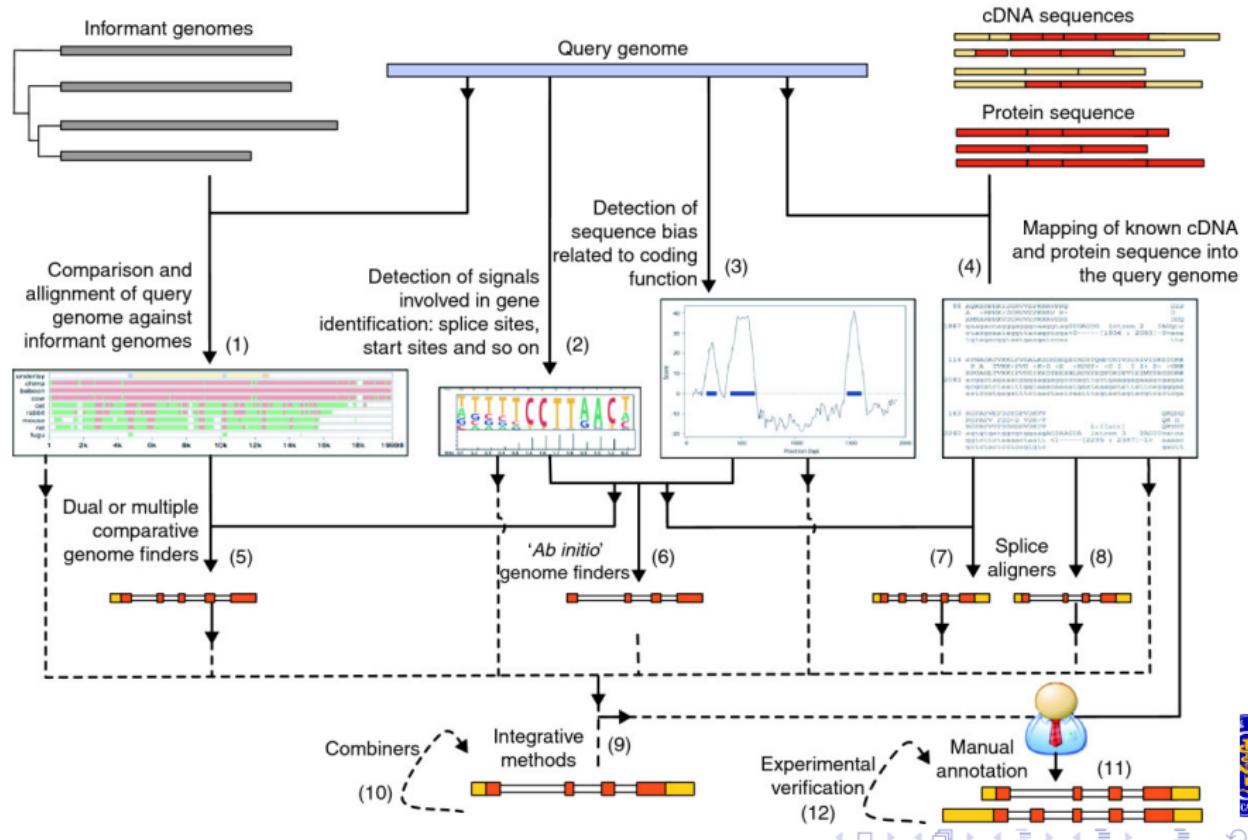
总结

- 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	
geneld	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
- [List of gene prediction software\(Wikipedia\)](#)
- [Computational prediction of eukaryotic protein-coding genes, Box 2, Useful internet resources](#)



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 EMBOSS
- 8 总结与答疑
- 9 引言
- 10 重复序列分析

- 11 基因识别
- 12 查找数据库与分析工具
- 13 总结与答疑
- 14 引言
- 15 mRNA 选择性剪接
- 16 miRNA 及其靶基因预测
- 17 lncRNA
- 18 学习数据库与分析工具的使用
- 19 总结与答疑
- 20 复习思考题



查找数据库与分析工具

- 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



- 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 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

EMBOSS

8

总结与答疑

9

引言

10

重复序列分析

11

基因识别

12

查找数据库与分析工具

13

总结与答疑

14

引言

15

mRNA 选择性剪接

16

miRNA 及其靶基因预测

17

lncRNA

18

学习数据库与分析工具的使用

19

总结与答疑

20

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

技能——查找数据库与分析工具

- 借鉴文献、收集专刊、请教专家、搜索网络
- 数据库有其时效性
- 分析工具有其适用范围



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

EMBOSS

8

总结与答疑

9

引言

10

重复序列分析

11

基因识别

12

查找数据库与分析工具

13

总结与答疑

14

引言

15

mRNA 选择性剪接

16

miRNA 及其靶基因预测

17

lncRNA

18

学习数据库与分析工具的使用

19

总结与答疑

20

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 识别子

更多内容



● 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
- lncRNA



- 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
- lncRNA



- 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
- lncRNA



- 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
- lncRNA



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.

