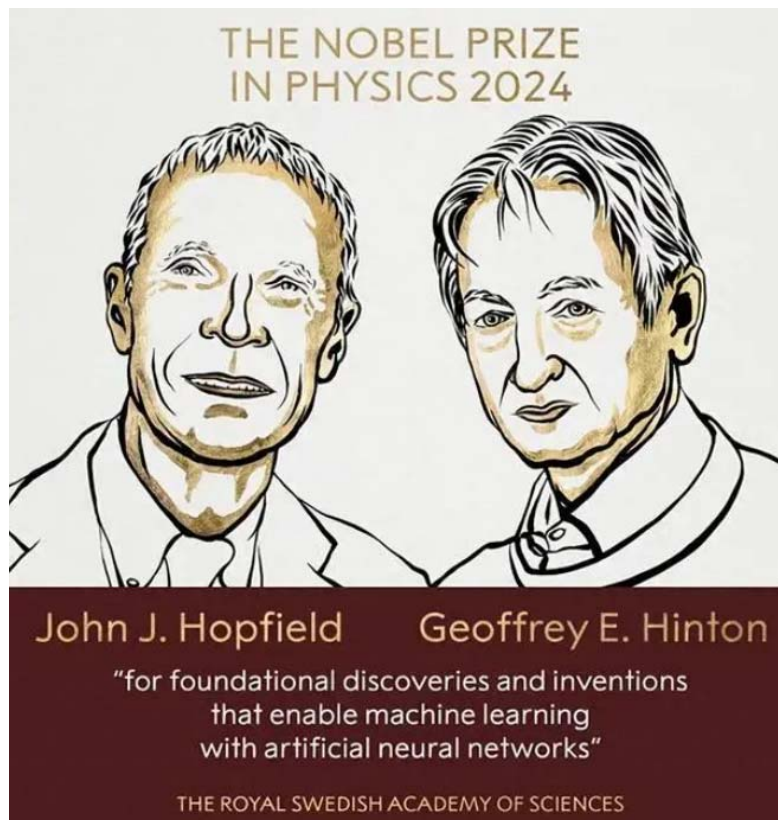




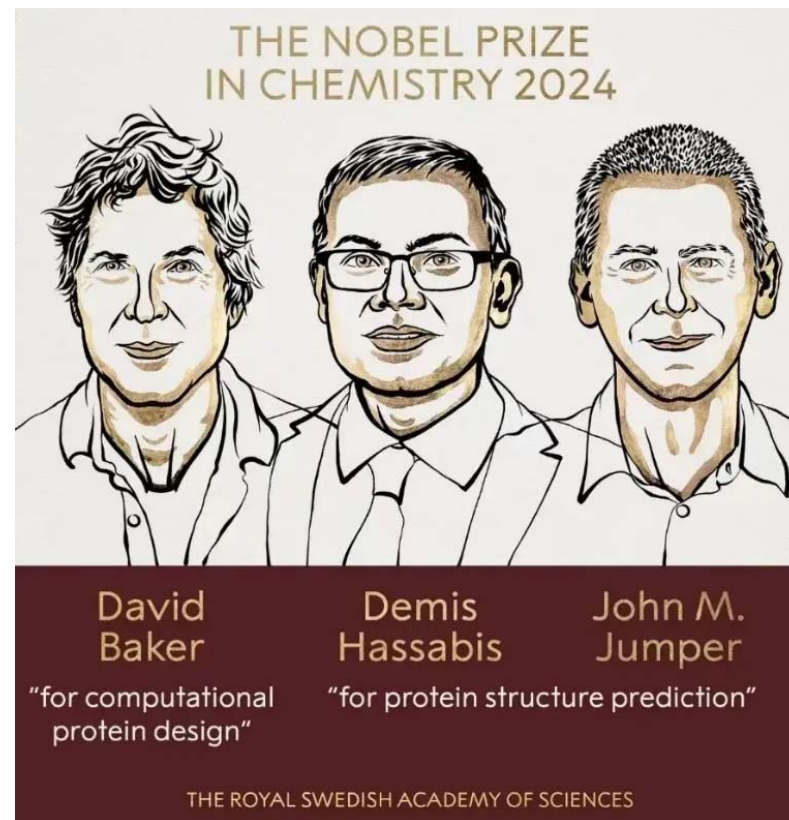
生物信息学

第一章 历史简介和概论

2024年诺贝尔物理学奖 & 化学奖



“基于人工神经网络实现机器学习的基础性发现和发明”



“计算蛋白质设计”、“蛋白质结构预测”

概率模型 vs. 语言模型



Primer

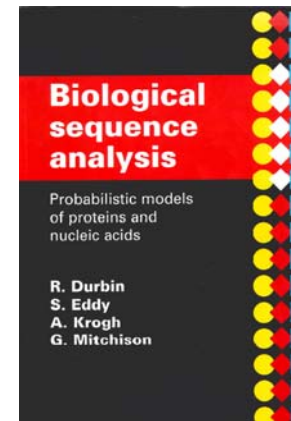
<https://doi.org/10.1038/s41587-024-02123-4>

Designing proteins with language models

Foundations of protein language models

Fundamentally, protein language models aim to predict how likely we are to observe a particular protein sequence S given all the protein sequence data collected thus far. We denote a protein sequence $S = (s_1, s_2, \dots, s_N)$, where s_i represents the amino acid at position i in the sequence. As a first approximation, we might consider the probability of observing a protein as the joint probability of observing each of its constituent amino acids. Under this model, referred to as unigram, we calculate the probability of a sequence S as

$$P(S) = \prod_i^N P(s_i)$$



生物信息学 1.0

In practice, to compute $P(S)$, we simply tabulate the frequency of each amino acid occurring in our sequence database and multiply the probabilities for the specific sequence S . However, proteins are not unordered collections of amino acids. Rather, the specific order in which we observe the amino acids is a critical determinant of structure and function. To capture this order dependency, we can use the preceding residues to inform the probability of the next amino acid. In an n -gram model, we multiply these contextualized probabilities to form the overall probability of the sequence:

$$P(S) = \prod_i^N P(s_i | s_{i-(n-1)}, \dots, s_{i-1})$$

AI生物学/生物信息学2.0

Nat Biotechnol, 2024, 42, 200-202

HUST

应该学生物信息学的N个理由



- ❑ 0. 计算是21世纪生物学研究的核心技能
- ❑ 1. 计算技能是高度可转移的
- ❑ 2. 计算能够帮助提高你的核心科学技能
- ❑ 3. 应当在博士或博后期间获得新的技能
- ❑ 4. 能够在生物学里建立更独特的技能
- ❑ 5. 可以发表更多的论文
- ❑ 6. 研究能有更大的灵活性
- ❑ 7. 工作场所不受限制
- ❑ 8. 计算研究的性价比高
- ❑ 9. 成功的科学家死在办公室
- ❑ 10. 能够知道为啥这个列表是从0开始



现代分子遗传学

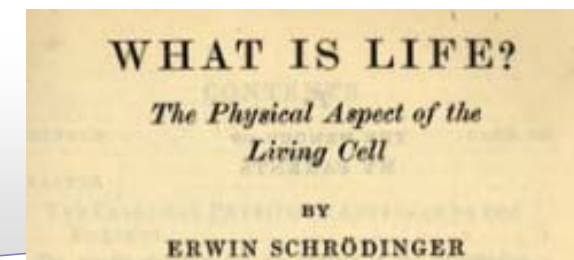


□ 分子遗传学的三大学派

✿ 信息学派、结构学派和功能学派

□ Erwin Schrödinger, 1944年, *What Is Life?*

- ✿ 从信息学的角度提出遗传密码的概念, 认为基因是“非周期性晶体” (大分子), 生物功能的基础是分子而不是原子
- ✿ 从量子力学的角度论证了基因的持久性和遗传模式长期稳定的可能性
- ✿ 提出了生命“以负熵为生”, 从环境中抽取“序”来维持系统的组织的概念 – 生命的热力学基础



噬菌体教堂的三主教



□ Max Delbrück, Salvador Luria, Alfred Hershey

- ✿ 利用噬菌体研究基因的功能
- ✿ 分子水平的孟德尔遗传学说
- ✿ 1969年诺贝尔奖

□ Max Delbrück (1906-1981)

- ✿ 1930s, Niels Bohr
- ✿ 1943, fluctuation test (波动实验)
- ✿ 1944, What is life?



信息学派的发展



❑ 量子力学/量子生物学/电子生物学

- ✿ Bernard Pullman (1919~1996)

- ✿ 建立International Academy of Quantum Molecular Science

❑ 分子动力学

- ✿ 整体牛顿力学，局部量子电动力学



WELCOME
NEWS
BUREAU
MEMBERS
HISTORY
STATUTES
AWARDS
CONGRESS

TOP NEWS

The International Academy of Quantum Molecular Sciences congratulates Martin Karplus (Univ. of Strasbourg and Harvard Univ.), Michael Levitt (Stanford Univ.), and Arieh Warshel (Univ. of Southern California) on winning the 2013 Nobel Prize in Chemistry for their work on Multiscale Models for Complex Chemical Systems.

