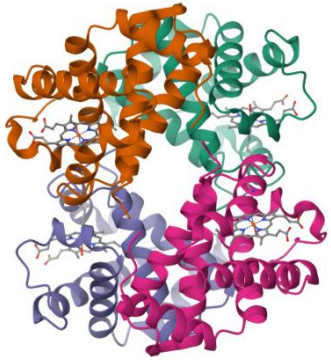
生物数据处理技术

生命健康信息科学与工程学院

解增言

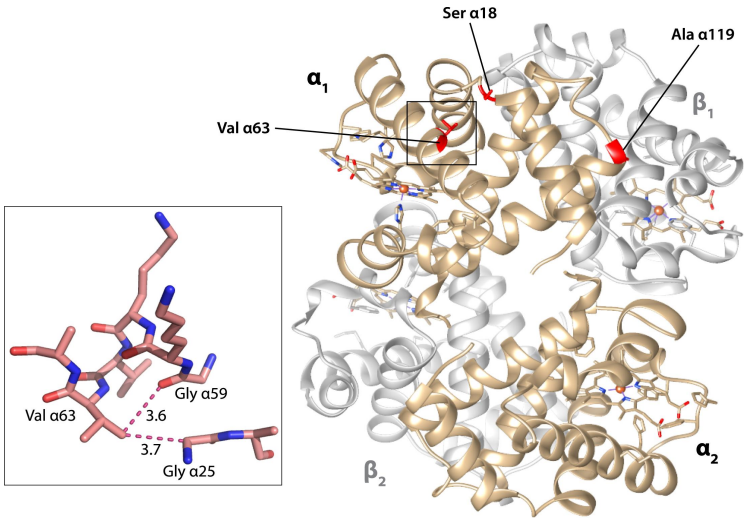
斑头雁（*Anser indicus*）



灰雁（*Anser anser*）

>ACT80359.1 hemoglobin alpha A subunit [Anser indicus]
MVLSAADKTNVKGVFSKI**S**GHAEEYGAETLERMFTAYPQTKTYFPHFDLQHGSAQIKAHGKKV**V**AALVEA
VNHIDDIAGALSKLSDLHAQKLRVDPVNFKFLGHCFLVVVAIHHPSALT**A**EVHASLDKFLCAVGTVLTAK
YR

>ACT80855.1 hemoglobin alpha A subunit [Anser anser]
MVLSAADKTNVKGVFSKI**G**GHAEEYGAETLERMFTAYPQTKTYFPHFDLQHGSAQIKAHGKKV**A**AALVEA
VNHIDDIAGALSKLSDLHAQKLRVDPVNFKFLGHCFLVVVAIHHPSALT**P**EVHASLDKFLCAVGTVLTAK
YR



教材与参考书

教材:

解增言，

Linux与生物信息学数据处理.

自编讲义

参考书1:

鸟哥，

鸟哥的Linux私房菜

基础学习篇（第四版）.

人民邮电出版社，2018

参考书2:

刘遄，

Linux就该这么学（第2版）.

人民邮电出版社，2021

课程资源

- 请查看雨课堂课件

考核方式

- **平时成绩（100分×60%）：**

  雨课堂（课前预习、课堂练习）：20分

  课后作业：20分（每次2分，共10次）

  随堂测试：20分（每次5分，共4次）

  课程网站学习情况：20分

  学银慕课学习情况：20分

  缺课：每次-2分

  附加任务：教材找错（每处错误确认后+1分，第一个找到的同学有效，最高至平时成绩满分）

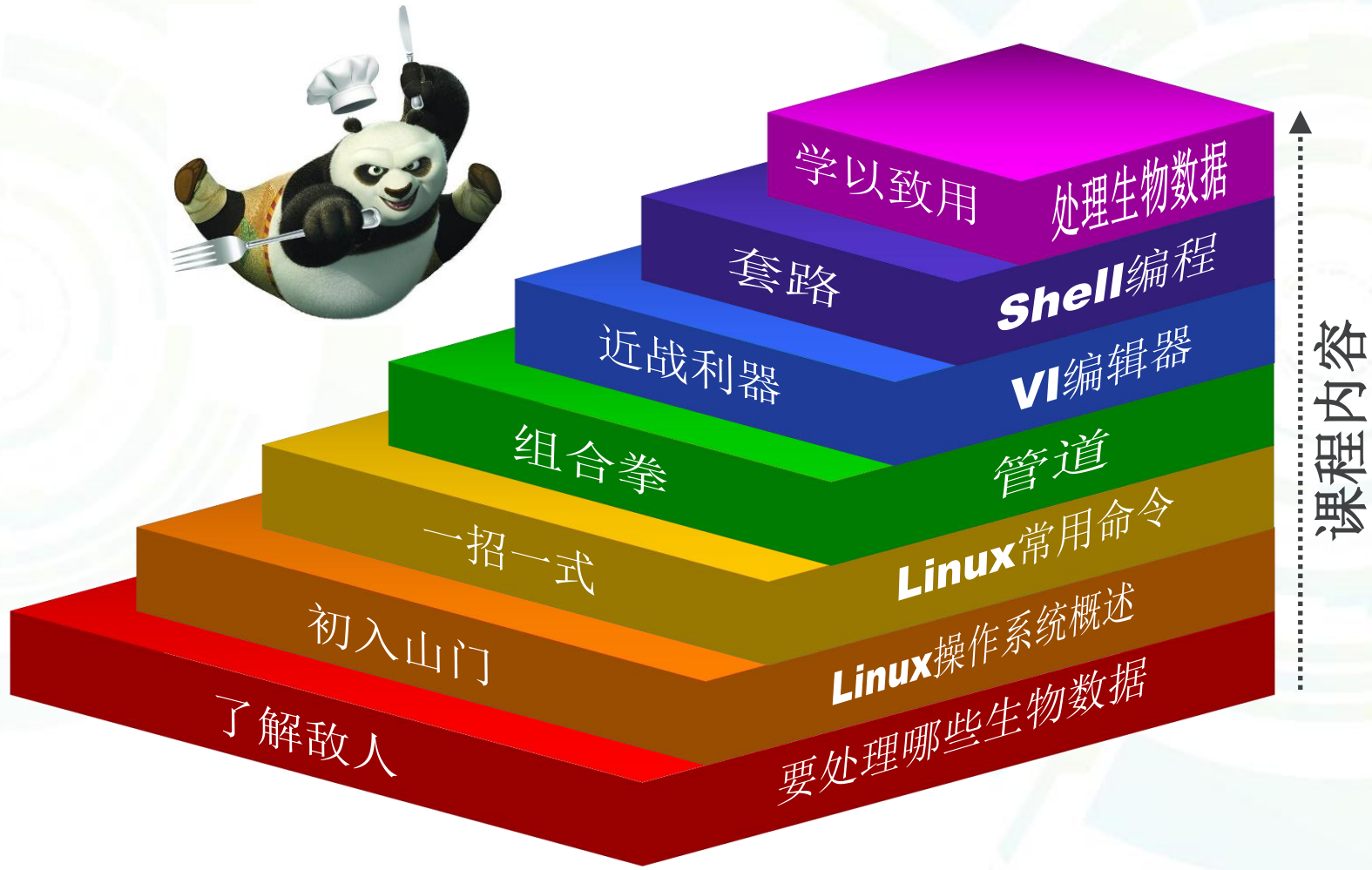
- **期末考试（100分×40%）：**

  单选、多选、填空、简答（题目来自课程网站www.linuxstudio.cn题库）

课程简介

- Linux与生物信息数据处理课程主要内容是学习利用Linux命令、管道及Shell编程，处理生物学数据（主要是分子生物学数据）。

课程内容


了解敌人
初入山门
一招一式
组合拳
近战利器
套路
学以致用
要处理哪些生物数据
Linux操作系统概述
Linux常用命令
管道
VI编辑器
Shell编程
处理生物数据
课程内容

课时安排

2学时 课程简介、生物中的信息（数据）

2学时 Linux系统概述、Linux文件系统

10学时 Linux常用命令

4学时 VI编辑器

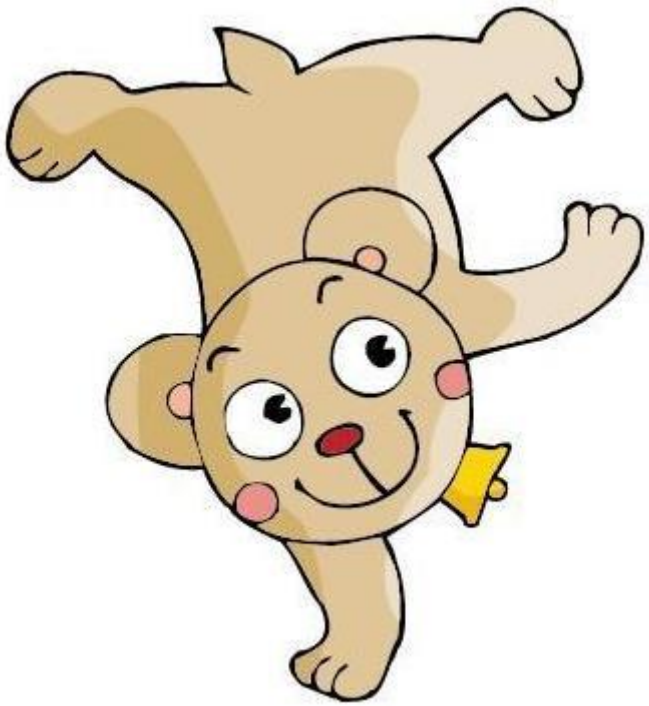
8学时 Shell编程基础

4学时 应用实例

上课形式

翻转课堂

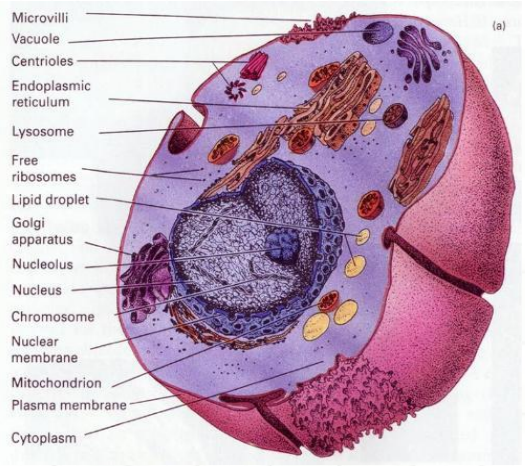
- ✓ 课前：观看教学视频，预习教材文字内容及课件，提前练习命令
- ✓ 课内：练习、作业、测试、解答问题、重点讲解、总结
- ✓ 课后：作业、复习、练习、测试



