01生物信息学简介

高建召

大纲

- 课程简介
- 生物信息学简介

学习生物信息学的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

课程简介

- 教师： 高建召
  - 数学科学学院327室
  - 飞书
  - 邮箱： gaojz@nankai.edu.cn

- 考试成绩：
  - 平时成绩40%+期末成绩60%

课件参考资源

- http://xue.biocuckoo.org/course.html （薛宇华中科技大学）
- https://liulab-dfci.github.io/bioinfo-combio/ （刘小乐-哈佛大学）
- http://readiab.org/introduction.html# （python脚本）
- https://www.ncrnalab.org/courses/#bioinfo3 （鲁志-清华大学）
- https://github.com/ossu/bioinformatics （一些关于生信课程资源）

参考教材

- 生物信息学（第二版），樊龙江，2021，科学出版社。
- 生物信息学：序列与基因组分析 （第二版），David Mount，2006，科学出版社
- 生物序列分析：蛋白质和核酸的概 率论模型，Richard Durbin等，清华大学出版社
- 生物信息学（第2版），李霞等主编， 2016，人民卫生出版社
- 机器学习，周志华 著，2016，清华大学出版社

其他参考书目

- Python for bioinformatics, Sebastian Bassi, CRC Press
- 统计学习方法，李航，清华大学出版社
- 深度学习，[美] 伊恩·古德费洛 等，人民邮电出版社
- 分子进化与系统发育, [美] Masatoshi Nei, Sudhir Kumar, 高等教育出版社
- 生物统计学基础, [美] 伯纳德·罗斯纳，科学出版社
- 结构生物信息学, [美] P.E.波恩 & H. 魏西希，化学工业出版社
- 生物信息学与功能基因组学，[美] 乔纳森·佩夫斯纳，化学工业出版社
- 生物信息学（第二版），陈铭 等译，科学出版社

课外阅读

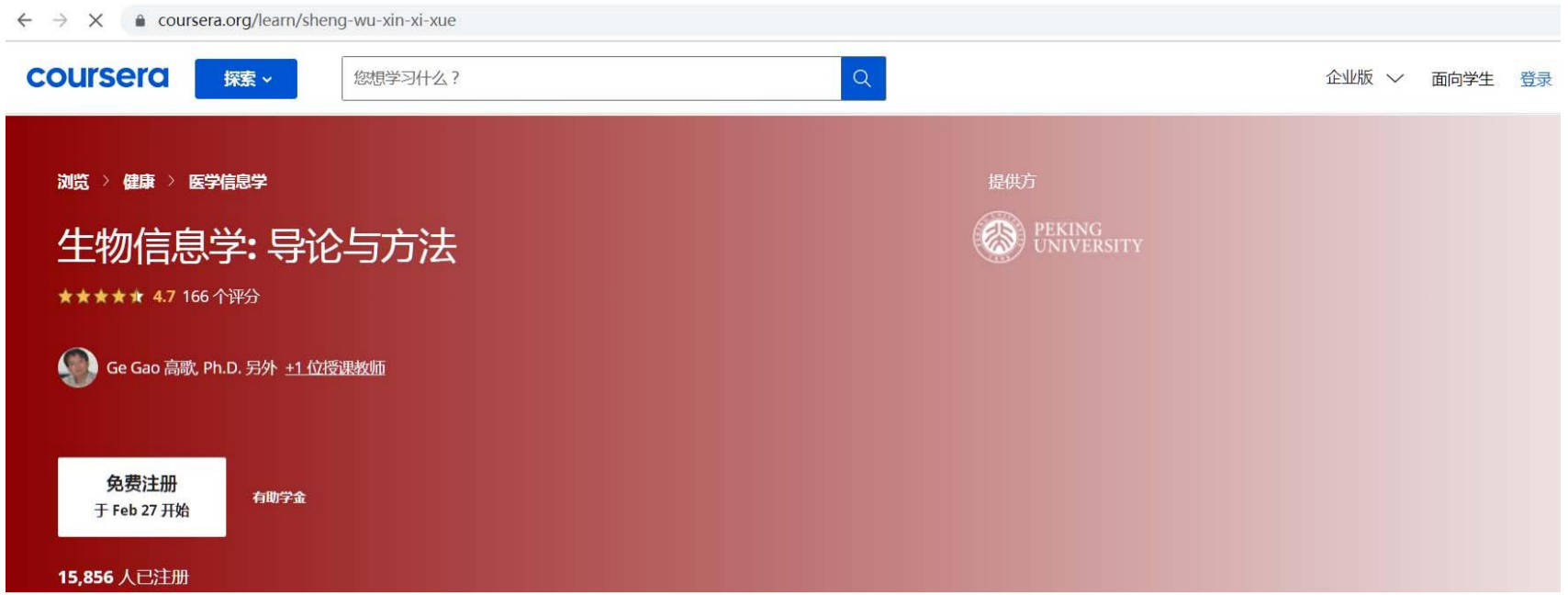
- 人工智能简史，尼克，人民邮电出版社
- 生物信息学札记，樊龙江
- 女士品茶，萨尔斯伯格，中国统计出版社
- 生命是什么，[奥] 埃尔温·薛定谔 ，湖南科学技术出版社
- 自私的基因，道金斯，中信出版社
- 世界观：科学史与科学哲学导论，理查德·德威特，电子工业 出版社
- 生命科学史，[美] 洛伊斯·N·玛格纳，上海人民出版社
- 生命的语言，[美] 弗朗西斯·S.柯林斯，湖南科学技术出版社
- 生命的线索，约翰·苏尔斯顿 等，中信出版社
- 师从天才，[美] 罗伯特·卡尼格尔，上海科技教育出版社