Josef Michl
President

The Nobel Prize in Chemistry 2013



Photo: A. Mahmoud
Martin Karplus
Prize share: 1/3



Photo: A. Mahmoud
Michael Levitt
Prize share: 1/3



Photo: A. Mahmoud
Arieh Warshel
Prize share: 1/3

中国早期的理论生命科学研究



- 用食物中的aa频率与人/小鼠的aa频率来评价食物的营养价值
- 输入负熵，多余的aa会被转化为其它aa
- 小鼠的结果未必能完全应用到人

第2卷 第1期
1962年3月

生物化学与生物物理学报
ACTA BIOCHIMICA et BIOPHYSICA SINICA

Vol. 2, No. 1
March, 1962

生物体负熵输入的計算 (以蛋白質营养問題为例)

徐京華

(中国科学院生物化学研究所, 上海)

表8 各种蛋白質对青年成人信息輸入与
生物價值*的比較

		淨 信 息	生物价值*
鸡 蛋		1.6171	96
牛 奶		1.1793	90
酵 母		0.8773	79
牛 肉		0.8114	76
明 胶		0.2131	25

* 实验数据取自 B. L. Oser^[8]。

徐京华同志生平简介



中国共产党党员，1936 年参加革命，原中国科学院上海生物化学研究所研究员、享局级离休干部徐京华同志因病长期医治无效，于 2003 年 1 月 23 日 18 时 30 分不幸逝世，享年 80 岁。

徐京华，出生于 1922 年 9 月，上海市人。1936 年 9 月参加中华民族解放先锋队，1937 年 5 月参加中国共产党，长期从事党的地下工作。1938 年在桂林读高中，1944 年毕业于国立西南联合大学（昆明）理学院化学系。1945 年至 1950 年担任国立中央研究院医学研究所助理员，从事神经肌肉生理研究。

解放后，徐京华同志先后担任中国科学院上海生理生化研究所助理研究员、副研究员；中国科学院上海分院生物物理研究所筹备处研究员；中国科学院上海生物化学研究所研究员、研究室主任、课题组长。先后兼任中国科学院生物物理所、理论物理所、浙江大学生物工程系教授。曾担任中国生物物理学会副理事长、上海生物物理学会理事长、上海分子科学学会理事长、上海非线性动力研究学会理事长等职。

DNA测序

□ 1968年，吴瑞（Ray Wu） 发明第一个DNA测序方法

- ✿ “引物延伸法”（Primer extension methods）
- ✿ 首次测定噬菌体λ粘端的12个碱基



Victor Ling
林重庆



洪国藩

REVIEW

DNA sequencing at 40: past, present and future

Jay Shendure^{1,2}, Shankar Balasubramanian^{3,4}, George M. Church⁵, Walter Gilbert⁶, Jane Rogers⁷, Jeffery A. Schloss⁸ & ...

Nature, 2017, 550, 345-353

Technical milestones

1953: Sequencing of insulin protein²

1965: Sequencing of alanine tRNA⁴

1968: Sequencing of cohesive ends of phage lambda DNA⁶

1977: Maxam–Gilbert sequencing⁹

1977: Sanger sequencing⁸

1981: Messing's M13 phage vector¹²

1986–1987: Fluorescent detection in electrophoretic sequencing^{14,15,17}

1987: Sequenase¹⁸

1988: Early example of sequencing by stepwise dNTP incorporation¹³⁹

1990: Paired-end sequencing²³

1992: Bodipy dyes¹⁴⁰

1993: *In vitro* RNA colonies³⁷

1996: Pyrosequencing⁴⁴

1999: *In vitro* DNA colonies in gels³⁸

2000: Massively parallel signature sequencing by ligation⁴⁷

2003: Emulsion PCR to generate *in vitro* DNA colonies on beads⁴²

2003: Single-molecule massively parallel sequencing-by-synthesis^{33,34}

2003: Zero-mode waveguides for single-molecule analysis⁵⁷

2003: Sequencing by synthesis of *in vitro* DNA colonies in gels⁴⁹

2005: Four-colour reversible terminators^{51–53}

2005: Sequencing by ligation of *in vitro* DNA colonies on beads⁴¹

2007: Large-scale targeted sequence capture^{93–96}

2010: Direct detection of DNA methylation during single-molecule sequencing⁶⁵

2010: Single-base resolution electron tunnelling through a solid-state detector¹⁴¹

2011: Semiconductor sequencing by proton detection¹⁴²

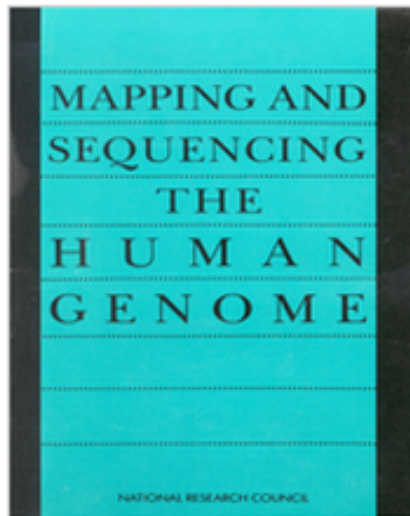
2012: Reduction to practice of nanopore sequencing^{143,144}

2012: Single-stranded library preparation method for ancient DNA¹⁴⁵

人类基因组计划



- ❑ 20年，\$200,000,000/年
- ❑ 人类基因组测序促进人类生物学和医学的发展
- ❑ 测序方法的发展是应用
- ❑ 信息和材料的收集、分析以及分配
- ❑ 比较遗传学方法：解释人类基因组的信息



Mapping and Sequencing the Human Genome

National Research Council (US) Committee on Mapping and Sequencing the Human Genome.

Washington (DC): [National Academies Press \(US\)](#); 1988.

ISBN-10: 0-309-03840-5

[Copyright and Permissions](#)

[Hardcopy Version at National Academies Press](#)

Search this book

<http://www.ncbi.nlm.nih.gov/books/NBK218252/>

生物信息学的发展历程



- ❑ 什么是生物信息学？
- ❑ Bioinformatics is a new subject of genetic **data** collection, analysis and dissemination to the research community (1987)
- ❑ Bioinformatics refers to **database-like** activities, involving persistent sets of data that are maintained in a consistent state over essentially indefinite periods of time (1994)



Hwa A. Lim

林华安

<http://www.dtrends.com/HAL.html>

什么是生物信息学？ (2)



- ❑ **Bioinformatics is the field of science in which biology, computer science, and information technology merge into a single **discipline**. The ultimate goal of the field is to enable the discovery of new biological insights as well as to create a global perspective from which unifying principles in biology can be discerned**

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

from NCBI's science primer

<https://www.ncbi.nlm.nih.gov/Class/MLACourse/Modules/MolBioReview/bioinformatics.html>

广义生物信息学观点



- ❑ Biology may be viewed as the study of transmission of information: from mother cell to daughter cell, from one cell or tissue type to another, from one generation to the next, and from one species to another. This informational viewpoint is termed **bioinformatics**
- ❑ 生物学研究可以被看成是研究信息的传递：从DNA经转录翻译到蛋白质，从细胞质中到细胞核内，从母细胞到子细胞，从一个细胞或一个组织到另一个细胞或另一个组织，从一代到下一代，从一个物种到另一个物种的进化演变。**这种信息论的观点可称为生物信息学** (Eisenberg *et al.*, 2006)

生物信息学发展过程中的里程碑性事件



时间	重大事件
1962	鲍林的分子进化理论
1965	Margaret Dayhoff 构建蛋白质序列图谱
1970	Needleman-Wunsch 算法开发
1977	DNA 测序以及使用利用 Staden 软件进行分析
1981	Smith-Waterman 算法开发
1981	序列模体的概念 (motif)
1983	序列数据库搜索算法的开发 (Wilbur-Lipman)
1985	快速序列相似性搜索工具 FASTP/FASTN 的开发
1990	快速序列相似性搜索工具 BLAST 的开发
1996	酵母基因组的完全测序
1997	PSI-BLAST 工具的开发
1998	秀丽线虫基因组的完全测序
1999	果蝇基因组的完全测序

基因组、计算和应用的里程碑性事件



Genome milestones

1977: Bacteriophage Φ X174 (ref. 72)
1982: Bacteriophage lambda¹³
1995: *Haemophilus influenzae*²⁶
1996: *Saccharomyces cerevisiae*²⁷
1998: *Caenorhabditis elegans*²⁸
2000: *Drosophila melanogaster*³²
2000: *Arabidopsis thaliana*¹⁴⁶
2001: *Homo sapiens*²⁹⁻³¹
2002: *Mus musculus*¹⁴⁷
2004: *Rattus norvegicus*¹⁴⁸
2005: *Pan troglodytes*¹⁴⁹
2005: *Oryza sativa*¹⁵⁰
2007: *Cyanidioschyzon merolae*¹²⁶
2009: *Zea mays*¹⁵¹
2010: Neanderthal⁸⁸
2012: Denisovan¹⁴⁵
2013: The HeLa cell line^{152,153}
2013: *Danio rerio*¹⁵⁴
2017: *Xenopus laevis*¹⁵⁵

Computational milestones

1981: Smith-Waterman¹⁵⁶
1982: GenBank (<https://www.ncbi.nlm.nih.gov/genbank/>)
1990: BLAST¹⁶
1995: TIGR assembler²⁴
1996: RepeatMasker
1997: GENSCAN¹⁵⁷
1998: phred, phrap, consed²²
2000: Celera assembler²⁵
2001: Bioconductor
2001: EULER⁷⁴
2002: BLAT¹⁵⁸
2002: UCSC Genome Browser¹⁵⁹
2002: Ensembl¹⁶⁰
2005: Galaxy¹⁶¹
2007: NCBI Short Read Archive
2008: ALLPATHS¹⁶²
2008: Velvet⁷⁵
2009: Bowtie⁸³
2009: BWA⁸²
2009: SAMtools⁸⁴
2009: BreakDancer¹⁶³
2009: Pindel¹⁶⁴
2009: TopHat¹¹⁵
2010: SOAPdenovo¹⁶⁵
2010: GATK⁸⁵
2010: Cufflinks¹¹⁶
2011: Integrated Genomics Viewer¹⁶⁶
2013: HGAP/Quiver¹⁶⁷
2017: Canu⁸¹

