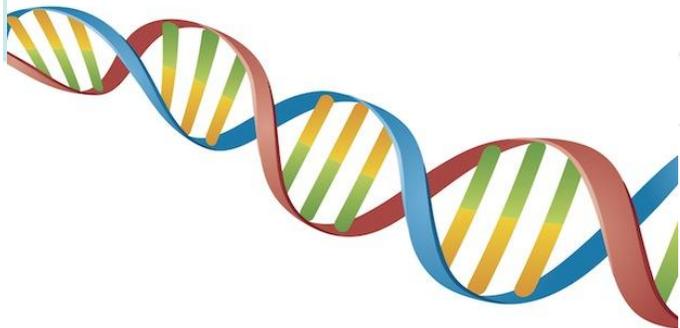




Linux与生物信息数据处理

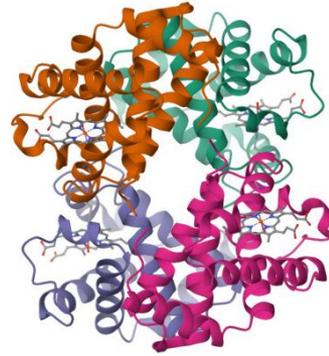
解增言

生物信息学院





斑头雁 (*Anser indicus*)



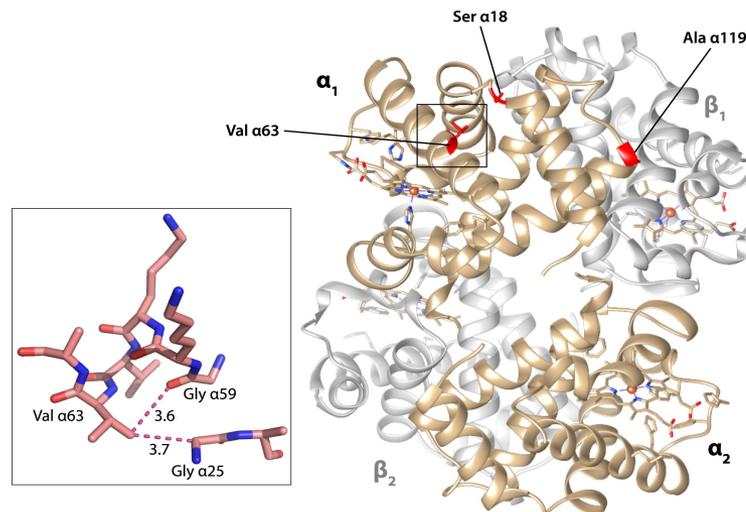
灰雁 (*Anser anser*)

>ACT80359.1 hemoglobin alpha A subunit [*Anser indicus*]

MVLSAADKTNVKGVFSKISGHAE EYGAETLERMFTAYPQTKTYFP HFDLQHGS AQIKAHGKKVVAALVEA
 VNHIDDIAGALSKLSDLHAQKLRVDPVNFKFLGHCFLVVVAIHHPSALTAEVHASLDKFLCAVGTVLTAK
 YR

>ACT80855.1 hemoglobin alpha A subunit [*Anser anser*]

MVLSAADKTNVKGVFSKIGGHAE EYGAETLERMFTAYPQTKTYFP HFDLQHGS AQIKAHGKKVVAALVEA
 VNHIDDIAGALSKLSDLHAQKLRVDPVNFKFLGHCFLVVVAIHHPSALTP EVHASLDKFLCAVGTVLTAK
 YR



教材与参考书



参考书1

鸟哥的Linux私房菜
(基础篇第四版)

参考书2

Daniel J. Barrett,
Linux命令速查手册 (第三版) .
中国电力出版社, 2018



课程网站与上机账号



- **课程网站:**

- <https://www.xueyinonline.com/detail/236158134>

- <http://www.linuxstudio.cn>（免费注册）

- **作业及练习服务器:**

- www.linuxstudio.cn

- **用户名:** 姓名全拼（如wangbin）
- **密码:** 姓名全拼2023（如wangbin2023）。请及时登录并用passwd命令更改密码



考核方式



○ 平时成绩（100分×60%）：

作业：20分（每次2分，共10次）

随堂测试：20分（每次5分，共4次）

雨课堂：20分

课程网站学习情况：20分

学银慕课学习情况：10分

打字准确性与速度：10分

缺课：每次-5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（除简答外，题目来自课程网站www.linuxstudio.cn题库）



课程简介



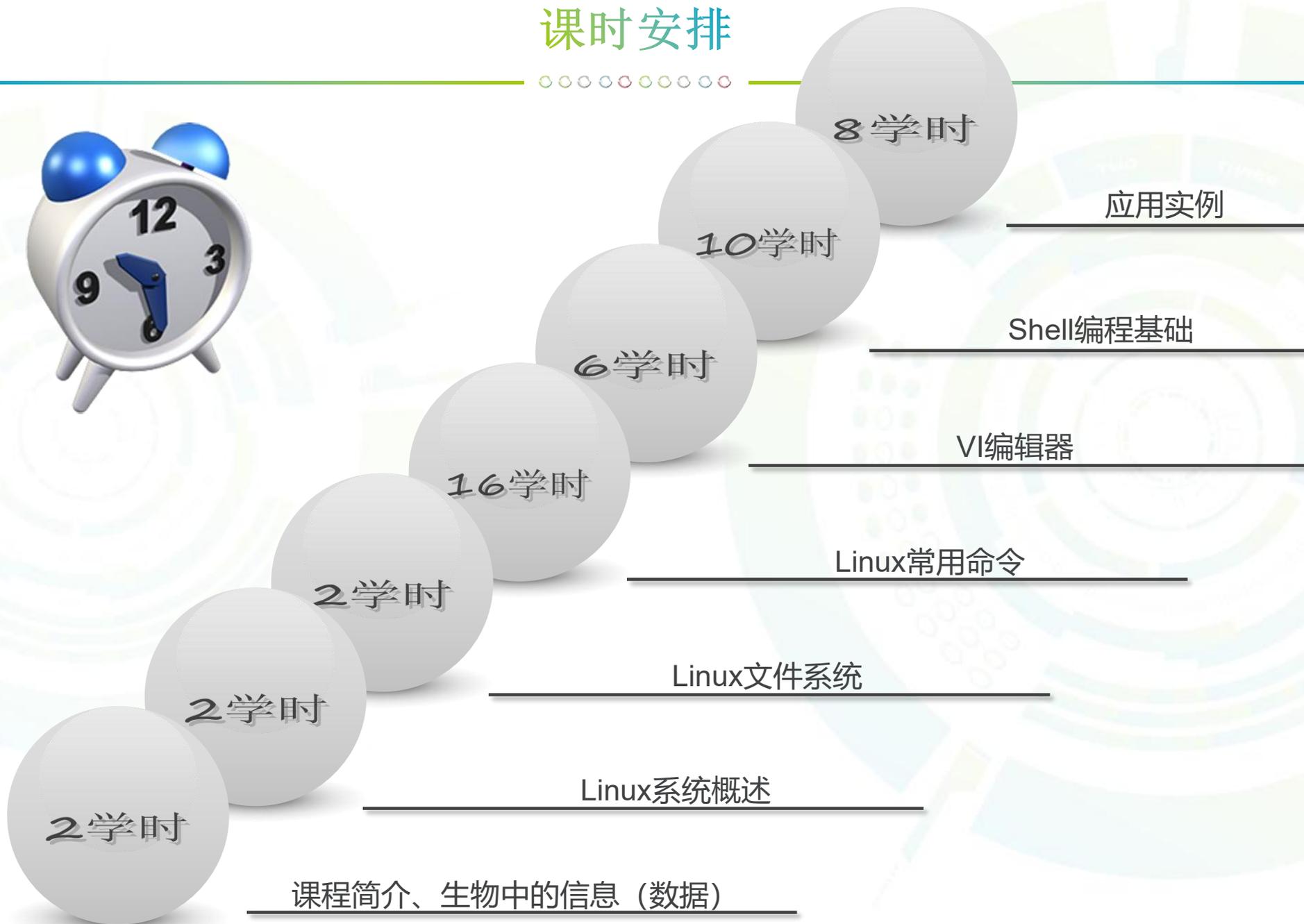
- Linux与生物信息数据处理课程主要内容是学习利用Linux命令、管道及Shell编程，处理生物学数据（主要是分子生物学数据）。



