生物信息学与系统生物学

章祥荪　　吴凌云　　王勇　　张世华

中国科学院数学与系统科学研究院

生物信息学

基础知识

吴凌云

中国科学院数学与系统科学研究院

大纲

- 序列比较与分析
- 基因与Motif预测
- 单体型组装与推断
- 蛋白质结构与功能
- 基因组变异分析
- 高通量技术

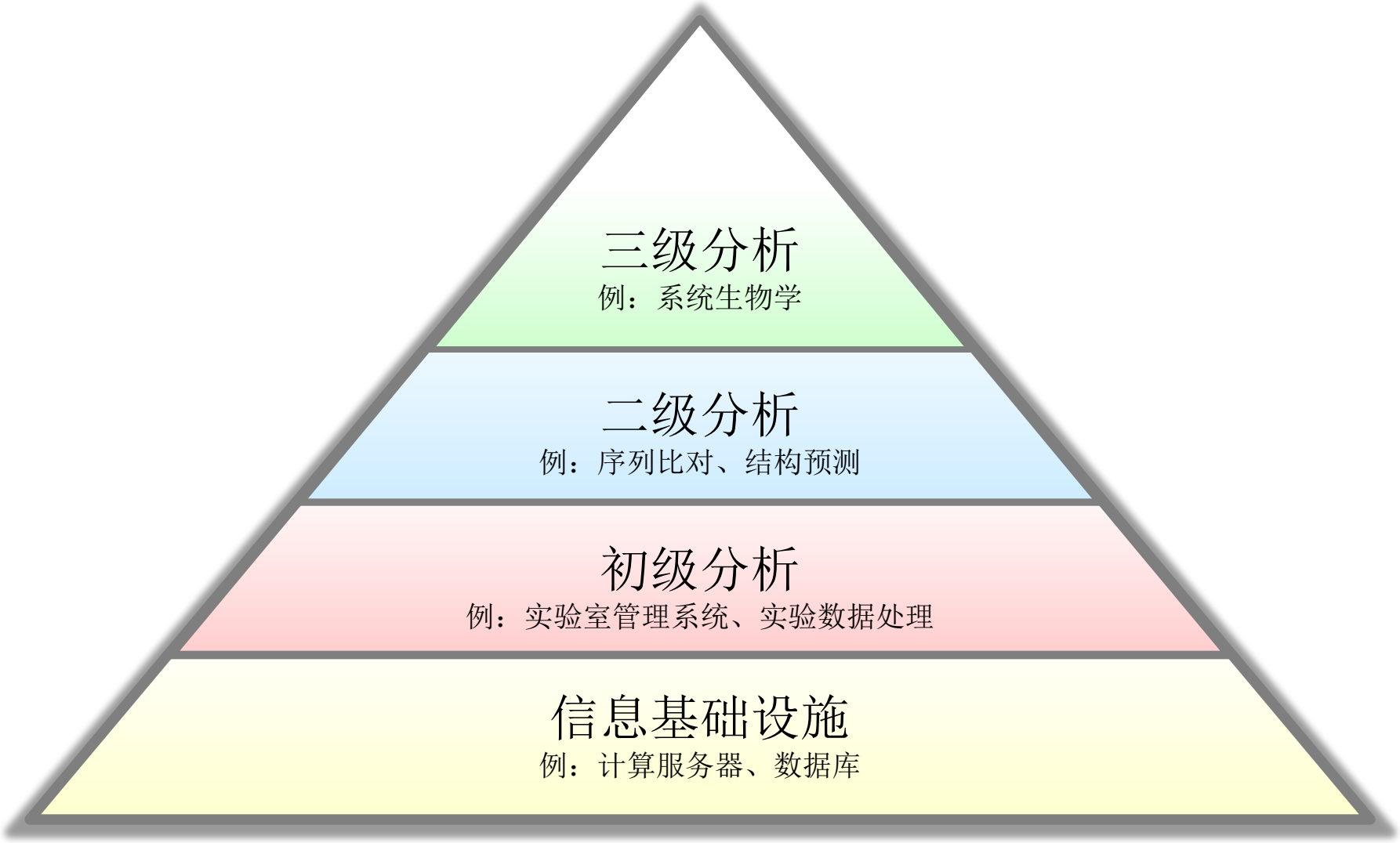
课程信息

- http://doc.aporc.org/wiki/Course001

课程目标

- 什么问题（what problems）
- 如何解决（how to solve）
- 关键技术（key computational techniques）
- 多大作用（how much help）

生物信息学与系统生物学



研究对象

- DNA and RNA
- Protein
- Functions
  - Regulatory
  - Interactions
- Signals
- Experimental Data

数学模型

- Sequences
  - DNA, RNA, Protein
- Structures
  - DNA, RNA, Protein
- Networks
  - Protein-Protein Interactions, Gene Regulatory, Signaling Pathway

数学方法 (1)

- Statistics and Probabilistic, e.g. hidden Markov model (HMM) in sequence analysis and gene finding
- Operations Research, e.g. dynamic programming (DP) in sequence alignment
- Mathematical Programming, e.g. linear programming and non-linear programming in protein structure prediction and docking problem

数学方法 (2)

- Topology, e.g. in protein structure comparison
- Functional Theory, e.g. Fourier transform and Wavelet in protein structure comparison
- Information Theory, e.g. in sequence alignment and protein structure prediction

数学方法 (3)

- Computational Mathematics, e.g. differential equation in molecular dynamics and protein folding problem
- Groups Theory, e.g. in genetic codes and DNA sequence analysis
- Combinatory, e.g. in molecular evolution and DNA rearrangement

Journals (1)

- Explicitly including word “Bioinformatics” in journal name
  - Bioinformatics
  - BMC Bioinformatics
  - Applied Bioinformatics
  - Briefings in Bioinformatics
  - Journal of Bioinformatics and Computational Biology
  - Genomics, Proteomics & Bioinformatics
  - …

Journals (2)

- Other biological journals
  - Nucleic Acids Research
  - Genome Research
  - Genomics
  - Journal of Molecular Biology
  - Journal of Computational Biology
  - Computational Biology and Chemistry
  - Molecular Biology and Evolution
  - BioTechniques
  - BioTechnology Software
  - …

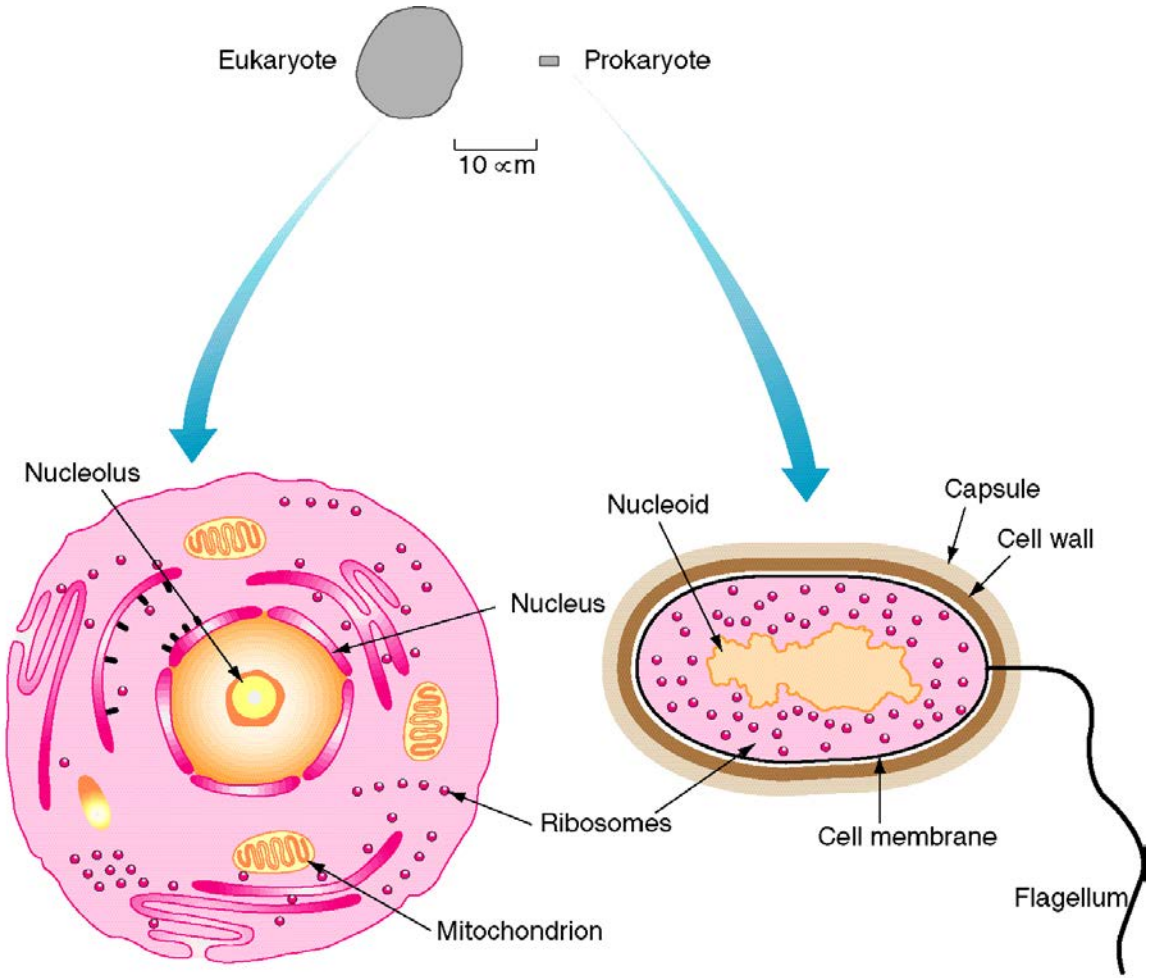
International Conferences

- WABI: International Workshop on Algorithms in Bioinformatics
- RECOMB: International Conference on Research in Computational Molecular Biology
- ISMB: International Conference on Intelligent Systems for Molecular Biology
- ICSB: International Conference on Systems Biology
- CSB: IEEE Computational Systems Bioinformatics Conference
- PSB: Proceedings of Pacific Symposium on Biocomputing
- …

