

# 生物信息学

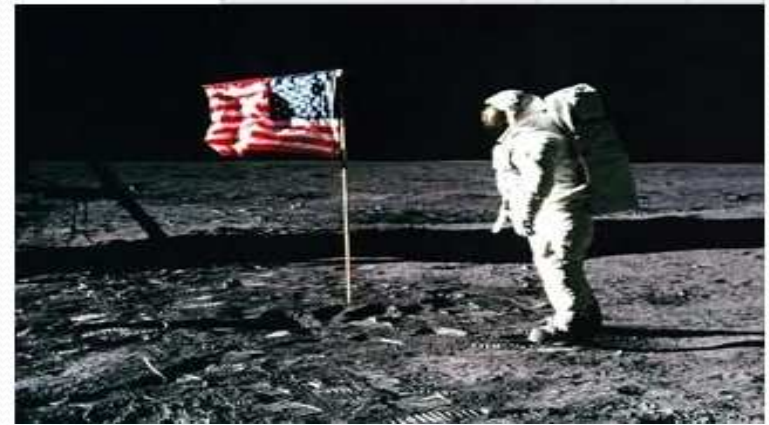
郑丽沙

Lishazheng@buaa.edu.cn

# 生物信息学产生的背景



曼哈顿计划

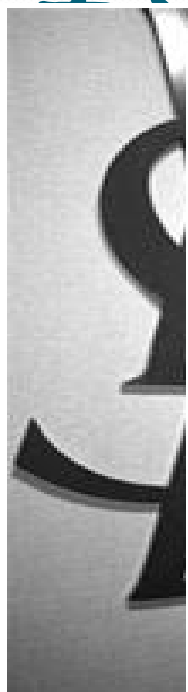


登月计划



人类基因组计划

# 癌



时装大师



苹果创始人



委内瑞拉前总统查韦斯死于未知癌症

著名男高音歌唱家帕瓦罗蒂死于胰腺癌





尼克松总统在1971年提出的对癌症宣战，号称10年内  
攻克癌症

- 2006年5月18日，人类“生命之书”中最长也是最后的一章一号染色体的测序完成标志着人类基因组计划的执行和完成。不仅测定了人类基因30亿个基因组的基对的序列，而且在进行过程中，为大规模模式生物基因组的测序工作可以完成得更经济、快捷，迄今(2009年)已完成积累经验和技术储备，使得其他生物基因组的测序工作可以完成得更经济、快捷，迄今(2009年)已完成了81种生物的组测序工作。

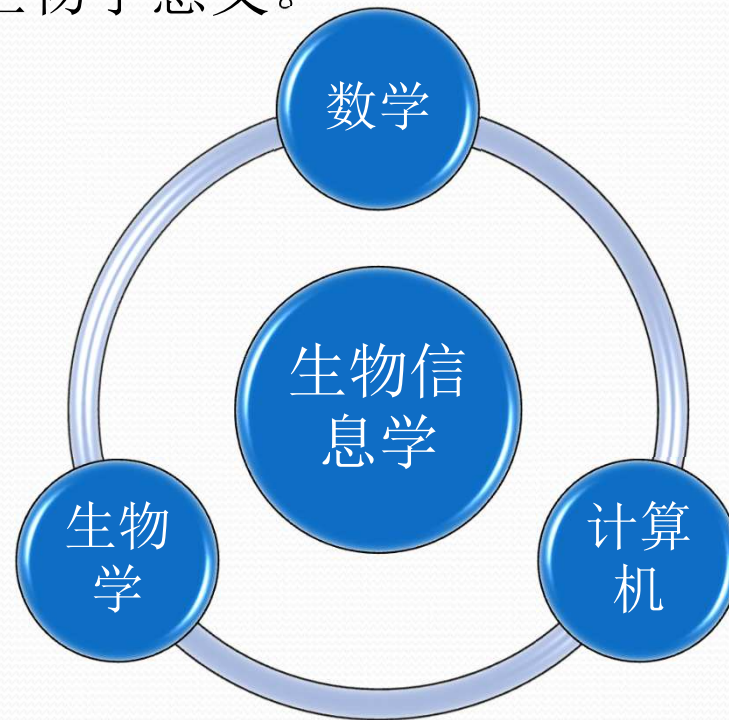




- 人类基因组计划标志着科学进入一个全新时代。生物学家从一个一个去发现、研究自己“喜欢”的基因，到从整个基因组的规模去认识、研究一个物种的所有基因以及物种之间基因的比较。



- 生物信息学(bioinformatics)是一门交叉科学，它包含的所有方面，它综合运用数学、计算机科学和生物学的各种工具，来阐明和理解大量数据所包含的生物学意义。



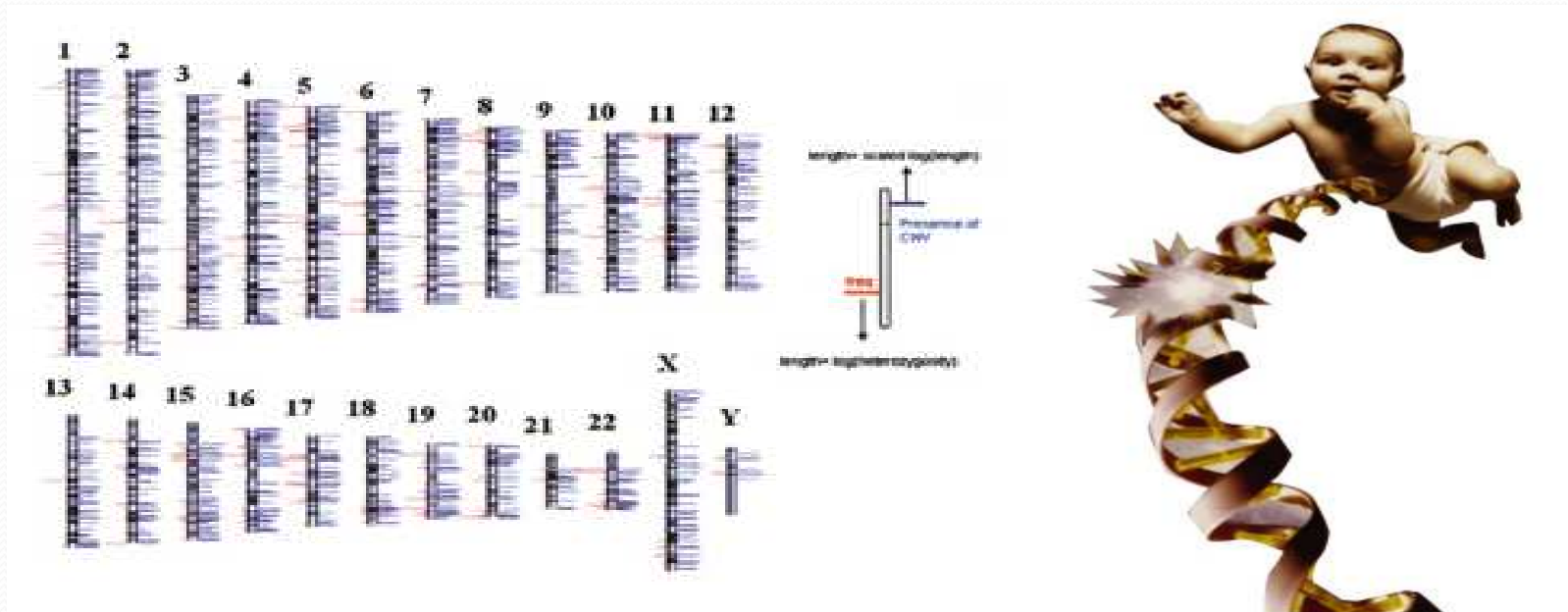


- 基因组信息学的关键是“**读懂**”基因组的核苷酸顺序，即全部基因在染色体上的确切位置以及各DNA片段的功能；然后依据特定蛋白质的功能进行药物设计，了解基因表达在调控中的作用，描述人类疾病的诊断、治疗的内在规律。



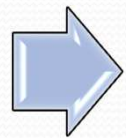


- 通俗地讲，生物信息学就是要弄清楚词、语法和词法，从而认识 and 了解生命的奥秘。从组成“生命之书”的数据中，找到它的书写规律、逻辑关系、语法和词法，以了解生命的奥秘。

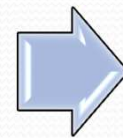




收集数据（采  
样、实验设计）



分析数据（建  
模、知识发现）



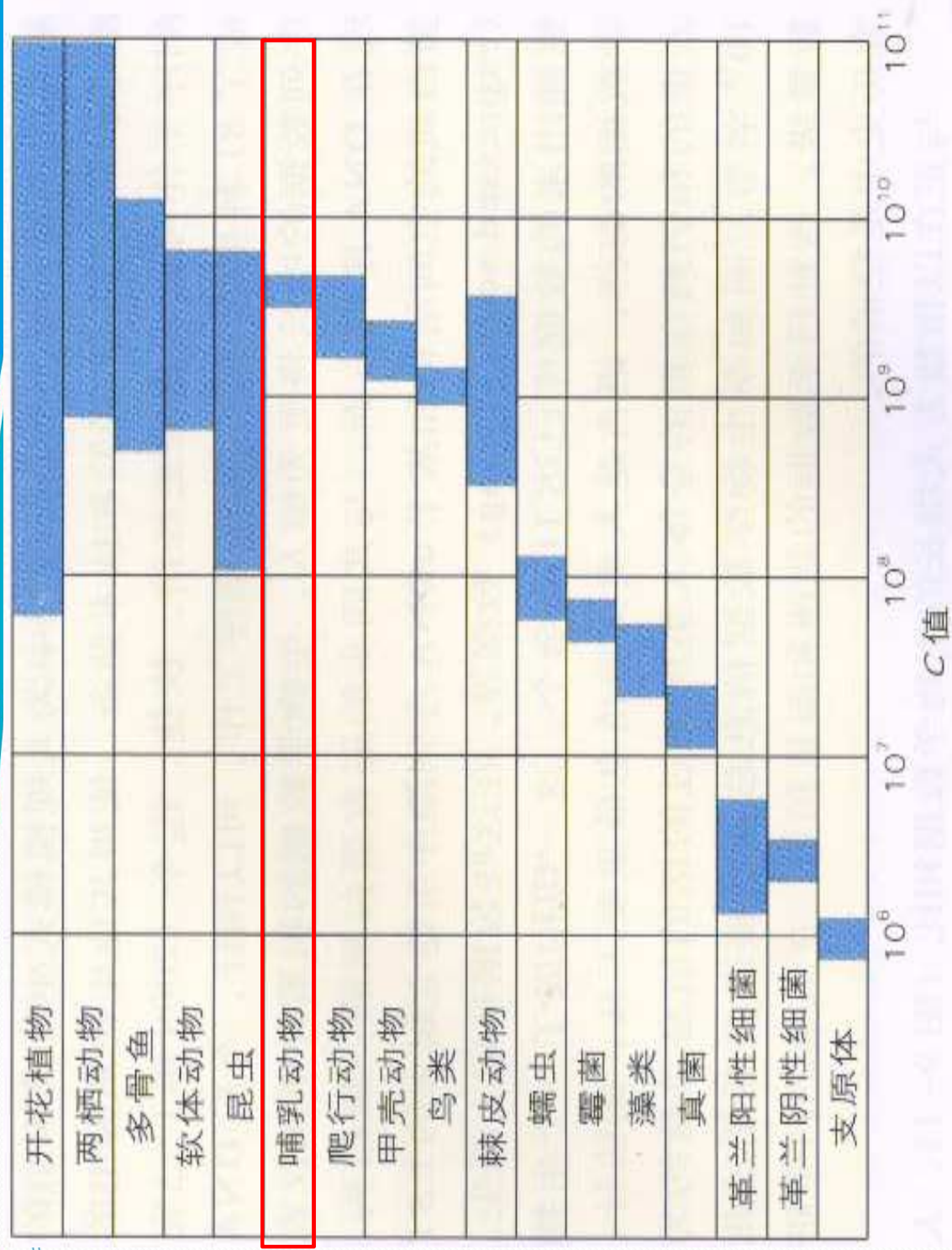
推理（预测、  
分类）




## 基因组结构特点

- 基因组(genome)是指生物体内的细胞中一套完整的遗传信息。科学家在1948年发现，一个生物体的任何一个细胞的DNA数量相同，这种细胞内DNA总数量的量度总称为基因组的C值(C values)。








