基因识别

伊现富（Yi Xianfu）

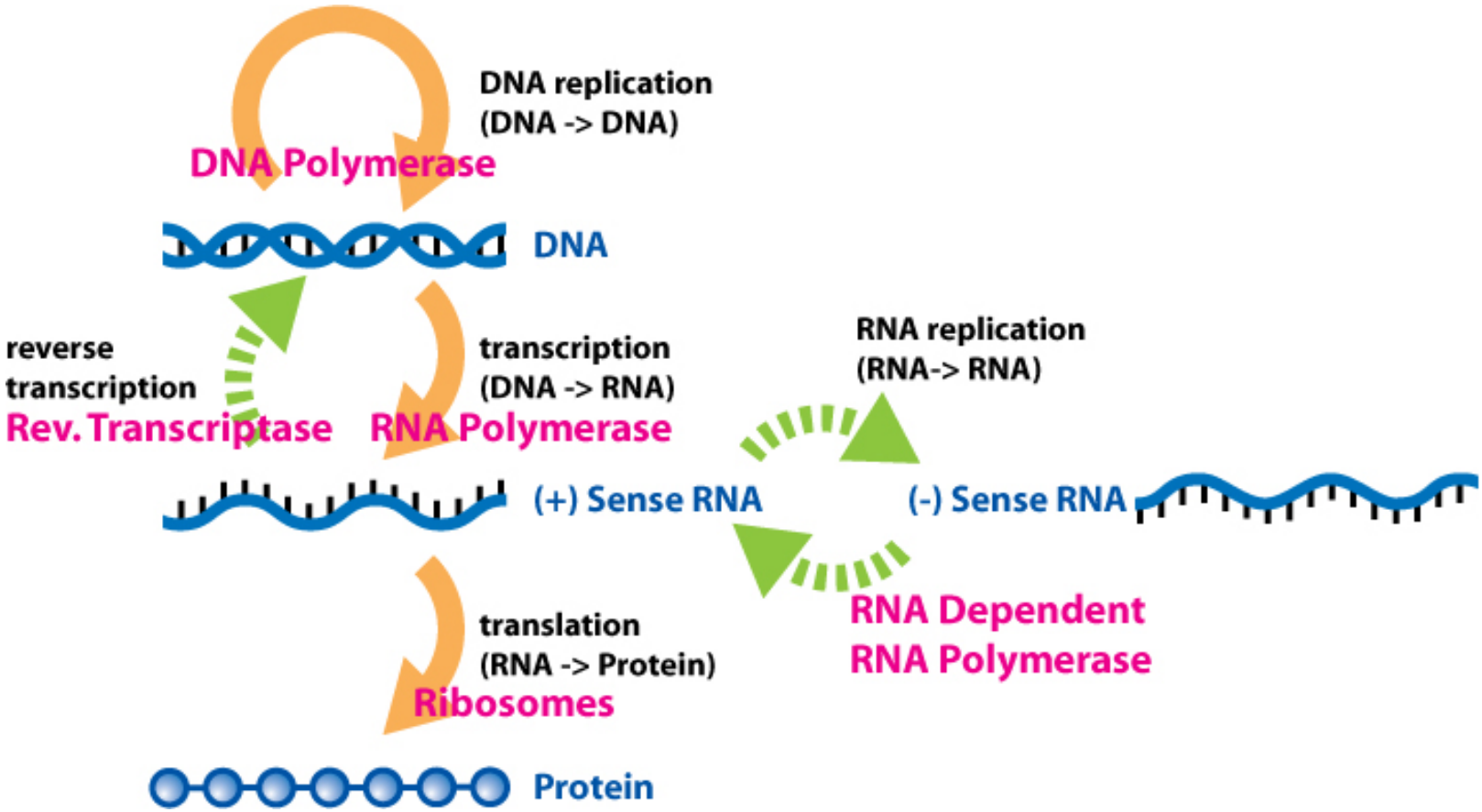
天津医科大学（TIJMU）
生物医学工程学院

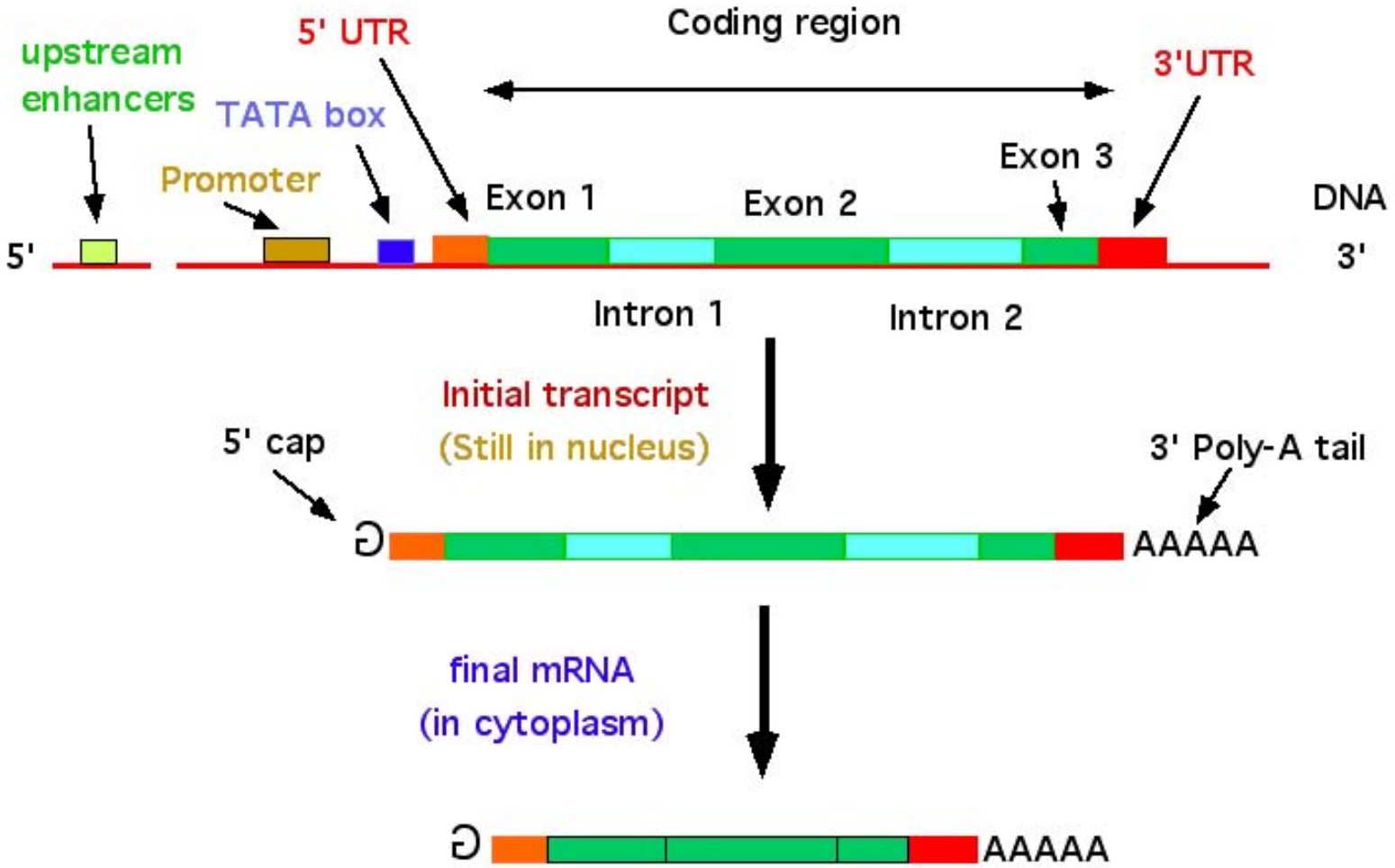
2014 年 5 月 23 日

教学提纲

教学提纲

