生物信息学

天津医科大学
生物医学工程与技术学院

2017-2018 学年上学期（秋）
公共选修课

第四章 核酸序列分析

伊现富（Yi Xianfu）

天津医科大学（TIJMU）
生物医学工程与技术学院

2017 年 9 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
3. 扩展知识：EMBOSS，相关数学与算法，解决问题的思路

4.2 基因组结构注释分析

1. 重复序列
2. 基因识别
3. 扩展知识：查找数据库与分析工具

4.3 RNA 序列分析

1. mRNA 选择性剪接
2. miRNA 及其靶基因
3. 扩展知识：lncRNA，学习数据库与工具的使用

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. 功能位点分析
6. 启动子分析
7. CpG 岛识别
8. EMBOSS
9. 序列分析中的算法
10. 总结与答疑
11. 引言
12. 重复序列分析
13. 基因识别
14. 查找数据库与分析工具
15. 总结与答疑
16. 引言
17. mRNA 选择性剪接
18. miRNA 及其靶基因预测
19. lncRNA
20. 学习数据库与分析工具的使用
21. 总结与答疑
22. 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 功能位点分析
6 启动子分析
7 CpG 岛识别
8 EMBOSS
9 序列分析中的算法
10 总结与答疑
11 引言
12 重复序列分析
13 基因识别
14 查找数据库与分析工具
15 总结与答疑
16 引言
17 mRNA 选择性剪接
18 miRNA 及其靶基因预测
19 lncRNA
20 学习数据库与分析工具的使用
21 总结与答疑
22 复习思考题

引言 | 大千世界

鸟类
哺乳类
被子植物
爬行类
裸子植物
鱼类
园口类
原索动物
藻类
节肢动物
苔藓
棘皮动物
软体动物
菌类
环节动物
腔肠动物
扁形动物
藻类
原生动物
真核生物
原核生物
细菌
兰藻

引言 | 中心法则

DNA replication
(DNA -> DNA)
DNA Polymerase
DNA
reverse
transcription
Rev. Transcriptase
transcription
(DNA -> RNA)
RNA Polymerase
RNA replication
(RNA-> RNA)
(+) Sense RNA
(-) Sense RNA
RNA Dependent
RNA Polymerase
translation
(RNA -> Protein)
Ribosomes
Protein

引言 | DNA

base pairs:
C≡G
T=A
hydrogen bond
sugar phosphate
backbone
exon
intron

引言 | DNA & RNA

Cytosine
C
Guanine
G
Adenine
A
Uracil
U
Nucleobases of RNA
Nucleobases
Base pair
helix of sugar-phosphates
RNA
Ribonucleic acid
DNA
Deoxyribonucleic acid
Cytosine
C
Guanine
G
Adenine
A
Thymine
T
Nucleobases of DNA

引言 | ACGT

引言 | ACGT⇒ 生信

upstream
enhancers
5' UTR
Coding region
3'UTR
TATA box
Promoter
Exon 1
Exon 2
Exon 3
DNA
5'
3'
Intron 1
Intron 2
Initial transcript
(Still in nucleus)
5' cap
3' Poly-A tail
AAAAA
final mRNA
(in cytoplasm)
AAAAA

- 遗传信息的分布不是随机的
- 遗传信息分布的模式可以遗传
- 不同物种间对应相同或类似功能或结构的遗传信息的分布模式具有相似性

核酸序列有与生物学功能相对应的规律和特征！

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。

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。

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。

引言 | 序列分析

cDNA
EST-genome alignment
Genomic sequence
Nucleotide BLAST
Open reading frame prediction
Identification of new/related sequences
Translation
PSI-BLAST
Protein BLAST
Multiple or pairwise
P alignments N
Boxshade
Protein
SMART
ProDom
Interpro
Prints
Pfam
Prosite
Integrated in NIX:
(registration needed)
Nucleotide BLAST
Repeat identification
Gene finders:
Grail
Genefinder
Genemark
Fex
Fgene
Hexon
tRNAscan
PolyAH
Promoter Inspector
MatInspector
TRANSFAC

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 功能位点分析
6 启动子分析
7 CpG 岛识别
8 EMBOSS
9 序列分析中的算法
10 总结与答疑
11 引言
12 重复序列分析
13 基因识别
14 查找数据库与分析工具
15 总结与答疑
16 引言
17 mRNA 选择性剪接
18 miRNA 及其靶基因预测
19 lncRNA
20 学习数据库与分析工具的使用
21 总结与答疑
22 复习思考题

查戈夫法则

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

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 ～ 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 ～ 50000 个核苷酸

DNA 序列 | 序列长度 | 基因组

DNA 序列 | 碱基组成

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

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）：$\frac{A+T}{G+C}$
- 结合滑动窗口进行计算

DNA 序列 | GC 含量 | 分析

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 | 分析

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

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

		Coding "o" regions [a]		Intron regions [b][c]			Difference
Code	Yeast	GC	AT	GC	AT	n	Intron-exon [b]
	S. cerevisiae	39.6	60.4	33.4	66.6	260	6.2
AT	*S. servazzii*	34.7	65.3	27.6	72.4	22	7.1
AU	*S. kluyveri*	41.5	58.5	36.8	63.2	27	4.7
AZ	*K. marxianus*	42.3	57.7	34.5	65.5	13	7.8
BD	*C. tropicalis*	34.5	65.5	26.9	73.1	7	7.6
BC	*D. hansenii*	36.5	63.5	33.6	66.4	12	2.9
BB	*P. angusta*	48.5	51.5	41.8	58.2	29	6.7
AW	*Y. lipolytica*	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	*Escherichia coli* plasmid pC15-1a	NC_005327.1	63.66	52.6
	8877592	*Klebsiella pneumoniae* plasmid pKF3-140	NC_013951.1	63.21	52.5
	7886608	*Salmonella enterica* plasmid pAM04528	NC_012693.1	62.43	51.9
	7003405	*Haemophilus influenzae* plasmid ICEhin1056	NC_011409.1	43.36	39.1
	2653967	*Serratia marcescens* plasmid R478	NC_004989.1	43.28	36.9
	4927413	*Yersinia pestis* biovar Orientalis str. IP275 pIP1202	NC_009141.1	57.63	52.9
	6002612	*Acinetobacter baumannii* AYE	NC_010410.1	63.21	39.3
	1794537	*Rhodopirellula baltica* SH 1	NC_005027.1	57.55	55.4
	3433250	*Corynebacterium jeikeium* K411	NC_007164.1	68.50	61.40
	2797858	*Listeria monocytogenes* serotype 4b str. F2365	NC_002973.6	42.30	38.00
					$p = 0.02$
transferase	1238790	*Klebsiella pneumoniae* plasmid pJHCMW1	NC_003486.1	52.97	49
	13919580	*Providencia stuartii* plasmid pTC2	NC_019375.1	53.03	52.5
	9487131	*Klebsiella pneumoniae* plasmid pNL194	NC_014368.1	53.03	53.1
	9487121	*Klebsiella pneumoniae* plasmid pNL194	NC_014368.1	52.9	53.1
	1055588	*Citrobacter freundii* plasmid pCTX-M3	NC_004464.2	51.89	51
	7156160	*Escherichia coli* UMN026	NC_011751.1	58.37	50.6
	6810778	*Salmonella enterica* subsp. *enterica* serovar Paratyphi A str. AKU_12601	NC_011147.1	54.11	52.2
	6455499	*Cupriavidus taiwanensis* LMG 19424, Chr 2	NC_010530.1	70.94	67
	6928720	*Burkholderia cenocepacia* J2315, Chr 2	NC_011001.1	71.33	66.9
	2662391	*Bordetella bronchiseptica* RB50	NC_002927.3	73.45	68.1
					$p = 0.2$