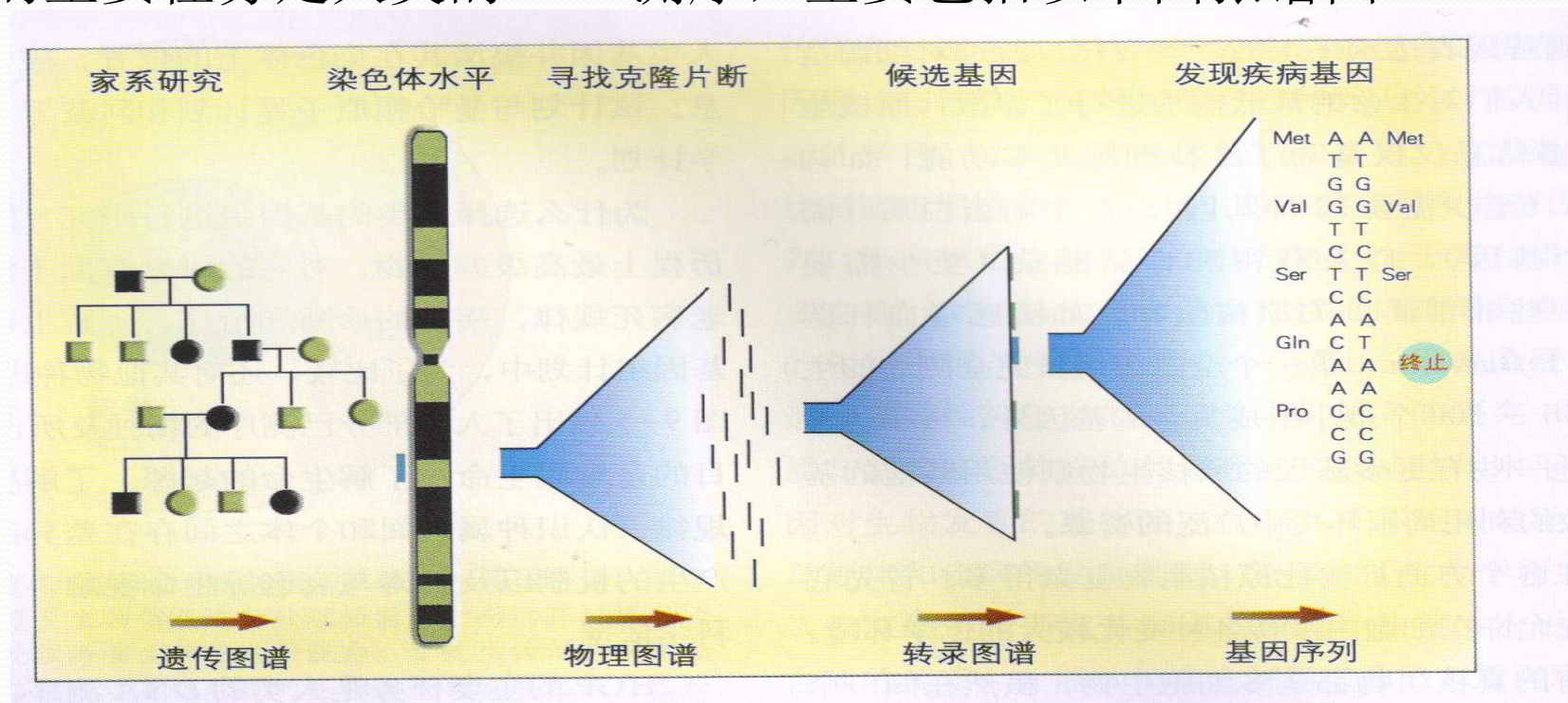
- 
- 有趣的是，一个物种的基因组大小是不变的，但是不同物种间基因组的大小有很大变化，但是基因组的大小与生物体的复杂度并没有很好的相关性。在基因组的复杂度和大小之间缺乏很好的相关性的现象通常叫做C值悖论( C-value paradox)。

可能的解释有哪些呢？

- 
- 生物体的高等还是低等并不能光从染色体的多少或者DNA的多少来衡量，而应该看他们的有用基因的数量。因为染色体中很多内含子、junkDNA和重复序列在的研究看来对生物的性状是不必要的。
  - C值的大小并不能完全说明生物进化的程度和遗传复杂性的高低。也在计算C值时是根据所有的表达产物来估计的。比如有1000种蛋白质产物就意味着会有大约1000种mRNA，就是1000个基因，而实际上还有非编码的，所以少估计了。
  - 生物类群中C值变化范围宽就意味着在某些生物中有些DNA是冗余的，不能编码有功能的活性物质。DNA总量变化范围的产生至少有一个原因，即在染色体上存在着不同数目的重复序列，这些重复序列是不表达的。



- HGP的主要任务是人类的DNA测序，主要包括以下四张谱图

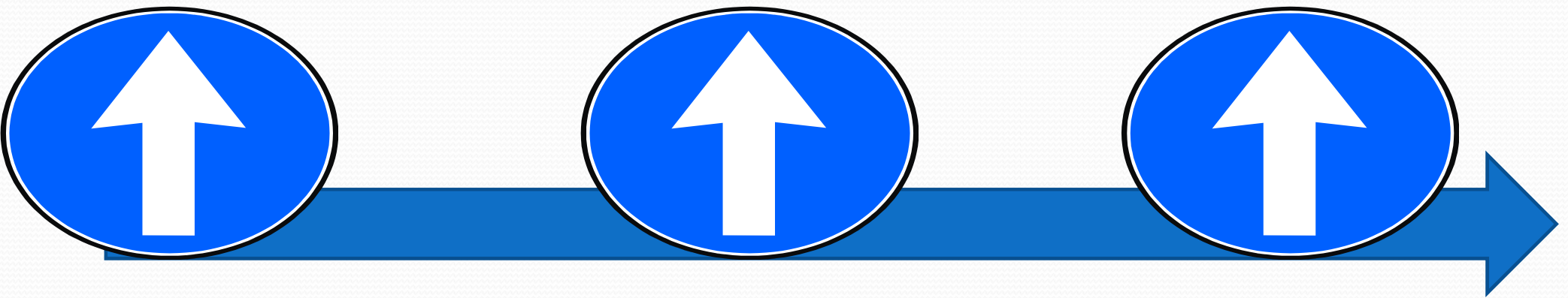


# 遗传图谱(genetic map)

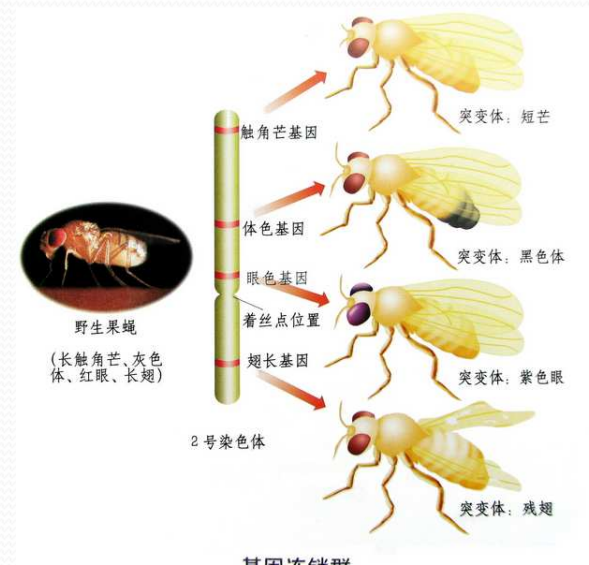
- 又称连锁图谱（linkage map），它是以具有遗传多态性（在一个遗传位点上具有一个以上的等位基因，在群体中的出现频率皆高于1%）的遗传标记为“路标”，以遗传学距离（在减数分裂事件中两个位点之间进行交换、重组的百分率，1%的重组率称为1 cM，厘摩）为图距的基因组图。
- 早期使用的多态性标志有RFLP(限制性酶切片段长度多态性)、RAPD(随机引物扩增多态性DNA)、AFLP(扩增片段长度多态性)；80年代后出现的有STR(短串联重复序列，又称微卫星)DNA遗传多态性分析和90年代发展的 SNP(单个核苷酸的多态性)分析。



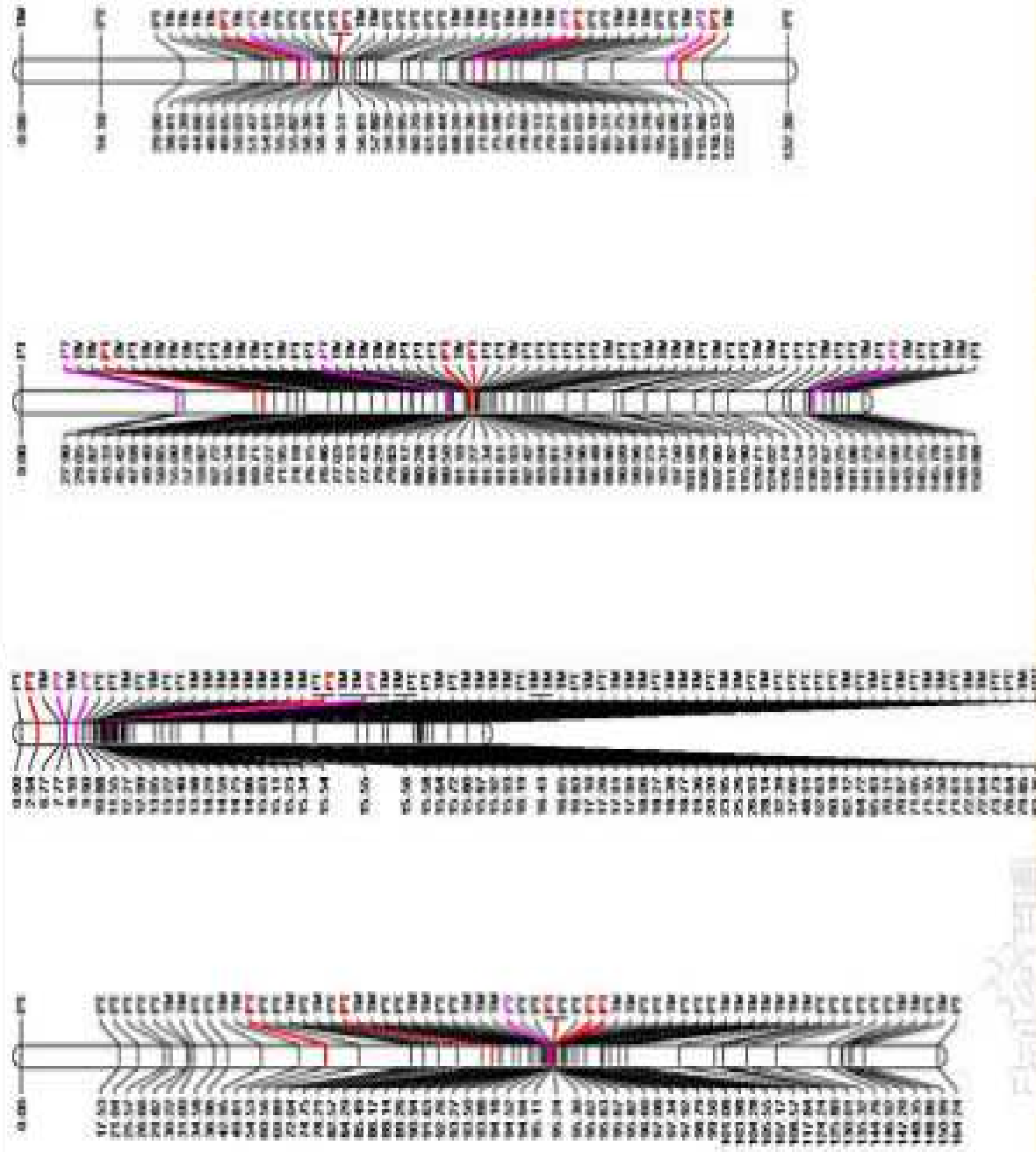
- 人基因组全长约3300cM，如两个标记之间相距1cM，则需3300个标记，如相距2~5cM，则需660~1650个标记。
- 显示所知的基因和/或遗传标记的相对位置，而不是在每条染色体上特殊的物理位置。