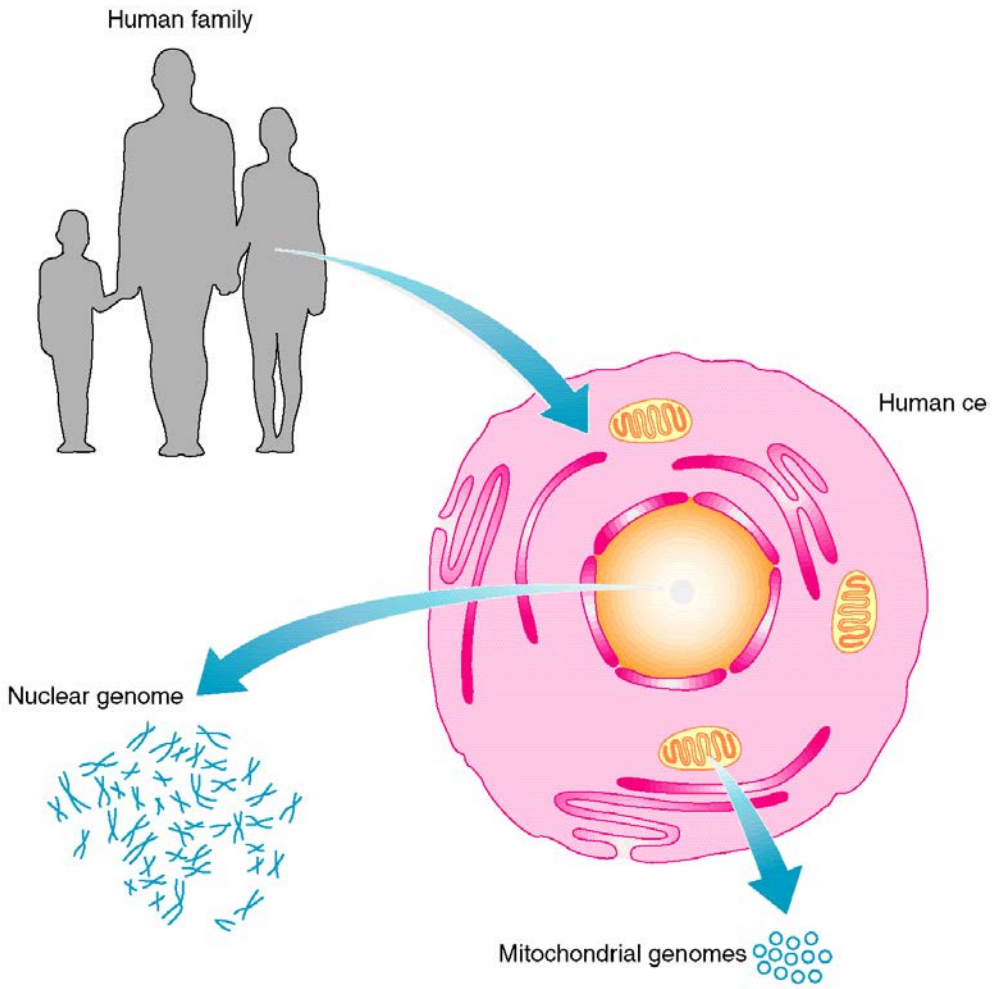
目录

- 染色体
- DNA
- 基因
- 蛋白质

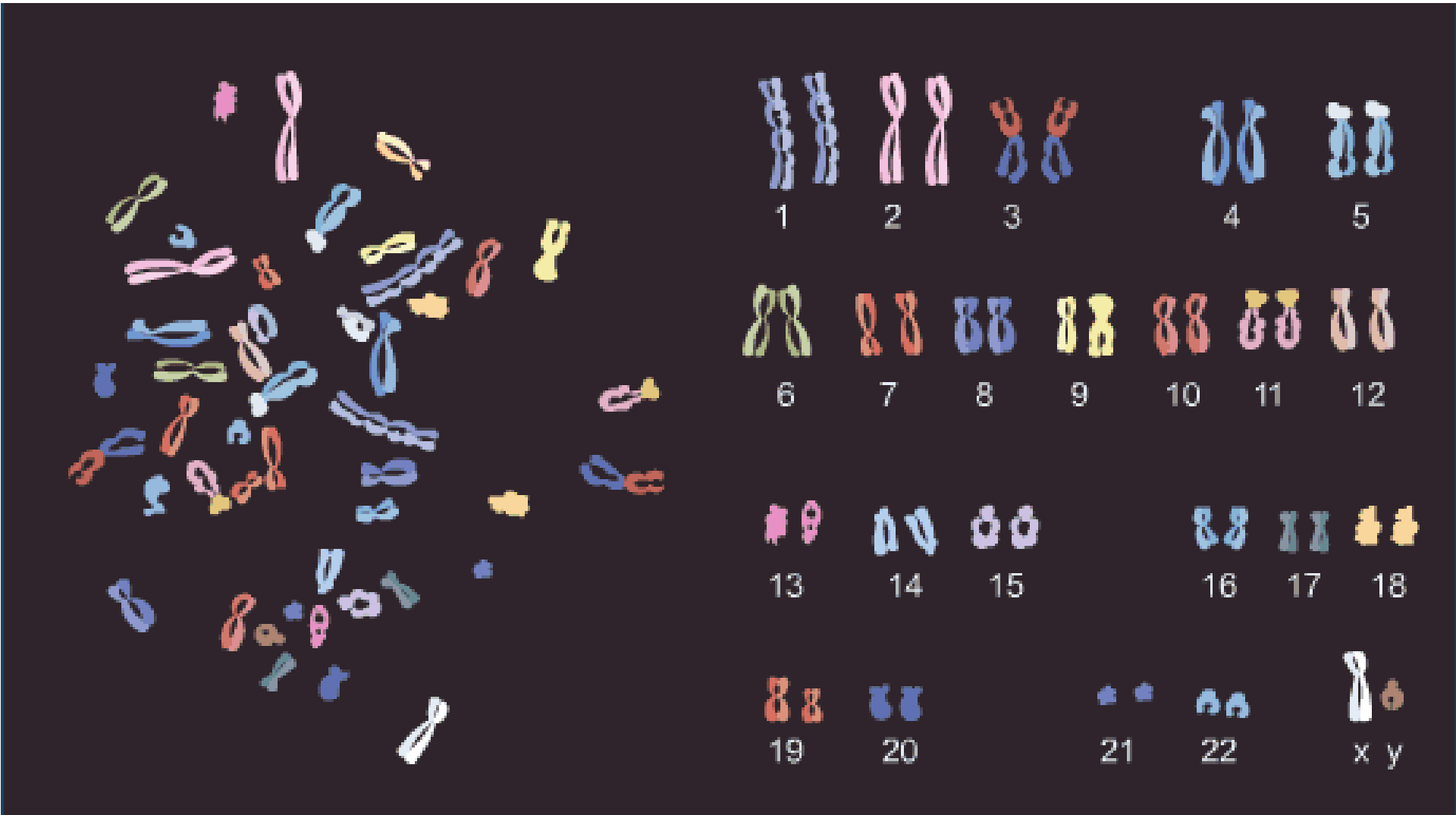
细胞结构


Eukaryote
Prokaryote
10 ∝m
Nucleolus
Nucleus
Ribosomes
Mitochondrion
Nucleoid
Capsule
Cell wall
Cell membrane
Flagellum