DNA 序列 | GC 含量 | 滑动窗口

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' $\Longrightarrow$ 3' [右]
- 多肽/蛋白质：[左] N 端（氨基端）$\Longrightarrow$ 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' $\Longrightarrow$ 3' [右]
- 多肽/蛋白质：[左] N 端（氨基端）$\Longrightarrow$ C 端（羧基端）[右]

DNA 序列 | N 联核苷酸

简介

N（2,3,4,……）个连续出现的核苷酸，也叫 *k*-mer。

常见

- 二联核苷酸：$4 \times 4 = 16$
- 三联核苷酸：$4 \times 4 \times 4 = 64$
- 三联核苷酸 $\Longrightarrow$ 密码子

DNA 序列 | N 联核苷酸

简介

N（2,3,4,……）个连续出现的核苷酸，也叫 *k*-mer。

常见

- 二联核苷酸：$4 \times 4 = 16$
- 三联核苷酸：$4 \times 4 \times 4 = 64$
- 三联核苷酸 $\Longrightarrow$ 密码子

second base in codon

first base in codon	U	C	A	G	third base in codon
U	UUU Phe	UCU Ser	UAU Tyr	UGU Cys	U
	UUC Phe	UCC Ser	UAC Tyr	UGC Cys	C
	UUA Leu	UCA Ser	UAA stop	UGA stop	A
	UUG Leu	UCG Ser	UAG stop	UGG Trp	G
C	CUU Leu	CCU Pro	CAU His	CGU Arg	U
	CUC Leu	CCC Pro	CAC His	CGC Arg	C
	CUA Leu	CCA Pro	CAA Gln	CGA Arg	A
	CUG Leu	CCG Pro	CAG Gln	CGG Arg	G
A	AUU Ile	ACU Thr	AAU Asn	AGU Ser	U
	AUC Ile	ACC Thr	AAC Asn	AGC Ser	C
	AUA Ile	ACA Thr	AAA Lys	AGA Arg	A
	AUG Met	ACG Thr	AAG Lys	AGG Arg	G
G	GUU Val	GCU Ala	GAU Asp	GGU Gly	U
	GUC Val	GCC Ala	GAC Asp	GGC Gly	C
	GUA Val	GCA Ala	GAA Glu	GGA Gly	A
	GUG Val	GCG Ala	GAG Glu	GGG Gly	G

DNA 序列 | N 联核苷酸 | 密码子

- 密码子（codon）：编码多肽链中某氨基酸（共 20 种）的三联核苷酸（共 64 种）
- 密码子的简并（degeneracy）：每种氨基酸（M、W 除外）都对应 2 种以上的密码子，最多有 6 种
- 密码子使用偏好性（codon usage bias）：不同物种、不同个体、不同基因中，同义密码子用法（如出现频率等）存在差异
- 蛋白三级结构、功能与密码子用法有关
- 对于同一类型的基因，由物种引起的同义密码子使用偏好性的差异较小
- 密码子使用偏好性的分析：Codon Adaptation Index, CAI

$$\mathrm{CAI} = \exp\left(1/L \sum_{l=1}^{L} \log\left(w_i(l)\right)\right)$$

$$w_i = \frac{f_i}{\max(f_j)} \qquad i, j \in [\text{synonymous codons for amino acid}]$$

DNA 序列 | 分析工具

- SeqTools
- DNA Sequence Reverse and Complement Online Tool
- DNA/RNA GC Content Calculator
- Oligo Calculator
- BioWord (A Microsoft Word add-in for biological sequence manipulation)
- SMS2 (Sequence Manipulation Suite)
- EMBOSS (The European Molecular Biology Open Software Suite)
- ...

DNA 序列 | 透过表象看本质

- 计数：计算字符出现的次数
- 反转：反转字符串
- 互补：字符替换
- 计算：简单的四则运算

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 功能位点分析
6 启动子分析
7 CpG 岛识别
8 EMBOSS
9 序列分析中的算法
10 总结与答疑
11 引言
12 重复序列分析
13 基因识别
14 查找数据库与分析工具
15 总结与答疑
16 引言
17 mRNA 选择性剪接
18 miRNA 及其靶基因预测
19 lncRNA
20 学习数据库与分析工具的使用
21 总结与答疑
22 复习思考题

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

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	*Escherichia*	genus
co	*coli*	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

《题金山寺》，北宋 • 苏轼

潮随暗浪雪山倾，远浦渔舟钓月明。
桥对寺门松径小，槛当泉眼石波清。
迢迢绿树江天晓，霭霭红霞晚日晴。
遥望四边云接水，雪峰千点数鸥轻。

轻鸥数点千峰雪，水接云边四望遥。
晴日晚霞红霭霭，晓天江树绿迢迢。
清波石眼泉当槛，小径松门寺对桥。
明月钓舟渔浦远，倾山雪浪暗随潮。

回文

- 上海自来水来自海上
- 山东落花生花落东山
- 画上荷花和尚画

《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

- 上海自来水来自海上
- 山东落花生花落东山
- 画上荷花和尚画

《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

- 上海自来水来自海上
- 山东落花生花落东山
- 画上荷花和尚画

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 | 末端

(a) Blunt ends (RsaI)
Sugar-phosphate backbone
RsaI
5'
3'
A A T G T A C C G G
T T A C A T G G C C
(b) Sticky 5' ends (EcoRI)
EcoRI
C G A A T T C A T
G C T T A A G T A
5' overhangs
(c) Sticky 3' ends (KpnI)
KpnI
3' overhangs
C A G G T A C C T T
G T C C A T G G A A

限制酶 | II 型 | 末端 | 黏性末端

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

限制酶 | II 型 | 末端 | 平滑末端

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

限制酶 | 数据库与分析工具

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

检索

- 搜索引擎：Google，Bing，Yahoo! Search，DuckDuckGo，搜狗，360 搜索，百度，……
- 检索技巧：关键词（英文 vs. 中文）；表达式（与、或、非）；……
- 翻墙方法：Lantern（蓝灯），shadowsocks，VPN，……

限制酶 | 数据库与分析工具

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

检索

- 搜索引擎：Google，Bing，Yahoo! Search，DuckDuckGo，搜狗，360 搜索，百度，……
- 检索技巧：关键词（英文 vs. 中文）；表达式（与、或、非）；……
- 翻墙方法：Lantern（蓝灯），shadowsocks，VPN，……

字符串搜索

已知限制酶识别位点的前提下，在 DNA 序列这个长的字符串中搜索识别位点对应的子序列这个短字符串。

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 功能位点分析
6 启动子分析
7 CpG 岛识别
8 EMBOSS
9 序列分析中的算法
10 总结与答疑
11 引言
12 重复序列分析
13 基因识别
14 查找数据库与分析工具
15 总结与答疑
16 引言
17 mRNA 选择性剪接
18 miRNA 及其靶基因预测
19 lncRNA
20 学习数据库与分析工具的使用
21 总结与答疑
22 复习思考题

