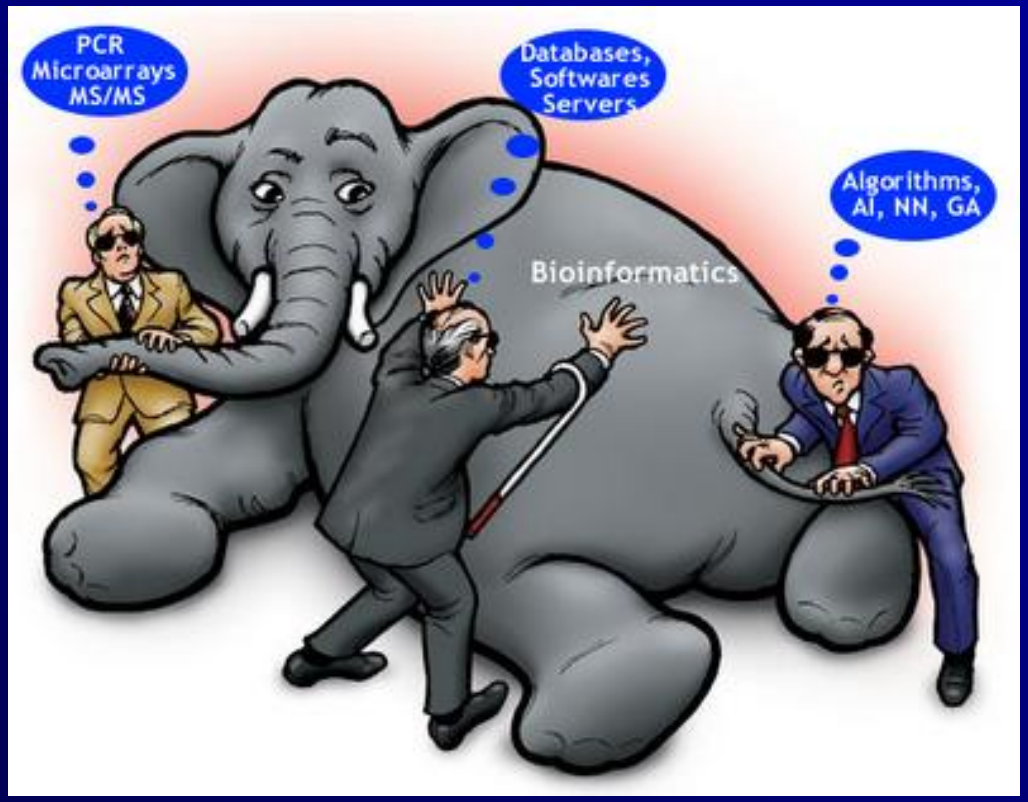
生物信息学基础

李 余 动
lyd@zjsu.edu.cn

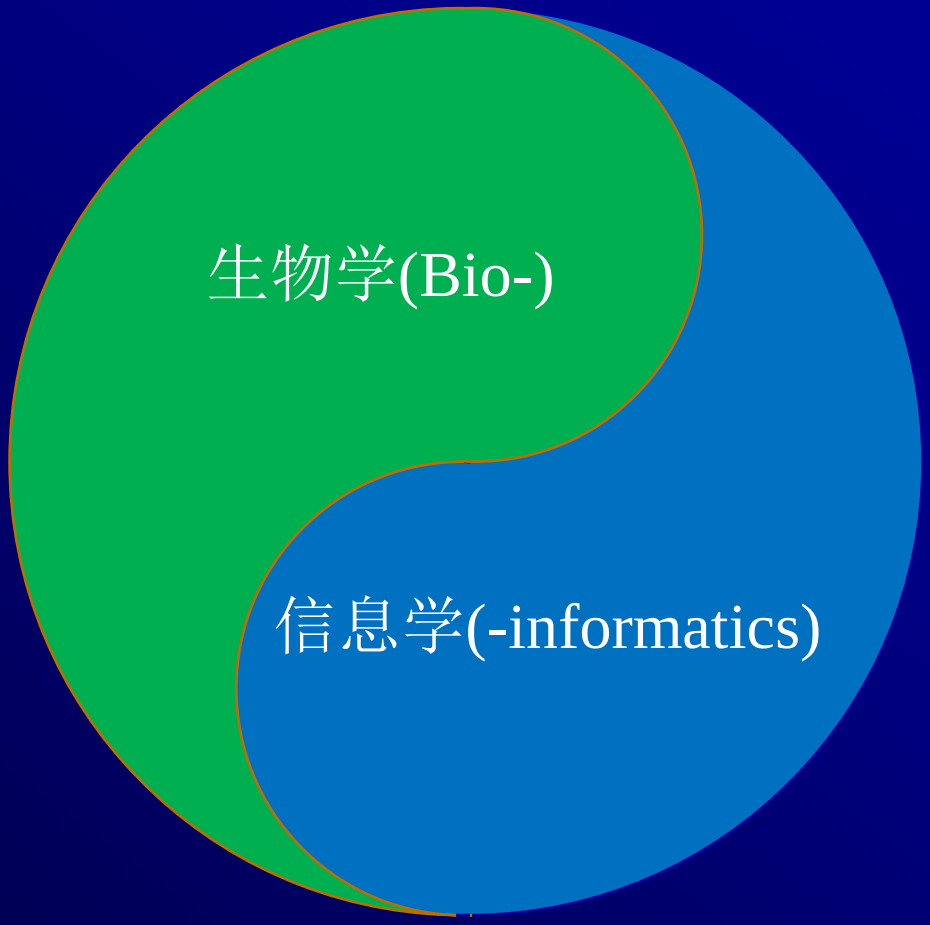
What is Bioinformatics?



生物信息学(Bioinformatics)是一门多学科交叉技术，又名计算生物学(Computational Biology)。

什么是生物信息学？

DNA
RNA
蛋白质
细胞
...



计算机
数据库
算法
模型
...

Bioinformatics is an interdisciplinary field that develops and applies computational methods and tools for understanding biological data.

生物信息学与基因组数据

- 基因组测序产生的海量生物分子数据是生物信息学的源泉

人类基因组

If you print 100 characters per line and 50 lines per page,
you'll fill **600,000 pages**, stacked **60 meters** high.

If you read one base per second, nonstop,
it will take you **100 years.**

"The massive quantities of data generated by genomic research have given rise to the field of bioinformatics--an emerging discipline that seeks to integrate computer science with applications derived from molecular biology. We are swimming in a rapidly rising sea of data...how do we keep from drowning?"

Roos DS. Bioinformatics——Trying to swim in a sea of data. Science, 2001,291(5507):1260-1261

什么是生物信息学？

生物信息学（Bioinformatics）是一门集数学、计算机科学和生物学的方法和技术于一体，以理解海量的生物学数据为目的的学科，涵盖了生物信息的获取、处理、存储、分配、分析和阐述等各个方面。

from NCBI's science primer:
www.ncbi.nlm.nih.gov/about/primer/bioinformatics.html

最早的序列分析：胰岛素蛋白质

Insulin Chain A: 8-10位存在着不同（牛，ASV；猪，TSI；羊，AGV）(Brown *et al.*, 1955**)。**

```
Bovine    GIVEQCCASVCSLYQLENYCN
Pig       GIVEQCCTSICSLYQLENYCN
Sheep     GIVEQCCAGVCSLYQLENYCN
Human     GIVEQCCTSICSLYQLENYCN
```

（Linus Carl Pauling，1901－1994）

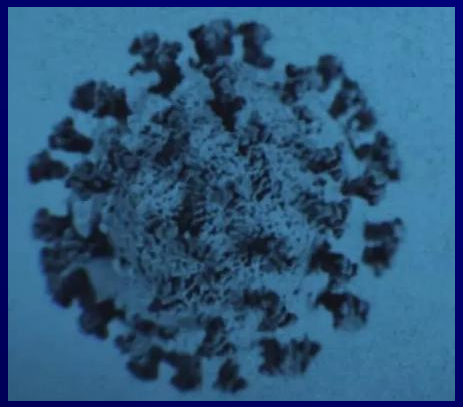
现代人类基因组序列分析

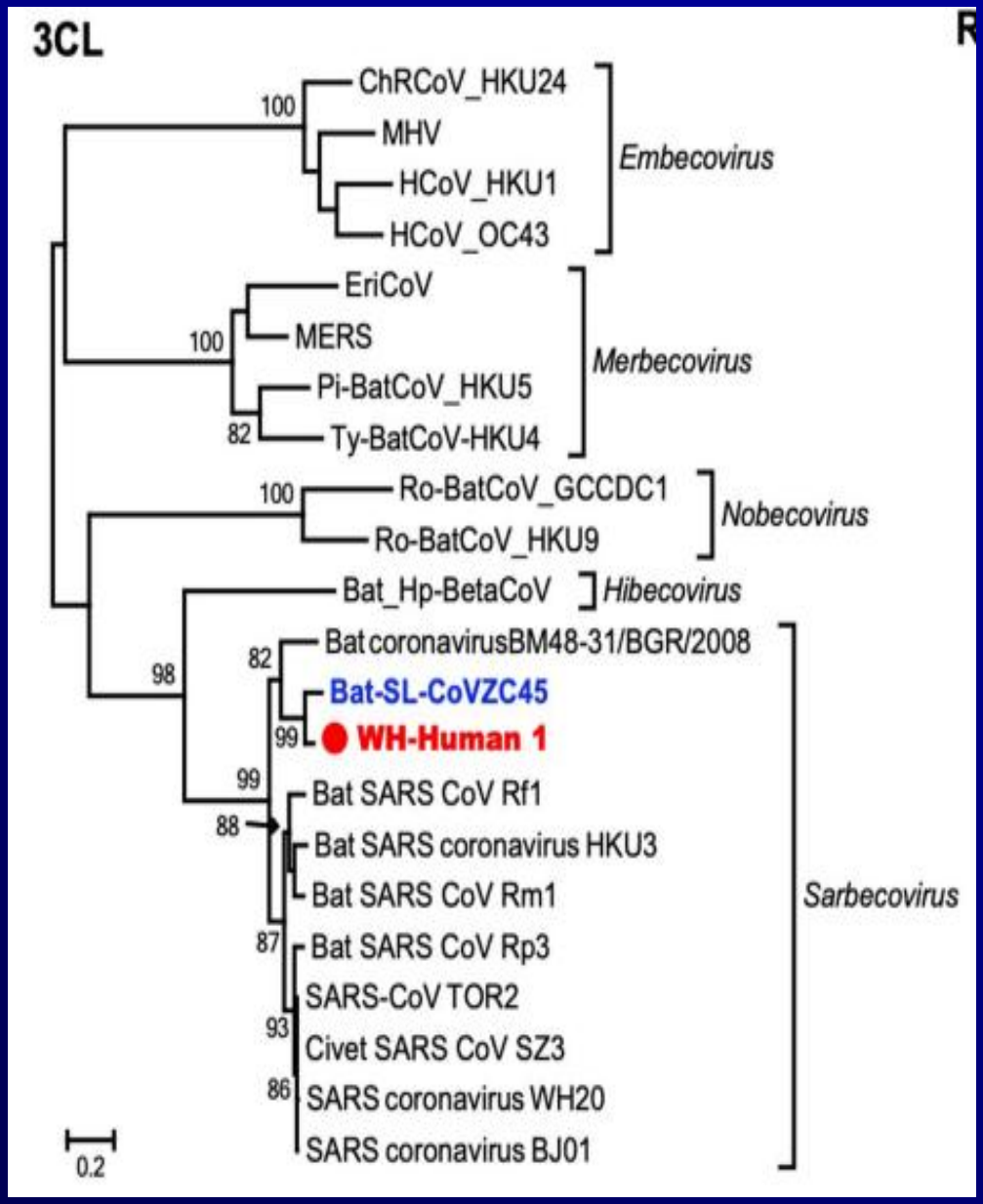
1.5% DNA difference

我们在形态学、行为学、认知上和黑猩猩不一样。但人类和黑猩猩的基因组只存在1.3%的差异（如此小的差距令我们真受打击☹）。

新冠病毒与蝙蝠SARS样冠状病毒序列相似度最高

COVID-19





生物信息学学科先驱/创始人

Margaret Dayhoff

Michael Waterman

以Dayhoff的PAM替换矩阵和Waterman的序列比对算法为代表，它们的出现代表了生物信息学的诞生。它们实际组成了生物信息学的一个最基本的内容和思路：序列比较。