Application milestones

1977: Genome sequencing⁷²
1982: Shotgun sequencing¹³
1983, 1991: Expressed sequence tags^{107,108}
1995: Serial analysis of gene expression¹⁰⁹
1998: Large-scale human SNP discovery¹⁶⁸
2004: Metagenome assembly¹²²
2005: Bacterial genome resequencing with NGS^{40,41}
2007: Chromatin immunoprecipitation followed by sequencing (ChIP-seq) using NGS¹¹⁷
2007–2008: Human genome and cancer genome resequencing using NGS^{55,90-92}
2008: RNA-seq using NGS¹¹⁰⁻¹¹⁴
2008: Chromatin accessibility using NGS¹¹⁸
2009: Exome resequencing using NGS⁹⁷
2009: Ribosome profiling using NGS¹¹⁹
2010: Completion of Phase I of the 1000 Genomes Project⁹⁸
2010: *De novo* assembly of a large genome from short reads¹⁶⁹
2011: Haplotype-resolved human genome resequencing using NGS^{170,171}
2016: Human genome *de novo* assembly with PacBio¹⁷²
2017: Human genome *de novo* assembly with nanopore⁶⁴

分子进化理论：鲍林

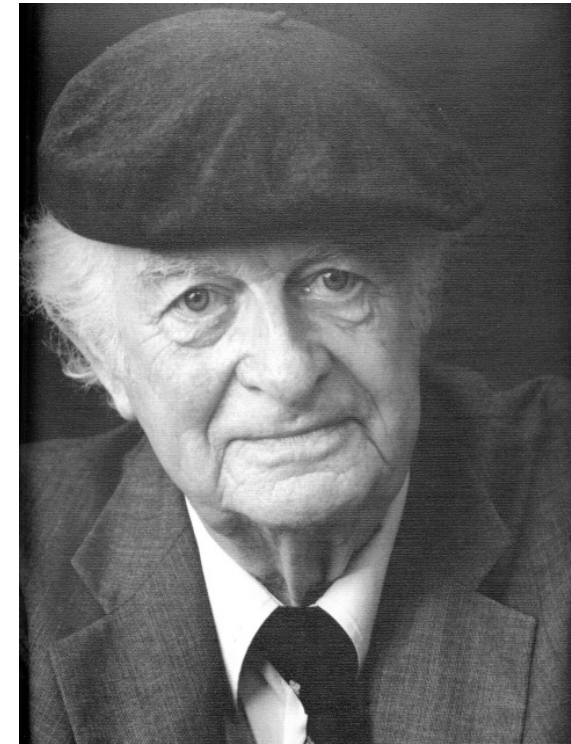


Molecular Disease and Evolution

by

Linus Pauling

(Rudolf Virchow Lecture, 5 November 1962)



1901 ~ 1994

I believe that it is likely that a human being manufactures 50,000 or 100,000 different kinds of protein molecules. A representative protein molecule, such as hemoglobin, is built of about 10,000 atoms. It has a well defined structure; for most of the protein molecules not a single atom is out of place.

Dayhoff 打分矩阵



- ❑ 1978年, 34个蛋白质超家族
- ❑ 序列相似性 $\geq 85\%$
- ❑ 1572个点突变
- ❑ PAM: Accepted Point Mutation, 可接受的点突变
- ❑ 马尔可夫模型 (Markov Model): 位点突变速率独立均等
- ❑ PAM1: 序列分歧 $\sim 1\%$ 时的氨基酸替代打分矩阵

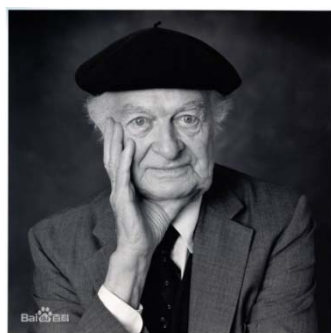


Margaret Dayhoff,
1925 ~ 1983

生物信息学：爸爸/妈妈是哪位？



弗雷德里克·桑格
是我？



莱纳斯·卡尔·鲍林
不是我？



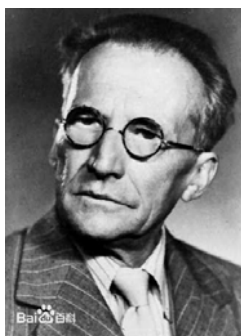
坦布尔·斯密
右边说是我



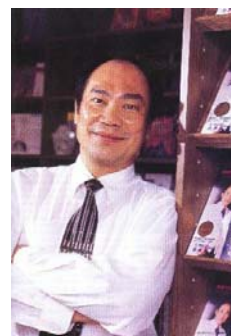
乔治·贝尔
左边这位！



鲍琳·霍奇维格
我都说是我了



薛定谔
关我啥事啊？



林华安
其实应该是我



玛格丽特·奥克雷·戴霍夫
那我呢？

国际四大核酸序列数据库



- ❑ 1974年，George I. Bell等人收集DNA序列，构建GenBank数据库，1982开发第一个版本
- ❑ 1980年，欧洲EBI-ENA数据库建立
- ❑ 1984年，日本DDBJ数据库建立成立
- ❑ 2016年，中科院北京基因组所BIG Data Center at Beijing Institute of Genomics (BIGD) 数据库建立；中科院上海营养与健康研究所Bio-Med Big Data Center (BMDC)数据库建立
- ❑ 2019年，BIGD改为**NGDC** (National Genomics Data Center)；BMDC改成NGDC/NODE (National Omics Data Encyclopedia)

NGDC



☐ <https://ngdc.cncb.ac.cn/>

CNCB NGDC

Databases Tools Standards Publications About

国家基因组科学数据中心
National Genomics Data Center

Nstn
国家科技资源共享服务平台

登录 语言 / Language ▾

面向我国人口健康和社会可持续发展的重大战略需求，建立生命与健康大数据汇交存储、安全管理、开放共享与整合挖掘研究体系，研发大数据前沿交叉与转化应用的新方法和新技术，建成支撑我国生命科学发展、国际领先的基因组科学数据中心。

All databases

Find a bioproject, biosample, gene, protein, tool, database...

Q Search

e.g., PRJCA000126; SAMC000385; tp53; EGFR; human; KaKs_Calculator; GenBank

国家基因组科学数据中心发布人类癌症可变剪接知识库AScancer Atlas, 欢迎使用!

提交

科学项目数据汇交

人类遗传资源信息管理备份

序列搜索比对

新冠病毒信息库

文献库

» 资源

热门资源

原始数据

基因组和变异

基因表达

非编码RNA

表观基因组

BioCode 生物工具软件

BioProject 生物项目库

BioSample 生物样本库

GSA 组学原始数据归档库

GSA-Human 人类组学原始数据归档库

OMIX 多元数据归档库

2019新型冠状病毒信息库



☐ <https://ngdc.cncb.ac.cn/ncov/>

首页 基因组序列 ▾ 基因组变异 ▾ 基因组及变异知识 ▾ 监测预警 ▾ 临床信息 文献情报 在线工具 ▾ 关于 ▾ 语言 / Language ▾

RCoV19

Version 4.0

2019新型冠状病毒信息库

New

Q 搜索

高级搜索

数据提交

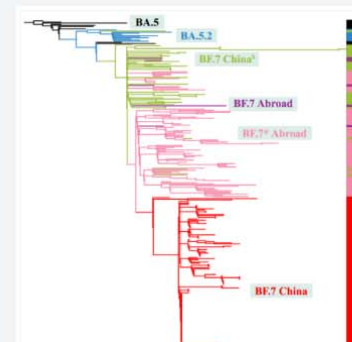
数据下载

原始序列

2022年10-12月份中国流行变异株的报告 (2023-01-09)

自2022年10月以来,中国大陆测序公开817个SARS-CoV-2序列,涵盖79个Pangolin谱系(<https://ngdc.cncb.ac.cn/ncov/monitoring/country/China>). 最大的谱系之一是BF.7 (278, 34%), 其序列样本来自2022年10月至12月的北京 (208, 74.8%)、上海 (30, 10.8%) 和福建 (18, 6.5%)。序列比较分析发现, 中国主要传播流行的BF.7序列与国际BF.7及其18个亚系 (截至2023年1月7日) 完全不同, 相较于国际BF.7序列增加了五个新突变, 即ORF1a:V274L、ORF1b:L238F、S:C1243F、ORF7a:H47Y和ORF10:C29632T。这表明, 在中国有一个独立的BF.7亚系在传播, 近期也被国际科学家提出并命名为BF.7.14(<https://github.com/cov-lineages/pango-designation/issues/1470>)。

