第一章　生物信息学概述

陈　铭

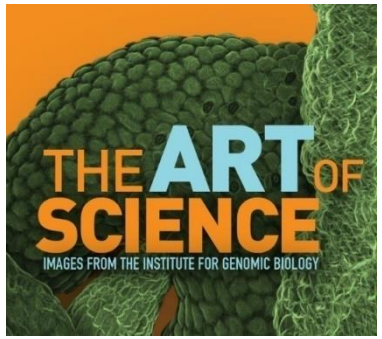
浙江大学

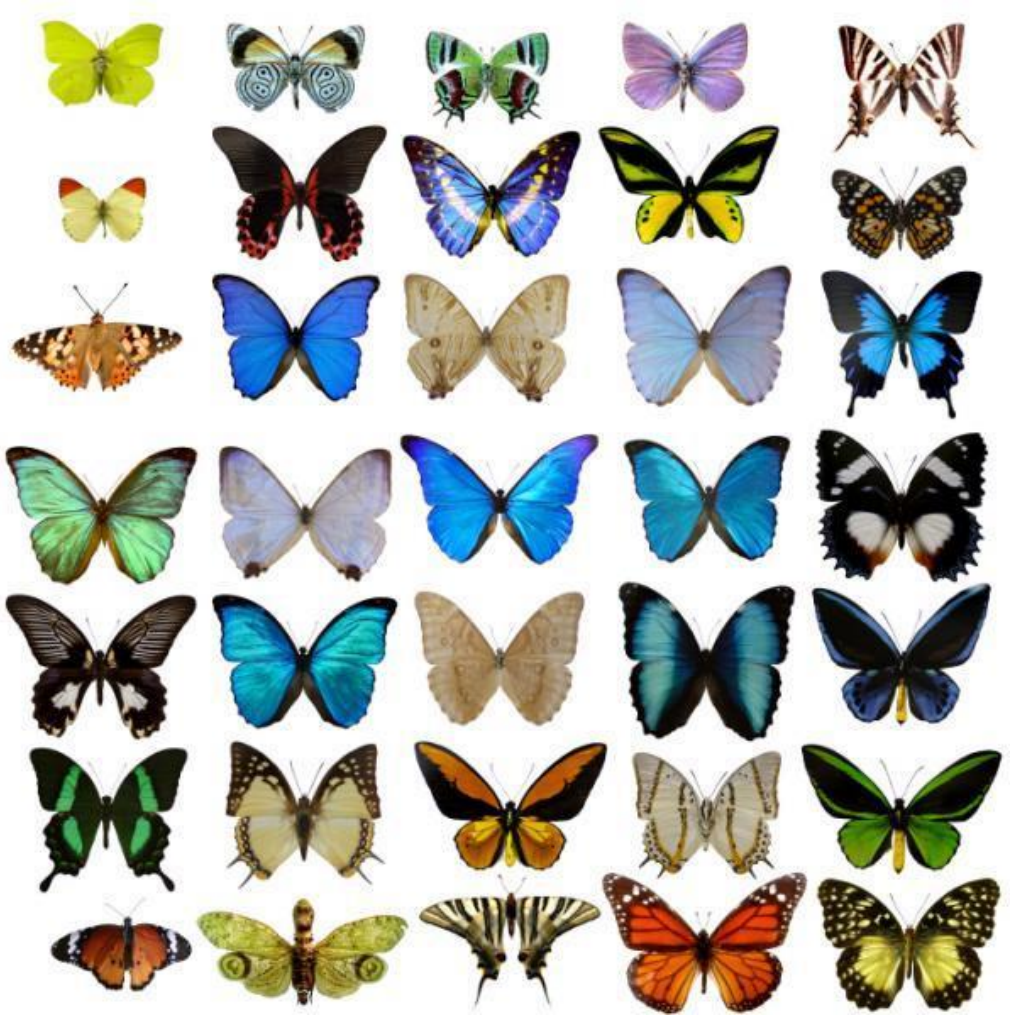
本章内容

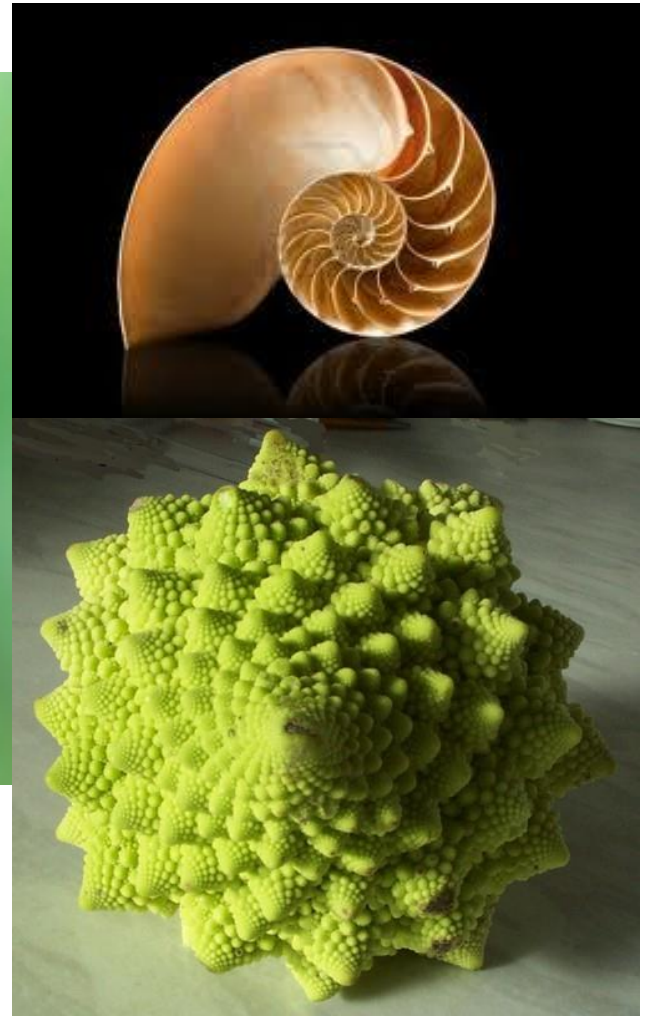
1. 生物信息学的基本概念与学科发展
2. 生物信息学的发展历史与趋势
3. 生物信息学的研究领域与内容
4. 生物信息学的算法基础
5. 生物信息学的机遇与挑战
6. 生物信息学的教育教学

第一节　生物信息学的基本概念与学科发展

生物信息学：生物学探究



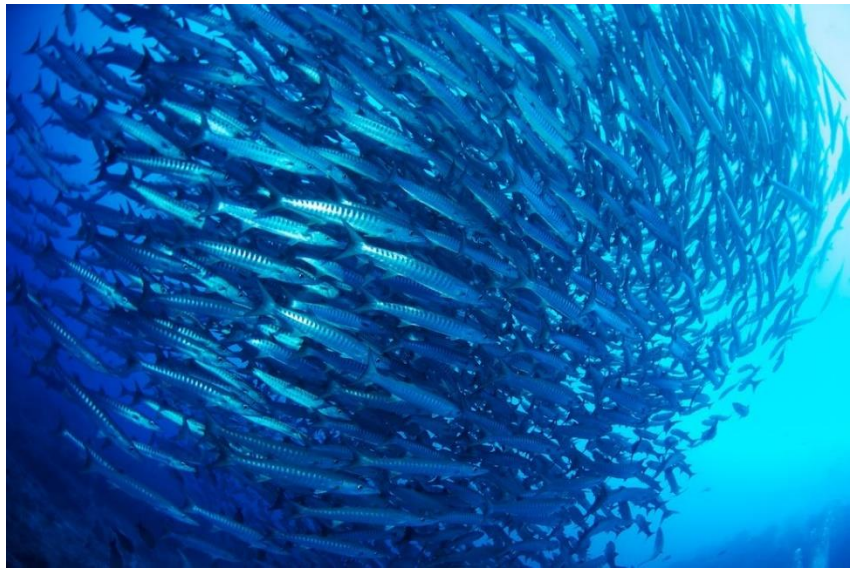




ScienceDaily (June 19, 2008)

生命科学中的patterns

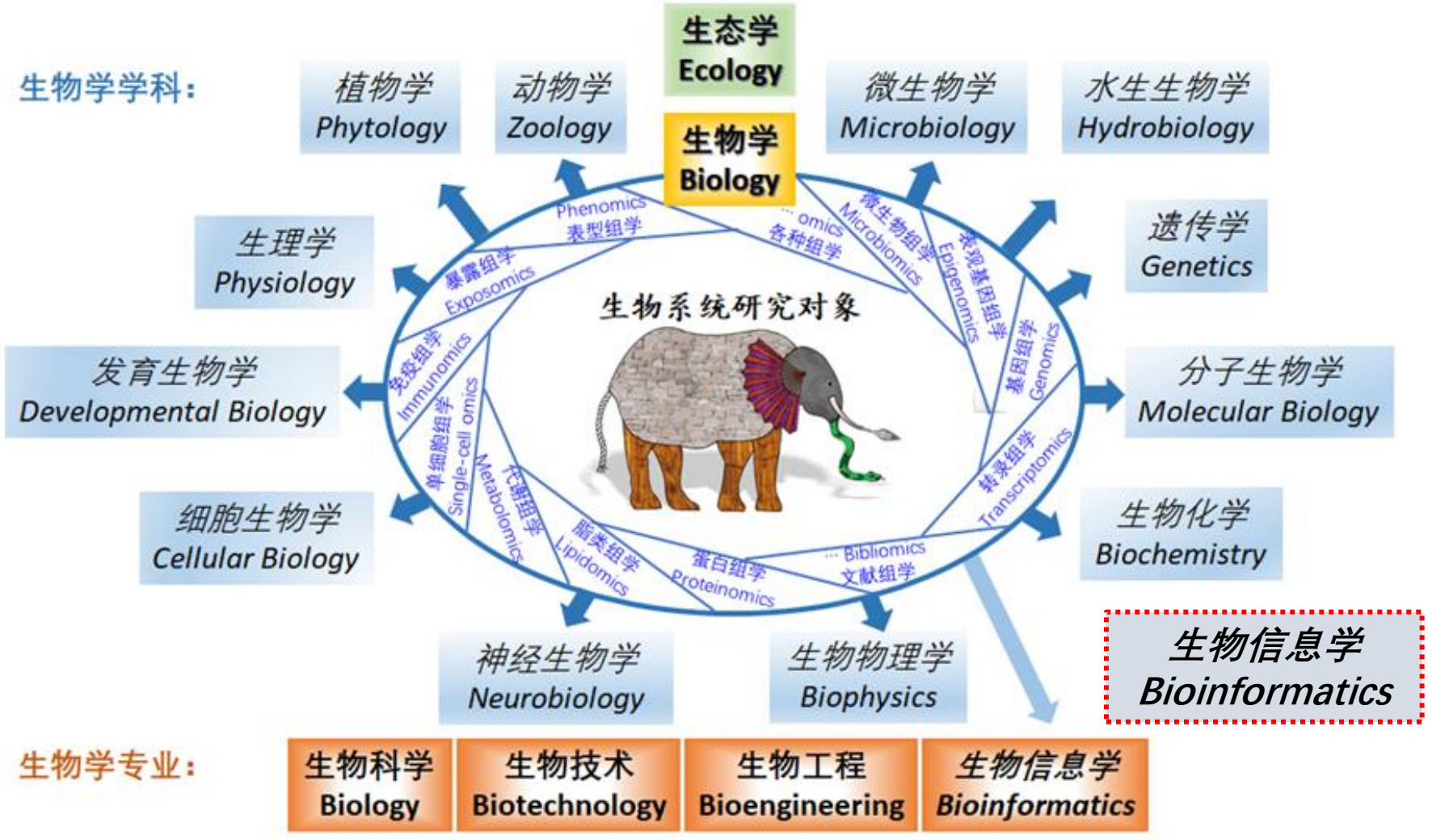
生物信息学：生物学探究



定量生物学！

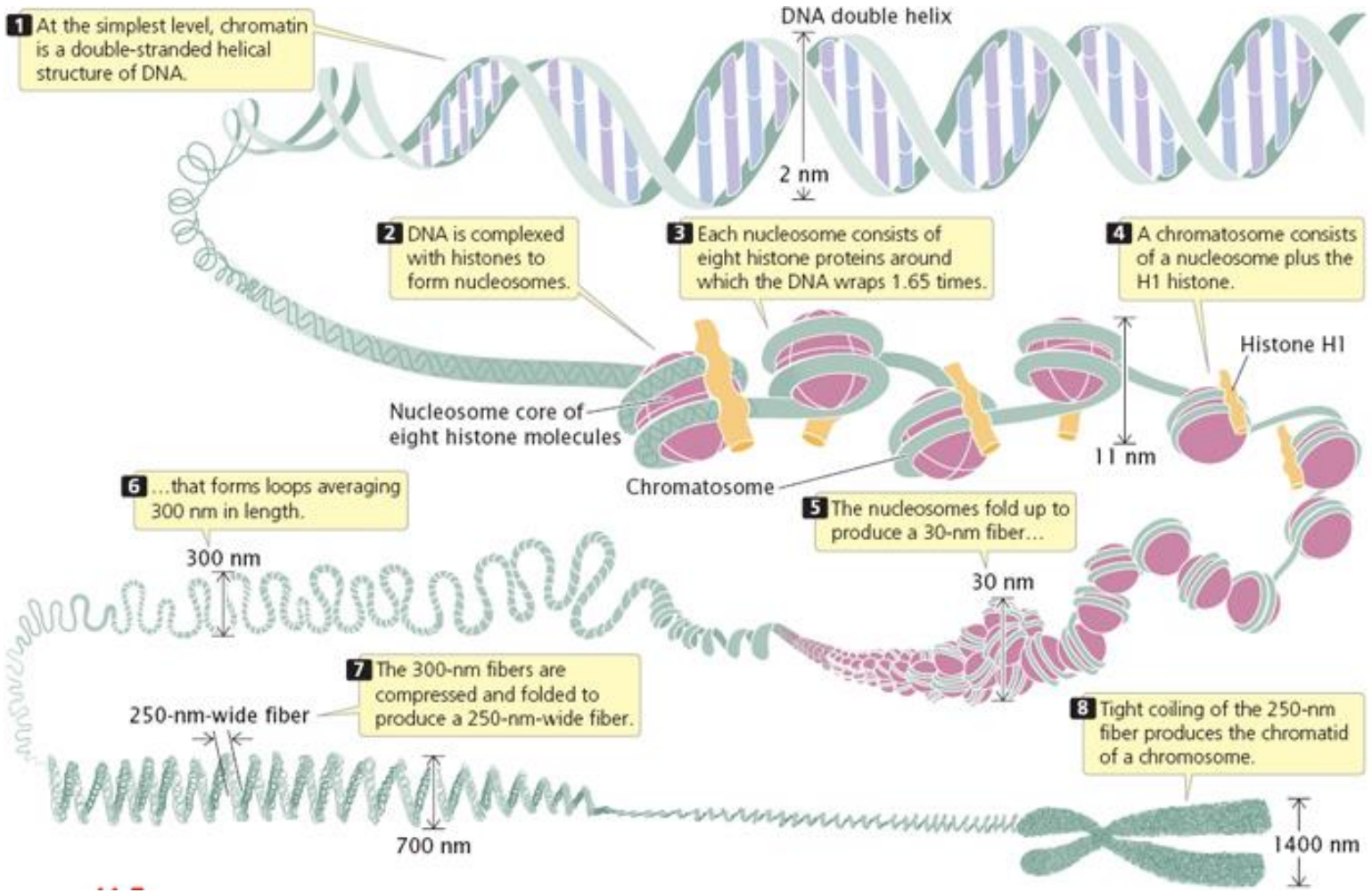
生物信息学的学科与专业地位

从生物学学科领域的角度，生物信息学涉及生命科学以及其他相关研究的各个领域。



- 生物化学与分子生物学
- 细胞生物学
- 遗传与发育生物学
- 生理学
- 微生物学
- 神经生物学
- 生物物理学
- 生态学
- 医学
- 药学
- 动植物学

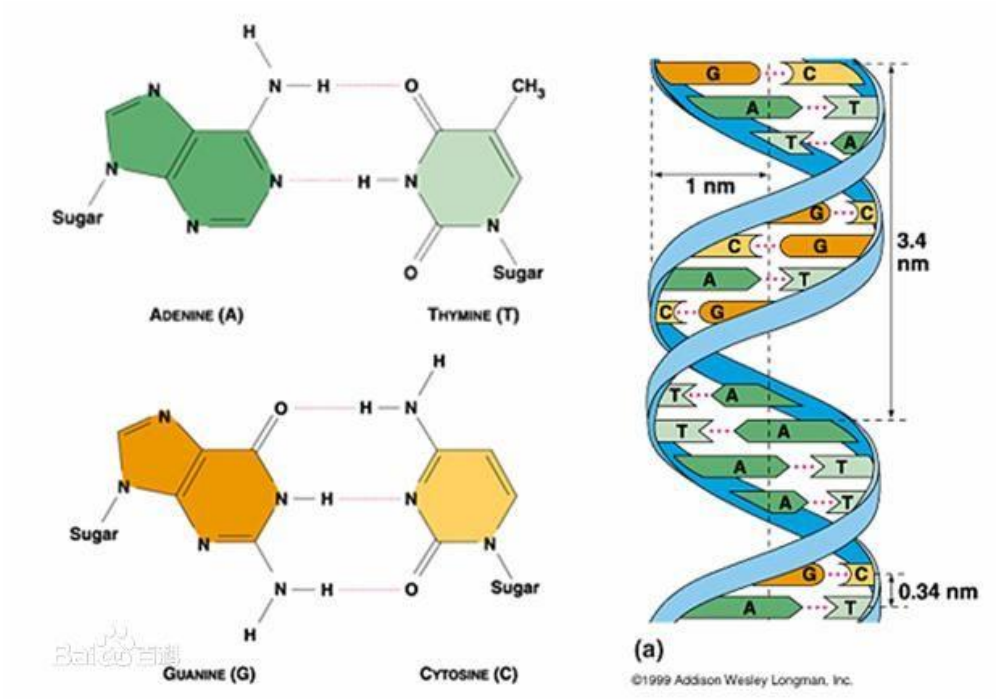
生物信息学：生物学探究



http://www.whfreeman.com/

生物信息学：生物学探究

生物编码：为什么是ATCG？它们的结构和功能？



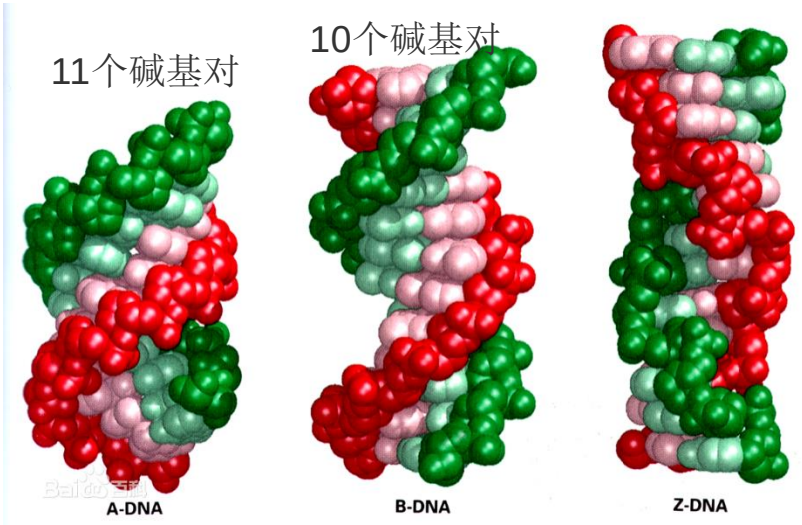
DNA（Deoxyribonucleic acid）脱氧核糖核酸
RNA（Ribonucleic acid）核糖核酸
PNA（Peptide nucleic acids）肽核酸