生物信息学发展过程中的里程碑性事件

时间	事件
1962	**Pauling提出分子进化理论**
1967	**Dayhoff构建蛋白质序列数据库**
1970	**Needleman-Wunsch算法被提出**
1981	**Smith-Waterman算法出现**
1982	**Genbank数据库公开**；EMBL创立
1983	Wilbur和Lipman提出序列数据库的搜索算法
1985	快速序列相似性搜索程序FASTP/FASTN发布
1988	**NCBI成立**
1990	**BLAST发布**
1991	表达序列标签（EST）被提出，开创EST测序
1993	Sanger中心建立
1994	欧洲生物信息学研究所成立
1995	**第一个细菌基因组测序完成**
1996	**酵母基因组测序完成**
1997	PSI-BLAST(BLAST系列程序之一)发布
2001	人类基因组草图公布
2005	**第二代高通量测序技术出现**
2008	基于高通量测序的转录组测序技术RNA-Seq等出现
2009	第三代高通量测序技术出现
2018	DeepMind公司发布AlphaFold预测蛋白质结构软件

Major research areas

Genome Sequencing

Sequence Alignment

Protein Structure prediction

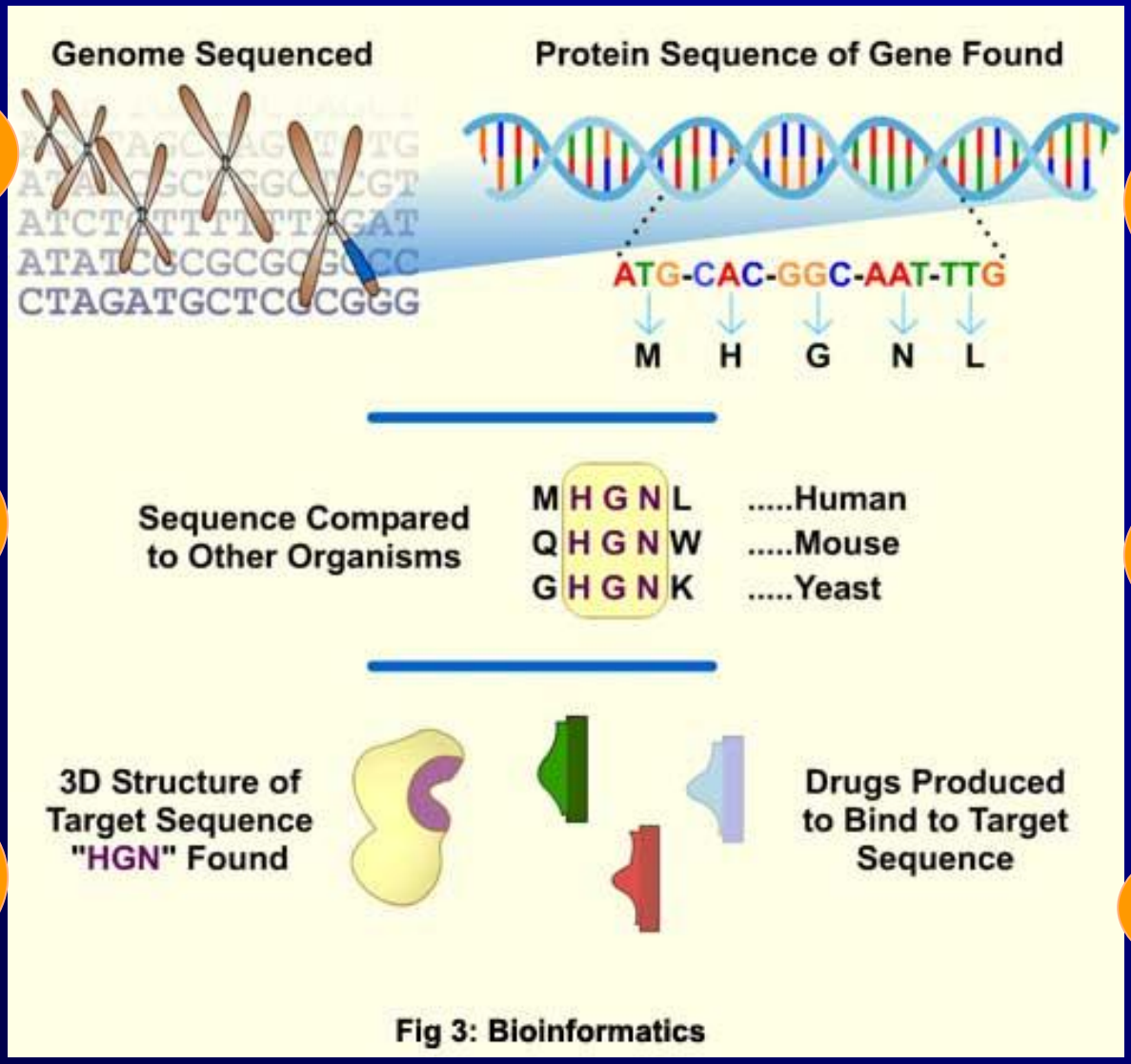


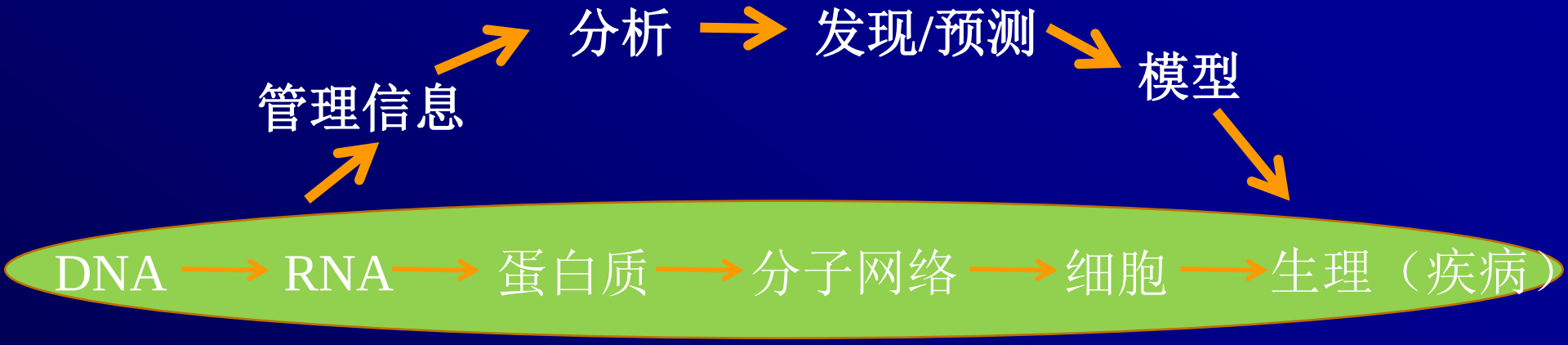
Fig 3: Bioinformatics

Gene finding

Molecular Evolution

Protein Docking

研究内容



序列比对

发现新基因、物种进化树、DNA拼接等

RNA表达水平差异、非编码RNA鉴定

蛋白质鉴定、结构预测、比对等

构建蛋白质相互作用网络、转录调控网络、代谢网络、信号传导网络等

细胞模拟、虚拟生命

人类遗传学研究、
药物设计与筛选

Science Paradigms(模式)

“……新的生物学研究模式的出发点应该是理论的。科学家将从理论推测出发，然后再返回到实验中去，追踪或验证这些理论假设。……生物学家不仅必须成为计算机学者，而且也要改变他们研究生命现象的途径。”

——W. Gilbert, Towards A Paradigm Shift in Biology, Nature, 349(1991)

Biology in the 21st century is being transformed from a purely lab-based science to an information science as well.

走向真正的科学

物理学

物理现象的观测

经验阶段

物理学的描述

理性阶段

数学的描述

$F=ma$

$E=MC^2$

生命科学

生命现象的观测

经验阶段

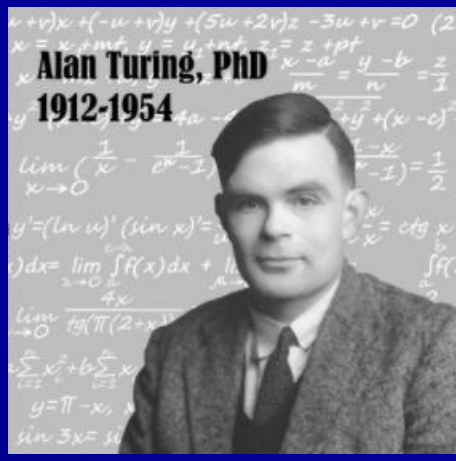
生物学的描述

理性阶段

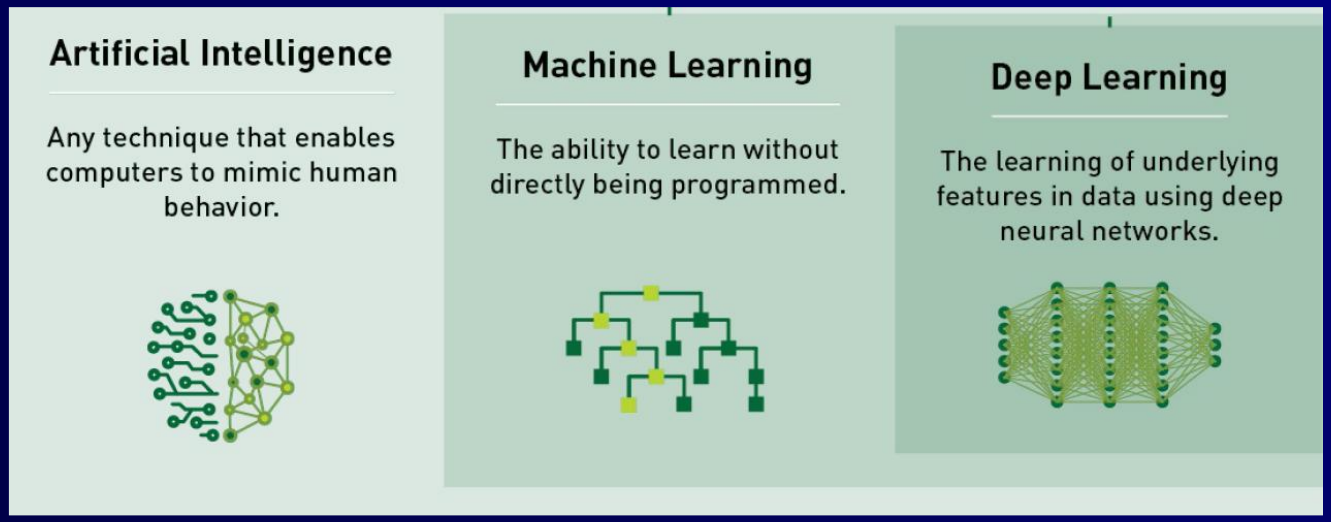
数学的描述

？

人工智能基础

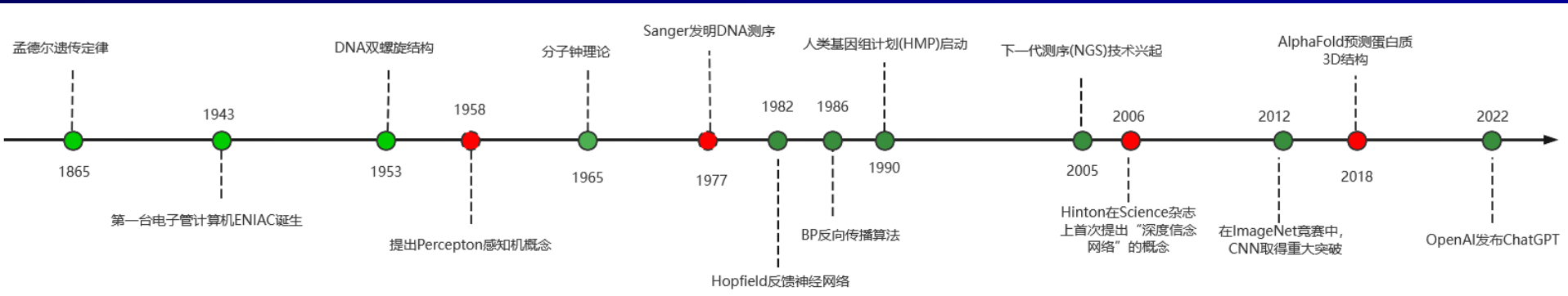


- 人工智能：用机器模拟人的智能
  - John McCarthy，1956年达特茅斯会议
- 机器学习：让机器从数据中学习
  - Alan Turing (阿兰•图灵)，1950年发表论文“Computing Machinery and Intelligence”
- 深度学习：深层神经网络模型
  - Geoffrey Hinton，2024年诺贝尔物理奖