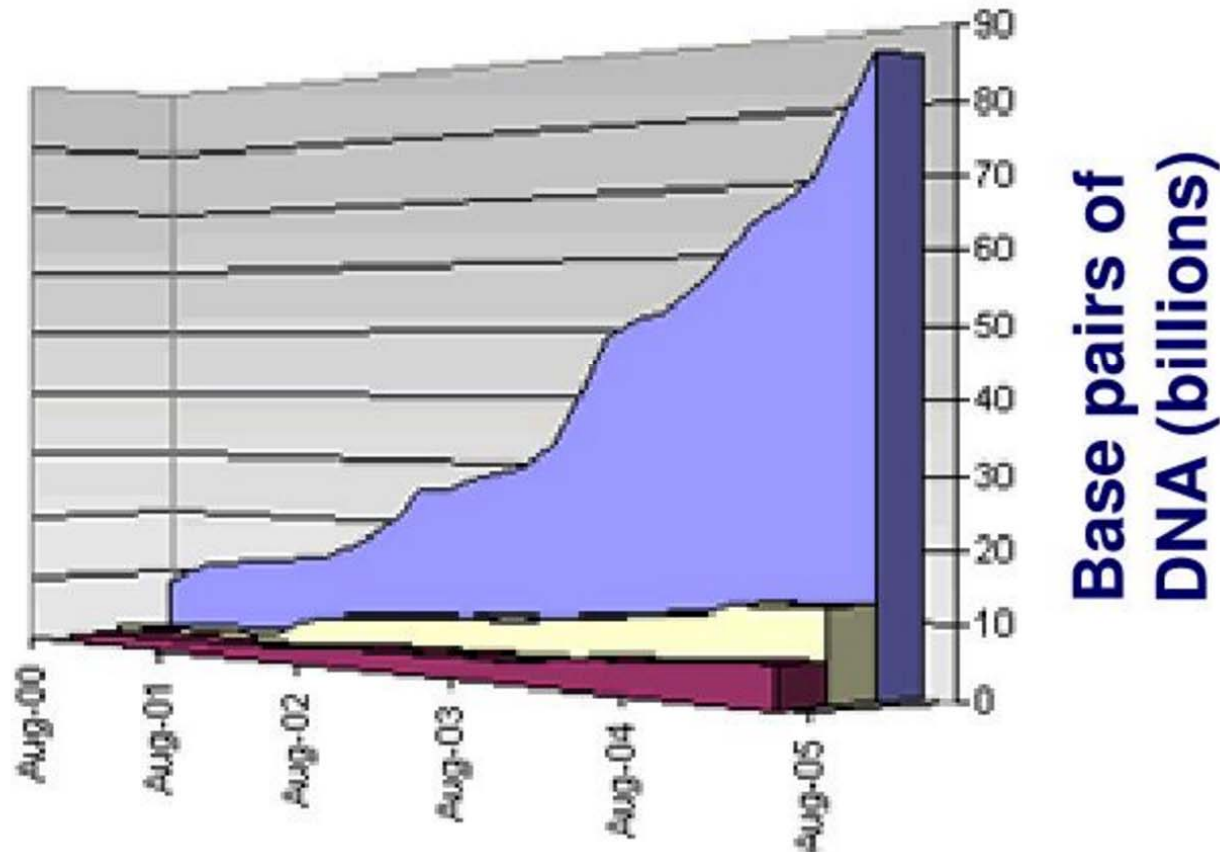
» 更多信息



核酸数据库的数据增长



Growth of the International Nucleotide Sequence Database Collaboration



Base pairs contributed by GenBank ■ EMBL ■ DDBJ ■

获取序列及检索公共数据库



- ❑ Entrez的开发, David J. Lipman等人
- ❑ 提供关键字的搜索的方法
- ❑ “硬搜索”: 包含关键字的, 完全匹配的结果
- ❑ “软搜索”: 与查询内容相关的信息
- ❑ 查询内容: 基因/蛋白质的名称、标识符, 文献、蛋白质结构等

Search NCBI



❏ <https://www.ncbi.nlm.nih.gov/search/>

 **National Library of Medicine**
National Center for Biotechnology Information

Log in

Search NCBI

Search NCBI

Search

COVID-19 resources

[COVID-19 news](#) >



COVID-19 treatment

Coronavirus Disease 2019 (COVID-19) Treatment Guidelines (National Institutes of Health)



SARS-CoV-2 genomes

Download viral genome and protein sequences, annotation and a data report



NCBI Virus

The most up-to-date set of SARS-CoV-2 nucleotide and protein sequences

Literature

PubMed

PubMed® comprises more than 35 million citations for biomedical literature from MEDLINE, life science journals, and online books. Citations may include links to full text content from PubMed Central and publisher web sites.

PubMed.gov

Example searches Search for titles, citations, identifiers and more

[Cancer Statistics 2021](#)

[Tunyasuvunakool, Nature 2021](#)

[Revealing protein-protein interactions by transcriptome sequencing](#)

[GeneReviews](#)

Literature databases

Bookshelf

Books and reports

MeSH

Ontology used for PubMed indexing

NLM Catalog

Books, journals and more in the NLM Collections

PubMed

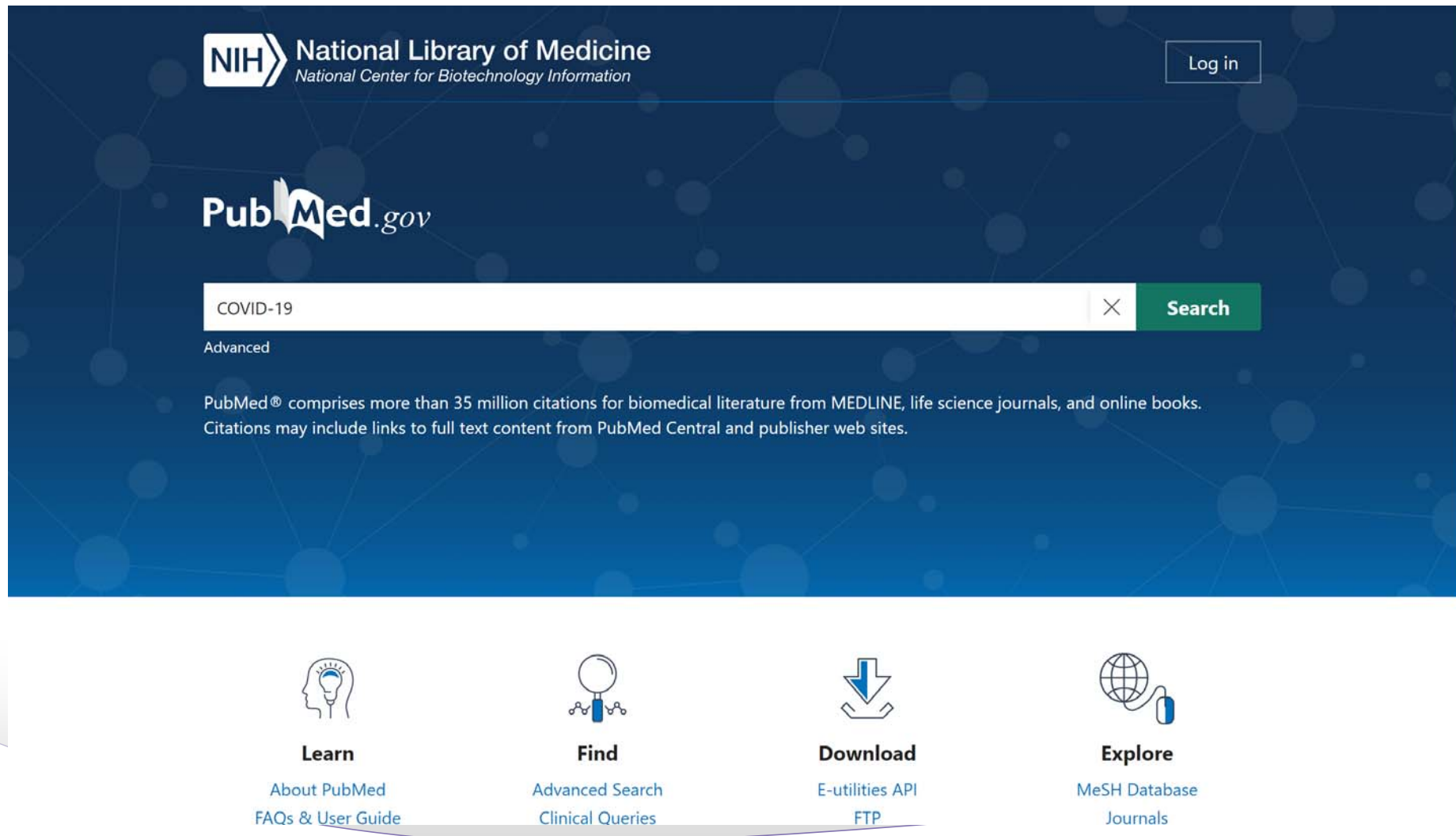
Scientific and medical abstracts/citations

PubMed Central

PubMed: 文献检索



□ 1992年, Entrez项目之一



The screenshot shows the PubMed.gov homepage with a dark blue background and a network diagram. At the top left is the NIH logo and the text "National Library of Medicine National Center for Biotechnology Information". At the top right is a "Log in" button. The "PubMed.gov" logo is prominently displayed. Below it is a search bar containing "COVID-19" and a green "Search" button. Under the search bar, the word "Advanced" is visible. A paragraph of text states: "PubMed® comprises more than 35 million citations for biomedical literature from MEDLINE, life science journals, and online books. Citations may include links to full text content from PubMed Central and publisher web sites." At the bottom, there are four columns of links: "Learn" (About PubMed, FAQs & User Guide), "Find" (Advanced Search, Clinical Queries), "Download" (E-utilities API, FTP), and "Explore" (MeSH Database, Journals).

NIH National Library of Medicine
National Center for Biotechnology Information

Log in

PubMed.gov

COVID-19 Search

Advanced

PubMed® comprises more than 35 million citations for biomedical literature from MEDLINE, life science journals, and online books. Citations may include links to full text content from PubMed Central and publisher web sites.

Learn
About PubMed
FAQs & User Guide

Find
Advanced Search
Clinical Queries

Download
E-utilities API
FTP

Explore
MeSH Database
Journals

PubMed: 文献检索



- ☐ Abstract
- ☐ Free full text
- ☐ Full text

Use COVID-19 filters from PubMed Clinical Queries to refine your search

[Treatment](#) [Mechanism](#) [Transmission](#) [More filters](#)

[See more SARS-CoV-2 literature, sequence, and clinical content from NCBI](#)



COVID-19 diagnosis -A review of current methods.

1

Yüce M, Filiztekin E, Özkaya KG.

Cite

Biosens Bioelectron. 2021 Jan 15;172:112752. doi: 10.1016/j.bios.2020.112752. Epub 2020 Oct 24.

PMID: 33126180

[Free PMC article.](#)

[Review.](#)

Share

A fast and accurate self-**testing** tool for **COVID-19** diagnosis has become a prerequisite to comprehend the exact number of cases worldwide and to take medical and governmental actions accordingly. **SARS-CoV-2** (formerly, **2019-nCoV** ...



□ 2011年，新的疾病分类方法 (New Taxonomy)



Toward Precision Medicine

[< Prev](#)[Next >](#)

Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease

National Research Council (US) Committee on A Framework for Developing a New Taxonomy of Disease.