开放阅读框

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

开放阅读框 | 分析工具

- 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

字符串搜索

根据对应物种的密码子表，在给定的 DNA 序列中找到起始密码子，依次向后寻找终止密码子，并计算两者之间的距离，保留满足长度要求的或者最长的，即是最终预测的 ORF。

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 功能位点分析
6 启动子分析
7 CpG 岛识别
8 EMBOSS
9 序列分析中的算法
10 总结与答疑
11 引言
12 重复序列分析
13 基因识别
14 查找数据库与分析工具
15 总结与答疑
16 引言
17 mRNA 选择性剪接
18 miRNA 及其靶基因预测
19 lncRNA
20 学习数据库与分析工具的使用
21 总结与答疑
22 复习思考题

功能位点 | 简介

功能位点（functional site）

DNA 序列中，除基因外，还包含其它信息，存放这些信息的 DNA 片段称为功能位点。它们与功能相关，是功能单元。又称功能序列（functional sequence）、序列模式/模体/基元/基序（motif）、信号（signal）等。如，启动子 (promoter)、基因终止序列 (terminator sequence）、剪切位点（splice site）等。

Motif model

TTGACA
TCGACA
TTGACA
TTGAAA
ATGACA
TTGACA
GTGACA
TTGACT
TTGACC
TTGACA

Consensus Pattern →

TTGACA

Positional Weight Matrix (PWM) →

	alignment position					
nucleotide	1	2	3	4	5	6
A	0.1	0	0	**1**	0.1	**0.8**
C	0	0.1	0	0	**0.9**	0.1
G	0.1	0	**1**	0	0	0
T	**0.8**	**0.9**	0	0	0	0.1

Motif can be described in two ways based on the binding sites discovered

功能位点 | 共有序列

共有序列（consensus sequence）

又称一致性片段，描述了功能位点每个位置上进化的保守性。例如：NTATN。

共有序列的局限

- 关于序列特征的一种定性描述
- 能说明每个位置可能出现的碱基类型，但不能准确说明各碱基出现的可能性

共有序列与功能位点

1. 构造共有序列。
2. 利用共有序列在给定的核酸序列上搜寻功能位点，并计算所找到的功能位点的可靠性。

功能位点 | 共有序列

共有序列（consensus sequence）

又称一致性片段，描述了功能位点每个位置上进化的保守性。例如：NTATN。

共有序列的局限

- 关于序列特征的一种定性描述
- 能说明每个位置可能出现的碱基类型，但不能准确说明各碱基出现的可能性

共有序列与功能位点

1. 构造共有序列。
2. 利用共有序列在给定的核酸序列上搜寻功能位点，并计算所找到的功能位点的可靠性。

功能位点 | 共有序列

共有序列（consensus sequence）

又称一致性片段，描述了功能位点每个位置上进化的保守性。例如：NTATN。

共有序列的局限

- 关于序列特征的一种定性描述
- 能说明每个位置可能出现的碱基类型，但不能准确说明各碱基出现的可能性

共有序列与功能位点

1. 构造共有序列。
2. 利用共有序列在给定的核酸序列上搜寻功能位点，并计算所找到的功能位点的可靠性。

功能位点 | 共有序列 | 构造

1. 初始化共有序列为一系列可变位置，以“N”代表。
2. 在可变位置寻找出现次数最多的核苷酸，并将该位置转化为保守位置。
3. 对当前所得到的共有序列进行特异性检查，若通过检查，转 (5)，否则转 (4)。
4. 形成与当前共有序列一致的位点子集，转 (2)。
5. 从原位点集合中删除与当前共有序列一致的位点，若还有剩余位点，则转 (1)，构造另外的共有序列。

功能位点 | 共有序列 | 构造

[1]
TTATG
ATATA
TACGC
TTGTC
TCCAC
NNNNN
[2]
TTATG
ATATA
TACGC
TTGTC
TCCAC
TNNNN
[3]
TNNNN
非特异
剩余位点
TTATG
ATATA
[4]
tTATG
tACGC
tTGTC
tCCAC
[2]
tTATG
tACGC
tTGTC
tCCAC
TNNNC
[3]
TNNNC
非特异
[4]
tACGc
tTGTc
tCCAc
[2]
tACGc
tTGTc
tCCAc
TNSNC
[3]
TNSNC
特异
[5]
Consensus1
TNSNC
[5]
Consensus2
NTATN

加权矩阵

用权系数（weight coefficient）描述功能位点各位置上每种核苷酸的相对重要性。加权矩阵的大小为 $4 \times n$（碱基种类数目 $\times$ 功能位点长度）。矩阵的每一个元素 M(a,n) 的值代表第 a 种核苷酸在功能位点第 n 个位置上出现的得分。其中，$a \in \{A,T,G,C\}$。

	1	2	3	4	5	6	...
A	1	8	22	7	-3	19	
T	2	6	14	2	-1	0	
G	3	11	0	-5	0	-19	
C	5	-9	16	8	8	0	

对于 ATTGCA 来说，得分 $W = 1 + 6 + 14 - 5 + 8 + 19 = 43$。

功能位点 | PWM

PWM

A position weight matrix (PWM), also known as a position-specific weight matrix (PSWM) or position-specific scoring matrix (PSSM), is a commonly used representation of motifs (patterns) in biological sequences.

PWMs are often derived from a set of aligned sequences that are thought to be functionally related and have become an important part of many software tools for computational motif discovery.

功能位点 | PWM

Both strands used to compute Sig Value.

PWM:

a	0.0	0.0	0.0	9.0	3.0	3.0	5.0	2.0	0.0	0.0	0.0	2.0	3.0	2.0	0.0	0.0	0.0
t	0.0	0.0	0.0	0.0	1.0	1.0	2.0	2.0	13.0	2.0	10.0	1.0	1.0	6.0	0.0	0.0	0.0
g	0.0	13.0	13.0	0.0	8.0	7.0	4.0	0.0	0.0	3.0	0.0	3.0	5.0	1.0	0.0	0.0	13.0
c	13.0	0.0	0.0	4.0	1.0	2.0	2.0	9.0	0.0	8.0	3.0	7.0	4.0	4.0	13.0	13.0	0.0

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 功能位点分析
6 启动子分析
7 CpG 岛识别
8 EMBOSS
9 序列分析中的算法
10 总结与答疑
11 引言
12 重复序列分析
13 基因识别
14 查找数据库与分析工具
15 总结与答疑
16 引言
17 mRNA 选择性剪接
18 miRNA 及其靶基因预测
19 lncRNA
20 学习数据库与分析工具的使用
21 总结与答疑
22 复习思考题

启动子 | 转录调控

- 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 | 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

Prokaryotic promoter
Consensus sequences for RNA polymerase binding
Transcription start site
−35
−10
+1
TTGACA
TATAAT
−35 region
Pribnow box
Structural genes
DNA template (coding) strand
RNA polymerase
Eukaryotic promoter
−75
−25
+1
GGXCAATCT
TATAAA
CAAT box
TATA box
DNA template (coding) strand (x=any base)

启动子 | TF&TFBS

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 | TF&TFBS

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 | 数据库与分析工具

- 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)