怎样获得一个好成绩

- 按时**预习**：上课前完成指定的内容，完成雨课堂推送的课件预习；
- 完成**课堂任务**：上课时专心听讲，及时完成课堂练习和课堂作业；
- 按时完成**课后作业**：根据要求，按时完成服务器上的课后作业；
- **学银慕课**：按时完成慕课教学视频学习及测试，积极参加慕课平台的讨论；
- **课程网站**：按时完成课程网站的文字材料学习及练习，多做自主测试，参加课程阶段测试；按时上课：按时签到，不迟到，不旷课。
- **最重要的一点：多练习、理解Linux命令，并用来解决实际问题**。

第1章

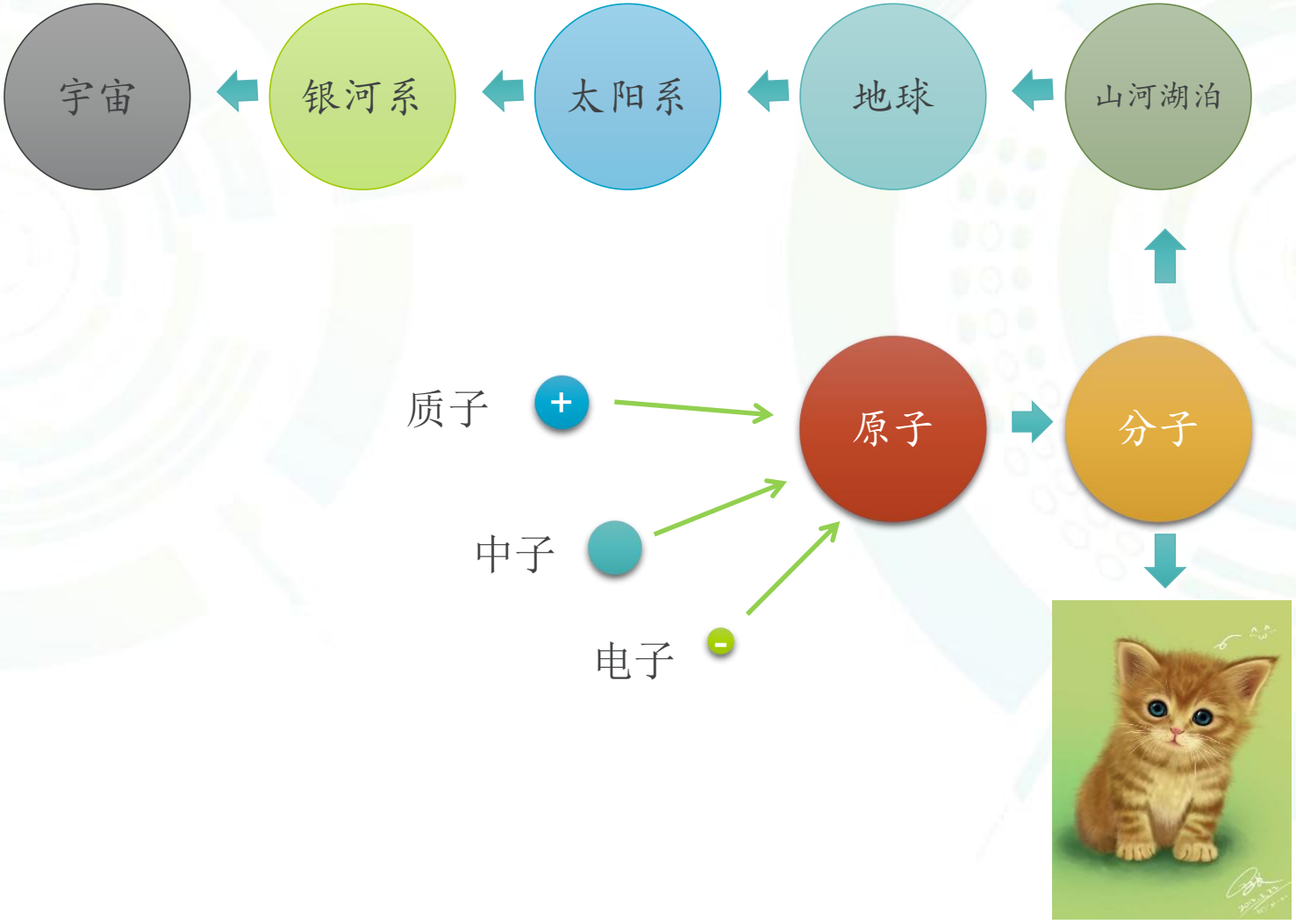


生物学中的数据

解增言

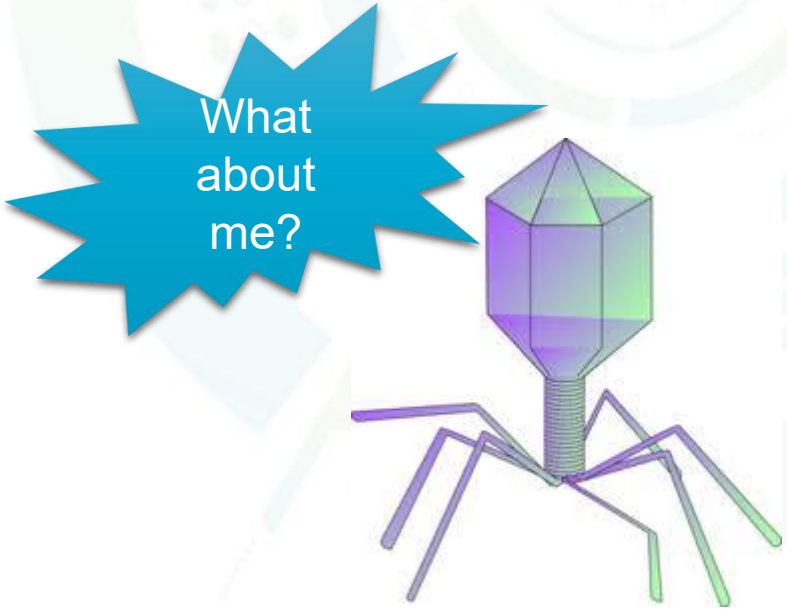
生命健康信息学院

从宇宙到生物



生物的特征（与非生物的区别）

- 细胞是结构和功能的基本单位
- 新陈代谢
- 生长与发育
- 应激性
- 繁殖
- 遗传和变异
- 适应与进化
- 与外界进行物质交换
- 呼吸



生物类群

- **原核生物**
  - 真细菌
  - 古细菌
- **真核生物**
  - 原生生物
  - 真菌
  - 动物
  - 植物

多选题 1分

设置

下面哪些是原核生物?