课程内容



课时安排

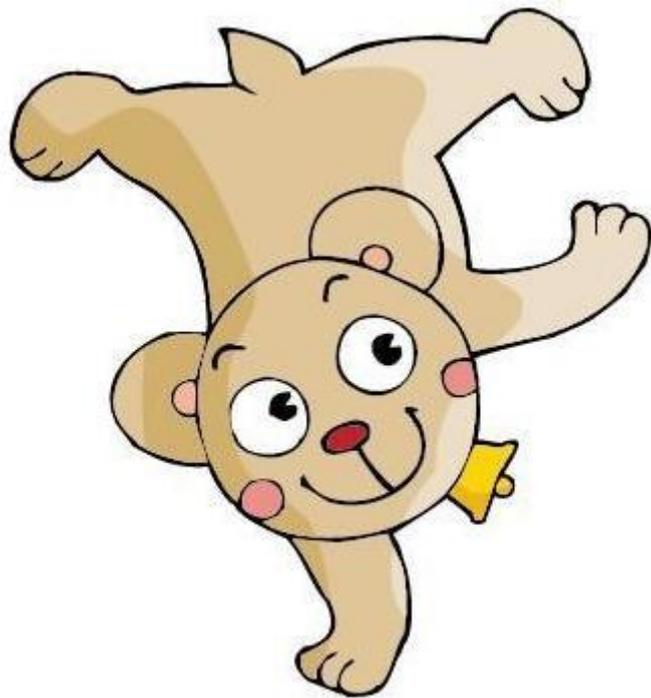


上课形式



○ 翻转课堂

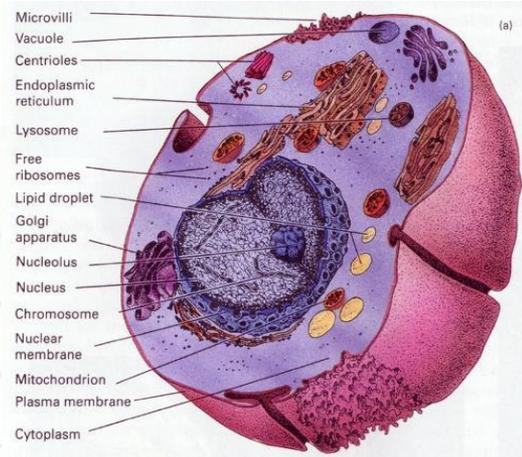
- ✓ 课前：观看教学视频，提前练习命令
- ✓ 课内：练习、作业、测试、解答问题、重点讲解、总结
- ✓ 课后：复习



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，及时完成雨课堂的随堂练习；
- 按时完成**作业**：根据要求，按时完成服务器上的作业练习；
- **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；
- **打字练习**：养成良好的打字习惯，提高打字的规范性、准确性及速度；
- 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。**