字符串搜索

- 在特定范围内进行搜索
- 子字符串由权重不等的一组字符串构成

"AAGTTAATGGATCTG"

$$\begin{pmatrix} 14 & 16 & 4 & 0 & 1 & 19 & 20 & 1 & 4 & 13 & 4 & 4 & 13 & 12 & 3 \\ 3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 7 & 3 & 1 & 0 & 3 & 1 & 12 \\ 4 & 3 & 17 & 0 & 0 & 2 & 0 & 0 & 9 & 1 & 3 & 0 & 5 & 2 & 2 \\ 0 & 2 & 0 & 21 & 20 & 0 & 1 & 20 & 1 & 4 & 13 & 17 & 0 & 6 & 4 \end{pmatrix}$$

字符串搜索

- 在特定范围内进行搜索
- 子字符串由权重不等的一组字符串构成

"AAGTTAATGGATCTG"

$$\begin{pmatrix} 14 & 16 & 4 & 0 & 1 & 19 & 20 & 1 & 4 & 13 & 4 & 4 & 13 & 12 & 3 \\ 3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 7 & 3 & 1 & 0 & 3 & 1 & 12 \\ 4 & 3 & 17 & 0 & 0 & 2 & 0 & 0 & 9 & 1 & 3 & 0 & 5 & 2 & 2 \\ 0 & 2 & 0 & 21 & 20 & 0 & 1 & 20 & 1 & 4 & 13 & 17 & 0 & 6 & 4 \end{pmatrix}$$

启动子 | 透过表象看本质 | Promoter 2.0

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 功能位点分析
6 启动子分析
7 CpG 岛识别
8 EMBOSS
9 序列分析中的算法
10 总结与答疑
11 引言
12 重复序列分析
13 基因识别
14 查找数据库与分析工具
15 总结与答疑
16 引言
17 mRNA 选择性剪接
18 miRNA 及其靶基因预测
19 lncRNA
20 学习数据库与分析工具的使用
21 总结与答疑
22 复习思考题

CpG 岛 | 特征

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 | 特征

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%
   - $\frac{Num\ of\ CpG}{Num\ of\ C \times Num\ of\ G} \times Total\ number\ of\ nucleotides\ in\ the\ sequence$
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%
   - $\frac{Num\ of\ CpG}{Num\ of\ C \times Num\ of\ G} \times Total\ number\ of\ nucleotides\ in\ the\ sequence$
4. 500bp，55%，65%

CpG 岛 | 分析工具

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

CpG 岛 | 透过表象看本质

- 分段：使用滑动窗口将长序列分段
- 计数：长度、G 和 C
- 计算：含量、比率
- 比较：和标准进行比较

```
"Sliding window" example

window 1 : score = 0/9, perfect match
ATCTTCAGCCAAAGATGAAGTT
ATCTTCAGC

window 2 : score = 7/9 , terrible match
ATCTTCAGCCAAAGATGAAGTT
 ATCTTCAGC
```

CpG 岛 | 透过表象看本质

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 功能位点分析
6 启动子分析
7 CpG 岛识别
8 EMBOSS
9 序列分析中的算法
10 总结与答疑
11 引言
12 重复序列分析
13 基因识别
14 查找数据库与分析工具
15 总结与答疑
16 引言
17 mRNA 选择性剪接
18 miRNA 及其靶基因预测
19 lncRNA
20 学习数据库与分析工具的使用
21 总结与答疑
22 复习思考题

EMBOSS | 简介

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）

EMBOSS | 主要程序

- 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。

EMBOSS | 演示

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 功能位点分析
6 启动子分析
7 CpG 岛识别
8 EMBOSS
9 序列分析中的算法
10 总结与答疑
11 引言
12 重复序列分析
13 基因识别
14 查找数据库与分析工具
15 总结与答疑
16 引言
17 mRNA 选择性剪接
18 miRNA 及其靶基因预测
19 lncRNA
20 学习数据库与分析工具的使用
21 总结与答疑
22 复习思考题

序列分析中的算法 | 滑动窗口（Sliding Window）

参数

- window size：窗口大小
- step：步长

序列分析中的算法 | 滑动窗口 | 窗口大小

动态规划

动态规划（Dynamic programming，简称 DP）是一种在数学、管理科学、计算机科学、经济学和生物信息学中使用的，通过把原问题分解为相对简单的子问题的方式求解复杂问题的方法。动态规划常常适用于有重叠子问题和最优子结构性质的问题。

动态规划背后的基本思想非常简单。大致上，若要解一个给定问题，我们需要解其不同部分（即子问题），再合并子问题的解以得出原问题的解。通常许多子问题非常相似，为此动态规划法试图仅仅解决每个子问题一次，从而减少计算量：一旦某个给定子问题的解已经算出，则将其记忆化存储，以便下次需要同一个子问题解之时直接查表。这种做法在重复子问题的数目关于输入的规模呈指数增长时特别有用。

Dynamic programming and backtracking

source

target

		-		A		T		T
-		0	-1	-1	-2	-2	-3	-3
T		-1	-1	-2	0	-3	-1	-4
		-1	-2	-1	-2	0	-1	-1
T		-2	-2	-2	0	-1	+1	-2
		-2	-3	-2	-3	0	-1	+1
C		-3	-3	-3	-3	-1	-1	0
		-3	-4	-3	-4	-1	-2	0

	-	A	T	T
-	0	-1	-2	-3
T	-1	-1	0	-1
T	-2	-2	0	+1
C	-3	-3	-1	0

$S_{i-1,j-1}$ $+s(a_i,b_j)$	$S_{i,j-1}$ $+s(-,b_j)$
$S_{i-1,j}$ $+s(a_i,-)$	$S_{i,j}$

scoring scheme :

- $s(a_i,b_i) = +1$ if $a_i = b_i$
- $s(a_i,b_i) = -1$ if $a_i = b_i$
- $s(a_i,-) = -1$
- $s(-,b_i) = -1$

A T T –

– T T C

机器学习

机器学习理论主要是设计和分析一些让计算机可以自动“学习”的算法。机器学习算法是一类从数据中自动分析获得规律，并利用规律对未知数据进行预测的算法。

分类

监督学习，无监督学习，半监督学习，增强学习

算法

人工神经网络，决策树，线性判别分析，最近邻居法，支持向量机，……

序列分析中的算法 | 机器学习

机器学习

机器学习理论主要是设计和分析一些让计算机可以自动“学习”的算法。机器学习算法是一类从数据中自动分析获得规律，并利用规律对未知数据进行预测的算法。

分类

监督学习，无监督学习，半监督学习，增强学习

算法

人工神经网络，决策树，线性判别分析，最近邻居法，支持向量机，……

机器学习

机器学习理论主要是设计和分析一些让计算机可以自动“学习”的算法。机器学习算法是一类从数据中自动分析获得规律，并利用规律对未知数据进行预测的算法。

分类

监督学习，无监督学习，半监督学习，增强学习

算法

人工神经网络，决策树，线性判别分析，最近邻居法，支持向量机，……

序列分析中的算法 | 机器学习

序列分析中的算法 | 机器学习

machine learning

The classical machine learning workflow can be broken down into four steps: data pre-processing, feature extraction, model learning and model evaluation.

Supervised machine learning methods relate input features x to an output label y, whereas unsupervised method learns factors about x without observed labels.

Raw input data are often high-dimensional and related to the corresponding label in a complicated way, which is challenging for many classical machine learning algorithms. Alternatively, higher-level features extracted using a deep model may be able to better discriminate between classes.

