生物信息学：大数据科学

这些数据KB~TB量级

你处理的数据可能来自于：

人脸

虹膜

指纹

掌纹

语音信号

视频信号

图像，文本，矩阵

卫星图像

自然场景

运动场景

自动驾驶

医学影像

文字文本

我们的数据TB~PB量级

我们的数据：

基因组

细胞

披着“生物”外衣的ACGT字符串序列

我们要做的是：

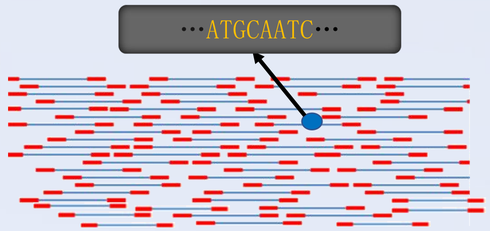
海量字符串的检索、比对与分析

生物信息学：算法密集型科学

我们挖掘

PB级数据

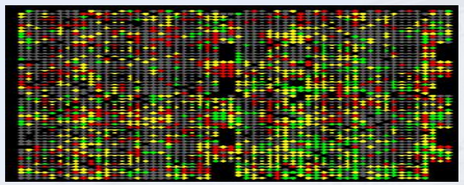
- 字符串序列索引、比对与分析



Set=$\{read_1 read_2, \ldots, read_k\}, k \epsilon (10^9, 10^{14})$

$read \epsilon \Sigma \{A, C, G, T\}^n, 200 < n < 10000, n \epsilon N^+$

- 大型特征矩阵分析与分类



$$\begin{bmatrix} 010000 & \cdots & 000000 \\ 000000 & & 100000 \\ \vdots & \ddots & \vdots \\ 100000 & & 000000 \\ 010000 & \cdots & 000000 \end{bmatrix}$$

我们创造

PB级数据结构

新后缀树

新图索引

新哈希

分布式数据结构

内外存混合结构

…

稀疏数组

稀疏矩阵

…

PB级数据处理算法

- **新**动态规划算法
- **新**贪心算法
- **新**分治算法
- **新**排序算法
- **新**搜索算法
- **新**近似算法
- **新**随机算法

……

- 启发式算法
- 亚线性算法
- 分布式计算
- 机器学习

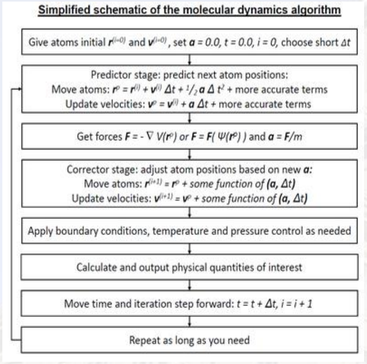
……

生物信息学：改变世界的科学

诺贝尔奖中的生物信息学算法

化学分子结构数据分析算法

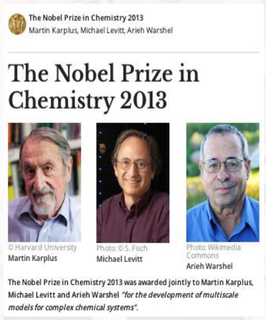
图论+分治算法



(2013年诺贝尔奖)

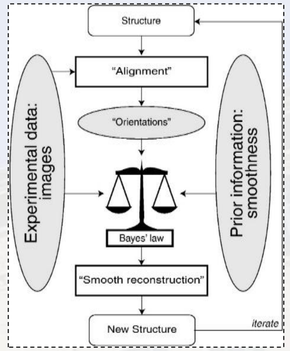
诺奖评审委员表示：

他们把化学实验搬进了计算机虚拟世界，计算机和试管变得同样重要



冷冻电镜数据分析算法

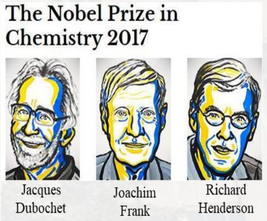
图论+并行算法



(2017年诺贝尔奖)