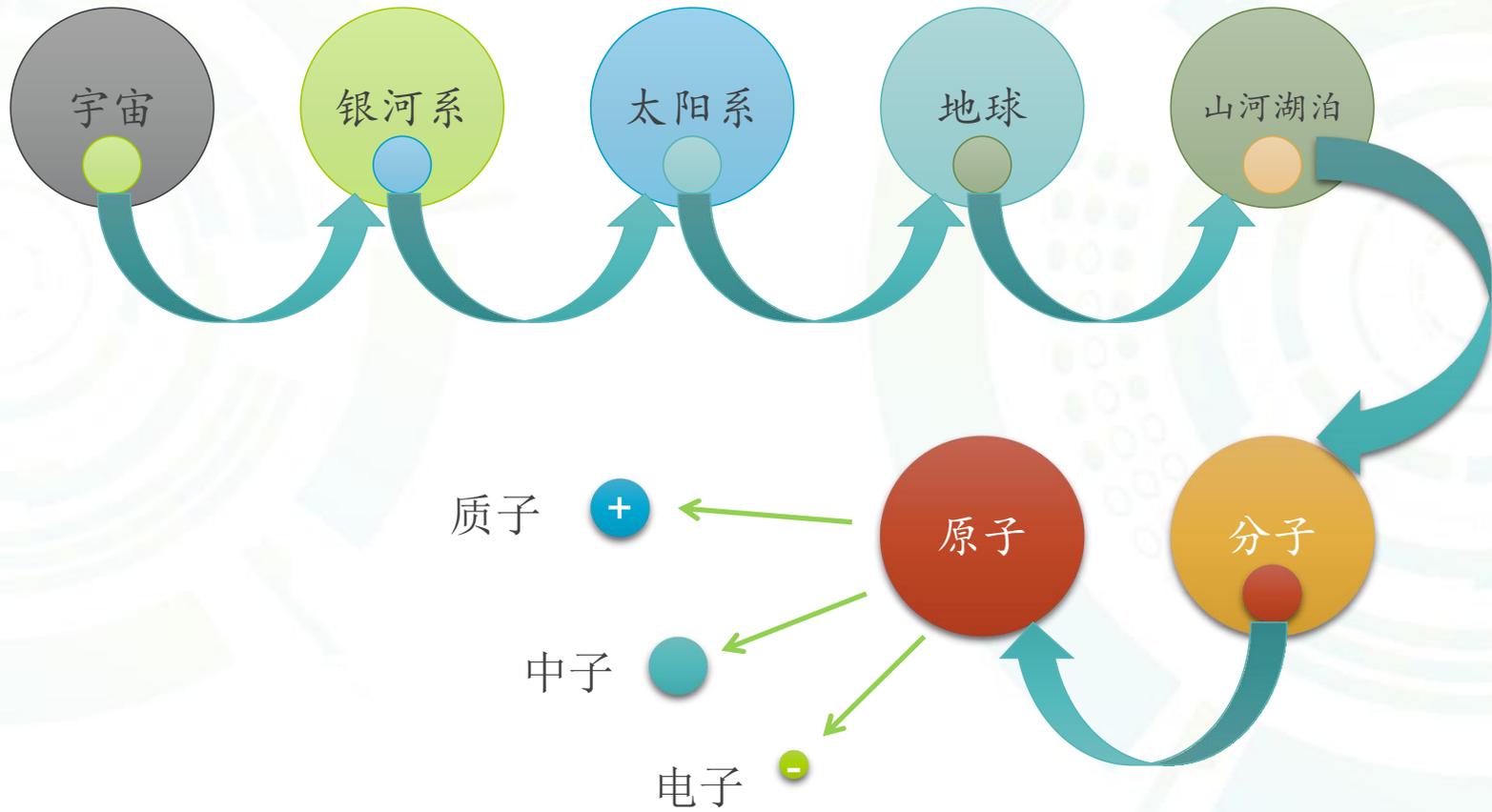
生物学中的数据

解增言

生物信息学院



从宇宙到生物



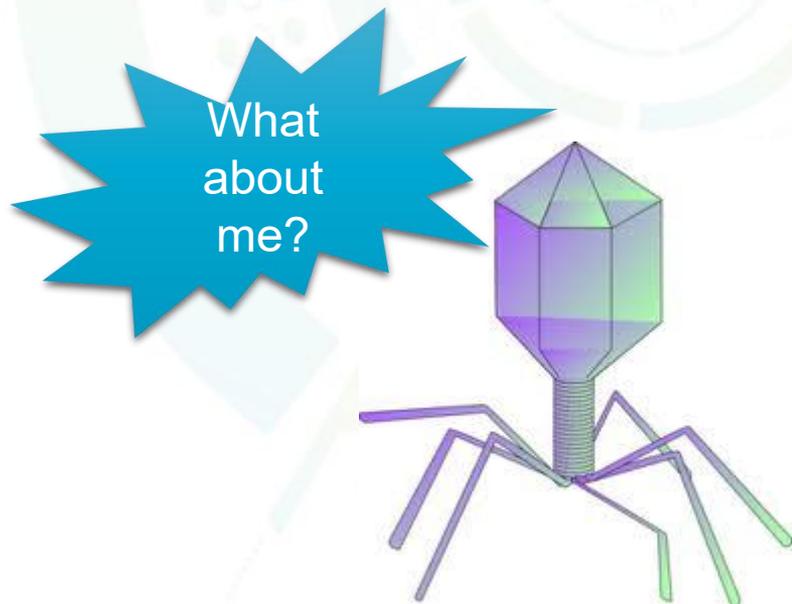
从宇宙到生物



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



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）

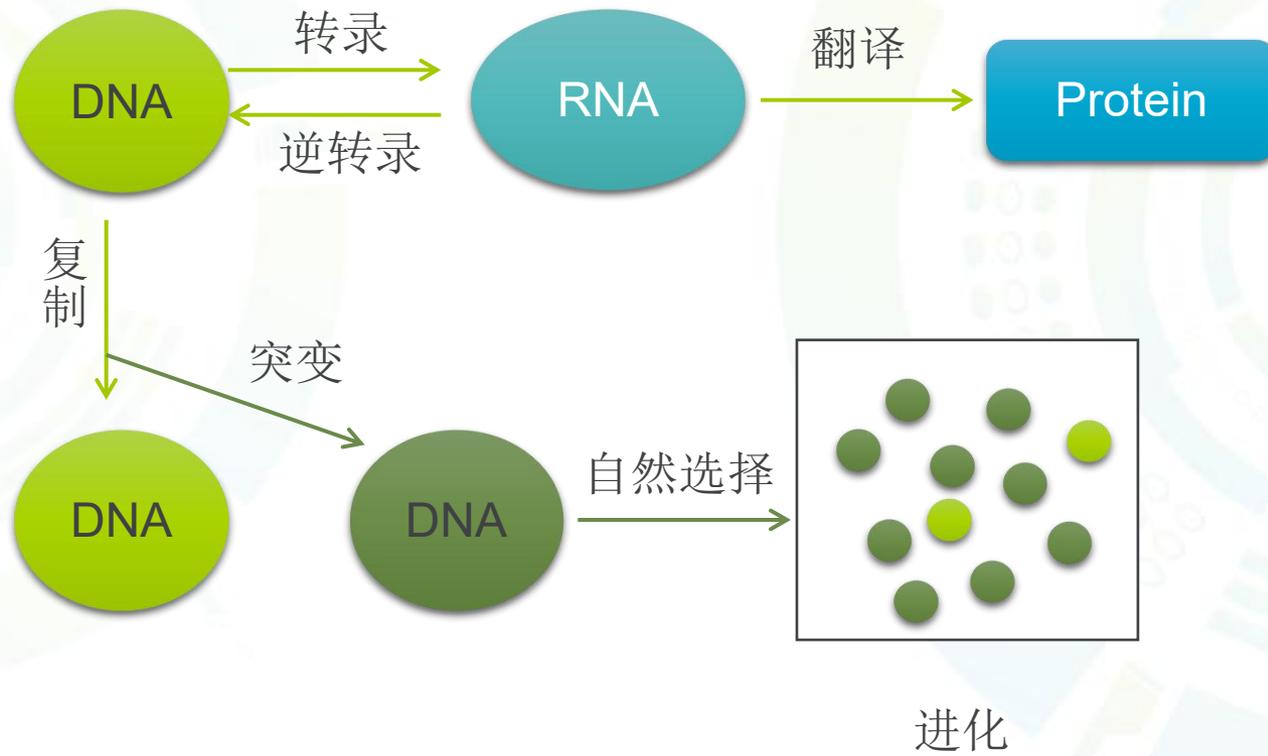
RNA（核糖核酸）

- **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



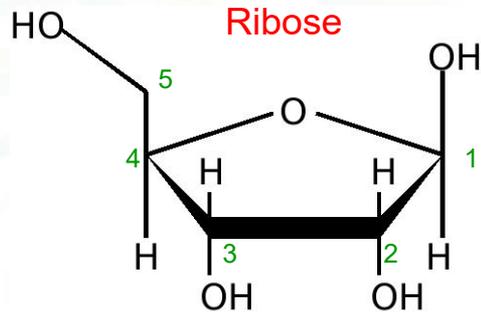
- 主要生物单分子：

核糖核苷酸

脱氧核糖核苷酸

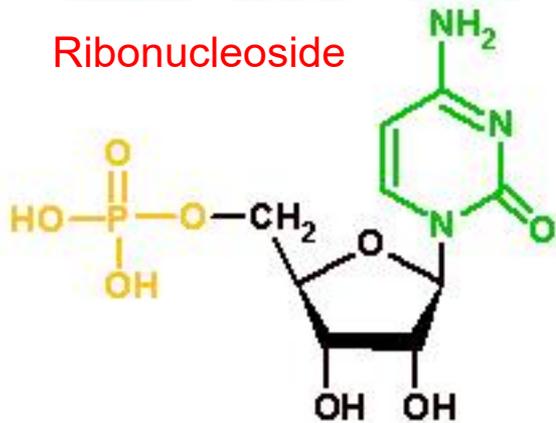
氨基酸

RNA构成



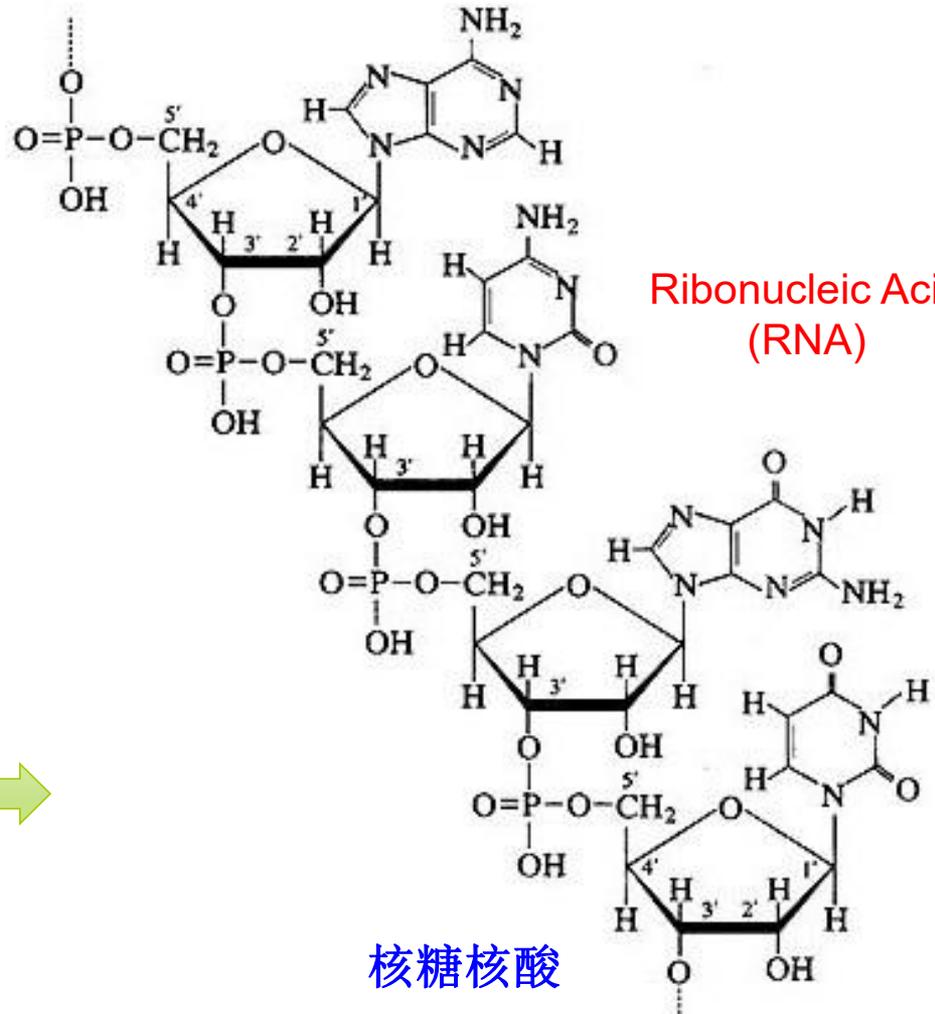
Ribose

核糖



Ribonucleoside

核糖核苷酸



Ribonucleic Acid (RNA)