- 遗传图谱的建立为**基因识别**和**完成基因定位**创造了条件。6 000多个遗传标记已经能够把人的基因组分成6 000多个区域，使得连锁分析法可以找到某一致病的或表现的基因与某一标记邻近（紧密连锁）的证据，这样可把这一基因定位于这一已知区域，再对基因进行分离和研究。对于疾病诊治而言，找基因和分析基因是个关键。







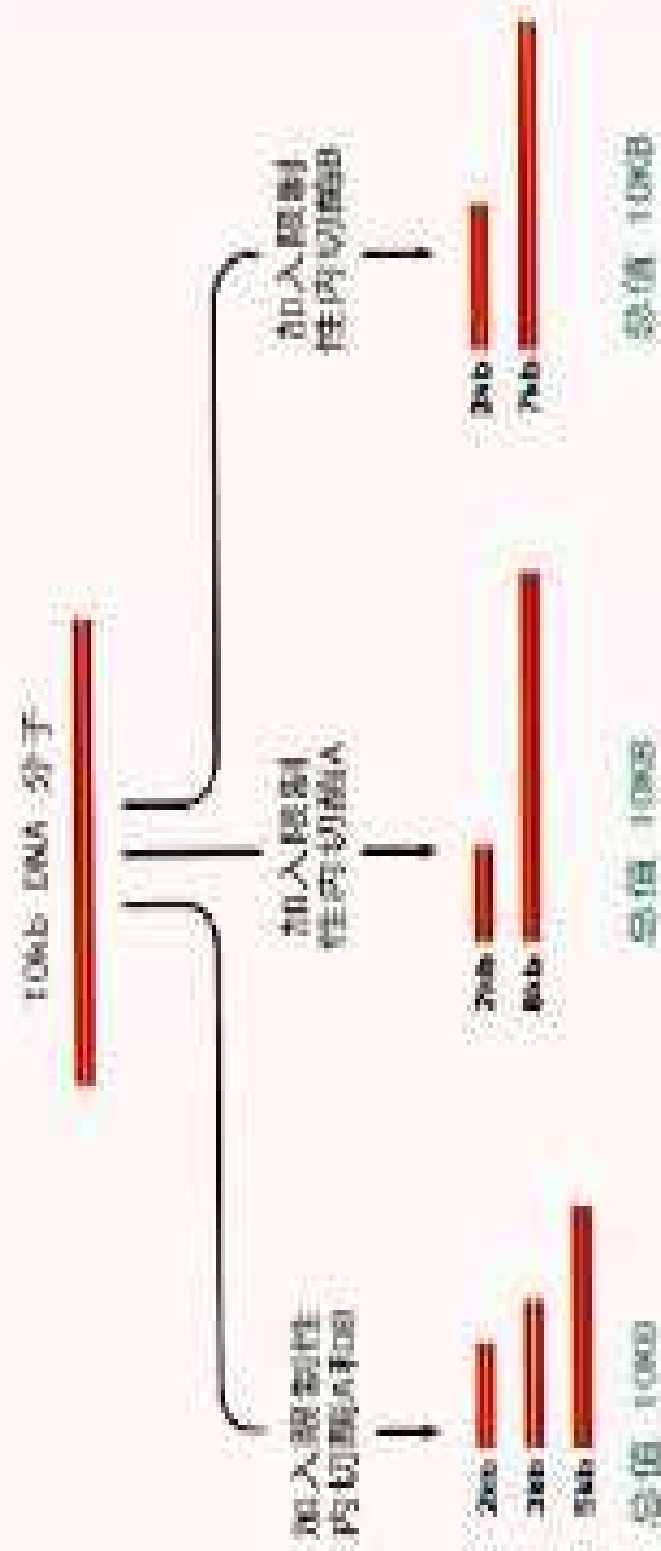
烟草分子标记遗传连锁图谱（部分）

# 物理图谱(physical map)

- 物理图谱是指有关构成基因组的全部基因的排列和间距的信息，它是通过对构成基因组DNA分子进行测定而绘制的。
- 绘制物理图谱的目的是把有关基因的遗传信息及其在每条染色体上的相对位置线性而系统地排列出来。
- 物理图谱是利用限制性内切酶将染色体切成片段，再根据重叠序列确定片段间连接顺序，以及遗传标志之间物理距离碱基对(bp) 或千碱基(kb) 或兆碱基(Mb)的图谱。

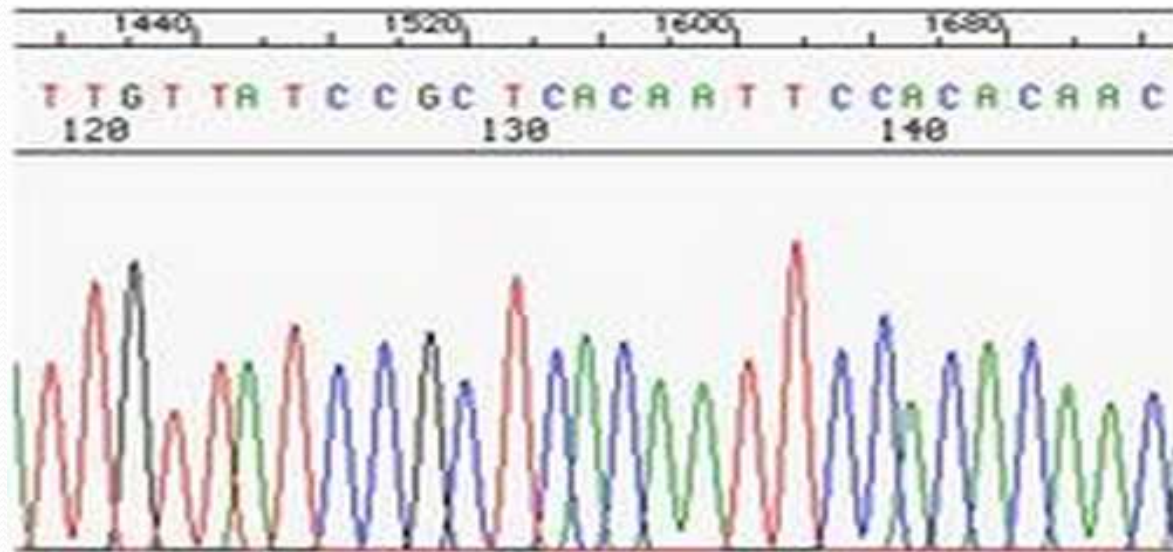
DNA物理图谱是指DNA链的限制性酶切片段的排列顺序，即酶切片段在DNA链上的定位。





# 序列图谱(sequence map)

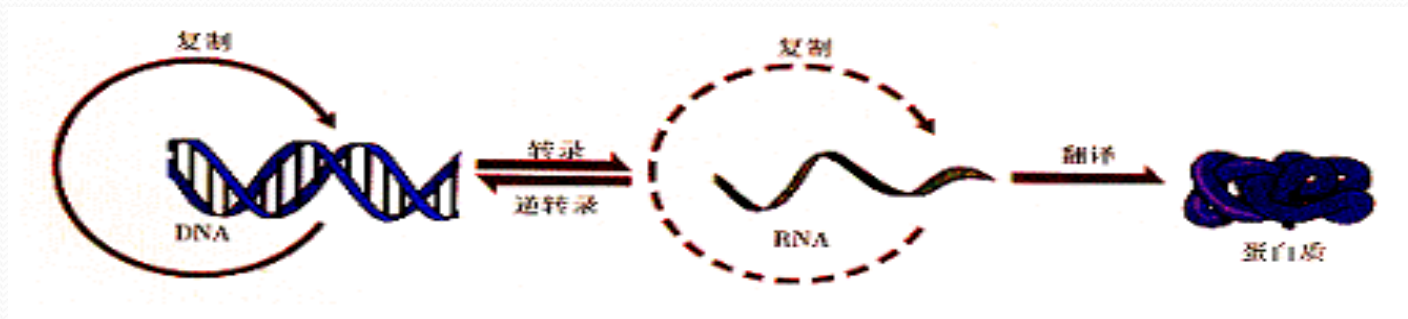
- 随着遗传图谱和物理图谱的完成，测序就成为重中之重。DNA序列分析技术包括制备DNA片段化及碱基分析、DNA信息翻译的多阶段的过程。通过测序得到基因组的序列图谱。





# 基因图谱(gene map)

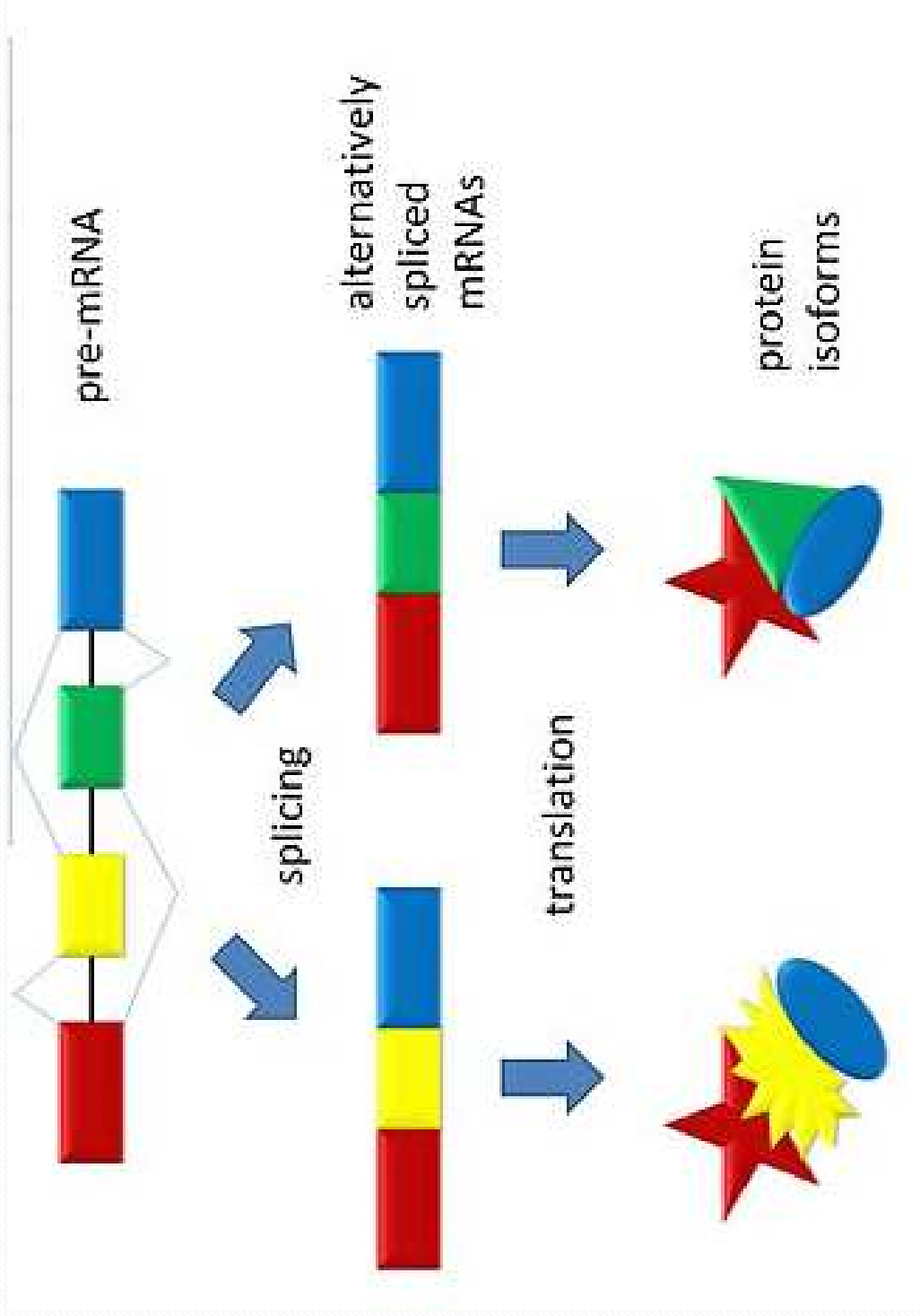
- 基因图谱是在识别基因组所包含的蛋白质编码序列的基础上绘制的结合有关基因序列、位置及表达模式等信息的图谱。
- 采用的最主要的方法是通过基因表达产物的mRNA反追到染色体上的位置。
- 真正读懂DNA，不仅要知道哪一段A、T、G、C的顺序，而且哪些排列组合表示一个基因（有些排列不表示任何基因）。




