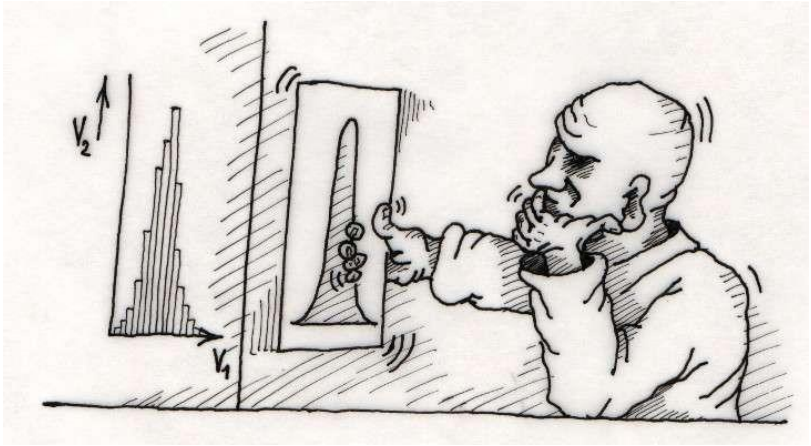
生物统计学：
生物信息中的概率统计模型

2019年秋



有关信息

- 授课教师：宁康
  - Email: ningkang@hust.edu.cn
  - Office: 华中科技大学东十一楼504室
  - Phone: 87793041, 18627968927