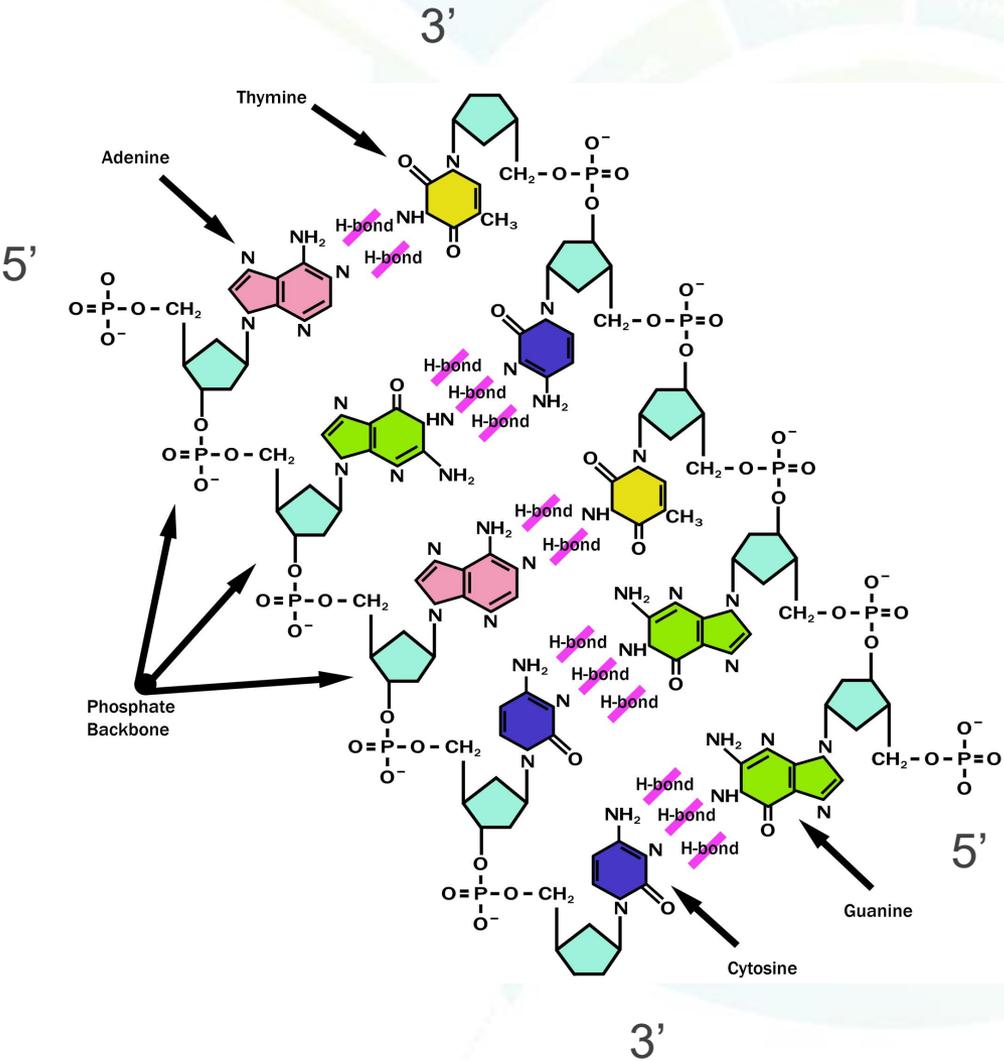
核糖核酸

DNA简写

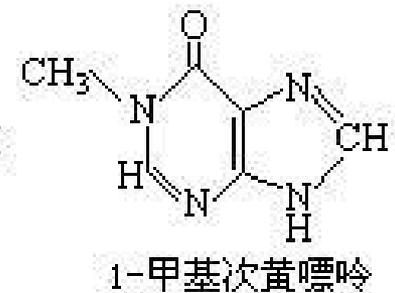
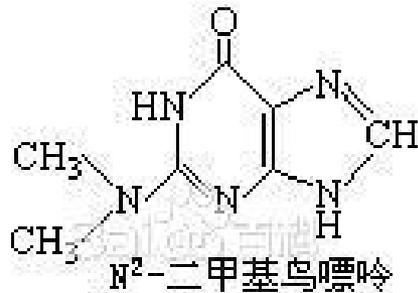
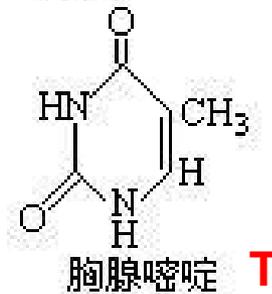
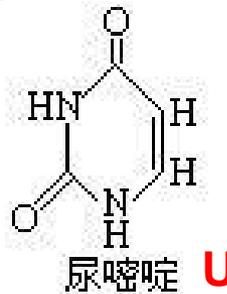
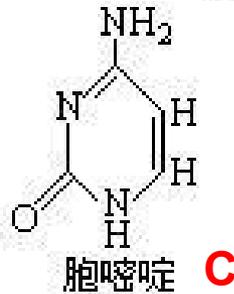
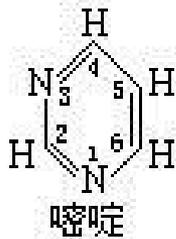
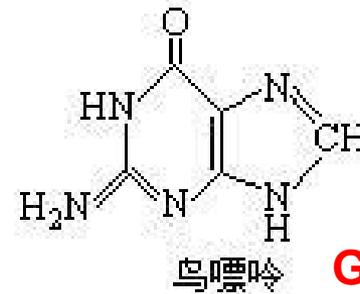
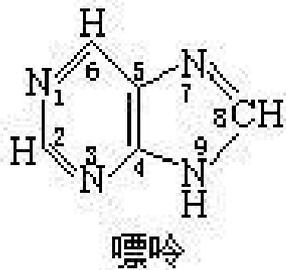


5' AGACC 3'
3' TCTGG 5'

5' ATGGCCAGGTAA 3'



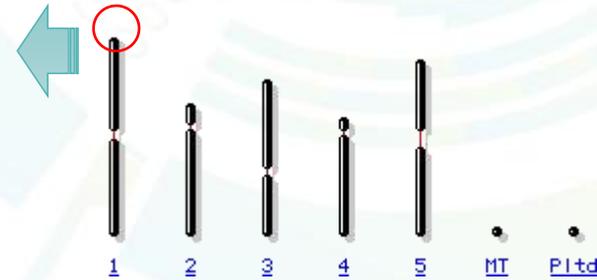
常见碱基



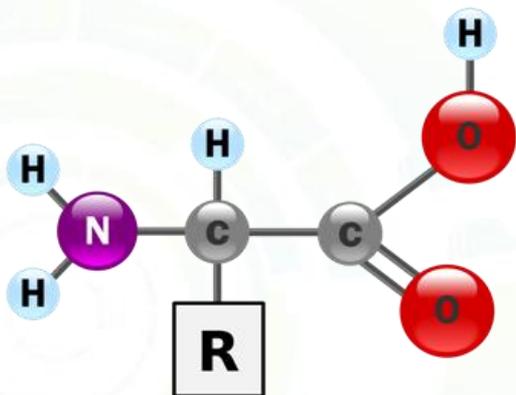
拟南芥第一条染色体部分序列

- >Chr1 CHROMOSOME dumped from ADB: Feb/3/09 16:9;
last updated: 2007-12-20

```
CCCTAAACCCTAAACCCTAAACCCTAAACCTCTGAATCCTTAATCCC  
TAAATCCCTAAATCTTTAAATCCTACATCCATGAATCCCTAAATACC  
TAATTCCTAAACCCGAAACCGGTTTCTCTGGTTGAAAATCATTGTG  
TATATAATGATAATTTTATCGTTTTTATGTAATTGCTTATTGTTGTG  
TGTAGATTTTTTAAAAATATCATTTGAGGTCAATACAAATCCTATTT  
CTTGTGGTTTTCTTTCCTTCACTTAGCTATGGATGGTTTATCTTCAT  
TTGTTATATTGGATACAAGCTTTGCTACGATCTACATTTGGGAA
```



蛋白质构成



氨基酸结构

21种氨基酸

⊕ Positive ⊖ Negative
+ Side chain charge at physiological pH 7.4

Twenty-One Amino Acids

A. Amino Acids with Electrically Charged Side Chains

Positive			Negative	
Arginine (Arg) R	Histidine (His) H	Lysine (Lys) K	Aspartic Acid (Asp) D	Glutamic Acid (Glu) E
<chem>NC(CCC[NH2+])C(=O)O</chem> pKa 2.03, 9.00, 12.10	<chem>NC(Cc1c[nH]cn1)C(=O)O</chem> pKa 1.70, 6.04, 9.09	<chem>NC(CCCC[NH3+])C(=O)O</chem> pKa 2.15, 9.16, 10.67	<chem>NC(CC(=O)[O-])C(=O)O</chem> pKa 1.95, 3.71, 9.66	<chem>NC(CCC(=O)[O-])C(=O)O</chem> pKa 2.16, 4.15, 9.58