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


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

教学提纲

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。

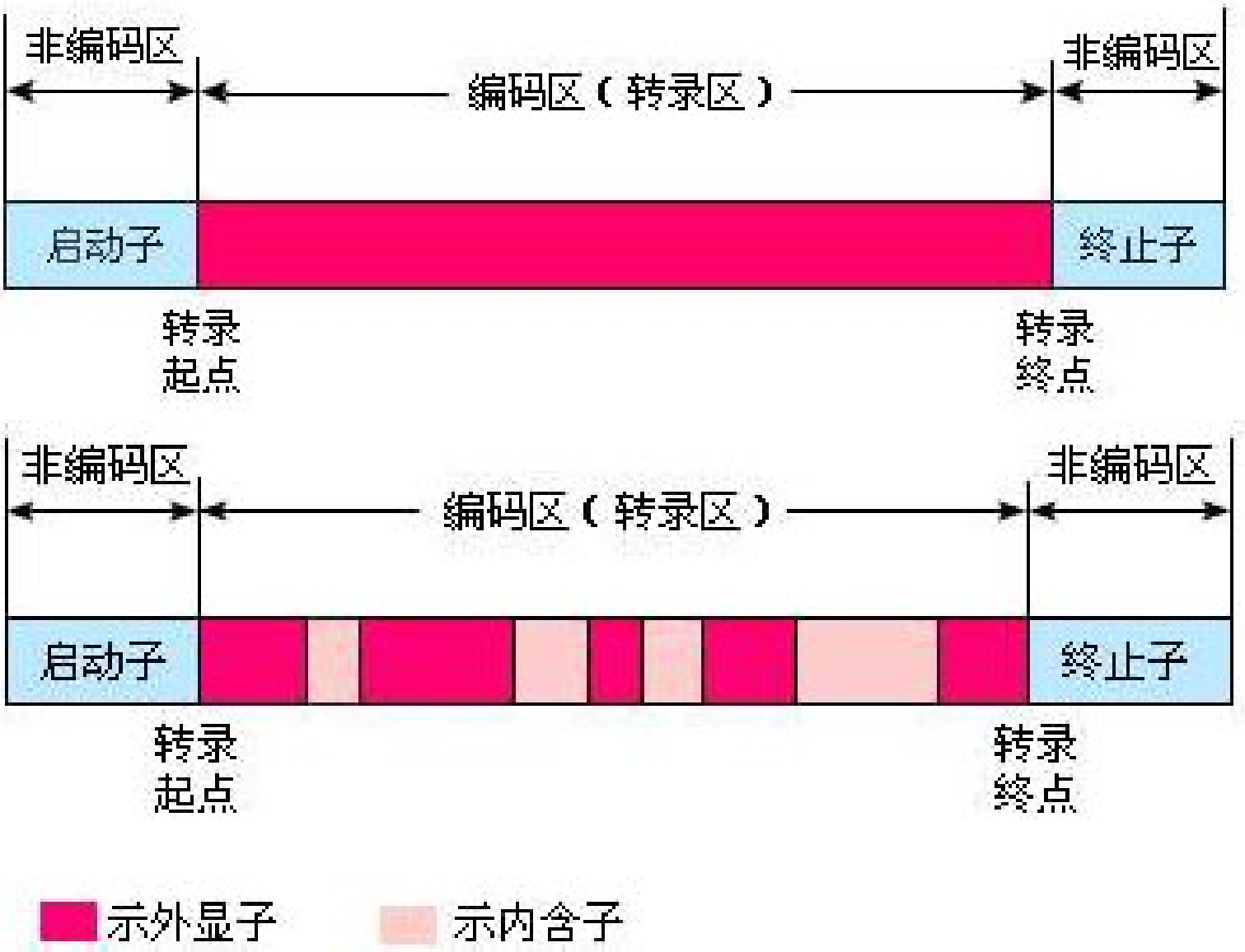
基因（gene）

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


非编码区
编码区（转录区）
非编码区
启动子
终止子
转录
起点
转录
终点
非编码区
编码区（转录区）
非编码区
启动子
终止子
转录
起点
转录
终点
示外显子
示内含子