诺奖评审委员表示：

依靠该技术，很快细胞在我们面前将再无秘密



生物信息学：改变世界的科学

BLAST: 走向诺贝尔奖的字符串查找算法（如下，查找字符串N在字符串M中的位置）

字符串N = CGAGCATGCAG……　　长度：10^2~10^3

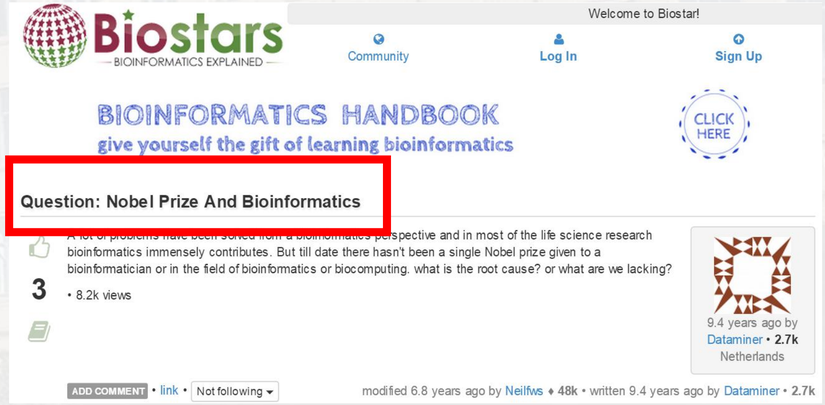
字符串M = ACGAGCAATTGCCGCCGTTTTCAGCAGTTATGTCATC……　　长度 >> 3×10^9

算法核心：哈希+动态规划

引用总数：83437次

BLAST算法被Nature列入：
20世纪科学计算发展的里程碑

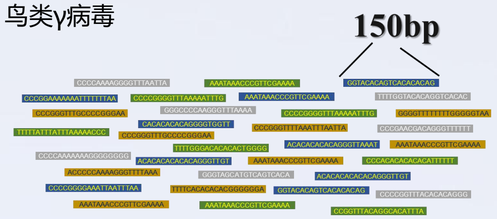
Question: Nobel Prize And Bioinformatics