Deep networks use a hierarchical structure to learn increasingly abstract feature representations from the raw data.

聚类分析（cluster analysis）

聚类是把相似的对象通过静态分类的方法分成不同的组别或者更多的子集（subset），这样让在同一个子集中的成员对象都有相似的一些属性。

判别分析

线性判别分析（Linear Discriminant Analysis），简称判别分析，是统计学上的一种分析方法，用于在已知的分类之下遇到有新的样本时，选定一个判别标准，以判定如何将新样本放置于哪一个类别之中。

人工神经网络（artificial neural network，ANN）

简称神经网络（neural network，NN），是一种模仿生物神经网络的结构和功能的数学模型或计算模型。神经网络由大量的人工神经元联结进行计算。大多数情况下人工神经网络能在外界信息的基础上改变内部结构，是一种自适应系统。现代神经网络是一种非线性统计性数据建模工具，常用来对输入和输出间复杂的关系进行建模，或用来探索数据的模式。

深度学习

深度学习（deep learning）是机器学习拉出的分支，它试图使用包含复杂结构或由多重非线性变换构成的多个处理层对数据进行高层抽象的算法。

深度学习是机器学习中一种基于对数据进行表征学习的方法。深度学习的好处是用非监督式或半监督式（Semi-supervised learning）的特征学习和分层特征提取高效算法来替代手工获取特征（Feature）。

至今已有数种深度学习框架，如深度神经网络、卷积神经网络和深度置信网络（Deep belief network）和递归神经网络已被应用于计算机视觉、语音识别、自然语言处理、音频识别与生物信息学等领域并获取了极好的效果。

序列分析中的算法 | 机器学习 | 深度学习

Deep neural networks learn hierarchical feature representations

马尔可夫性质

当一个随机过程在给定现在状态及所有过去状态情况下，其未来状态的条件概率分布仅依赖于当前状态；换句话说，在给定现在状态时，它与过去状态（即该过程的历史路径）是条件独立的，那么此随机过程即具有马尔可夫性质。

马尔可夫过程（Markov process）

一个具备了马尔可夫性质的随机过程。马尔可夫过程是不具备记忆特质的。换言之，马可夫过程的条件概率仅仅与系统的当前状态相关，而与它的过去历史或未来状态，都是独立、不相关的。

马尔可夫性质

当一个随机过程在给定现在状态及所有过去状态情况下，其未来状态的条件概率分布仅依赖于当前状态；换句话说，在给定现在状态时，它与过去状态（即该过程的历史路径）是条件独立的，那么此随机过程即具有马尔可夫性质。

马尔可夫过程（Markov process）

一个具备了马尔可夫性质的随机过程。马尔可夫过程是不具备记忆特质的。换言之，马可夫过程的条件概率仅仅与系统的当前状态相关，而与它的过去历史或未来状态，都是独立、不相关的。

马尔可夫链（Markov chain）

具备离散状态的马尔可夫过程，通常被称为马尔可夫链。又称离散时间马尔可夫链（discrete-time Markov chain，DTMC），是马尔可夫过程中的一个特例，为具备马尔可夫性质与离散时间状态的随机过程。该过程中，在给定当前知识或信息的情况下，只有当前的状态用来预测将来，过去（即当前以前的历史状态）对于预测将来（即当前以后的未来状态）是无关的。

隐马尔可夫模型（Hidden Markov Model，HMM）

统计模型，用来描述一个含有隐含未知参数的马尔可夫过程。其难点是从可观察的参数中确定该过程的隐含参数。然后利用这些参数来作进一步的分析，例如模式识别。

马尔可夫链（Markov chain）

具备离散状态的马尔可夫过程，通常被称为马尔可夫链。又称离散时间马尔可夫链（discrete-time Markov chain，DTMC），是马尔可夫过程中的一个特例，为具备马尔可夫性质与离散时间状态的随机过程。该过程中，在给定当前知识或信息的情况下，只有当前的状态用来预测将来，过去（即当前以前的历史状态）对于预测将来（即当前以后的未来状态）是无关的。

隐马尔可夫模型（Hidden Markov Model，HMM）

统计模型，用来描述一个含有隐含未知参数的马尔可夫过程。其难点是从可观察的参数中确定该过程的隐含参数。然后利用这些参数来作进一步的分析，例如模式识别。

一个具有两个转换状态的马尔可夫链.

隐马尔可夫模型状态变迁图（例子）
x — 隐含状态
y — 可观察的输出
a — 转换概率（transition probabilities）
b — 输出概率（output probabilities）

A General Definition of HMM

$$HMM = (S, V, B, A, \Pi)$$

N states

$$S = \{s_1, \ldots, s_N\}$$

M symbols

$$V = \{v_1, \ldots, v_M\}$$

Initial state probability:

$$\Pi = \{\pi_1, \ldots, \pi_N\} \quad \sum_{i=1}^{N} \pi_i = 1$$

$$\pi_i : prob\ of\ starting\ at\ state\ s_i$$

State transition probability:

$$A = \{a_{ij}\} \quad 1 \le i, j \le N \quad \sum_{j=1}^{N} a_{ij} = 1$$

$$a_{ij} : prob\ of\ going\ s_i \rightarrow s_j$$

Output probability:

$$B = \{b_i(v_k)\} \quad 1 \le i \le N, 1 \le k \le M \quad \sum_{k=1}^{M} b_i(v_k) = 1$$

$$b_i(v_k) : prob\ of\ "generating" v_k\ at\ s_i$$

三个典型问题

1. 已知模型参数，计算某一特定输出序列的概率。通常使用 forward 算法解决。
2. 已知模型参数，寻找最可能的能产生某一特定输出序列的隐含状态的序列。通常使用 Viterbi 算法解决。
3. 已知输出序列，寻找最可能的状态转移以及输出概率。通常使用 Baum-Welch 算法以及 Reversed Viterbi 算法解决。

最近的一些方法使用 Junction tree 算法来解决这三个问题。

序列分析中的算法 | 机器学习 | 隐马尔可夫模型

A = 0.25
C = 0.25
G = 0.25
T = 0.25
A = 0.05
C = 0
G = 0.95
T = 0
A = 0.4
C = 0.1
G = 0.1
T = 0.4
Start
1.0
E
0.1
5
1.0
I
0.1
End
0.9
0.9
Sequence: CTTCATGTGAAAGCAGACGTAAGTCA
State path: EEEEEEEEEEEEEEEEEEEEE5IIIIIIII
log P
−41.22
−43.90
Parsing:
−43.45
−43.94
−42.58
−41.71
46%
28%
Posterior decoding:
11%

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 功能位点分析
6 启动子分析
7 CpG 岛识别
8 EMBOSS
9 序列分析中的算法
10 总结与答疑
11 引言
12 重复序列分析
13 基因识别
14 查找数据库与分析工具
15 总结与答疑
16 引言
17 mRNA 选择性剪接
18 miRNA 及其靶基因预测
19 lncRNA
20 学习数据库与分析工具的使用
21 总结与答疑
22 复习思考题