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，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

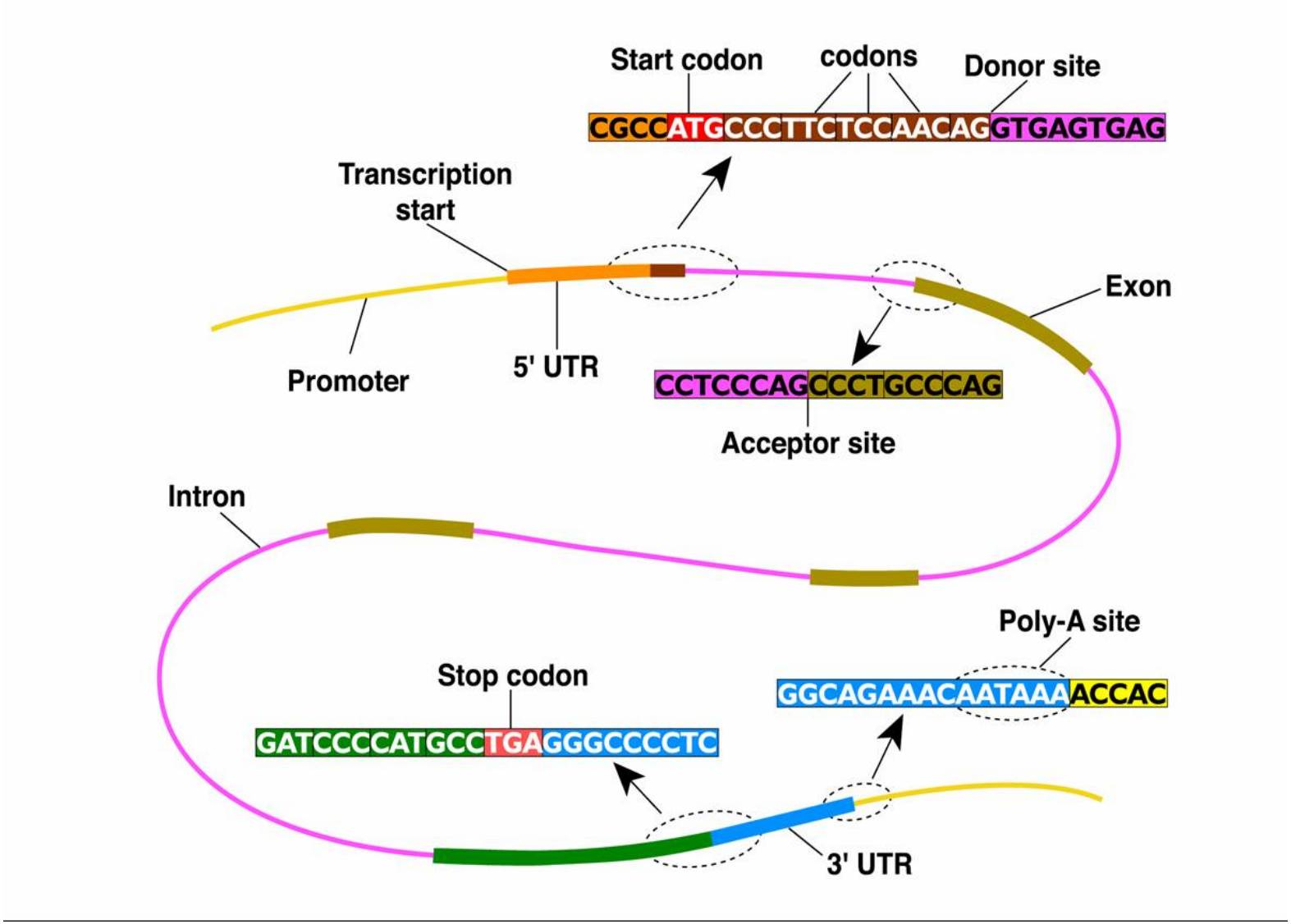
- 不同长度的扩展序列，没有一致性序列，但具有把自己与周围 DNA 区分开来的保守特征
- 密码子使用偏好性（codon usage bias），双联密码子出现频率，基因组等值区（isochore）