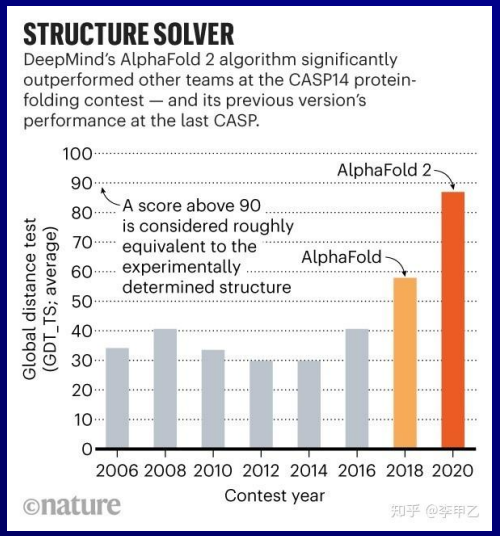


人工智能与生物信息学

- 人工智能一直与生物信息学研究交叉融合发展。近年来，随着高通量实验技术的发展，机器学习（深度学习）被广泛应用于生物大数据分析。



- Alphafold2预测了几乎所有蛋白质(2.14亿)结构（2022）
  - 人工智能技术提高药物研发效率。

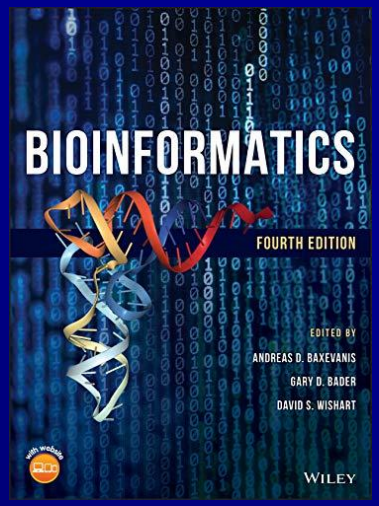


本课程主要内容

1. 生物学与计算机基础：**BioEdit, Linux**
2. 生物数据库：GenBank, PBD
3. 序列比对：**ClustalW, BLAST**
4. 分子进化与系统发育树：MEGA
5. 蛋白质结构预测：**PyMOL**
6. **PCR**引物设计：**Primer Premier5**
7. 基因组学：DNA测序、组装与注释
8. 下一代测序 **(NGS)**：**WGS, RNA-seq, Microbiome**
9. 统计与编程基础：**R/Python**
10. 人工智能在生物信息学中的应用

参考教材&配套资源

- 李余动编著，生物信息学与基因组分析入门，2021，浙江大学出版社
- Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins, 4th Edition, 2020
- 教材网站：https://bigbook.pages.dev/
  - 课程PPT、数据或软件下载



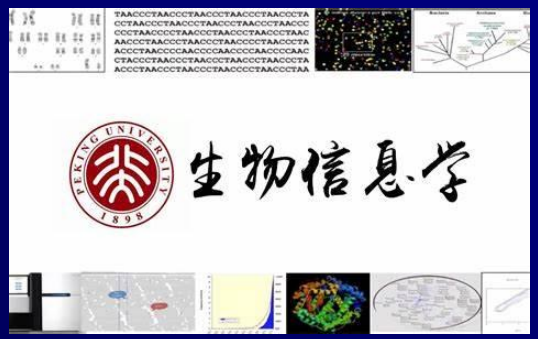
超星学习通慕课：

- https://mooc1.chaoxing.com/course/240919024.html

生物信息学MOOC课程

- 生物信息学(山东大学）
  - **中国大学MOOC精品课程：**

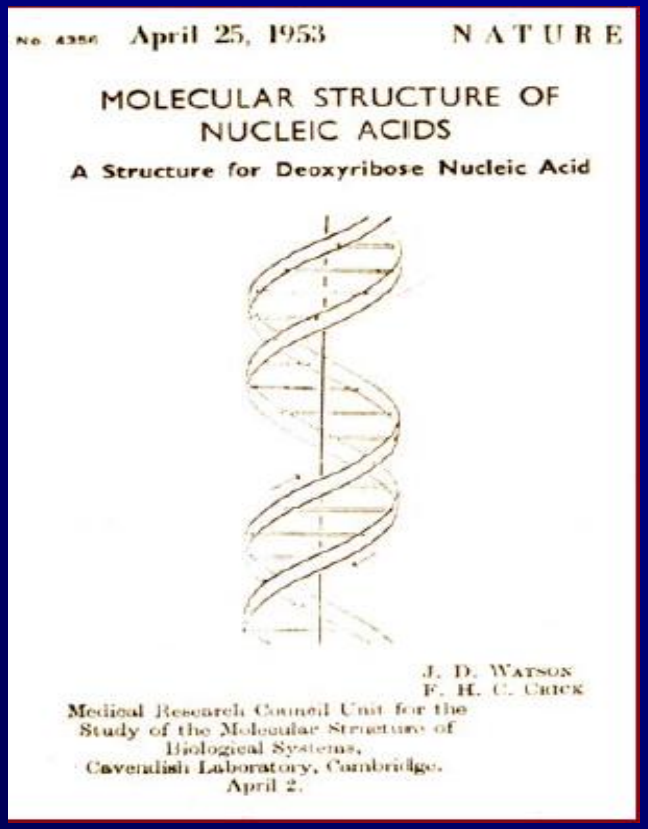
    https://www.icourse163.org/course/SDU-1001907001
  - **生信基础与应用**
- 生物信息学-导论与方法(北京大学)
  - 华文慕课: **http://www.chinesemooc.org/mooc/4393**
  - **下一代测序与算法**



课程考核方式

- 平时作业
  - 每周一次课堂/课后作业（超星学习通）。
  - 在线讨论：学习问题、教学问题等。
- 课程报告
  - 查阅相关文献资料，介绍一块生物信息学内容，如软件、算法、数据库等；
  - 结合自己的兴趣，构建一个生信相关的AI智能体（字节COZE、百度文心等平台）。
- 评价标准
  - 平时成绩（50%）：课后作业+在线讨论
  - 期末成绩（50%）：小组课程报告(2~3人/组)

Primer of Molecular Biology



No. 4356 April 25, 1953 NATURE

MOLECULAR STRUCTURE OF NUCLEIC ACIDS

A Structure for Deoxyribose Nucleic Acid

J. D. Watson
F. H. C. Crick

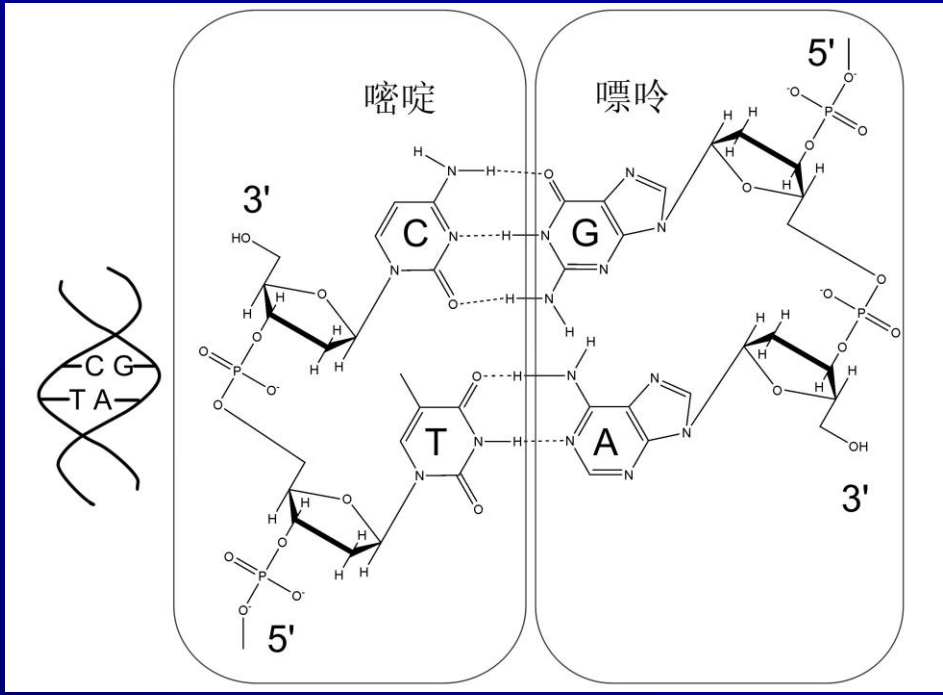
Medical Research Council Unit for the Study of the Molecular Structure of Biological Systems, Cavendish Laboratory, Cambridge. April 2.

Molecules of Life

- Types of molecules in cells
  - DNA: genetic material
  - Protein: 3D structures
  - RNA: intermediary between DNA and proteins
- Role of molecules in cells
  - Pass on the instructions for making an organism
  - Perform various chemical reactions necessary for life

DNA: genetic material

- Double Helix
- Basic unit = nucleotide
  - 5-Carbon Sugar
  - Phosphate
  - Base: (A, G, T, C)



- Repeating backbone of sugar and phosphate
  - 10bp per turn (34 Å)
  - Each bp = 3.4 Å. (linear information storage density is ~6 $\times 10^8$ bits/cm)

Why is DNA double-stranded?

- Base pairs (A-T, G-C) are complementary, Known as Watson-Crick bps
- A double-stranded DNA sequence can be represented by strings of letters.