A 酵母

B 蓝藻

C 大肠杆菌

D 极端嗜热菌

提交

生物学科的层次

- 生态学、进化生物学
- 动物学、植物学、微生物学
- 发育生物学、解剖学、遗传学
- 细胞生物学
- 生物化学
- 分子生物学、分子遗传学

生物学数据的类型

- 生态学和生物多样性数据
- 分子生物学数据
- 生理生化数据
- ……

生物大分子

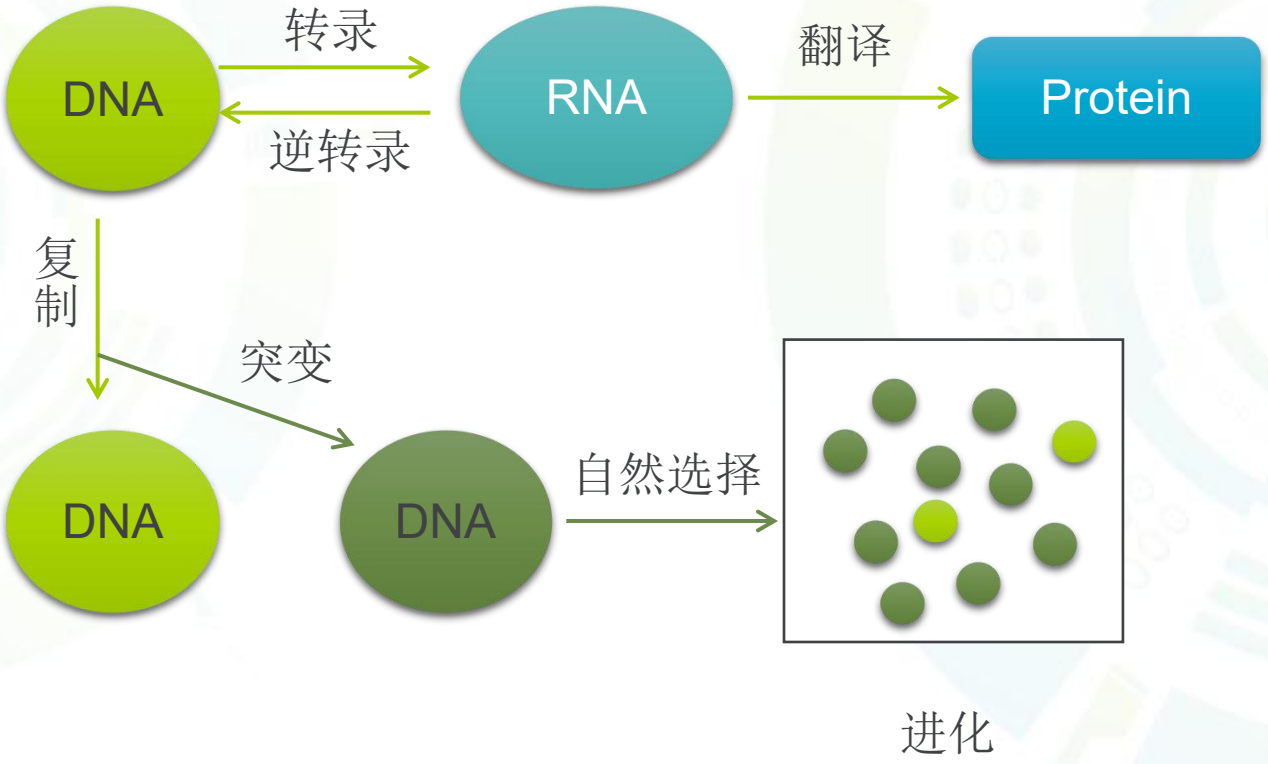
- 核酸

  DNA（脱氧核糖核酸）

  RNA（核糖核酸）

- 蛋白质
- 多糖

生物信息传递



分子生物学数据的类型

- 核酸一二三级结构
- 蛋白质一二三四级结构
- 基因表达数据
- 基因组数据
- 基因/蛋白质相互作用数据
- 基因/蛋白质修饰（甲基化等）数据

组成生物大分子的单位：生物单分子

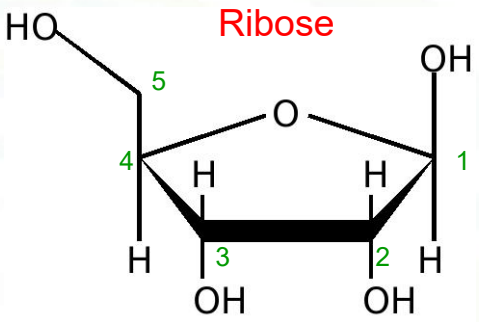
- 主要生物单分子：

  核糖核苷酸

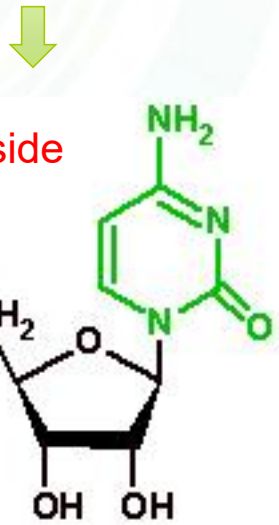
  脱氧核糖核苷酸

  氨基酸

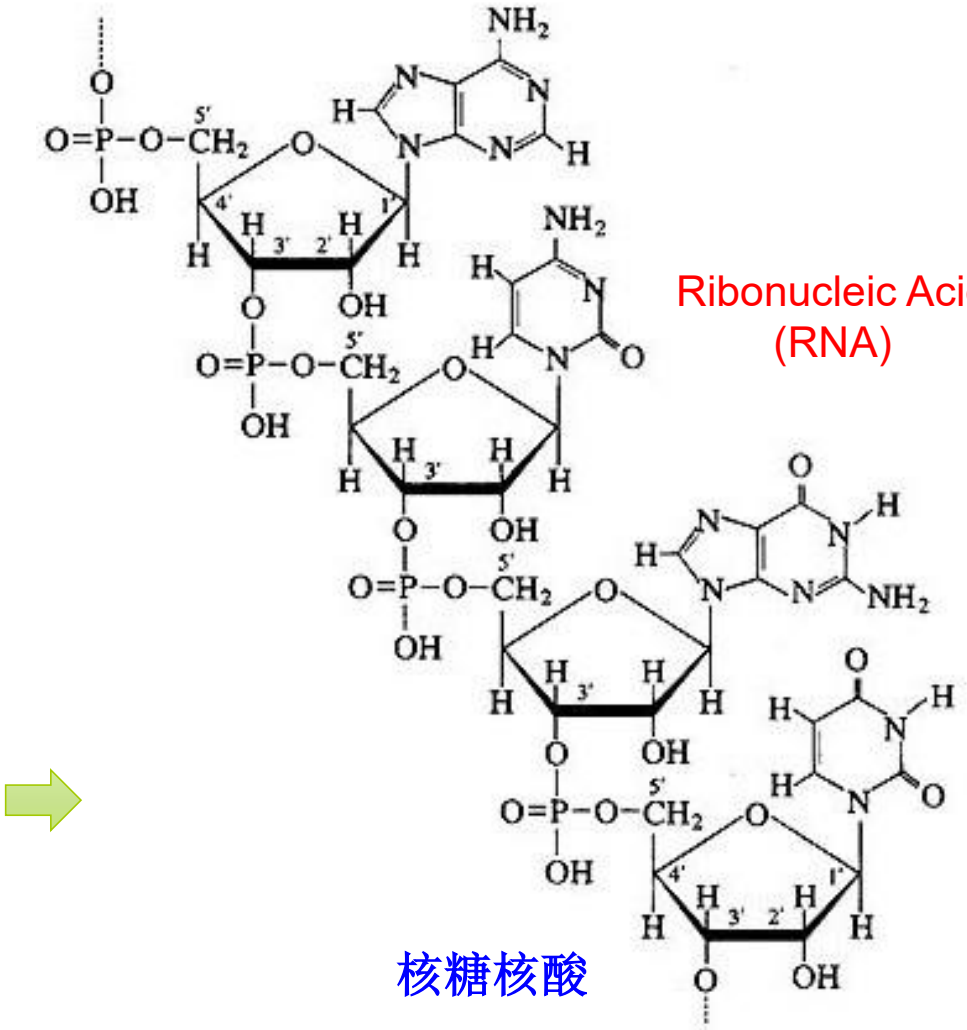
RNA构成



核糖

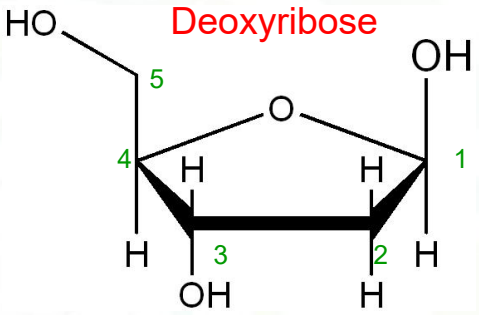


核糖核苷酸



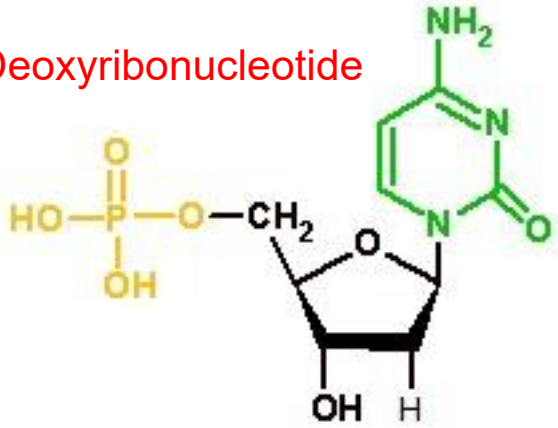
核糖核酸

DNA构成

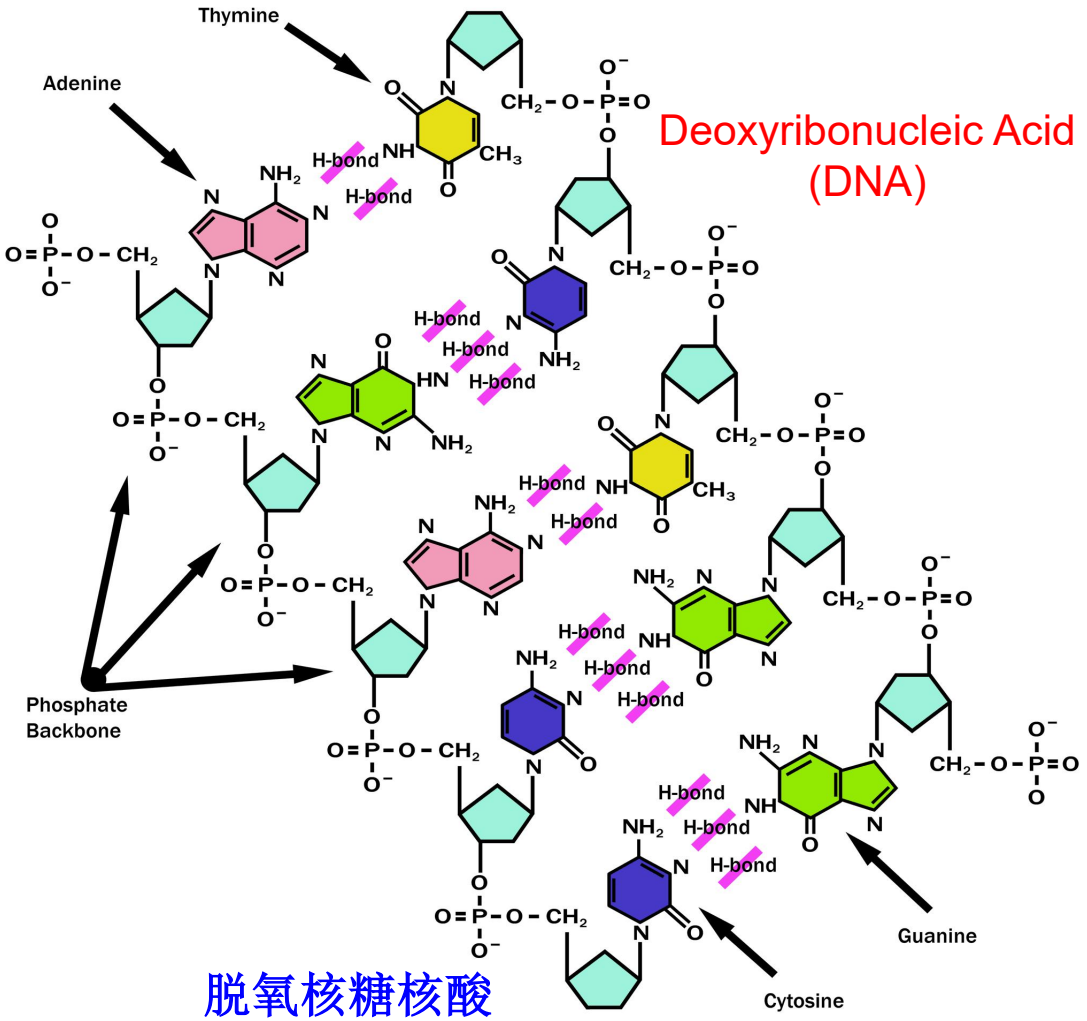


脱氧核糖

Deoxyribonucleotide



脱氧核糖核苷酸



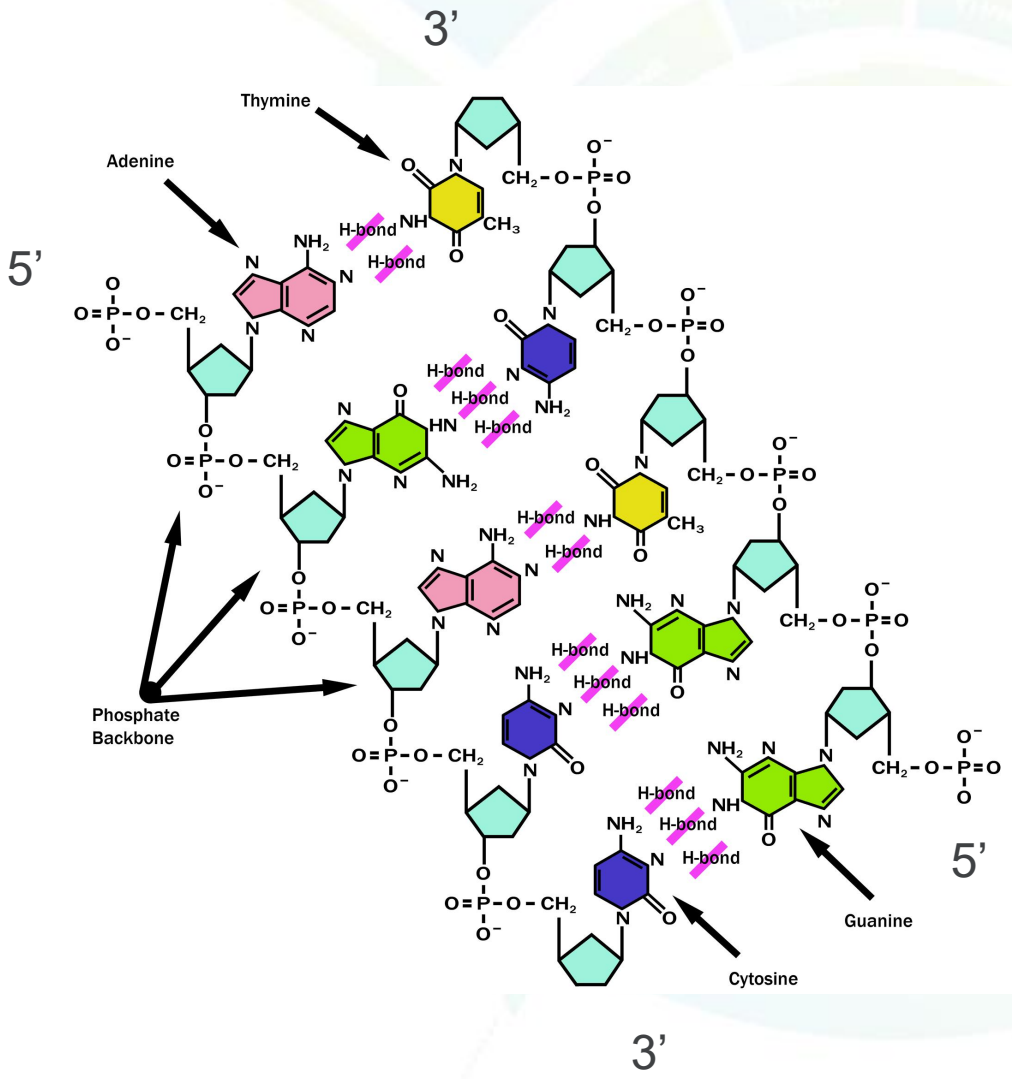
脱氧核糖核酸

DNA简写

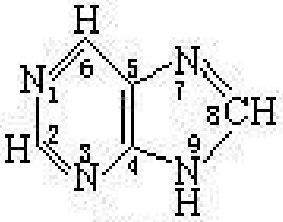
5’ AGACC 3’

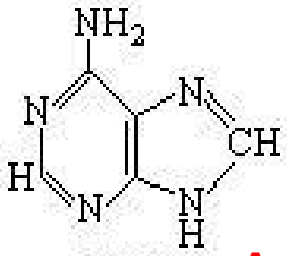
3’ TCTGG 5’

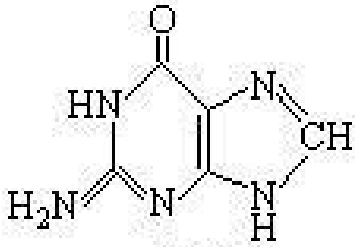
5’ ATGGCCAGGTAA 3’

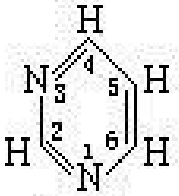