相关知识基础

- 生物学背景：细胞生物学、分子生物学、结构生物学
- 分子进化理论：MP, NJ, ML…
- 生物统计学
- 计算能力/编程能力：Perl/Python, R, PHP+MySQL,JAVA…
- 机器学习/深度学习算法

网上课程（1）

- 生物信息学：导论与方法
  - 教师：高歌、魏丽萍，北京大学
  - https://www.coursera.org/learn/sheng-wu-xin-xi-xue

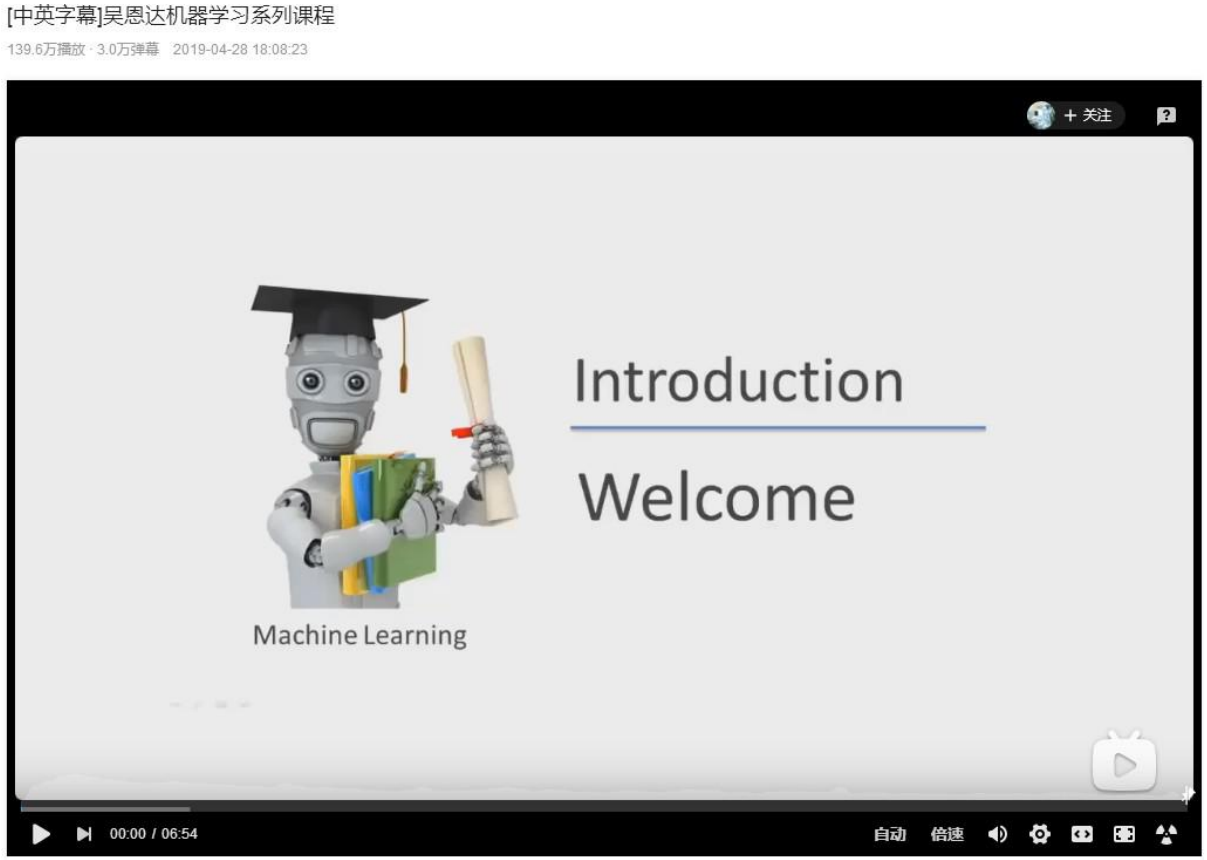


网上课程（2）

- 李宏毅2020机器学习深度学习
  - https://www.bilibili.com/video/BV1JE411g7XF

网上课程（3）

- 吴恩达机器学习系列课程
  - https://www.bilibili.com/video/BV164411b7dx



网上课程（4）

❒Keras深度学习快速简明教程

- https://www.bilibili.com/video/BV1gE411R7jd

Keras深度学习快速简明教程 最易学的深度学习入门课程 人人都可以学的人工智能入门

4.1万播放 · 222弹幕 2019-09-30 11:00:50

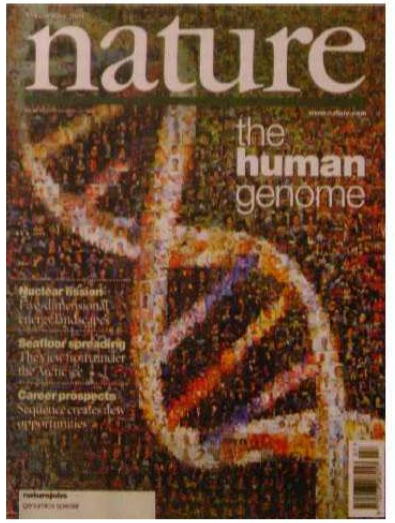
生物信息学发展历程

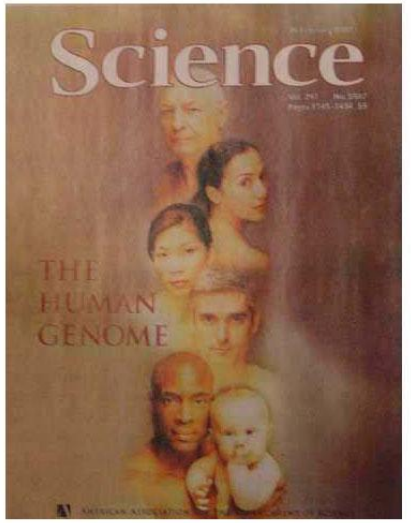
早期生物信息学发展的主要事件

- 1962 Pauling 提出分子进化理论
- 1965 Dayhoff构建蛋白质序列数据库
- 1970 Needleman-Wunsch算法、
- 1977 第一个基因组（吞噬体 ΦX174）被测序