Washington (DC): [National Academies Press \(US\)](#); 2011.

ISBN-13: 978-0-309-22222-8 ISBN-10: 0-309-22222-2

[Copyright and Permissions](#)

[Hardcopy Version at National Academies Press](#)

Search this book

Excerpt

A Knowledge Network of Disease could embrace and inform rapidly expanding efforts by the biomedical research community to define at the molecular level the disease predispositions and pathogenic processes occurring in individuals. This network has the potential to play a critical role across the globe for the public-health and health-care-delivery communities by enabling development of a more accurate, molecularly-informed taxonomy of disease.

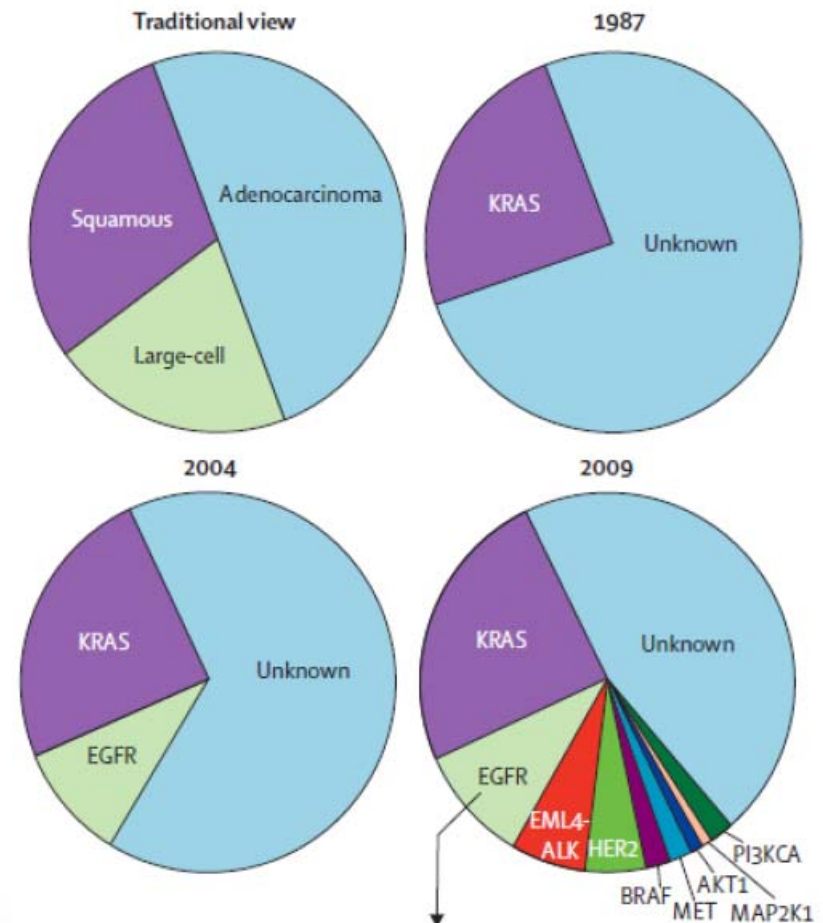
<http://www.ncbi.nlm.nih.gov/books/NBK91503/>

新的疾病分类方法



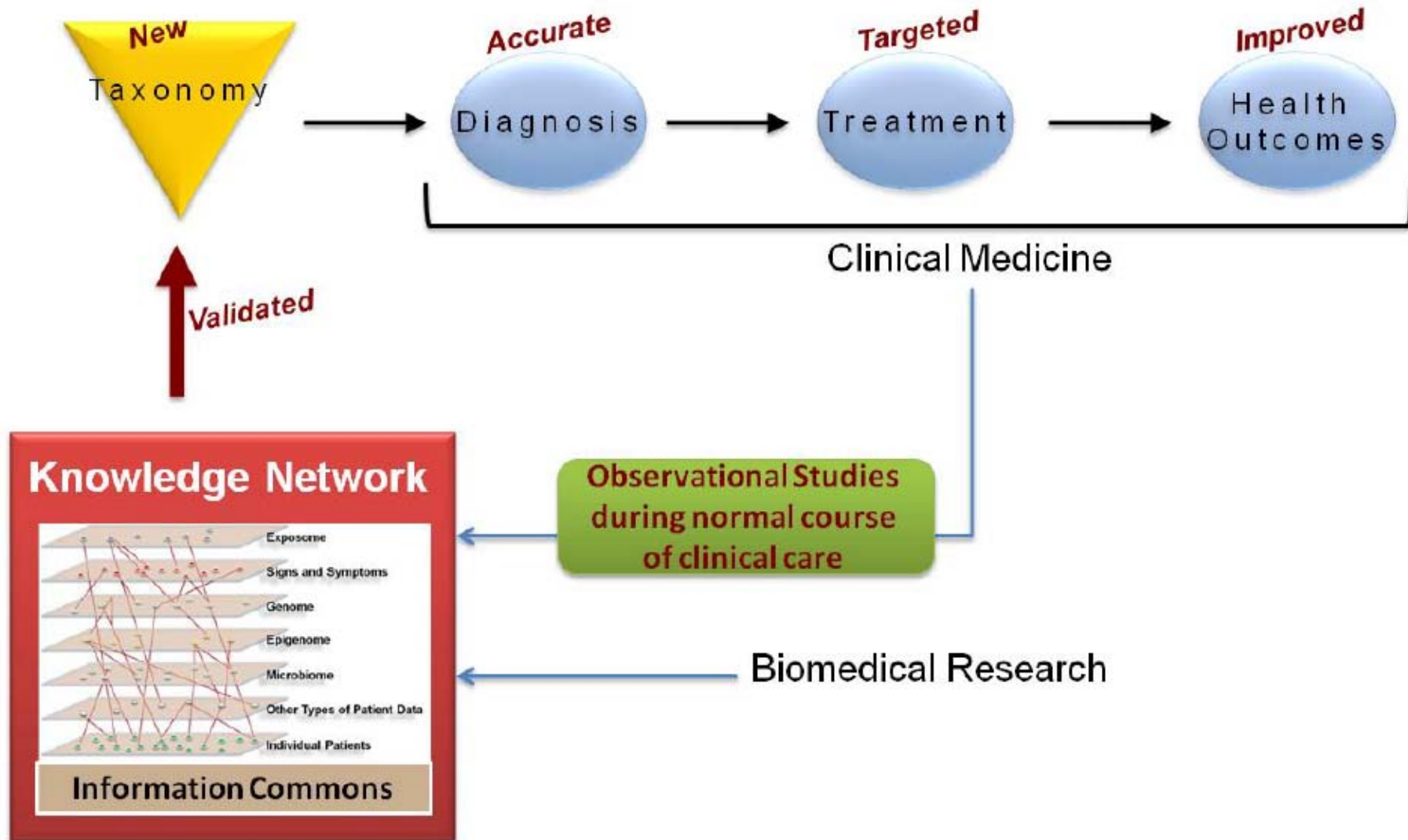
□ 分子组学数据

- 🌸 Genome (基因组)
- 🌸 Transcriptome (转录组)
- 🌸 Proteome (蛋白质组)
- 🌸 Metabolome (代谢组)
- 🌸 Lipidome (脂质组)
- 🌸 Epigenome (表观组)
- 🌸 Meta-genome (宏基因组)



- Mutations associated with drug sensitivity
EGFR Gly719X, exon 19 deletion, Leu858Arg, Leu861Gln
- Mutations associated with primary drug resistance
EGFR exon 20 insertions
- Mutations associated with acquired drug resistance
EGFR Thr790Met, Asp761Tyr, Leu747Ser, Thr854Ala

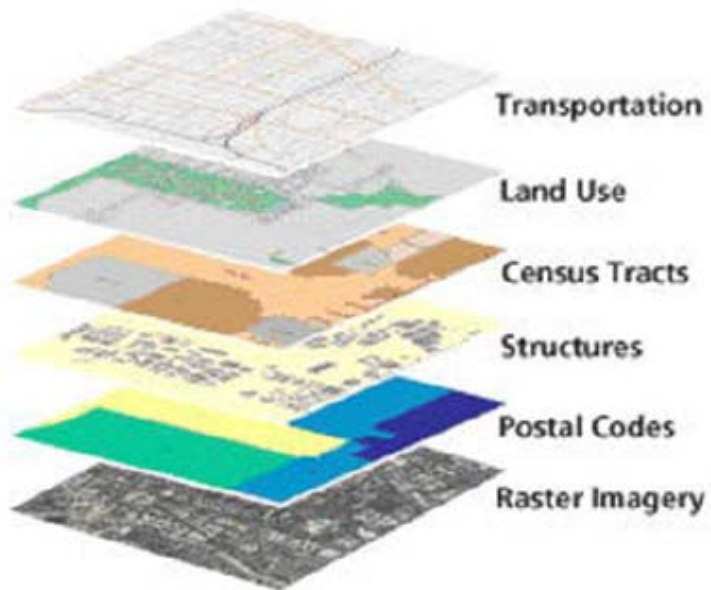
信息共享 (Information Commons)