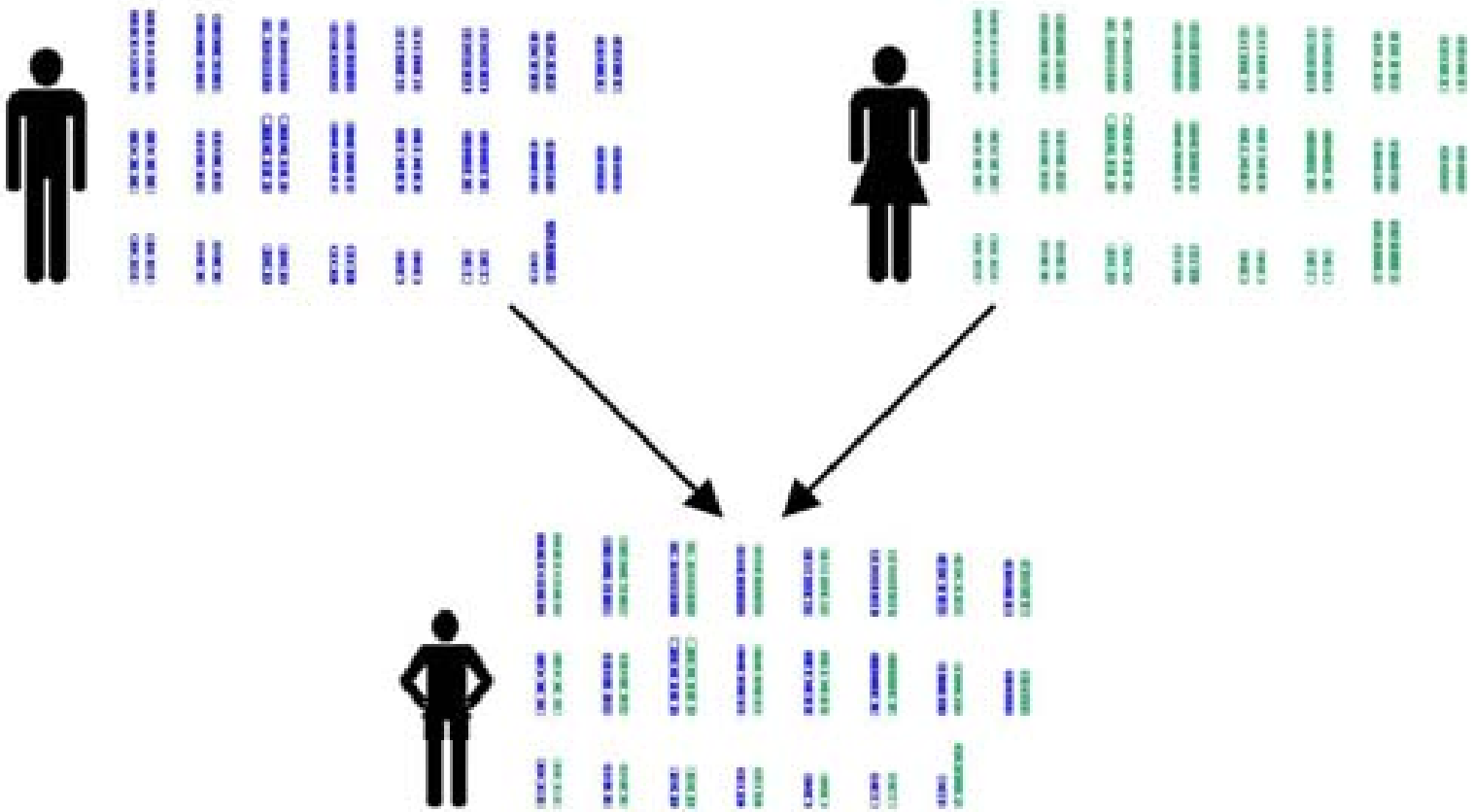
染色体与基因组


Human family
Human ce
Nuclear genome
Mitochondrial genomes

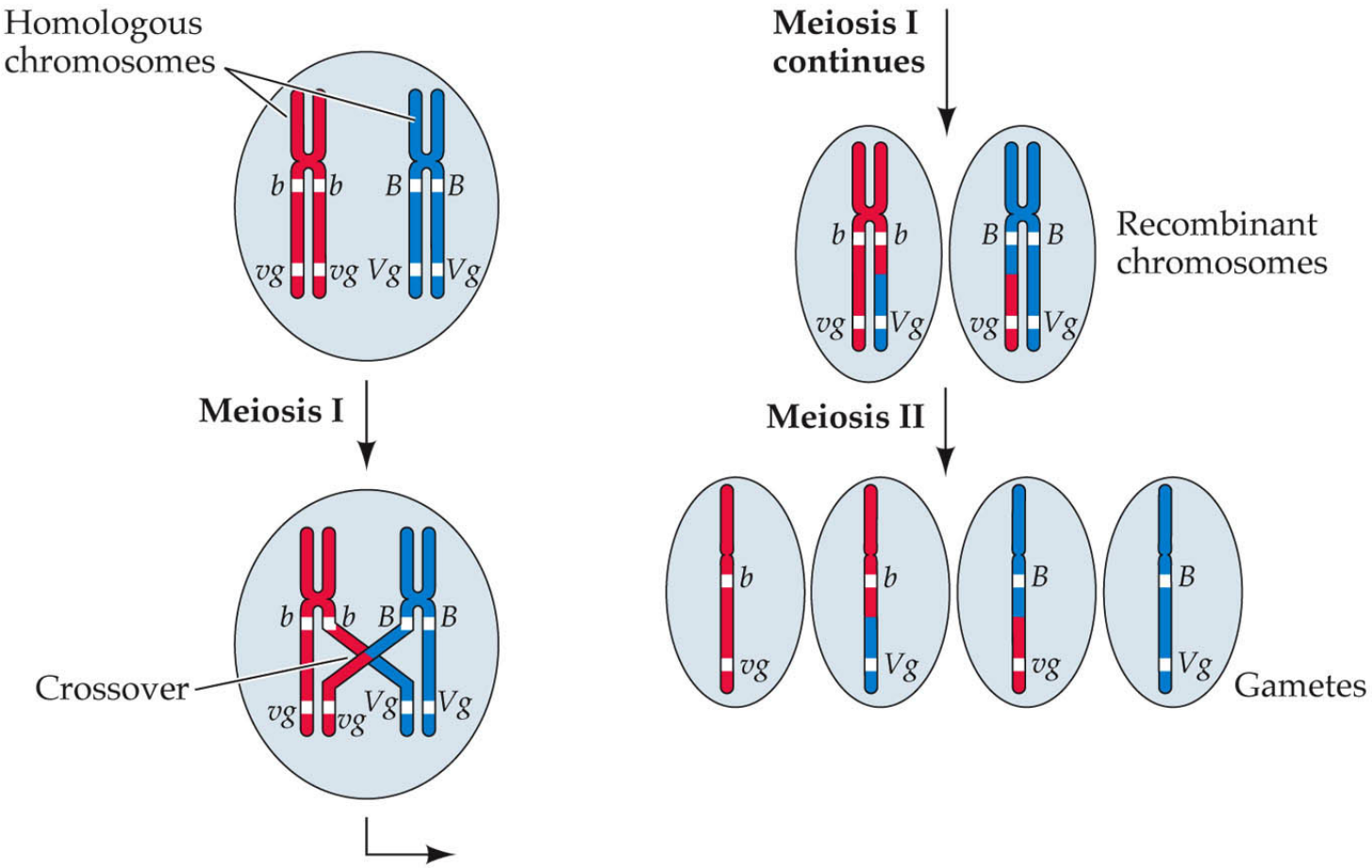
人的23对染色体



染色体的遗传

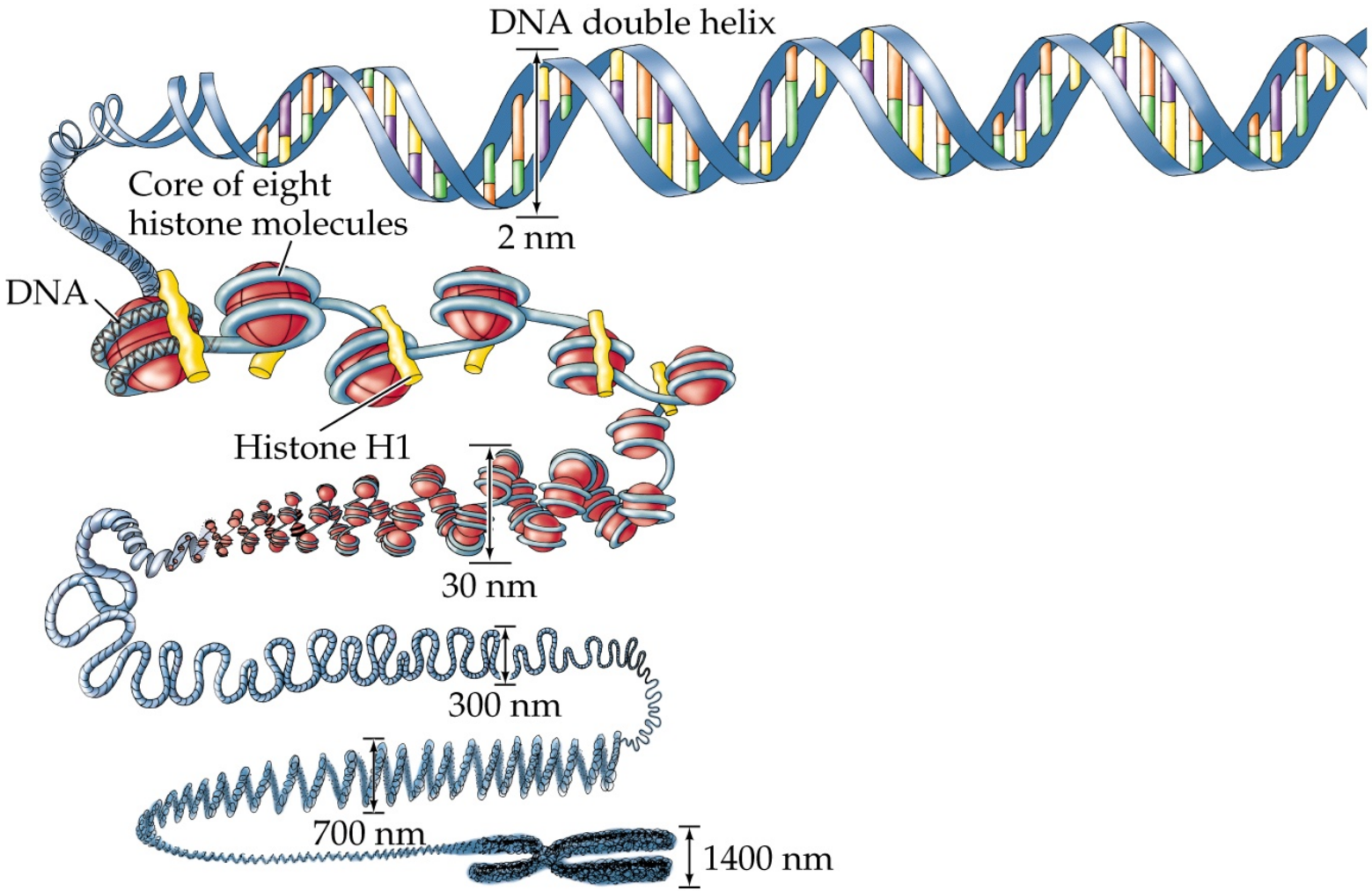


染色体的重组



LIFE: THE SCIENCE OF BIOLOGY, Seventh Edition, Figure 10.19 Crossing Over Results in Genetic Recombination