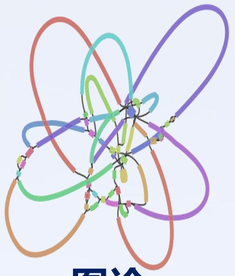


生物信息学：改变世界的科学

抗击新冠

新冠病毒　老鼠β病毒
人类α病毒　鸟类γ病毒





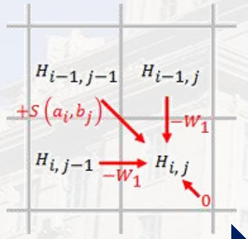
图论

聚类

短序列拼接

……ATGCAATCAGTAACGGTCCA……

长度为30000bp的新冠病毒基因组



动态规划算法
计算序列相似性

物种间相似序列

新冠独有特征序列

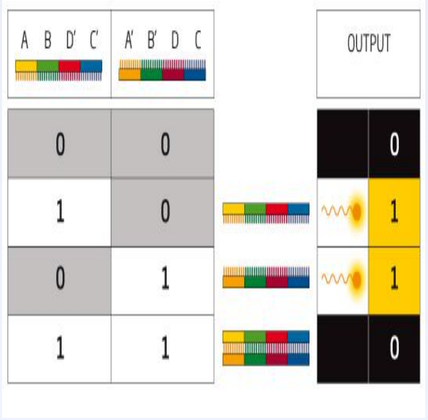
有效识别新冠独有特征序列及蛋白
病毒检测试剂盒

生物信息学：通往未来的科学

生物信息学正在颠覆传统，创造新的财富

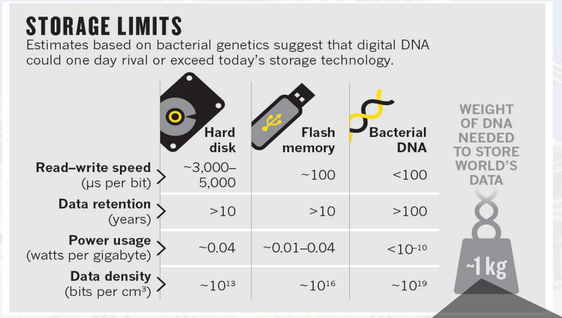
DNA计算机

通过DNA编码和匹配，进行海量计算操作



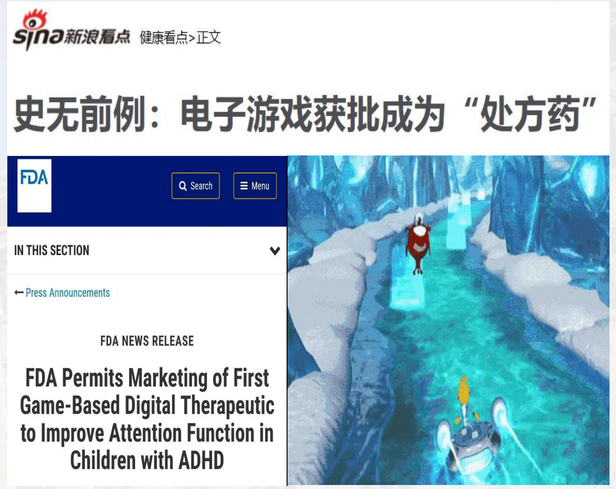
DNA存储

光盘将由苹果来代替



一公斤DNA可以存储全世界的数据

游戏处方药



胡安·恩里克斯 著

过去的财富由0/1世界创造，

未来***将进入ACGT世界创造财富的时代***。

生物信息学：通往未来的科学

比尔•盖茨直言：

- “如果让他重新选专业，他会选择这三个专业：AI，Energy，和Biosciences。”

生物信息学满足了比尔•盖茨的两个愿望：

AI + Biosciences

马云直言：

- “中国下一个首富，一定在大健康领域。”