2019生物统计学

扫一扫二维码，加入该群。

- 课程网页
  - http://www.microbioinformatics.org/teach/#
  - QQ群：882140516

考评

课程成绩

=

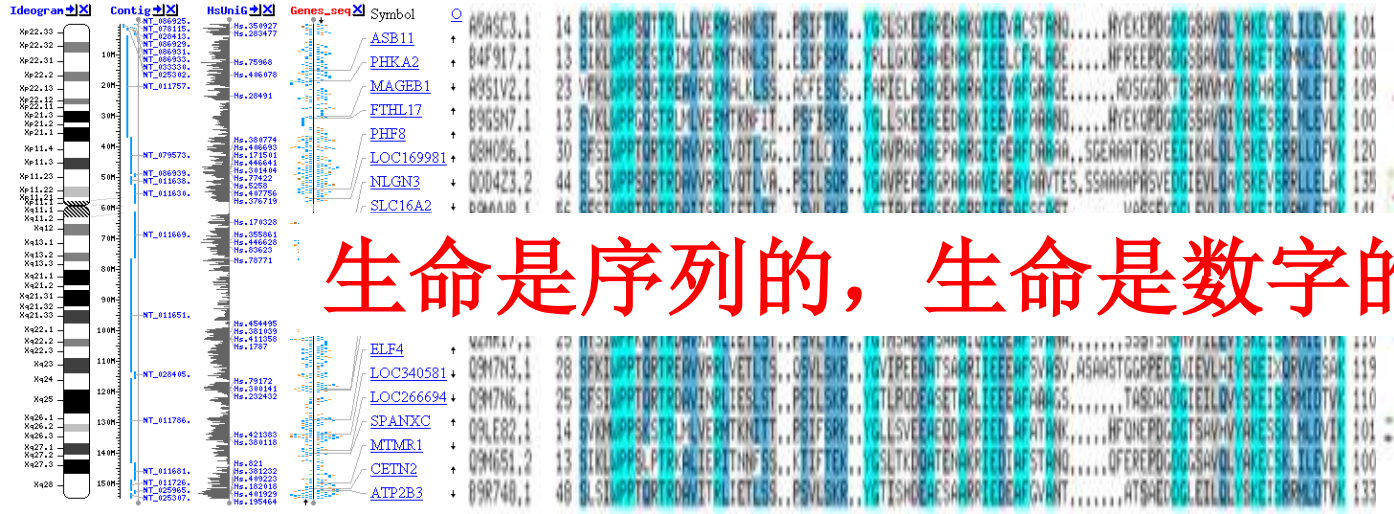
课堂讨论（10%）

+课后作业&随堂测验（20%）

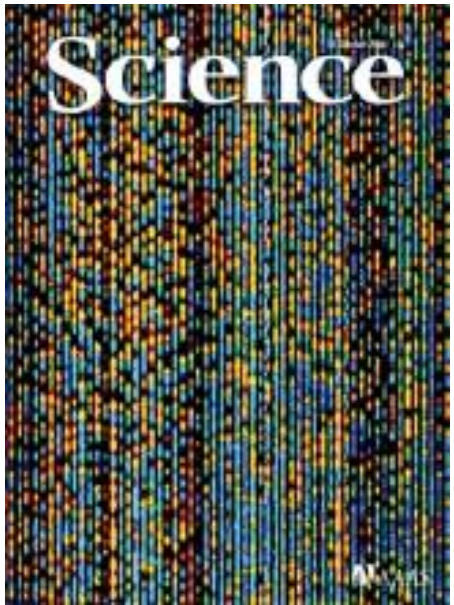
+终结性考试（70%）

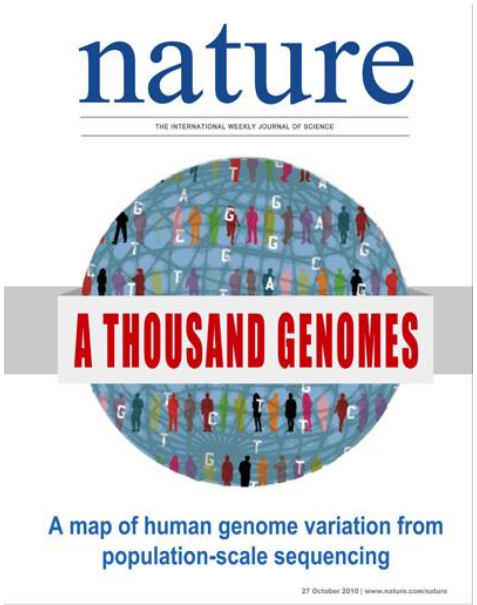
生物统计学：生物学视角