染色体的结构



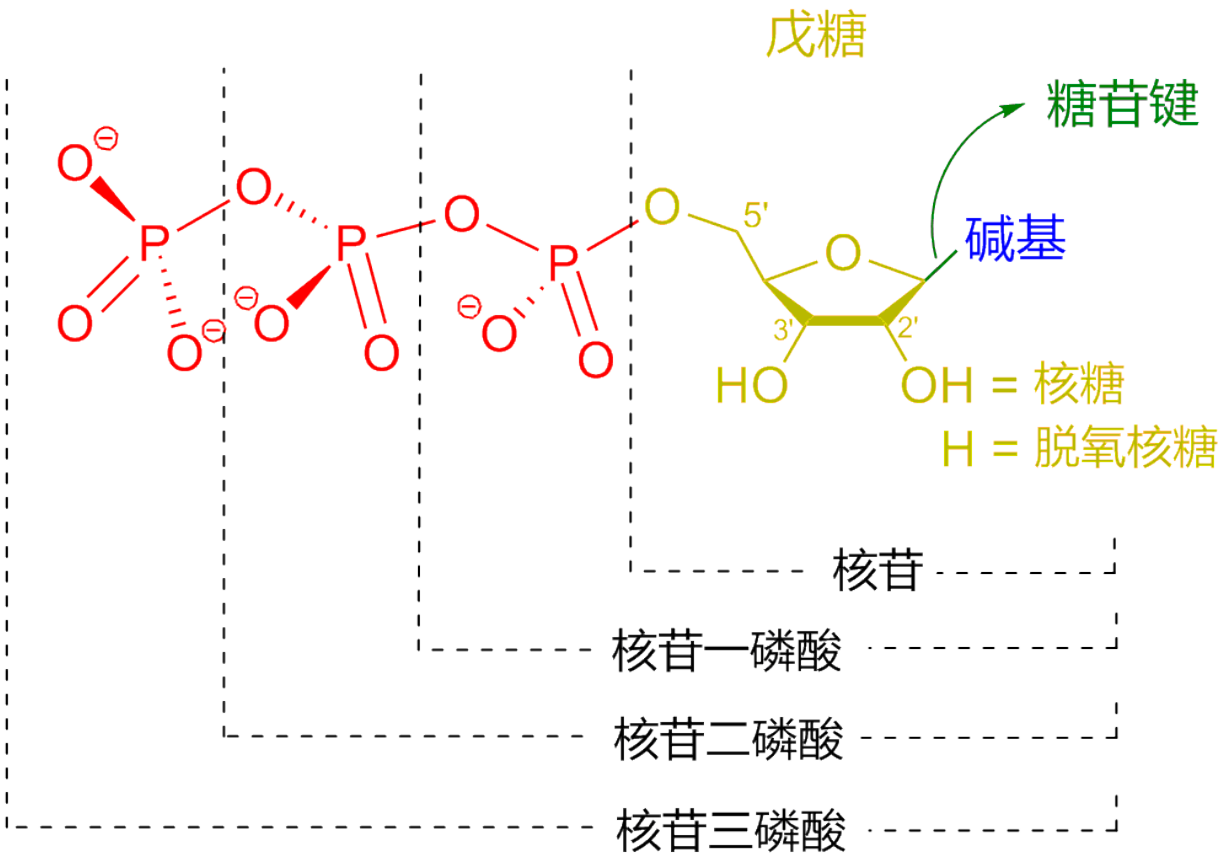
染色体相关概念

- 染色体（chromosome）
- 同源染色体（homology chromosome）
- 等位基因（allele）

目录

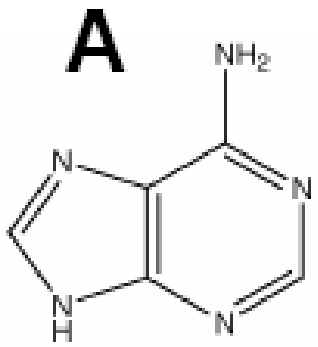
- 染色体
- DNA
- 基因
- 蛋白质

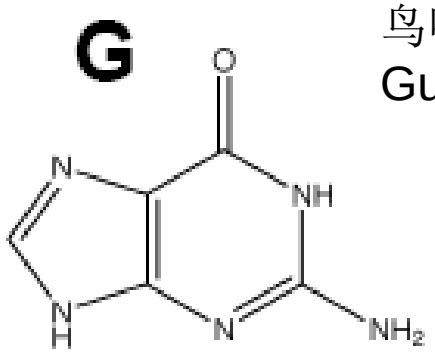
脱氧核糖核苷酸



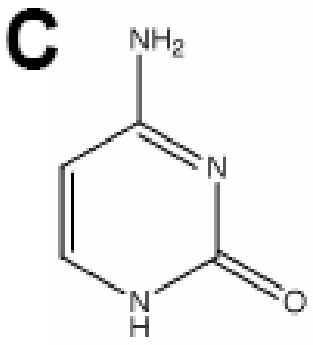
碱基 (Bases)

腺嘌呤
Adenine

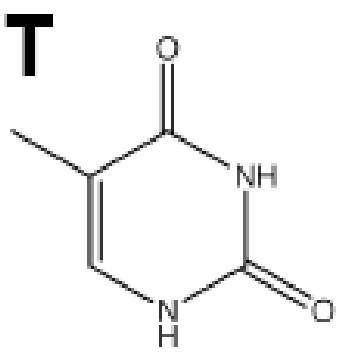




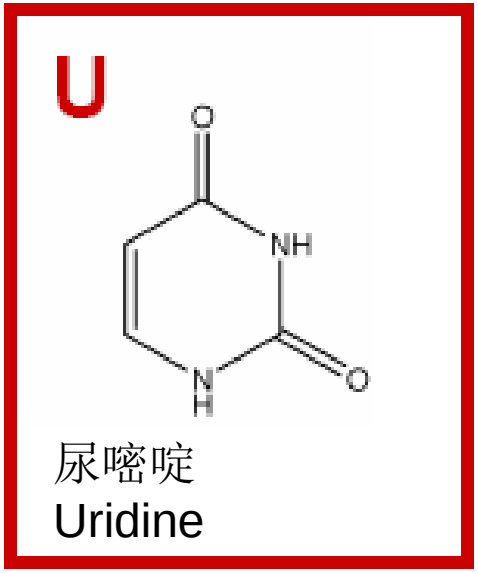
鸟嘌呤
Guanine



胞嘧啶
Cytosine

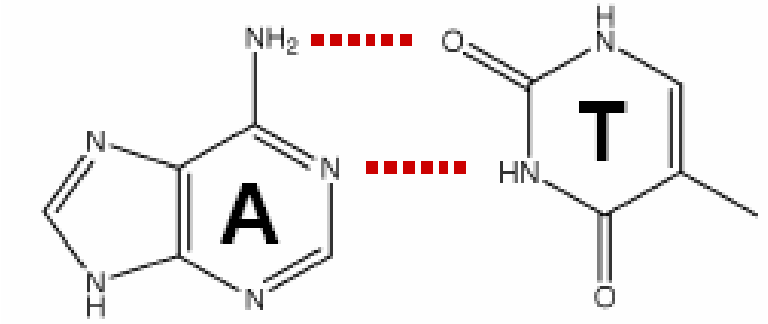


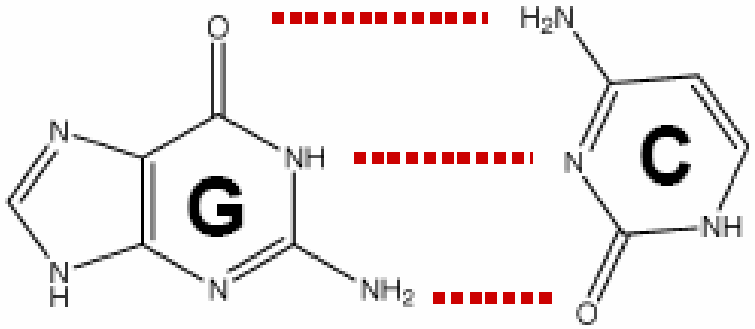
胸腺嘧啶
Thymine



尿嘧啶
Uridine

碱基配对 (Base Paring)