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



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

EMBOSS

8

总结与答疑

9

引言

10

重复序列分析

11

基因识别

12

查找数据库与分析工具

13

总结与答疑

14

引言

15

mRNA 选择性剪接

16

miRNA 及其靶基因预测

17

lncRNA

18

学习数据库与分析工具的使用

19

总结与答疑

20

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。



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

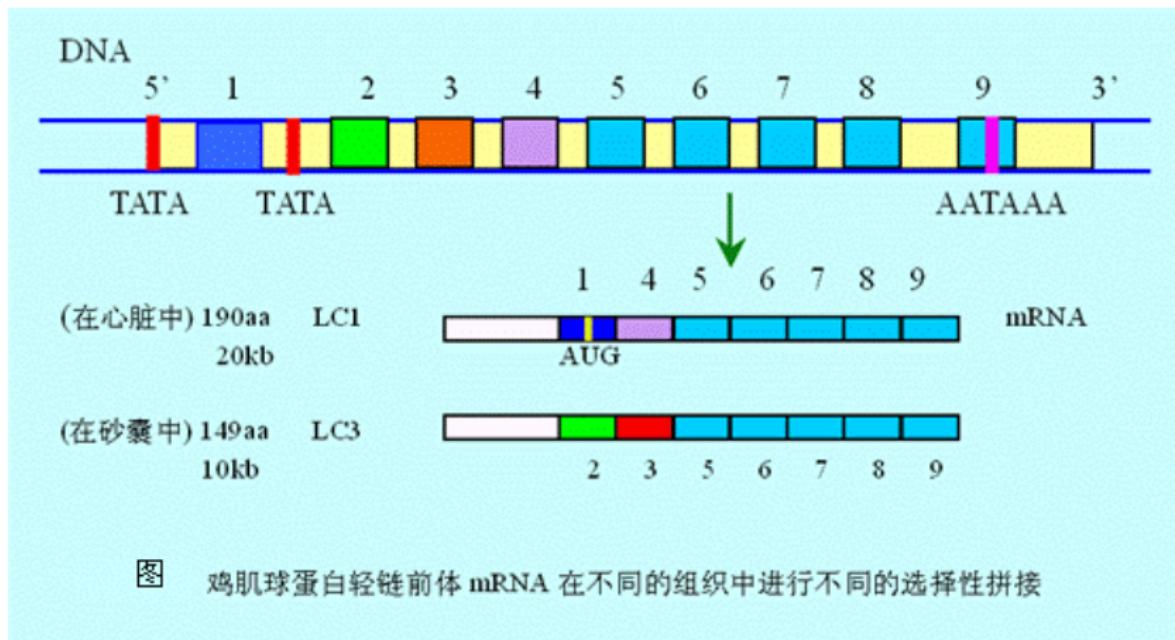
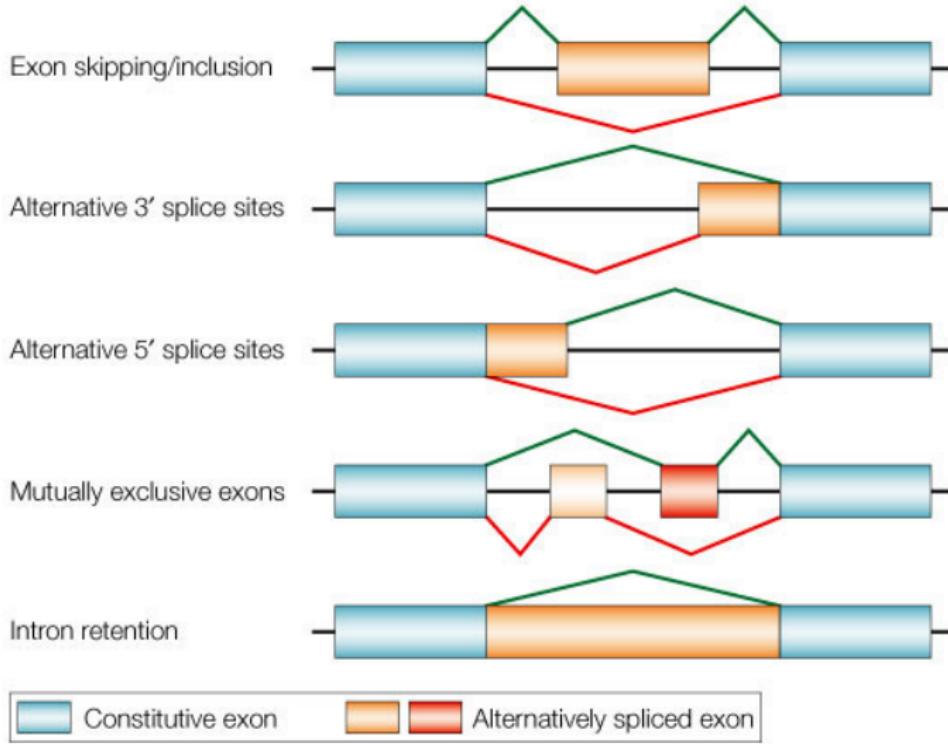