甲基胞嘧啶（mC）

甲基腺嘌呤（mA）

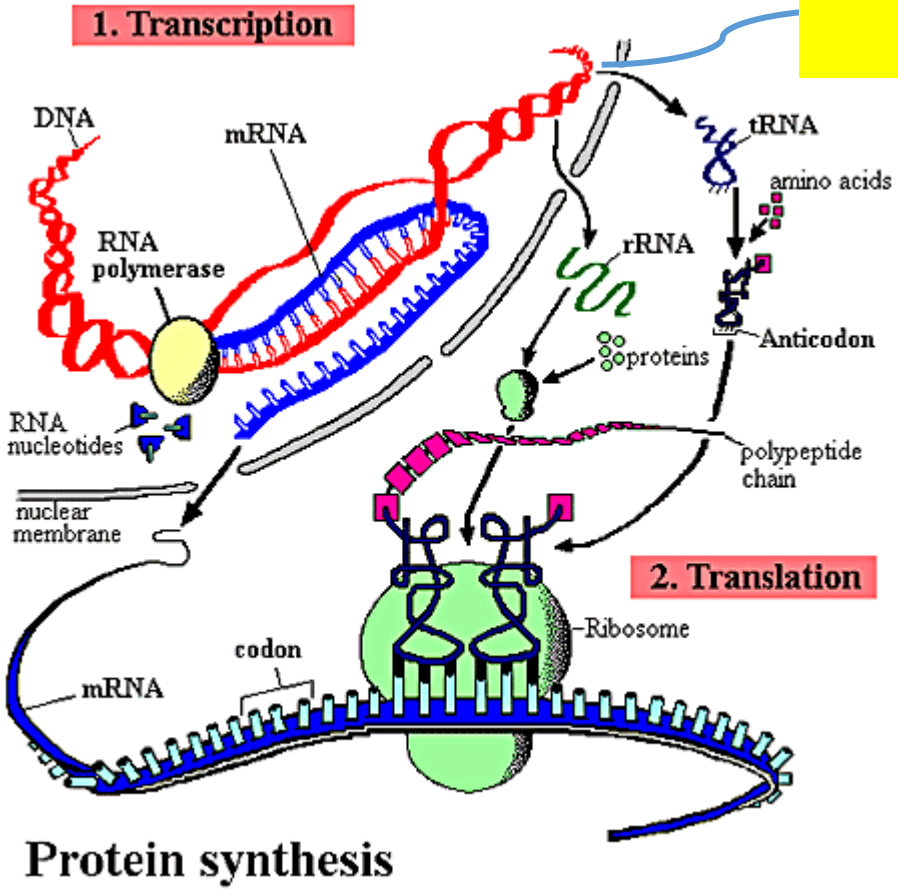
5-胞嘧啶甲酰（5-formylcytosine），
5-胞嘧啶羧基（5-carboxylcytosine）

溶液中的DNA分子比B-DNA分子的螺旋程度更高，平均每螺周有10.5个碱基对。



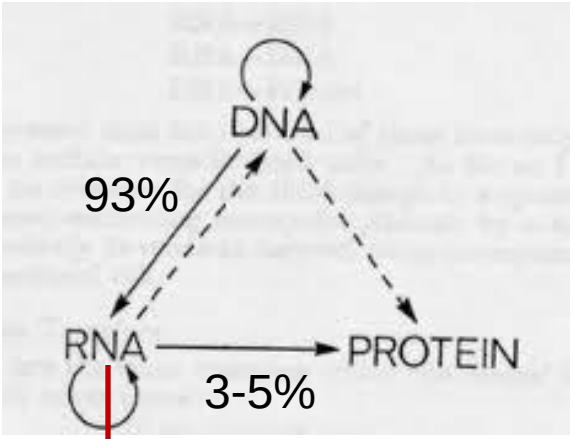
生物信息学：生物学探究

Non-coding RNAs
非编码RNA



Protein synthesis

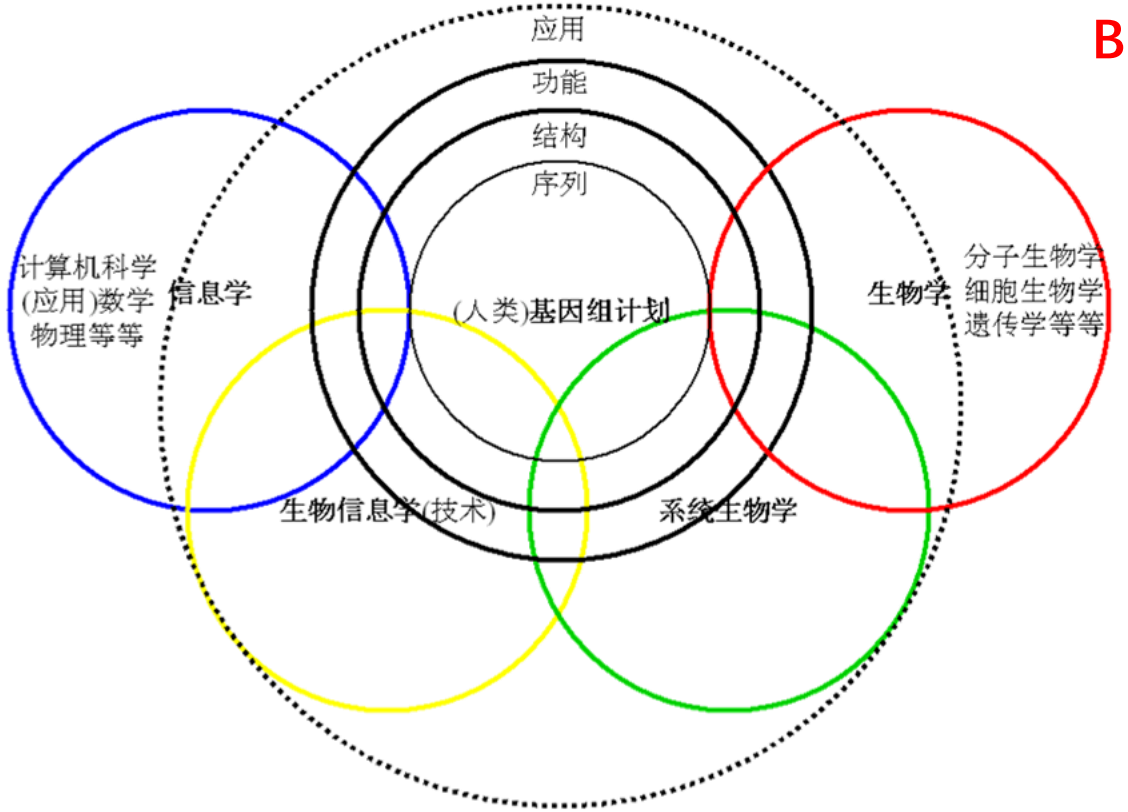
Francis Crick, 1958; *Nature* **227** , 1970



ncRNA (75% plants)

生物信息学的定义与定位

生物信息学是生物科学与信息科学的交叉学科，是利用计算机科学（信息学）的技术手段来研究生物学的数据，如对生物数据进行获取(retrival)，存储(storage)，传输(transfer)，计算(manuipulation)，分析(analysis)，模拟(simulation)，预测(prediction)等等的一门新兴学科，是21世纪科学发展的热点之一。



Biology + Informatics = Bioinformatics

But

1 + 1 =/= 2

生物信息学定位

- 从生物中来，到生物中去
- 各领域方向个性数据分析
- 各领域方向的共性方法与技术
- 各组学的整合以把握全局的特征进而发现新规律

第二节　生物信息学的历史与趋势

生物信息学历史发展

生物信息学发展的各个时期

博物生物学时期（1859-1952）

- **1859：**达尔文《物种起源》出版
- **1865：**孟德尔遗传定律
- **1869：**首次分离得到DNA
- **1879：**弗莱明观察到有丝分裂
- **1900：**孟德尔遗传规律的重新发现
- **1902：**疾病可以有序遗传、遗传的染色体理论
- **1909：**词语“gene”的出现
- **1911：**染色体理论在果蝇中得到验证
- **1941：**一个基因编码一个酶
- **1943：**DNA的X射线衍射
- **1944：**DNA可以改造细胞的特性；跳跃基因的发现；薛定谔发表《生命是什么》
- **1952：**DNA构成基因

- **1998：**中国人类基因组研究北方中心（北京）和南方中心（上海）成立；亚太生物信息学网络（APBioNet）成立；美国Celera遗传公司成立，开展人类基因组测序；线虫基因组完成；CABIOS期刊更名为Bioinformatics
- **1999：**“北京华大基因研究中心”的成立；中国获准加入HGP，承担1%测序工作；人类22号染色体序列完成
- **2000：**中国科学院上海生命科学研究院生物信息中心(SIBI)成立；我国完成基因组计划的1%；德、日等国科学家宣布基本完成人体第21对染色体的测序工作；果蝇基因组完成
- **2001：**首届全国生物信息学会议（CCB）举行；中国完成籼稻基因组工作框架图；美、日、德、法、英、中6国科学家和美国Celera公司联合公布人类基因组图谱及初步分析结果

分子生物学发展时期（1953-1975）

- **1975：**DNA测序工作的开启
- **1973：**第一个动物基因被克隆
- **1972：**伯格发现第一个重组DNA
- **1968：**发现第一个限制性酶
- **1965：**中国人工合成牛胰岛素结晶
- **1961：**mRNA将信息从细胞核内传递到细胞质
- **1959：**染色体异常致病被发现
- **1958：**梅塞尔森和斯塔尔证实DNA的半保留复制；克拉克提出中心法则
- **1956：**血红蛋白的一个氨基酸改变可以导致镰状细胞性贫血
- **1955：**人类的46条染色体的确定、DNA聚合酶的发现、第一个蛋白质序列(牛胰岛素)被测定
- **1953：**克拉克、沃森、富兰克林和威尔金斯等人发现DNA双螺旋结构

基因组学时期（1976-2001）

- **1976：**第一个遗传工程公司成立
- **1977：**桑格研究小组完成了第一个全基因组噬菌体Φ-X174的测序；内含子的发现
- **1981：**中国实现酵母内氨酸转移核糖核酸的人工合成
- **1985：**穆利斯创立PCR技术；生物信息学专业期刊（CABIOS）创刊；德国生物信息学会议（GCB）举行
- **1986：**中国开始实施“863计划”；日本核酸序列数据库DDBJ诞生；蛋白质数据库SWISS-PROT建立
- **1988：**美国国家生物技术信息中心（NCBI）成立
- **1990：**国际人类基因组计划（HGP）启动
- **1993：**中国人类基因组计划（CHGP）正式启动；HGP新5年计划；欧洲生物信息学研究所（EBI）获准成立；第一届ISMB国际会议美国国家医学图书馆（NLM）举行
- **1994：**威尔金斯提出蛋白质组（Proteome）的概念、细菌基因组计划
- **1995：**人类基因组物理图谱完成；日本信息生物学中心（CIB）成立；流感嗜血杆菌（Haemophilus influenzae Rd）全基因组序列发布
- **1996：**北京大学蛋白质工程和植物遗传学工程国家实验室加入欧洲分子生物学网络（EMBnet）；Affymetrix生产商用DNA芯片
- **1997：**北京大学生物信息学中心（CBI）成立；中国科学院召开“DNA芯片的现状与未来”和“生物信息学”香山会议；大肠杆菌基因组完成

后基因组时期（2002-2012）

- **2012：**CRISPR/Cas基因编辑系统的应用；ENCODE发表阶段性研究成果；英国启动十万人基因组计划；欧洲植物表型网络、国际植物表型联盟成立
- **2010：**外显子测序；三代测序技术的出现
- **2009：**黄瓜、高粱和两个玉米品种的基因组测序；汤富酬发表单细胞转录组测序技术
- **2008：**千人基因组测序计划启动；拟南芥1001株系测序启动
- **2007：**世界首份“个人版”基因图谱完成；人体微生物组计划（HMP）启动；澳大利亚成立第一个植物表型研究中心（APPF）
- **2006：**中国科学家参与人类3号染色体的DNA完成序列与详尽分析；癌症基因组图谱计划（TCGA）启动
- **2005：**大猩猩和狗基因组发表；HapMap项目完成；二代测序技术的出现
- **2004：**大鼠和鸡基因组草图完成
- **2003：**中国科学院北京基因组研究所（现国家生物信息中心）成立；HGP完成；DNA元素百科全书（ENCODE）计划启动；人类表观基因组计划启动
- **2002：**小鼠基因组完成；人类基因组单体型图（HapMap）计划启动

大数据与人工智能时期（2013-）

- **2013：**人类脑计划启动
- **2014：**HMP第二期计划（iHMP/HMP2）启动
- **2015：**人类表观基因组计划发布阶段性表观基因组图谱；精准医疗的提出
- **2016：**深圳国家基因库 CNGB 正式运营；中科院绘制人类脑网络组图谱；空间转录组学技术发展
- **2017：**“中国十万人基因组计划”启动；人类表型组学计划；人类细胞图谱计划启动；首次合成包含两种人工碱基的生命体
- **2018：**单细胞水平细胞谱系追踪技术；世界首例单条染色体真核细胞出现；小麦基因组图谱历经13年绘制完成；英国发起五百万人基因组计划
- **2019：**国家基因组科学数据中心（NGDC）成立；北京基因组所加挂“国家生物信息中心”；DNA 显微镜研制成功；新型人造 DNA 结构信息密度可加倍；HMP2发表阶段性研究成果
- **2020：**首张人类细胞图谱公布；AlphaFold 2在在蛋白质结构预测大赛CASP 14成绩优异；TCGA发表阶段性研究成果
- **2021：**成功利用AlphaFold2破译整个人类蛋白质组结构（98.5%的人类蛋白质）
- **2022：**OpenAI开发的人工智能聊天机器人程序ChatGPT正式推出；T2T Consortium填补人类基因组计划最后8%的空缺
- **2023：**科技部启动“人工智能驱动的科学研究”专项部署工作；首次完成对人类Y染色体的完整测序；AlphaFold AI准确预测几乎所有已知蛋白质结构；人工智能在药物发现和开发中的应用
- **2024：**AlphaFold 3发布

生物技术与信息技术的协同发展

1950-1970

1970-1980

生物信息萌芽期

- 第一代测序技术 Sanger 法（链终止法）：Sanger F（1977）
- Bioinformatics 概念首次提出：Hogeweg P 和 Hesper B（1978）
- 第一款DNA序列分析软件 Staden：Bonfield J 和 Staden R（1979）

1980-1990

生物信息起源期

- 分子进化理论：Pauling L（1962）
- 第一个生物序列数据库 Atlas of Protein Sequences：Dayhoff M (1965)
- 第一个生物信息学软件 COMPROTEIN：Dayhoff M 和 Ledley RS（1958-1962）
- Needleman-Wunsch 序列比对算法：Needleman SB 和 Wunsch CD（1970）
- PAM矩阵： Dayhoff M及其同事（1978）

生物信息初创期

- GenBank 释放（1982）
- Bioinformatics 杂志前身被创建（1985）
- NCBI 成立（1988）
- BLAST 算法（1990）

1990-2000

生物信息发展期：基因组学兴起

- Sanger 中心成立（1993）
- EMBL 核酸数据库（1993）
- Pubmed 数据库（1997）
- 人类基因组计划（HGP）
- 模式物种基因组：酵母（1996）、果蝇（1999）、线虫（1998）、拟南芥（2000）
- 基因组拼接软件

2000-2010

生物信息成熟期：高通量时代

- 二代测序技术（~2005）
  - Roche/454、Illumina Solexa、ABI SOLiD
- 功能基因组计划
  - HapMap 计划（2002）
  - ENCODE 计划（2003）
  - 表观基因组计划（2003）
  - 癌症基因组计划（TCGA，2006）
  - 宏基因组计划（HMP，2007）
  - 1000基因组计划（2008）

2010-至今

生物信息黄金期：大数据时代

- 三代测序技术（2010）
  - PacBio（2010）、ONT（2014）
- 生物大数据科学计划
  - 英国10万人基因组计划（2012）
  - 精准医疗（2015）、人类表型组计划（2017）
- 国家基因组科学数据中心成立（2016）
- AlphaFold（2018）、AlphaFold3（2024）
- GeneFormer、scGPT、scFoundation（2023

整合生物学
系统生物学

分子生物学的发展

生物技术与信息技术的协同发展

1950-1970 生物信息起源期

- Fortran：Backus J（1957）
- LISP：McCarthy J（1958）
- BASIC：Gates B等（1964）
- PASCAL：Wirth N（1970）

1970-1980 生物信息萌芽期

- C：Ritchie D（1972）
- SQL：Boyce R 和 Chamberlain D（1972）
- Smalltalk：Kay A, Goldberg A 和 Ingalls D（1972）

1980-1990 生物信息初创期

- R：Gentleman R 和 Ihaka R（~1980）
- C++：Stroustrup B（1983）
- Objective C：Cox B 和 Love T（1983）
- GNU 协议：Stallman R（1985）
- Perl：Wall L（1987）
- Python：Rossum G（1989）
- WWW 技术：Berners-Lee T（~1990）