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）


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	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

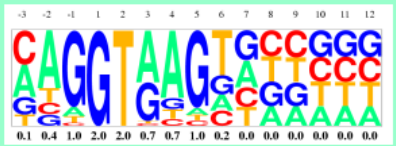
内容

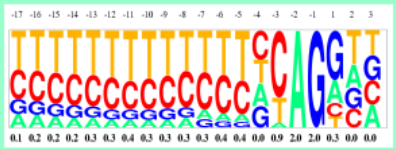
密码子使用偏好性，双联密码子出现频率，基因组等值区

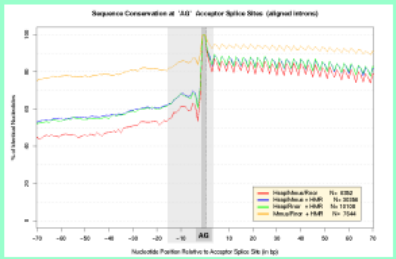
总结

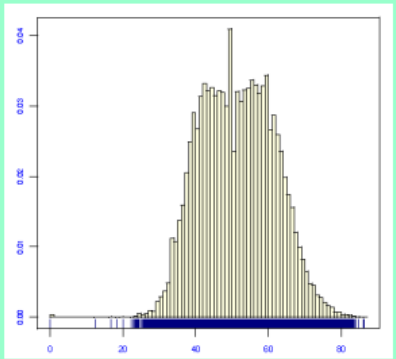
- 综合信号信息确定外显子的边界，识别编码区域
- 通过内容统计值区分外显子、内含子和基因间区域
- 信号复杂，内容难判别，预测相当有挑战性
- 联合信号和内容检测以及同源性搜索，提高识别效率