图 鸡肌球蛋白轻链前体 mRNA 在不同的组织中进行不同的选择性拼接

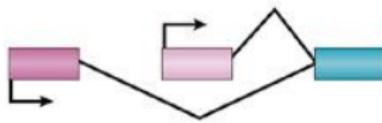


选择性剪接 | 机制 | 五种

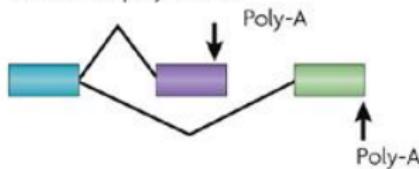


选择性剪接 | 机制 | 七种

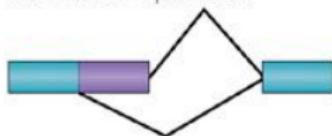
Alternative promoters



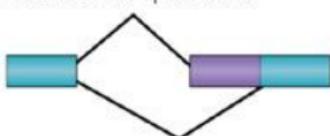
Alternative poly-A sites



Alternative 5' splice sites



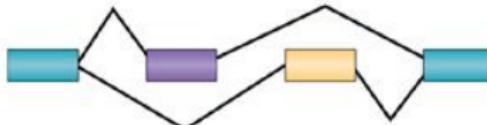
Alternative 3' splice sites



Cassette exon



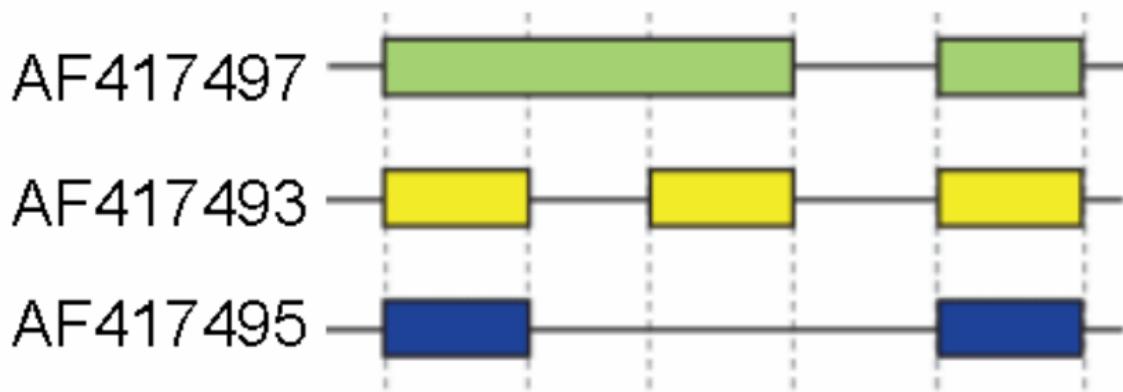
Mutually exclusive exons



Retained intron



选择性剪接 | 机制 | 复杂实例



选择性剪接 | 数据库与分析工具

- 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 EMBOSS
- 8 总结与答疑
- 9 引言
- 10 重复序列分析

- 11 基因识别
- 12 查找数据库与分析工具
- 13 总结与答疑
- 14 引言
- 15 mRNA 选择性剪接
- 16 miRNA 及其靶基因预测
- 17 lncRNA
- 18 学习数据库与分析工具的使用
- 19 总结与答疑
- 20 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。