1990-2000 生物信息发展期：基因组学兴起

- Linux： Torvalds L（1991）
- Visual Basic：用户图形界面 GUI（1991）
- Ruby：Matsumoto Y（1993）
- Java：Gosling J（1995）
- JavaScript：Eich B（1995）
- PHP：Lerdorf R（1995）
- C#：Microsoft（2000）

2000-2010 生物信息成熟期：高通量时代

- Scala：Odersky M（2003）
- 云计算（Cloud Computing）
  - 亚马逊 AWS（2006）
  - Google App Engine（2008）
  - 微软 Azure（2009）
- 区块链（Blockchain，2008）
- GO：Google（2009）

Robert Gentleman

2010-至今 生物信息黄金期：大数据时代

- Swift：Apple（2014）
- 深度学习提出（2012）
  - CNN（2012）
  - 深度强化学习（2014）
  - Transformer（2017）
- 深度学习主流框架
  - TensorFlow（2015）
  - PyTorch（2016）
- ChatGPT 3.0（2020）
- 人工智能时代崛起

整合生物学
系统生物学

计算机科学的发展

Linus

生物信息学的历史

Frederick Sanger将**胰岛素的氨基酸序列**完整**地定序出来，同**时证明**蛋白**质具有明确构造。

1955

1953

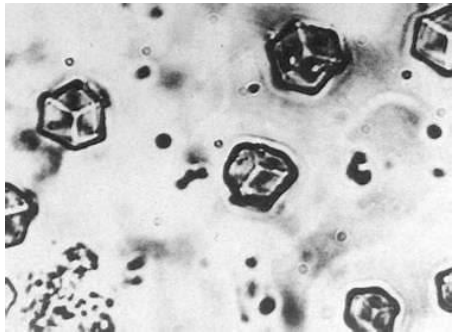
Francis Crick，James Watson和Maurice Wilkins 发现DNA**双螺旋**结构。

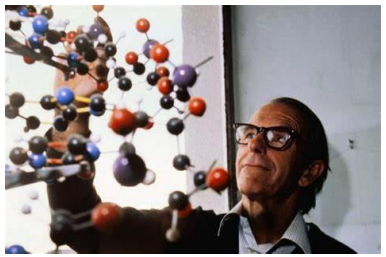
1958

中国科学院上海生物化学研究所提出人工合成胰岛素。同年年底该项目被**列入1959年国家科研**计划，并获得**国家机密研究**计划代号“601”，**意**为“六十年代第一大任**务”。

1965

中国团队在世界上第一次人工全**合成了与天然牛胰**岛素分子化学结构相同并具有完整生物活性的**蛋白**质，且生物活性达到天然牛**胰**岛素的80%。





1975

Sanger发展出一种称为链终止法（chain termination method）**的技**术来测定DNA**序列**，这种方法也称作“**双去氧**终止法”**或是“桑格法”**。

生物信息学的历史

国际人类基因组计划（HGP）启动。其宗旨在于测定组成人类染色体（指单倍体）中所包含的30亿个碱基对组成的核苷酸序列。

1990

罗静初教授

1993

由国家自然科学基金委员会生命科学部组织的以谈家桢教授为组长的专家组，在沪论证并通过了《中华民族基因组中若干位点基因结构的研究》重大项目，标志着我国HGP正式启动。

北京大学生物信息中心（CBI）成立，EMBnet的中国国家节点。当时国内数据库种类最多，数据量最大的生物信息站点，为国内外用户提供了多项生物信息服务。

1997

1998/1999

1998年，“国家基因组北方研究中心”和“国家基因组南方研究中心”成立。
1999年，“北京华大基因研究中心”的成立。我国参加了HGP，并承担了人类3号染色体短臂3000万碱基（约占人类基因组全部碱基序列的1%）的测序任务。

中、美、日、德、法、英等6国科学家和美国Celera公司在克雷格·文特尔的带领下联合发表“人类基因组计划”的结果——人类基因组草图及初步分析。

2001

杨焕明院士

国际人类基因组计划

生物信息学的历史

由美国、中国和德国等国科学家组成的合作小组，在《自然》杂志发表**人类3号染色体的DNA完成序列与**详尽分析。这是国际“人类基因组计划”发表的已完成分析的最大染色体之一。

2006

国家基因组科学数据中心（National Genomics Data Center，简称NGDC）经科技部、财政部通知公布，由中国科学院北京基因组研究所（国家生物信息中心）作为依托单位，联合中国科学院生物物理研究所和中国科学院上海营养与健康研究所共同建设。

2019

2003

中国科学院北京基因组研究所（现国家生物信息中心）成立，重点研究基因组结构、变异、功能及其演化规律，加强基因组学与其他学科的交叉融合，发展基因组学的新理论、新方法和新技术。

HGP完成了其主要的测序工作

2012

Jennifer Doudna和Emmanuelle Charpentier等研究人员揭示**CRISPR-Cas9系统的功能机制**。Cas9蛋白可以利用RNA分子作为引导，识别并切割特定的DNA序列。这项发现标志着CRISPR技术在精确基因编辑方面的巨大潜力。

生物信息学的历史

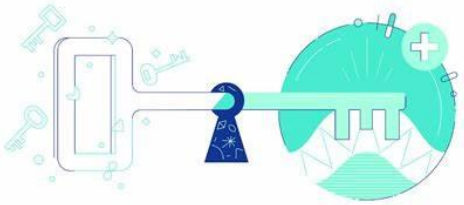
2020

AlphaFold 2在蛋白质结构预测大赛CASP 14**中**，对大**部分蛋白**质结构的预测与真实结构只差一个原子的宽度，**达到了人**类利用冷冻电子显微镜等复杂仪器观察预测**的水平**。

2021

成功利用AlphaFold 2预测98.5%**的人**类蛋白质组结构。

2022



DeepMind和EMBL-EBI在2022**年7月**联合宣布，AlphaFold**已**经预测了超过2亿个蛋白质的结**构，覆盖了地球上几乎所有已知的蛋白**质。

美国人工智能研究实验室OpenAI**新推出一种人工智能技**术驱动的自然语言处理工具——**ChatGPT**（Chat Generative Pre-trained Transformer）

2024

DeepMind发布了**AlphaFold 3**，这是该系列**的最新版本**，进一步推进了**蛋白**质结构预测的边界。

我国生物信息学的发展

中国生物信息学终身成就奖


陈润生院士


李衍达院士


郝柏林院士


罗辽复教授


张春霆院士


孙之荣教授

- 首批中国生物信息学终身成就奖得主简介
- 生信史一：顾孝诚教授与北京大学生物信息中心（罗静初）
- 生信史二：The international Human Genome Project（杨焕明）
- 生信史三：Early bioinformatics research in China（陈润生）
- 国家基因组科学数据中心（国家生物信息中心）：整合中国组学资源，打破数据孤岛

More…
101课程网站
虚拟教研室线
上资源

我国生物信息学的发展

- 中国生物信息学终身成就奖得主简介之**陈润生院士**

陈润生，1941年6月18日出生，天津人，生物信息学家，中国科学院生物物理研究所研究员、博士生导师。2007年当选为中国科学院院士、2014年当选为国际欧亚科学院院士。

科研成就：陈润生曾参加中国第一个完整基因组泉生热袍菌B4基因组序列的组装和基因标识，曾参加人类基因组1%和水稻基因组工作草图的研究；构建了收录非编码RNA及其基因的数据库NONCODE，以及收录非编码RNA与其它生物大分子相互作用的数据库NPInter。

陈润生课题组主要研究方向包括但不限于：利用生物信息学手段并结合实验室多年积累的RNA组学技术，深入开展在肿瘤发生、发展以及干细胞重编程过程中长非编码RNA的系统发现和功能机制研究；非编码RNA数据库（NONCODE）、非编码RNA与各种生物大分子（DNA、RNA及蛋白质）相互作用数据库（NPInter）以及其它专家数据库的构建与升级；RNA研究新技术、新方法；长非编码RNA翻译复杂性的理论与功能机制研究；着丝粒及组成型异染色质建成过程中长非编码RNA的功能与调控机制研究；小鼠早期胚胎发育过程中长非编码RNA的功能与调控机制研究。

- 生信史三：Early bioinformatics research in China（陈润生）

我国生物信息学的发展

- **杨焕明院士**

杨焕明，1952年10月出生于浙江乐清，基因组学家，中国科学院院士、发展中国家科学院院士、非洲科学院院士、印度国家科学院外籍院士、德国国家科学院外籍院士、美国国家科学院外籍院士，中国医学科学院教授，华大基因理事长、华大基因学院院长、东南大学生命健康高等研究院名誉院长

科研成就：杨焕明和他的团队为“国际人类基因组计划”、“国际人类单体型图计划”、“国际千人基因组计划”、“国际癌症基因组计划”等国际合作的基因组计划，以及第一个亚洲人基因组、人类泛基因组学、古人类基因组、肠道Meta基因组的研究做出了重要贡献。作为联合创始人，创建了华大基因和中国科学院北京基因组研究所。

杨焕明领导华大中心经过艰苦的拼搏，在世界上首次利用全基因组“霰弹法”策略对大型植物基因组进行测序，独立完成了超级杂交水稻父本籼稻“9311”基因组（大小约为4.6亿个碱基对）的“工作框架图”。该项目的完成，建立和完善了基因组学、生物信息学和蛋白质组学研究的多个技术平台，其水平与发达国家齐步，使中国成为继美国之后的第二个具有全面测定和分析大型全基因组能力的国家，而且从无到有地开发了重复序列识别及注释系统，并率先公布了数据库，促进了“国际联盟”的工作进程，在国际上产生了巨大而深远的影响。杨焕明及其团队所承担的人类基因组、水稻基因组以及家猪、家鸡、家蚕基因组等重大项目使中国的基因组研究得以跻身于世界前沿。杨焕明还特别关注基因组研究的社会影响和基因知识的普及。

我国生物信息学的发展

- **张泽民院士**

张泽民，1967年7月出生于河南省驻马店市，中国科学院院士，北京大学生物医学前沿创新中心（BIOPIC）研究员、主任，北京大学未来基因诊断高精尖创新中心研究员，北京大学生命科学学院教授，北大-清华生命科学联合中心高级研究员。 2023年当选为中国科学院院士。

科研成就：张泽民应用前沿基因组学和生物信息学技术来研究肿瘤免疫，干/湿实验结合来探究肿瘤微环境；应用单细胞技术研究肿瘤浸润免疫细胞的组成和功能状态，探究各类细胞的特异属性、相互关系、及在治疗中的动态变化，从而开拓克服肿瘤的新方法；开发原创性的生物信息学工具来进行组学数据的分析、整合和可视化，以辅助科学发现。

张泽民的团队应用前沿单细胞测序技术和生物信息学技术在肿瘤免疫研究中取得系列重大突破。揭示肝癌组织内耗竭性CD8+T细胞及Treg的特征表达基因，进一步描述了肝癌微环境的免疫组分和状态及肿瘤浸润免疫细胞跨组织的动态过程；开发生物信息方法STARTRAC，并发现结直肠癌肿瘤微环境和T细胞受体共同影响了肿瘤浸润CD8+ T细胞的状态转化，揭示了新的治疗靶点；结合癌症病人和小鼠肿瘤模型，发现特定的巨噬细胞和树突状细胞亚群在结直肠癌的相互作用中对调节肿瘤免疫起到关键影响。

我国生物信息学的发展


首页 - 国家基因组科学数据中心
https://ngdc.cncb.ac.cn
国家生物信息中心
China National Center for Bioinformation
Data Resources
Computing Analysis
Data Network
Standards
国家基因组科学数据中心
National Genomics Data Center
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...
Search
e.g., PRJCA000126; SAMC000385; tp53; EGFR; human; KaKs_Calculator; GenBank
国家基因组科学数据中心非编码RNA资源团队招聘启事
提交
科技计划项目数据汇交
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
数据资源
原始数据
基因组和变异
基因表达
非编码RNA
表观基因组
单细胞组学
热门资源
BioCode
生物工具软件
BioProject
生物项目库
BioSample
生物样本库
BIT
生物信息在线分析平台
Database Commons
生物数据库目录
GEN
基因表达数据库
GenBase
基因序列库
GSA
组学原始数据归档库
GSA-Human
人类组学原始数据归档库


中心
息前沿交叉研究与转化应用
检索
34 个
软件工具
45 家
合作伙伴

生物信息学的研究趋势与进展

10K、
100K、1M

Pan-omics

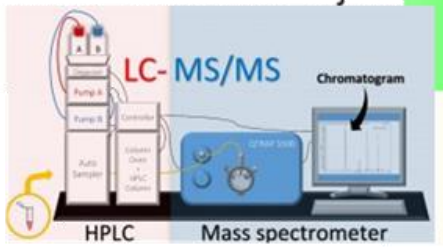
Transcriptomics

RNA-seq,
more

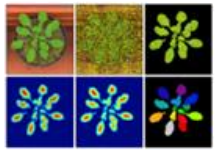
Single-cell omics

HPP Human Proteome Project

Proteomics



Metabolomics



Phenomics

AI application

Big data !

Multi-model, Multi-level

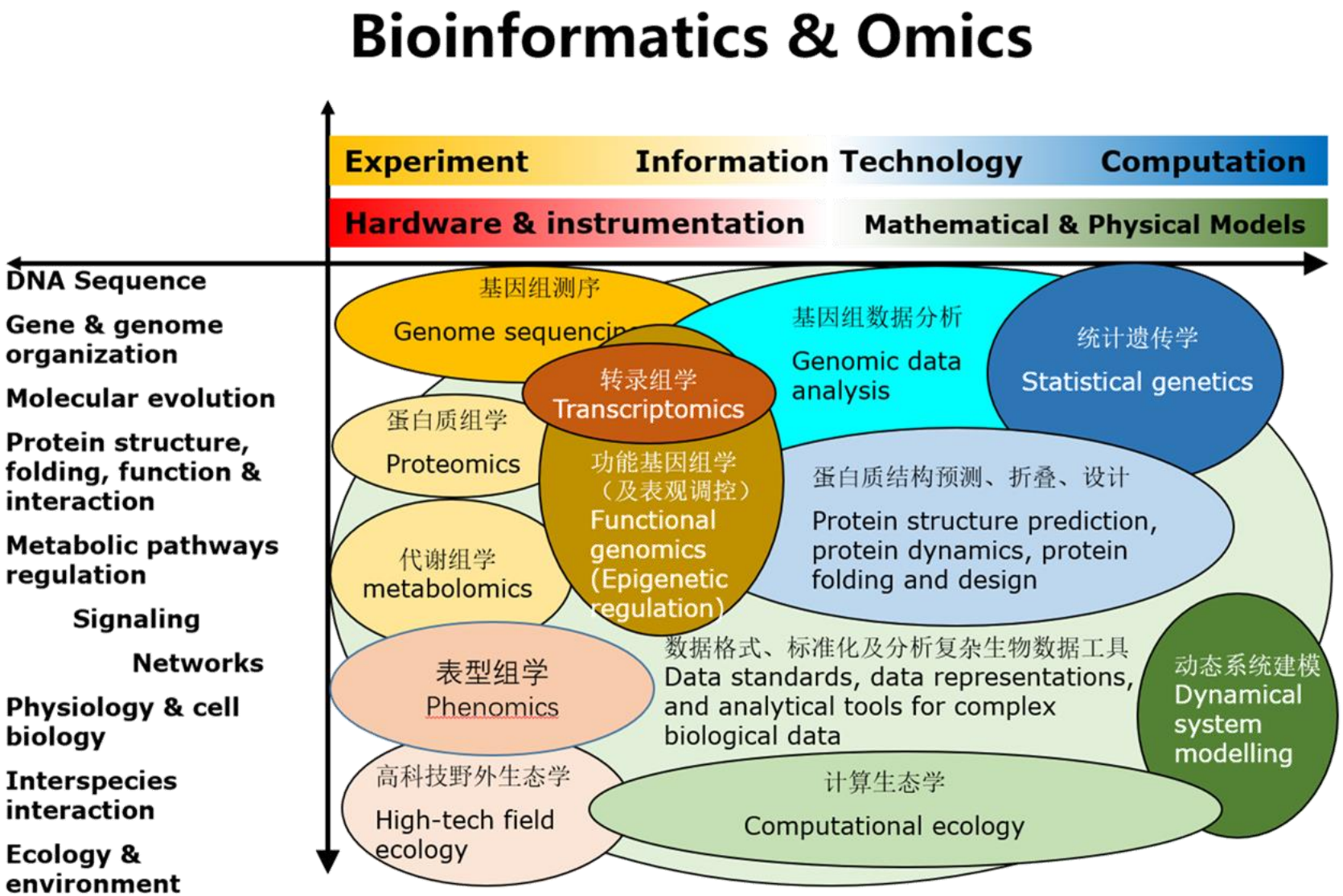
Systems Biology

1990 1995 2000 2005 2010 2015 2020 2025

GWAS、AlphaFold、ChatGPT带给我们的思考!