## 人类基因组给予我们的启发

- 一个最惊人的发现是人类基因组仅含有( 2-2.5)万个编码蛋白质的基因，而不是预测中的10万个基因。预测主要基于估计人类细胞制造接近(10~15)万种蛋白质。实际的基因数量比预料中的基因数量少得多的原因之一是，大量有关联功能的基因家族的发现。
- 另外，已经发现很多基因通过选择性剪接作用编码了多种多样的蛋白质。据估计，超过一半的人类基因可以通过选择性剪接产生大量的蛋白质。





- 
- ①人类基因组含有大约31亿对碱基对，其中28.5亿碱基对已经全部完成测序。
  - ②所有的个体大约有99.9%的基因组是相同的。
  - ③不到2%的基因组是编码基因。
  - ④我们的DNA大多数不编码蛋白质，重复DNA序列至少占非编码基因的50%。
  - ⑤基因组含有大约(2-2.5)万编码蛋白质的基因。
  - ⑥很多人类基因可以产生不止一种蛋白质，使得人类细胞能够仅由(2-2.5)万个基因制造大约（8~10）万个蛋白质。



- ⑦超过一半的人类基因的作用是不知道的。
- ⑧1号染色体含有最多数量的基因，Y染色体含有最少数量的基因。
- ⑨人类基因组和其他有机体内的基因显示了很高的序列相似性。
- ⑩成千上万的人类疾病基因已经被识别，并绘制出它们在染色体中的位置。



## 后基因组计划


- 在HGP完成之后，更主要、更艰巨的工作是解读破译这些基因结构和功能，以及基因之间的相互作用关系与调控作用等。
- 以基因组功能研究为主要内容的后基因组时代即功能基因组学时代已经到来。





# 人类基因组计划引发的“组学”革命

- ①蛋白质组学——研究细胞中所有的蛋白质。
- ②代谢组学——研究参与细胞新陈代谢的蛋白质和酶的代谢途径。
- ③糖组学——研究细胞内的碳水化合物。
- ④相互作用组学——研究细胞内蛋白质网络复杂的相互作用。
- ⑤转录组学——研究细胞内所有基因的表达（转录）。

- 
- 跨物种、跨群体的比较模式生物与人类基因组的DNA序列、物种起源、进化、基因功能演化、差异表达和定位、克隆人类疾病基因的**比较基因组学**
  - 以基因组学与临床医学，医药产业的结合为特征，以基因治疗为突破口，研究不同个体疾病、药物反应与DNA多态关系的**医学基因组学**(medicalgenomics)和**药物基因组学**(pharmacogenomics)等等。



# 蛋白质组学



# 比较基因组学



- HGP还涉及绘制和测序大量的模型生物有机体的基因组，包括大肠杆菌、拟南芥、酿酒酵母、黑腹果蝇、秀丽线虫、小家鼠和其他物种。
- 我们和其他物种共有基因的数量很多，范围从酵母里的30%，到老鼠基因的80%，再到黑猩猩基因的95%是相同的。最近，作为“人类最好的朋友”的狗的基因测序也被完成了，结果显示人类和狗的基因75%是相同的。
- 以这种研究将会引发对人类进化更深层次的了解。它可以使科学家们能够研究这些有机生物体内的基因结构和功能，旨在理解其他物种包括人类体的基因结构和功能。



## 应用

- 2007年，康涅狄格454生命科学公司和贝勒大学的研究人员们排列了詹姆斯·沃森的基因组，几乎花费了100万美元。454生命科学公司正在开发一个创新性的DNA测序仪，断定对DNA结构的共同发现者的基因进行测序的“Jim计划”是一个发展和促进测序技术高姿态的方式。詹姆斯·沃森在2005年就给454生命科学公司的科学家们提供了他的血样，2007年中旬，他们给了沃森博士两个含有他的基因组序列的光盘。沃森将他的基因组序列公开给了研究人员，除了载脂蛋白E基因（ApoE）。该基因突变能显示出老年痴呆症的倾向。



- 第一个黑色三叶杨树（杨树的一种）的基因组被测序了。迄今为止，杨树的45555个基因是基因组中发现的最多数量。科学家们期望利用这个计划的数据来帮助林业产业生产出更好的产品，包括生物燃料，甚至是能够从大气中捕捉更多二氧化碳的基因工程杨树。
- 最近蜜蜂的基因组也完成了测序，该项目的信息将除了被用于了解蜜蜂的遗传和行为以帮助蜂蜜生产行业，还将用于推进对蜜蜂毒素如何产生过敏反应的理解。





# 生物学数据库

- 生物信息数据库可以分为一级数据库和二级数据库
  - 一级数据库是进行简单的归类整理和注释；
  - 二级数据库是在一级数据库、实验数据和理论分析的基础上针对特定目标衍生而来，是对生物学知识和信息的进一步整理，



# 著名数据库

- 国际上著名的一级核酸数据库有GenBank数据库、EMBL核酸库和DDBJ库等；蛋白质序列数据库有SWISS-PROT、PIR等；蛋白质结构库有PDB等。
- 国际上二级生物学数据库非常多，它们因针对不同的研究内容和需要而各具特色。
  - 人类基因组图谱库GDB；
  - 转录因子和结合位点库TRANSFAC；
  - 蛋白质结构家族分类库SCOP等。







# GenBank

- GenBank数据库包含了所有已知的核酸序列和蛋白质序列，以及与它们相关的文献著作和生物学注释。

<http://www.ncbi.nlm.nih.gov/genbank>

 NCBI Resources 

GenBank

[GenBank](#) [Submit](#) [Genomes](#) [WGS](#) [HTGs](#) [EST/GSS](#) [Metagenomes](#) [TPA](#) [TSA](#) [INSDC](#)

## GenBank Overview

### What is GenBank?

GenBank<sup>®</sup> is the NIH genetic sequence database, an annotated collection of all publicly available DNA sequences ([Nucleic Acids Research, 2013 Jan;41\(D1\):D36-42](#)). GenBank is part of the [International Nucleotide Sequence Database Collaboration](#), which comprises the DNA DataBank of Japan (DDBJ), the European Molecular Biology Laboratory (EMBL), and GenBank at NCBI. These three organizations exchange data on a daily basis.

The complete [release notes](#) for the current version of GenBank are available on the NCBI ftp site. A new release is made every two months. GenBank growth [statistics](#) for both the traditional GenBank divisions and the WGS division are available from each release.

An [annotated sample GenBank record](#) for a *Saccharomyces cerevisiae* gene demonstrates many of the features of the GenBank flat file format.

### Access to GenBank

There are several ways to search and retrieve data from GenBank.

- Search GenBank for sequence identifiers and annotations with [Entrez Nucleotide](#), which is divided into three divisions: [CoreNucleotide](#) (the main collection), [dbEST](#) (Expressed Sequence Tags), and [dbGSS](#) (Genome Survey Sequences).
- Search and align GenBank sequences to a query sequence using [BLAST](#) (Basic Local Alignment Search Tool). BLAST searches CoreNucleotide, dbEST, and dbGSS independently; see [BLAST info](#) for more information about the numerous BLAST databases.
- Search, link, and download sequences programmatically using [NCBI e-utils](#).

### GenBank Data Usage

## GenBank Resources

- [GenBank Home](#)
- [Submission Types](#)
- [Submission Tools](#)
- [Search GenBank](#)
- [Update GenBank Records](#)

- GenBank数据记录包含了对序列的简要描述、科学命名、物种分类名称、参考文献、序列特征表以及序列本身。序列特征表里包含对序列的生物学特征注释，如：编码区、元、重复区域、突变位点或修饰位点等。

NCBI Resources How To Sign in to NCBI

Nucleotide Nucleotide Search Advanced Help

Display Settings: GenBank Send:

### Drosophila melanogaster chromosome X

GenBank: AE014298.5

[FASTA](#) [Graphics](#)

LOCUS AE014298 23542271 bp DNA linear INV 15-AUG-2014

DEFINITION Drosophila melanogaster chromosome X.

ACCESSION AE014298 AE002566 AE002593 AE002611 AE002620 AE002629 AE002634  
AE002636 AE002694 AE002799 AE002813 AE002822 AE002925 AE003122  
AE003224 AE003327 AE003417-AE003451 AE003484-AE003513  
AE003568-AE003574 CM000460 DS483593 DS483747

VERSION AE014298.5 GI:667695275

DBLINK BioProject: [PRJNA13812](#)  
BioSample: [SAMN02803731](#)

KEYWORDS .

SOURCE Drosophila melanogaster (fruit fly)

ORGANISM [Drosophila melanogaster](#)  
Eukaryota; Metazoa; Ecdysozoa; Arthropoda; Hexapoda; Insecta;  
Pterygota; Neoptera; Endopterygota; Diptera; Brachycera;  
Muscomorpha; Ephydroidea; Drosophilidae; Drosophila; Sophophora.

Change region shown

Customize view

Basic Features

☐ Default features

☒ Gene, RNA, and CDS features only

Display options

☐ Show sequence

☐ Show reverse complement

Update View

Analyze this sequence

Run BLAST

Pick Primers

Highlight Sequence Features



- EMBL数据库由欧洲生物信息学研究所(EBI)维护。由于与GenBank和DDBJ的数据合作交换，它也是全面的核酸序列库。数据管理维护、查询、检索可以通过互联网上的序列提取系统(SRS)完成。

## The European Bioinformatics Institute

Part of the European Molecular Biology Laboratory

Thursday, 16th October 2014

**Please Note:** Between **20th October 2014** and **23rd December 2014** EMBL-EBI will be consolidating its web infrastructure into a single data centre. Services are expected to continue operating as normal during this period, but some disruption may be experienced.

EMBL-EBI provides freely available data from life science experiments, performs basic research in computational biology and offers an extensive user training programme, supporting researchers in academia and industry.

Find a gene, protein or chemical:

Search

Examples: blast keratin bfl1...