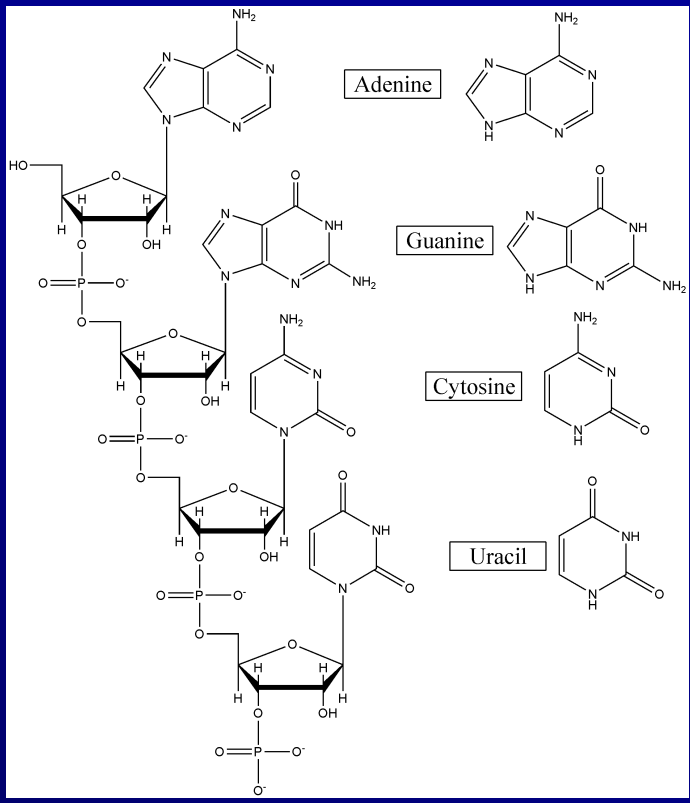
  5' ... TACTGAA ... 3'

  3' ... ATGACTT ... 5’
- Length of DNA in bps (e.g. 100kbp)
- 任意长度大于4的一串核苷酸被称作一个序列，例如序列AAAGTCTGAC。

bp = base pair(碱基对)
kbp = kilo base pair(千碱基对)

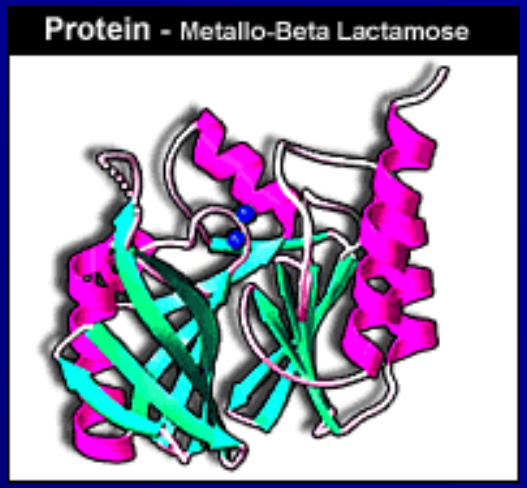
RNA: intermediary between DNA and proteins

- Chemically, RNA is very similar to DNA. There are some main differences:
  - RNA uses the sugar ribose instead of deoxyribose in its backbone.
  - RNA uses the base Uracil (U) instead of Thymine (T). U is also complementary to A.
  - RNA tends to be single-stranded.
- Example of types of RNA:
  - tRNA, mRNA, rRNA, microRNA, circRNA…



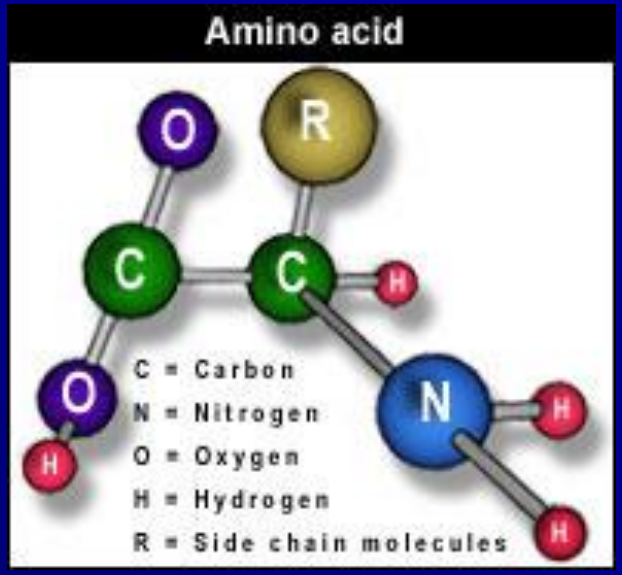
Proteins: the building blocks of life

- Different Roles of Proteins
  - Enzymes
  - Carry signals
  - Transport small molecules
  - Form cellular structures (tissues)
  - Regulate cell processes (such as defense mechanisms)

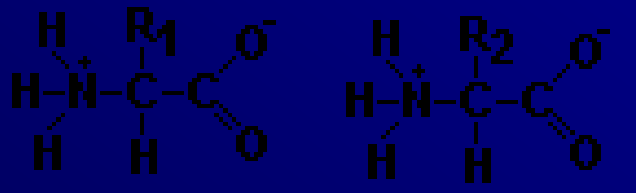


Proteins made of Amino Acids

- Proteins consist of amino acids linked by peptide bonds
- Each amino acid consists of:
  - a central carbon atom
  - an amino group
  - a carboxyl group
  - a side chain (Differences in side chains distinguish the various amino acids)
- 20 different amino acids found in nature:
  - alphabet of 20 letters



- Peptide bond formation:
  - Loss of water leaves residue of original amino acid – thus, protein has residues

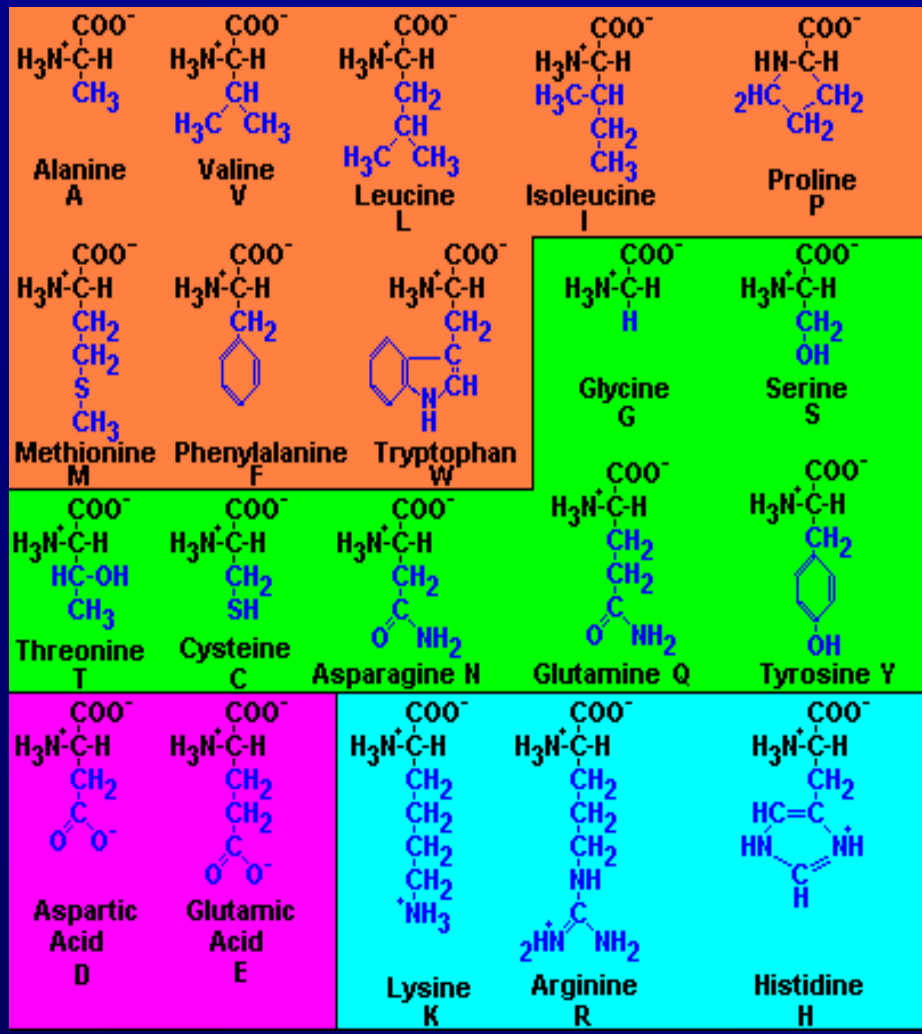


20 amino acids

- Convention: 3-letter or 1-letter
- Amino Acid side chains vary in:
  - Size
  - Shape
  - Polarity

- 8: nonpolar and hydrophobic

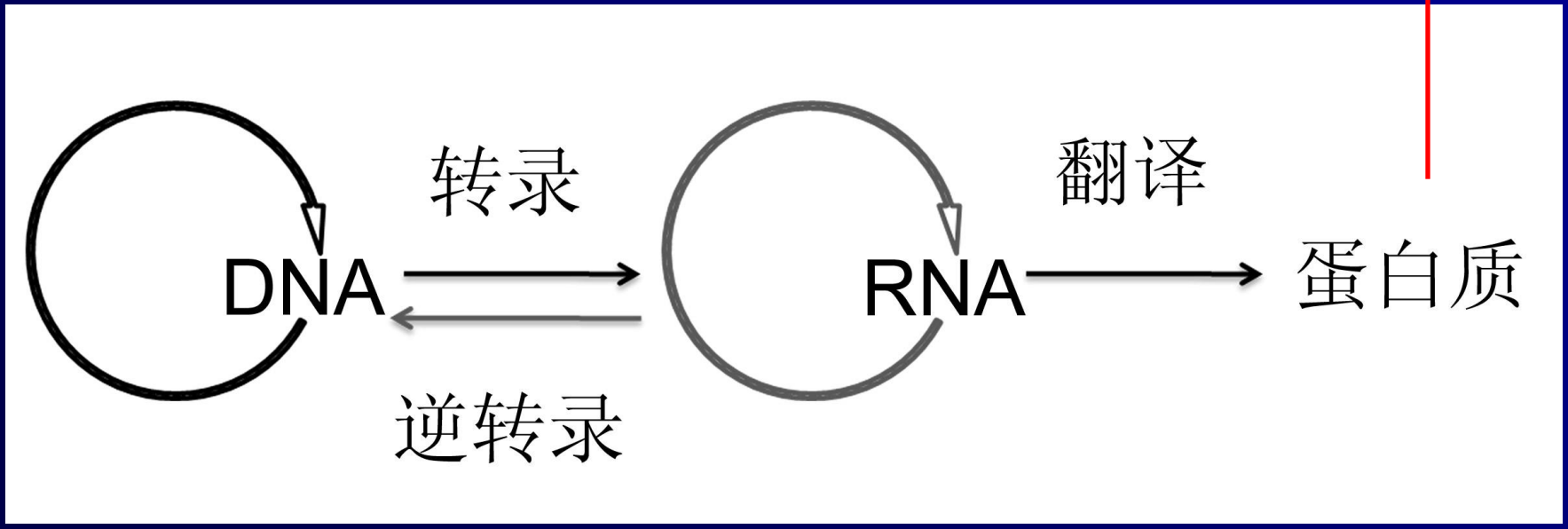
- 12 are polar and hydrophilic
  - 2: acidic, negatively charged
  - 3: basic, positively charged