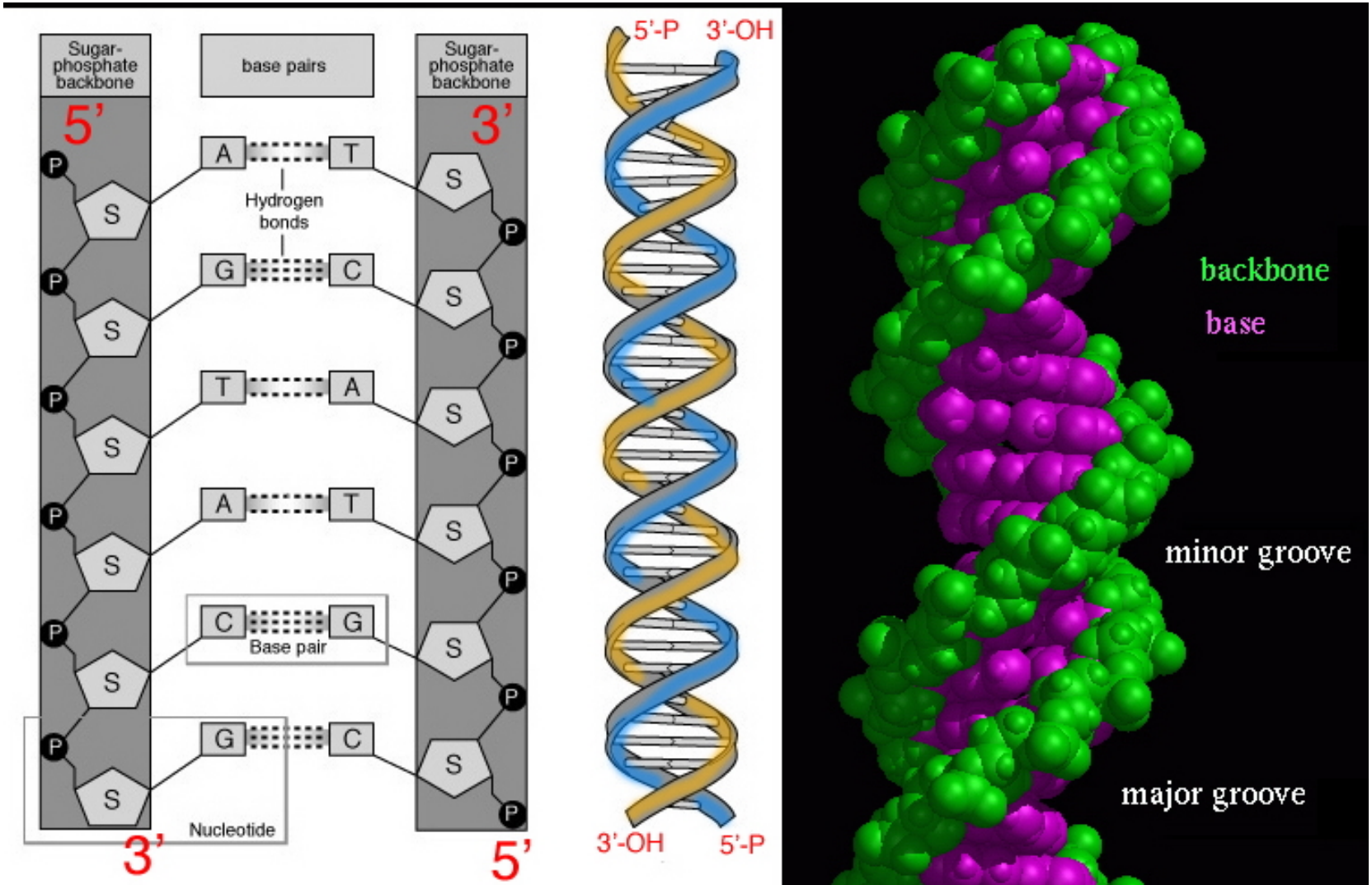




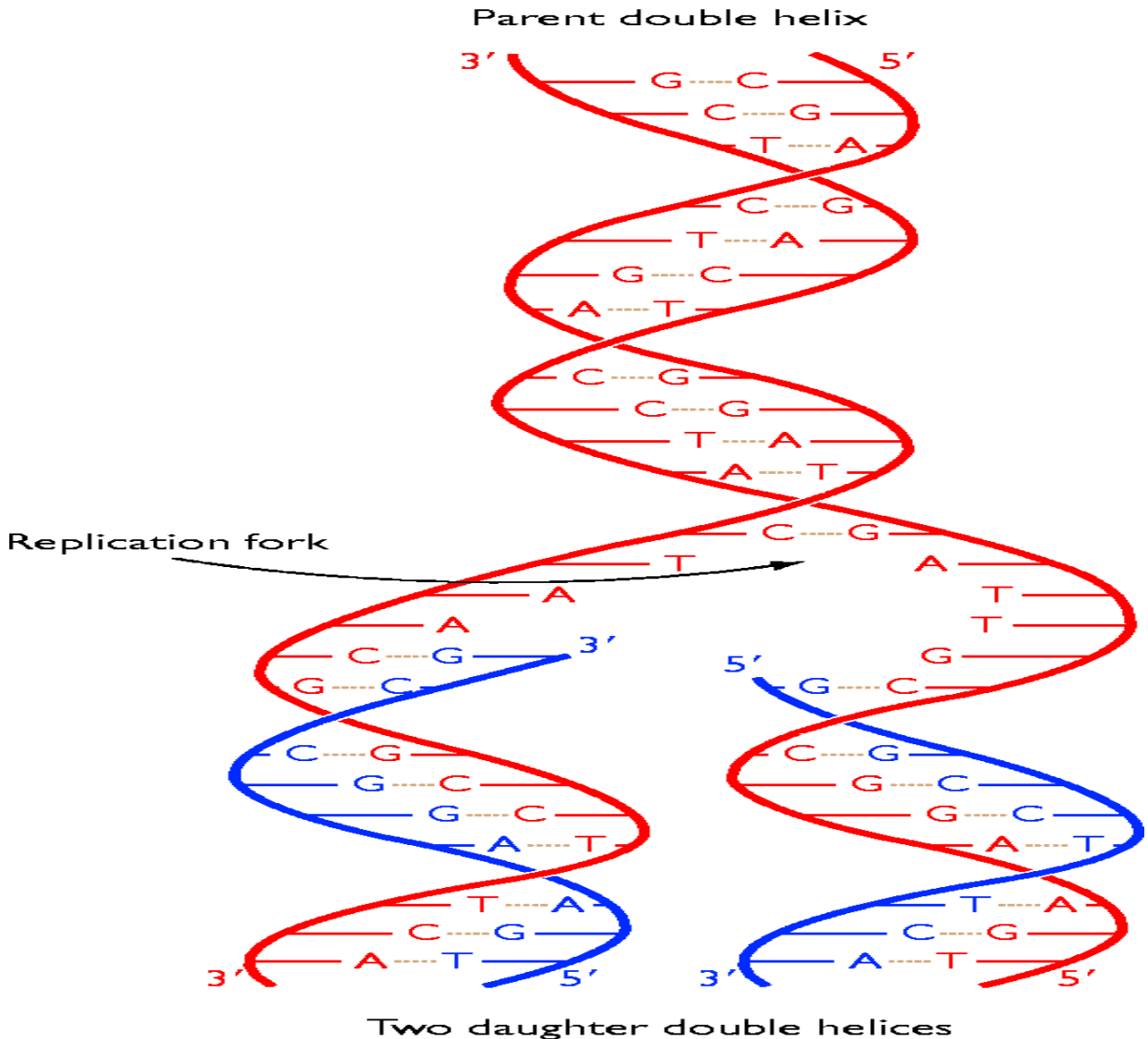
碱基分类

- 嘌呤（purine）
  - A：adenine，腺嘌呤，氨基，弱键
  - G：guanine，鸟嘌呤，酮基，强健
- 嘧啶（pyrimidine）
  - C：cytosine，胞嘧啶，氨基，强健
  - T：thymine，胸腺嘧啶，酮基，弱键
- 氨基: A, C / 酮基: T, G
- 弱键: A, T / 强键: C, G

DNA的双螺旋结构


Sugar-phosphate backbone
base pairs
Sugar-phosphate backbone
5'
3'
A
T
Hydrogen bonds
G
C
T
A
A
T
C
G
Base pair
G
C
Nucleotide
3'
5'
5'-P
3'-OH
3'-OH
5'-P
backbone
base
minor groove
major groove

DNA的复制


Parent double helix
3′
5′
Replication fork
3′
5′
3′
5′
Two daughter double helices

DNA的相关概念

- DNA双螺旋结构（double helix；double strand）
- DNA链的方向（orientation）：3’和5’
- 碱基互补配对（Watson-Crick base pairing）：A/T C/G
- 碱基对（base pair）

目录

- 染色体
- DNA
- **基因**
- 蛋白质

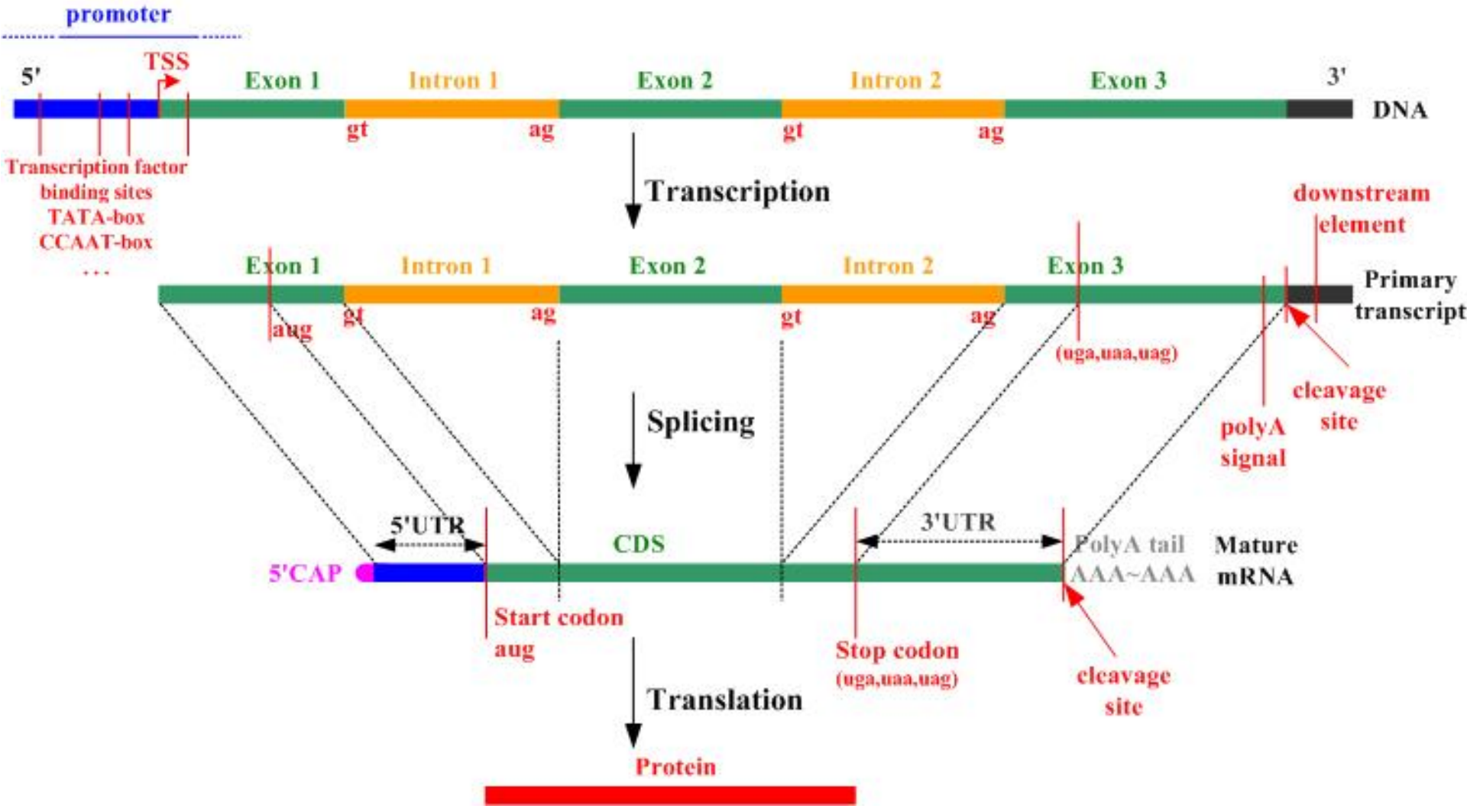
基因和基因组

- DNA上具有特定功能的一个片断，负责一种特定性状的表达。一般来讲，一个基因只编码一个蛋白质。
- 任何一条染色体上都带有许多基因，一条高等生物的染色体上可能带有成千上万个基因，一个细胞中的全部基因序列及其间隔序列统称为基因组（genome）。