第三节　生物信息学的研究领域和内容

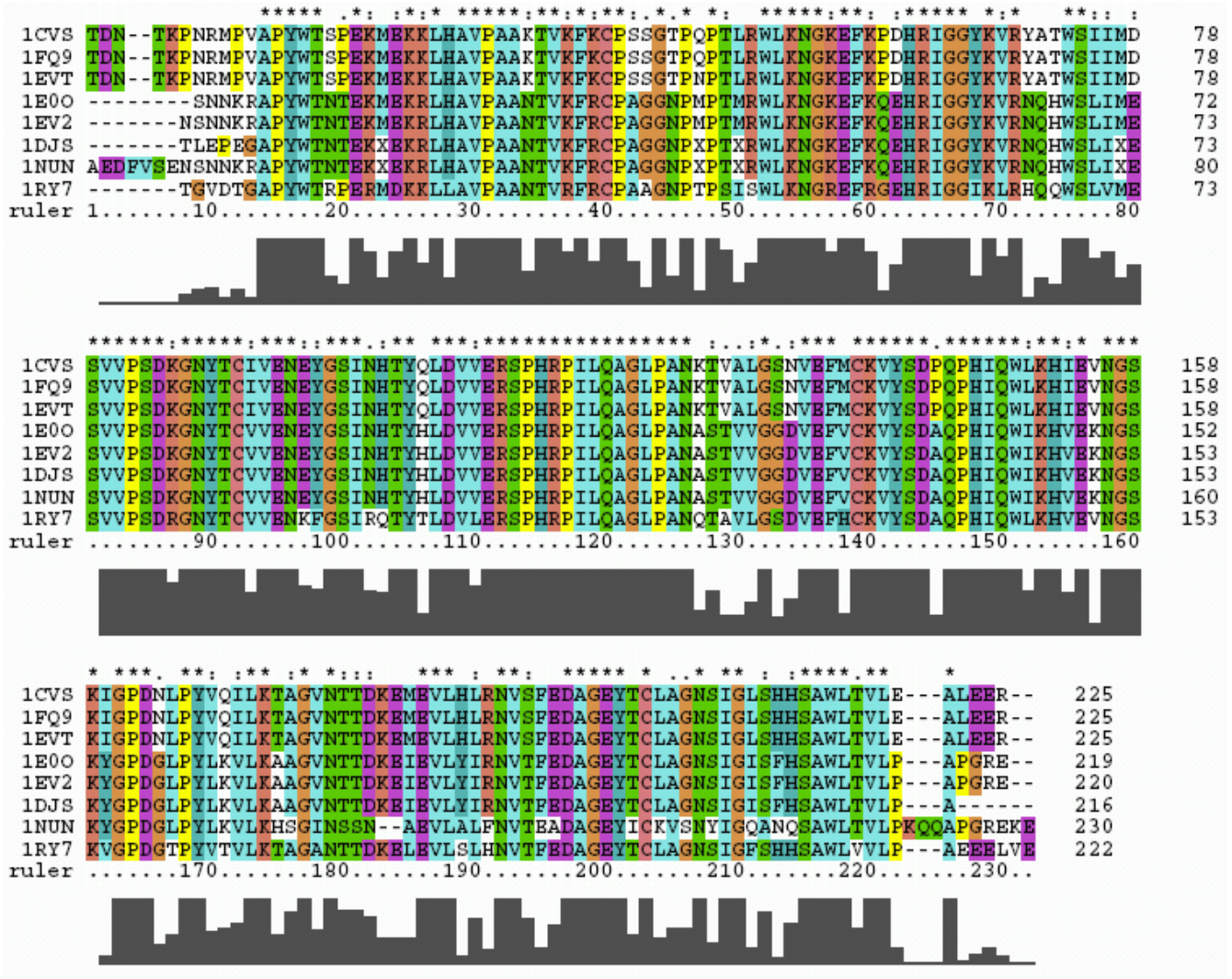
生物信息学的研究内容



陈铭，后基因组时代的生物信息学，2004　　陈铭：整合生物信息学（2006）

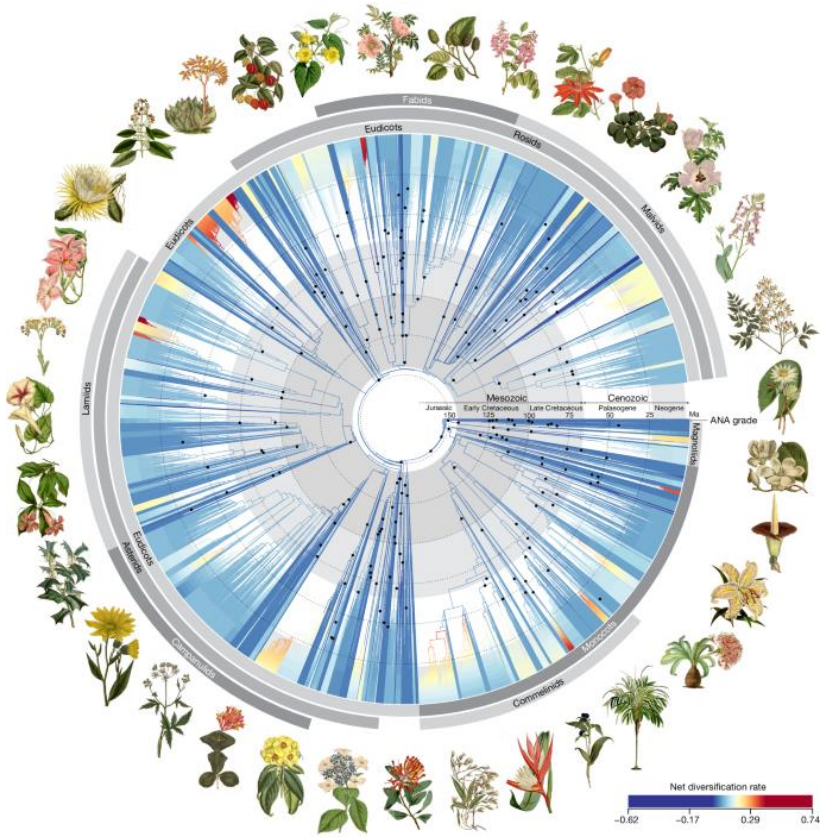
生物信息学的研究内容

- 序列测定与分析
- 比较基因组学
- 转录组学
- 转录调控与表观遗传学
- 蛋白质组学
- 代谢组学
- 表型组学
- 系统生物学
- 整合生物信息学
- 面向国家战略需求的研究
- 生物信息学资源平台建设

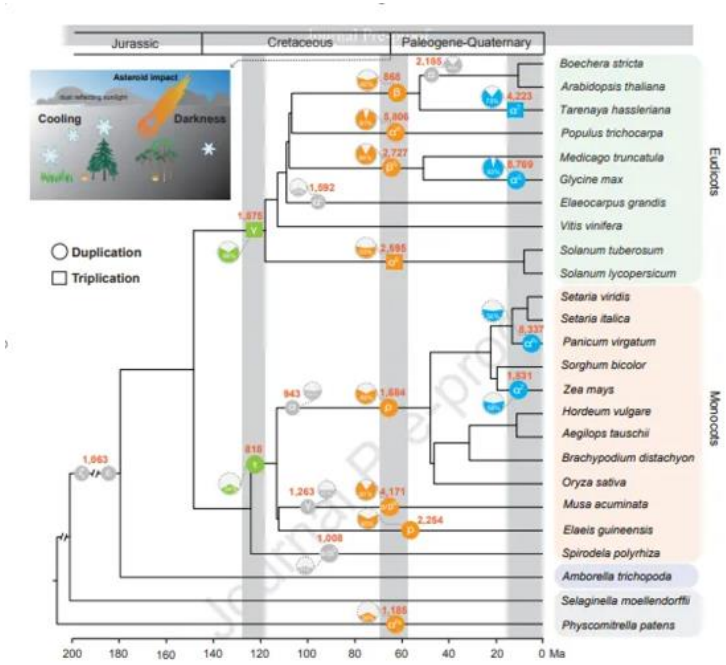


生物信息学的研究内容

- 序列测定与分析
- 比较基因组学
- 转录组学
- 转录调控与表观遗传学
- 蛋白质组学
- 代谢组学
- 表型组学
- 系统生物学
- 整合生物信息学
- 面向国家战略需求的研究
- 生物信息学资源平台建设



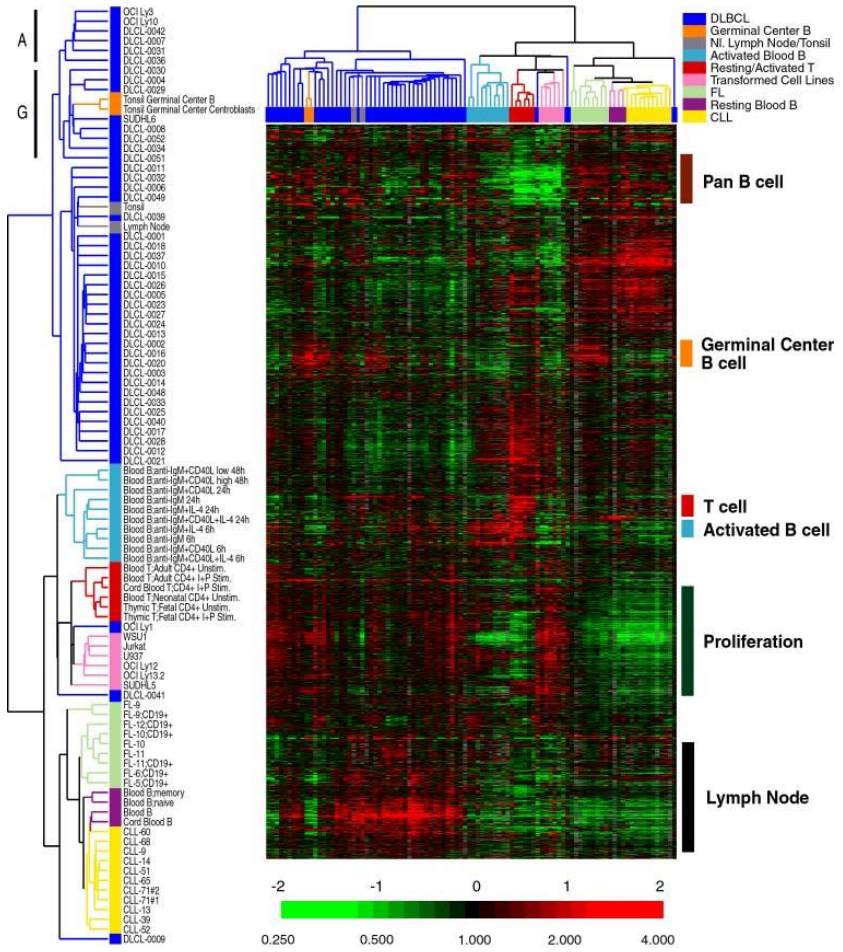
被子植物的“生命之树”
（Zuntini *et al., Nature, 2024*）



陆地植物基因组复制历史
（Wu S et al., Molecular plants. 2019）

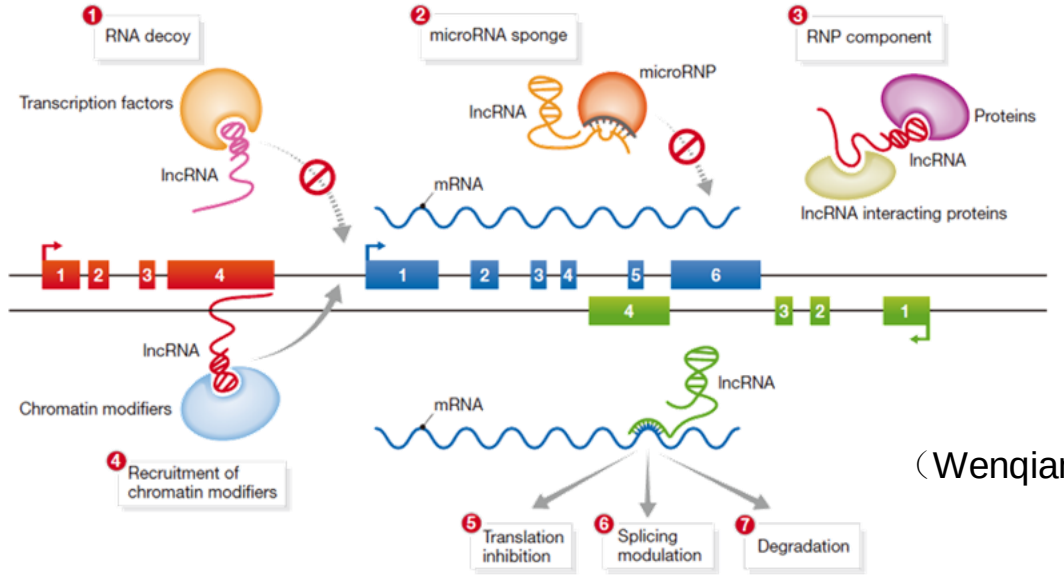
生物信息学的研究内容

- 序列测定与分析
- 比较基因组学
- 转录组学
- 转录调控与表观遗传学
- 蛋白质组学
- 代谢组学
- 表型组学
- 系统生物学
- 整合生物信息学
- 面向国家战略需求的研究
- 生物信息学资源平台建设

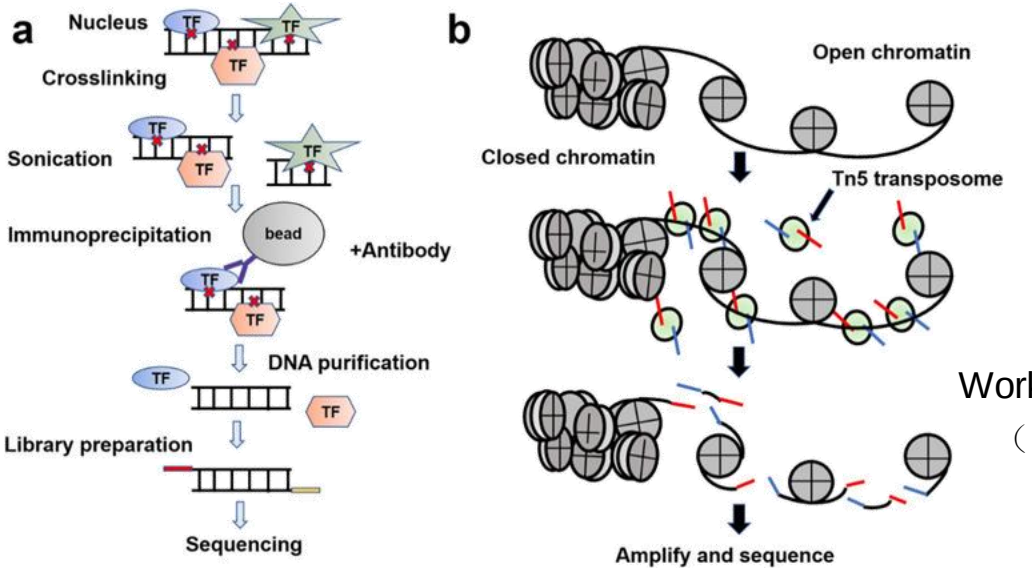


生物信息学的研究内容

- 序列测定与分析
- 比较基因组学
- 转录组学
- 转录调控与表观遗传学
- 蛋白质组学
- 代谢组学
- 表型组学
- 系统生物学
- 整合生物信息学
- 面向国家战略需求的研究
- 生物信息学资源平台建设



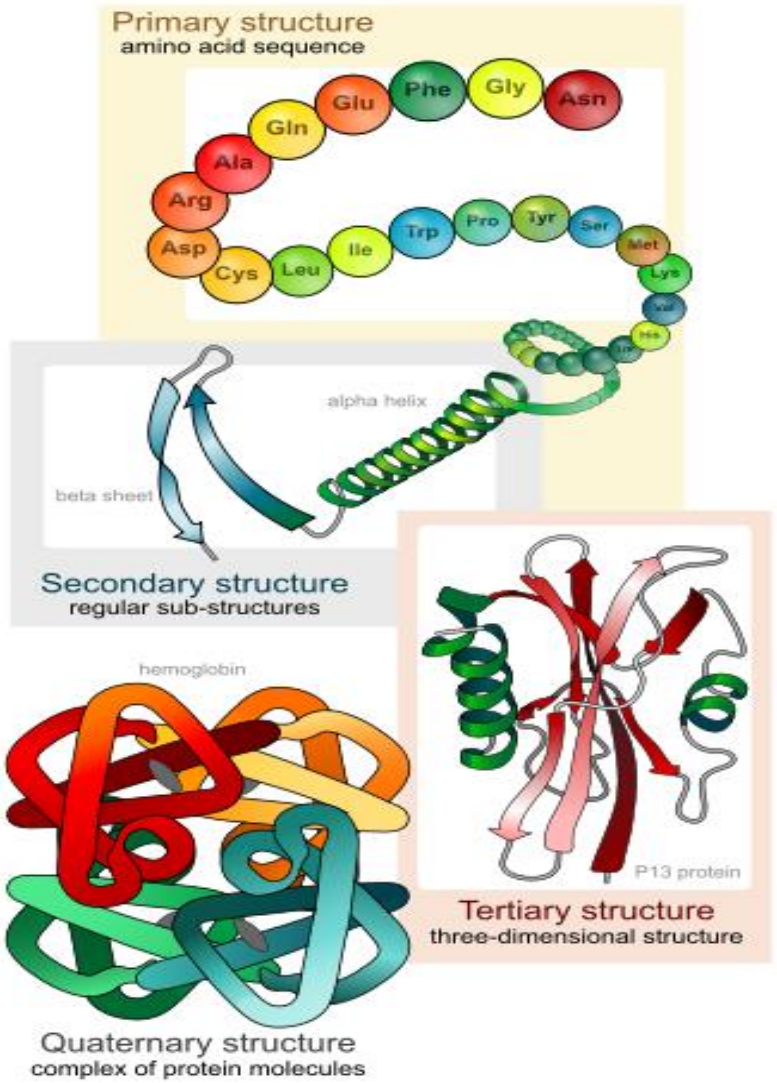
lncRNA的调控
（Wenqian Hu, et al., EMBO reports. 2012）

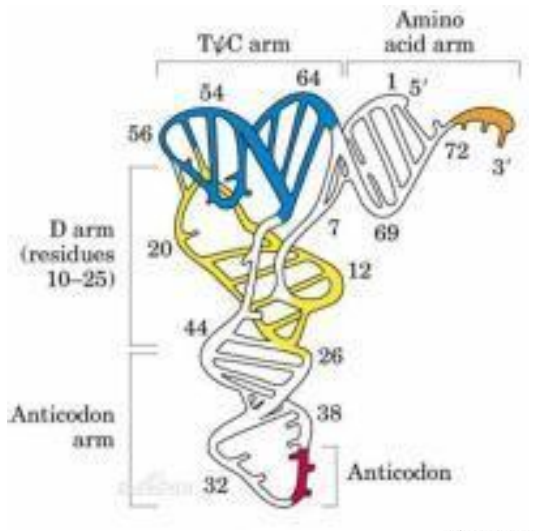


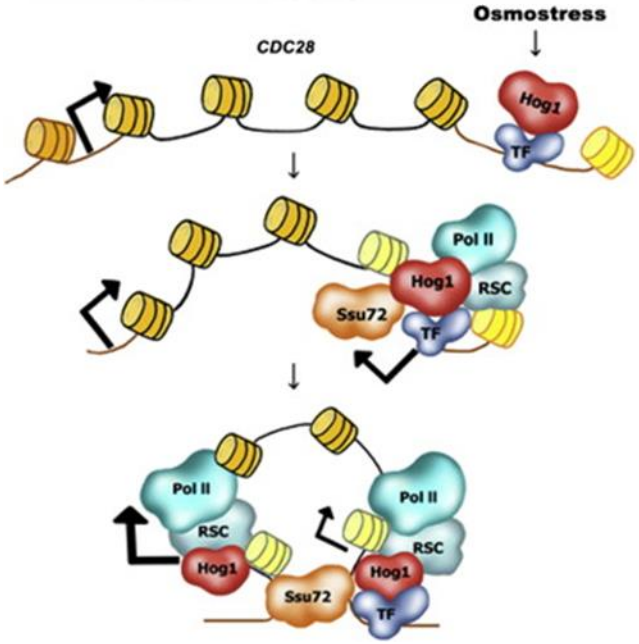
Workflows of ChIP-seq and ATAC-seq.
（Ma, et al., Mol Biomed. 2020）