How genes encode proteins

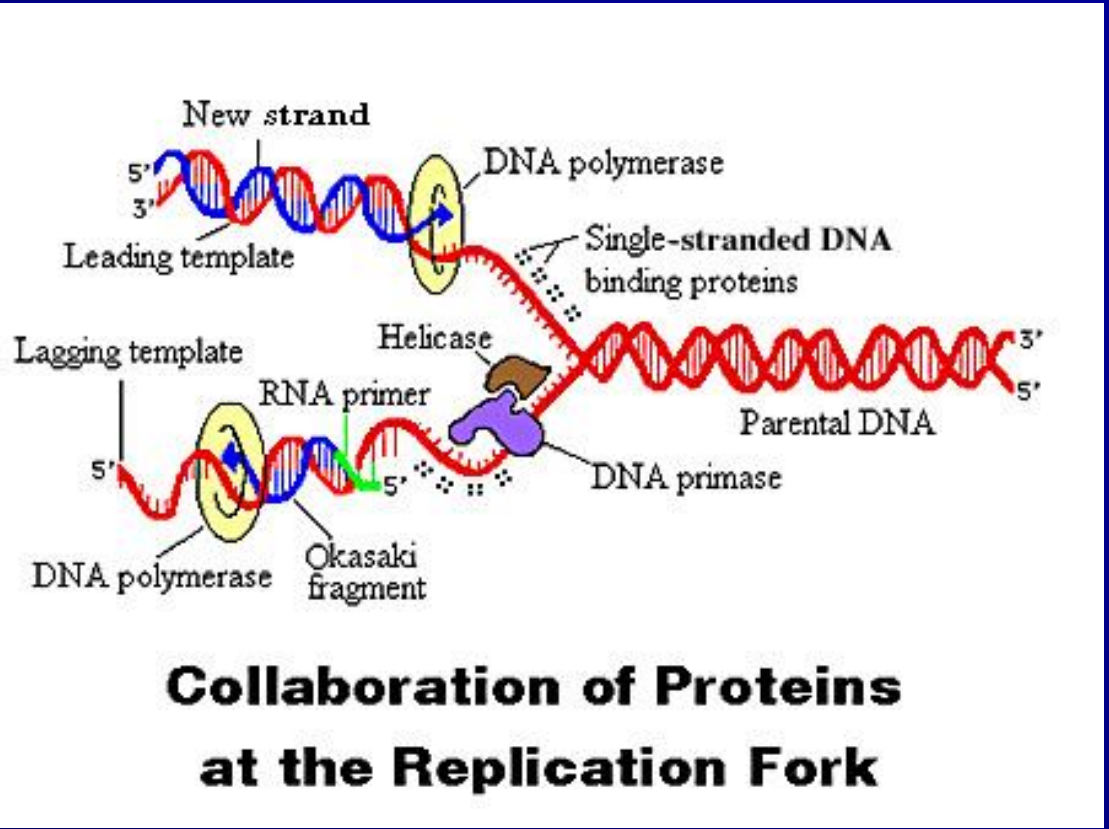
- Central Dogma （中心法则）
  - Replication: DNA synthesis
  - Transcription: RNA synthesis
  - Translation: Protein synthesis

Traits/Phenotypes

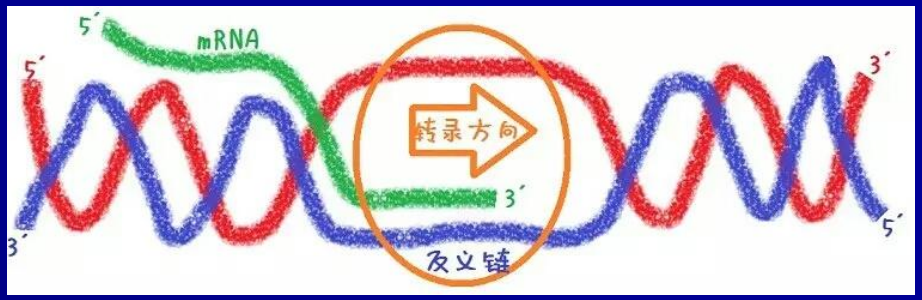


DNA replication

- DNA → DNA
- DNA polymerase



Transcription

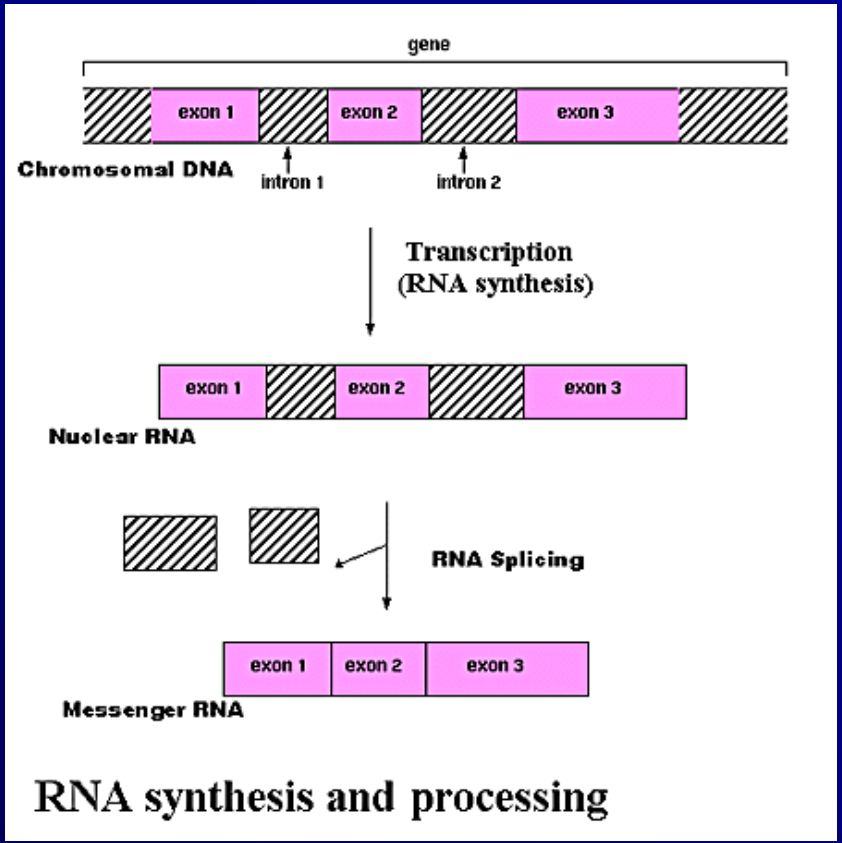


- DNA → RNA (Replace T with U)

  5‘ ... TACTGAA ... 3’ (模板链)

  5‘ ... UACUGAA ... 3’ (编码链)

- RNA splicing
  - Introns are removed
  - Exons are retained and joined

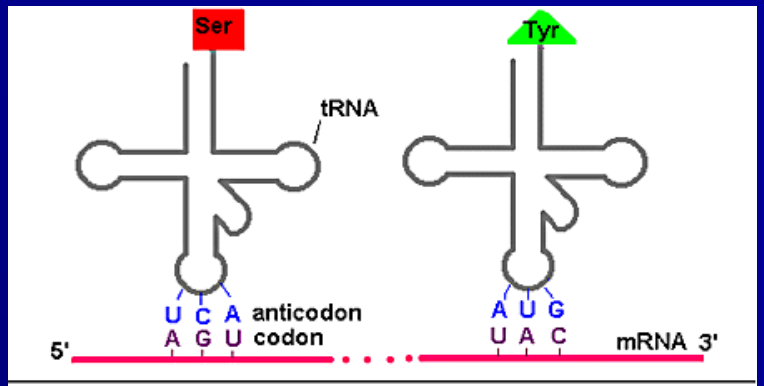


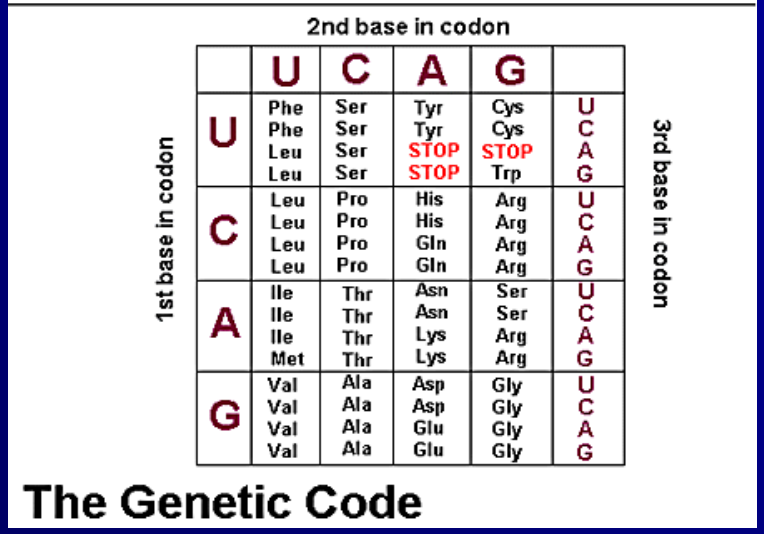
RNA synthesis and processing

Translation: protein synthesis

- mRNA → Protein
- Ribosome
- Codon
  - unit of 3-nucleotide in mRNA
- Anticodon
  - unit of 3-nuclotide in tRNA
- Genetic code: 64 triplet codons
  - 3 stop codon
  - 61 codons to specify only 20 different amino acids -- most of the aa's are represented by ≥ 1 codon.

Genetic code is degenerate!!!





ORF: open reading frame

- 开放阅读框：从起始密码子开始，直到终止密码子结束的密码子序列。
- 阅阅读框是由起始密码子决定的，只有当核糖体在正确的相位或阅读框(reading frame)中阅读，才能够准确地翻译。

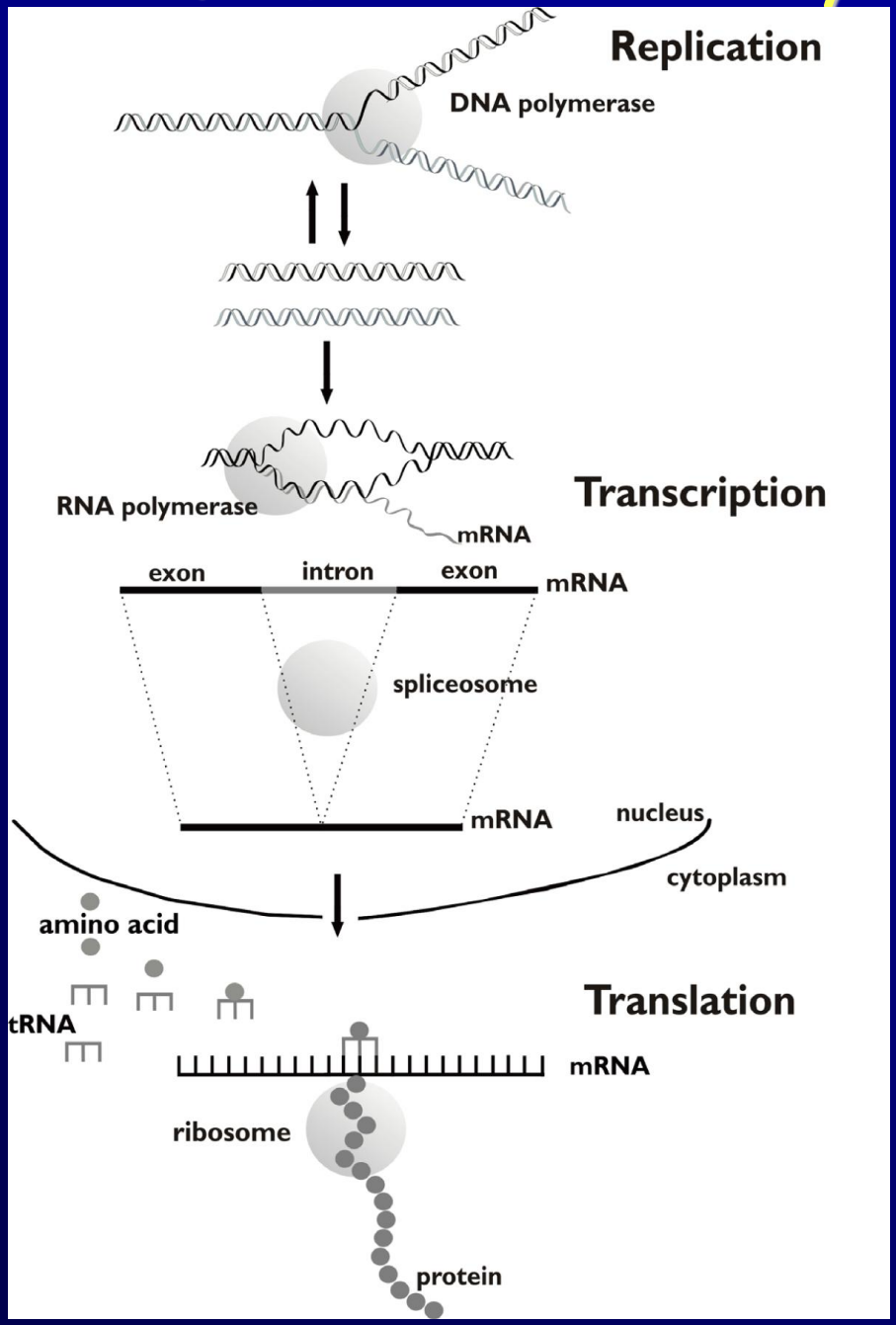
5' **ATG**CCCAAGCTGAATAGCGTAGAGGGGTTTTCATCATTTGAG**TAA** 3'