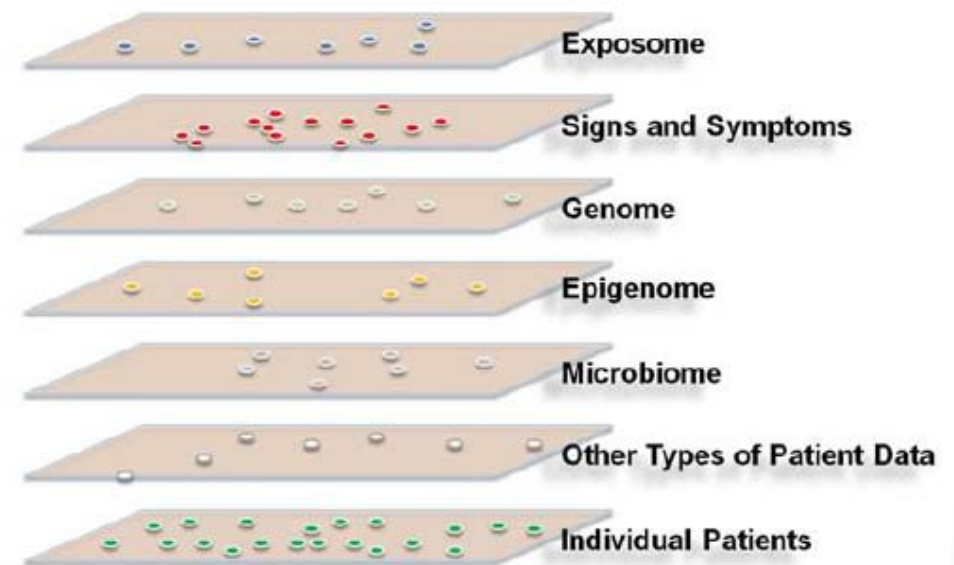
信息共享 (Information Commons)



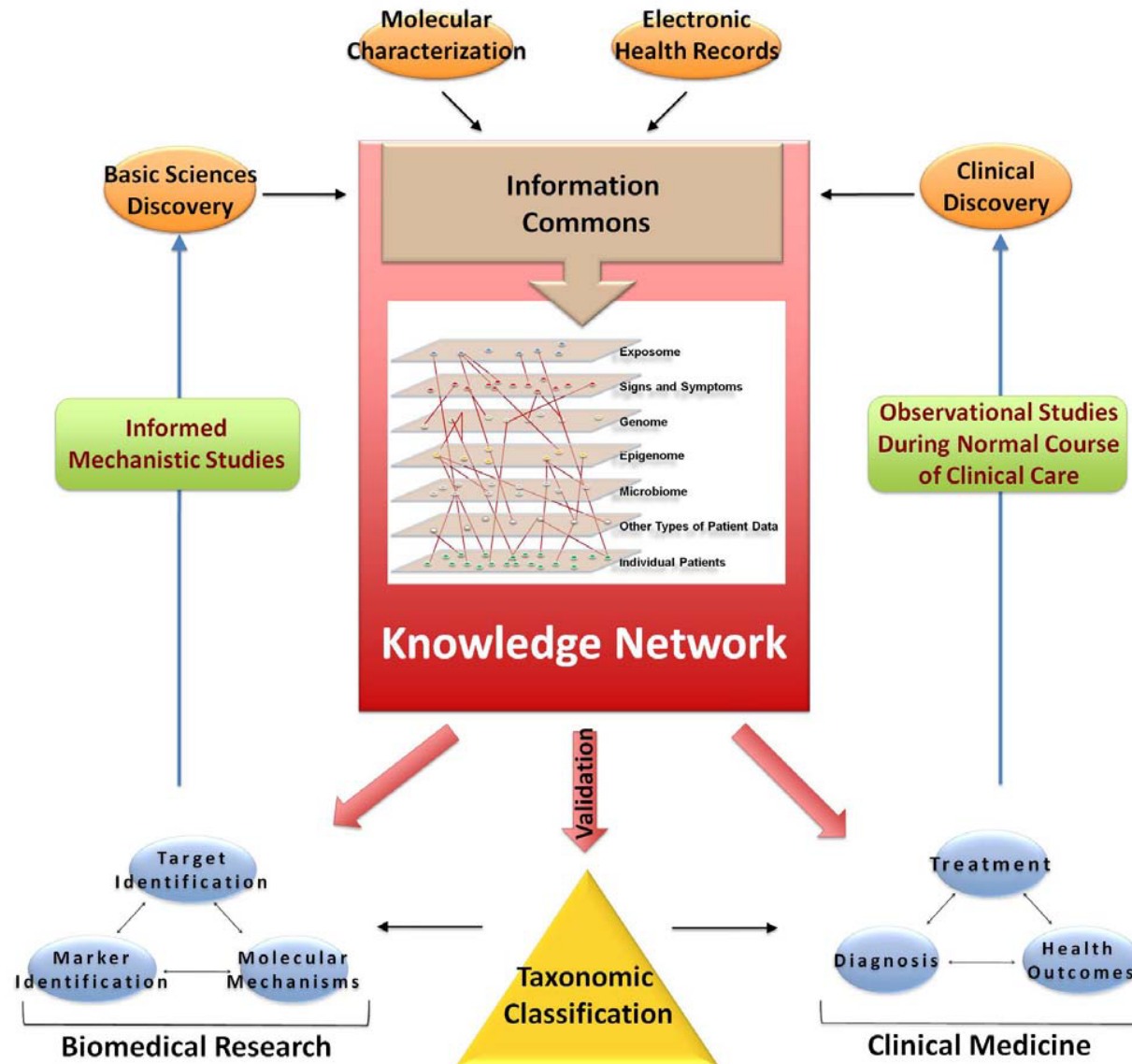
Google Maps: GIS layers
Organized by Geographical Positioning



Information Commons
Organized Around Individual Patients



Biomedical Knowledge Network



生物信息学主要研究内容与方向



- 开发新的**算法及统计学方法**来揭示**大数据**之间的联系
- 开发、设计相关的**数据库和工具**，能够方便有效的获取、管理以及使用各种类型的数据和信息
- **分析和解释**各种类型的生物学数据，包括核酸、氨基酸序列、蛋白质功能结构域以及蛋白质三级结构等

<http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/milestones.html>

中国生物信息学的研究方向



- ❑ 1. 基因组信息学
- ❑ 2. 转录组信息学
- ❑ 3. 表观组信息学
- ❑ 4. 蛋白质组信息学
- ❑ 5. 修饰组信息学
- ❑ 6. 结构信息学
- ❑ 7. 单细胞生物信息学
- ❑ 8. 三维基因组信息学
- ❑ 9. 微生物组信息学
- ❑ 10. 计算癌症基因组学
- ❑ 11. 计算系统生物学
- ❑ 12. 药物信息学
- ❑ 13. 人工生物系统的设计与控制
- ❑ 14. 生物信息数据挖掘方法
- ❑ 15. 生命与健康大数据科学
- ❑ 16. 精准健康信息学
- ❑ 17. 群体遗传与计算演化基因组学
- ❑ 18. 生物影像信息学
- ❑ 19. 植物信息学
- ❑ 20. 生物信息学研究新方向、新技术与新方法
- ❑ 21. 人工智能与生物学

中国生信相关的学会、协会（二级）



- ❑ 中国生物信息学学会（筹）
- ❑ 中国细胞生物学学会功能基因组信息学与系统生物学分会
- ❑ 中国生物工程学会计算生物学与生物信息学专业委员会
- ❑ 中国计算机学会生物信息学专委会
- ❑ 中国遗传学会基因组学专业委员会
- ❑ 中国生物物理学会生物信息学与理论生物物理专业委员会
- ❑ 中国生物化学与分子生物学会分子系统生物学专业委员会
- ❑ 中国运筹学会计算系统生物学分会
- ❑ 中国人工智能学会生物信息学与人工生命专业委员会
- ❑ 中国电子学会生物计算与生物信息处理专业委员会
- ❑ 中国医药生物技术协会生物医学信息技术分会
- ❑ 中国交叉科学学会生物信息学专业委员会
- ❑ 中国系统仿真学会生命系统建模仿真专业委员会
- ❑ 中国生物物理学会人工智能生物学分会

湖北省生物信息学会



- ❑ 2018年10月16日，湖北省科学技术协会批准
- ❑ 2018年12月8日，湖北省生物信息学会成立大会、第四届湖北省生物信息学青年学者论坛暨在汉四校两所联合生物信息研究生年会
- ❑ 2019年3月29-31日，第六届华中-华东地区生物信息学研讨会，宜昌，三峡大学
 - ✿ 终身成就奖 & 杰出贡献奖
- ❑ 征集学会的logo
- ❑ 生物信息分析认证体系：等级考试



国内相关专业介绍



- 华中科技大学
 - ✿ 生命学院：宁康、郭安源、薛宇、陈卫华
 - ✿ 物理学院：肖奕、黄胜友、刘士勇
 - ✿ 其他学院：潘林强、何西淼、王超龙、夏天
- 武汉大学：刘娟、赵华斌、周宇、邹秀芬
- 华中农业大学：张红雨、陈玲玲、李国亮、谢为博
- 北京大学：魏丽萍、高歌、李川昀、来鲁华、欧阳颀、邓明华、席瑞斌、崔庆华、陆剑、李程
- 清华大学：张学工、汪小我、李梢、鲁志、杨雪瑞、张强锋、曾坚阳
- 上海同济大学：刘小乐、江赐忠、张勇、曹志伟、刘琦

985、211高校



- ❑ 厦门大学：纪志梁
- ❑ 中国科学技术大学：刘海燕、李骛、瞿昆
- ❑ 复旦大学：赵兴明、倪挺、田卫东
- ❑ 天津大学：高峰
- ❑ 浙江大学：陈铭、樊龙江
- ❑ 上海交通大学：沈红斌、王侃侃
- ❑ 中山大学：贺雄雷、任间、谢志、骆观正、王金凯
- ❑ 南方医科大学：张镇海、王栋
- ❑ 四川大学：沈百荣、谢丹
- ❑ 中国农业大学：王向峰

中科院系统 & 研究所



- ❑ 北京中科院生物物理所：陈润生
- ❑ 北京基因组所：章张、王前飞、张治华、方向东
- ❑ 北京中科院遗传所：王秀杰、钱文峰
- ❑ 北京中科院计算所：赵屹
- ❑ 北京中科院数学与系统科学研究院：张世华、王勇
- ❑ 北京生命科学研究院：赵方庆、吴金雨
- ❑ 北京动物所：张勇
- ❑ 上海生物信息中心：李亦学、谢鹭
- ❑ 上海系统生物学实验室：陈洛南
- ❑ 上海马普所：王泽峰、韩敬东、徐书华、杨力、李海鹏
- ❑ 中国医学科学院基础医学研究所：蒋太交

知名企业：华大基因



大平台-研究平台

大科学-重大项目

人类遗传学研究

炎黄计划
国际千人基因组计划
糖尿病关联基因及变异研究
国际癌症基因组计划
古人类基因组
中华千种单基因病研究计划
人类蛋白质组计划

微生物研究

人体肠道微生物基因组研究计划
万种微生物基因组计划

动植物研究

千种植物基因组参考序列图谱构建计划
万种脊椎动物基因组计划
国际大熊猫基因组计划
黄瓜基因组计划
兰花基因组计划
桑蚕基因组重测序
大豆基因组重测序
千种植物转录组研究
“世界三极”动物基因组计划

表观遗传学研究

人类基因组甲基化研究
家蚕基因组甲基化研究

研究机构

生物化学技术研究所
生物智能技术研究所
精准健康研究所
数字化地球研究所
超级细胞研究所

研究成果

发表文章

学术期刊

GigaScience



[2018-11-23] 华大与密歇根大学医院签署谅解备忘录 建立教育、科学、临床实践合作关系

更多 >

知名企业：诺禾致源



诺禾致源 2019 校园招聘

你好, unique

诺禾未来 致源 有你!