生物信息学的研究内容

- 序列测定与分析
- 比较基因组学
- 转录组学
- 转录调控与表观遗传学
- 蛋白质组学
- 代谢组学
- 表型组学
- 系统生物学
- 整合生物信息学
- 面向国家战略需求的研究
- 生物信息学资源平台建设

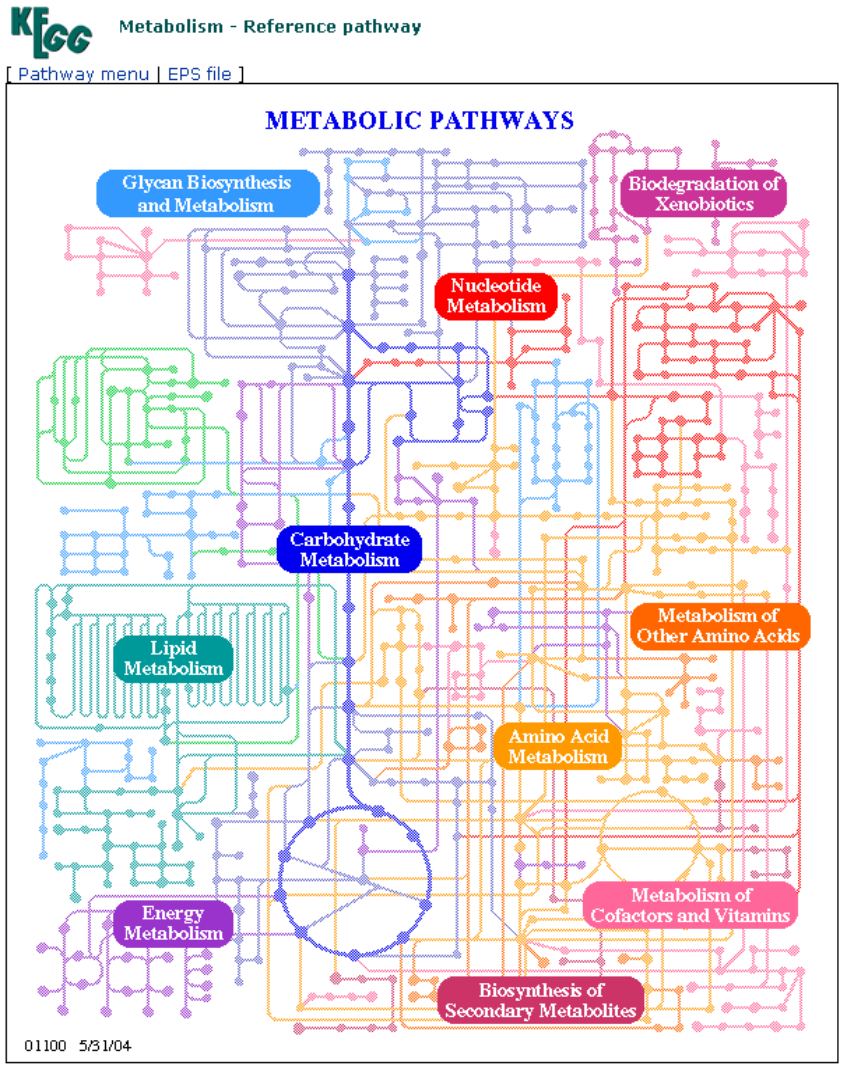






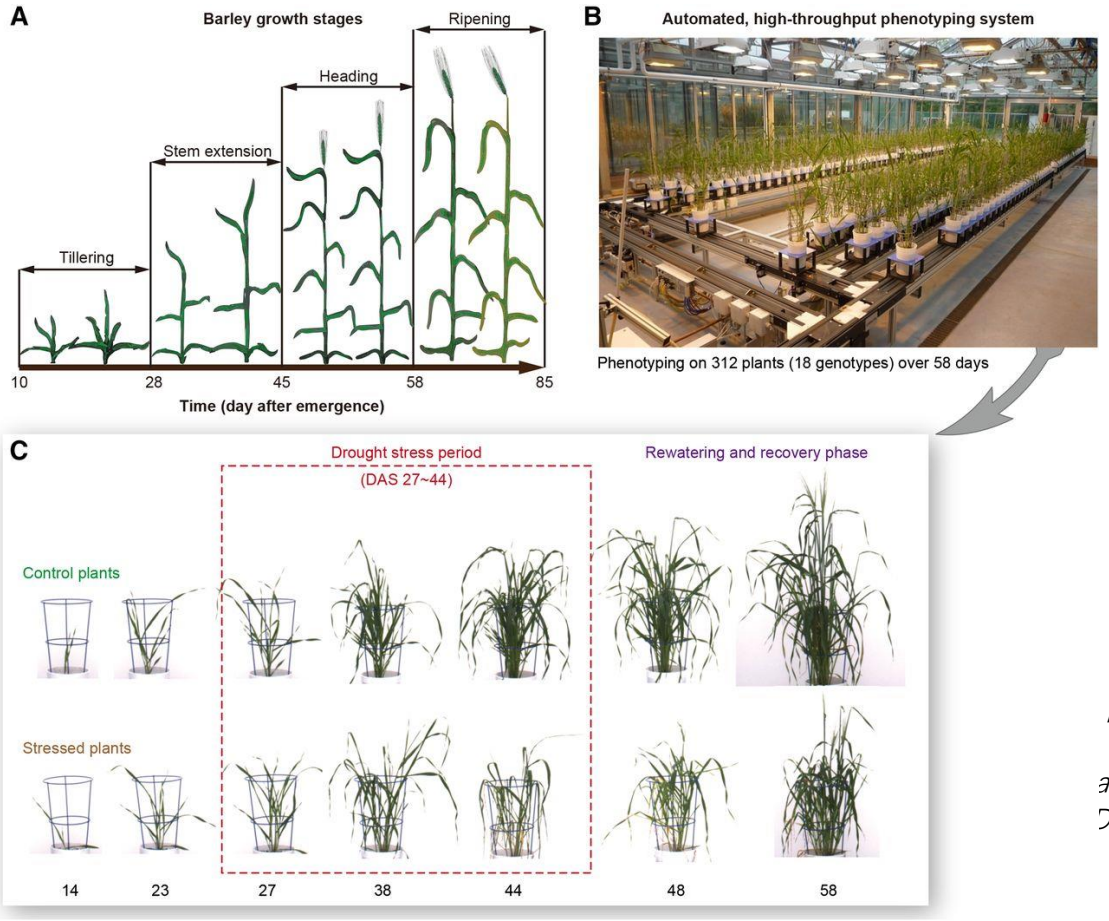
生物信息学的研究内容

- 序列测定与分析
- 比较基因组学
- 转录组学
- 转录调控与表观遗传学
- 蛋白质组学
- 代谢组学
- 表型组学
- 系统生物学
- 整合生物信息学
- 面向国家战略需求的研究
- 生物信息学资源平台建设



生物信息学的研究内容

- 序列测定与分析
- 比较基因组学
- 转录组学
- 转录调控与表观遗传学
- 蛋白质组学
- 代谢组学
- 表型组学
- 系统生物学
- 整合生物信息学
- 面向国家战略需求的研究
- 生物信息学资源平台建设



植物表型组学的五大支柱

Data acquisition
Sensors, physics, calibration, metrology, vehicles, robotics

Data management
Ontology, Standards, Metainformation, Information system, Data sharing

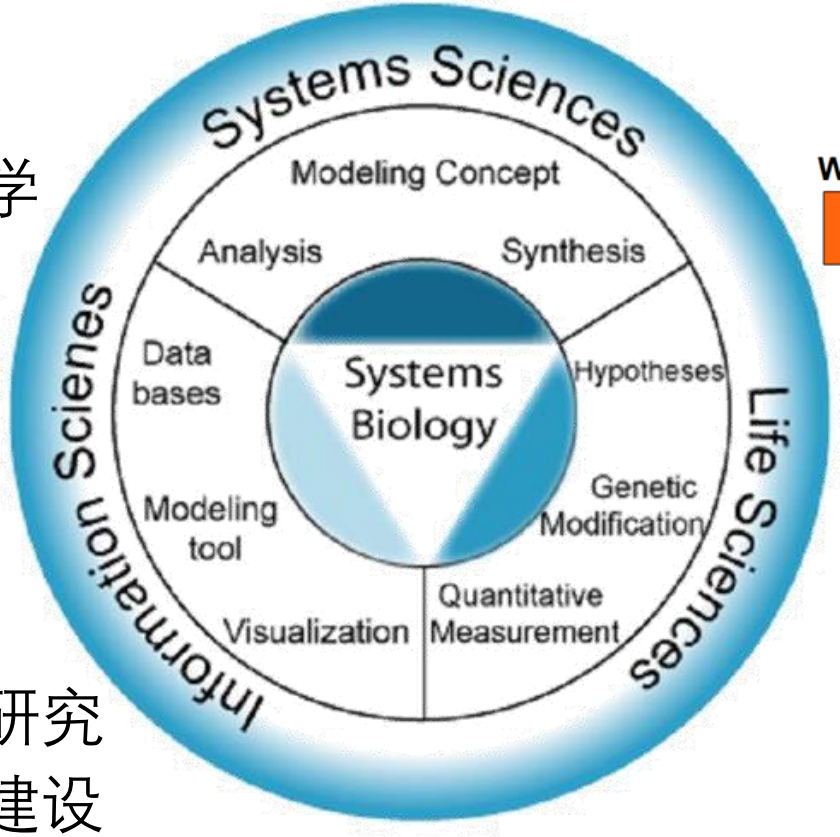
Data interpretation
Computer vision, Machine learning, Statistics, Signal processing, Data fusion, Scaling

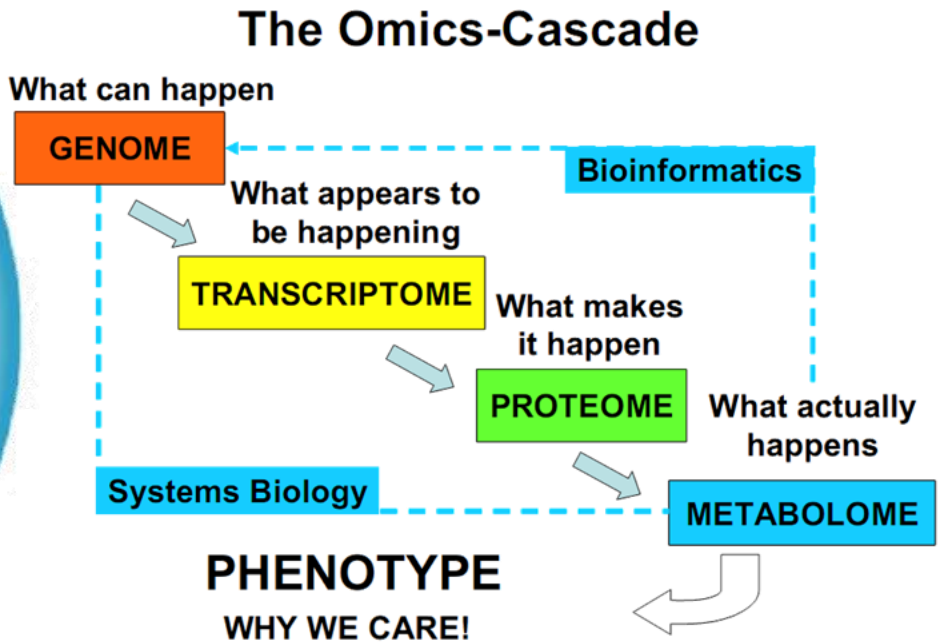
Modeling
Ecophysiological processes, Dynamics of plant structure, Meta-analysis, Data assimilation, Sensitivity analysis

Application
Plant breeding
$G \times E \times M$
agement, Precision agricul
Decision support

生物信息学的研究内容

- 序列测定与分析
- 比较基因组学
- 转录组学
- 转录调控与表观遗传学
- 蛋白质组学
- 代谢组学
- 表型组学
- 系统生物学
- 整合生物信息学
- 面向国家战略需求的研究
- 生物信息学资源平台建设





生物信息学的研究内容

- 序列测定与分析
- 比较基因组学
- 转录组学
- 转录调控与表观遗传学
- 蛋白质组学
- 代谢组学
- 表型组学
- 系统生物学
- 整合生物信息学
- 面向国家战略需求的研究
- 生物信息学资源平台建设

1. 整合生物信息学的研究领域
2. 生物数据挖掘与整合
3. 生命科学与生信技术的整合
4. 学科、人才的整合

陈铭，整合生物信息学，《*计算机教育*》，2006

生物信息学的研究内容

- 序列测定与分析
- 比较基因组学
- 转录组学
- 转录调控与表观遗传学
- 蛋白质组学
- 代谢组学
- 表型组学
- 系统生物学
- 整合生物信息学
- 面向国家战略需求的研究
- 生物信息学资源平台建设

生物经济五年规划

- 2022年5月10日，国家发展改革委印发《“十四五”生物经济发展规划》（以下简称《规划》）《规划》明确提出，推动基因检测、生物遗传等先进技术与疾病预防深度融合，开展遗传病、出生缺陷、肿瘤、心血管疾病、代谢疾病等重大疾病早期筛查，为个体化治疗提供精准解决方案和决策支持。
- 我国生物经济创新发展还面临不少挑战。比如，原始创新能力仍较为薄弱，基础生命科学理论、底层关键共性技术、高端仪器和试剂、生物信息资源等积累不够，以企业为主体、市场为导向、产学研深度融合的技术创新体系仍不完善

四大重点发展领域

- **一是**顺应“以治病为中心”转向“以健康为中心”的新趋势，发展面向人民生命健康的生物医药。

健康中国

- **二是**顺应“解决温饱”转向“营养多元”的新趋势，发展面向农业现代化的生物农业。

健康农业

- **三是**顺应“追求产能产效”转向“坚持生态优先”的新趋势，发展面向绿色低碳的生物质替代应用。

绿色发展（合成生物学）

- **四是**顺应“被动防御”转向“主动保障”的新趋势，加强国家生物安全风险防控和治理体系建设。

高质量发展

生物信息学应用

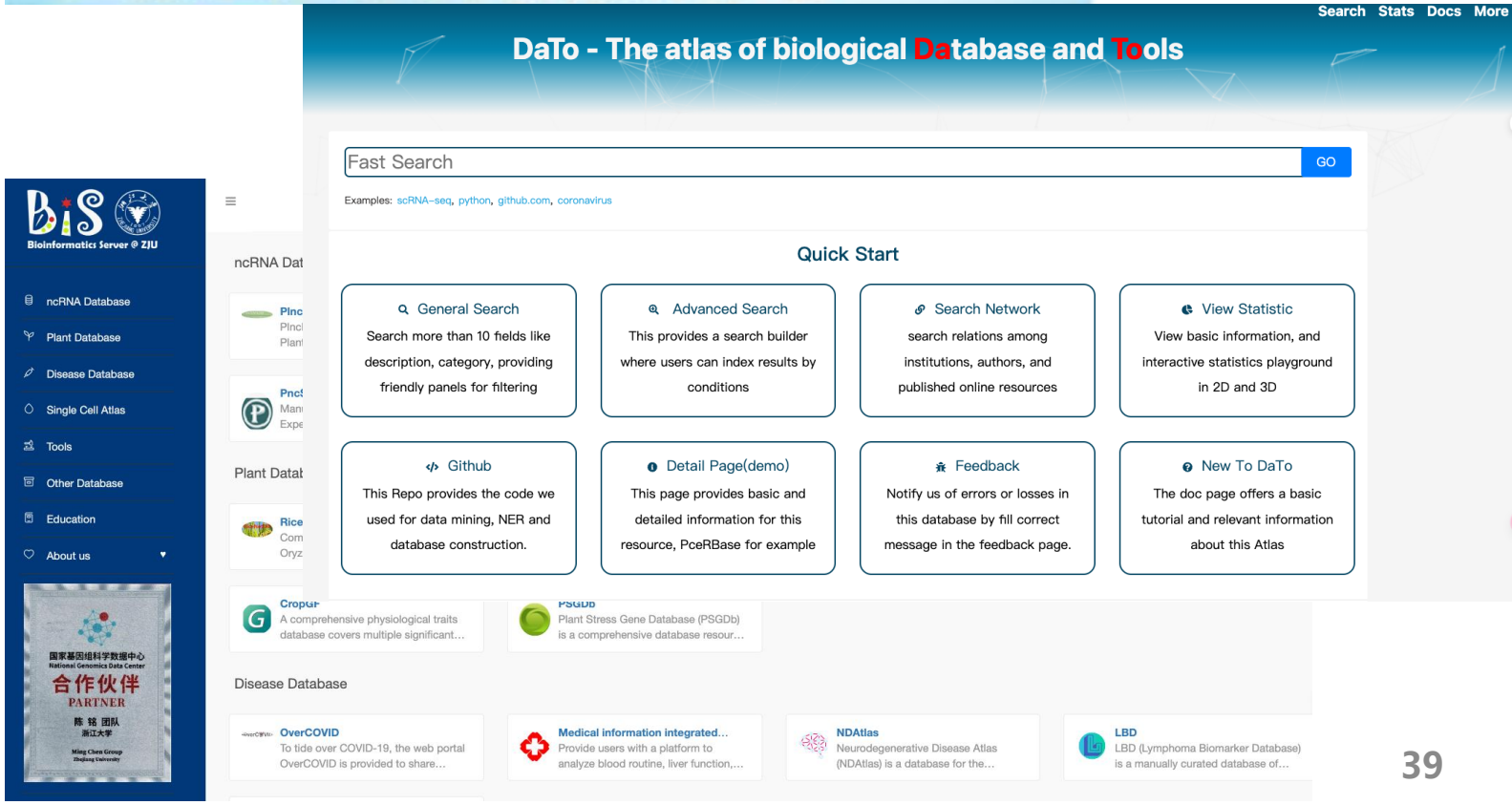
- Molecular medicine分子医学
- Personalised medicine个性医学
- Preventative medicine预防医学
- Gene therapy基因治疗
- Drug development 药物研发