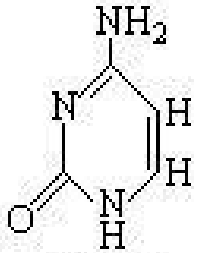
常见碱基

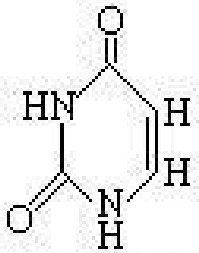

嘌呤

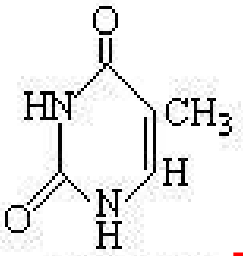

腺嘌呤 A

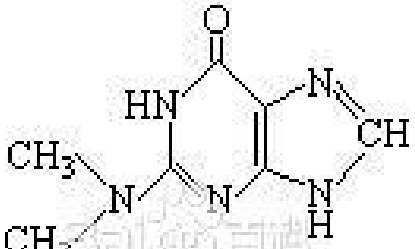

鸟嘌呤 G

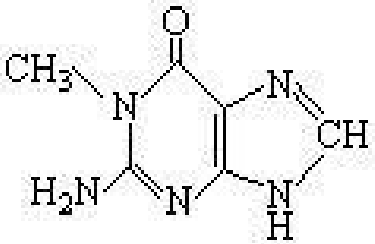

嘧啶

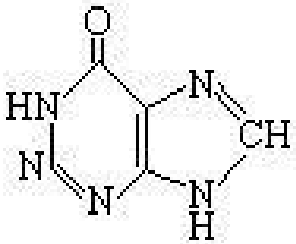

胞嘧啶 C

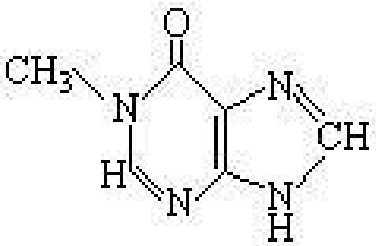

尿嘧啶 U


胸腺嘧啶 T


N^2-二甲基鸟嘌呤


1-甲基鸟嘌呤


次黄嘌呤


1-甲基次黄嘌呤

单选题 1分

设置

关于DNA和RNA，下面说法正确的是：

A 尿嘧啶只存在于DNA中

B RNA中没有胸腺嘧啶

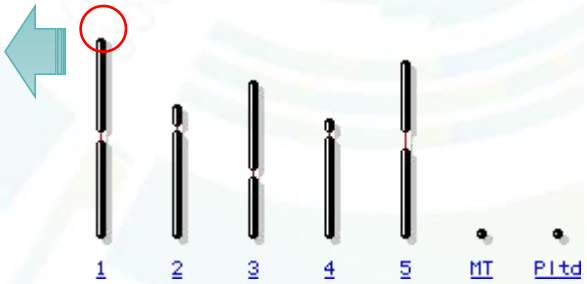
C 组成RNA的碱基只有4种

D RNA比DNA稳定

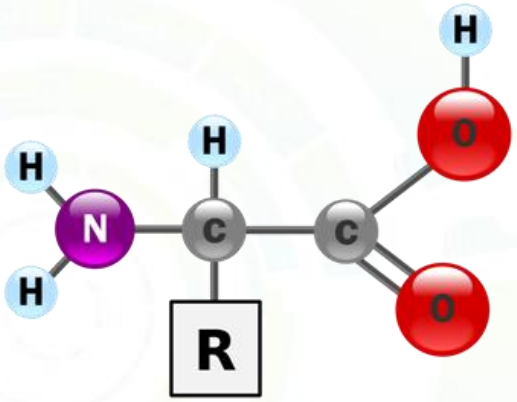
提交

拟南芥第一条染色体部分序列

- >Chr1 CHROMOSOME dumped from ADB: Feb/3/09 16:9; last updated: 2007-12-20
  CCCTAAACCCTAAACCCTAAACCCTAAACCTCTGAATCCTTAATCCC
  TAAATCCCTAAATCTTTAAATCCTACATCCATGAATCCCTAAATACC
  TAATTCCCTAAACCCGAAACCGGTTTCTCTGGTTGAAAATCATTGTG
  TATATAATGATAATTTTATCGTTTTTATGTAATTGCTTATTGTTGTG
  TGTAGATTTTTTAAAAATATCATTTGAGGTCAATACAAATCCTATTT
  CTTGTGGTTTTCTTTCCTTCACTTAGCTATGGATGGTTTATCTTCAT
  TTGTTATATTGGATACAAGCTTTGCTACGATCTACATTTGGGAA