B. Amino Acids with Polar Uncharged Side Chains

Serine (Ser) S	Threonine (Thr) T	Asparagine (Asn) N	Glutamine (Gln) Q
<chem>NC(CO)C(=O)O</chem> pKa 2.13, 9.05	<chem>NC(C(C)O)C(=O)O</chem> pKa 2.20, 8.96	<chem>NC(CC(N)=O)C(=O)O</chem> pKa 2.10, 8.70	<chem>NC(CCC(N)=O)C(=O)O</chem> pKa 2.18, 9.00

C. Special Cases

Cysteine (Cys) C	Selenocysteine (Sec) U	Glycine (Gly) G	Proline (Pro) P
<chem>NC(CS)C(=O)O</chem> pKa 1.91, 10.28, 8.14	<chem>NC(CSe)C(=O)O</chem> pKa 1.9, 10	<chem>NC(=O)C(=O)O</chem> pKa 2.34, 9.58	<chem>C1CCNC1</chem> pKa 1.95, 10.4

D. Amino Acids with Hydrophobic Side Chain

Alanine (Ala) A	Valine (Val) V	Isoleucine (Ile) I	Leucine (Leu) L	Methionine (Met) M	Phenylalanine (Phe) F	Tyrosine (Tyr) Y	Tryptophan (Trp) W
<chem>NC(C)C(=O)O</chem> pKa 2.33, 9.71	<chem>NC(C(C)C)C(=O)O</chem> pKa 2.27, 9.52	<chem>NC(C(C)CC)C(=O)O</chem> pKa 2.26, 9.60	<chem>NC(C(C)CC)C(=O)O</chem> pKa 2.32, 9.58	<chem>NC(CSC)C(=O)O</chem> pKa 2.16, 9.08	<chem>NC(Cc1ccccc1)C(=O)O</chem> pKa 2.18, 9.09	<chem>NC(Cc1ccc(O)cc1)C(=O)O</chem> pKa 2.24, 9.04	<chem>NC(Cc1c[nH]c2ccccc12)C(=O)O</chem> pKa 2.38, 9.34

NONPOLAR, HYDROPHOBIC

POLAR, UNCHARGED

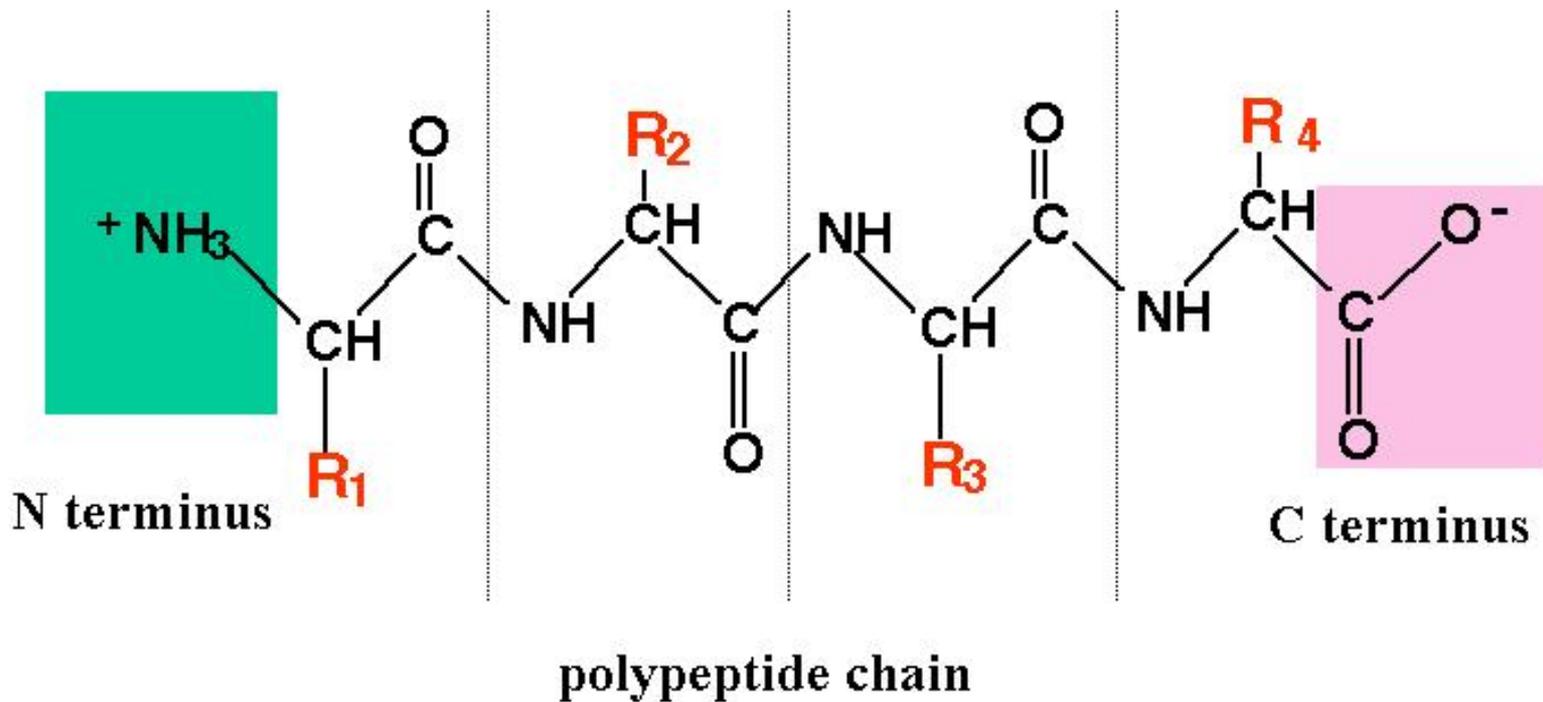
R GROUPS

Alanine Ala A MW = 89	$\begin{matrix} ^- \text{OOC} \\ \\ \text{H}_3\text{N}^+ - \text{CH} - \text{CH}_3 \end{matrix}$		$\begin{matrix} \text{H} - \text{CH} - \text{COO}^- \\ \\ \text{N}^+ \text{H}_3 \end{matrix}$	Glycine Gly G MW = 75
Valine Val V MW = 117	$\begin{matrix} ^- \text{OOC} \\ \\ \text{H}_3\text{N}^+ - \text{CH} - \text{CH} \begin{matrix} / \text{CH}_3 \\ \backslash \text{CH}_3 \end{matrix} \end{matrix}$		$\begin{matrix} \text{HO} - \text{CH}_2 - \text{CH} - \text{COO}^- \\ \\ \text{N}^+ \text{H}_3 \end{matrix}$	Serine Ser S MW = 105
Leucine Leu L MW = 131	$\begin{matrix} ^- \text{OOC} \\ \\ \text{H}_3\text{N}^+ - \text{CH} - \text{CH}_2 - \text{CH} \begin{matrix} / \text{CH}_3 \\ \backslash \text{CH}_3 \end{matrix} \end{matrix}$		$\begin{matrix} \text{OH} \\ \\ \text{CH}_3 - \text{CH} - \text{CH} - \text{COO}^- \\ \\ \text{N}^+ \text{H}_3 \end{matrix}$	Threonine Thr T MW = 119
Isoleucine Ile I MW = 131	$\begin{matrix} ^- \text{OOC} \\ \\ \text{H}_3\text{N}^+ - \text{CH} - \text{CH} \begin{matrix} / \text{CH}_3 \\ \backslash \text{CH}_2 - \text{CH}_3 \end{matrix} \end{matrix}$		$\begin{matrix} \text{HS} - \text{CH}_2 - \text{CH} - \text{COO}^- \\ \\ \text{N}^+ \text{H}_3 \end{matrix}$	Cysteine Cys C MW = 121
Phenylalanine Phe F MW = 131	$\begin{matrix} ^- \text{OOC} \\ \\ \text{H}_3\text{N}^+ - \text{CH} - \text{CH}_2 - \text{C}_6\text{H}_5 \end{matrix}$		$\begin{matrix} \text{HO} - \text{C}_6\text{H}_4 - \text{CH}_2 - \text{CH} - \text{COO}^- \\ \\ \text{N}^+ \text{H}_3 \end{matrix}$	Tyrosine Tyr Y MW = 181
Tryptophan Trp W MW = 204	$\begin{matrix} ^- \text{OOC} \\ \\ \text{H}_3\text{N}^+ - \text{CH} - \text{CH}_2 - \text{C}_8\text{H}_6\text{N}_2 \end{matrix}$		$\begin{matrix} \text{NH}_2 \\ \\ \text{O} = \text{C} - \text{CH}_2 - \text{CH} - \text{COO}^- \\ \\ \text{N}^+ \text{H}_3 \end{matrix}$	Asparagine Asn N MW = 132
Methionine Met M MW = 149	$\begin{matrix} ^- \text{OOC} \\ \\ \text{H}_3\text{N}^+ - \text{CH} - \text{CH}_2 - \text{CH}_2 - \text{S} - \text{CH}_3 \end{matrix}$		$\begin{matrix} \text{NH}_2 \\ \\ \text{O} = \text{C} - \text{CH}_2 - \text{CH}_2 - \text{CH} - \text{COO}^- \\ \\ \text{N}^+ \text{H}_3 \end{matrix}$	Glutamine Gln Q MW = 146
Proline Pro P MW = 115	$\begin{matrix} ^- \text{OOC} \\ \\ \text{CH} - \text{CH}_2 \\ \quad \quad \quad \\ \text{HN} - \text{CH}_2 \end{matrix}$		POLAR BASIC $\begin{matrix} \text{NH}_3^+ - \text{CH}_2 - (\text{CH}_2)_3 - \text{CH} - \text{COO}^- \\ \\ \text{N}^+ \text{H}_3 \end{matrix}$	Lysine Lys K MW = 146
Aspartic acid Asp D MW = 133	$\begin{matrix} ^- \text{OOC} \\ \\ \text{H}_3\text{N}^+ - \text{CH} - \text{CH}_2 - \text{C}(=\text{O})\text{O}^- \end{matrix}$		$\begin{matrix} \text{NH}_2 \\ \\ \text{N}^+ \text{H}_2 = \text{C} - \text{NH} - (\text{CH}_2)_3 - \text{CH} - \text{COO}^- \\ \\ \text{N}^+ \text{H}_3 \end{matrix}$	Arginine Arg R MW = 174
Glutamine acid Glu E MW = 147	$\begin{matrix} ^- \text{OOC} \\ \\ \text{H}_3\text{N}^+ - \text{CH} - \text{CH}_2 - \text{CH}_2 - \text{C}(=\text{O})\text{O}^- \end{matrix}$		$\begin{matrix} \text{HN} = \text{C} - \text{CH}_2 - \text{CH} - \text{COO}^- \\ \\ \text{N}^+ \text{H}_3 \end{matrix}$	Histidine His H MW = 155