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

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

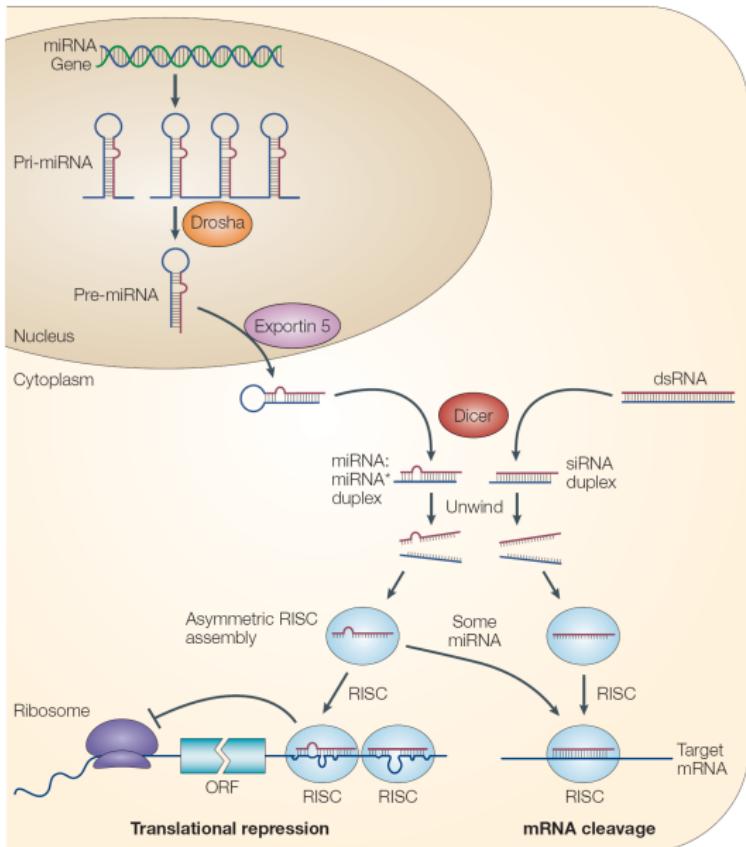
物理位置

倾向于成簇地出现在染色体上

进化

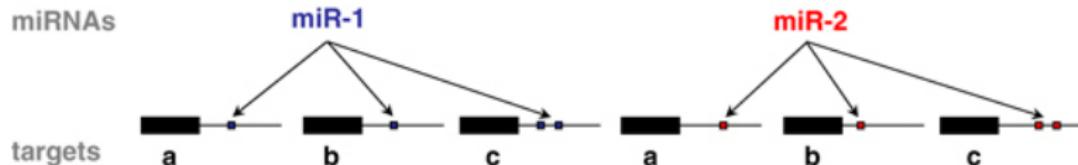
在物种间高度保守

miRNA | 生成

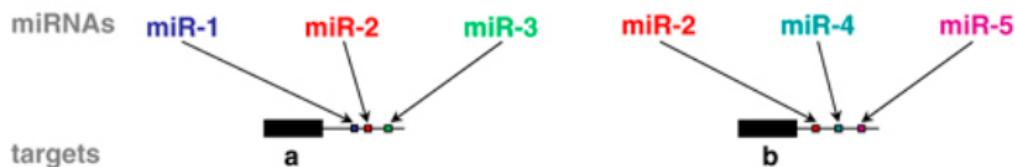