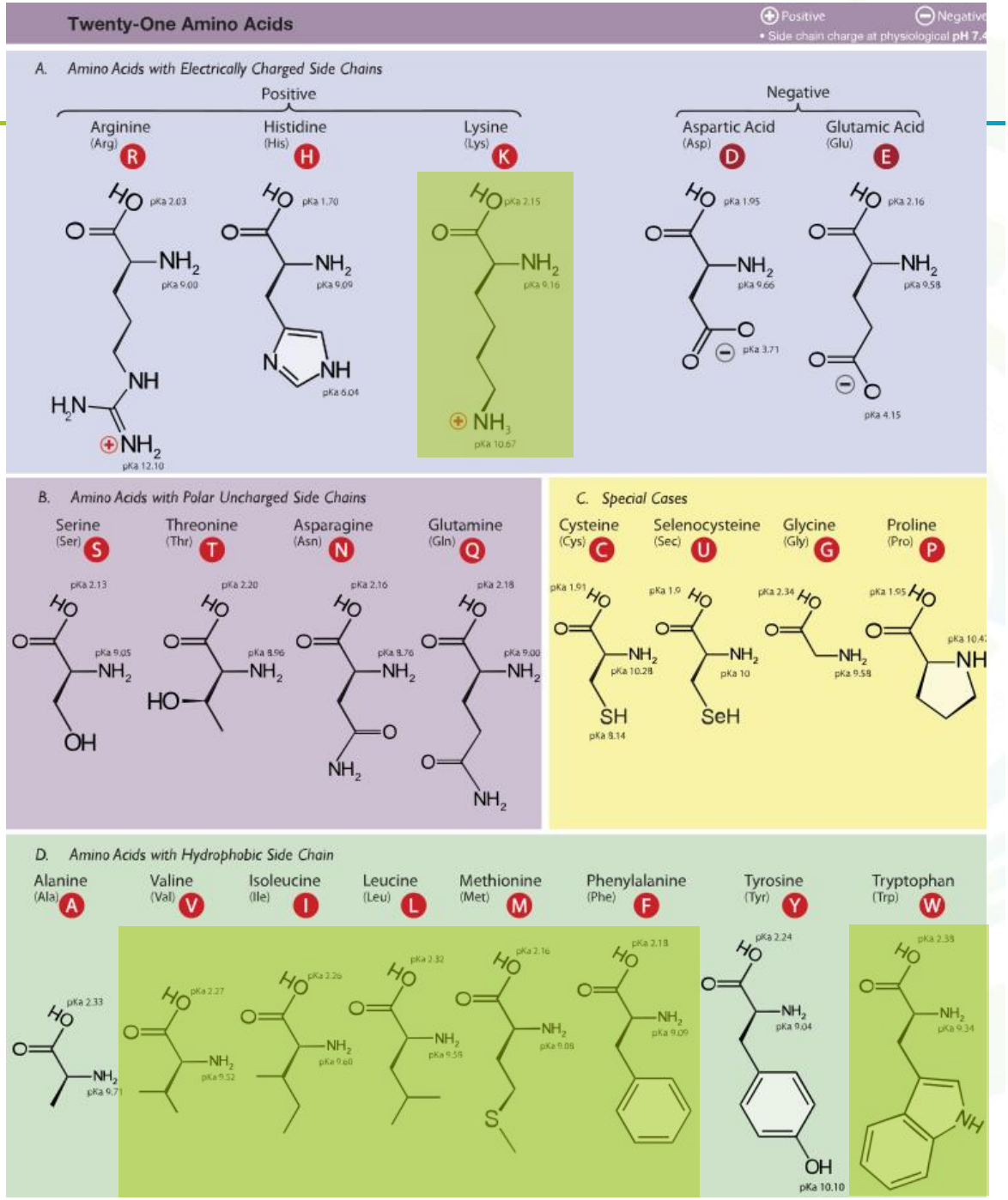


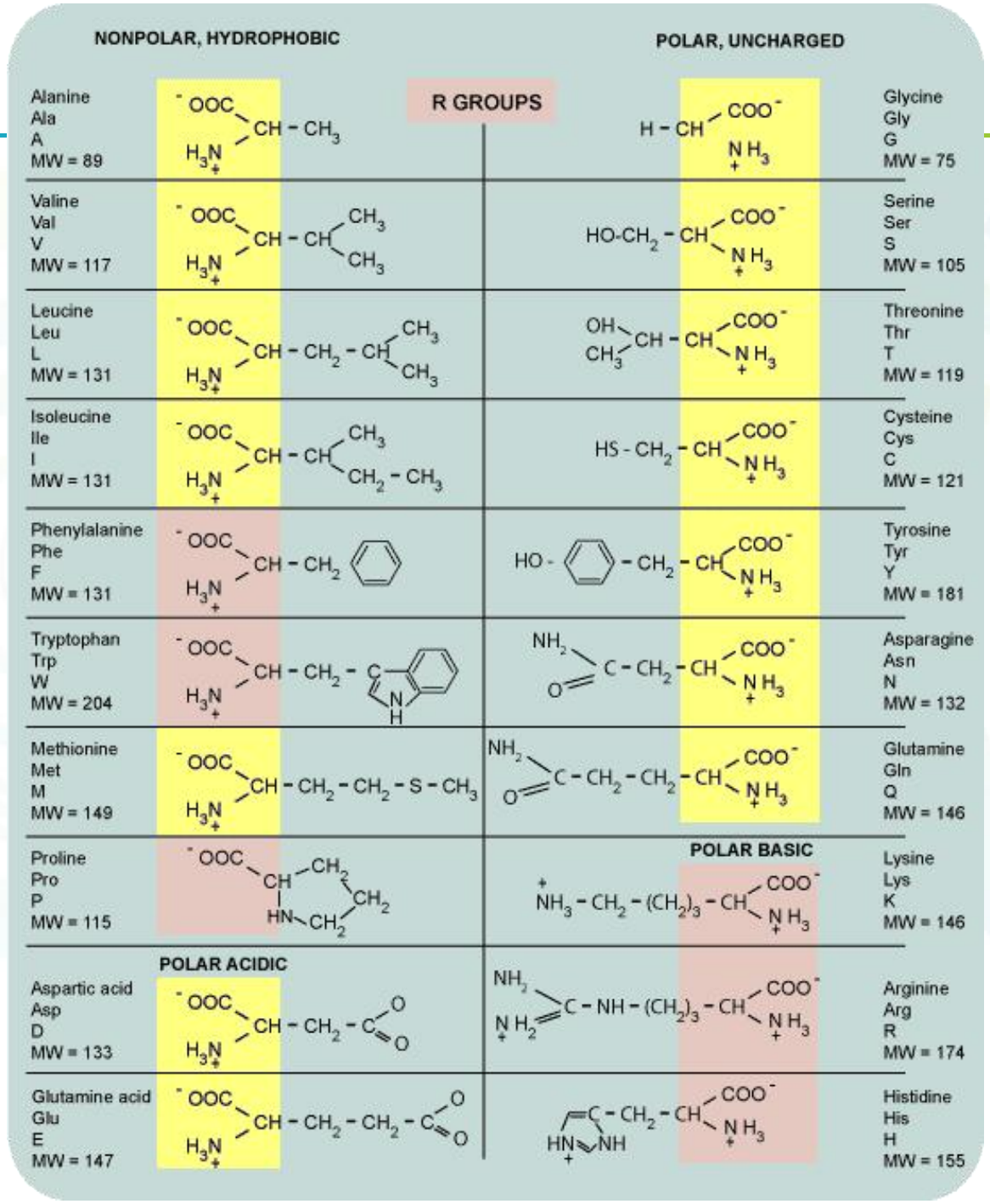
蛋白质构成



氨基酸结构

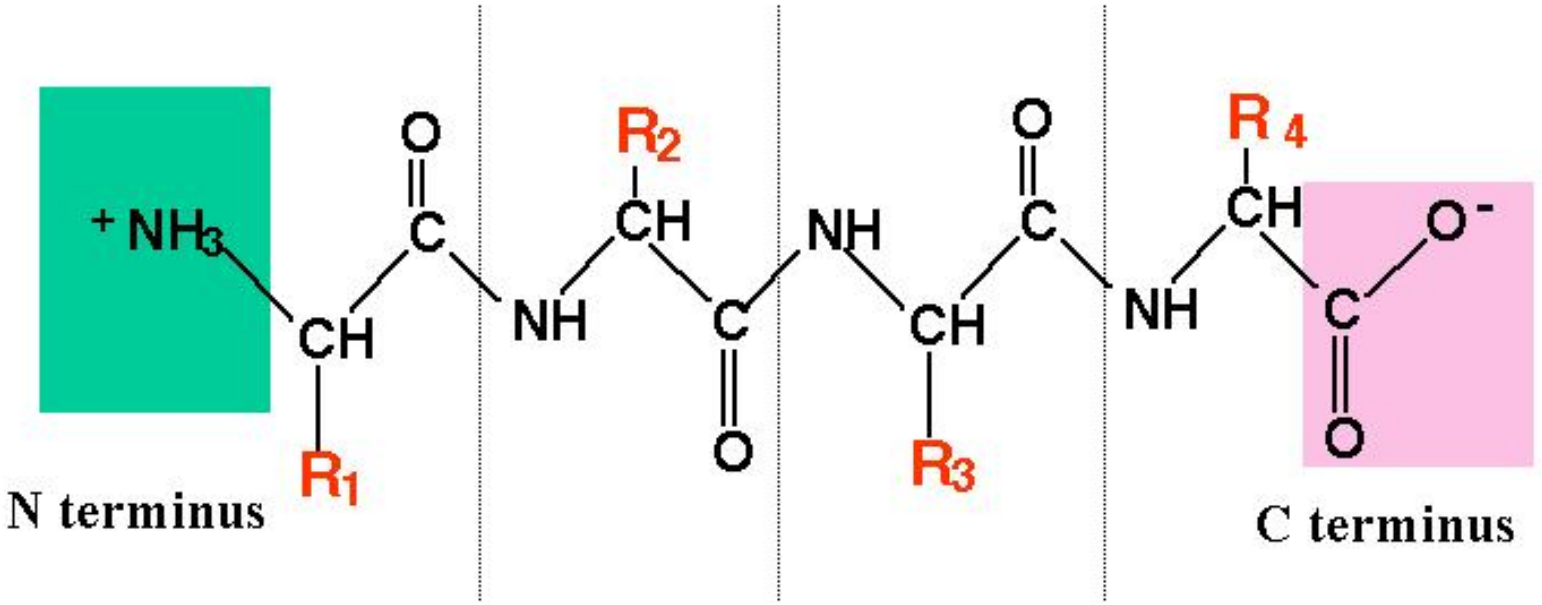
21种氨基酸




NONPOLAR, HYDROPHOBIC
POLAR, UNCHARGED
R GROUPS
Alanine
Ala
A
MW = 89
Glycine
Gly
G
MW = 75
Valine
Val
V
MW = 117
Serine
Ser
S
MW = 105
Leucine
Leu
L
MW = 131
Threonine
Thr
T
MW = 119
Isoleucine
Ile
I
MW = 131
Cysteine
Cys
C
MW = 121
Phenylalanine
Phe
F
MW = 131
Tyrosine
Tyr
Y
MW = 181
Tryptophan
Trp
W
MW = 204
Asparagine
Asn
N
MW = 132
Methionine
Met
M
MW = 149
Glutamine
Gln
Q
MW = 146
Proline
Pro
P
MW = 115
POLAR BASIC
Lysine
Lys
K
MW = 146
POLAR ACIDIC
Aspartic acid
Asp
D
MW = 133
Arginine
Arg
R
MW = 174
Glutamine acid
Glu
E
MW = 147
Histidine
His
H
MW = 155