### Popular

 [Services](#)


 [Research](#)


 [Training](#)

 [News](#)

 [Jobs](#)

 [Visit us](#)

 [EMBL](#)

 [Contacts](#)

European Molecular Biology Laboratory

Visit **EMBL.org**

- 日本DNA数据DDBJ也是一个全面的核酸序列数据库，与GenBank和EMBL核酸库合作交换数据。




#### DDBJ をご利用の皆様へ（2014.11.28 – 2014.12.4）

国立遺伝学研究所の停電に伴い、以下の日程で NIG Supercomputer、DDBJ の全ネットワークサービスを停止いたします。

サービス名	停止期間
•D-way •DRA Search •BioProject Search •BioSample Search •JGA Submission/Download Tool •JGA Meta Viewer	11月28日(金) 13:00 ~ 12月4日(木) 10:00
•NIG Supercomputer	11月28日(金) 17:00 ~ 12月3日(水) 24:00
•DDBJ Homepage •DDBJ塩基配列登録システム •getentry •ARSA •BLAST •ClustalW •DDBJ Read Annotation Pipeline	11月28日(金) 17:00 ~ 12月4日(木) 10:00



- 
- **SWISS-PROT**蛋白质序列数据库，由欧洲生物信息学研究所(EBI)维护。数据库包含两部分，**SWISS-PROT**和**TrEMBL**。**SWISS-PROT**是经过注释的蛋白质序列数据库，数据库由蛋白质序列条目构成，每个条目包含蛋白质序列、引用文献信息、分类学信息、注释等，注释中包括蛋白质的功能、转录后修饰、特殊位点和区域、二级结构、结构、与其他序列的相似性、序列残缺与疾病的关系、序列变异体和冲突等信息。

- 蛋白质数据库(PDB)是国际上唯一的生物大分子结构 数据库。PDB收集的数据包括X光晶体衍射和核磁共振（NMR）的蛋白质结构数据。2009年3月为止，共收集了56751个蛋白质结构，其中48618个来自X光晶体射方法，7776个来自NMR方法，230个来自电镜，1271个来自其他方法。

**RCSB PDB**  
PROTEIN DATA BANK

An Information Portal to  
105212 Biological  
Macromolecular Structures

Search by PDB ID, author, macromolecule, sequence, or ligands

Go

[Advanced Search](#) | [Browse by Annotations](#)



Welcome

Deposit

Search

Visualize

Analyze

## A Structural View of Biology

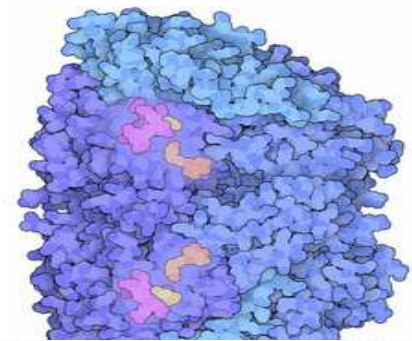
This resource is powered by the Protein Data Bank archive - information about the 3D shapes of proteins, nucleic acids, and complex assemblies that helps students and researchers understand all aspects of biomedicine and agriculture, from protein synthesis to health and disease.

As a member of the wwPDB, the RCSB PDB curates and annotates PDB data.

The RCSB PDB builds upon the data by creating tools and resources for research and education in molecular biology, structural biology, computational biology, and beyond.

Use this website to access curated and integrated biological macromolecular information in the context of function, biological processes, evolution, pathways, and disease states.

## November Molecule of the Month





# 生物信息学研究内容

- 序列比对
- 系统进化分析
- 计算机辅助基因识别
- 非编码区分析和DNA语言研究
- 比较基因组学
- 序列重叠群装配
- 蛋白质结构预测与分子设计
- 基于结构的药物设计
- 药物基因组学与药物蛋白质组学
- 其他

# 序列比对

- 意义在于从核酸、氨基酸的层次分析序列的相似性，推测其结构功能及进化上的联系，是基因识别、分子进化、生命起源研究的基础。

```

A A A A A G C G C G T C G T A G T C C G G A T C G G A G T C T G C A A C T C G A C T C G T G A A A T C G G A A T C G C T A G T A A T C G C A A A T C A G A A T G T T
A A A A A G C G C G T C G T A G T C C G G A T C G G A G T C T G C A A C T C G A C T C G T G A A A T C G G A A T C G C T A G T A A T C G C A A A T C A G A A T G T T
A T A A A A C C G T T C T C A G T T C G G A T T G T A G G C T G C A A C T C G C C T A C A T G A A G C T G G A A T C G C T A G T A A T C G C G G A T C A G C A T G C C
A C A A A T C T G T T C T C A G T T C G G A T C G C A G T C T G C A A C T C G A C T G C G T G A A G C T G G A A T C G C T A G T A A T C G C G G A T C A G C A T G C C
A T A A A G T T G T T C T C A G T T C G G A T T G T A G T C T G C A A C T C G A C T A C A T G A A G C T G G A A T C G C T A G T A A T C G T A G A T C A G C A T G C T
T T A A A G C C A G T C T C A G T T C G G A T T G T A G G C T G C A A C T C G C C T A C A T G A A G T C G G A A T C G C T A G T A A T C G C G G A T C A G C A C G C C
T T A A A G C C A G T C T C A G T T C G G A T T G T A G G C T G C A A C T C G C C T A C A T G A A G T C G G A A T C G C T A G T A A T C G C G G A T C A G C A C G C C
A T A A A A C C G A T C G T A G T C C G G A T C G C A G T C T G C A A C T C G A C T G C G T G A A G T C G G A A T C G C T A G T A A T C G T G A A T C A G A A T G T C
A T A A A A C C G A T C G T A G T C C G G A T C G C A G T C T G C A A C T C G A C T G C G T G A A G T C G G A A T C G C T A G T A A T C G T G A A T C A G A A T G T C
A T A A A A C C G A T C G T A G T C C G G A T C G C A G T C T G C A A C T C G A C T G C G T G A A G T C G G A A T C G C T A G T A A T C G T G A A T C A G A A T G T C
A G A A A G T G C A T C T A A G T C C G G A T T G G A G T C T G C A A C T C G A C T C C A T G A A G T C G G A A T C G C T A G T A A T C G C A A A T C A G A A T G T T
A G A A A G T G C A T C T A A G T C C G G A T T G G A G T C T G C A A C T C G A C T C C A T G A A G T C G G A A T C G C T A G T A A T C G C A A A T C A G A A T G T T
A T A A A G T A C G T C G T A G T C C G G A T T G G A G T C T G C A A C T C G A C T C C A T G A A G T C G G A A T C G C T A G T A A T C G T G G A T C A G A A T G C C
A T A A A G T A C G T C G T A G T C C G G A T T G G A G T C T G C A A C T C G A C T C C A T G A A G T C G G A A T C G C T A G T A A T C G T G G A T C A G A A T G C C
A T A A A G T G C G T C G T A G T C C G G A T T G G A G T C T G C A A C T C G A C T C C A T G A A G T C G G A A T C G C T A G T A A T C G T G G A T C A G A A T G C C
A T A A A G T G C G T C G T A G T C C G G A T T G G A G T C T G C A A C T C G A C T C C A T G A A G T C G G A A T C G C T A G T A A T C G T G G A T C A G A A T G C C
A T A A A G T G C G T C G T A G T C C G G A T T G G A G T C T G C A A C T C G A C T C C A T G A A G T C G G A A T C G C T A G T A A T C G T A G A T C A G A A T G C T
A T A A A G T G C G T C G T A G T C C G G A T T G G A G T C T G C A A C T C G A C T C C A T G A A G T C G G A A T C G C T A G T A A T C G T G G A T C A G A A T G C C

```



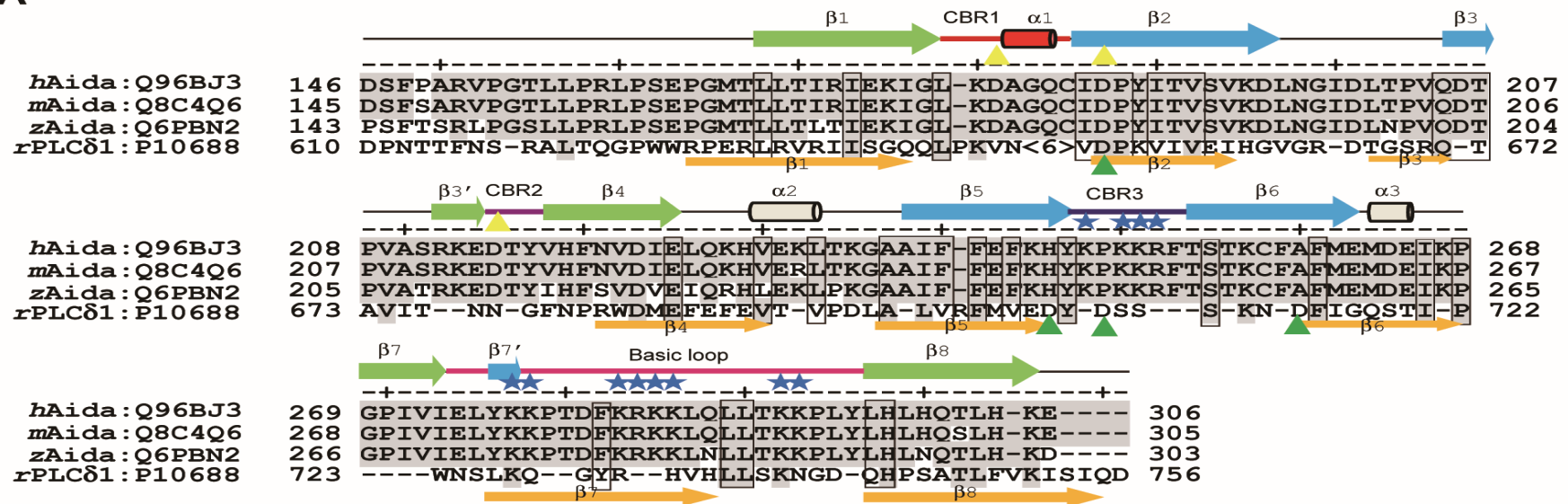
## 序列比对的用途和内容

- 序列发生分析：
- 通过序列比对，可以寻找序列间的同源性（相似性），这种同源相似性是序列间进化关系的一种反映，所构建的数据矩阵成为系统发生分析的基础。

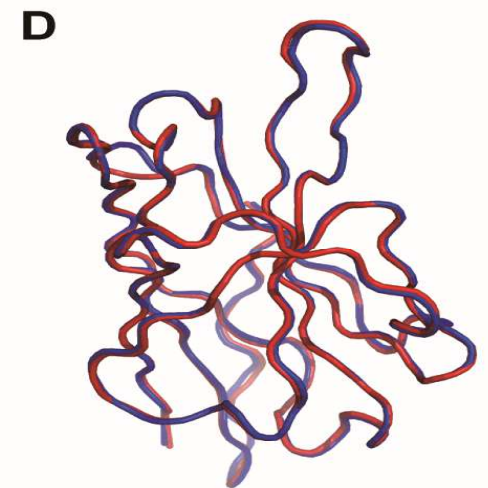
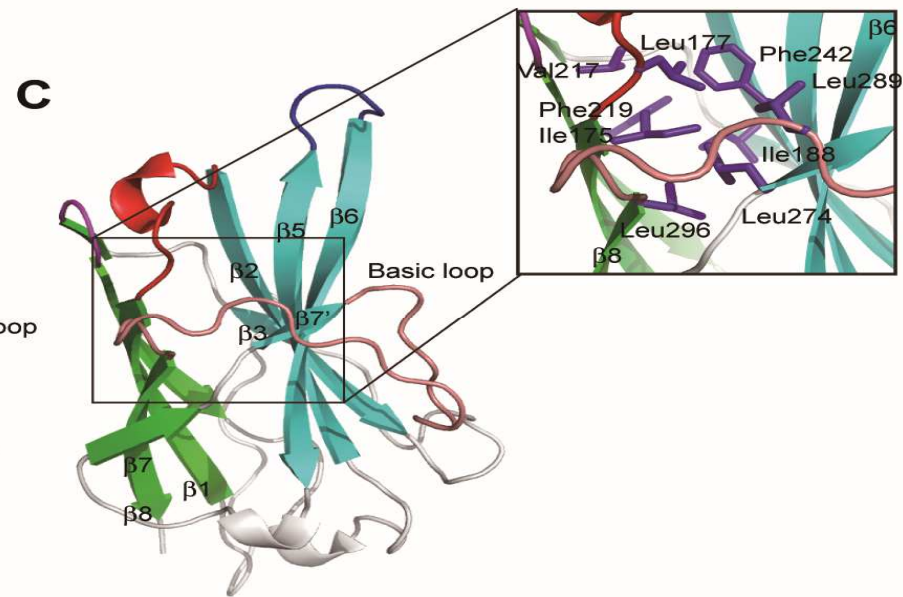
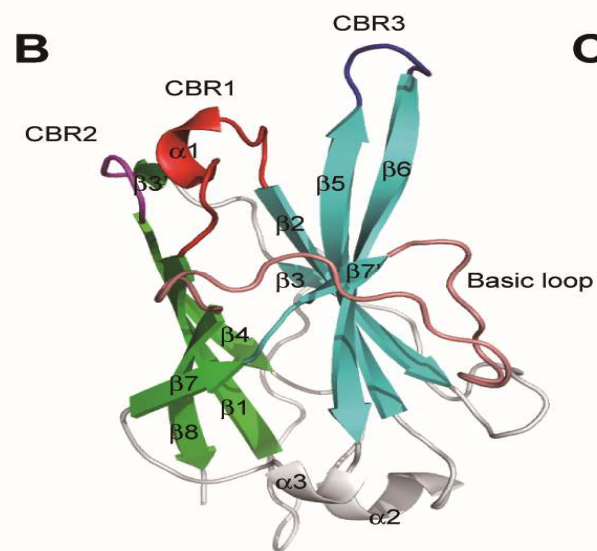


- 结构预测
- 将新获得的蛋白质序列与已知结构的序列进行比对，可以通过序列同源性来粗略地结构预测。

**A**







2014 *FEBS J* Lisha Zheng et al.