•创新认知生物系统，破译生命信息
•改变生物学的研究方式，指导服务实验科学
•促进农林、医药学等及相关产业发展

- Microbial genome applications微生物基因组应用
  - Waste cleanup, Climate change Studies, Alternative energy sources, Biotechnology, Antibiotic resistance, Forensic analysis of microbes 法医检定, The reality of bioweapon creation
- Evolutionary studies进化研究
- Crop improvement作物改良
  - Insect resistance, Improve nutritional quality, Development of Drought resistance varieties
- Vetinary Science 兽医科学
- Comparative Studies比较研究

生物信息学的研究内容

- 序列测定与分析
- 比较基因组学
- 转录组学
- 转录调控与表观遗传学
- 蛋白质组学
- 代谢组学
- 表型组学
- 系统生物学
- 整合生物信息学
- 面向国家战略需求的研究
- 生物信息学资源平台建设



生物信息学的研究内容

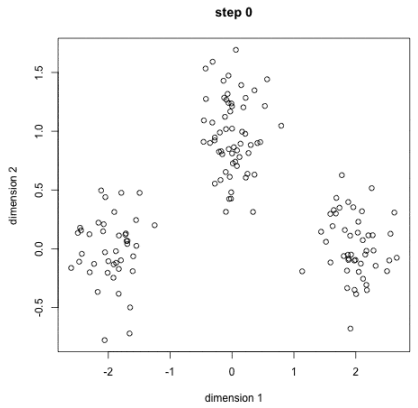

NCBI
EMBL-EBI
DDBJ
DNA Data Bank of Japan
SIB
UniProt
CLUSTAL
MEGA
Molecular Evolutionary Genetics Analysis
国家基因组科学数据中心
National Genomics Data Center
KEGG
GEO
Gene Expression Omnibus
Swiss Institute of Bioinformatics
MaxQuant
MQ
China National GeneBank 国家基因库
CNGB
GENEONTOLOGY
Unifying Biology
Sequence Read Archive
Bioinformatics >> Prediction Servers >> SignalP-5.0
SignalP-5.0 Server
RCSB PDB
PROTEIN DATA BANK
GROMACS
FAST. FLEXIBLE. FREE.
blastx
translated nucleotide ▶ protein
MUSCLE
HUMAN PROTEOME MAP
Pfam
RxList
The Internet Drug Index
BOW TIE
Bowtie 2
Fast and sensitive read alignment
MAFFTash
ALIGNMENT OF MULTIPLE SEQUENCES AND STRUCTURES
e!
I-TASSER
Protein Structure & Function Predictions
DRUGBANK
Trinity
Chrysalis
Inchworm
Butterfly
Cufflinks
Galaxy
PROJECT
GWAS Catalog
ENCODE
1987 2019
BRENDA
The Comprehensive Enzyme Information System
TCGA
NONCODE
NCI | NHGRI
miRBase
TargetScan
Prediction of microRNA targets
BioCyc
Database Collection
DAVID
BIOINFORMATICS DATABASE
GENCODE
OMIM
reactome
PubMed
gatk
GTExPortal
ImageJ
Image Processing and Analysis in Java
igv
Integrative Genomics Viewer
CCLE
Cancer Cell Line Encyclopedia
CGA
Cancer Genome Analysis
RAxML
STRING

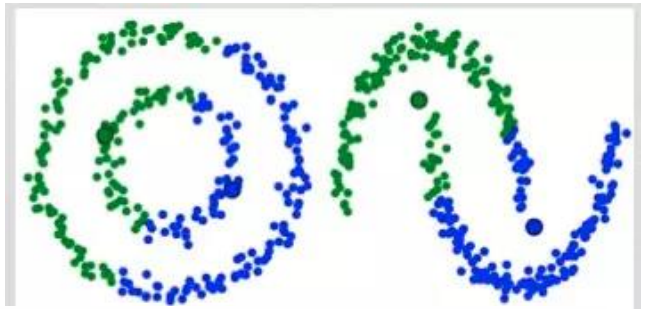
第四节　生物信息学算法基础

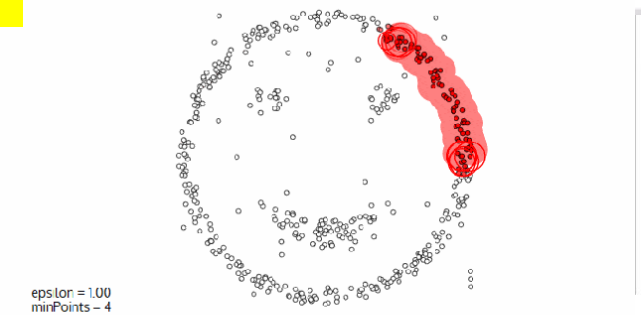
生物信息学算法基础

- 同
- 回归
- 分类
- 聚类
- 关联
- 。 。 。

大数据
特征
降维

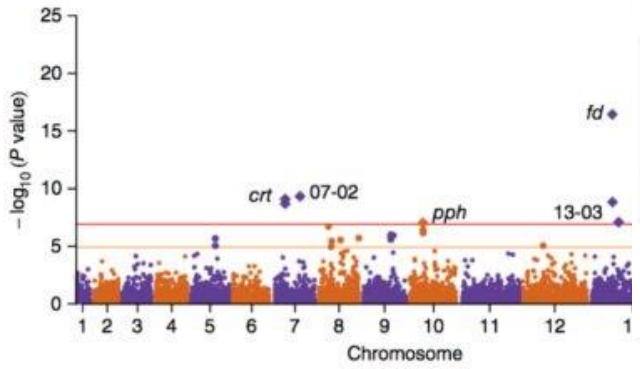


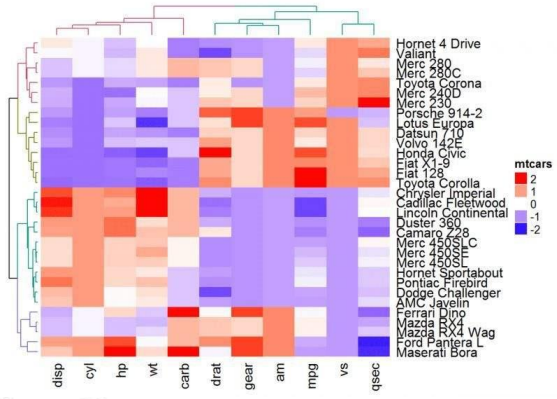




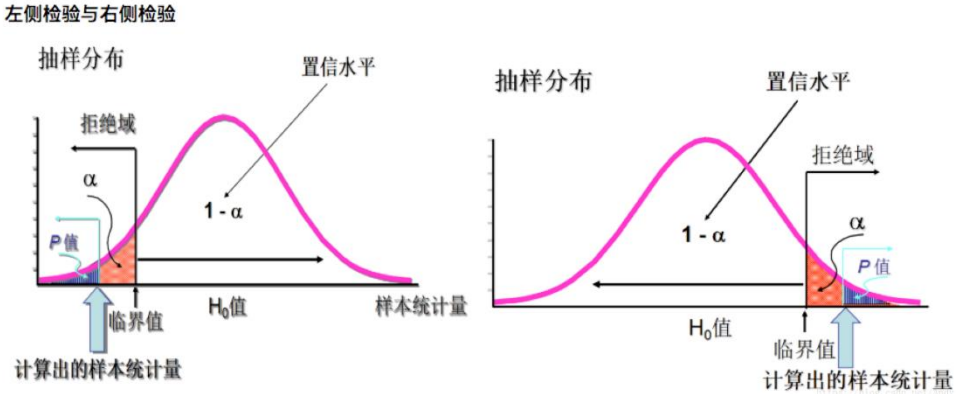
- 不同

- 同的另一面





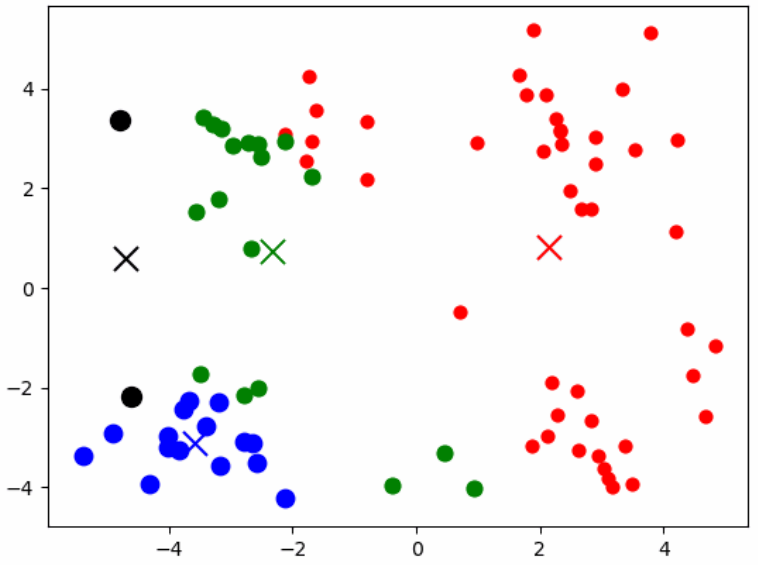
- 差异分析方法：统计学差异是否存在差异？ 方差分析（F检验）、t检验、卡方检验 **P值**



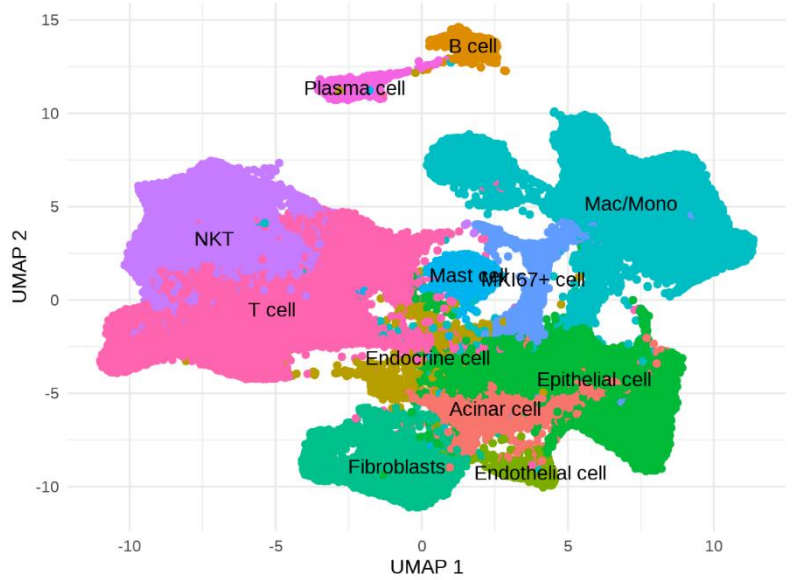
- **聚类算法**

聚类：按照数据内在的相似性将未标注的数据集划分为多个类别，以寻找数据中隐含的范式；同一类别内的数据相似度较高，不同类别的数据相似度较低

生物信息学应用：如基因表达谱的样本分组、蛋白质序列的家族分类等



K-means聚类



聚类应用：

单细胞聚类，

细胞类型注释

划分聚类：将数据集划分为不同的聚簇，如K-means算法、K-medoids算法等

层次聚类：构建一棵聚类树来创建多层次的分层结构

密度聚类：将聚簇定义为高密度区域之间的低密度区域

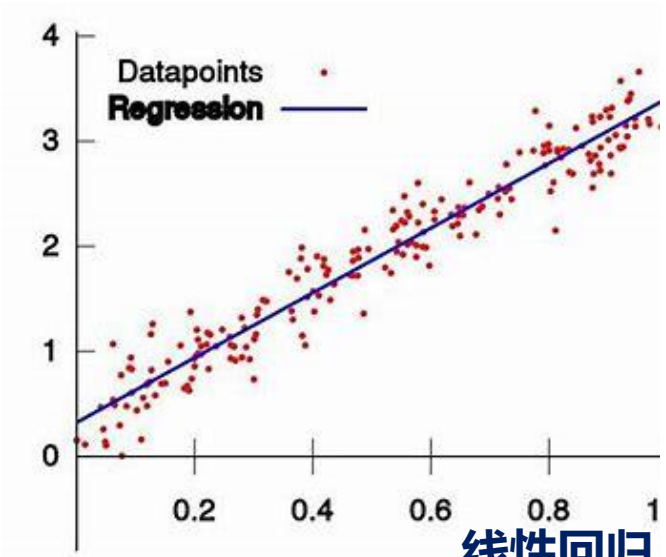
网格聚类：将数据空间划分为网格，然后将数据分配到对应的网格中

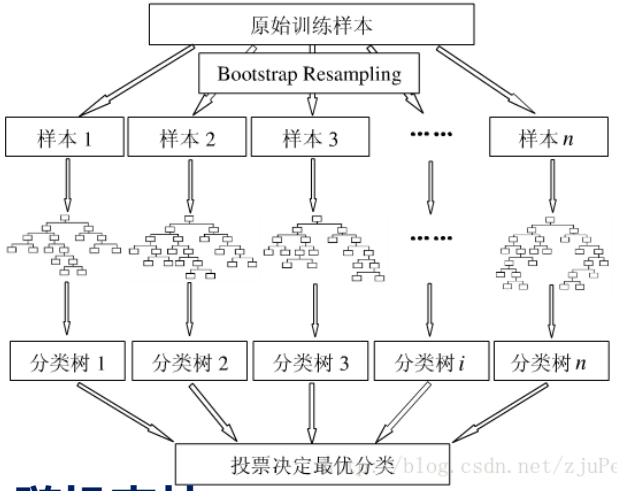
混合聚类：将多个聚类算法结合起来，以便更好地捕捉不同数据结构

生物信息学算法基础

- 回归算法

回归：监督学习方法，对数值型连续随机变量进行建模和预测

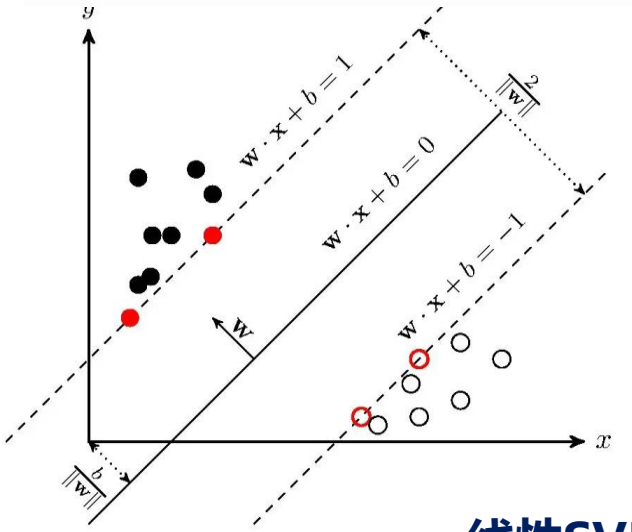


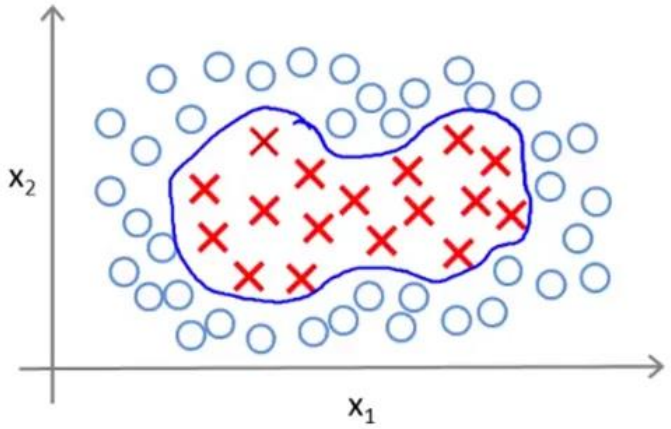


线性回归 和 随机森林

- 分类算法

分类：监督学习方法，对有标签样本学习，从而将未知类别的样本分类到预定义的类别或标签中





线性SVM 和 核SVM

- **降维算法**

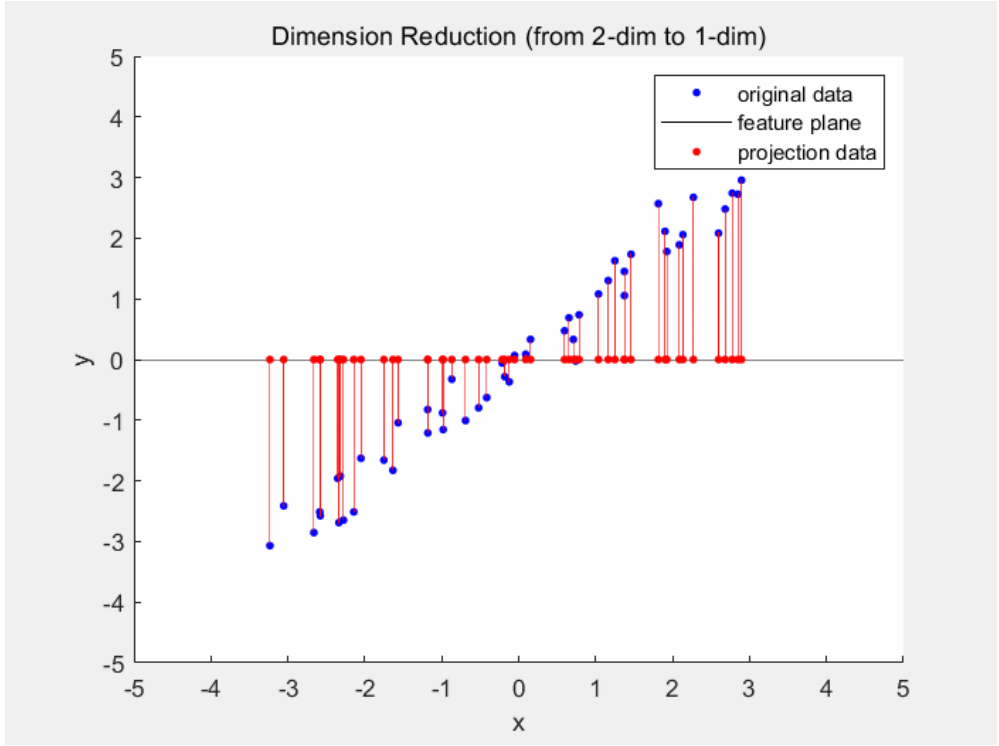
生物大数据——多模态，高维度，包含大量的冗余和噪声（例如单细胞转录组表达矩阵上万个细胞×上万个细胞）

降维算法——从高维数据中提取出最具代表性的特征，保留主要信息的同时减少冗余信息和噪声，将数据降维到二维或三维还有助于直观的数据可视化

基于特征选取的方法：挑选最具信息量的特征以减少数据维度，如缺少值比率（Missing Value Ratio）、随机森林（Random Forest）等

基于组分因素的方法：试图找到数据中的主要成分或因素以实现降维，如因素分析（Factor Analysis）、主成分分析（PCA）等

基于映射的方法：将数据映射到新的低维空间中，如等距映射 ISOMAP、t-SNE和UMAP等



PCA降维：将n维的向量映射到k维上，这k维是全新的正交特征称为主成分Principal Component，每个主成分都保留了原始数据一部分的变异

第五节　生物信息学的机遇与挑战

生物信息学的机遇

数据量爆炸式增长

- 高通量测序技术的突破性进展；
- 丰富的数据资源；
- 各种组学技术的兴起；

人工智能与算法创新

- 人工智能的不断发展；
- 深度学习等算法的应用；

跨学科合作的加强

- 与计算机科学、数学、统计学之间更加紧密的合作；
- 深化生物现象的理解，为其他领域的发展也带来新的思路和方法；

高维度、高数量级的生物大数据

Big Data!