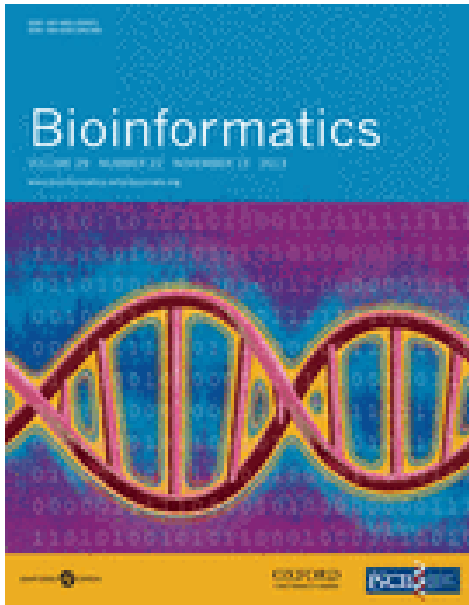
生物信息学@HUST



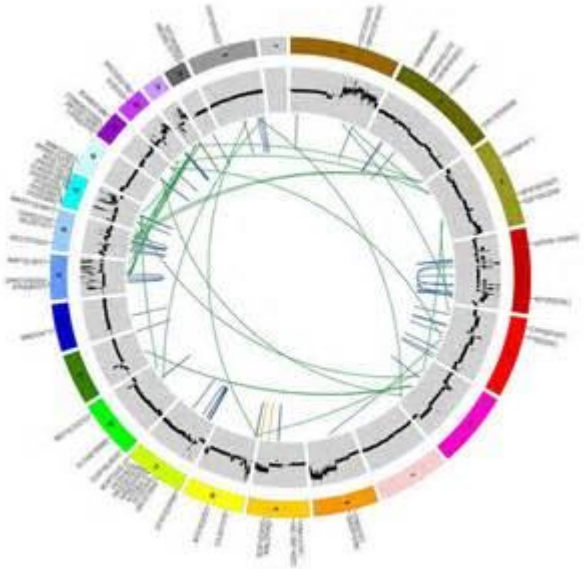
生命是序列的，生命是数字的！

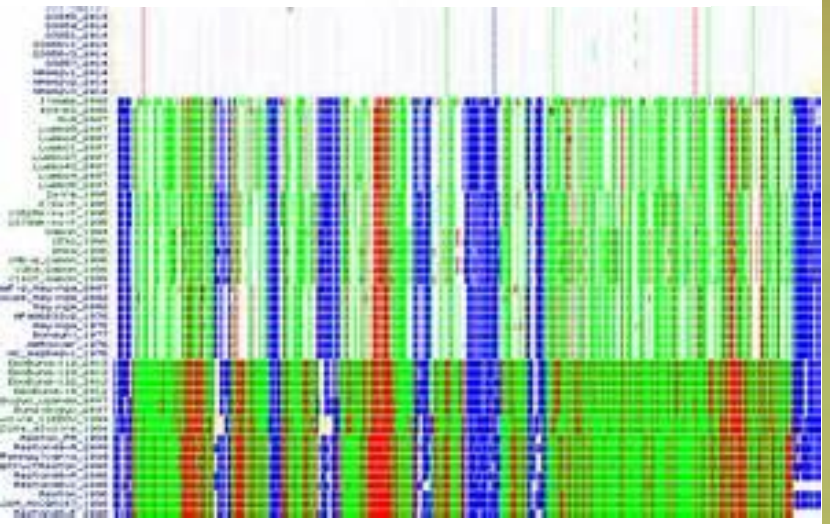


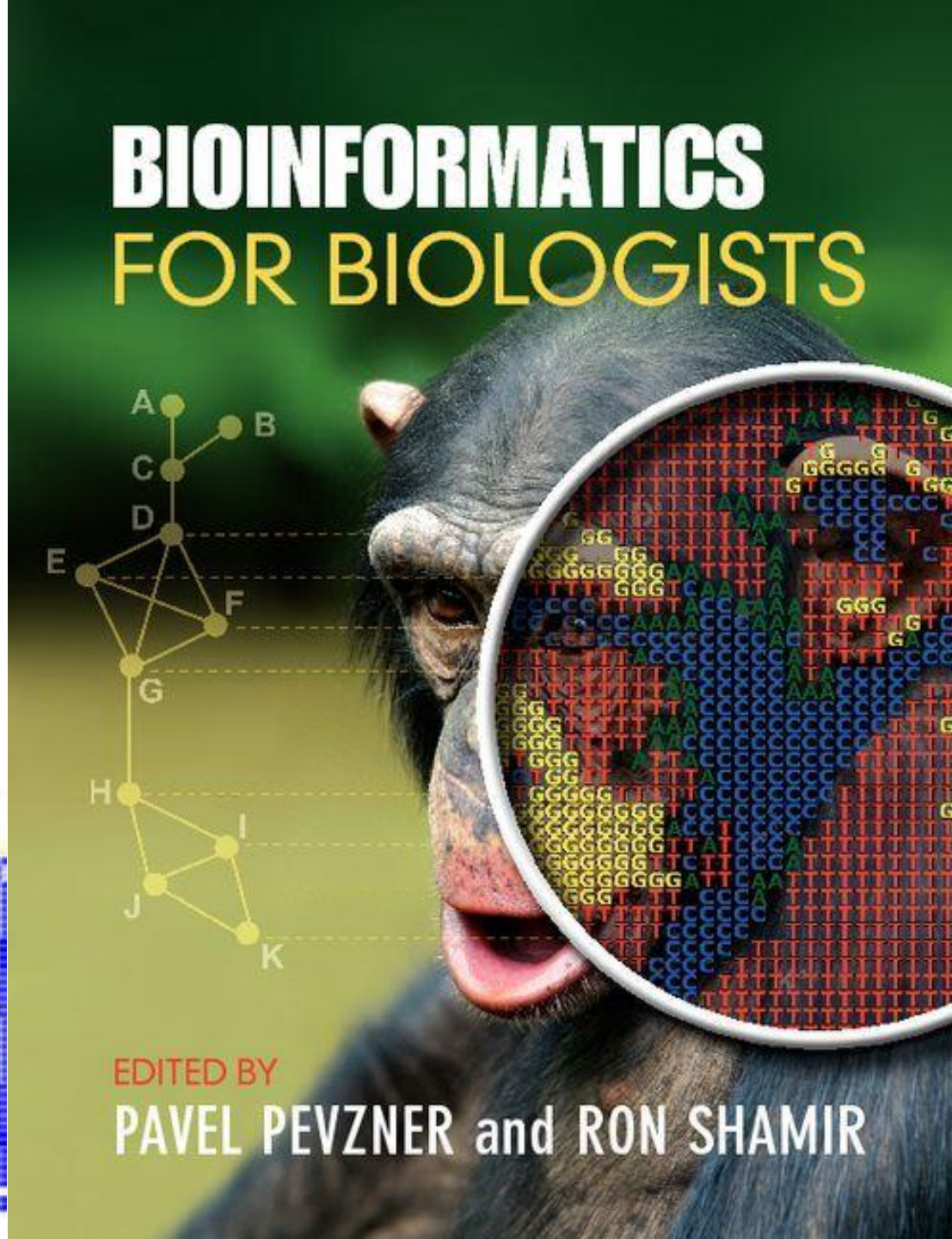




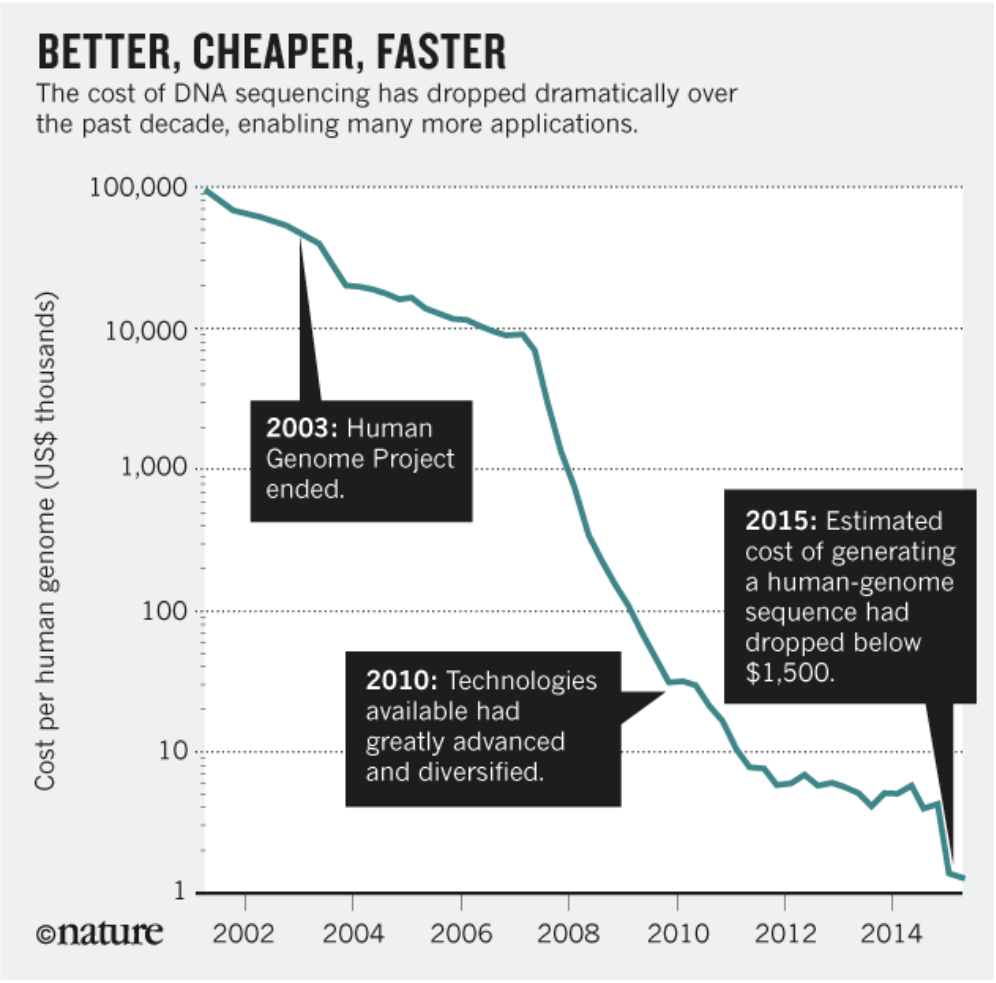
生命是序列的，
生命是数字的！

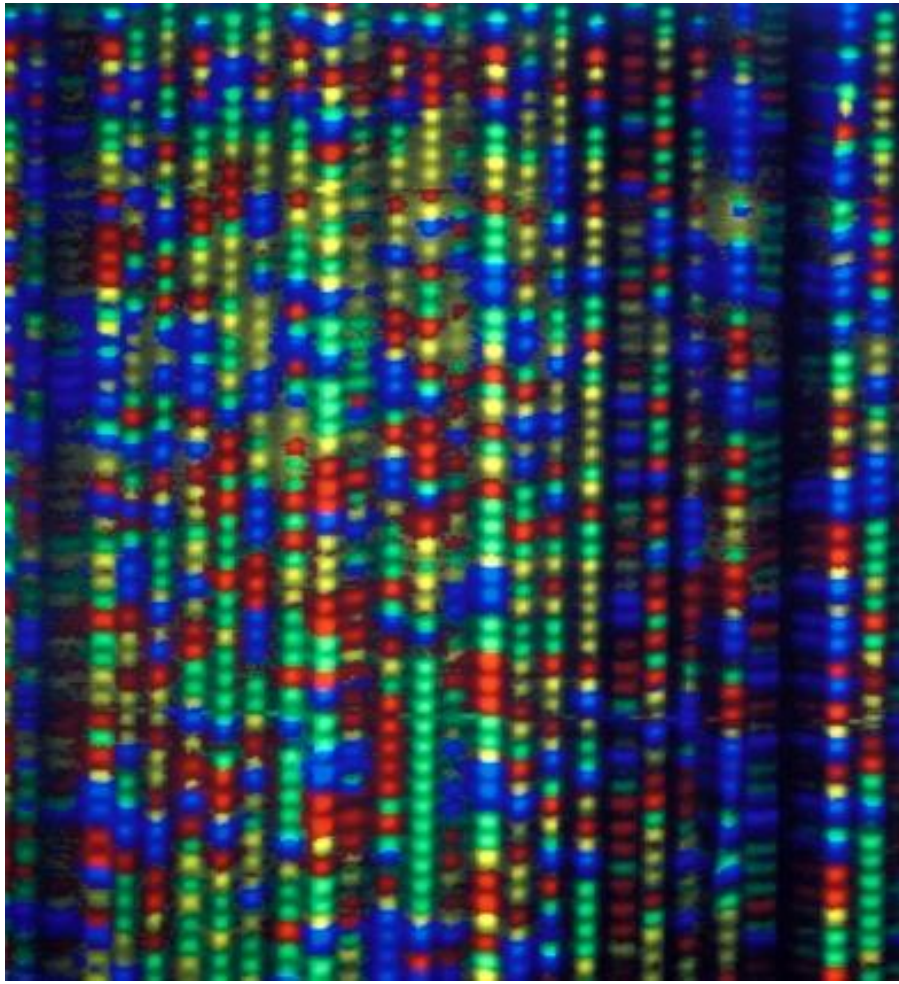