DNA和基因

- 基因间的DNA（junk DNA）
- 基因中的内含子（intron）
- 基因中的外显子（exon）
- 启动子（promoter）

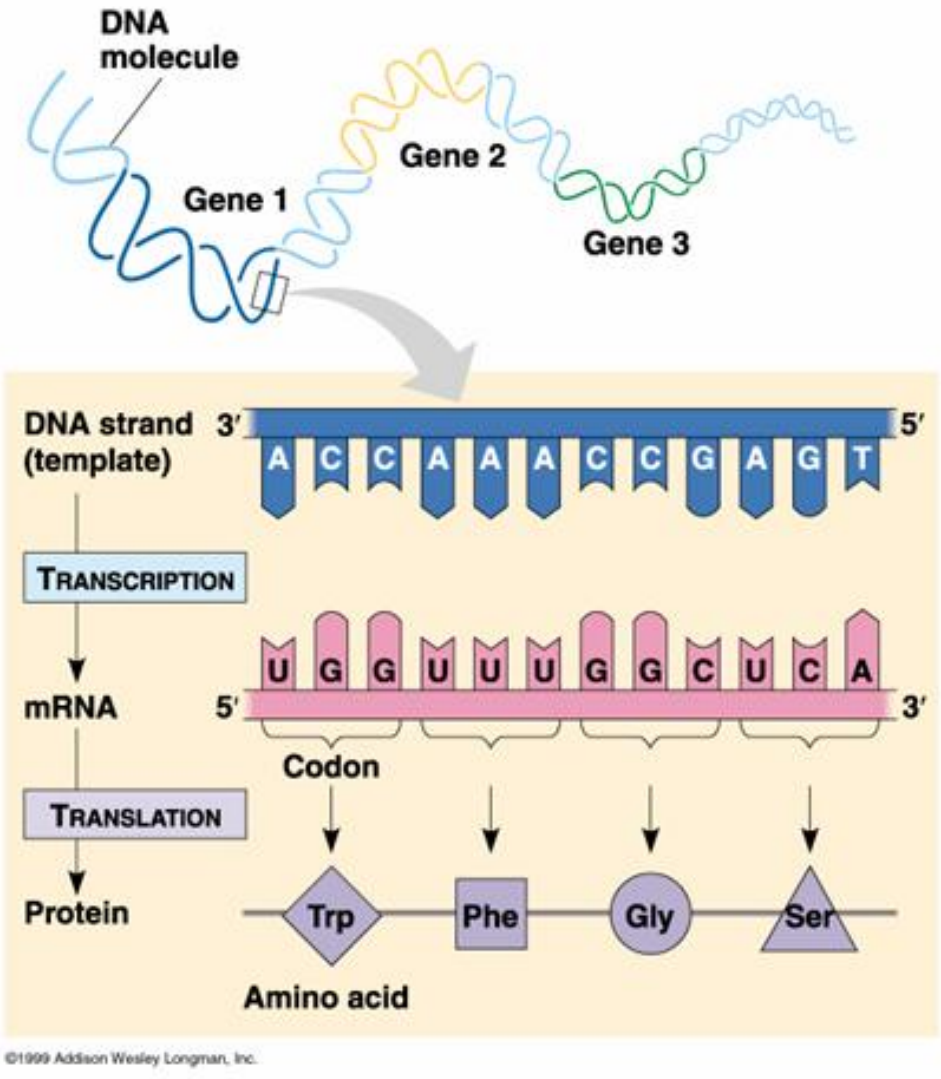
中心法则（Central Dogma）



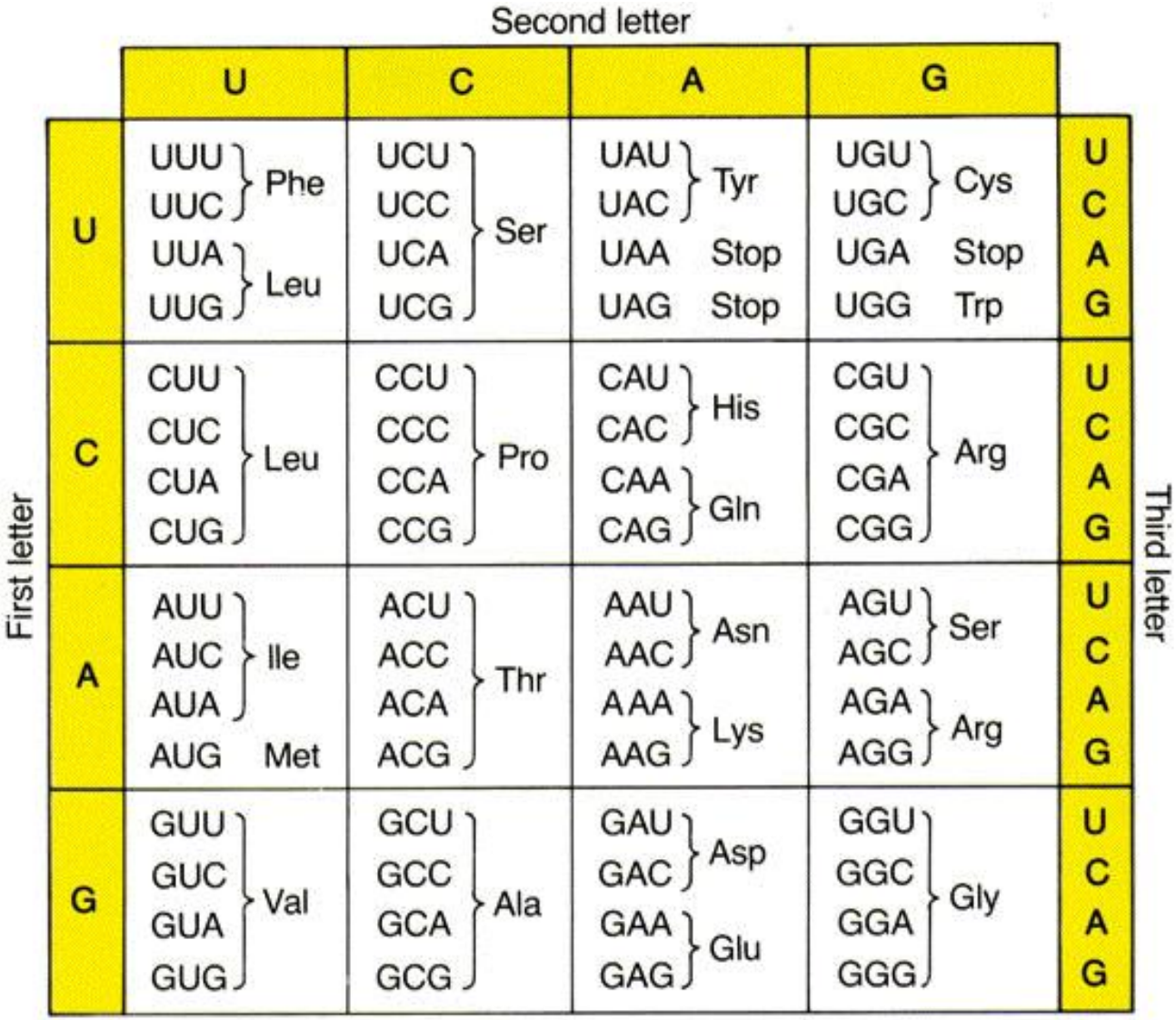
目录

- 染色体
- DNA
- 基因
- **蛋白质**

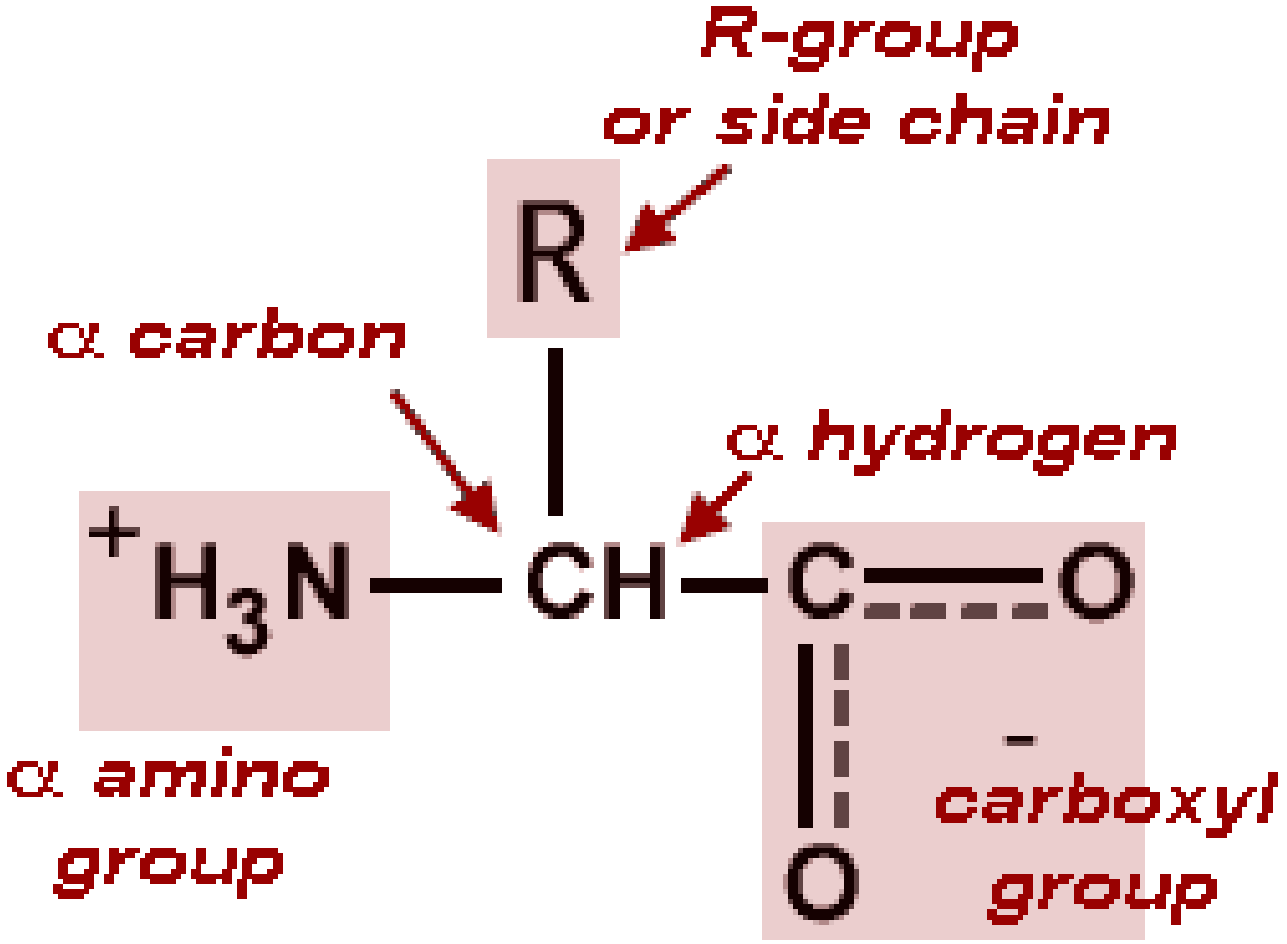
蛋白质的合成


DNA molecule
Gene 1
Gene 2
Gene 3
DNA strand (template)
3′
5′
A C C A A A C C G A G T
TRANSCRIPTION
mRNA
5′
3′
U G G U U U G G C U C A
Codon
TRANSLATION
Protein
Trp
Phe
Gly
Ser
Amino acid
©1999 Addison Wesley Longman, Inc.