阅读框：

```
1 atg ccc aag ctg aat agc gta gag ggg ttt tca tca ttt gag taa
   M   P   K   L   N   S   V   E   G   F   S   S   F   E   *
2 tgc cca agc tga ata gcg tag agg ggt ttt cat cat ttg agt
   C   P   S   *   I   A   *   R   G   F   H   H   L   S
3 gcc caa gct gaa tag cgt aga ggg gtt ttc atc att tga gta
   A   Q   A   E   *   R   R   G   V   F   I   I   *   V
```

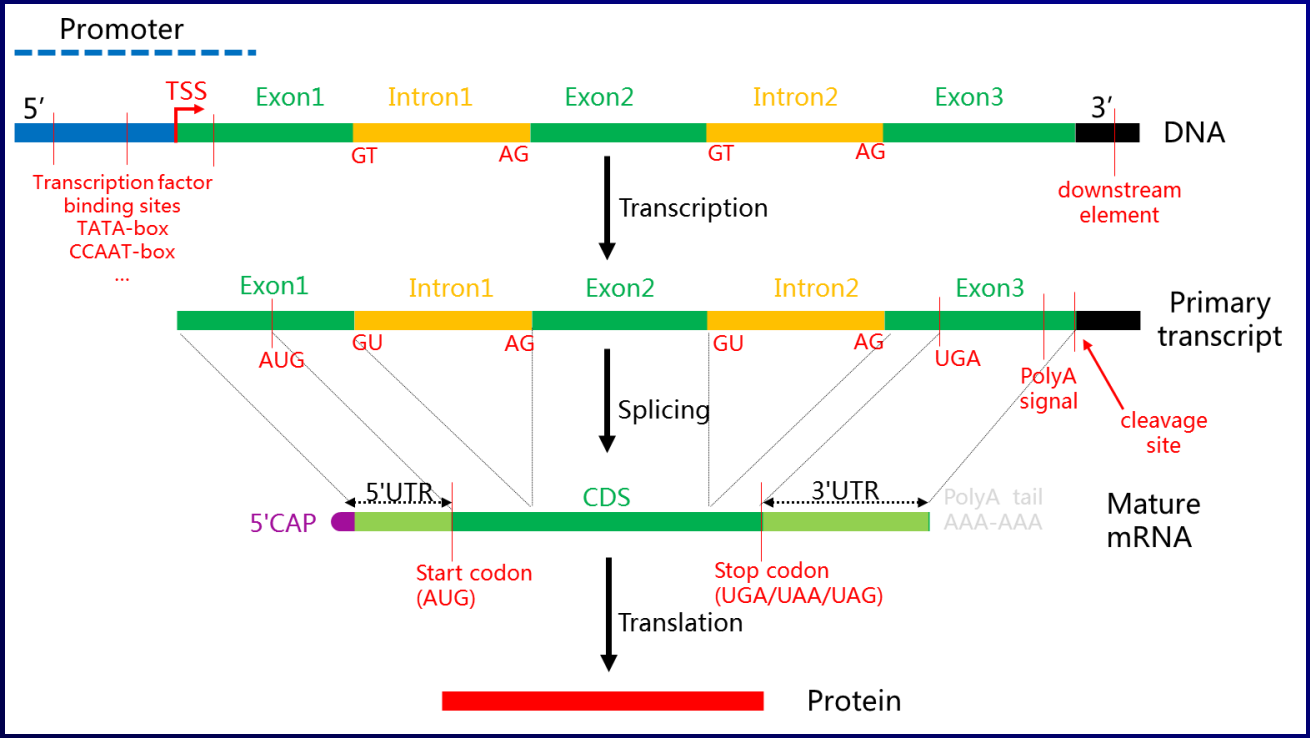
思考：ORF与基因的区别？

Flow of genetic information in a eukaryotic cell



真核生物的基因结构

- 内含子(Intron)与外显子(Exon)
- 非翻译区（UnTranslated Region，UTR）、启动子（promoter）、增强子(enhancer)区域等



DNA sequence

```
1   gtcgacccac gcgtccgtct tgaaagaata tgaagttgta aagagctggt aaagtggtaa
61  taagcaagat gatggaatct ggggctccta tatgccatac ctgtggtgaa caggtggggc
121 atgatgcaaa tggggagcta tttgtggctt gccatgagtg tagctatccc atgtgcaagt
181 cttgtttcga gtttgaaatc aatgagggcc ggaaagtttg cttgcggtgt ggctcgccat
241 atgatgagaa cttgctggat gatgtagaaa agaaggggtc tggcaatcaa tccacaatgg
301 catctcacct caacgattct caggatgtcg gaatccatgc tagacatatc agtagtgtgt
361 ccactgtgga tagtgaaatg aatgatgaat atgggaatcc aatttggaag aatcgggtga
421 agagctgtaa ggataaagag aacaagaaga aaaagagaag tcctaaggct gaaactgaac
```

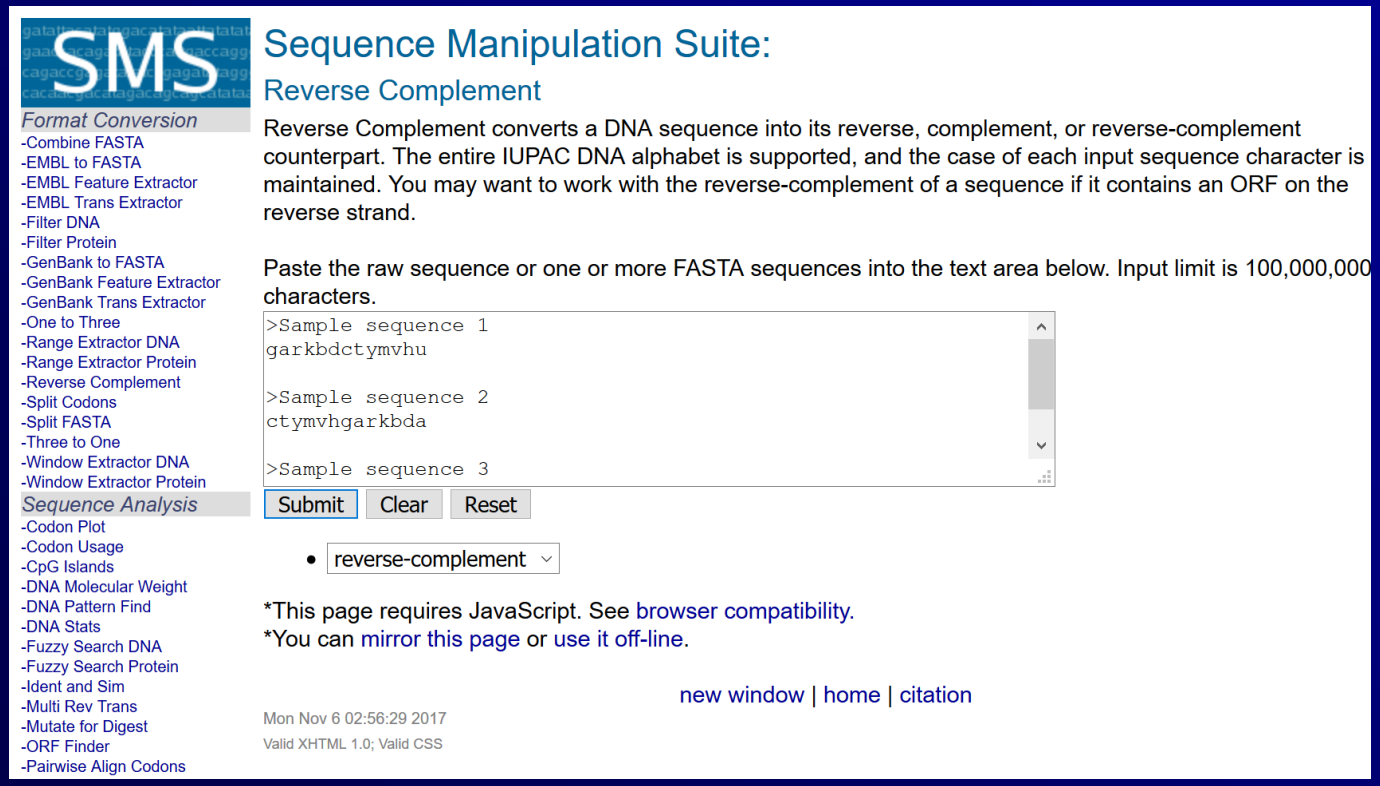
- **Protein coding regions of Genes begin with ATG and end with either TAG, TGA or TAA**

- atg atg gaa tct ggg gct cct… use genetic code..
- M M E S G A P …….*

- Study function of proteins and expression of genes in different organs and tissues

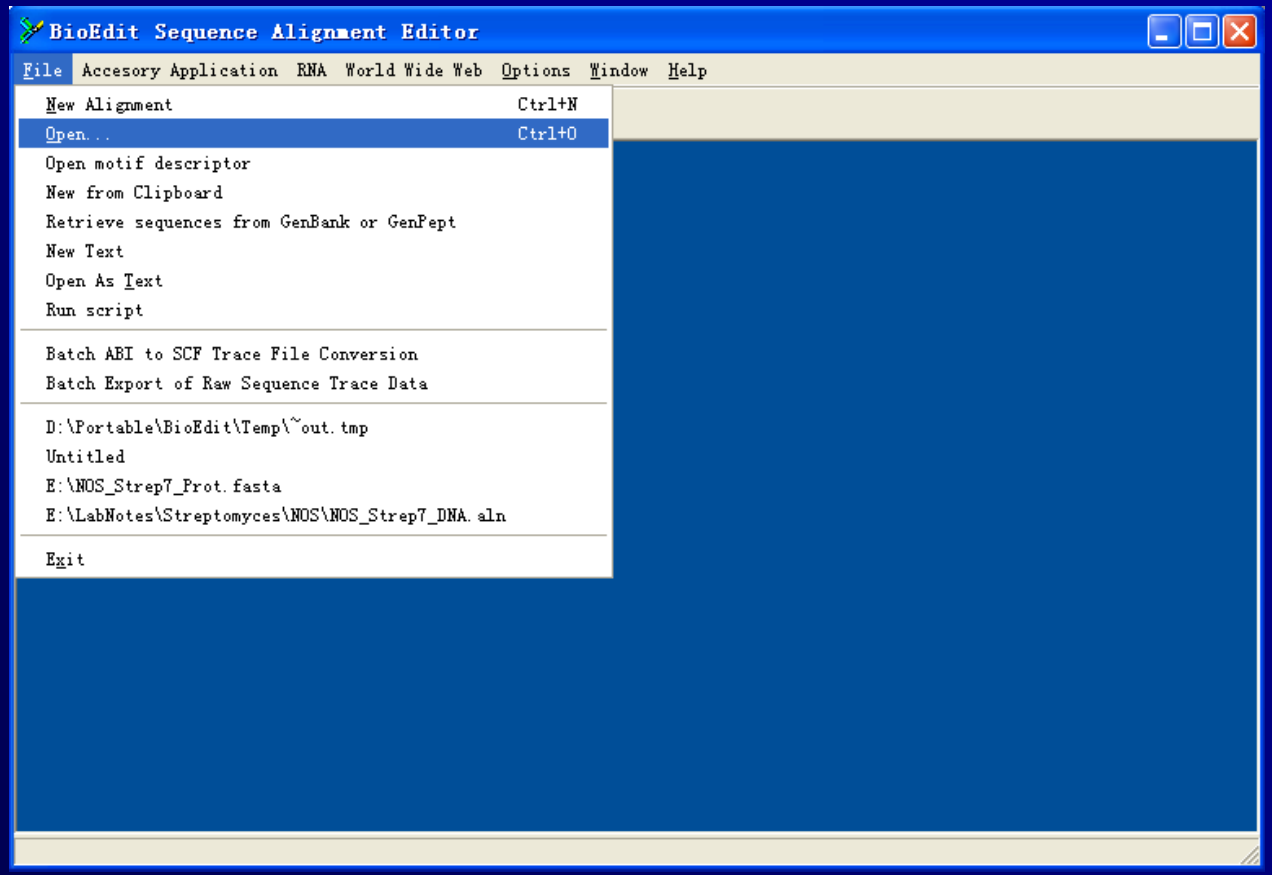
在线工具SMS (Sequence Manipulation Suite)