- 1978 “bioinformatics” 出现
- 1981 Smith-Waterman算法
- 1982 GeneBank成立
- 1987 邻接法（进化方法）
- 1988 NCBI，EMBnet创立
- 1990 数据库搜索工具BLAST发布
- 1997 PSI-BLAST(BLAST系列程序之一）发布
- 1998 多细胞线虫基因组测序完成。

早期生物信息学发展的主要事件

- 1999 果蝇基因组测序完成
- 2000 拟南芥基因组测序完成
- 2001 人类基因组草图公布
- 2002 UCSC Genome Browser
- 2005 第二代高通量测序仪（454）面世
- 2007 ChIP-Seq 技术
- 2008 RNA-Seq技术
- 2010 第三代高通量测序仪（PacBio）面世





左一Craig Venter克雷格·温特

右一 Francis Collins佛兰西斯·柯林斯

人类基因组基本数据

1Gb (Gigabases) = 1 000 000 000bp = 10^9bp

- 人类基因组大约有30多亿（3Gb）个碱基对（A,C,G,T）.
- 大约20000个基因；蛋白质的编码序列，只占总长度的约1.5%。
- 清楚功能的3%; 不清楚功能约97%（“暗物质”）。

已测序最大的动物基因组(2021-1-25)：澳大利亚肺鱼：Australian lungfish (Neoceratodus forsteri) 430亿碱基对（43Gb）

Genomic record breaking: Largest animal genome sequenced – The Oxford Scientist (oxsci.org)

生物信息学发展历程

(1)萌芽期(60-70年代):