- 序列模体鉴定(sequence motif identification):
- 局部排列鉴定蛋白质和核苷酸序列中潜在的序列和功能模体。

Kinase

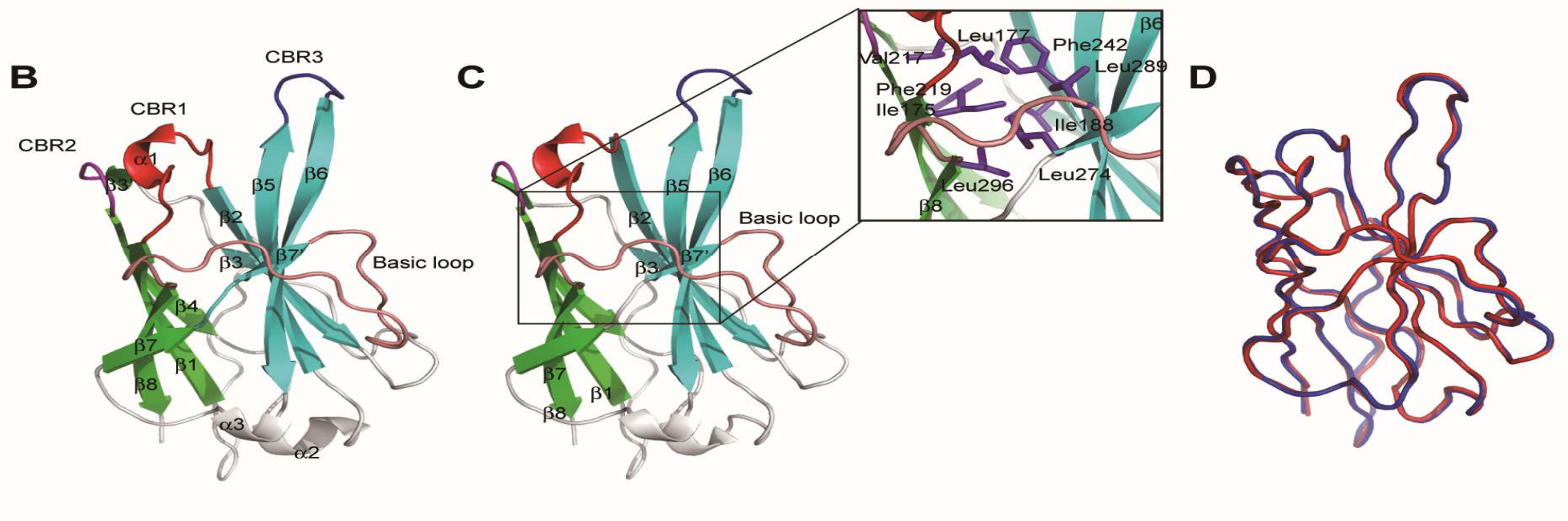
Kinase


Kinase

Kinase



- 功能预测：
- 蛋白质序列间高度的相似性通常意味着功能的相似性



- 
- 数据库搜索：
  - 比对是数据库搜索算法的基础，将查询序列与整个数据库的所有序列进行比对，从数据库中获得与其最相似序列的已有数据，能最快速获得查询序列的大量有价值的参考信息，结构和功能都会有很大的帮助。



- 序列数据库搜索中最著名且常用的工具之一是BLAST（ Basic Local Alignment Search Tool ）。Blast是一种基本的局部对位排列搜索工具，是现在应用最广泛的序列相似性搜索工具。
- 首先将查询序列打碎成一个个单词，或者说是定长的子序列（单词默认长度为4）。查询序列中所有可能的单词是通过在查询序列上滚动与单词等长的窗口来获得。

BLAST®

BLAST finds regions of similarity between biological sequences. [more...](#)

**New** Try [SmartBLAST](#) for an improved protein-protein search

### BLAST Assembled Genomes

Find Genomic BLAST pages:

Enter organism name or id-completions will be suggested

<input type="checkbox"/> Human	<input type="checkbox"/> Rabbit	<input type="checkbox"/> Zebrafish
<input type="checkbox"/> Mouse	<input type="checkbox"/> Chimp	<input type="checkbox"/> Clawed frog
<input type="checkbox"/> Rat	<input type="checkbox"/> Guinea pig	<input type="checkbox"/> Arabidopsis
<input type="checkbox"/> Cow	<input type="checkbox"/> Fruit fly	<input type="checkbox"/> Rice
<input type="checkbox"/> Pig	<input type="checkbox"/> Honey bee	<input type="checkbox"/> Yeast
<input type="checkbox"/> Dog	<input type="checkbox"/> Chicken	<input type="checkbox"/> Microbes

### Basic BLAST

Choose a BLAST program to run.

<a href="#">nucleotide blast</a>	Search a <b>nucleotide</b> database using a <b>nucleotide</b> query <i>Algorithms: blastn, megablast, discontinuous megablast</i>
<a href="#">protein blast</a>	Search <b>protein</b> database using a <b>protein</b> query <i>Algorithms: blastp, psi-blast, phi-blast, delta-blast</i>
<a href="#">blastx</a>	Search <b>protein</b> database using a <b>translated nucleotide</b> query
<a href="#">tblastn</a>	Search <b>translated nucleotide</b> database using a <b>protein</b> query
<a href="#">tblastx</a>	Search <b>translated nucleotide</b> database using a <b>translated nucleotide</b> query

- 例如，一个查询序列AILVPTV有4个不同的单词：AILV、ILVP、LVPT、VPTV。最常见氨基酸组成的单词会被省略，然后从数据库中搜索余下的单词出现的情况。每当从数据库找出一个单词的匹配，那么就从单词两端延伸该匹配，直到比对打分低于给定的阈值。





NCBI Home
Resource List (A-Z)
All Resources
Chemicals & Bioassays
Data & Software
DNA & RNA
Domains & Structures
Genes & Expression
Genetics & Medicine
Genomes & Maps

## Welcome to NCBI

The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

[About the NCBI](#) | [Mission](#) | [Organization](#) | [Research](#) | [NCBI News](#)

## Get Started

- [Tools](#): Analyze data using NCBI software
- [Downloads](#): Get NCBI data or software
- [How To's](#): Learn how to accomplish specific tasks at NCBI
- [Submissions](#): Submit data to GenBank or other NCBI databases

## Popular Resources

[PubMed](#)  
[Bookshelf](#)  
[PubMed Central](#)  
[PubMed Health](#)  
[BLAST](#)  
[Nucleotide](#)  
[Genome](#)  
[SNP](#)  
[Gene](#)  
[Protein](#)

<http://www.ncbi.nlm.nih.gov/>

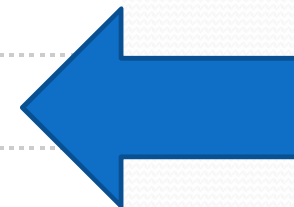
PubMed Central

PubMed Health

BLAST

Nucleotide

Genome



BLAST finds regions of similarity between biological sequences. [more...](#)

**New**

**DELTA-BLAST**, a more sensitive protein-protein

## BLAST Assembled Genomes

Find Genomic BLAST pages:

**GO**

- ☐ [Human](#)
- ☐ [Mouse](#)
- ☐ [Rat](#)
- ☐ [Cow](#)
- ☐ [Pig](#)
- ☐ [Dog](#)
- ☐ [Rabbit](#)
- ☐ [Chimpanzee](#)
- ☐ [Guinea Pig](#)
- ☐ [Fruit Fly](#)
- ☐ [Honey Bee](#)
- ☐ [Chicken](#)

## Basic BLAST

Choose a BLAST program to run.

<a href="#">nucleotide blast</a>	Search a <b>nucleotide</b> database using a <b>nucleotide</b> query <i>Algorithms:</i> <b>blastn</b> , <b>megablast</b> , <b>discontiguous megablast</b>
<a href="#">protein blast</a>	Search <b>protein</b> database using a <b>protein</b> query <i>Algorithms:</i> <b>blastp</b> , <b>psi-blast</b> , <b>phi-blast</b> , <b>delta-blast</b>
<a href="#">blastx</a>	Search <b>protein</b> database using a <b>translated nucleotide</b> query
<a href="#">tblastn</a>	Search <b>translated nucleotide</b> database using a <b>protein</b> query
<a href="#">tblastx</a>	Search <b>translated nucleotide</b> database using a <b>translated nucleotide</b> query



**blastn**

[blastp](#)

[blastx](#)

[tblastn](#)

[tblastx](#)

BLASTN programs search

### Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s)

[Clear](#)

GGCCACACACGGAGGCAGGGA

TTATACAGGGCGTACACTTTC

Fi

Or, upload file

选择文件

未选择文件

Job Title

Enter a descriptive title for your BLAST search

☐ Align two or more sequences

### Choose Search Set

Database

☒ Human genomic + transcript ☐ Mouse genomic + transcript ☐ Other

◆ Human genomic plus transcript (Human G+T)

Exclude  
Optional

☐ Models (XM/XP) ☐ Uncultured/environmental sample sequences

Limit to  
Optional

☐ Sequences from type material

Entrez Query

Optional

Enter an Entrez query to limit search

[YouTube](#)

[Create](#)



BLAST®

Basic Local Alignment Search Tool

- Home
- Recent Results
- Saved Strategies
- Help

► NCBI/BLAST/blastn suite/ Formatting Results - 7S7H9ZE1014

**?** Your search parameters were adjusted to search for a short input sequence.

[Edit and Resubmit](#) [Save Search Strategies](#) [Formatting options](#) [Download](#)

## Nucleotide Sequence (42 letters)

**RID** 7S7H9ZE1014 (Expires on 12-02 14:52 pm)  
**Query ID** lcl|7803  
**Description** None  
**Molecule type** nucleic acid  
**Query Length** 42

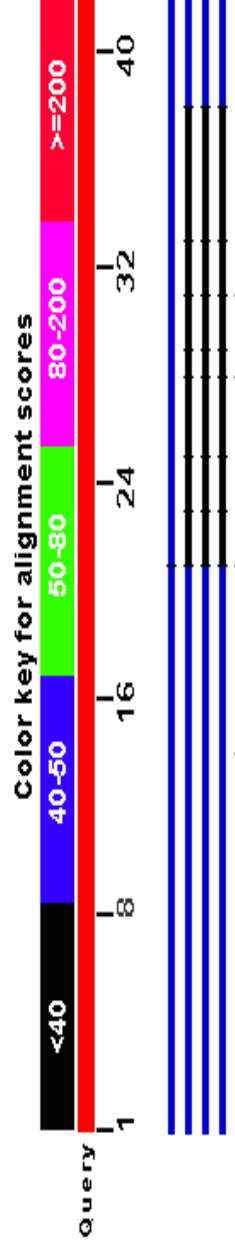
**Database Name** Human G+T (2 databases)  
**Description** [► See details](#)  
**Program** BLASTN 2.2.30+ [► Citation](#)

Other reports: [► Search Summary](#) [\[Taxonomy reports\]](#) [\[Distance tree of results\]](#) [\[Genome view\]](#)