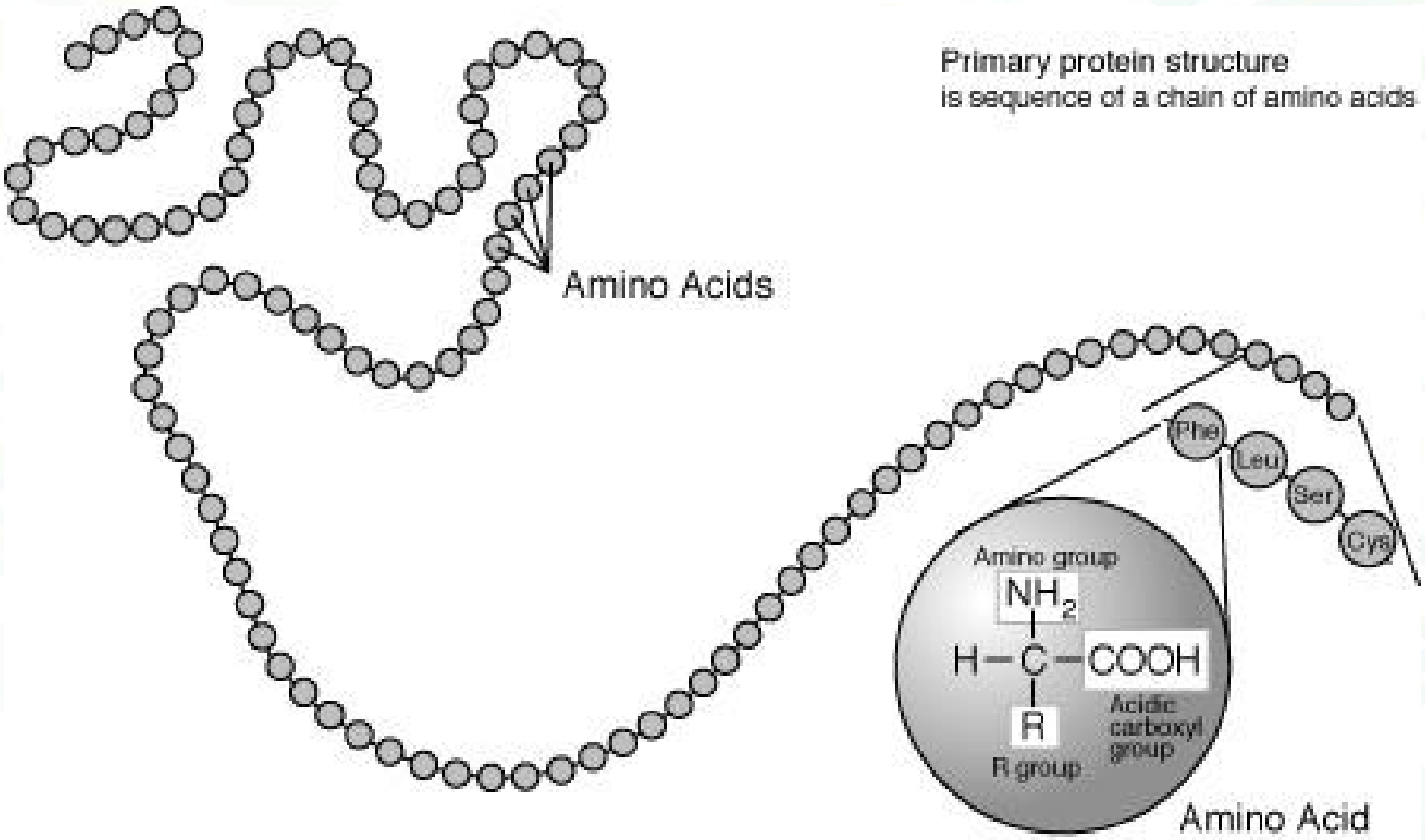
多肽结构

Peptide = chain of amino acids



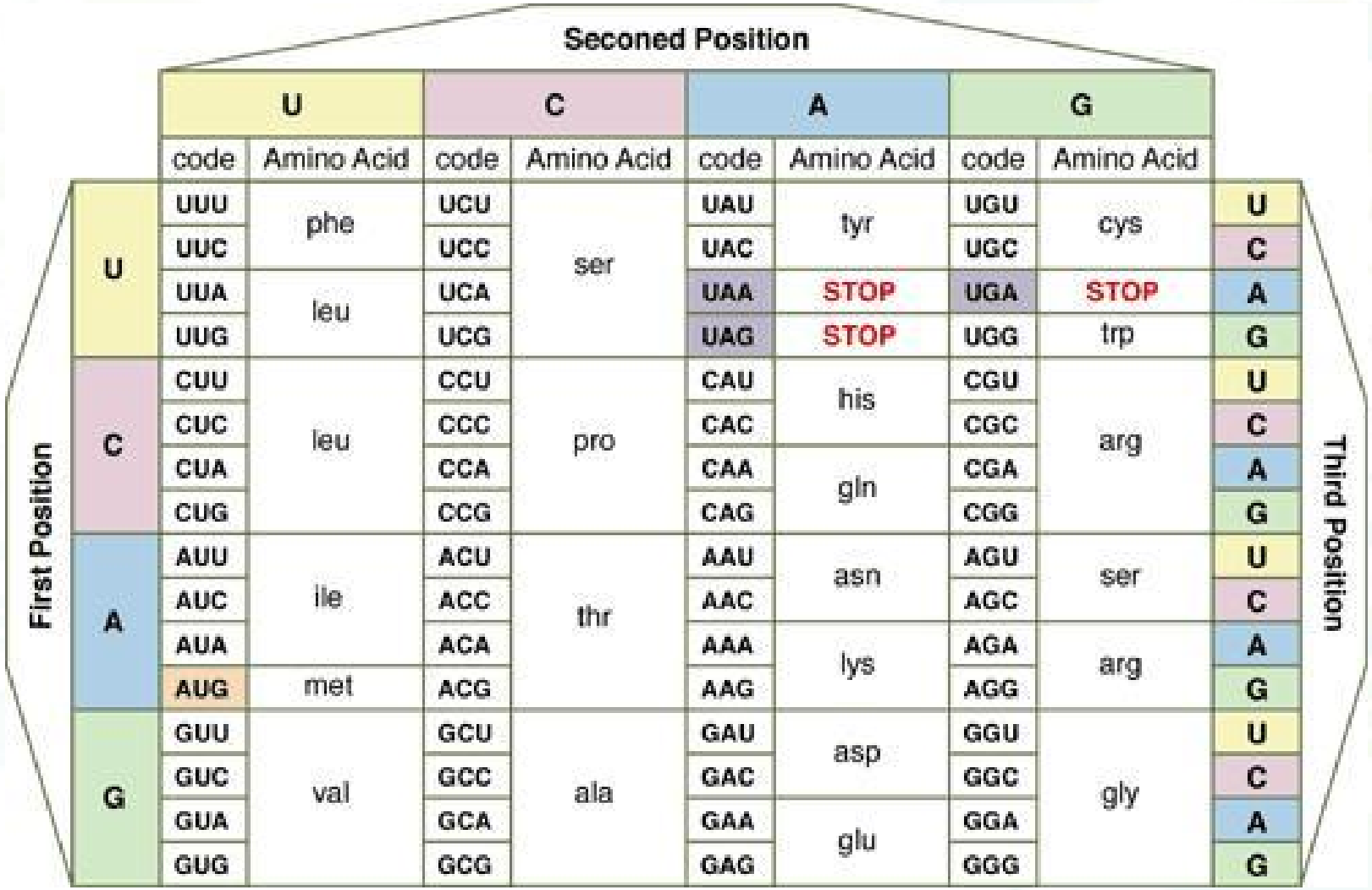
polypeptide chain

蛋白质一级结构


Primary protein structure
is sequence of a chain of amino acids
Amino Acids
Phe
Leu
Ser
Cys
Amino group
NH2
H—C—COOH
R
R group
Acidic carboxyl group
Amino Acid

Asp	D	Aspartic acid	Ile	I	Isoleucine
Thr	T	Threonine	Leu	L	Leucine
Ser	S	Serine	Tyr	Y	Tyrosine
Glu	E	Glutamic acid	Phe	F	Phenylalanine
Pro	P	Proline	His	H	Histidine
Gly	G	Glycine	Lys	K	Lysine
Ala	A	Alanine	Arg	R	Arginine
Cys	C	Cysteine	Trp	W	Tryptophan
Val	V	Valine	Gln	Q	Glutamine
Met	M	Methionine	ASN	N	Asparagine

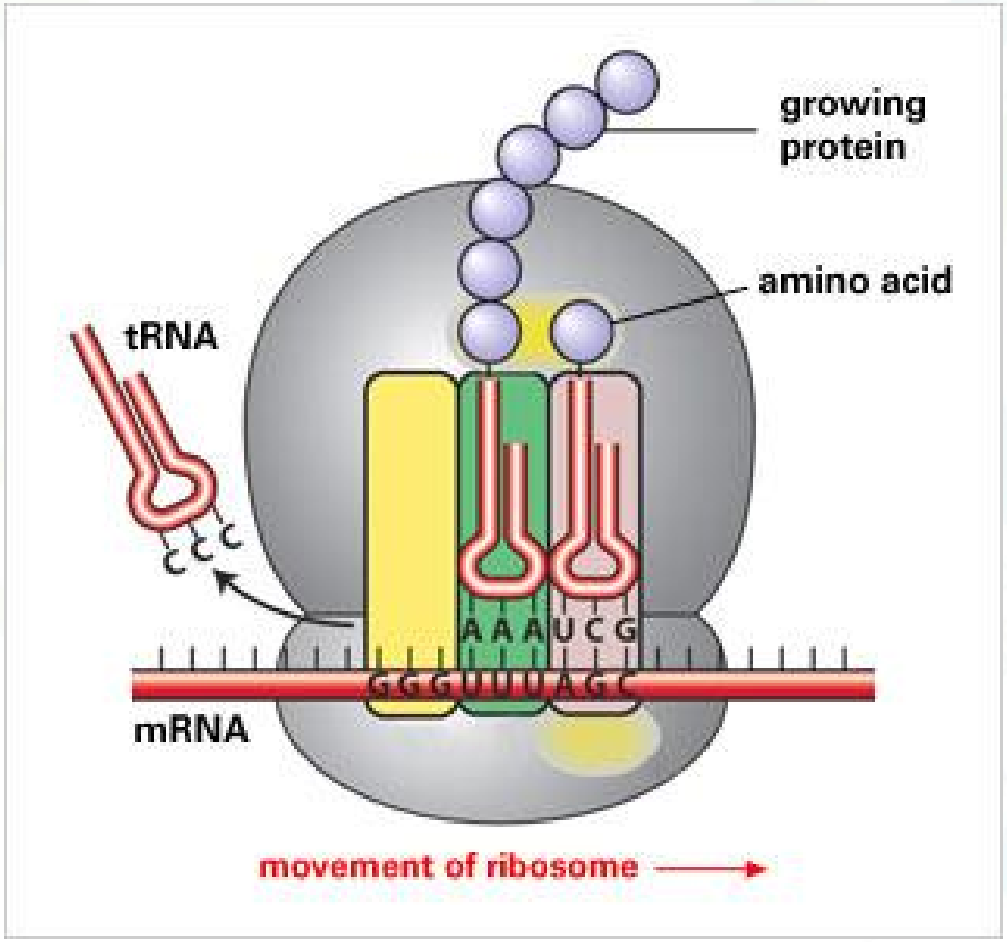
氨基酸编码



First Position	U				C				A				G				Third Position
	U	code	Amino Acid		U	code	Amino Acid		U	code	Amino Acid		U	code	Amino Acid		
U		UUU	phe			UCU	ser			UAU	tyr			UGU	cys		U
		UUC				UCC				UAC				UGC			C
		UUA	leu			UCA				UAA	STOP			UGA	STOP		A
		UUG				UCG				UAG	STOP			UGG	trp		G
C		CUU	leu			CCU	pro			CAU	his			CGU	arg		U
		CUC				CCC				CAC				CGC			C
		CUA				CCA				CAA	gln			CGA			A
		CUG				CCG				CAG				CGG			G
A		AUU	ile			ACU	thr			AAU	asn			AGU	ser		U
		AUC				ACC				AAC				AGC			C
		AUA				ACA				AAA	lys			AGA	arg		A
		AUG	met			ACG				AAG				AGG			G
G		GUU	val			GCU	ala			GAU	asp			GGU	gly		U
		GUC				GCC				GAC				GGC			C
		GUA				GCA				GAA	glu			GGA			A
		GUG				GCG				GAG				GGG			G