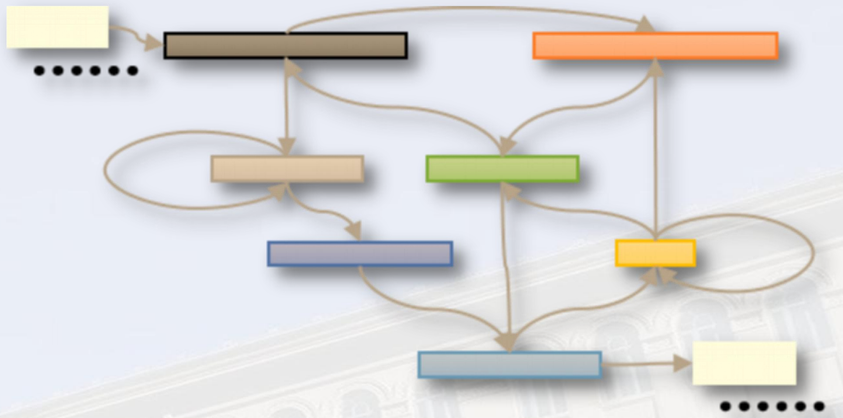
生物信息学：通往未来的科学

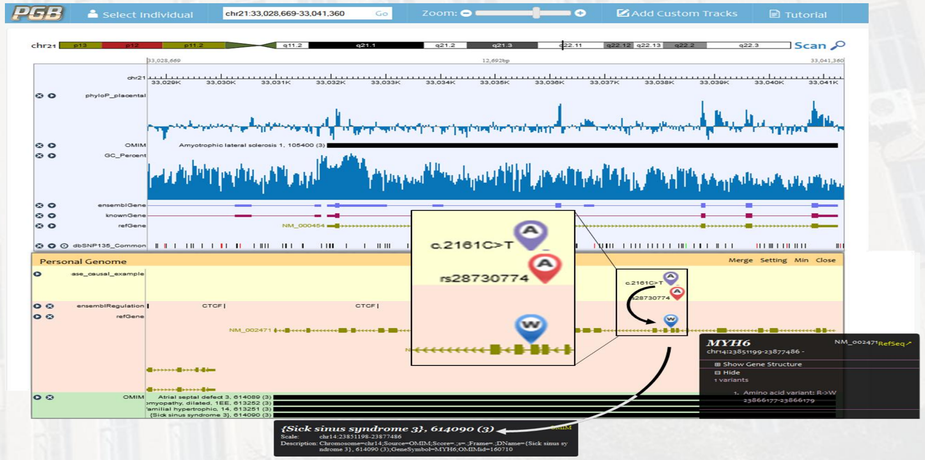
- 基因测试的先驱：23andme
- 新药研发工具：ADImab
- 云端基因数据库：DNAnexus
- 远程医疗的平台：DoctorOnDemand
- 肿瘤大数据探索者：FlatironHealth
- 基因疗法：iPierian
- 医保大数据的典范：Predilytics
- 胰岛素使用的颠覆者：RaniTherapeutics
- 自闭症检测的突破者：SynapDx
- 研究高花费的终结者：Transcriptic
- 医疗模式变革先驱：Onemedical
- 基因测癌能上市：FoundationMedicine

我们创造了世界领先的算法和系统

代表性学术成果

- 基因组序列非线性表示理论、索引和比对算法
- 融合序列与家系遗传信息的罕见变异检测算法
- 基于生物网络的疾病非编码micro-RNA预测模型
- 个人基因组变异和家系信息的基因组可视化算法





我们创造了世界领先的算法和系统

国际同行的学术评价

学术成果（一）

Michael Schatz
约翰霍普金斯大学教授

在**Nature Review Genetics期刊**发文，将完成人提出的算法列为**基因组片段比对的9个代表性算法之一**

Table 2 (cont.) | Bioinformatics methods for long-range sequencing and mapping

Bioinformatics analysis	Approach	Selected methods	URL	Refs
Long-read mapping	Seed-and-extend method using clusters of short seeds	BLASR (P)	https://github.com/PacificBiosciences/blasr	75
		BWA-MEM (O and P)	http://bio-bwa.sourceforge.net/	76
		DALIGNER (P)	https://github.com/thegenemyers/DALIGNER	43
		GraphMap (O and P)	https://github.com/isovic/graphmap	141
		Kart (O and P)	https://github.com/hsinnan75/Kart	142
		LASMSA (O and P)	https://github.com/hitbc/LAMSA	143
		LAST (O and P)	http://last.cbrc.jp/	80
		Minimap2 (O and P)	https://github.com/lh3/minimap	77
		NGMLR (O and P)	https://github.com/philres/nextgenmap-lr	73

评价出处：
Nat Rev., 2018, 19(6): 329-346.

Jan Fostier
比利时根特大学教授

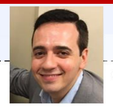
完成人提出的算法是**当前最先进的比对算法之一**

Graph aligner tools

The performance of BrownieAligner is compared with the state-of-the-art graph aligners BGREAT and deBGA. A de Bruijn graph is first constructed from the reference genome, followed by the alignment of reads to the cor-

评价出处：
BMC Bioinformatics, 2018, 19:311

学术成果（三）



Dario Veneziano
俄亥俄州立大学教授

完成人提出了**一种创新性的方法**，该方法是**第一个**本体驱动的lncRNA功能分析方法

Finally, an innovative approach in the functional annotation of lncRNAs is provided by Jiang et al. (2015). *LncRNA2Function* provides the first ontology-driven user-friendly web system based on the idea that similar expression patterns across multiple conditions may share similar functions and biological pathways. The tool functionally annotates a single or a set of lncR-

评价出处： Front Bioeng Biotechnol. 2015; 3: 77

Gorodkin Jan
哥本哈根大学教授
生物信息中心主任

将完成人的数据集**作为金标准**，用于评估其方法的性能

To benchmark our method for predicting miRNA–disease associations, we used the manually curated gold standard set of direct associations between miRNAs and diseases from Jiang *et al.* (2010). To allow for direct comparison with the datasets described earlier in the text, we mapped the miRNAs and diseases to miRbase and Disease Ontology identifiers, respectively. After mapping, the set consisted of 236 associations among 92 miRNAs and 48 diseases.