miRNA | 作用网络

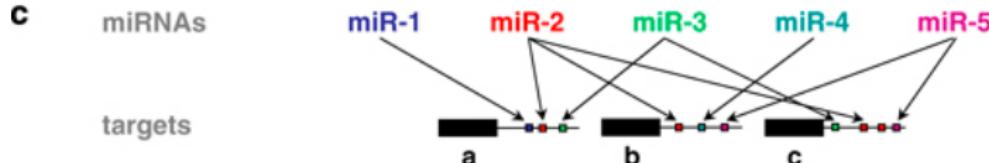
a

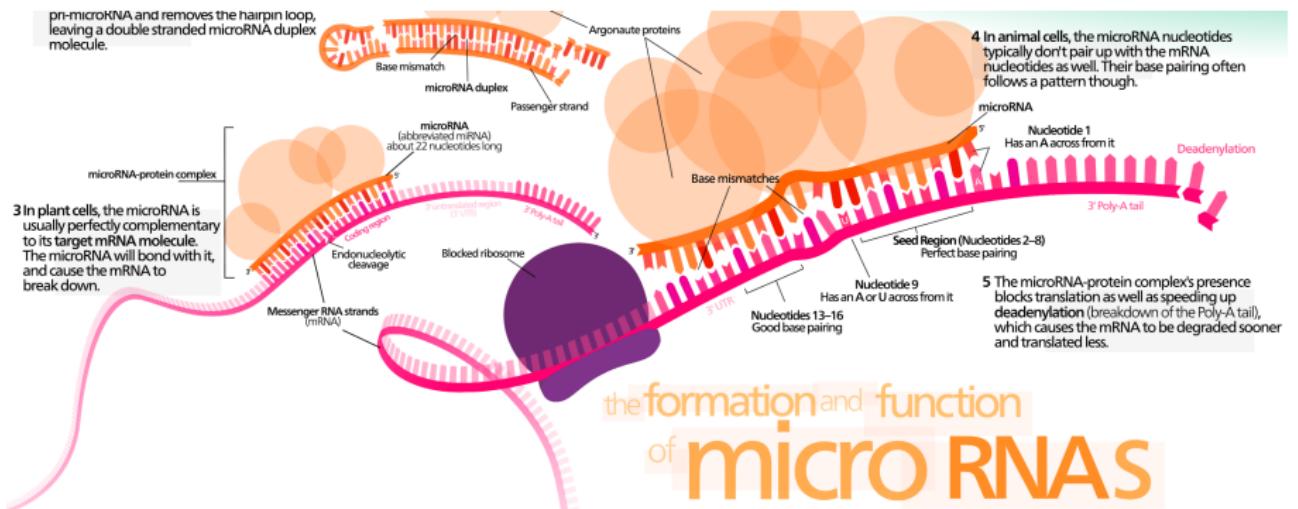


b



c

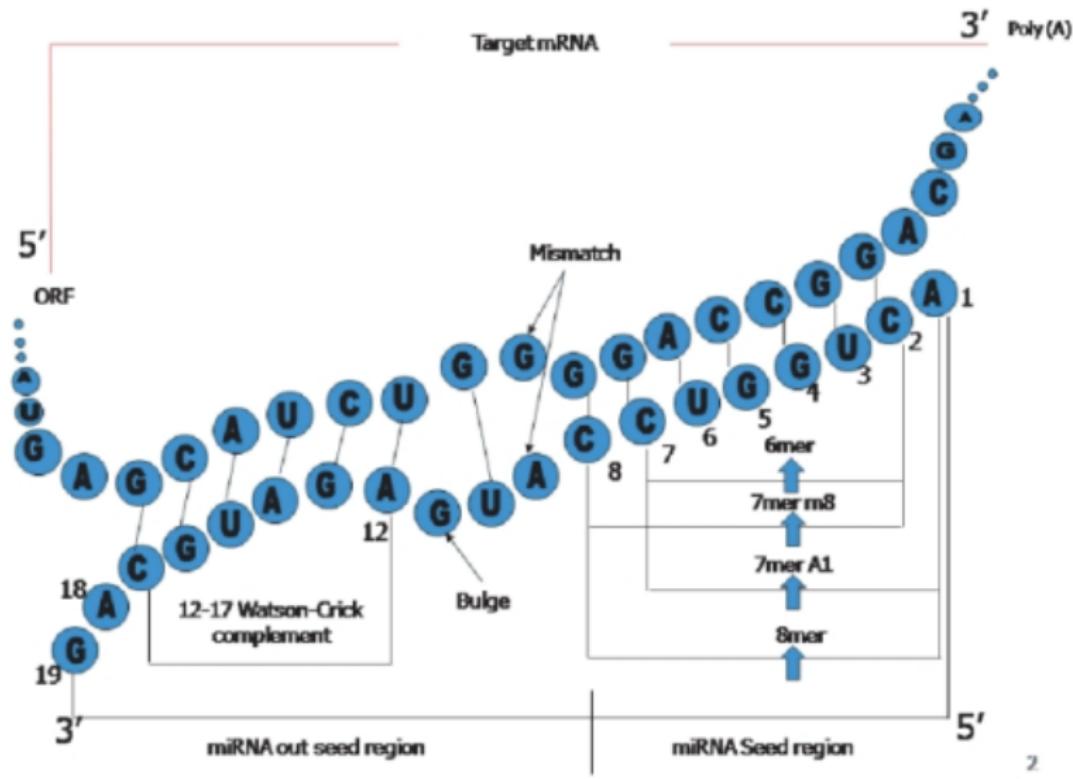




- ① 同源片段搜索方法
- ② 基于比较基因组学的预测方法
- ③ 基于序列和结构特征打分的预测方法
- ④ 结合作用靶标的预测方法
- ⑤ 基于机器学习的预测方法