- 估计地球上生活着$4.6*10^{30}$个细菌细胞（地球质量$2.9*10^{28}$克)
- 细菌的种数肯定超过10^6(每种动物包括昆虫都有特异的共生细菌)
- 细菌总含碳量约3500-5500亿吨，是植物含碳量的60-100%
- 细菌总含氮量约850-1300亿吨，总含磷量约90-140亿吨，两者都比植物多10倍
- 单个细菌基因组大小：$0.5*10^6$-$1.4*10^7$bp (so far)
- 人体内外大约有100万亿个细菌分布，占体重的1-3%（2斤左右）
- 人体细胞公约40-60万亿个，平均直径10-20mm

- Even in *Escherichia coli*, for instance, there are 225,000 proteins, 15,000 ribosomes, 170,000 tRNA-molecules, 15,000,000 small organic molecules and 25,000,000 ions inside the a few µm cell.
- There are estimated 10^{14}-10^{16} biochemical reactions in a cell.

部分数据来自郝柏林院士相关报告

生物信息学的挑战

数据管理与分析的复杂性

- 数据本身的准确性；
- 数据的质量和标准化；
- 如何确保数据的可靠性和可比性；

常用工具与创新方法

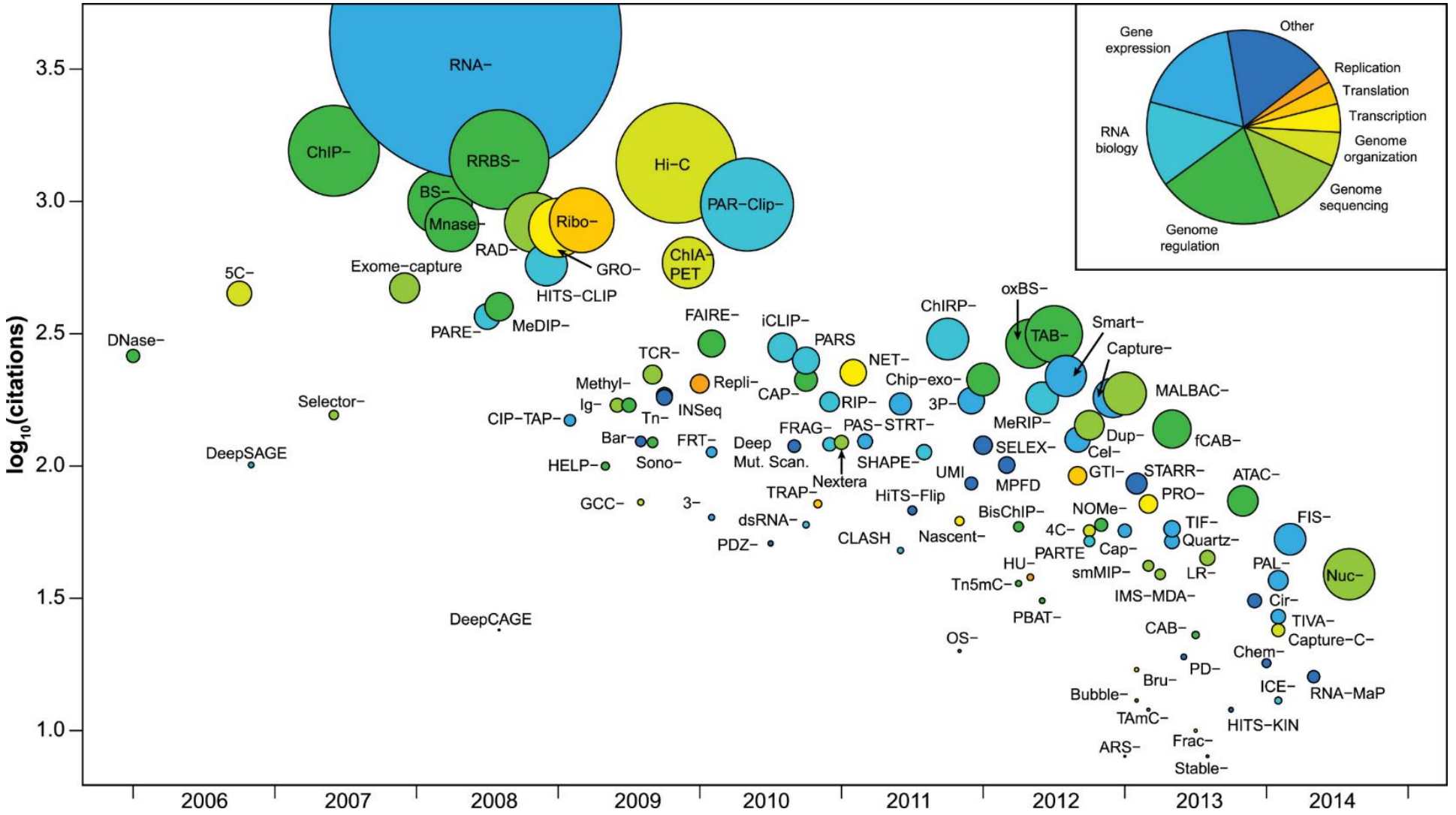
- 开发的新工具与人们常用的经典工具；
- 鼓励尝试新工具，推动创新和发展；

多尺度数据的整合与交叉分析

对于多种尺度的数据，例如多种组学数据，开发更为复杂的方法和工具进行整合和交叉分析；

生物信息学的挑战

New Technology: Detectable & Precise, “Seeable”!



Global vs Local
Statics vs dynamics
General vs Specific

What we have seen?
Good and sufficient enough? (curation, clean)
Modeling and Prediction (features, samples)

How can we see more?
High resolution
Multi-dimension
Dynamics…

Data Challenge: to see what we can not see!

Different methods
Different results

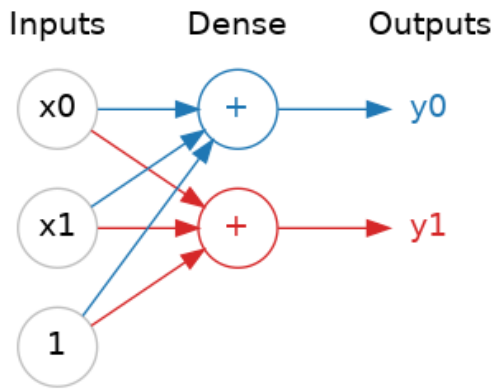
生物信息学的挑战

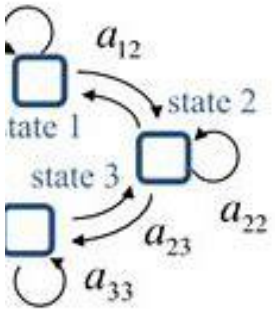
Data Mining

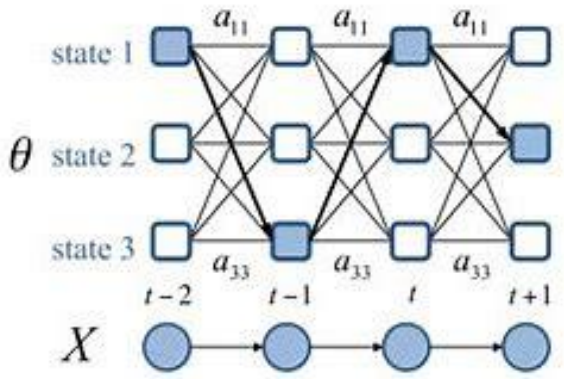
- 归纳逻辑程序（Inductive Logic Programming）
- 遗传算法（Genetic Algorithm）
- 神经网络（Neural Network）
- 统计方法（Statistical Methods）

常用的数据挖掘技术方法

- 贝叶斯方法（Bayesian Methods）
- 决策树（Decision Tree）
- 隐马尔可夫模型（Hidden Markov Model）
- 文本挖掘
- 网络挖掘





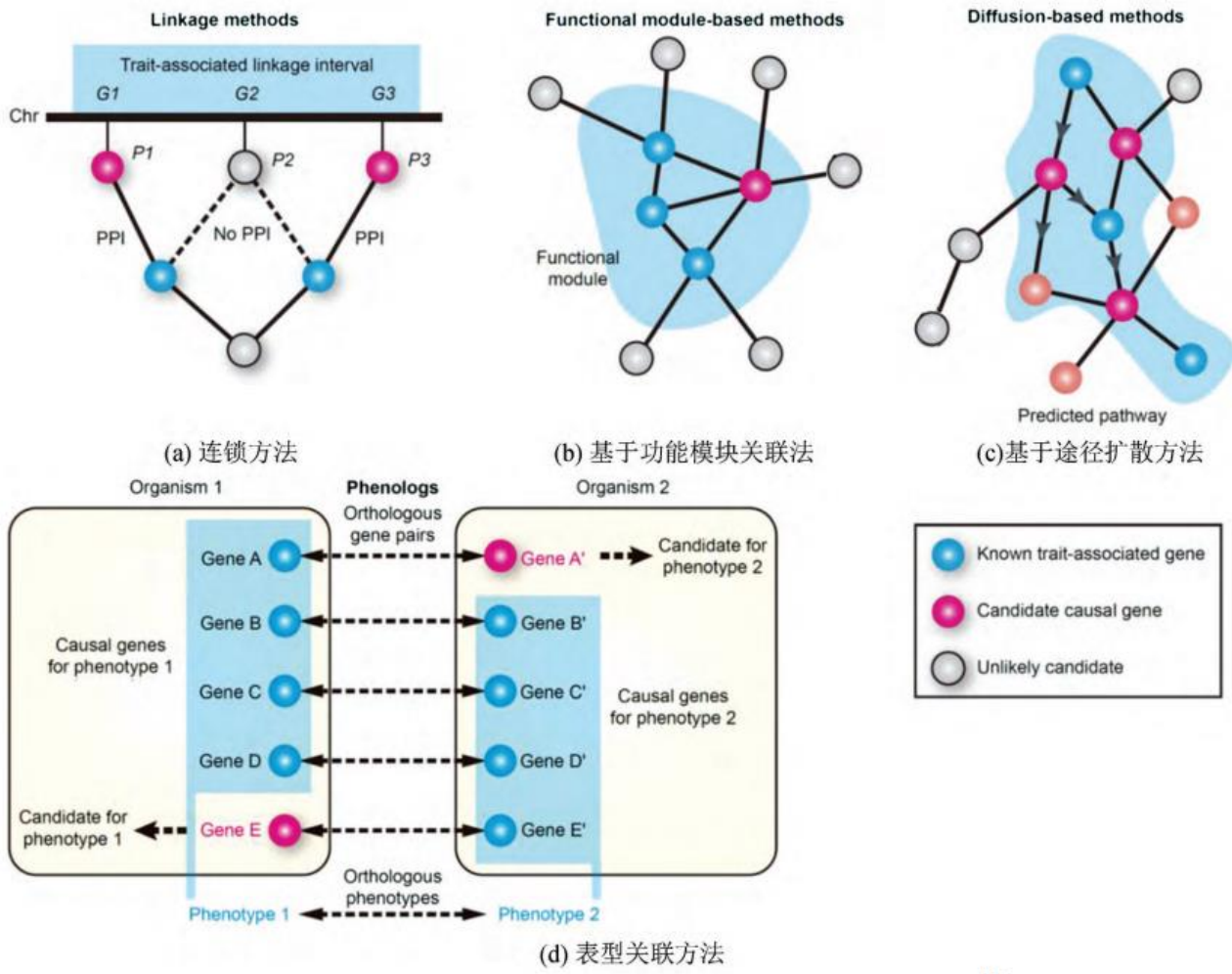


More…

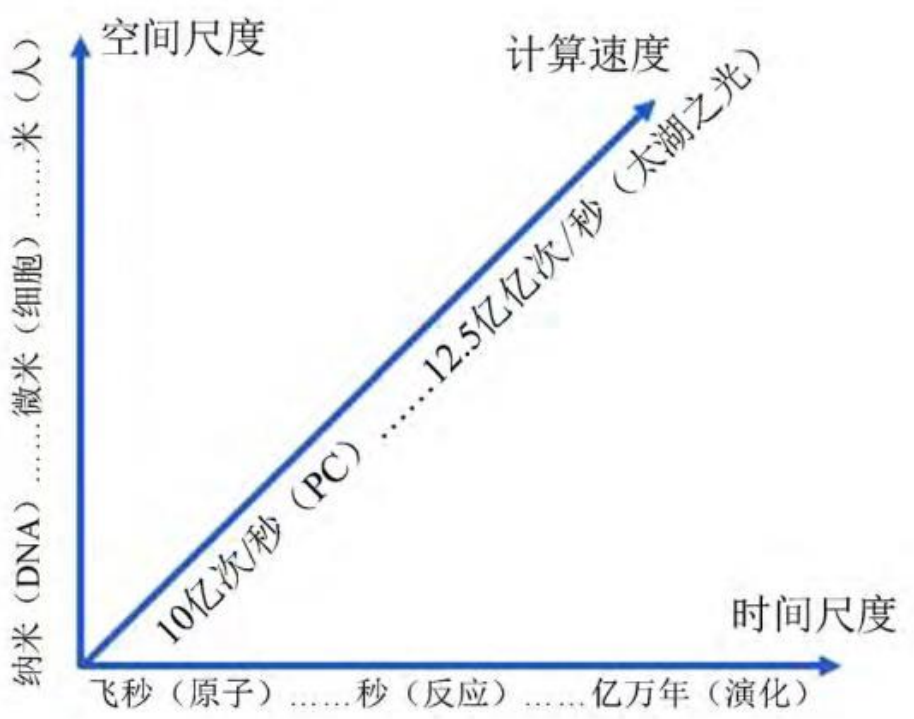
第二章

第三章

生物信息学的挑战



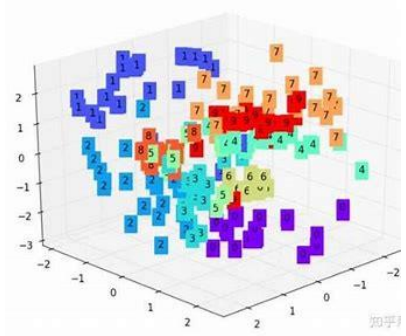
生物信息学相关性关系网络建立示意图



生物系统的时空复杂性与计算速度

陈铭，大数据时代的整合生物信息学，《生物信息学》，2022

生物信息学的挑战：方法与技术



特征（维度） → **整合、再整合**

经验模型（单纯主要特征）

机器学习模型（主要特征筛选）

深度学习模型（特征一把抓）

现代AI模型（多维细节特征）

AI+经验模型（全局细节特征）

第20卷 第2期 生物信息学 Vol.20 No.2
2022年06月 Chinese Journal of Bioinformatics Jun. 2022

DOI:10.12113/202110002

大数据时代的整合生物信息学

陈 铭
（浙江大学 生命科学学院,生物信息学系,杭州 310058）

摘 要:随着生物数据测量技术的不断发展,生物数据的类型、内容、复杂度不断增加,生物信息学已迈入大数据时代。面对大数据时代多模态、多层次、高维度、非线性的复杂生物数据,生物信息学需要发展相应的方法和技术进行有效整合生物信息学研究与应用。本文对大数据时代整合生物信息学所涉及的数据整合、方法整合、系统整合及相关问题进行梳理和探讨。

关键词:整合生物信息学;生物系统;组学;大数据;问题;对策

各种特征的“线路图”整合设计CNN为代表，能拟合复杂的非线性关系

考虑各种特征关系与贡献，Attention为代表，考量不同特征，能处理更复杂的数据、完成更困难的任务

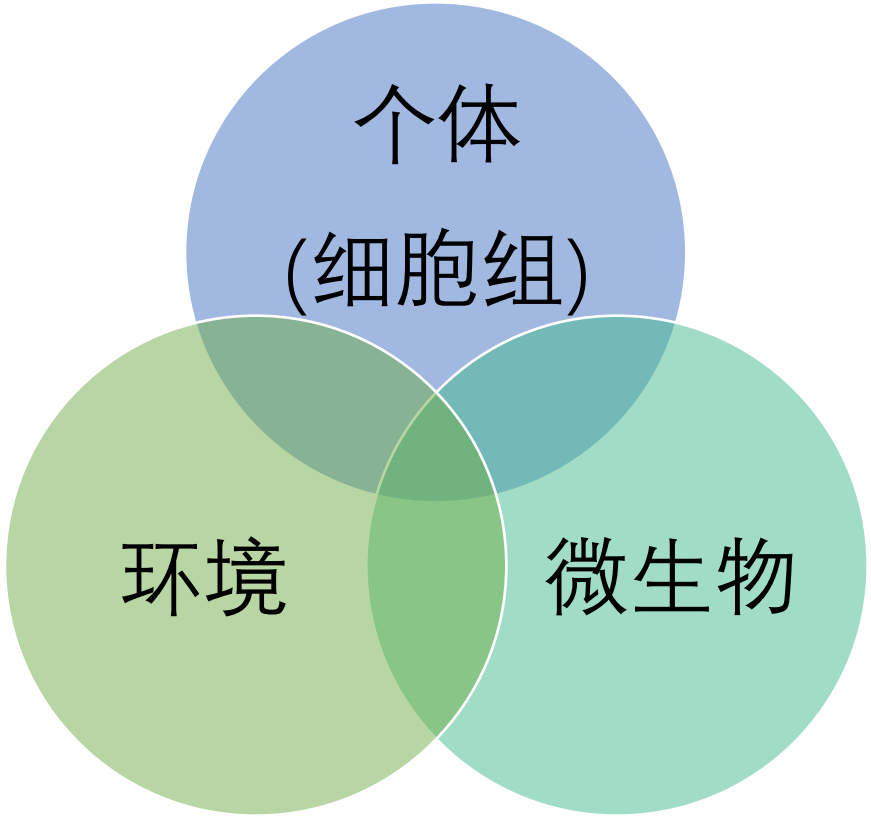
大模型（多任务多模态）、强人工智能、智能机器，能力更强，潜力巨大

生物信息学的挑战：整合生物信息学

基因组学（Genomics）
表观基因组学（Epigenomics）
转录组学（Transcriptomics）
蛋白组学（Proteinomics）
代谢组学（Metabolomics）
脂类组学（Lipidomics）
糖组学（Glycomics）
免疫组学（Immunomics）
微生物组学（microbiomics）
表型组学（Phenomics）
暴露组学（Exposomics）
辐射组学（Radiomics）
…
单细胞组学（Single-cell omics）
文献组学（Bibliomics）

？小科技

组学(Omics) → 整合生物信息学 → 解决大问题



Multiple Omics
Various Levels

Coding! or
Non-coding?!