以Dayhoff的替换矩阵和Neelleman-Wunsch算法为代表，它们实际组成了生物信息学的一个最基本的内容和思路：序列比较。它们的出现，代表了生物信息学的诞生(虽然“生物信息学”一词很晚才出现)；

(2)形成期(80年代):

以分子数据库和FASTA等相似性搜索程序为代表。在这一阶段，生物信息学作为一个新兴学科已经形成，并确立了自身学科的特征和地位；

(3)基因组测序时代(90年代-至今):

以模式基因组测序与BLAST为代表；

(4)高通量测序时代（2005—）：

以第二和三代测序技术和基因组重测序为代表。

生物信息学定义

什么是生物信息学?

- **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)

生物信息学研究内容和发展方向

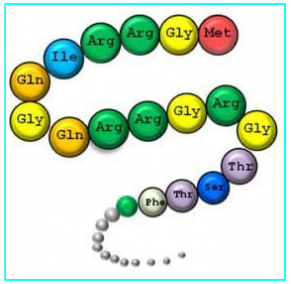
生物信息学主要研究内容与方向

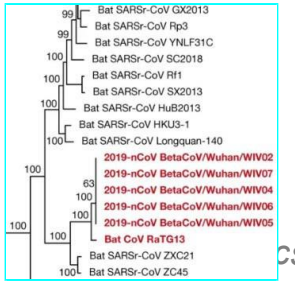
- (数据)开发、设计相关的数据库和工具，能够方 便有效的获取、管理以及使用各种类型的 数据和信息
- (算法)开发新的算法及统计学方法来揭示大数据 之间的联系
- (应用)分析和解释各种类型的生物学数据，包括 核酸、氨基酸序列、蛋白质功能结构域以 及蛋白质三级结构等

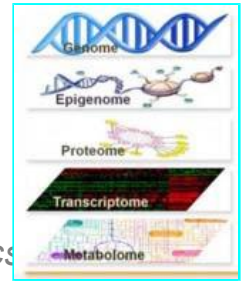
http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/milestones.html

生物信息学的数据类型

- 序列数据：**DNA、RNA、蛋白质序列**
- 结构数据：**NMR、X-ray、Cryo-EM/Cryo-ET**
- 遗传/进化距离数据
- 谱数据：基因芯片、蛋白质-蛋白质相互作用
- 影像数据: 图像、视频（**CT、MRI**、超声）
- 文本数据：科学文献、电子病历
- 混合数据：二代测序数据（序列、谱）







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. 人工智能生物学

相关学会

- 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

相关期刊

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/

生物信息学研究者具备基本条件

- 具备分子生物学的核心知识
- 分子生物学的中心法则知道得一清二楚。
- 至少1到2个用于序列分析的主要分子生物软件了如指掌。
- 用计算机命令行环境下轻松工作
- 能用C/C++计算机语言或Python或Perl脚本语言进行编程。

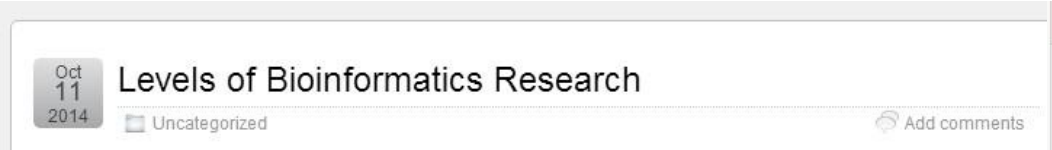
--摘自：Gibas, Jambeck 《Developing Bioinformatics Computer Skills》

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！



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

http://blog.sciencenet.cn/blog-404304-834869.html

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/

测序仪

测序仪：荧光自动测序仪（第一代）；高通量测序仪（第二代）

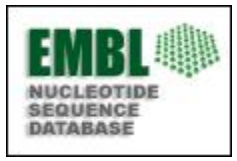
测序仪

2012年2月英国牛津纳米孔公司发布公告称，两年内将推出U盘测序仪产品MinION（第三代测序仪），个人基因组的测序将在15分钟内完成。2014年初产品进入试用期，价格为1000美金左右。截止2018年，全球共售出近万台。目前实际测序长度超过150kb。

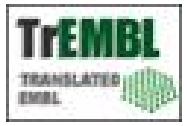
生物数据库

1. 核酸数据库

国际核苷酸序列数据库



2. 蛋白质数据库



=

3. 蛋白质结构数据库

4. 其他专项生物数据库

End