评价出处： Bioinformatics, 2014, 30(3): 392–397

学术成果（三）

Evan Eichler
美国科学院院士

美国科学院院士Evan Eichler领导的国际自闭症基因组计划**利用完成人提出的算法进行罕见变异检测**，发现自闭症易感基因和变异，结果发表在权威遗传学期刊**AJHG**上

bands versus seven in siblings). We initially identified 5,609 de novo SNVs and indels by using DNMFilter (70.1 ± 12.0 variants per individual). De novo variants showed evidence of nonrandom clustering (Figure S5), and the number of new mutations significantly correlated

评价出处： AJHG, 2016, 98, 58–74

Cynthia Hawkins
多伦多儿童医院教授

多伦多儿童医院**Cynthia Hawkins教授**在脑癌基因组研究中**利用完成人提出的算法进行结构变异检测**，发现脑癌相关结构变异，研究成果发表在权威遗传学期刊**Nature Genetics**上

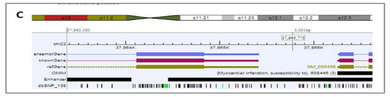
duplicates were removed from alignments using Picard 1.77 (**Supplementary Table 6**). Structural variants were identified using PRISM 1.1.6 and PRISM CTX 1.0.1 (ref. 26). Recurrent structural variants (**Supplementary Table 7**) were found after subtracting common structural variants[27] and all structural variants found in normal tissue and were visually validated using the DNAC

评价出处： Nat Genet, 2014, 46, 451–56

学术成果（四）

Wang Junwen
美国梅奥医学中心

国际权威生物数据库**GWASdb**
将完成人提出的个人基因组可视化方法**用于可视化重要的基因组注释信息**



or flat files indexed by Tabix (41). We used Circos (42) to generate global GWASdb v2 SNPs Manhattan plot and personal genome browser (PGB) (43) to display important annotation tracks.

评价出处： Nucleic Acids Res, 2016, D1, 869–876

Ioannis Iliopoulos
希腊克里特大学教授

将完成人提出的方法与Ensembl、UCSC Genome Browser等经典方法并列为**代表性的基因组可视化工具**

Table 5 Genome browsers

Tool and references	Description	URL	Web-based or standalone
...			
Ensembl [105, 106]	Annotation, analysis and display of vertebrates and other eukaryotic species	http://www.ensembl.org/	Web-based
...	tool for interactive genomic datasets	http://www.broadinstitute.org/igv/	Standalone
The personal genome browser [292]	Visualizing functions of genetic variants	http://www.pgbrowser.org/	Web-based
UCSC cancer genomics Browser [251, 252]	Integration of clinical data	http://genome-cancer.ucsc.edu/	Web-based
UCSC Genome Browser [103]	Rapid linear visualization, examination and querying of the data at many levels	http://genome.ucsc.edu/cgi-bin/hgGateway	Web-based

评价出处： GigaScience, 2015, 4:38

顶尖的生物信息学基础设施

- 拥有生物大数据教育部重点实验室
- 拥有PB级基因组大数据研究平台
- 计算资源：10000CPU核，5PB存储
- 测序仪：10台Illumina NovaSeq