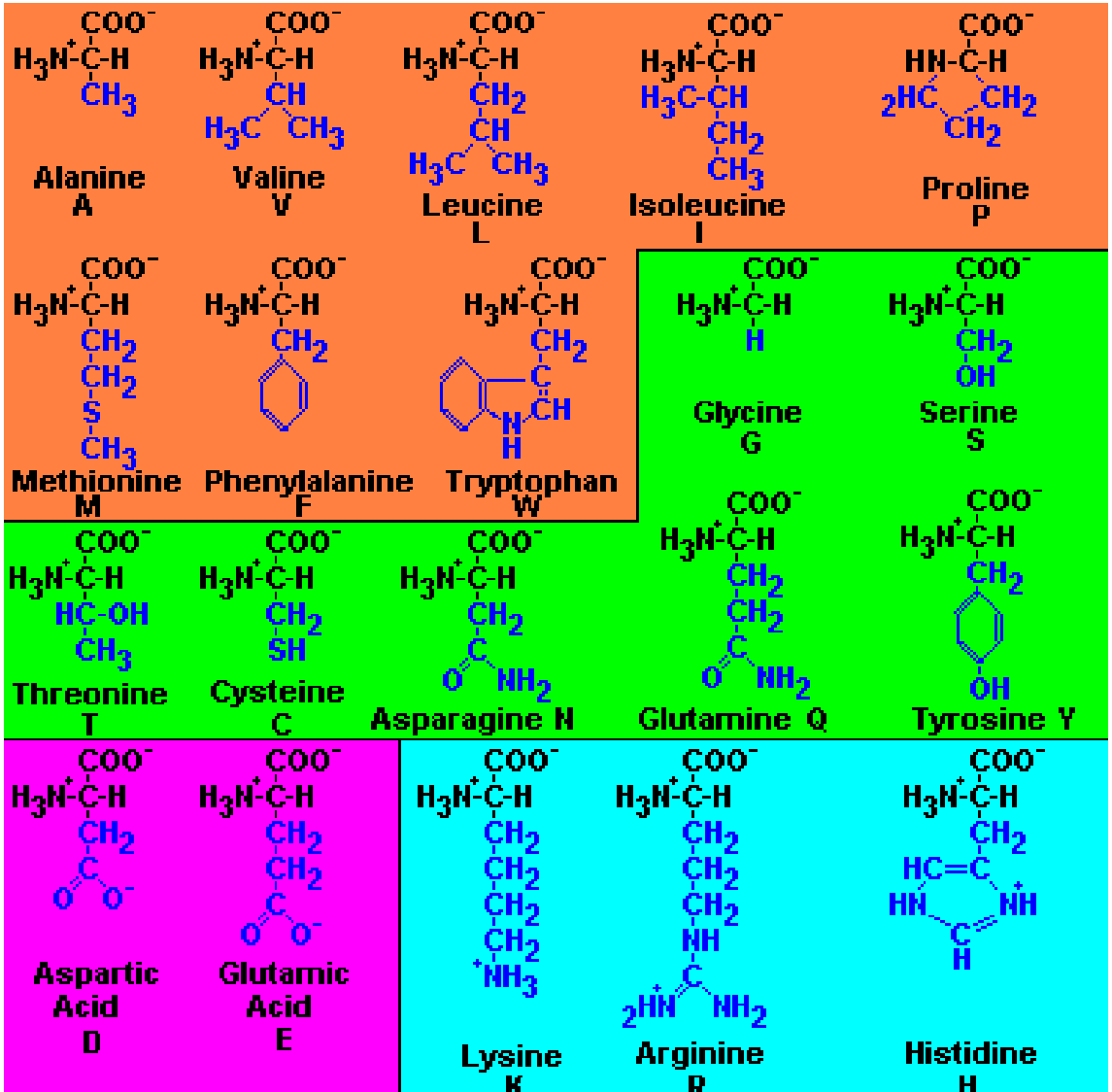
遗传密码（Genetic Code）



氨基酸的结构

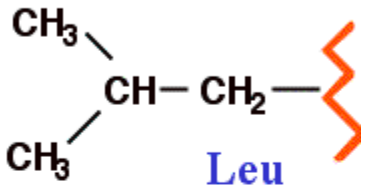
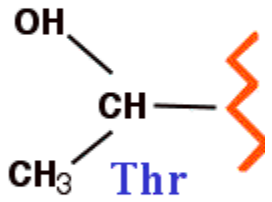
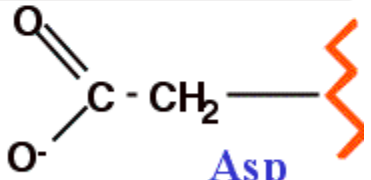
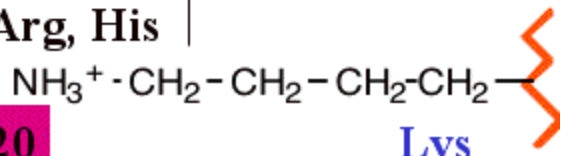

R-group
or side chain
R
α carbon
α hydrogen
+H3N
CH
C
O
O
-
α amino
group
carboxyl
group

氨基酸


Non-polar
Polar
Acidic
Basic
Alanine A
Valine V
Leucine L
Isoleucine I
Proline P
Methionine M
Phenylalanine F
Tryptophan W
Glycine G
Serine S
Threonine T
Cysteine C
Asparagine N
Glutamine Q
Tyrosine Y
Aspartic Acid D
Glutamic Acid E
Lysine K
Arginine R
Histidine H

氨基酸的分类

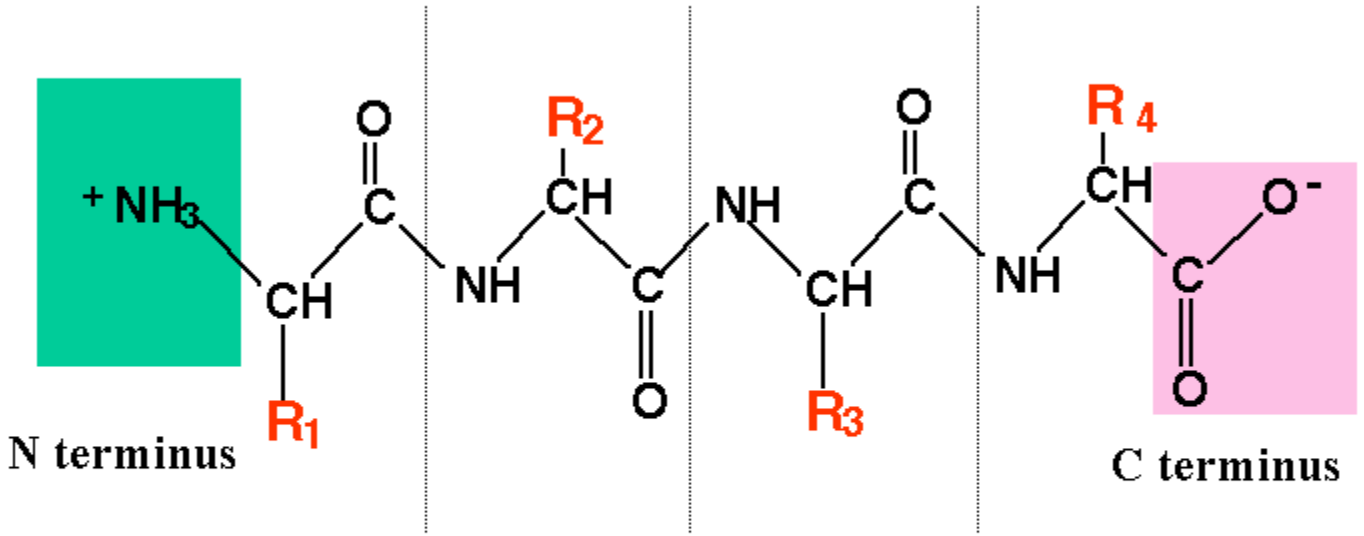
Amino acids groups

Group	Characteristics	Names	Example (-Rx)
non-polar	hydrophobic	Ala, Val, Leu, Ile, Pro, Phe Trp, Met	 Leu
polar	hydrophilic (non-charged)	Gly , Ser, Thr, Cys, Tyr, Asn Gln	 Thr
acidic	negatively charged	Asp, Glu	 Asp
basic	positively charged	Lys, Arg, His	 Lys

Total = 20

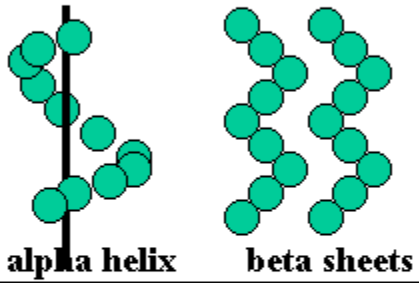
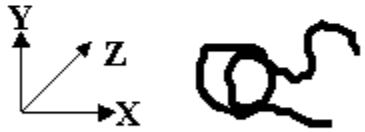
肽链（Peptide）

Peptide = chain of amino acids



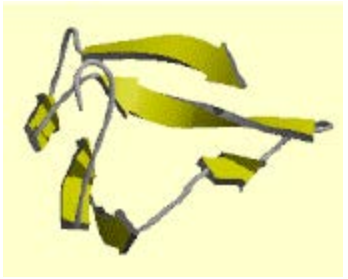
蛋白质结构

Protein Structure(Summary)

•Primary	The amino acid sequence	Glu-Arg-Phe-Gly
•Secondary	Characteristic structures that occur in many proteins (E.g. alpha helix , beta sheets)	
•Tertiary	Three dimensional structure of proteins	
•Quaternary	Three dimensional structure of proteins composed of multiple subunits	