Novogene 诺禾致源

首页 科技服务 肿瘤合作研究 遗传合作研究 样本系统 关于诺禾

科技服务: 400 658 1585 咨询客服

联系客服, 在线咨询

市场营销 生物信息 技术支持 数据转化

知名企业：贝瑞基因



www.berrygenomics.com

贝瑞基因

首页 贝瑞服务 贝瑞产品 服务支持 投资者关系 关于我们 EN

全外显子组检测

创新注解系统，提高临床诊断率

[了解详情](#)

贝比安

认识贝比安

贝比安能检测什么

哪些人可以选择贝比安

贝比安检测后应该做什么

贝比安plus

认识贝比安Plus

技术

技术原理

技术优势

临床

国内临床试验

产前检测

流产原因排查

染色体疾病检测

自闭症原因排查

资源

FAQ

参考文献

遗传疾病公共数据库

视频演示

科诺安

认识科诺安

科诺安能检测什么

哪些人可以选择科诺安

技术

技术原理

技术优势

临床

国内临床试验

产前检测

流产原因排查

染色体疾病检测

自闭症原因排查

资源

FAQ

参考文献

遗传疾病公共数据库

贝全安全外显子组检测和拷贝数变异检测

认识贝全安-WES

贝全安能检测什么

为何选择贝全安-WES

技术

技术原理

技术优势

临床

临床检测服务流程

临床应用方向

检测服务特点

样本采集要求

贝全安-WES+CNV-Seq

昂科益

认识昂科益

什么是肿瘤

肿瘤的相关基因

认识昂科益

昂科益检测后应该做什么

昂科益肿瘤分子诊断

非小细胞肺癌基因突变检测

遗传性肿瘤基因突变检测

BRCA1/2全外显子基因突变检测

结直肠癌基因突变检测

微卫星不稳定 (MSI) 检测

实体瘤全靶点基因检测

技术

cSMART技术

目标区域捕获测序

资源

参考文献

携心安

SMA携带者筛查

什么是SMA?

SMA分型

SMA的遗传机制

SMA携带者筛查的必要性

如何进行SMA携带者筛查

如果我是SMA携带者，怎么办?

FXS携带者筛查与患者确诊

什么是FXS携带者筛查与患者确诊

FXS携带者筛查与患者确诊的遗传机制

FXS携带者筛查与患者确诊的必要性

如何进行FXS携带者筛查与疾病确诊

如果我是FXS携带者，怎么办?

科技服务

概况

简介

文化理念

技术平台

计算资源

Illumina平台

基因组

转录组

表观

PacBio平台

PacBio平台介绍

PacBio全基因组de novo测序

全长转录组测序Iso-Seq

PacBio检测大片段结构变异SV

BioNano平台

BioNano平台介绍

BioNano辅助基因组组装

BioNano检测基因组大片段结构变异

10x Genomics平台

10x Genomics平台介绍

10x Genomics辅助基因组组装

10x Genomics 单细胞转录组测序

优势产品

人全基因组重测序 (WGS)

全外显子组测序 (WES)

全基因组de novo测序

资源

创新技术

www.berrygenomics.com/#

知名企业：武汉未来组



NextOmics 未来组

首页 科研服务 行业资讯 基因探索者招募计划 关于我们 简 希望组

动植物基因组

- ▶ 动植物基因组 de novo
- ▶ 动植物基因组组装升级
- ▶ 动植物全基因组重测序
- ▶ 动植物基因组结构变异 (SV) 分析

微生物及宏基因组

- ▶ 细菌基因组完成图
- ▶ 真菌基因组近完成图
- ▶ 全长核糖体小亚基 16S/18S/ITS 3.0
- ▶ 宏基因组 3.0
- ▶ 16S/18S/ITS 扩增子测序
- ▶ 宏基因组 2.0

转录组学研究

- ▶ 全长转录组 (iso-seq, 基于 Pacbio/Nanopore 平台)
- ▶ RNA-Seq (PE150, 20万 reads)

DNA 修饰图谱 3.0

- ▶ 全基因组 DNA 修饰图谱 3.0
- ▶ 富集 DNA 修饰图谱 3.0

小基因组完成图

- ▶ 小基因组完成图 3.0

基因组在线数据库

- ▶ 基因组在线数据库 (Genome database)