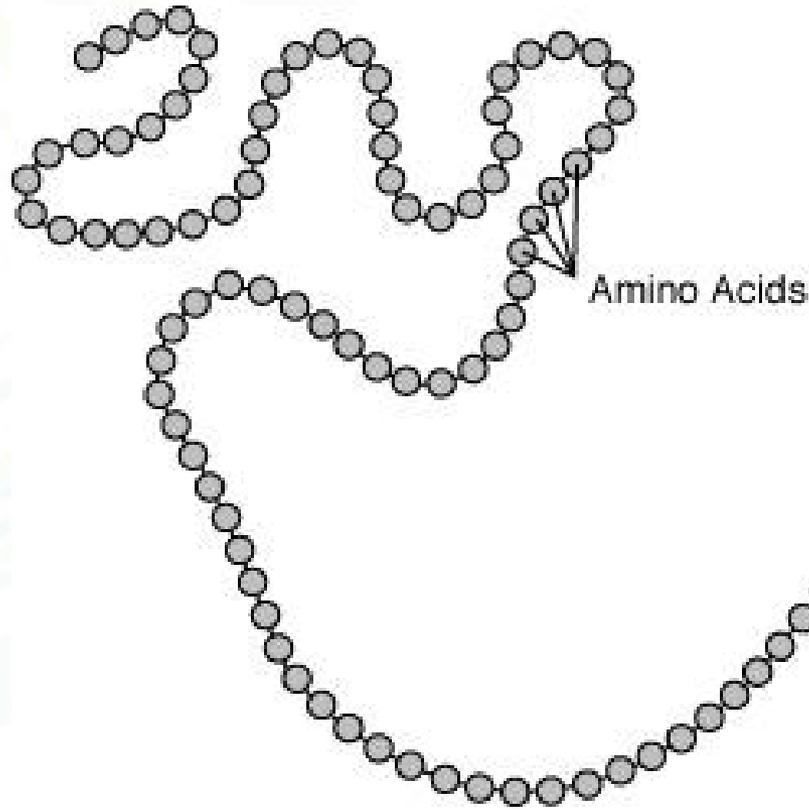
多肽结构



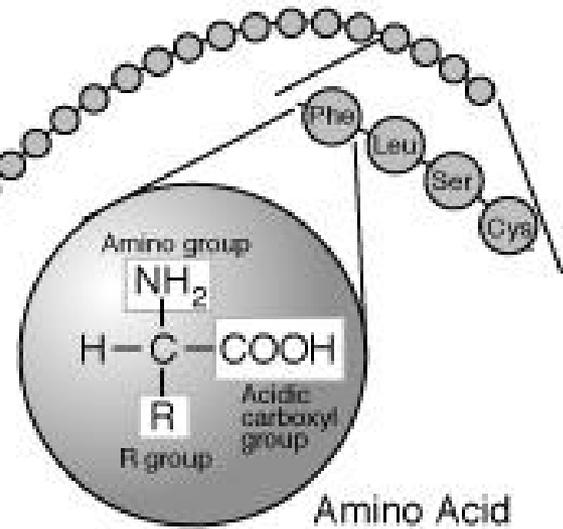
Peptide = chain of amino acids



蛋白质一级结构



Primary protein structure
is sequence of a chain of amino acids





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

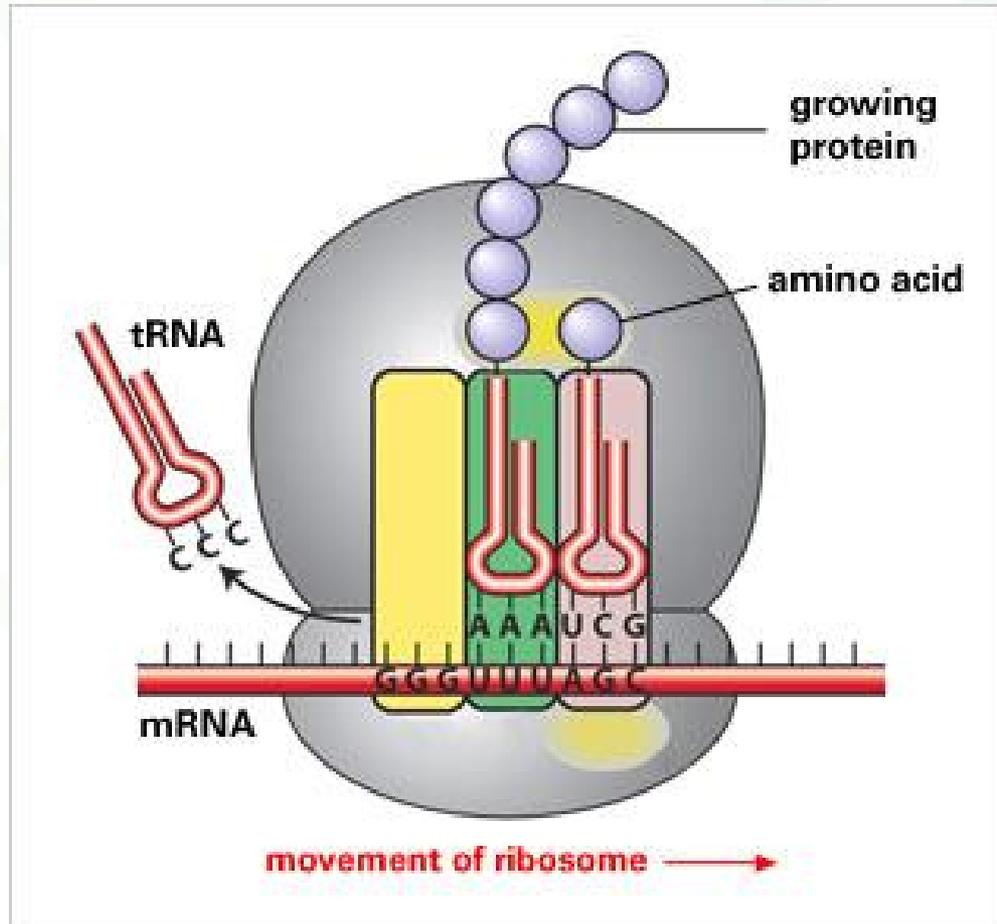
氨基酸编码



		Seconed Position									
		U		C		A		G			
		code	Amino Acid	code	Amino Acid	code	Amino Acid	code	Amino Acid		
First Position	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	
										Third Position	