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 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 EMBOSS
- 8 总结与答疑
- 9 引言
- 10 重复序列分析

- 11 基因识别
- 12 查找数据库与分析工具
- 13 总结与答疑
- 14 引言
- 15 mRNA 选择性剪接
- 16 miRNA 及其靶基因预测
- 17 lncRNA
- 18 学习数据库与分析工具的使用
- 19 总结与答疑
- 20 复习思考题



类似于 mRNA

- 大多被 RNA 聚合酶 II 所转录
- 有 5' 帽子和 3' 端的 poly(A) 尾巴
- 剪接现象
- 启动子区域和剪接位置具有保守性

独特性

- 长度偏短、外显子数目偏少
- 不存在较长的 ORF
- 密码子偏好性与内含子区域相似
- 二级结构中有丰富的长茎发夹结构
- 在不同物种间的保守性差
- 主要富集在细胞核

类似于 mRNA

- 大多被 RNA 聚合酶 II 所转录
- 有 5' 帽子和 3' 端的 poly(A) 尾巴
- 剪接现象
- 启动子区域和剪接位置具有保守性

独特性

- 长度偏短、外显子数目偏少
- 不存在较长的 ORF
- 密码子偏好性与内含子区域相似
- 二级结构中有丰富的长茎发夹结构
- 在不同物种间的保守性差
- 主要富集在细胞核

- 其表达具有时空特异性，与特定的生物过程相关
- 具有复杂的调控功能，在染色质改变、转录调控及后转录调控中发挥重要作用
- 复杂的代谢机制，大多数 lncRNA 是稳定的，半衰期的变化范围较大
- 与疾病存在密切关系，如肿瘤、阿尔兹海默病、心血管疾病等

数据库

长链非编码 RNA 数据库（维基百科）



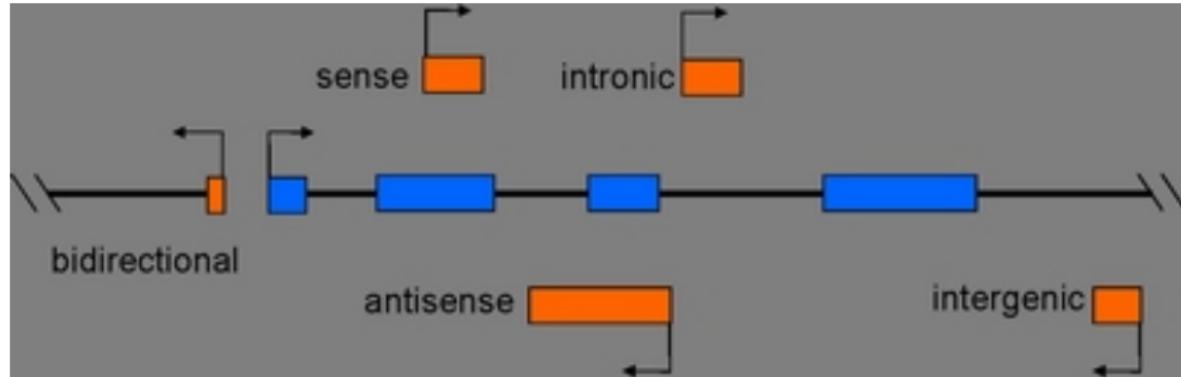
- 其表达具有时空特异性，与特定的生物过程相关
- 具有复杂的调控功能，在染色质改变、转录调控及后转录调控中发挥重要作用
- 复杂的代谢机制，大多数 lncRNA 是稳定的，半衰期的变化范围较大
- 与疾病存在密切关系，如肿瘤、阿尔兹海默病、心血管疾病等