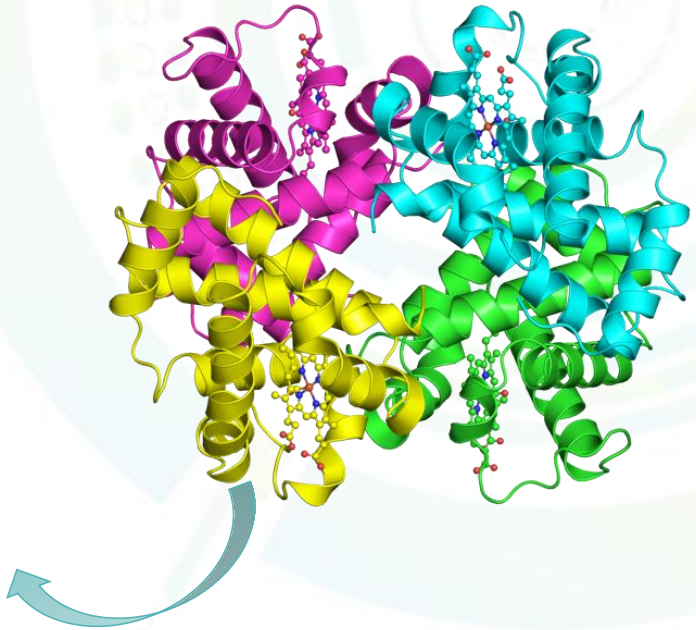
概念：密码子简并性

蛋白质合成


growing protein
amino acid
tRNA
C C C
A A A U C G
G G G U U U A G C
mRNA
movement of ribosome

人类血红蛋白α亚基序列

- >gi|4504345|ref|NP_000508.1| hemoglobin subunit alpha [Homo sapiens]
  MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPT
  TKTYFPHFDLSHGSAQVKGHGKKVADALTNAVAHVDDMP
  NALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAE
  FTPAVHASLDKFLASVSTVLTSKYR



多选题 1分

设置

关于蛋白质，下面哪些描述不正确？

A 蛋白质是在线粒体上合成的

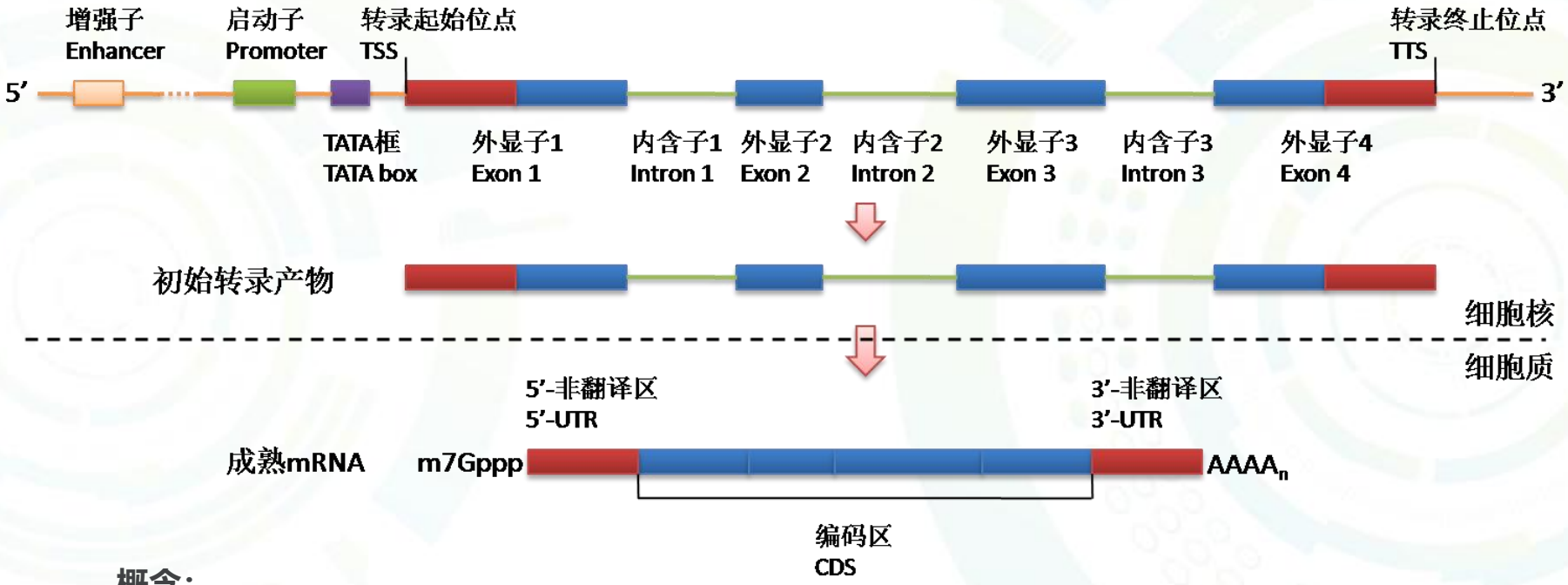
B 碱性氨基酸在pH=7时带正电荷

C 在水溶液中，球蛋白的疏水氨基酸一般位于蛋白质表面

D 64种密码子中，有3个是终止密码子

提交

基因结构（真核生物）



概念：

TSS（转录起始位点）

TTS（转录终止位点）

CDS（编码区）

UTR（非翻译区）

Exon（外显子）

Intron（内含子）

基因结构示例

>NC_003076.8:19315061-19315975 Arabidopsis thaliana chromosome 5 sequence

ATGGAACGTGGAGGCTTCCATGGCTACCGCAAGCTGTCCGTGAACAACACCACTCCTTCTCCACCAGgta
gtgccattctctatacccccctcttttcacaggctctcttcatttcagttgcatgcgaaaccattctctgc
aatccctccattgtcatgtctgtactcttttcatgacgaacagttaatgaaatagcttttcaatcttata
aaccgcgcatgcagacgtcatcgaagccattatgcactaaaacttccatttttcttatttttgttagGAT
TAGCAGCGAATTTTCTGATGGCAGAGGGCAGTATGCGTCCTCCAGAATTCAACCAGCCTAACAAAACCAG
TAATGGTGGTGAGGAGGAGTGCACGGTGAGGGAGCAAGACAGGTTCATGCCTATTGCCAACGTGATACGG
ATCATGCGGAGGATCTTACCTGCTCACGCCAAGATCTCAGATGACTCCAAGGAGACGATCCAAGAGTGTG
TTTCGGAGTACATCAGCTTCATAACAGGGGAGGCTAATGAGCGGTGCCAGCGGGAACAGCGCAAGACCAT
CACTGCTGAGGACGTCTTGTGGGCAATGAGCAAGCTCGGTTTTGATGACTACATCGAACCCCTCACGTTG
TACCTCCACCGCTACAGAGAGTTGGAAGGTGAAAGAGGGGTTAGCTGCAGTGCTGGGTCCGTTAGTATGA
CCAACGGCTTGGTGGTCAAGAGGCCTAATGGGACCATGACCGAGTATGGAGCCTACGGGCCTGTGCCAGG
GATTCACATGGCGCAGTACCATTATCGTCATCAGAACGGGTTTGTTTTCAGTGGTAACGAACCTAATTCT
AAGATGAGTGGTTCATCTTCAGGAGCAAGTGGCGCCAGAGTTGAAGTATTTCCGACTCAACAACATAAGT
ACTGA

拟南芥*LEAFY COTYLEDON 1*基因的编码区（CDS）包含一个内含子，上面序列部分小写字母为内含子，内含子一般以**GT**开始，以**AG**结束

TAIR数据库拟南芥基因组序列

Download - Araport11 blastsets

- Araport11_3_utr_20220505.gz 3,812 KB 2022-05-05
- Araport11_5_utr_20220505.gz 3,157 KB 2022-05-05
- Araport11_cdna_20220505.gz 19,275 KB 2022-05-05
- Araport11_cdna_20220505_representative_gene_model.gz 15,491 KB 2022-05-05
- Araport11_cds_20220505.gz 13,208 KB 2022-05-05
- Araport11_cds_20220505_representative_gene_model.gz 11,918 KB 2022-05-05
- Araport11_downstream_1000_20220505.gz 10,891 KB 2022-05-05
- Araport11_downstream_1000_translation_end_20220505.gz 10,606 KB 2022-05-05
- Araport11_downstream_3000_20220505.gz 28,985 KB 2022-05-05
- Araport11_downstream_3000_translation_end_20220505.gz 27,605 KB 2022-05-05
- Araport11_downstream_500_20220505.gz 5,812 KB 2022-05-05
- Araport11_downstream_500_translation_end_20220505.gz 5,820 KB 2022-05-05
- Araport11_intergenic_20220505.gz 16,834 KB 2022-05-05
- Araport11_intron_20220505.gz 11,152 KB 2022-05-05
- Araport11_pep_20220505.gz 8,441 KB 2022-05-05
- Araport11_pep_20220505_representative_gene_model.gz 8,009 KB 2022-05-05
- Araport11_seq_20220505.gz 29,101 KB 2022-05-05
- Araport11_seq_20220505_representative_gene_model.gz 24,593 KB 2022-05-05
- Araport11_upstream_1000_20220505.gz 10,882 KB 2022-05-05
- Araport11_upstream_1000_translation_start_20220505.gz 10,511 KB 2022-05-05
- Araport11_upstream_3000_20220505.gz 29,215 KB 2022-05-05
- Araport11_upstream_3000_translation_start_20220505.gz 27,591 KB 2022-05-05
- Araport11_upstream_500_20220505.gz 5,786 KB 2022-05-05
- Araport11_upstream_500_translation_start_20220505.gz 5,742 KB 2022-05-05

TAIR数据库（https://www.arabidopsis.org）的拟南芥基因组数据下载界面，注意其中的**UTR**（非翻译区）、**cDNA**（即mRNA）、**CDS**（编码区）、**InterGenic**（基因间序列）、**intron**（内含子）、**pep**（蛋白质，peptide）、**seq**（基因组核酸序列）的含义。