概念：密码子简并性

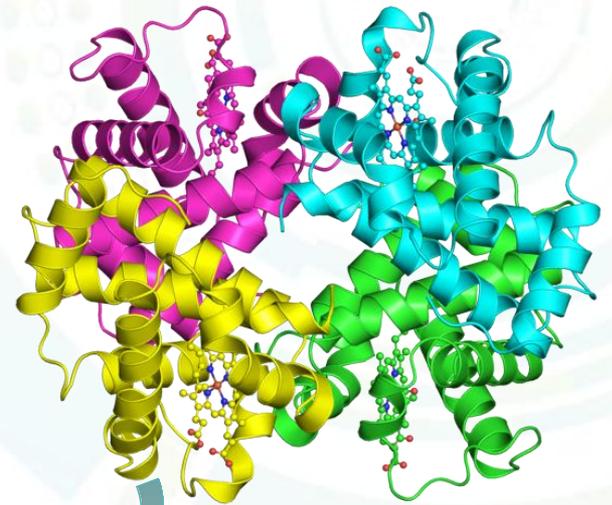
蛋白质合成



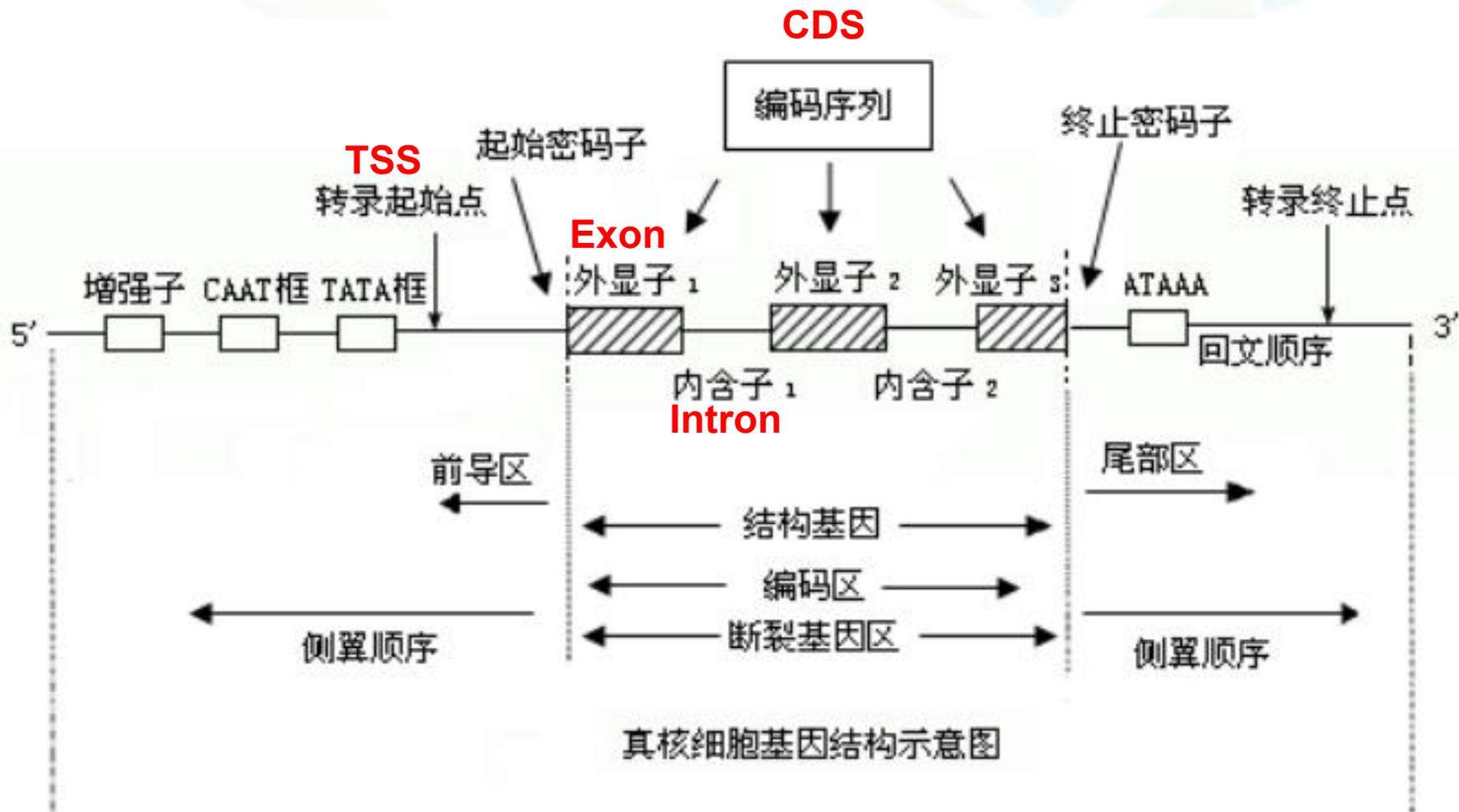
人类血红蛋白 α 亚基序列



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

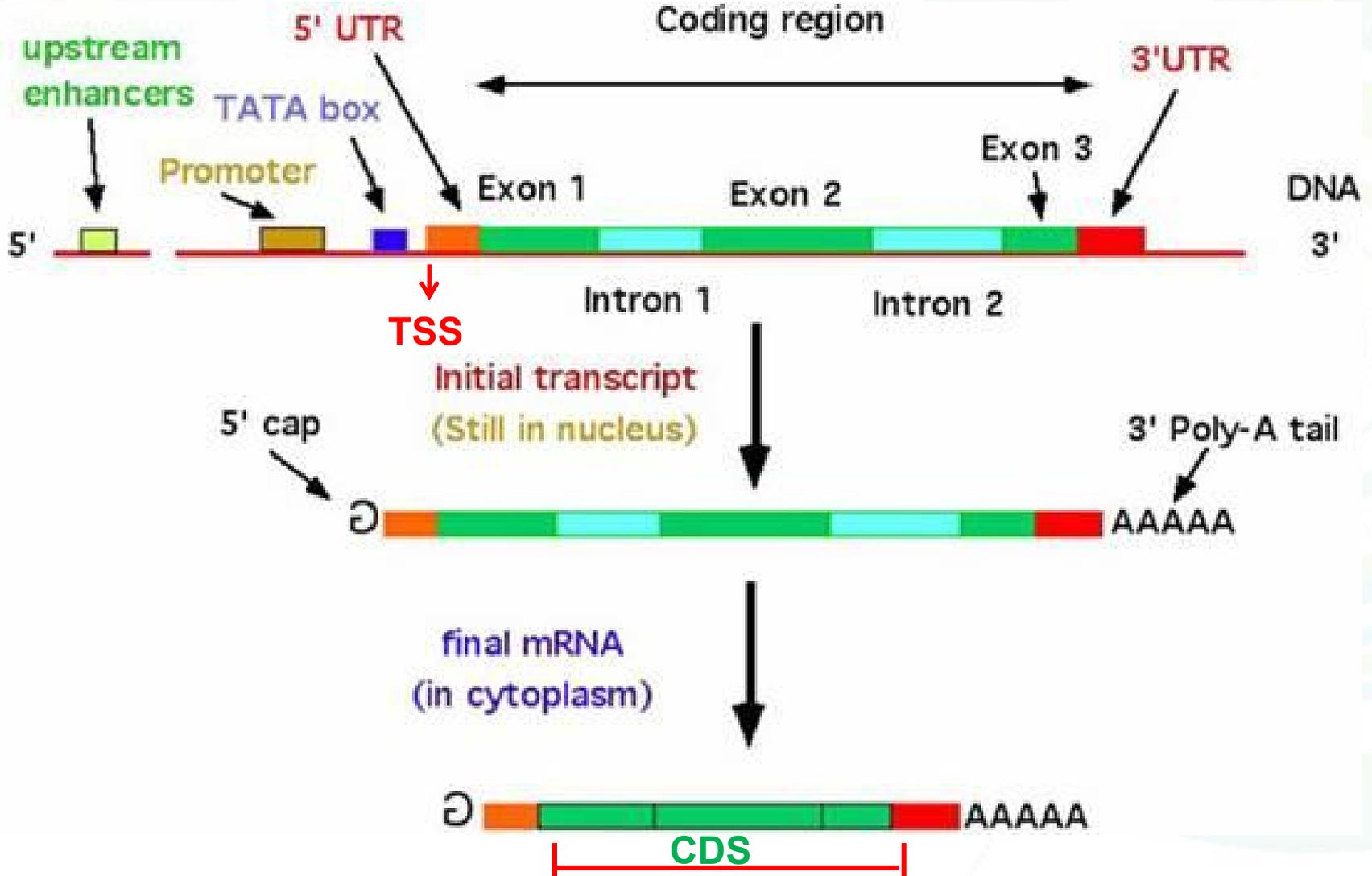


基因结构



真核细胞基因结构示意图

基因结构



基因结构示例



- >NC_003076.8:19315061-19315975 Arabidopsis thaliana chromosome 5 sequence
- ATGGAACGTGGAGGCTTCCATGGCTACCGCAAGCTGTCCGTGAACAACCACTCCTTCTCCACCAG~~gta~~
- ~~gtgccattctctataccccctcttttcacaggctctcttcatttcagttgcatgcgaaaccattctctgc~~
- ~~aatccctccattgtcatgtctgtactcttttcacaggaacagttaatgaaatagcttttcaatcttata~~
- ~~aaccgcgcatgcagacgtcatcgaagccattatgcactaaaacttccatttttcttatttttgttag~~GAT
- TAGCAGCGAATTTTCTGATGGCAGAGGGCAGTATGCGTCCTCCAGAATTC AACCAGCCTAACAAAACCAG
- TAATGGTGGT GAGGAGGAGTGCACGGT GAGGGAGCAAGACAGGTT CATGCCTATTGCCAACGTGATACGG
- ATCATGCGGAGGATCTTACCTGCTCACGCCAAGATCTCAGATGACTCCAAGGAGACGATCCAAGAGTGTG
- TTTCGGAGTACATCAGCTTCATAACAGGGGAGGCTAATGAGCGGTGCCAGCGGGAACAGCGCAAGACCAT
- CACTGCTGAGGACGTCTTGTGGGCAATGAGCAAGCTCGGTTTTGATGACTACATCGAACCCCTCACGTTG
- TACCTCCACCGCTACAGAGAGTTGGAAGGTGAAAGAGGGGTTAGCTGCAGTGCTGGGTCCGTTAGTATGA
- CCAACGGCTTGGTGGTCAAGAGGCCTAATGGGACCATGACCGAGTATGGAGCCTACGGGCCTGTGCCAGG
- GATTCACATGGCGCAGTACCATTATCGTCATCAGAACGGGTTTGTTCAGTGGTAACGAACCTAATTCT
- 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

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

TAIR数据库

(<https://www.arabidopsis.org>)

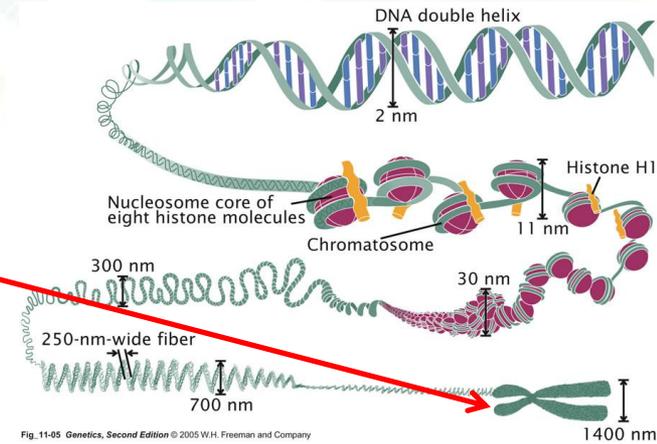
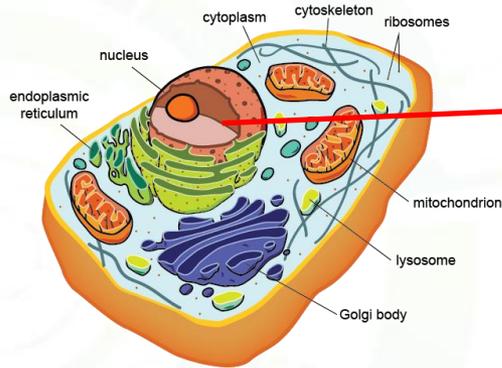
的拟南芥基因组数据下载界面，
注意其中的

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。

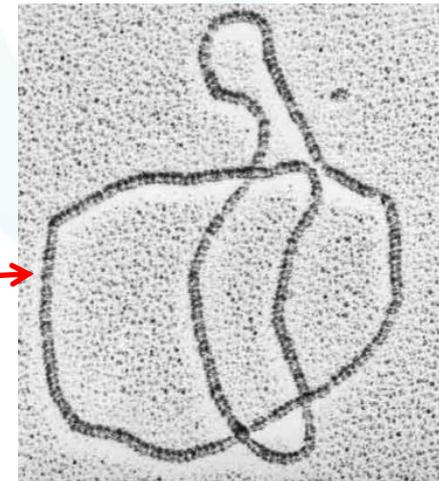
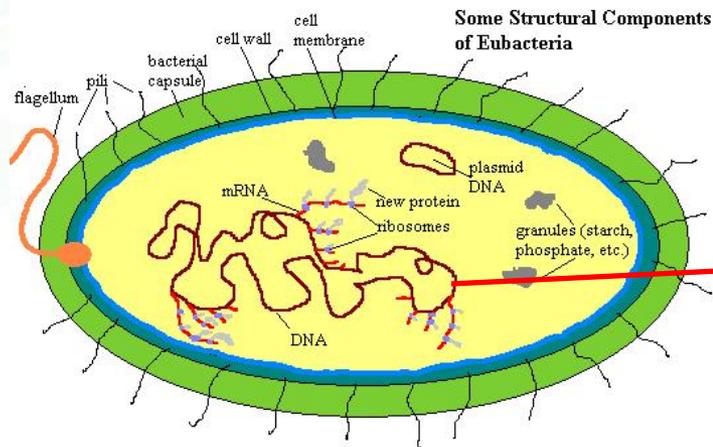
基因组结构



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



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