## Graphic Summary

### Distribution of 198 Blast Hits on the Query Sequence

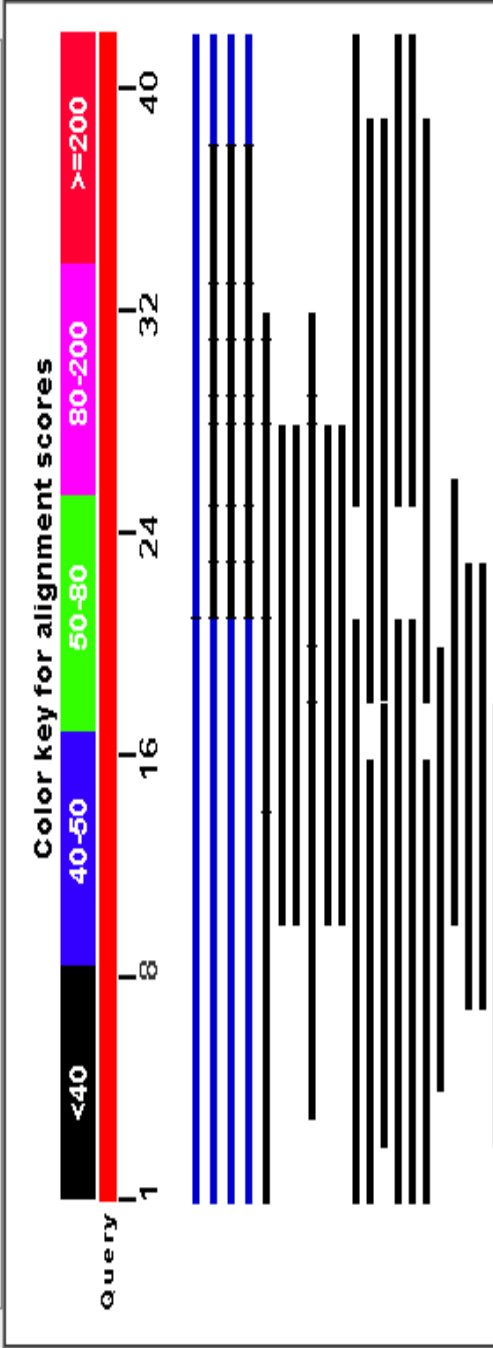
Mouse over to see the define, click to show alignments





Distribution of 198 Blast Hits on the Query Sequence

Mouse-over to show define and scores, click to show alignments



Sequences producing significant alignments:

Select: All None Selected 0

Download GenBank Graphics Sort by: E value

Homo sapiens CD80 molecule (CD80), mRNA  
Sequence ID: [reflNM\\_005191.3](#) Length: 2757 Number of Matches: 2

Range 1: 1242 to 1262 GenBank Graphics


Score	Expect	Identities	Gaps	Strand
42.1 bits(21)	0.045	21/21(100%)	0/21(0%)	Plus/Minus

Query 22 TTATACAGGGCGTACACTTTC 42  
|||||  
Sbjct 1262 TTATACAGGGCGTACACTTTC 1242

Range 2: 399 to 419 GenBank Graphics

Score	Expect	Identities	Gaps	Strand
42.1 bits(21)	0.045	21/21(100%)	0/21(0%)	Plus/Plus

Query 1 GGCCACACACGGAGGCAGGGA 21  
|||||  
Sbjct 399 GGCCACACACGGAGGCAGGGA 419

- 
- 目前使用最广泛的多序列比对程序是CLUSTAL W（它的PC版本是CLUSTAL X）。
  - CLUSTAL W是一种渐进的比对方法，先将多个序列两两比对构建距离矩阵，反映序列之间两两关系；然后根据距离矩阵计算产生系统进化指导树，对关系密切的序列进行加权；然后从最紧密的两条序列开始，逐步引入临近的序列并不断重新构建比对，直到所有序列都被加入为止。



- Blast只进行局部比对
- CLUSTAL W进行全局比对

[Download](#) [GenBank](#) [Graphics](#) Sort by: [E value](#)

Homo sapiens CD80 molecule (CD80), mRNA  
Sequence ID: [ref|NM\\_005191.3|](#) Length: 2757 Number

Range 1: 1242 to 1262 [GenBank](#) [Graphics](#)

Score	Expect	Identities
42.1 bits(21)	0.045	21/21(100%)

Query	22	TTATACAGGGCGTACACTTTC	42
Sbjct	1262	TTATACAGGGCGTACACTTTC	1242

Range 2: 399 to 419 [GenBank](#) [Graphics](#)

Score	Expect	Identities
42.1 bits(21)	0.045	21/21(100%)

Query	1	GGCCACACACGGAGGCAGGGA	21
Sbjct	399	GGCCACACACGGAGGCAGGGA	419

CLUSTAL W (1.60) multiple sequence alignment

```

hun-U1A  -----MAVPETRPNHTIYINNLNEKIKKDELKKSLEYAIFSQFGQILDILVSRSLKMRGQ
nse-U1A  MATIATMPVPETRANHTIYINNLNEKIKKDELKKSLEYAIFSQFGQILDILVSRIDKMRGQ
xla-U1A  -----MSIQEVRPMNTIYINNLNEKIKKDELKKSLEYAIFSQFGQILDILVSRNLKMRGQ
dae-U1A  -----MEMLPNQTIYINNLNEKIKKDELKKSLEYAIFSQFGQILDILVSRNLKMRGQ
* * *****,*****. . .*****

hun-U1A  AFVIFKEVSSATNALRSMQGFPFYDKPMRIQYAKTSDIIAKDKGTTFVERDRKR-EKRRK
nse-U1A  AFVIFKEVTSATNALRSMQGFPFYDKPMRIQYAKTSDIIAKDKGTTFVERDRKR-EKRRK
xla-U1A  AFVIFKETSSATNALRSMQGFPFYDKPMRIQYAKTSDIIAKDKGTTFVERDRKRQEKRRV
dae-U1A  AFVIFKEIGSASNALRTMQGFPFYDKPMQIAYSKSDSDIVAKIKGTFFKERPKKVKPPKPA
***** ** ,***, *****,* *,*,****,**,***, ** ,* .

```

# 系统进化分析

- 系统进化分析就是要推断或者评估这些进化关系。所推断出来的进化关系一般用分支图表（进化树）来描述，这个进化树就描述了同一谱系的进化关系，包括了分子进化（基因树）、物种进化以及分子进化和物种进化的综合。

Tree of life





- 用以描述类群系统发育关系的一种树状图。它由系统发生树 (phylogenetic tree)，也称为系统树，是表达种群（或序列）间系统发育关系的一种树状图。它由一系列点和分支组成，其中一个节点代表一个分类单元。分支末端的节点对应一个基因或者生物体。与外部节点对应，内部节点代表一个推断出的共同祖先，它在某个时候分歧出两个独立的分支。虽然内部节点可以多，但大多数树的内部节点都只有两个分支，因此称为二节点。标度树是指分支长度与相邻节点的差异程度。正是将所有的外部节点排成行，表示它们之间的亲缘关系，没有表示它们之间差异程度的信息。