单选题 1分

设置

关于基因结构，下面哪个说法不正确？

A 只有真核生物的基因才有内含子

B 非编码区也可能有内含子

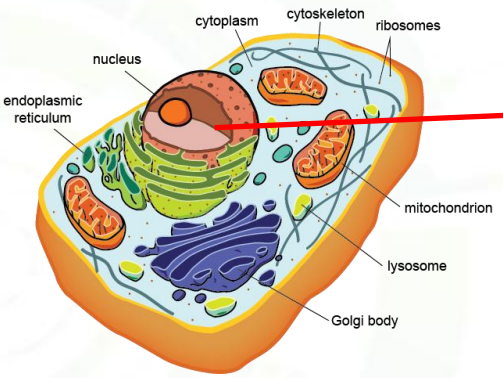
C 内含子的边界一般是GT-AG

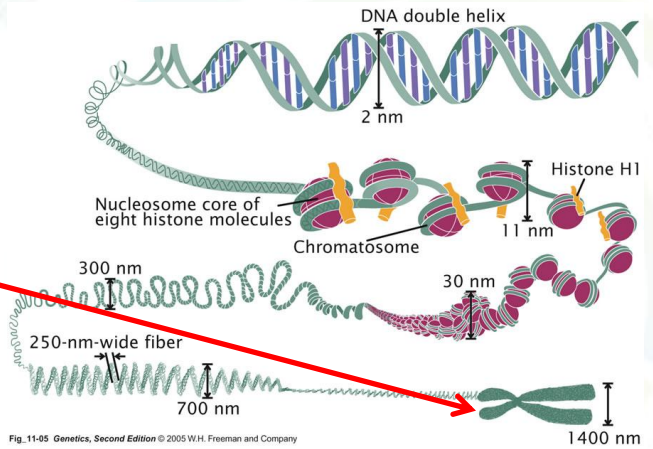
D 基因调控区一般位于基因上游

提交

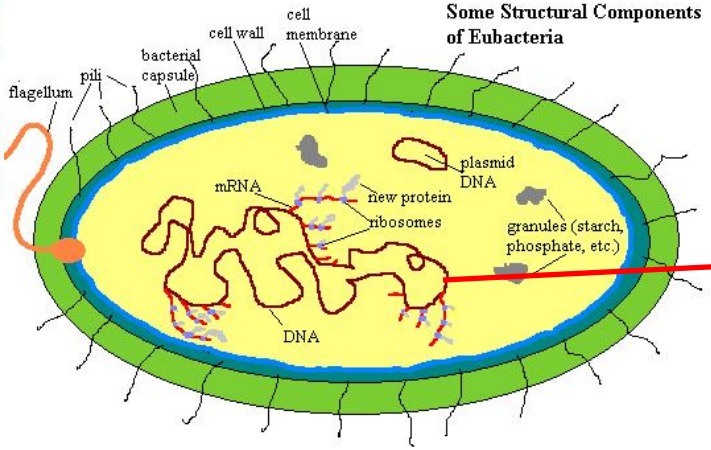
基因组结构

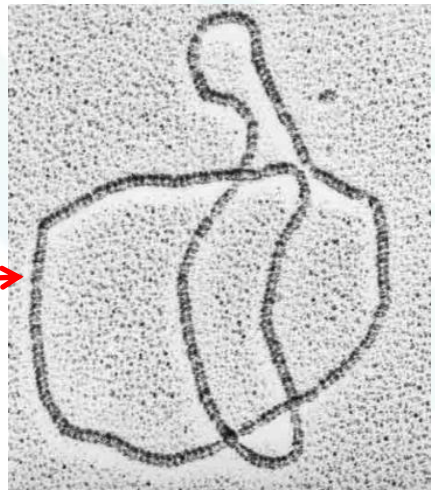
- 真核生物





- 原核生物





常见生物数据格式

- **FASTA**：最常用的蛋白质/核酸序列格式
- **FASTQ**：新一代测序数据格式
- **GenBank**（**GenPept**）：GenBank数据库的核酸和蛋白质序列格式
- **EMBL**：欧洲分子生物学实验室定义的核酸和蛋白质序列格式
- **GFF/GTF**：描述基因组测序数据中的序列特征的数据格式
- **PDB**：描述大分子空间结构的数据格式
- **NEXUS**：系统发育与分子进化数据格式
- **NWK**：描述进化树的数据格式

FASTA

- >gi|13810245|emb|CAC37410.1| globin [Sabella spallanzanii]
  MFRFALLCAFVADASAEGCCSMEDRQEVLNAWEALWSAEYTGRRVMIAQA
  AFQKLFEKAPDSKALFTRVNVDNIGSPQFRAHCIRVTNGFDTIINMAFDTDV
  LEELLTHLGNQHTKYQGMRAAYLTHFRESFAEILPQAIPCFNTAAWNRCITA
  MQDKIGASLAA

FASTQ

```
@SRR001666.1 071112_SLXA length=36
GGGTGATGGCCGCTGCCGATGGCGTCAAATCCCACC
+SRR001666.1 071112_SLXA length=36
IIIIIIIIIIIIIIIIIIIIIIIIIIIIIII9IG9IC
```

GenBank/GenPept

LOCUS CAC37410 165 aa linear INV 20-JUL-2001
DEFINITION globin [Sabella spallanzanii].
ACCESSION CAC37410
VERSION CAC37410.1 GI:13810245
DBSOURCE embl accession AJ131283.1
KEYWORDS .
SOURCE Sabella spallanzanii
ORGANISM Sabella spallanzanii
Eukaryota; Metazoa; Lophotrochozoa; Annelida; Polychaeta; Palpata;
Canalipalpata; Sabellida; Sabellidae; Sabella.
REFERENCE 1
AUTHORS Pallavicini,A., Negrisolo,E., Barbato,R., Dewilde,S.,
Ghiretti-Magaldi,A., Moens,L. and Lanfranchi,G.
TITLE The primary structure of globin and linker chains from the
chlorocruorin of the polychaete Sabella spallanzanii
JOURNAL J. Biol. Chem. 276 (28), 26384-26390 (2001)
PUBMED 11294828

FEATURES Location/Qualifiers

source 1..165
/organism="Sabella spallanzanii"
/db_xref="taxon:85702"
/tissue_type="haematopoietic"
/dev_stage="adult"
/note="Gmelin,1791"

Protein 1..165
/product="globin"
/function="respiratory pigment"

Region 28..159
/region_name="Mb_like"
/note="myoglobin_like; M family globin domain; cd01040"
/db_xref="CDD:271266"

Site order(65..66,82,85..86,89,110,113..114,120,124..125,128)
/site_type="other"
/note="heme binding site [chemical binding]"
/db_xref="CDD:271266"

CDS 1..165
/gene="globin"
/coded_by="AJ131283.1:54..551"
/experiment="experimental evidence, no additional details recorded"
/db_xref="GOA:Q9BHK3"
/db_xref="InterPro:IPR000971"
/db_xref="InterPro:IPR009050"
/db_xref="InterPro:IPR012292"
/db_xref="InterPro:IPR014610"
/db_xref="UniProtKB/TrEMBL:Q9BHK3"

```
ORIGIN
        1 mfrfallcaf vadasaegcc smedrqevln awealwsaey tgrrvmiaqa afqklfekap
       61 dskalftrvn vdnigspqfr ahcirvtngf dtiinmafdt dvleellthl gnqhtkyqgm
      121 raaylthfre sfaeilpqai pcfntaawnr citamqdkig aslaa
//
```

EMBL

```
ID SC10H5 standard; DNA; PRO; 4870 BP.
XX
AC AL031232;
XX
DE Streptomyces coelicolor cosmid 10H5.
XX
KW integral membrane protein.
XX
OS Streptomyces coelicolor
OC Eubacteria; Firmicutes; Actinomycetes; Streptomycetes;
OC Streptomycetaceae; Streptomyces.
XX
RN [1]
RP 1-4870
RA Oliver K., Harris D.;
RT ;
RL Unpublished.
```

```
FT   misc_feature    4769..4870
FT                   /note="overlap with cosmid 3A7 from 1 to 102"
XX
SQ   Sequence 4870 BP; 769 A; 1717 C; 1693 G; 691 T; 0 other;
     gatcagtaga cccagcgaca gcagggcggg gcccagcagg ccggccgtgg cgtagagcgc       60
     gaggacggcg accggcgtgg ccaccgacag gatggctgcg gcgacgcgga cgacaccgga      120
     gtgtgccagg gcccaccaca cgccgatggc cgcgagcgcg agtcccgcgc tgccgaacag      180
     ggcccacagc acactgcgca gaccggcggc cacgagtggc gccaggacgg tgcccagcag      240
     gagcagcagg gtgacgtggg cgcgcgctgc actgtggccg ccccgtccgc ccgacgcgcg      300
     cggctcgtca tctcgcggtc ccaccaccgg tcggccccat tactcgtcct caaccctgtg      360
     gcgactgacg ttccccggac aggtcgtacc gattgccgcc acgccccacc acgcacaggg      420
     cccagacgac gaagcctgac atggtgatca tgacgacgga ccacaccggg tagtacggca      480
     gcgagaggaa gttggcgatg atcaccagcc cggcgatggc gaccccggtg acacgtgccc      540
     acatcgccgt tttgagcagc ccggcgctga cgaccatggc gagcgcgccg agcgcgagat      600
     ggatccaccc ccacccggtg agatcgaact ggaaaacgta gttgggcgtg gtgacgaaga      660
     cgtcgtcctc ggcgatggcc atgatgcccc ggaagaggct gagcagcccg gcgaggaaga      720
     gcatcaccgc cgcgaaggcg gtaaggcccg tcgcccattc ctgcctcgcg gtgtgtgccg      780
     ggtggtgggt atgtgacgtg gtcatctcgg acctcgtttc gtggaatgcg gatgcttcag      840
     cgagcggagg cgccggtgcc cgccgcgccc gtgtgccctg ccgggccgtg accggacagg      900
     accaattcct tcgccttgcg gaactcctcg tccgtgatgg caccccggtc tcggatctcg      960
     gagagccggg ccagctcgtc gacgctgctg gacccgccgc ccacggtctt cctgatgtag     1020
```

GFF/GTF

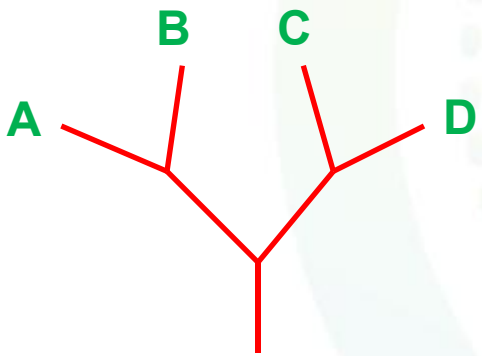
- **GFF格式包括以下字段（制表符分割）：**
  - **seqname - name of the chromosome or scaffold;**
  - **source - name of the program that generated this feature, or the data source**
  - **feature - feature type name, e.g. Gene, Variation, Similarity**
  - **start - Start position of the feature, with sequence numbering starting at 1.**
  - **end - End position of the feature, with sequence numbering starting at 1.**
  - **score - A floating point value.**
  - **strand - defined as + (forward) or - (reverse).**
  - **frame - One of '0', '1' or '2'.**
  - **attribute - A semicolon-separated list of tag-value pairs**

如：

```
pseudogene Ensembl gene 11869 14409 . + . ENSG00000223972
p_transcript Ensembl transcript 11869 14409 . + . ENSG00000223972
```

NWK(Newick)

- ((A,B),(C,D))



单选题 1分

设置

关于生物序列格式，下面说法正确的是：

A FASTA是最常用的蛋白质序列格式

B GenBank与FASTA格式不能相互转换

C GFF格式是描述进化树的数据格式

D PDB是描述蛋白质二级结构的数据格式

提交