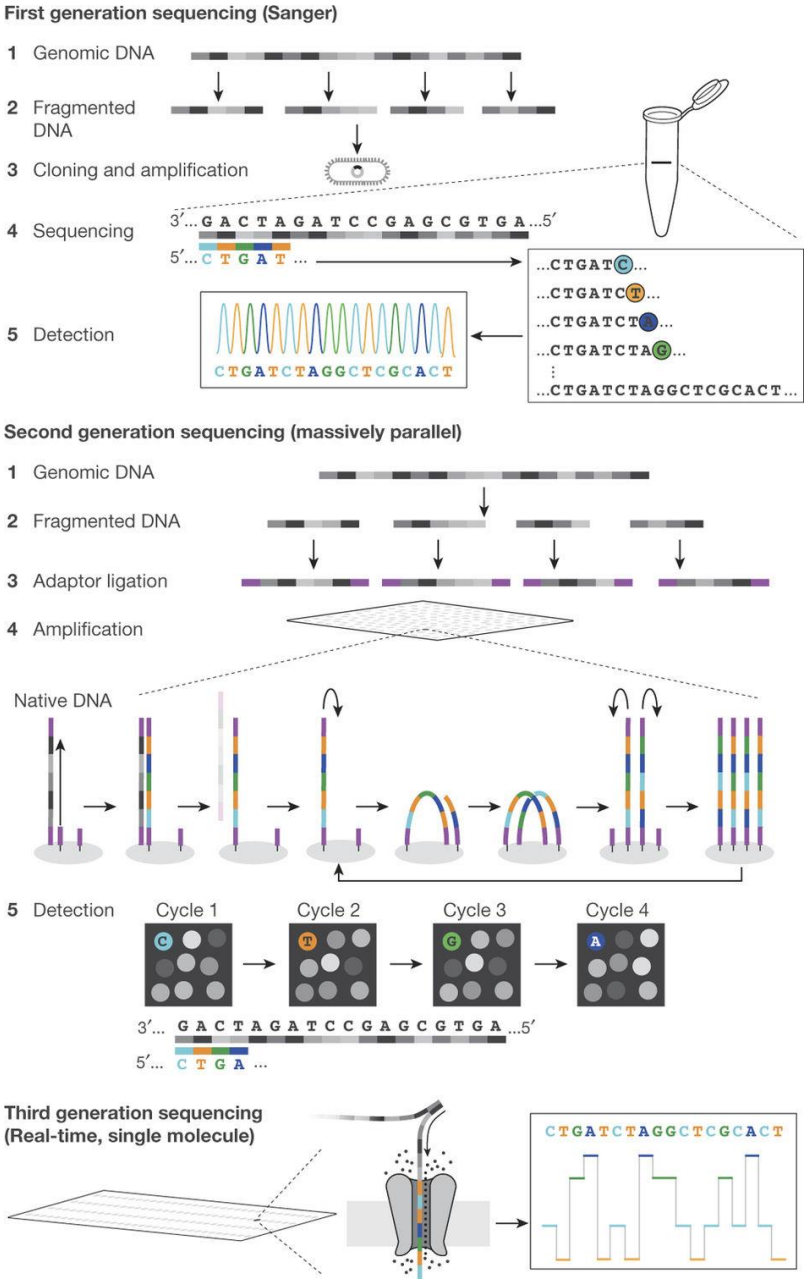


DNA sequencing and bioinforamtics

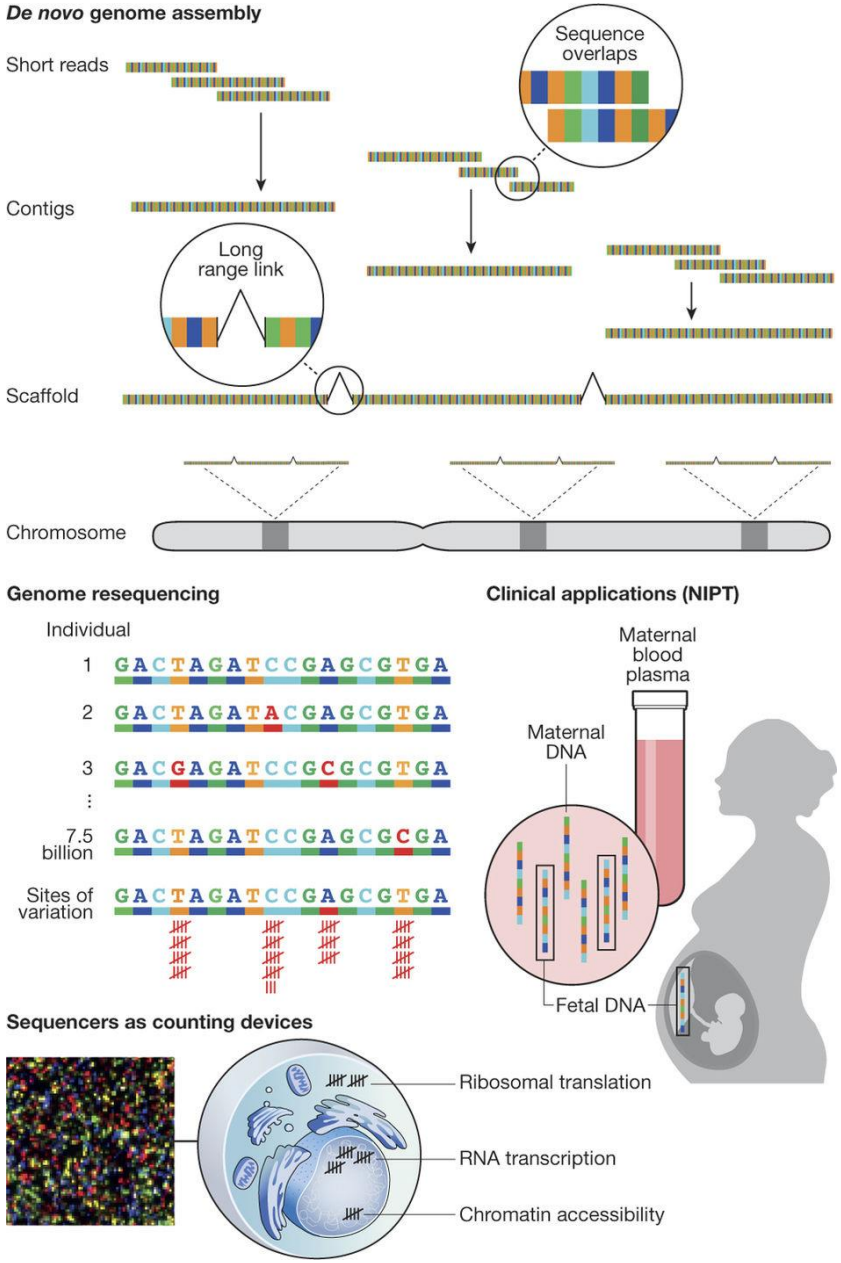




DNA Sequencing



Sequencing applications



在今天，DNA测序技术已经在诸多方面达到了临床应用的具体要求。

科学家估计，全世界每年大约有四百万到六百万孕妇正在通过外周血游离DNA对胎儿的21三体综合征进行诊断，而十年之内，这个数字将超过1500万。