数据库

长链非编码 RNA 数据库（维基百科）



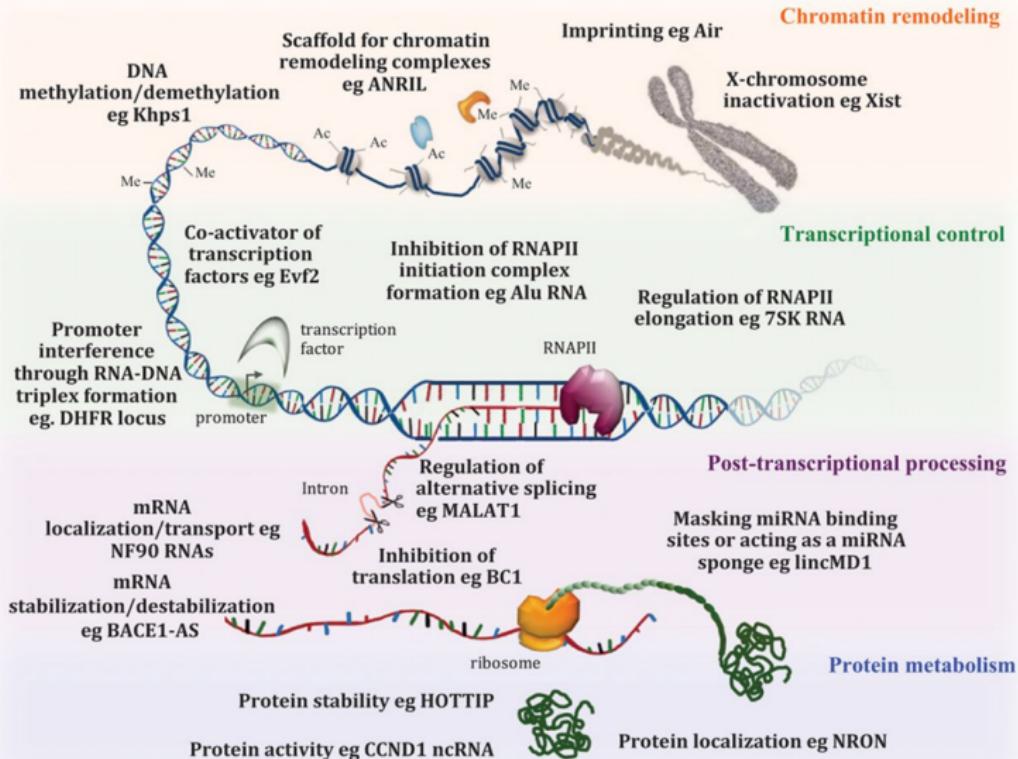
lncRNA | 类型



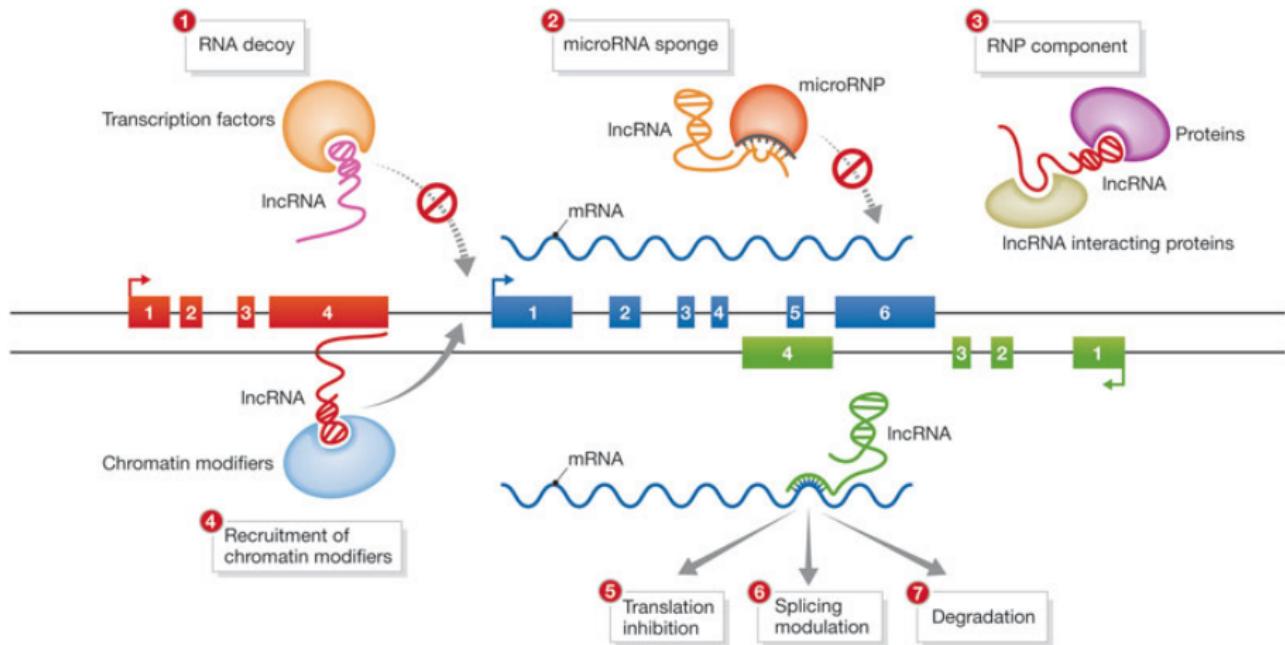
Long non-coding RNA	Symbol	References
Long intergenic non-coding RNA	LincRNA	[19,20]
Long intronic non-coding RNA		[14,21]
Natural antisense transcript	NAT	[22–24]
Promoter-associated long RNA	PALR	[25]
Promoter upstream transcript	PROMPT	[26]
Repetitive element-associated non-coding RNA		[27–29]
Transcribed pseudogene		[30,31]
Transcribed ultraconserved region	T-UCR	[32]
Enhancer-like non-coding RNA	eRNA	[33]