总结与答疑

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 功能位点分析
6 启动子分析
7 CpG 岛识别
8 EMBOSS
9 序列分析中的算法
10 总结与答疑
11 引言
12 重复序列分析
13 基因识别
14 查找数据库与分析工具
15 总结与答疑
16 引言
17 mRNA 选择性剪接
18 miRNA 及其靶基因预测
19 lncRNA
20 学习数据库与分析工具的使用
21 总结与答疑
22 复习思考题

引言 | 回顾

- 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

引言 | 回顾

- 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

引言 | 回顾

- 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

引言 | 回顾

- 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
  - 开放阅读框识别
  - 密码子偏好性分析
  - 转录因子结合位点分析
  - 转录模块分析
- 基因识别

引言 | 回顾

- 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
  - 序列模式识别
- 基因识别

引言 | 回顾

- 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

引言 | 回顾

- 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

引言 | 回顾

- 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

引言 | 回顾

- 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

引言 | 回顾

- 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

引言 | 回顾

- 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

引言 | 回顾

- 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 功能位点分析
6 启动子分析
7 CpG 岛识别
8 EMBOSS
9 序列分析中的算法
10 总结与答疑
11 引言
12 重复序列分析
13 基因识别
14 查找数据库与分析工具
15 总结与答疑
16 引言
17 mRNA 选择性剪接
18 miRNA 及其靶基因预测
19 lncRNA
20 学习数据库与分析工具的使用
21 总结与答疑
22 复习思考题

重复序列 | 基因组构成 | 基因数目

	Organism	# of protein-coding genes	# of genes naïve estimate: (genome size /1000)	BNID
viruses	HIV 1	9	10	105769
	Influenza A virus	10-11	14	105767
	Bacteriophage λ	66	49	105770
	Epstein Barr virus	80	170	103246
prokaryotes	*Buchnera sp.*	610	640	105757
	T. maritima	1,900	1,900	105766
	S. aureus	2,800	2,900	105763
	V. cholerae	3,900	4,000	105760
	B. subtilis	4,200	4,200	105753
	E. coli	4,200	4,600	105443
eukaryotes	*S. cerevisiae*	5,600	12,000	105444
	C. elegans	20,000	100,000	101364
	A. thaliana	27,000	120,000	100473
	D. melanogaster	14,000	180,000	100200
	F. rubripes	23,000	400,000	100280
	Z. mays	33,000	2,000,000	110565
	M. musculus	20,000	3,000,000	100310
	H. sapiens	19,000	3,000,000	105447
	T. aestivum (hexaploid)	90,000	20,000,000	105448

重复序列 | 基因组构成 | 原核

重复序列 | 基因组构成 | 真核

重复序列 | 分类

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 | 分类

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 | 分类

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

- 散在的重复性 DNA（转座子导致的重复）
  - 长末端重复（LTR）转座子
  - 长散布元件（LINEs）
  - 短散布元件（SINEs）
  - DNA 转座子
- 被修饰的假基因
- 简单重复序列
  - 微卫星序列
  - 小卫星序列
- 片段复制
- 串联重复序列块