在高收入国家，基因组测序已经广泛用于多种疾病的产前诊断，可以揭示大约30%的出生缺陷，同时，这一数字也正随着数据解读能力的成熟而逐渐上升。

在肿瘤学领域，液体活检在最近几年已经成为了肿瘤相关学术，产业以及投资界的新宠。基于DNA测序的液体活检被认为正逐步发展为癌症诊断与预后评估的标准方法，能够在可知的时间内逐步补充甚至取代传统的创伤性癌症诊断技术。

同样，手持DNA测序仪等设备的开发也使得流行病学家甚至能够在最为偏远的地区高效完成对人类样本，动物以及昆虫病原载体甚至是空气，水，食物的基因检测。

流行病学家和公共卫生专家也开始讨论如何通过对城市垃圾中微生物的DNA测序辅助传染性疾病的预防与控制。

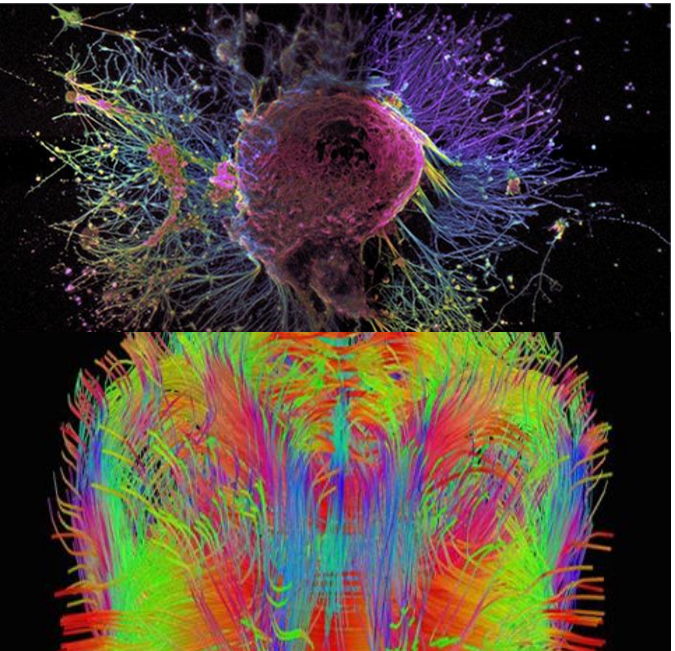
海洋生物学家也正在通过宏基因组学技术来对海洋的生态健康进行监测与研究。

在法医领域，便携式DNA测序仪可以将DNA分析带出法医实验室，使其成为一线警务工作人员的随身工具。帮助警方即使通过DNA监测确定嫌疑人，发展成为诸如酒精探测器一类的便捷工具。

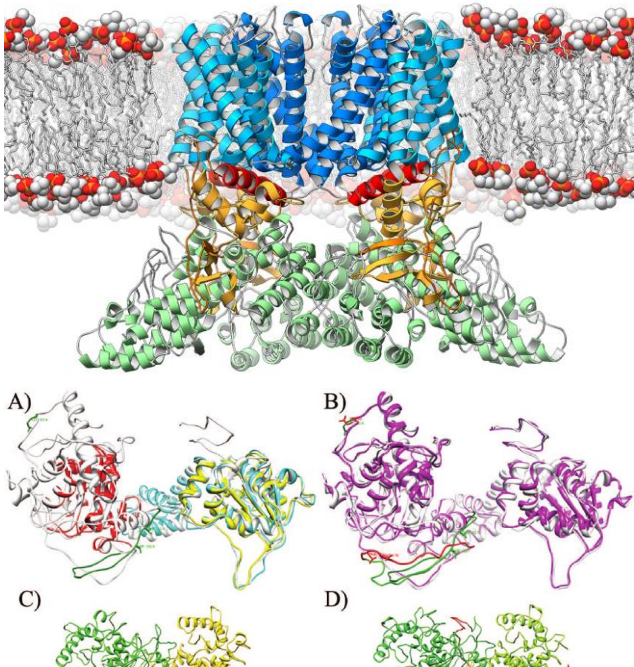
在人们的家里，DNA测序设备或许也可以成为下一个“智能”或“连接”设备，一些评论者甚至认定厕所是通过实时DNA测序监测家庭成员健康的理想场所。