末端节点

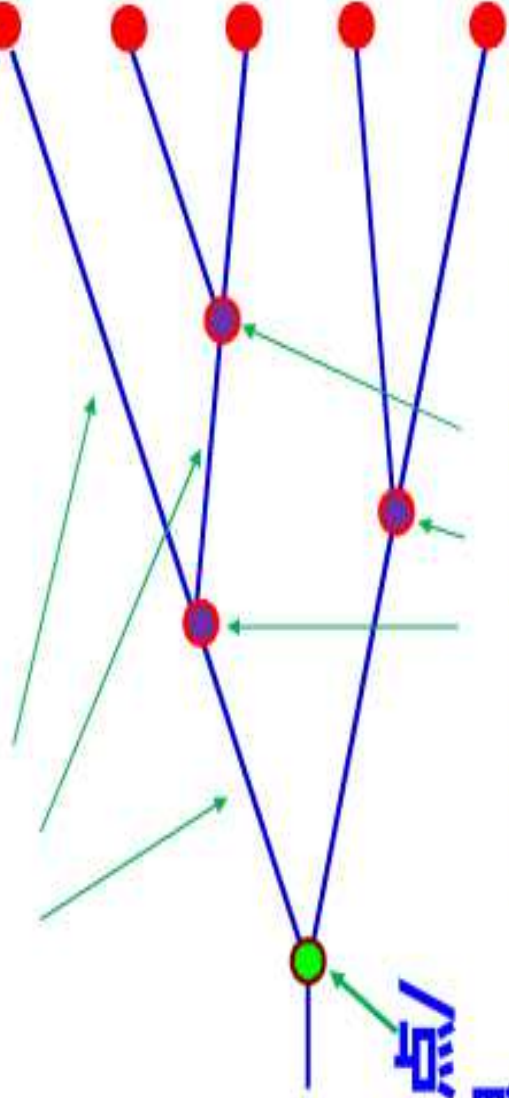
分支/世系

代表最终分  
类，可以是  
物种，群体  
，或者蛋白  
质、DNA、  
RNA分子等

A B C D E

祖先节点/  
树根

内部节点/分歧点，该  
分支可能的祖先结点

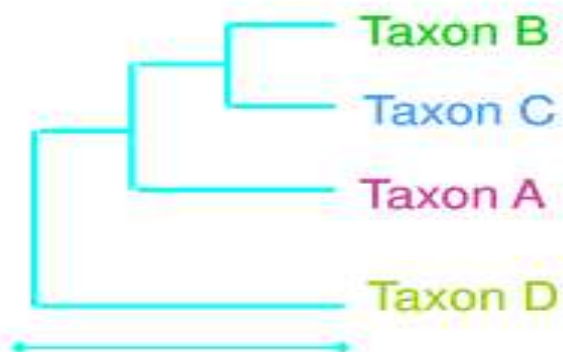




# 系统发育树：三种类型

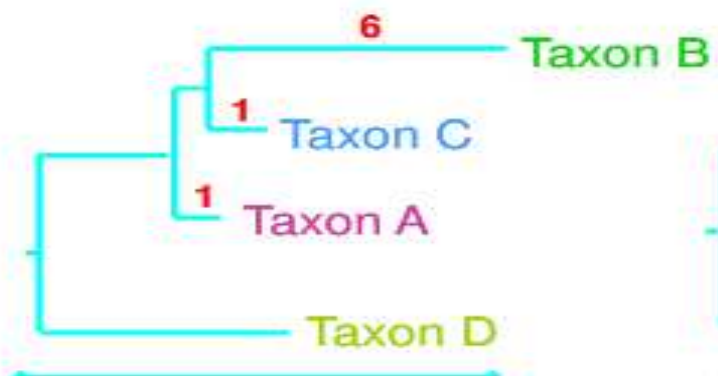


## 分支图



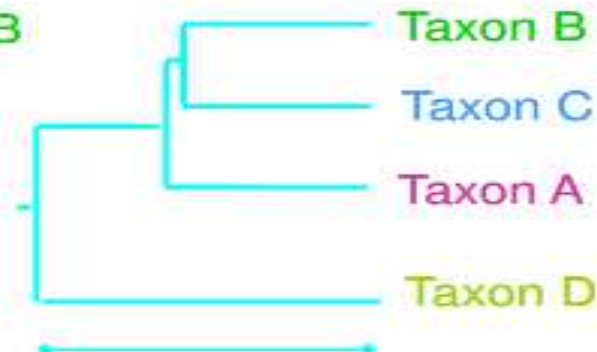
长度： 无意义

## 进化树




遗传变化

## 时间度量树

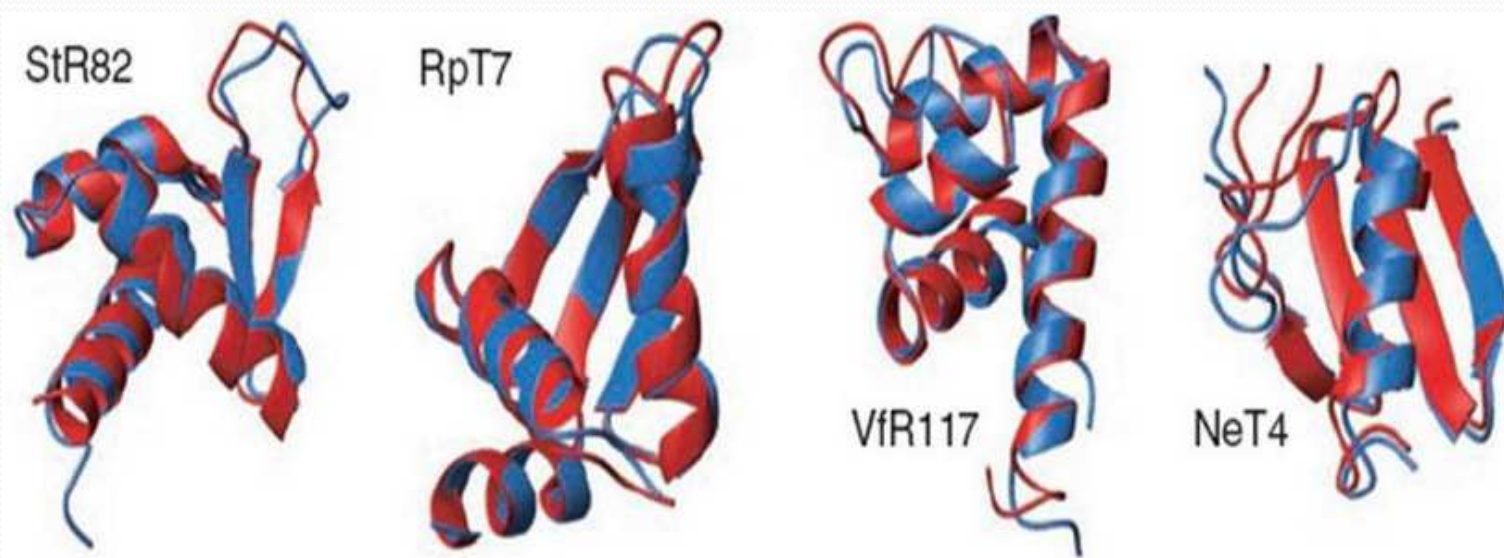


时间

- 
- 基于单个同源基因差异构建的系统发生树称为基因树(gene tree)，这种树与物种树(species tree)不同。
  - 物种树一般是从多个基因数据的分析中得到。它代表的仅仅是单个基因的进化历史，而不是它所在物种的进化历史。物种树一般是从多个基因数据的分析中得到。

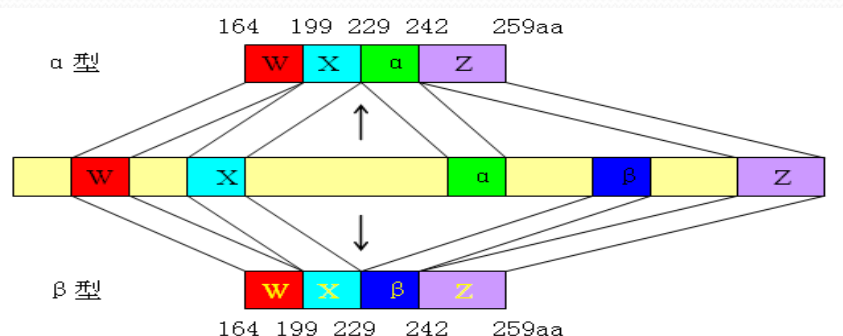


- 根据二级或者三级序列结构进行比对，比起直接利用一级序列进行比对的可信度要好，因为在同源性评估中，人们一直认为复杂结构的保守性高于简单特征（核苷酸、氨基酸）同源保守性，而且程序还可以搜索到一些特殊的关联点。



# 计算机辅助基因识别

- 基本问题是给定基因组序列后，正确识别基因的范围和在基因组序列中的精确位置.这是最重要的课题之一，而且越来越重要。经过20余年的努力，提出了数十种算法，有十种左右重要的算法和相应软件上网提供免费服务。原核生物计算机辅助基因识别相对容易些，结果好一些。从具有较多内含子的真核生物基因组序列中正确识别出起始密码子、剪切位点和终止密码子，是个相当困难的问题，研究现状不能令人满意，仍有大量的工作要做。





# 非编码区分析和DNA语言研究


- 在人类基因组中，编码部分进展总序列的3~5%，其它通常称为“垃圾”DNA，其实一点也不是垃圾，只是我们暂时还不知道其重要的功能。分析非编码区DNA序列需要大胆的想象和崭新的研究思路和方法。DNA序列作为一种遗传语言，不仅体现在编码序列之中，而且隐含在非编码序列之中。

你们觉得“垃圾”DNA可能有什么用处呢？

# 分子进化和比较基因组学

- 是最重要的课题之一。
- 早期的工作主要是利用不同物种中同一种基因序列的异同来研究生物的进化，构建进化树。既可以用DNA序列也可以用其编码的氨基酸序列来做，甚至于可通过相关蛋白质的结构比对来研究分子进化。以上研究已经积累了大量的工作。近年来由于较多模式生物基因组测序任务的完成，为从整个基因组的角度来研究分子进化提供了条件。可以设想，比较两个或多个完整基因组这一工作需要新的思路和方法，当然也渴望得到更丰硕的成果。这方面可做的工作是很多的。



- 
- 序列重叠群（**Contigs**）装配。
  - 一般来说，根据现行的测序技术，每次反应只能测出500 或更多一些碱基对的序列，这就有一个把大量的较短的序列全体构成了重叠群（**Contigs**）。逐步把它们拼接起来形成序列更长的重叠群，直至得到完整序列的过程称为重叠群装配。拼接EST数据以发现全长新基因也有类似的问题。

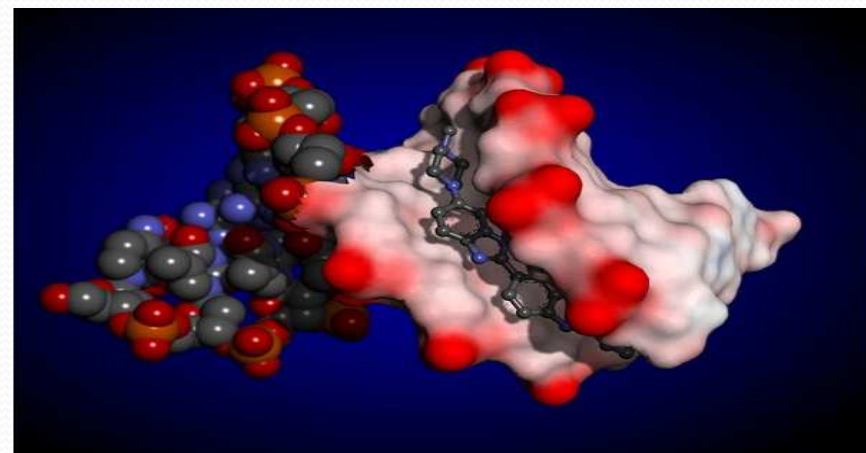
# 遗传密码的起源

- 遗传密码为什么是现在这样的？这一直是一个谜。一种最简单的理论认为，密码子与氨基酸之间的关系是生物进化历史上一次偶然的事件而造成的，并被固定在现代生物最后的共同祖先里，一直延续至今。不同于这种"冻结"理论，有人曾分别提出过选择优化、化学和历史等三种学说来解释遗传密码。随着各种生物基因组测序任务的完成，为研究遗传密码的起源和检验上述理论的真伪提供了新的素材。



# 基于结构的药物设计

- 人类基因组计划的目的是之一在于阐明人的约10万种蛋白质的结构、功能、相互作用以及与各种人类疾病之间的关系，寻求各种治疗和预防方法，包括药物治疗。基于生物大分子结构的药物设计是生物信息学中的极为重要的研究领域。为了抑制某些酶或蛋白质的活性，在已知其3级结构的基础上，可以利用分子对接算法，在计算机上设计抑制剂分子，作为候选药物。这种发现新药物的方法有强大的生命力，也有着巨大的经济效益。



## 与生物信息学关系密切的数学领域

- **统计学**，包括多元统计学，是生物信息学的数学基础之一；**概率论与随机过程理论**，如近年来兴起的隐马尔科夫链模型（HMM），在生物信息学中有重要应用；**运筹学**，如动态规划法是序列比对的基本工具，最优化理论与算法，在蛋白质空间结构预测和分子对接研究中有重要应用，**拓扑学**，这里指几何拓扑，在DNA超螺旋研究中是重要工具，在多肽链折叠研究中也有应用；**函数论**，如傅里叶变换和小波变换等都是生物信息学中的常规工具；**信息论**，在分子进化、蛋白质结构预测、序列比对中有重要应用，而人工神经网络方法则用途极为广泛；**计算数学**，如常微分方程数值解法是分子动力学的基本工具；群论，在研究遗传密码和DNA序列的对称性方面有重要应用；**组合数学**，在分子进化和基因组序列研究中十分有用。原则上讲，各种数学理论或多或少或直接或间接都应该在生物学研究中有各种各样的应用，其中包括生物信息学，这种情况正像过去的一、两个世纪，数学应用于物理学一样。而且，生物信息学的发展，又为数学的发展提供了一个新的机遇，可能会产生一些新的分支科学。



# 与生物信息学密切相关的计算机科学技术

- 首先是网络技术和数据库（特别是关系型数据库）管理技术，包括极为重要的实验室数据信息管理系统（LIMS）。其它诸如数据整合和可视化、数据挖掘（Data Mining）、基于Unix操作系统的各种软件包以及人工智能，和一些重要算法的复杂性研究。


# 生物信息学工业

- 生物信息学不仅具有重大的科学意义，而且具有巨大的经济效益。它既属于基础研究，以探索生物学自然学自然规律为己任；又属于应用研究，它的许多研究成果可以较快或立即产业化，成为价值很高的产品。生物信息学的这一特点在现有的许多学科中几乎是独一无二的。




- 这里仅举一个例子来说明生物信息学工业的潜力。据报导，只有50名员工的德国Lion生物信息学公司，将通过扫描公共数据库中的序列来发现500个可能的药物作用靶点，以一亿美元的价格预售给德国Bayer公司。



- 
- 又据报导，生物信息学产业的市场在1998年已经达到10亿美元，而到2002年估计可增长到2000亿美元以上。这是一笔巨大的财富，任何政府的科技决策人都不能对此视而不见。NIH已向美国国会建议投资160亿美元在美国建立5~20个将生物学与计算结合起来的中心。法国议会科技决策评估办公室，最近评估了基因工程、生物信息学和组合化学等学科的应用前景及法国的对策。美国出现了大批的基于生物信息学的公司，实施了许多生物信息学研究计划，主要与药物设计，基因工程药物，生物芯片，代谢工程与化学工程密切相关。生物信息学工业是知识经济的一个典型，潜力巨大。



- 
- 生物学是生物信息学的核心和灵魂，**数学与计算机技术则是它的基本工具**。这一点必须着重指出。预测生物信息学的未来主要就是要预测他对生物学的发展将带来什么样的根本性的突破。这种预测是十分困难的，甚至几乎不可能。但是人类科学研究史表明，科学数据的大量积累将导致重大的科学规律的发现。例如：对数百颗天体运行数据的分析导致了开普勒三大定律和万有引力定律的发现；数十种元素和上万种化合物数据的积累导致了元素周期表的发现；氢原子光谱学数据的积累促成了量子理论的提出，为量子力学的建立奠定了基础。历史的经验值得注意，有理由认为，今日生物学数据的巨大积累也将导致重大生物学规律的发现。

# 生物信息学与伦理

- 美国的HGP每年都拨出一定的款项资助伦理、法律和社会问题的研究 (ethical, legal. and social implications, 简称ELSI), 1998年的拨款为690万美元。
- 一个人的基因组应属于个人隐私, 其中含有什么致病基因或对某种疾病的易感基因, 用现代方法很容易查出, 查出结果泄漏出去, 被检对象就可能在就业、婚姻、保险等方面受到歧视。
- 试管婴儿、人工受精乃至克隆人等引发的伦理、法律和社会问题则更是公众十分关注的问题。
- 由人工进化产生的超级细菌、超级蛋白是否干扰了自然进化?



The background is a solid blue gradient. On the left side, there is a decorative wavy line that curves from the top to the bottom, with a lighter blue highlight on its inner curve.

Thank you !