- SMS是用JavaScript编写的序列操作工具，可在任何网络浏览器中运行，只要打开SMS网站即可使用。
- SMS主页左边栏列出了多种常用分析工具。



http://www.bioinformatics.org/sms2/index.html

生物软件：BioEdit



https://thalljiscience.github.io/

bioedit功能列表

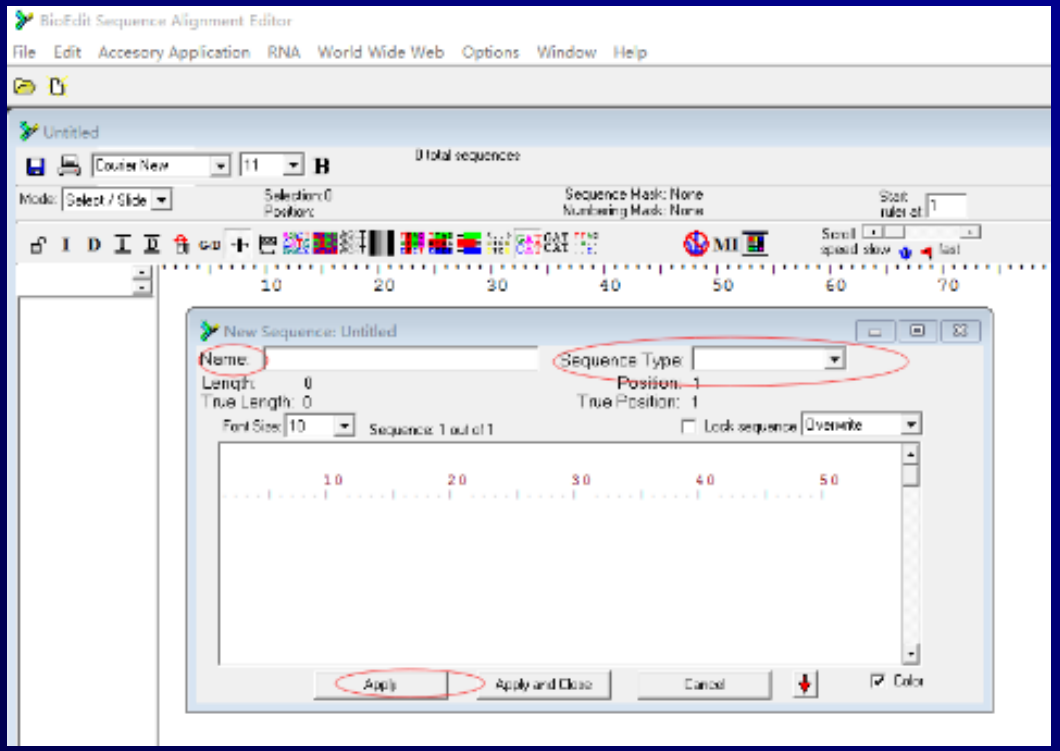
功能	描述
序列输入	多种序列输入方式;
序列分类	按标题、位置 、定义、参数、注释等分类;
成对排列	两序列的最佳排列及计算同一性和类似性;
序列屏蔽	仅采用联配中部分区域进行分析而排除其他。
核酸分析	组成、互补、反转、翻译、质粒、限制性内切酶;
蛋白质分析	氨基酸成分、疏水性轮廓、疏水力矩平均数
翻译或反翻译	把 DNA 或 RNA 翻译成蛋白质;
切换翻译	在核酸和编码蛋白质序列中切换核苷酸序列;
点图	相互比较两序列的矩阵，生成一个点图。
RNA 比较分析	共变; 潜在配对; 互交信息分析
BLAST	本地 Blast; BLAST INTERNET 客户端程序
Cluastal	多序列比对
进化分析	
使用互联网工具	HTML BLAST 网络浏览器 ;PSI-BLAST ;nnPredict …

1. 导入序列方法

1.File→Open…

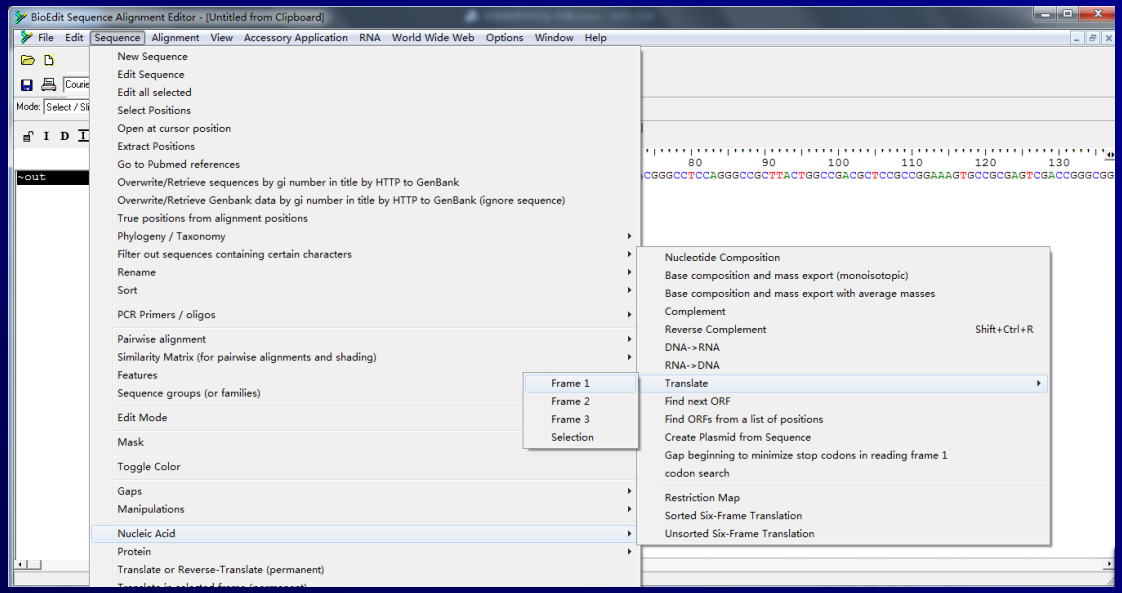
2.File→New from Clipboard

3.Sequence→New sequence



DNA序列翻译成蛋白质序列

- Sequence→Nucleic Acid→Translate



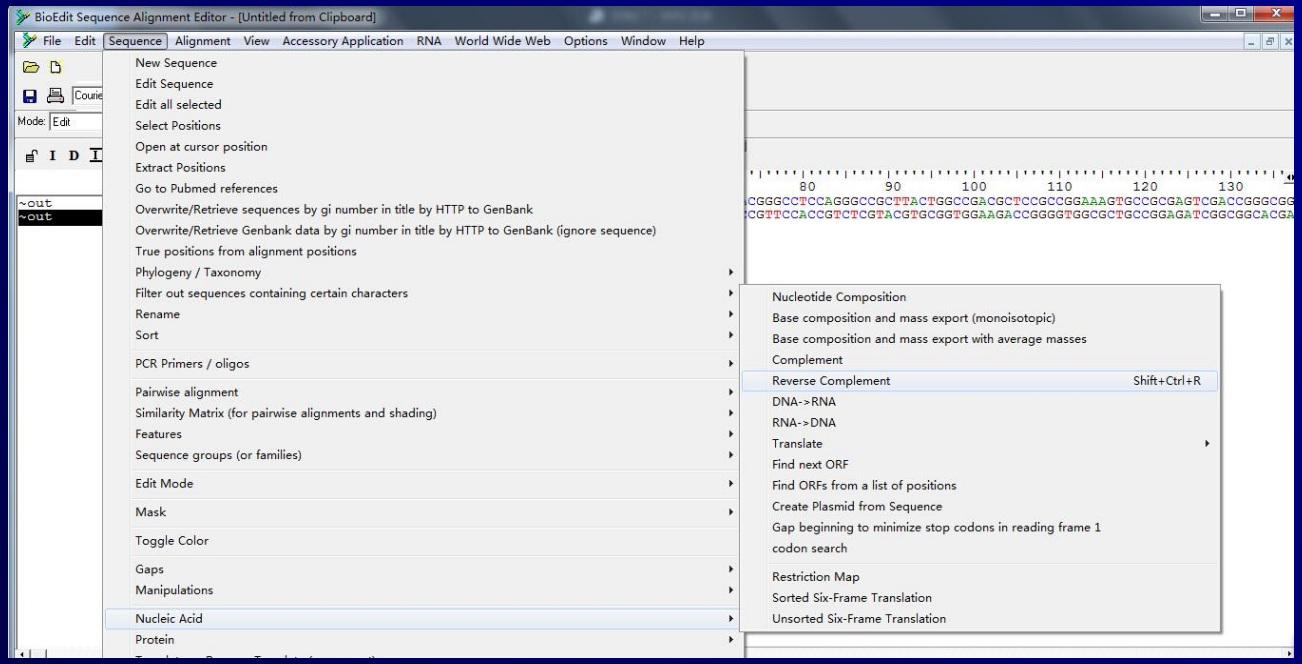
- 在菜单栏选择Sequence→Toggle Translation即可，也可使用快捷键Ctrl+G；

反向互补序列

5' ... TACTGAA ... 3'

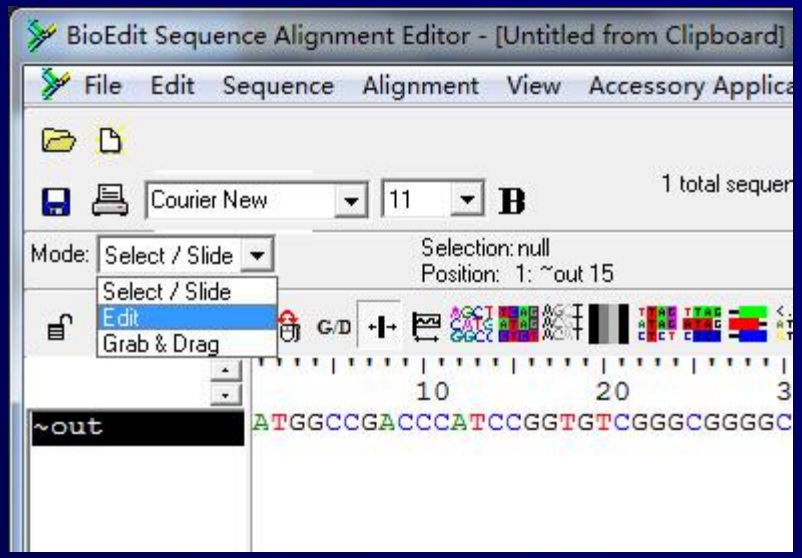
3' ... ATGACTT ... 5'