生物信息学@HUST

生命不只是序列的，但是生命始终是数字的！

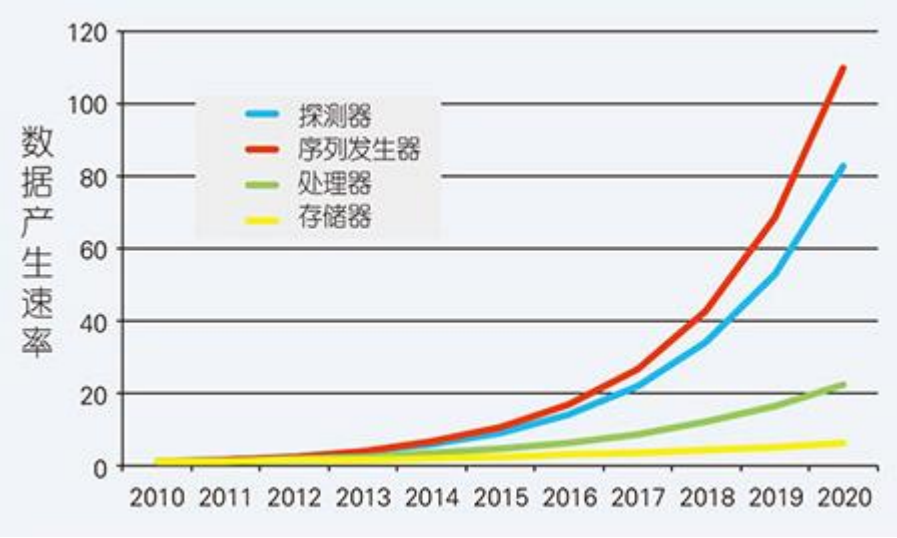


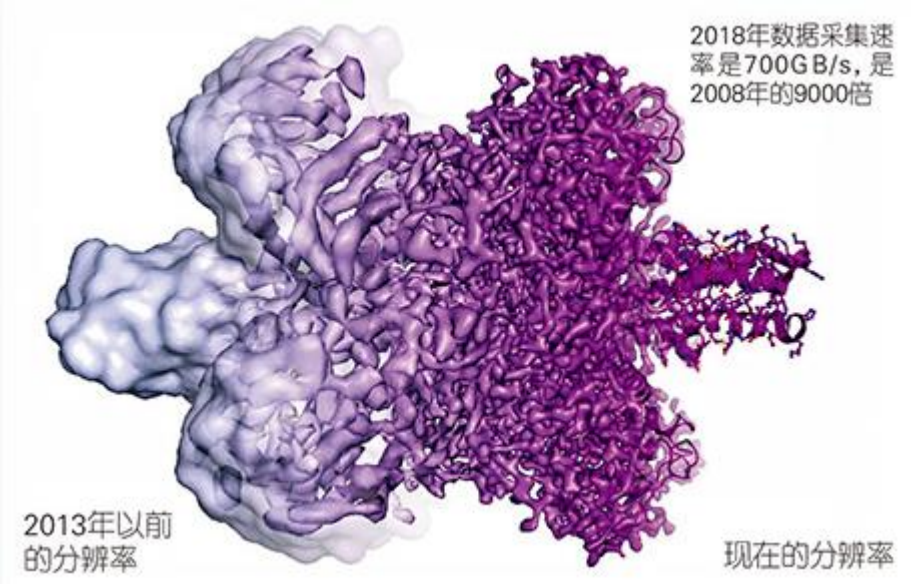
- 结构生物学（Structure biology）
- 生物图像（Bio-imaging）
- 。 。 。