生物大数据教育部重点实验室

基因组大数据计算平台

大规模基因组测序平台

一流的生物信息学人才培养能力

朱世佳(2007-2012)
美国德克萨斯大学助理教授
纽约大学博士后

滕明祥(2008-2012)
美国Moffitt癌症中心助理教授
哈佛大学博士后

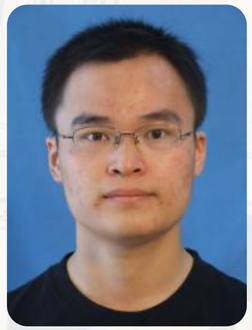
胡杨(2007-2013)
哈尔滨工业大学副教授
哈佛大学博士后

彭佳杰 (2010-2015)
西北工业大学副教授
哈佛大学博士后

姜玥(2008-2013)
加拿大多伦多儿童医院研究员
多伦多大学博士后

谭仁杰(2011-2017)
哥伦比亚大学
博士后

初砚硕(2013-2019)
MD Anderson
癌症中心博士后



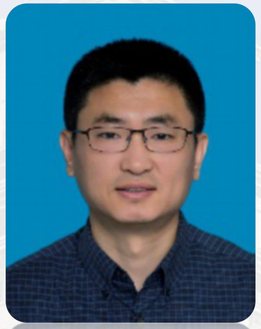
詹青(2011-2019)
阿普杜拉国王大学
博士后

王荣杰(2012-2018)
香港城市大学
博士后

一流的生物信息学人才培养能力

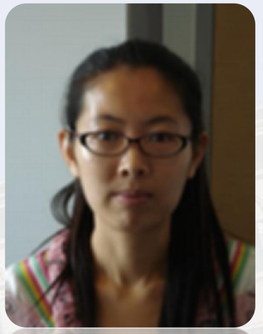
王纪轩
多伦多大学
攻读博士

苏俊豪
香港大学
攻读博士



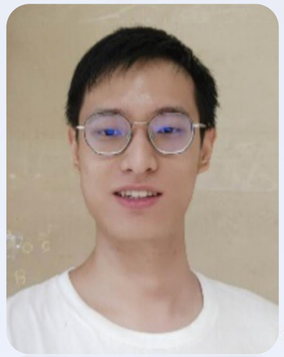
张新军
普渡大学
攻读博士

奇克
佐治亚理工
攻读博士



郝阳阳
印第安纳州立大学
攻读博士

一流的生物信息学人才培养能力



聂晨晞
苏黎世联邦理工学院
攻读博士

朱迪显
美国爱荷华大学
攻读博士

付译磊
美国莱斯大学
攻读博士

哈工大生物信息技术研究院现有教师

哈工大生物大数据教育部重点实验室主任
哈工大医学与健康学院执行院长
哈工大生物信息技术研究院院长
中国十万人基因组技术首席科学家

臧天仪
教授

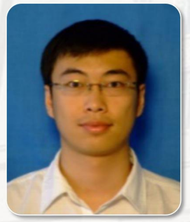
蒋庆华
教授

刘　博
教授

李　杰
副教授

胡　杨
副教授

李君一
副教授



隽立然
讲师

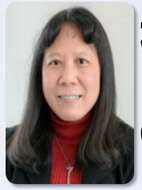
刘永壮
讲师

姜涛
讲师

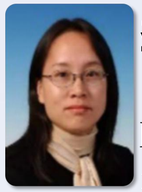
国宏哲
讲师

学术咨询委员会

主席 张学敏
中国科学院院士
NSFC医学部主任



林希虹
美国医学院院士
COPSS奖得主
哈佛大学



李蓬
中国科学院院士
NSFC生命学部主任
清华大学

陈润生
中国科学院院士
中科院生物物理所

赵国屏
中国科学院院士
中科院马普所

骆清铭
中国科学院院士
海南大学校长

王永雄
COPSS奖得主
斯坦福大学

詹启敏
中国工程院院士
北京大学副校长
医学部主任

金力
中国科学院院士
复旦大学副校长

顾东风
中国科学院院士
国家心血管病
中心副主任

梅宏
中国科学院院士
军事科学院副院长

Richard Durbin
英国皇家学会院士
剑桥大学

张学
中国工程院院士
哈尔滨医科大学
校长

饶子和
中国科学院院士
清华大学

李松
中国工程院院士
军事医学科学院