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., Negrisol, 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

source

Location/Qualifiers

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 smedrqlvln awealwsaey tgrrvmiaqa afqklfekap
61 dskalftrvn vdnigspqfr ahcirvtngf dtiinmafdt dvleellthl gnqhtkyqgm
121 raaylthfre sfaeilpqai pcfntaawnr citamqdkig aslaa

//



ID SC10H5 standard; DNA; PR0; 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 gccagcagg ccggccgtgg cgtagagcgc      60
gaggacggcg accggcgtgg ccaccgacag gatggctgcg gcgacgcgga cgacaccgga     120
gtgtgccagg gccaccaca cgccgatggc cgcgagcgcg agtcccgcgc tgccgaacag     180
ggcccacagc aactgcgca gaccggcggc cacgagtggc gccaggacgg tgcccagcag     240
gagcagcagg gtgacgtggg cgcgcgctgc actgtggccg ccccgctcgc ccgacgcgcg     300
cggtcgtca tctcgcggtc ccaccaccgg tcggcccat tactcgtcct caaccctgtg     360
gcgactgacg ttcccggac aggtcgtacc gattgccgcc acgcccacc acgcacaggg     420
cccagacgac gaagcctgac atggtgatca tgacgacgga ccacaccggg tagtacggca     480
gcgagaggaa gttggcgatg atcaccagcc cggcgatggc gaccccgggtg acacgtgcc     540
acatcgccgt tttgagcagc ccggcgctga cgaccatggc gagcgcgccg agcgcgagat     600
ggatccacc ccaccgggtg agatcgaact ggaaaacgta gttgggcgtg gtgacgaaga     660
cgtcgtcctc ggcgatggcc atgatgccc ggaagaggct gagcagcccg gcgaggaaga     720
gcatcaccgc cgcgaaggcg gtaaggcccg tcgccattc ctgcctcgcg gtgtgtgccg     780
ggtggtgggt atgtgacgtg gtcatctcgg acctcgttc gtggaatgcg gatgcttcag     840
cgagcggagg cgccggtgcc cgccgcgcc gtgtgccctg ccgggcccgtg accggacagg     900
accaattcct tcgccttgcg gaactcctcg tccgtgatgg caccocggtc tcggatctcg     960
gagagccggg ccagctcgtc gacgtgctg gaccgcgcgc ccacgggtctt 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))