- Sequence→Nucleic Acid→Reverse Complement



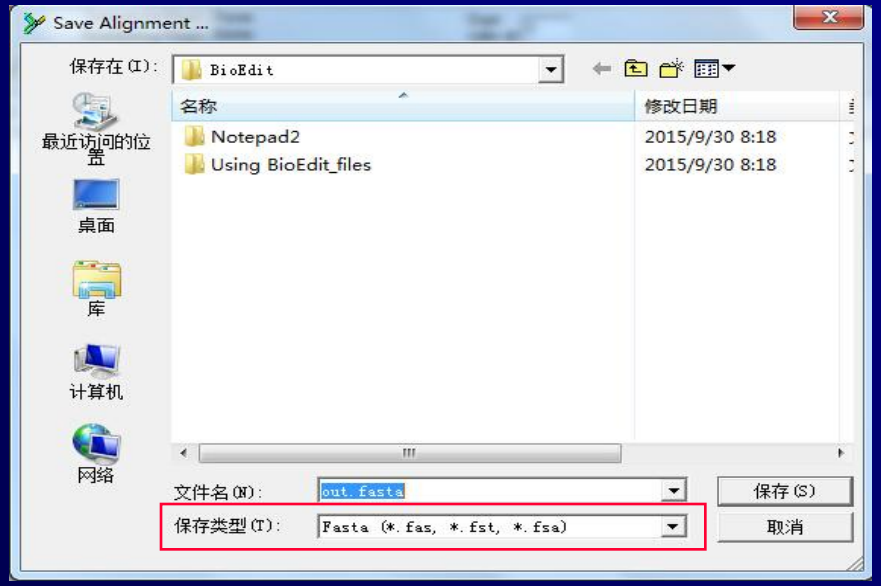
编辑序列

- 通过改变Mode为Edit可对选中序列进行编辑，分为Overwrite 和Insert两种形式。



保存序列

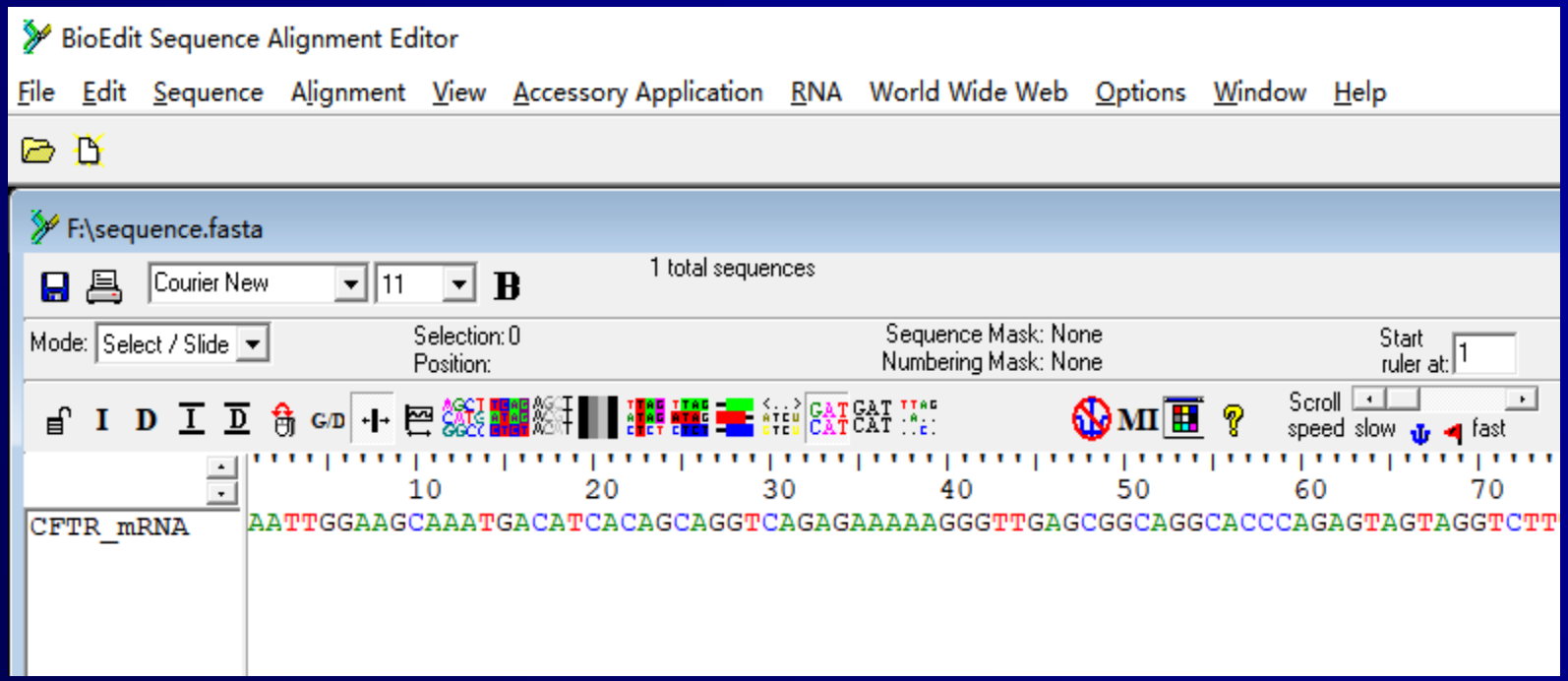
- 菜单选择File→Save as，保存文件名后加 .fasta，保存类型选Fasta (*.fas)。序列即保存为fasta格式。



BioEdit序列处理

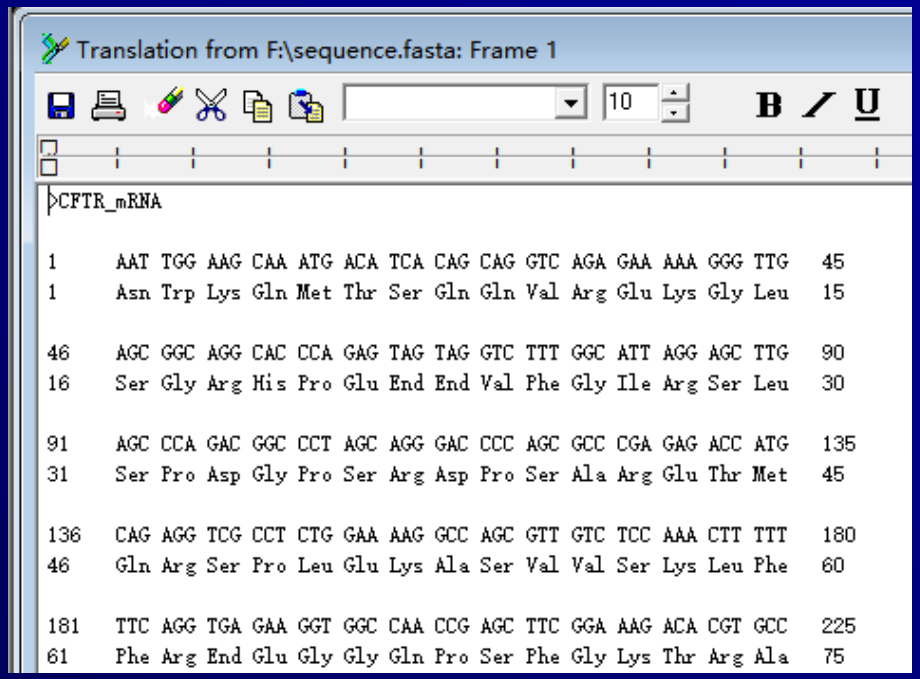
（1）从NCBI网站的gene数据库检索并下载CFTR基因的mRNA序列(索引号：NM_000492.3)，保存为fasta格式文件(CFTR_mRNA.fasta)，并用记事本打开把注释行改成:>CFTR_mRNA。

（2） 用BioEdit打开CFTR_mRNA.fasta文件。



（3） 如果想得到此mRNA的DNA 模板链(template)序列，可以通过BioEdit菜单 Sequence -> Nucleic acid -> Complement得到。

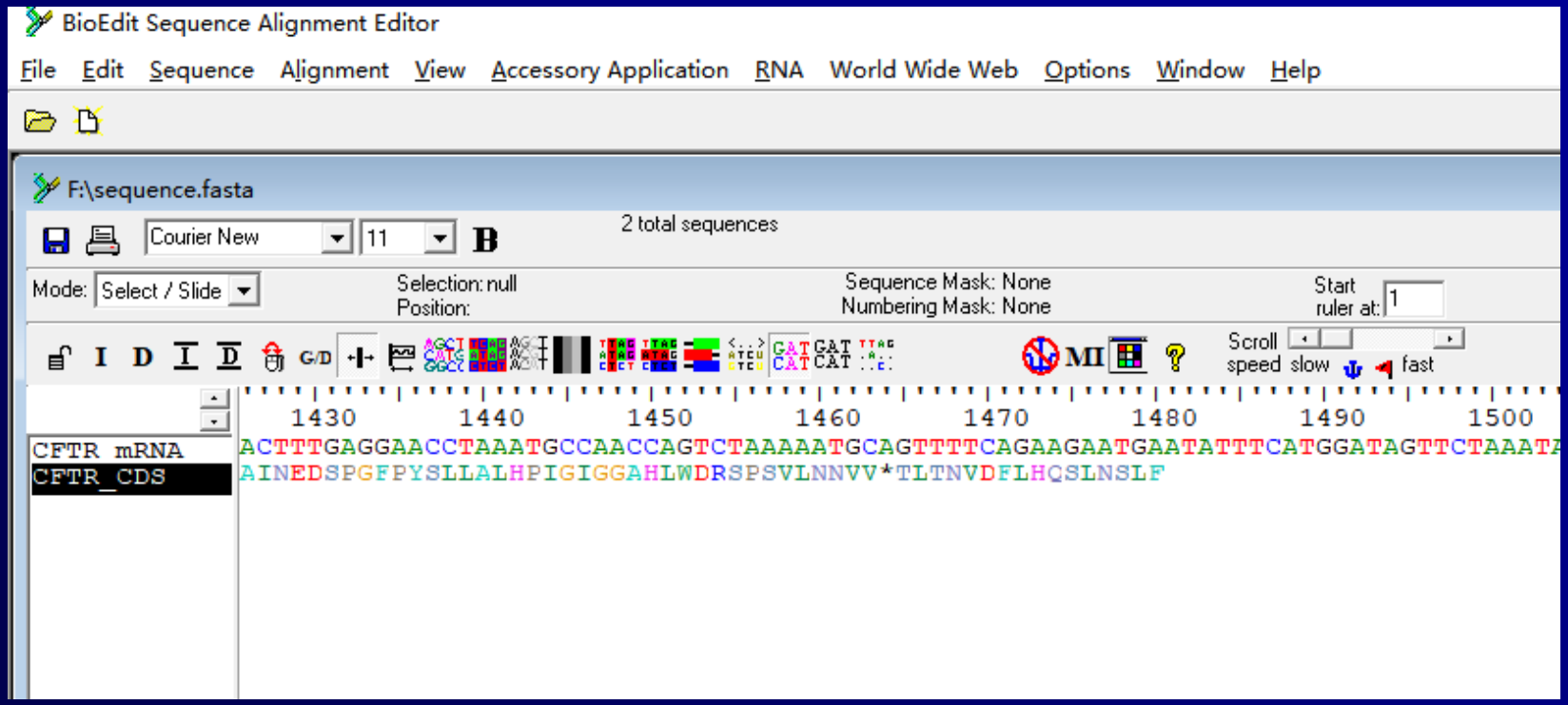
（4）将此mRNA翻译成蛋白质：Sequence -> Nucleic acid ->Translate



（5）通过氨基酸序列找到M（ATG）与终止密码子(*)的位置。菜单 Sequence->Select positions，拷贝得到编码蛋白的DNA序列(Edit->Copy)。检查起始密码ATG，与终止密码TAG。

注：此处CDS位置为133..4575。

（6）新建CDS序列：菜单Sequence->New sequence，在跳出的窗口中粘贴CDS序列(Ctrl-V)， 序列名称(Name)输入“CFTR_CDS”， Sequence type选“DNA”。



（7） 选择CFTR_CDS序列，再用Sequence->Toggle Translate得到CDS序列的氨基酸序列。如果操作都正确，最终的氨基酸序列是1480个氨基酸。

作业

- 学习一种生物软件的基本操作：
  - BioEdit (Win)
  - SnapGene (Mac/Win)