基因识别 | 真核基因









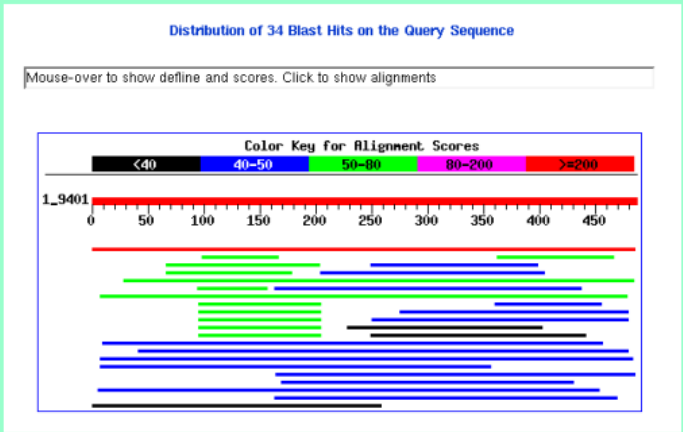
Homo sapiens

ANNOTATION DATABASE

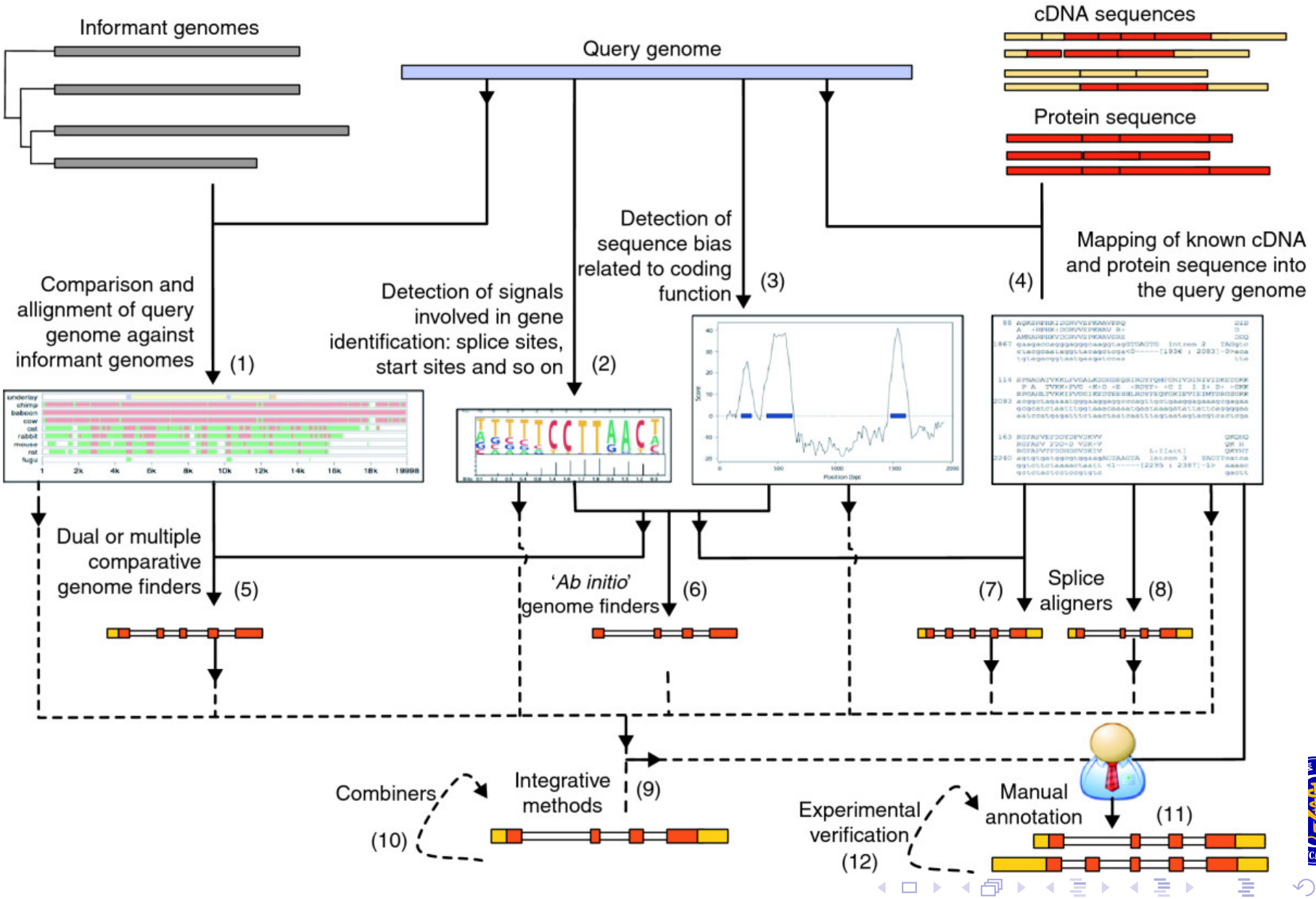
ab initio

homology

GENES



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	

工具列表

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

常见工具

- 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

教学提纲

总结与答疑

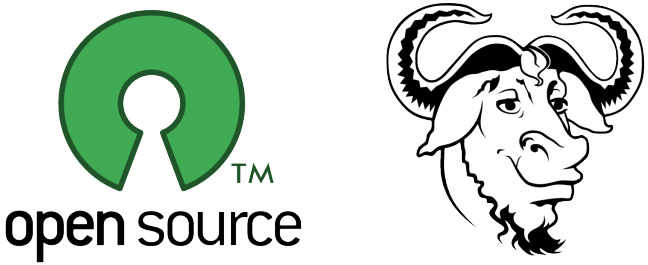
- 原核和真核的基因结构
- 基因识别的方法
- 基因预测中的信号 vs. 内容
- 基因识别的策略

教学提纲

复习思考题

1. 比较原核和真核的基因结构。
2. 简述基因识别的三大类方法。
3. 比较基因预测中的信号与内容。
4. 论述基因识别的主要策略。

Powered by



TEX LATEX XETEX Beamer