重复序列 | 分类

Sequence types	Repeat size(bp)	Array size (kb)	Copy number[1]	Functions, features of family members
Satellites — large tandem arrays		**10–25% of total DNA**		
Microsatellite	2–5	0.2–0.5	3 x 10^3	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		**35-40% of total DNA**		
Retrotransposons				
LTR-containing elements				
copia[2], *gypsy*[2]	~5 kb	NA	20–60	Can befound as free circular DNA Horizontal transfer of genes; can infect germline cells
Yeast Ty	6.3kb	NA	40	Ty1 and Ty3 transpose specifically to genes transcribed by RNA polymerase III; Repair of chromosomal breaks
Poly-A elements				
LINE1 (L1)	1–7 kb	NA	~10^5	Mutant sequences can promotecancer Some provide polyadenylation signals Some copies mobile
HeT-A, TART[2]	6–10 kb	5–10	~10^4	Maintenance of telomeres
SINEs				
Alu	300	NA	~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-ru
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 %
```

字符串搜索

在 DNA 序列这个长的字符串中搜索每个重复序列这个子字符串。

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 功能位点分析
6 启动子分析
7 CpG 岛识别
8 EMBOSS
9 序列分析中的算法
10 总结与答疑
11 引言
12 重复序列分析
13 基因识别
14 查找数据库与分析工具
15 总结与答疑
16 引言
17 mRNA 选择性剪接
18 miRNA 及其靶基因预测
19 lncRNA
20 学习数据库与分析工具的使用
21 总结与答疑
22 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

启动区
5' UTR
编码区
3' UTR
终止区
起始密码子
终止密码子
TATAAT盒
保守序列：TTGACA

增强子
启动区
5' UTR
exon
GT
AG
exon
GT
AG
exon
3' UTR
终止区
intron
intron
起始密码子
终止密码子
GC盒
TATA盒
CAAT盒

基因识别 | 方法

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 | 信号

Start codon
codons
Donor site
CGCCATGCCCTTCTCCAACAGGTGAGTGAG
Transcription start
Exon
Promoter
5' UTR
CCTCCCAGCCCTGCCCAG
Acceptor site
Intron
Poly-A site
Stop codon
GGCAGAAACAATAAAACCAC
GATCCCCATGCCTGAGGGCCCCTC
3' UTR

CODON USAGE IN *E. COLI* GENES[1]

	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

基因识别 | 原核基因

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

基因识别 | 原核基因

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

基因识别 | 真核基因

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

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

Homo sapiens
ANNOTATION DATABASE
ab initio
homology
GENES
Distribution of 34 Blast Hits on the Query Sequence
Mouse-over to show defline and scores. Click to show alignments
Color Key for Alignment Scores
<40
40-50
50-80
80-200
>=200
1_9401
0
50
100
150
200
250
300
350
400
450

基因识别 | 策略

Informant genomes
Query genome
cDNA sequences
Protein sequence
Comparison and allignment of query genome against informant genomes
(1)
Detection of signals involved in gene identification: splice sites, start sites and so on
(2)
Detection of sequence bias related to coding function
(3)
(4)
Mapping of known cDNA and protein sequence into the query genome
Dual or multiple comparative genome finders
(5)
'Ab initio' genome finders
(6)
(7)
Splice aligners
(8)
Combiners
(10)
Integrative methods
(9)
Experimental verification
(12)
Manual annotation
(11)

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.cbcb.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 *ab initio* 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 *Arabidopsis thaliana*	*Arabidopsis thaliana*
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	
geneid	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
GrailEXP	predicts exons, genes, promoters, polyas, CpG islands, EST similarities, and	

基因识别 | 工具列表

- 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

评价预测的准确性是用 cDNA 定位或已知基因结构作为基准的。

- 假阳性（False Positive，FP）：在非编码区预测出编码区
- 假阴性（False Negative，FN）：将编码区预测为非编码区
- 过界预测（Over Prediction，OP）：预测超出实际的边界（边界难准确定位）
- 片段化（Fragmentation）：内含子过大的基因，断裂成两个或多个基因
- 融合化（Fudion）：距离过近的基因，融合成一个大基因
- 只能预测出一种剪接形式，无法识别可变剪接
- 只能预测起始和终止密码子间的区域，不能预测 UTR 区域
- 大量的重复序列会对预测造成严重的影响

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

基因识别 | GENSCAN | 结果 | 解读

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，甚至可以直接将其忽略

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，甚至可以直接将其忽略

基因识别 | 透过表象看本质 | GeneMark

基因识别 | 透过表象看本质 | Glimmer

基因识别 | 透过表象看本质 | GENSCAN

基因识别 | 透过表象看本质 | GRAIL

基因识别 | 透过表象看本质 | 神经网络

基因识别 | 透过表象看本质 | 决策树（Decision trees）

d: donnor site score
a: acceptor site score
hex: in–frame hexamer frequency
asym: Fickett's position asymmetry statistic
donor: donor site score
leaf nodes: exon, pseudo–exon distribution in the training set

Integrated gene finding: Dynamic programming

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 功能位点分析
6 启动子分析
7 CpG 岛识别
8 EMBOSS
9 序列分析中的算法
10 总结与答疑
11 引言
12 重复序列分析
13 基因识别
14 查找数据库与分析工具
15 总结与答疑
16 引言
17 mRNA 选择性剪接
18 miRNA 及其靶基因预测
19 lncRNA
20 学习数据库与分析工具的使用
21 总结与答疑
22 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
- 图书馆

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
- 图书馆

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 功能位点分析
6 启动子分析
7 CpG 岛识别
8 EMBOSS
9 序列分析中的算法
10 总结与答疑
11 引言
12 重复序列分析
13 基因识别
14 查找数据库与分析工具
15 总结与答疑
16 引言
17 mRNA 选择性剪接
18 miRNA 及其靶基因预测
19 lncRNA
20 学习数据库与分析工具的使用
21 总结与答疑
22 复习思考题

总结与答疑

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 功能位点分析
6 启动子分析
7 CpG 岛识别
8 EMBOSS
9 序列分析中的算法
10 总结与答疑
11 引言
12 重复序列分析
13 基因识别
14 查找数据库与分析工具
15 总结与答疑
16 引言
17 mRNA 选择性剪接
18 miRNA 及其靶基因预测
19 lncRNA
20 学习数据库与分析工具的使用
21 总结与答疑
22 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

引言 | 回顾

- DNA 序列分析
  - 基本信息
  - 序列特征
  - 基因识别
- RNA 序列分析

引言 | 回顾

- DNA 序列分析
  - 基本信息
  - 序列特征
  - 基因识别
- RNA 序列分析
  - RNA 特征
  - 编码 RNA
  - 非编码 RNA

引言 | 回顾

- 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

引言 | 回顾

- 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

引言 | 回顾

- 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

引言 | 回顾

- 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

引言 | 回顾

- 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

引言 | RNA

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*	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

*All 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	Small ncRNA	21~25	Eukaryote	Translational regulation
siRNA	Small ncRNA	21~25	Eukaryote	Protection against viral infection
piRNA	Small ncRNA	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 功能位点分析
6 启动子分析
7 CpG 岛识别
8 EMBOSS
9 序列分析中的算法
10 总结与答疑
11 引言
12 重复序列分析
13 基因识别
14 查找数据库与分析工具
15 总结与答疑
16 引言
17 mRNA 选择性剪接
18 miRNA 及其靶基因预测
19 lncRNA
20 学习数据库与分析工具的使用
21 总结与答疑
22 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 | 剪接

DNA
Exon 1
Exon 2
Exon 3
Exon 4
Exon 5
RNA
Exon 1
Exon 2
Exon 3
Exon 4
Exon 5
Alternative Splicing
mRNA
1 2 3 4 5
1 2 4 5
1 2 3 5
Translation
Translation
Translation
Protein A
Protein B
Protein C

Within the intron, a donor site (5' end of the intron), a branch site (near the 3' end of the intron) and an acceptor site (3' end of the intron) are required for splicing.

A-G-[cut]-G-U-R-A-G-U (donor site) …intron sequence …Y-U-R-A-C (branch sequence 20-50 nucleotides upstream of acceptor site) … Y-rich-N-C-A-G-[cut]-G (acceptor site)

pre-mRNA
Exon
Intron
Exon
GU
A
AG
(1)
(2)
A
AG
GU
spliced
mRNA
+
A
AG
GU

选择性剪接 | 剪接 | 机制 | 过程 | 详观

选择性剪接 | 模式图

coding
non-coding
splice variants lead to protein diversity
protein A
protein B
protein C

选择性剪接 | 模式图

DNA
Exon 1
Exon 2
Exon 3
Exon 4
Exon 5
RNA
Exon 1
Exon 2
Exon 3
Exon 4
Exon 5
Alternative Splicing
mRNA
1 2 3 4 5
1 2 4 5
1 2 3 5
Translation
Translation
Translation
Protein A
Protein B
Protein C

选择性剪接 | 模式图

GENE
Genomic DNA
Transcription
exon
intron
exon
intron
exon
intron
exon
intron
exon
M7G
1
2
3
4
5
AAAA200
Precursor mRNA
Alternative Splicing
M7G
1
2
4
5
AAAA200
M7G
1
2
3
5
AAAA200
Mature mRNA isoform 1,
exon 3 skipped
Mature mRNA isoform 2,
Exon 4 skipped
Translation
Translation
M7G
AAAA200
M7G
AAAA200
Protein isoform 1, function A
Protein isoform 2, function B

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 | 实例

α–TM EXON GENE ORGANIZATION
5'UT 1–38
39–80
39–80
81–125
126–164