lncRNA | 生物功能



lncRNA | 作用机制



lncRNA | lncRNA 与疾病

LncRNA	Disease	References
aHIF	Multiple cancers	[61,62]
AK023948	Papillary thyroid carcinoma	[63]
ANRIL	Coronary artery disease; Multiple cancers	[64–68]
ASFMR1	Fragile X syndrome; Fragile X tremor ataxia syndrome	[69]
ATXN8OS	Spinocerebellar atrophy type 8	[70]
BACE1-AS	Alzheimer's disease	[71]
BC200	Alzheimer's disease; Multiple cancers	[72–74]
BIC	B-cell lymphoma	[75]
CUDR	Squamous carcinoma	[76]
DD3	Prostate cancer	[77,78]
FMR4	Fragile X syndrome; Fragile X tremor ataxia syndrome	[79]
GAS5	Breast cancer	[80]
H19	Multiple cancers	[81–85]
HOTAIR	Multiple cancers	[16,86]
HULC	Multiple cancers	[87,88]
Kcnq1ot1	Colon cancer	[89]
Kras1p	Prostate cancer	[30]
Linc-p21	Lung cancer	[43]
LOC285194	Osteosarcoma	[90]



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

EMBOSS

8

总结与答疑

9

引言

10

重复序列分析

11

基因识别

12

查找数据库与分析工具

13

总结与答疑

14

引言

15

mRNA 选择性剪接

16

miRNA 及其靶基因预测

17

lncRNA

18

学习数据库与分析工具的使用

19

总结与答疑

20

复习思考题



学习数据库与分析工具的使用

- 阅读官方的帮助手册
- 请教有使用经验的专家
- 查找简单的使用实例，并重复其操作步骤
- 使用 Google 等搜索引擎搜索相关资料
- 各种 protocols 期刊：*Nature protocols, Current Protocols (in Bioinformatics), SpringerProtocols, Methods in Molecular Biology*



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 EMBOSS
- 8 总结与答疑
- 9 引言
- 10 重复序列分析

- 11 基因识别
- 12 查找数据库与分析工具
- 13 总结与答疑
- 14 引言
- 15 mRNA 选择性剪接
- 16 miRNA 及其靶基因预测
- 17 lncRNA
- 18 学习数据库与分析工具的使用
- 19 总结与答疑
- 20 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

技能——学习数据库与分析工具的使用

- 阅读手册、请教专家、重复实例、搜索网络
- 历史资料使用的是历史版本



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 EMBOSS
- 8 总结与答疑
- 9 引言
- 10 重复序列分析

- 11 基因识别
- 12 查找数据库与分析工具
- 13 总结与答疑
- 14 引言
- 15 mRNA 选择性剪接
- 16 miRNA 及其靶基因预测
- 17 lncRNA
- 18 学习数据库与分析工具的使用
- 19 总结与答疑
- 20 复习思考题



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。
- 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。

技能

- ① 以计算 GC 含量为例，论述解决思路，即如何通过分析问题的属性确定相应的策略从而找到最合适的方法。
- ② 在解决生物信息学问题时，论述找到所需数据库和分析工具并掌握其使用方法的策略。

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。
- 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。

技能

- ① 以计算 GC 含量为例，论述解决思路，即如何通过分析问题的属性确定相应的策略从而找到最合适的方法。
- ② 在解决生物信息学问题时，论述找到所需数据库和分析工具并掌握其使用方法的策略。

Powered by



T_EX L^AT_EX X_ET_EX Beamer