人口 长寿与衰老
能源 生物能源
粮食 绿色农业 健康农业
环境 绿色发展

生物演化：从哪里来、到哪里去？我是谁？
分子育种、智慧育种、生物智造
生物安全、生态安全、同一健康

？大科学大工程

第六节　生物信息学的教育教学

如何成为生物信息学家？

当谈及生物信息学教学时，我们在讨论什么？

我们该从哪些方面培养合格的生物信息学家？

工作环境：
LINUX使用
环境的搭建
服务器运行维护
网站运行维护

基本的编程
Python
R

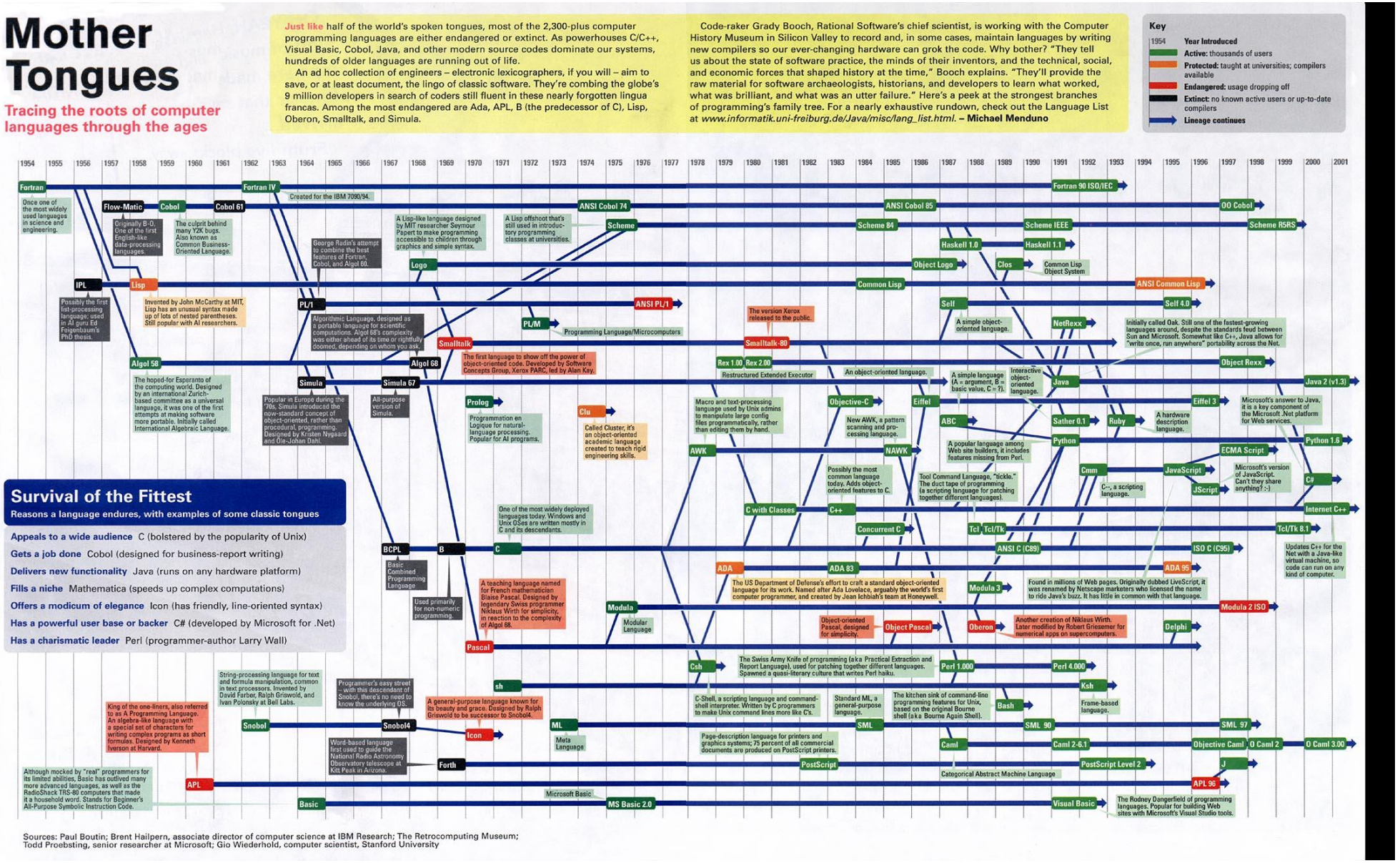
生物化学
分子生物学
生物物理学
“中心法则”

？

生物信息学工具：
比对工具
种群模拟
可视化
分子模拟

细分领域：
（系统生物学、计算生物学、群体遗传学等等）
细分技能：
（深度学习、数据库建设、生物统计等等）

Programming Languages



教育&科研定位导向供讨论

大科学计划、大科学工程为大国之王道！

新突破、新机制，引领发展！　　新技术、新工具　新算法、新理论

工程化

Level 5: 重大科学计划或突破性创新研究?

Level 4: 基于数据分析的生物学发现?

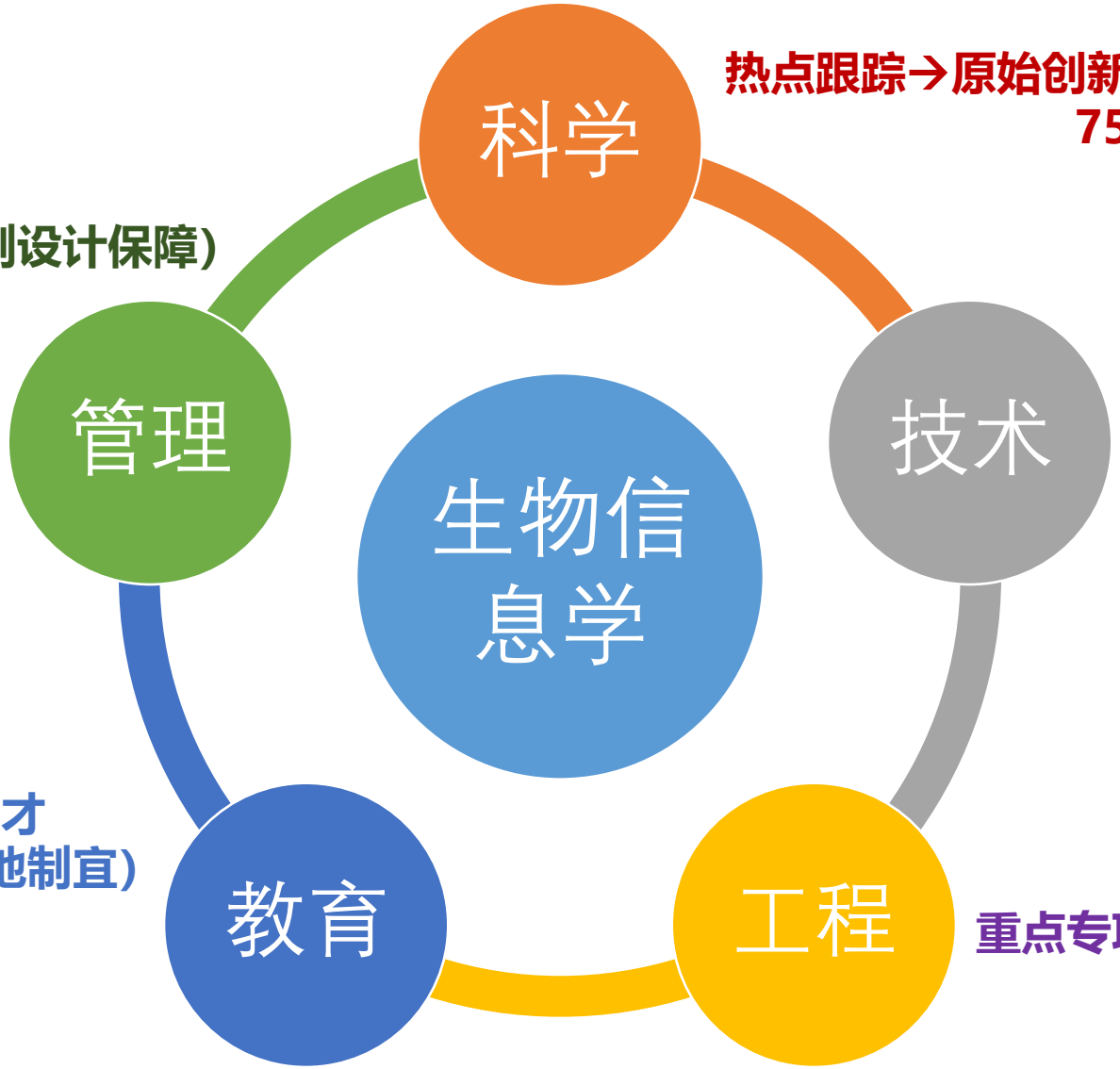
Level 3: 开发方法与软件工具，技术开发

Level 2: 简单数据分析，技术应用

Level 1: 为建模而建模，方法探索

问题驱动　　数据驱动

生物信息学重定位：STEEM



热点跟踪→原始创新（跟跑、并跑、领跑）
75-80%

mn→1+N（更新迭代生命力）
65-75%

重点专项→科学大工程（设计并引领）
50-60%

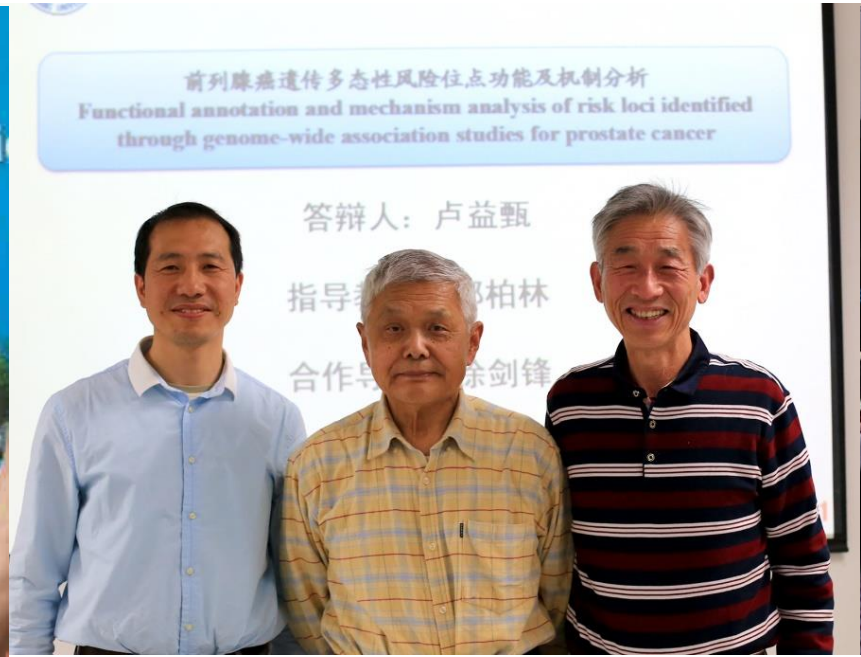
热门专业→核心人才
(各level因材施教、因地制宜)
80-90%

开始重视→科学管理（制度机制设计保障）
40-50%

（浙江大学）生信建设发展历程

- 1) 全国首批建设单位之一，2002年开始招生；
- 2) “本科—硕士—博士 ”专业、学科；生命科学学院
- 3) 沃森基因组科学研究院、浙江加州国际纳米技术研究院、生命科学研究院、转化医学研究院、定量生物中心。。。农学院、医学院、药学院、数学、计算机相关学院、海宁国际校区。

2006年学习调研单位:
哈尔滨医科大学
北京大学
清华大学
自动化系、生物学
上海交通大学
哈尔滨工业大学
厦门大学
。 。 。



教育部101计划生物信息学核心课程、核心教材、核心师资团队实施方案

联合全国33所“拔尖2.0”相关高校

国内外杰出科学家咨询

核心师资团队

一线专业课程教师研讨

核心教材

- 实验教材
- 简版教材
- 核心教材

生物信息学虚拟教研室

核心课程

- 理论课程
- 实验课程
- 在线课程
- 实践课程

科研、竞赛++产业基地

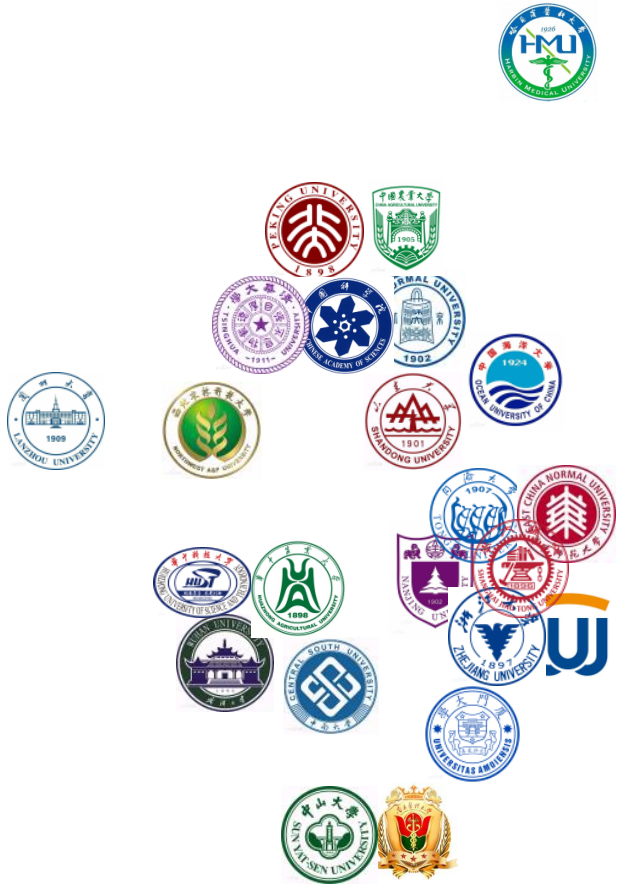
教学理念	团队建设	教材建设	质量标准
提升培养质量，激发潜力和学习动力，注意课程思政建设	参与意识强烈、理念先进、积极投入；任务分工合理	纸质化创新和数字化建设、应用与推广新模式	明确课程质量标准和教学大纲：课程目标、课程内容
课程模式	**教学方式**	**课程资源**	**质量管理**
理论+实验+实践+AI 线上线下；课内课外；现实虚拟	问题式、启发式、探究式、互动式：自主+批判+创新	数字化、试题库、案例库、代码库、云计算、竞赛平台	教学策略、教学内容、教学方法、教学过程、效果评价

生物信息学虚拟教研室

33所高校+若干其他优势领域高校

Attribution & Contribution

1+10（33）+N



全国更多生物信息学教学相关高校。

结语

“一场与工业革命和以计算机为基础的革命 有相同影响力的变化正在开始。下一个伟大时代将是基因组革命时代，它现在处于初 期阶段”。

——《第三次技术革命》

当前，多组学和生物信息学的发展已经进入又一个技术革命的时代——人工智能时代，生物信息学新方法新技术正在蓬勃发展，将对生物学、医学、药学、农业科学等领域产生巨大的影响，我们必定能够揭示各种生命现象的奥秘，并带动多个学科的跨越式发展，极有可能引发新的产业革命。

结语

本教材共十三章，可以构成六个内容模块及建议学时数。

模块一：生物信息学基础与资源
(第一章、第四章)

1.生物信息学大事记
2.生物信息学研究领域
3.我国生物信息学发展情况
4.生物信息学数据库

模块二：生物统计与人工智能
(第二章、第三章)

1.生物数据统计分析简介
2.参数估计与假设检验
3.统计建模
4.统计学习
5.高维统计分析
6.因果推断
7.深度学习
8.人工智能

模块三：序列、结构与功能的理论基础
(第五章、第九章部分、第十一章)

1.序列分析
2.分子进化和发育树构建
3.蛋白质结构分析
4.蛋白质分子动力学模拟
5.系统生物学方法

模块四：组学分析
(第六章、第七章、第八章、第九章部分、第十章)

1.基因组学
2.转录组学分析
3.表观组与转录调控分析
4.蛋白质组学分析
5.代谢组学
6.表型组学分析

模块五：应用与前沿
(第十二章)

1.精准医学
2.智能药学
3.智慧育种

模块六：生物信息学实验基础
(第十三章)

1.生物信息学编程与实验基础
3.数据分析流程搭建
4.数据库开发基础

参考教学时数

4学时 8学时 12学时
4学时 18学时 6学时

共享约束需知：使用条款

1. **版权声明**：本PPT及其所有内容（以下简称“本PPT”）仅用于教育和教学用途，版权归属于本PPT作者。
2. **使用要求**：任何使用本PPT的行为均须遵守以下条件：
   1) **致谢和标注**：若部分或全部使用本PPT的内容，请在使用内容的适当位置标注出处，并致谢本PPT作者。
   2) **修改和再分发**：未经作者书面许可，不得对本PPT进行修改或再分发。
3. **禁止商业化使用**：严禁将本PPT用于任何形式的商业化用途，包括但不限于：
   1) 通过网络或其他途径进行付费使用或分发；
   2) 在商业培训、广告或其他商业活动中使用本PPT的内容。
4. **法律责任**：任何违反上述条款的行为，作者保留追究法律责任的权利，包括但不限于：
   1) 要求停止侵权行为；
   2) 追究侵权使用者的经济赔偿责任。
5. **其他规定**：
   1) 本使用条款的解释权归本PPT作者所有。
   2) 作者保留随时更新本使用条款的权利，更新后的条款将即时生效。

谢谢大家！