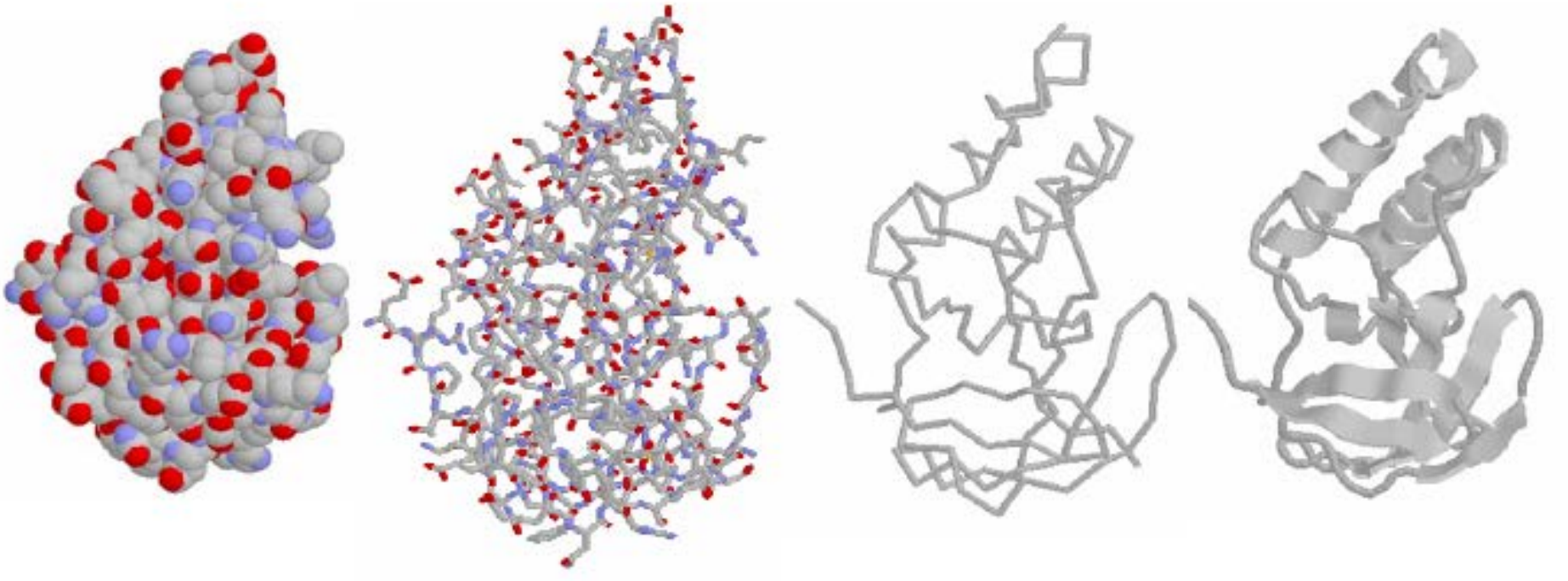
蛋白质二级结构

- α-helix (30-35%)
  α-螺旋
- β-sheet / β-strand (20-25%)
  β-折叠
- Coil (40-50%) 无规则卷曲
- Loop 环
- β-turn β-转角

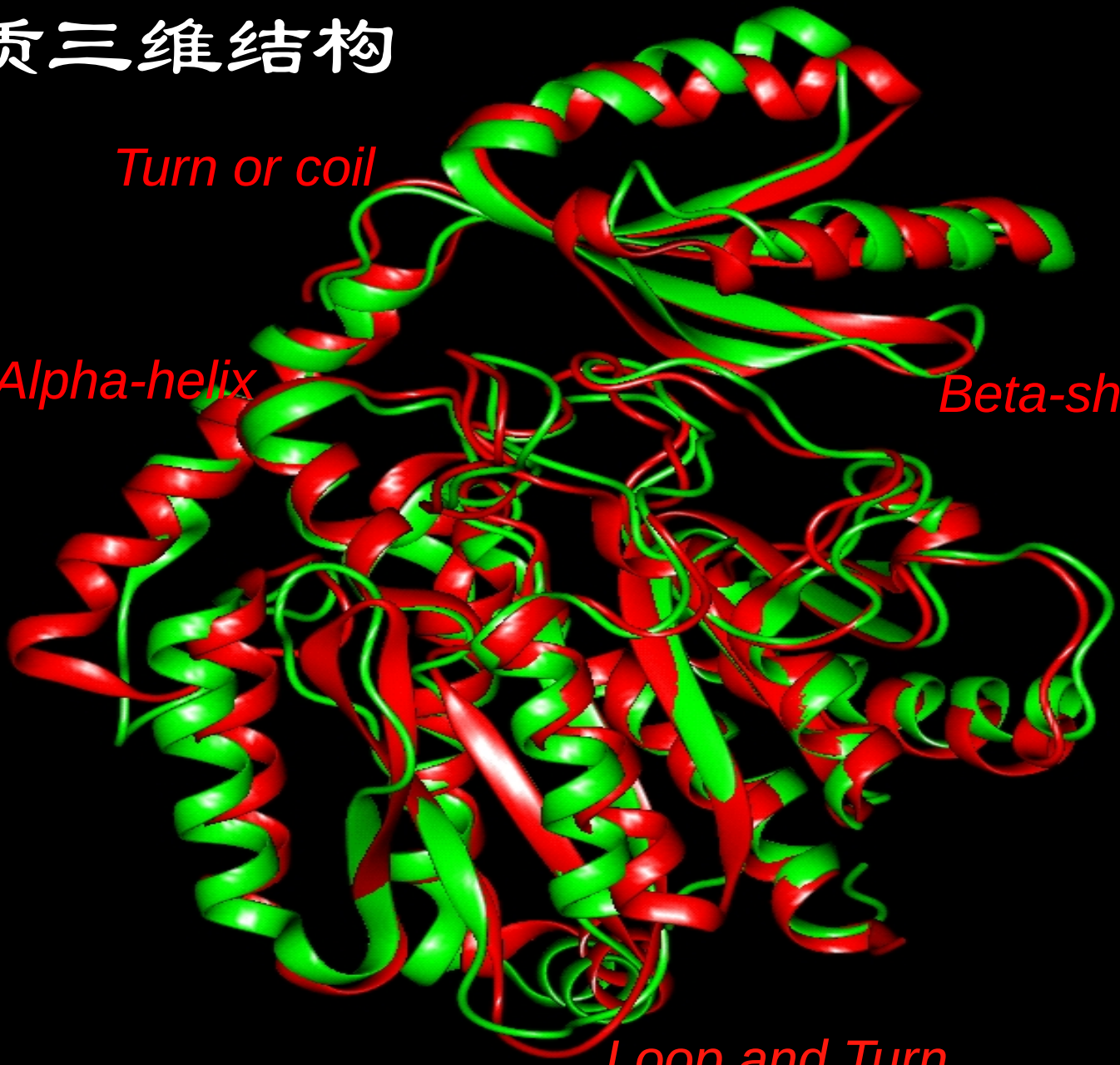


蛋白质的结构表示

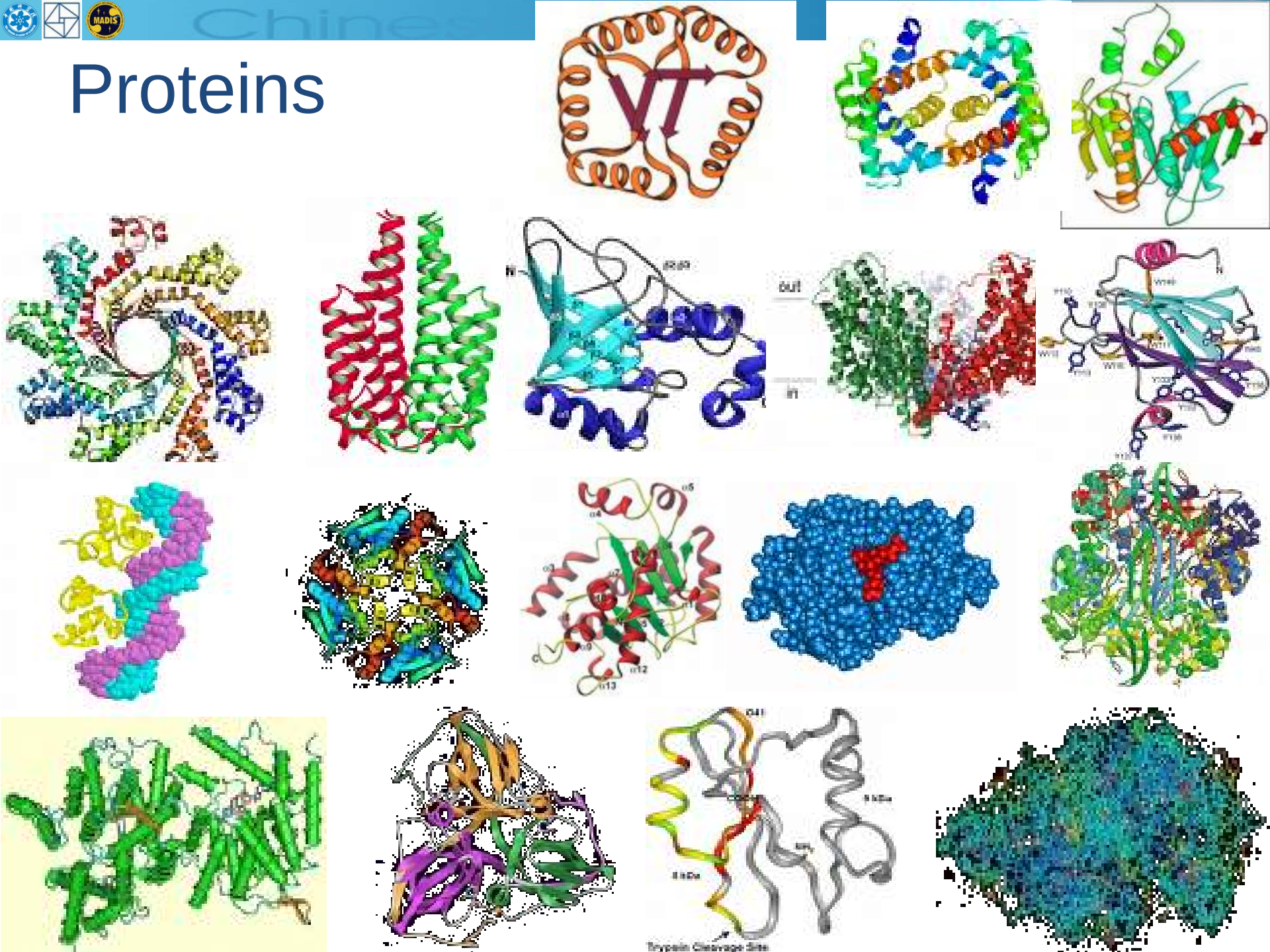
http://www.umass.edu/microbio/rasmol/



蛋白质三维结构



Proteins



蛋白质分类

Protein types

Type	Function	Examples
Structural	Give shape and structure to cell or organelles	Actin Tubulin
Enzymes	Catalyse biological reactions	Trypsin Adenylate cyclase
Receptors	Bind to other molecules and transmit signal	Glutamate R. Steroid R.
Other functional proteins	Have specific functions	Antibodies Nuclear factors Neuropeptides

Structural Classification Of Proteins

statistics from August, 2003

7 CLASSES

(a,b,a/b,a+b…)

800 FOLDS

domain structures

1,294 SUPERFAMILIES

possible evolutionary relationship

2,327 FAMILIES

strong sequence homology

54,745 DOMAINS

Murzin, AG, SE Brenner, T Hubbard, and C Chothia. "SCOP: A Structural Classification of Proteins Database for the Investigation of Sequences and Structures." *J Mol Biol.* 247, no. 4 (7 April 1995): 536-40.

CATH classification