明星产品

动植物基因组 de novo 3.0

重新定义大基因组组装指标 Contig N50>1M
Scaffold N50>5M

全长转录组

无需拼接的全长转录本
精确鉴定可变剪切和融合基因信息

细菌基因组 完成图 3.0

一个重叠群一染色体
无Gap, 无N碱基错误

Bioinformatics in China



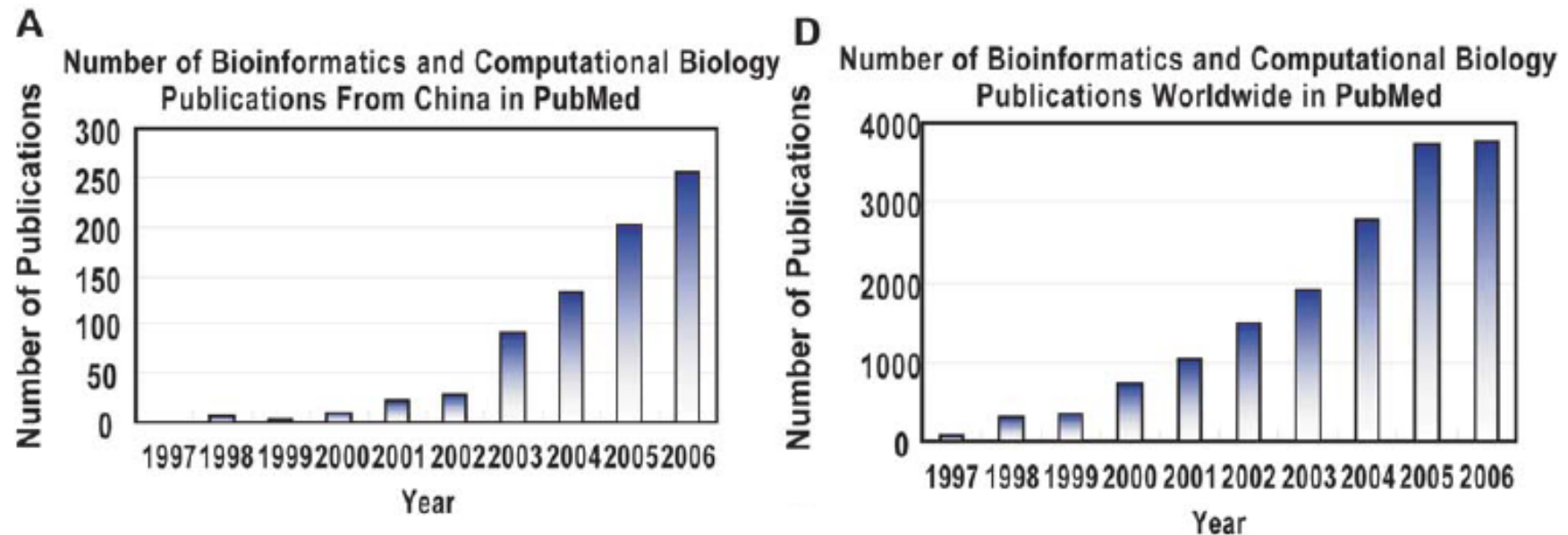
OPEN ACCESS Freely available online

PLOS COMPUTATIONAL BIOLOGY

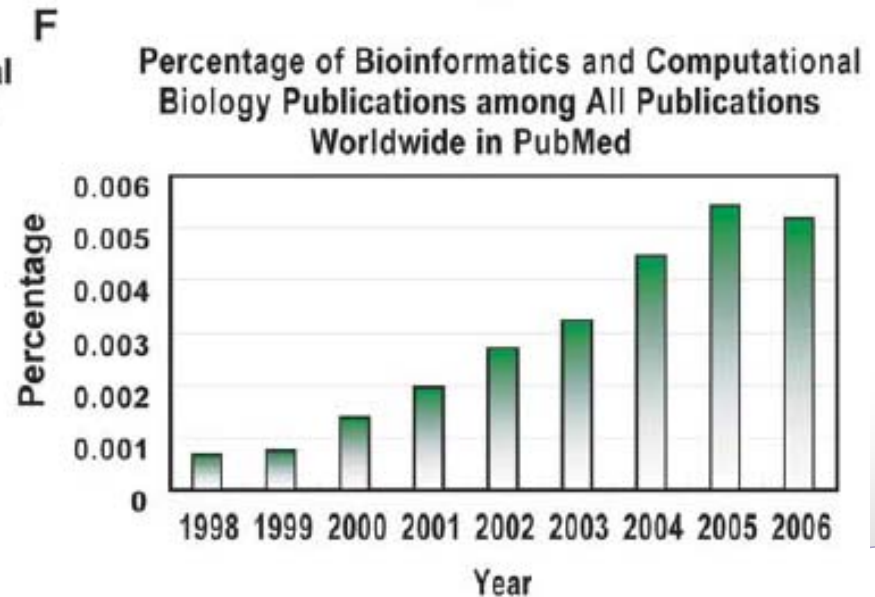
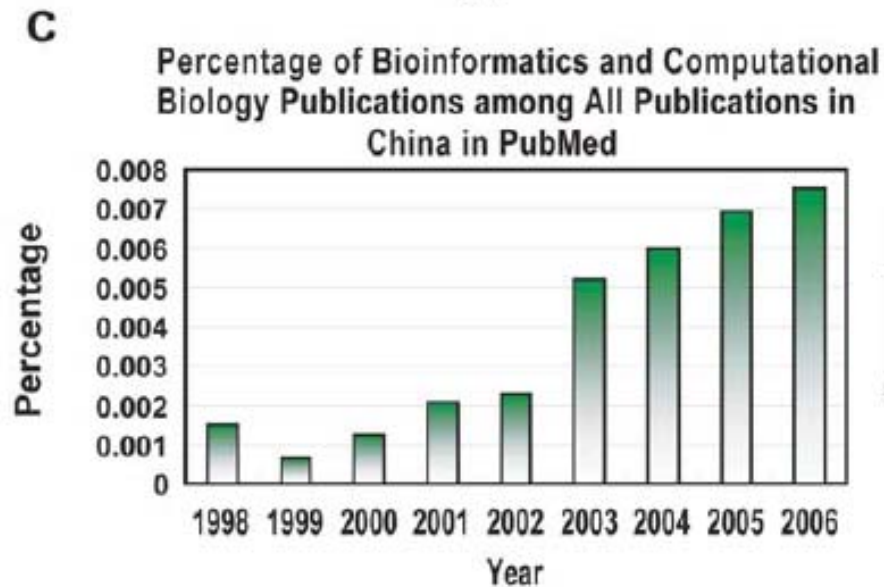
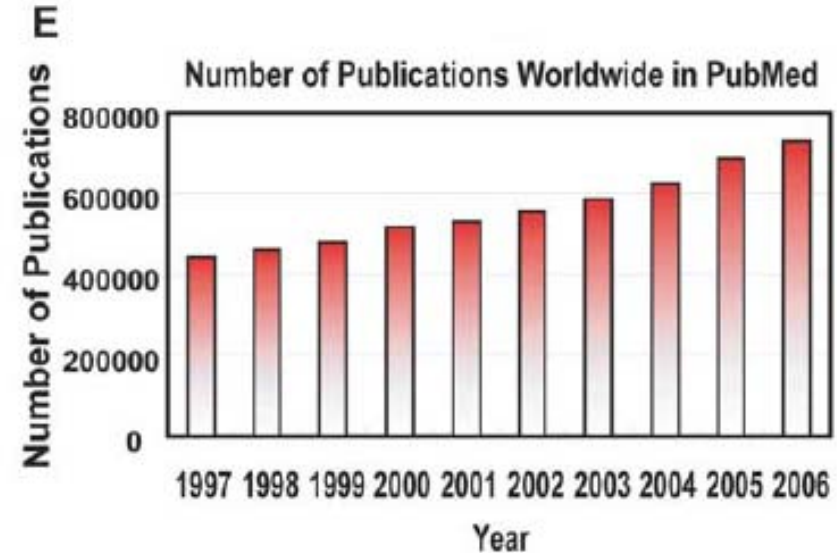
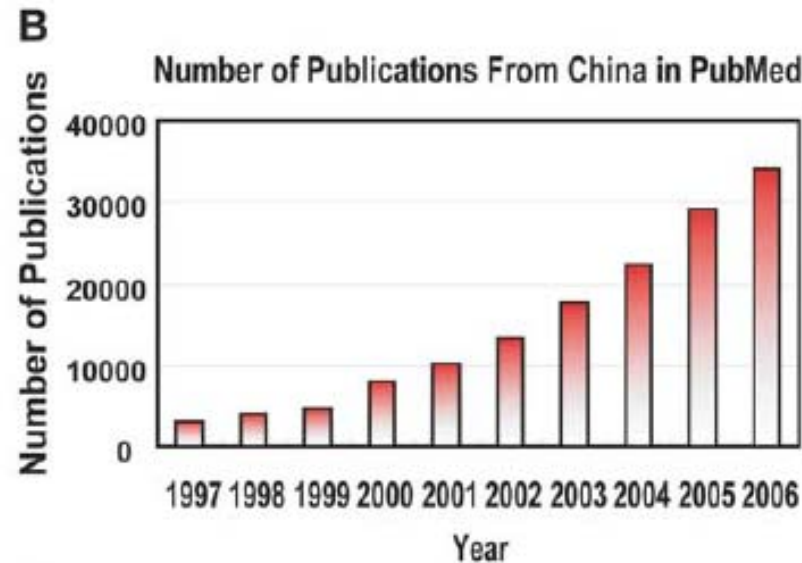
Perspective

Bioinformatics in China: A Personal Perspective

Liping Wei^{1*}, Jun Yu^{2*}



Bioinformatics in China



相关知识储备



- ❑ 生物学背景：e.g., 细胞生物学、分子生物学、发育生物学、分子神经生物学、生物化学，...
- ❑ 机器学习/深度学习/生成式AI
- ❑ 计算能力/编程能力：Perl/Python, R, PHP+MySQL, JAVA...
- ❑ 分子进化理论：MP, NJ, ML...
- ❑ 生物统计学
- ❑ 结构生物学

生物信息学的相关杂志



生物信息学相关期刊名称

网址

Bioinformatics	http://bioinformatics.oxfordjournals.org/
BMC Bioinformatics	http://www.biomedcentral.com/bmcbioinformatics/
Genome Biology	http://genomebiology.com/
Genome Research	http://www.genome.org/
Nucleic Acids Research	http://nar.oxfordjournals.org/
Briefings in Bioinformatics	http://www.henrystewart.com/briefings_in_bioinformatics/
FEBS letters	http://www.febsletters.org/
Biochemical and Biophysical Research Communications	http://www.sciencedirect.com/science/journal/0006291X
Molecular Systems Biology	http://www.nature.com/msb/index.html
Molecular Biology and Evolution	http://mbe.oxfordjournals.org/
PLoS Computational Biology	http://www.ploscompbiol.org/
PLoS ONE	http://www.plosone.org/
Protein Science	http://www.proteinscience.org/
Proteins	http://www3.interscience.wiley.com/cgi-bin/jhome/36176
Protein Engineering Design and Selection	http://peds.oxfordjournals.org/

GPB: Bioinformatics Commons



Volume: 16, Issue: 4



Volume: 16, Issue: 5

Preface

Bioinformatics Commons: The Cornerstone of Life and Health Sciences

Zhang Zhang, Yu Xue, Fangqing Zhao

[View abstract](#)

Page 223-225

Download 196

Database

CIRCpedia v2: An Updated Database for Comprehensive Circular RNA Annotation and Expression Comparison

Rui Dong, Xu-Kai Ma, Guo-Wei Li, Li Yang

[View abstract](#)

Page 226-233

Editorial

A Scientist Guerilla Fighter in the Frontiers of Bioinformatics—In Memory of Bailin Hao

Jun Yu

[View abstract](#)

Page 307-309

Download 39

Original Research

Polyphyly in 16S rRNA-based LVTree Versus Monophyly in Whole-genome-based CVTree

Guanghong Zuo, Ji Qi, Bailin Hao

[View abstract](#)

Page 310-319

如何评价生物信息学研究的水平？



- ❑ 0级 (Level 0): 为建模、而建模 (modeling for modeling's sake)
- ❑ 1级 (Level 1, 菜鸟级): 给数据、能分析
- ❑ 2级 (Level 2, 肉鸟级): 想新招、玩数据
- ❑ 3级 (Level 3, 顶级): 玩数据、作发现
- ❑ X级 (Level X, 神级): 玩科学、讲政治



刘小乐教授
哈佛大学

合适的研究体系、好的工具、百折不挠的毅力！

Oct
11
2014

Levels of Bioinformatics Research

Uncategorized

Add comments

<http://www.longwoodgenomics.org/2014/10/11/levels-of-bioinformatics-research/>

Homolog.us的评价体系



- ❑ Layer 1 - Using web to analyze biological data
- ❑ Layer 2 - Ability to install and run new programs
- ❑ Layer 3 - Writing own scripts for analysis in PERL, python or R
- ❑ Layer 4 - High level coding in C/C++/Java for implementing existing algorithms or modifying existing codes for new functionality
- ❑ Layer 5 - **Thinking mathematically**, developing own algorithms and implementing in C/C++/Java

Layer 1-4 = Level 1; Layer 5 = Level 2

A beginner's guide to
bioinformatics - part I

<http://www.homolog.us/blogs/blog/2011/07/22/a-beginners-guide-to-bioinformatics-part-i/>

A beginner's guide to
bioinformatics - part II

<http://www.homolog.us/blogs/blog/2011/07/22/a-beginners-guide-to-bioinformatics-part-ii/>

生物信息学要不要做实验？



- ❑ 生物信息学家应该做实验
- ❑ 最新的技术, **Computational biologists should stay at the cutting edge of technologies**
- ❑ 高通量
- ❑ 关注新技术产生的数据的质量
- ❑ 要回答具体的生物学问题
- ❑ 实验不能太难, 也不能太贵