165–188
189–213
214–234
235–257
258–284
3'UT
258–284
3'UT
α–TM mRNA TRANSCRIPTS
Striated muscle
Striated muscle'
Myoblast
Smooth muscle
Nonmuscle/ fibroblast
Hepatoma
Brain

选择性剪接 | 实例

a KCNMA1 pre-mRNA
5′
3′
b Dscam pre-mRNA
Exon 4
Exon 6
Exon 9
Exon 17
5′
3′
c mod(mdg4) gene
Common exons
Variable exons
5′
3′
3′
5′
Variable exons

选择性剪接 | 实例 | *Dscam*

$$12 \times 48 \times 33 \times 2 = 38016$$

选择性剪接 | 实例 | *Dscam*

$$12 \times 48 \times 33 \times 2 = 38016$$

选择性剪接 | 调控

- 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

a
5′ SS
Branch site
PPT
3′ SS
20–50 bp
b
U2
U1
U2
U1
U5
U6
U2
U4
c
hnRNP
SR
SR
hnRNP
ESS
ESE
ISE
ISS
PTB
Fox
TIA1
Nova

选择性剪接 | 调控

B
U2 snRNP
SR
hnRNP
U1
ISE
ISS
A
3'ss
ESE
ESS
5'ss
ISS
ISE
Branch site

选择性剪接 | 调控 | 激活 vs. 抑制

选择性剪接 | 机制 | 五种

Exon skipping/inclusion
Alternative 3′ splice sites
Alternative 5′ splice sites
Mutually exclusive exons
Intron retention
Constitutive exon
Alternatively spliced exon

选择性剪接 | 机制 | 七种

Alternative promoters
Alternative poly-A sites
Poly-A
Poly-A
Alternative 5′ splice sites
Alternative 3′ splice sites
Cassette exon
Mutually exclusive exons
Retained intron

选择性剪接 | 机制 | 七种

Cassette Exon
Mutually Exclusive Exons
Intron Retention
Alternative 5' or 3' Splice Sites
P1
P2
Alternative Promoters
AAAA...
AAAA...
Alternative Splicing and Polyadenlyation

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., tropo*myosin* 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 | 复杂实例

选择性剪接 | 机制 | 使用频率

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

选择性剪接 | 透过表象看本质

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 功能位点分析
6 启动子分析
7 CpG 岛识别
8 EMBOSS
9 序列分析中的算法
10 总结与答疑
11 引言
12 重复序列分析
13 基因识别
14 查找数据库与分析工具
15 总结与答疑
16 引言
17 mRNA 选择性剪接
18 miRNA 及其靶基因预测
19 lncRNA
20 学习数据库与分析工具的使用
21 总结与答疑
22 复习思考题

miRNA | 简介

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

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
Gene
Pri-miRNA
Drosha
Pre-miRNA
Exportin 5
Nucleus
Cytoplasm
dsRNA
Dicer
miRNA:
miRNA*
duplex
siRNA
duplex
Unwind
Asymmetric RISC
assembly
Some
miRNA
RISC
RISC
Ribosome
ORF
RISC
RISC
RISC
Target
mRNA
Translational repression
mRNA cleavage

miRNA | 作用网络

a
miRNAs
miR-1
miR-2
targets
a
b
c
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

pri-microRNA and removes the hairpin loop, leaving a double stranded microRNA duplex molecule.
Argonaute proteins
Base mismatch
microRNA duplex
Passenger strand
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
Deadenylation
microRNA
(abbreviated miRNA)
about 22 nucleotides long
microRNA-protein complex
Base mismatches
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.
3' untranslated region (3' UTR)
3' Poly-A tail
Coding region
3' Poly-A tail
Seed Region (Nucleotides 2–8)
Perfect base pairing
Endonucleolytic cleavage
Blocked ribosome
Nucleotide 9
Has an A or U across from it
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.
3' UTR
Nucleotides 13–16
Good base pairing
Messenger RNA strands
(mRNA)
the formation and function of micro RNAs

miRNA | 数据来源

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

miRNA | 预测

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

miRNA | 预测（续）

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，用异常严格的序列和结构标准筛选候选片段，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

Target mRNA
3′
Poly (A)
5′
ORF
Mismatch
6mer
7mer m8
7mer A1
8mer
12-17 Watson-Crick complement
Bulge
3′
miRNA out seed region
miRNA Seed region
5′

miRNA | 靶基因预测

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

miRNA | 数据库与分析工具

- 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。

miRNA | 数据库与分析工具

- 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。

选择性剪接 | 透过表象看本质 | MiRscan

选择性剪接 | 透过表象看本质 | miRFinder

选择性剪接 | 透过表象看本质 | HMM

选择性剪接 | 透过表象看本质 | miRanda

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 功能位点分析
6 启动子分析
7 CpG 岛识别
8 EMBOSS
9 序列分析中的算法
10 总结与答疑
11 引言
12 重复序列分析
13 基因识别
14 查找数据库与分析工具
15 总结与答疑
16 引言
17 mRNA 选择性剪接
18 miRNA 及其靶基因预测
19 lncRNA
20 学习数据库与分析工具的使用
21 总结与答疑
22 复习思考题

lncRNA | 序列结构特征

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

lncRNA | 序列结构特征

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

lncRNA | 生物功能特征

- 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、转录调控及转录后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 | 生物功能特征

- 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、转录调控及转录后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 | vs. mRNA

	mRNA	lncRNA
相同点	① specific temporal and spatial expression ② Formation of secondary structure ③ Post-transcriptional processing, such as the 5 'cap, polyadenylation, splicing ④ The important role in the development and disease	
不同点	code for proteins	regulatory function
	Highly conserved between species	Poorly conserved between species
	Present in the nucleus and cytoplasm	Found mainly in the nucleus
	Total 20,000-24,000 mRNA	3-100 times of mRNA
	The expression level: Low to High	The expression level: very low to moderate

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

Chromatin remodeling
Imprinting eg Air
Scaffold for chromatin remodeling complexes eg ANRIL
DNA methylation/demethylation eg Khps1
X-chromosome inactivation eg Xist
Me
Ac
Transcriptional control
Co-activator of transcription factors eg Evf2
Inhibition of RNAPII initiation complex formation eg Alu RNA
Regulation of RNAPII elongation eg 7SK RNA
Promoter interference through RNA-DNA triplex formation eg. DHFR locus
transcription factor
RNAPII
promoter
Post-transcriptional processing
Regulation of alternative splicing eg MALAT1
Intron
mRNA localization/transport eg NF90 RNAs
Masking miRNA binding sites or acting as a miRNA sponge eg lincMD1
Inhibition of translation eg BC1
mRNA stabilization/destabilization eg BACE1-AS
ribosome
Protein metabolism
Protein stability eg HOTTIP
Protein activity eg CCND1 ncRNA
Protein localization eg NRON

lncRNA | 作用机制

1 RNA decoy
Transcription factors
lncRNA
2 microRNA sponge
microRNP
lncRNA
mRNA
3 RNP component
Proteins
lncRNA
lncRNA interacting proteins
1 2 3 4
1 2 3 4 5 6
4 3 2 1
lncRNA
Chromatin modifiers
4 Recruitment of chromatin modifiers
mRNA
lncRNA
5 Translation inhibition
6 Splicing modulation
7 Degradation

lncRNA | 作用机制

sncRNAs
10
1
RNApII
5'
lncRNA
exon 1
exon 2
exon 3
lncRNA
lncRNA
3'
2
STOP
3'
lncRNA
miRNA
5'
epigenetic
regulation
lncRNA
5'
3'
pre-mRNA
3'
5'
miRNA
lncRNA
protein
9
alteration of
protein localization
8
3
4
5
6
7
structural and/or
organizational role
3'
5'
alternatively spliced mRNA
siRNAs
miRNA silencing
modulation of
protein activity

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 ataxia 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 功能位点分析
6 启动子分析
7 CpG 岛识别
8 EMBOSS
9 序列分析中的算法
10 总结与答疑
11 引言
12 重复序列分析
13 基因识别
14 查找数据库与分析工具
15 总结与答疑
16 引言
17 mRNA 选择性剪接
18 miRNA 及其靶基因预测
19 lncRNA
20 学习数据库与分析工具的使用
21 总结与答疑
22 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
- 参考相关的专业书籍
- 参加（免费/收费的）培训班/讲座
- 各种 protocols 期刊：*Nature protocols, Current Protocols (in Bioinformatics/Human Genetics/Molecular Biology/...), 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 功能位点分析
6 启动子分析
7 CpG 岛识别
8 EMBOSS
9 序列分析中的算法
10 总结与答疑
11 引言
12 重复序列分析
13 基因识别
14 查找数据库与分析工具
15 总结与答疑
16 引言
17 mRNA 选择性剪接
18 miRNA 及其靶基因预测
19 lncRNA
20 学习数据库与分析工具的使用
21 总结与答疑
22 复习思考题

总结与答疑

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

晚眺-苏轼

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 功能位点分析
6 启动子分析
7 CpG 岛识别
8 EMBOSS
9 序列分析中的算法
10 总结与答疑
11 引言
12 重复序列分析
13 基因识别
14 查找数据库与分析工具
15 总结与答疑
16 引言
17 mRNA 选择性剪接
18 miRNA 及其靶基因预测
19 lncRNA
20 学习数据库与分析工具的使用
21 总结与答疑
22 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。

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。

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

TEX LATEX XETEX Beamer