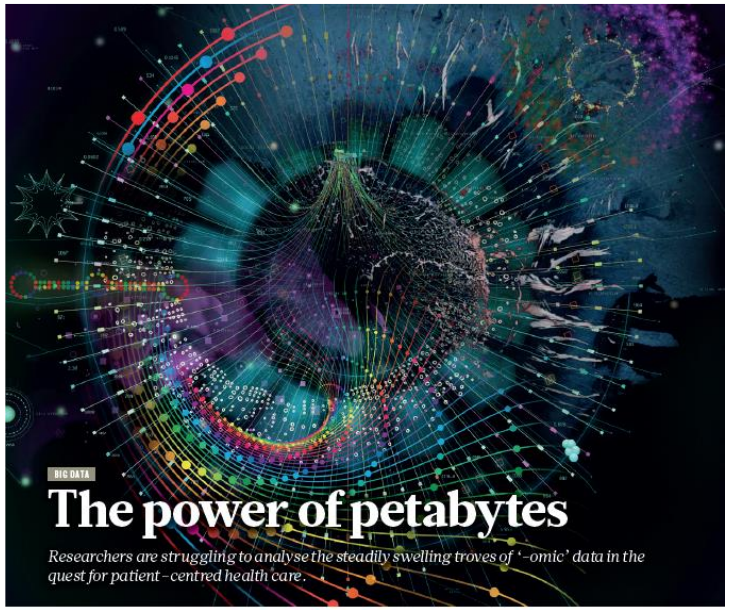
生物信息学@HUST

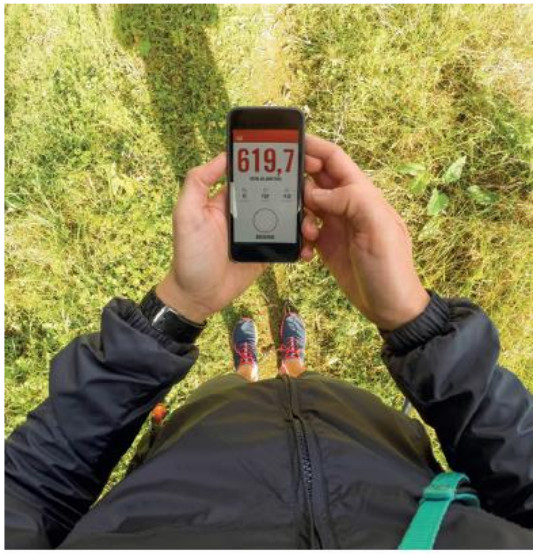
生命不只是序列的，但是生命始终是数字的！





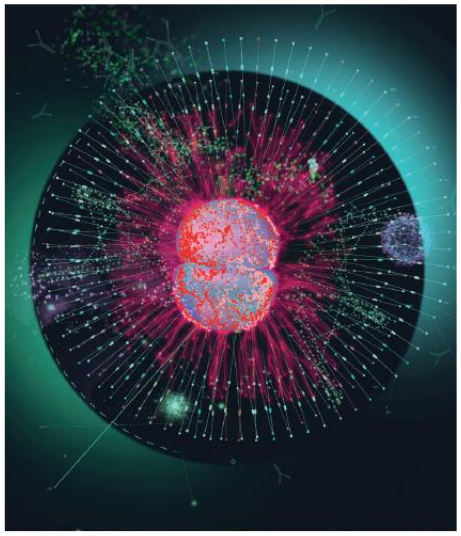
Big-data become popular…





Smartphone fitness apps enable researchers to gather health data from large numbers of people.

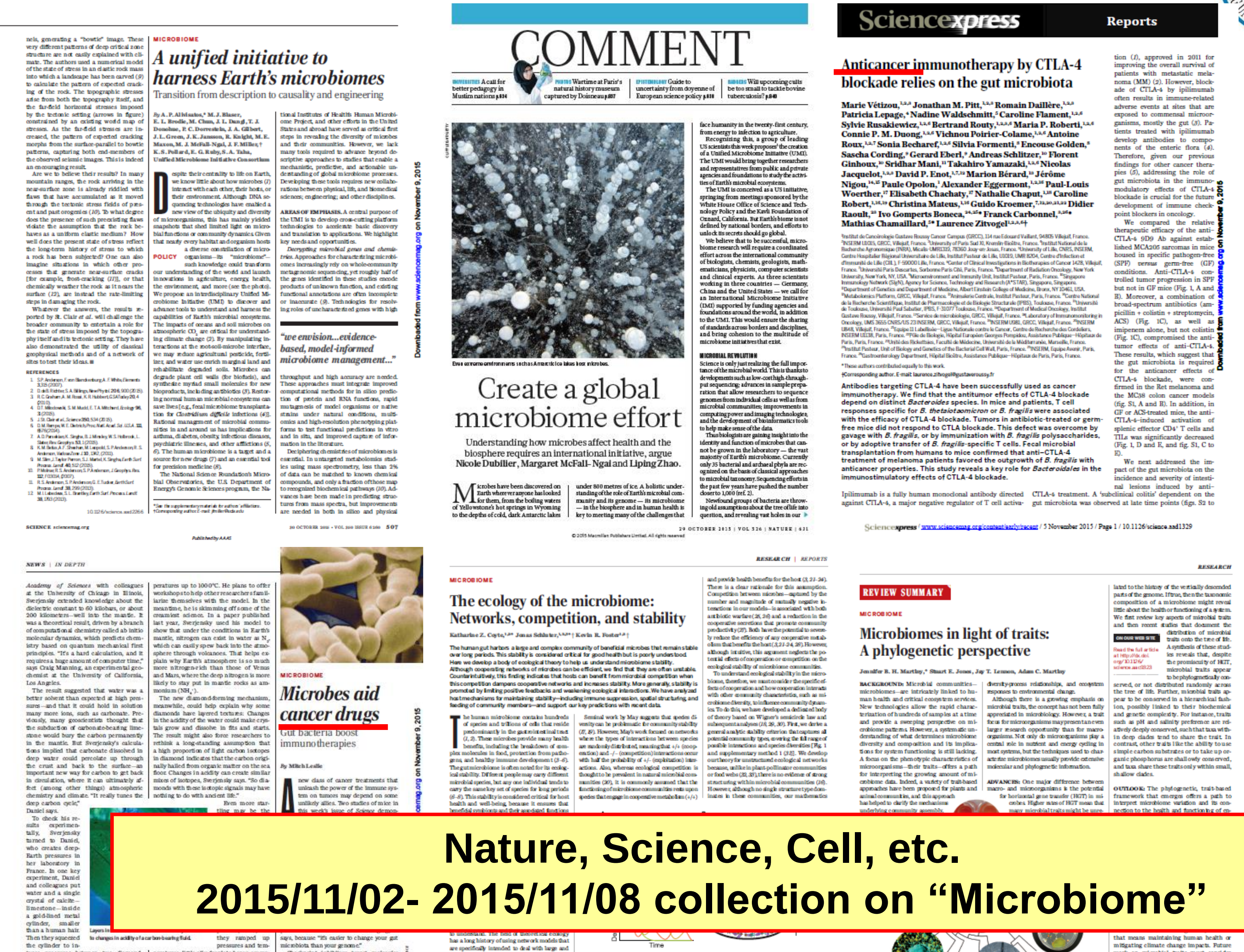
MOBILE DATA
Made to measure



COLLABORATIONS
Mining the motherlodes

Nature, 2015/11/05 collection on "Big-data in biomedicine"

Microbiome become popular as well…



Nature, Science, Cell, etc.

2015/11/02- 2015/11/08 collection on "Microbiome"

Microbiome and big-data…

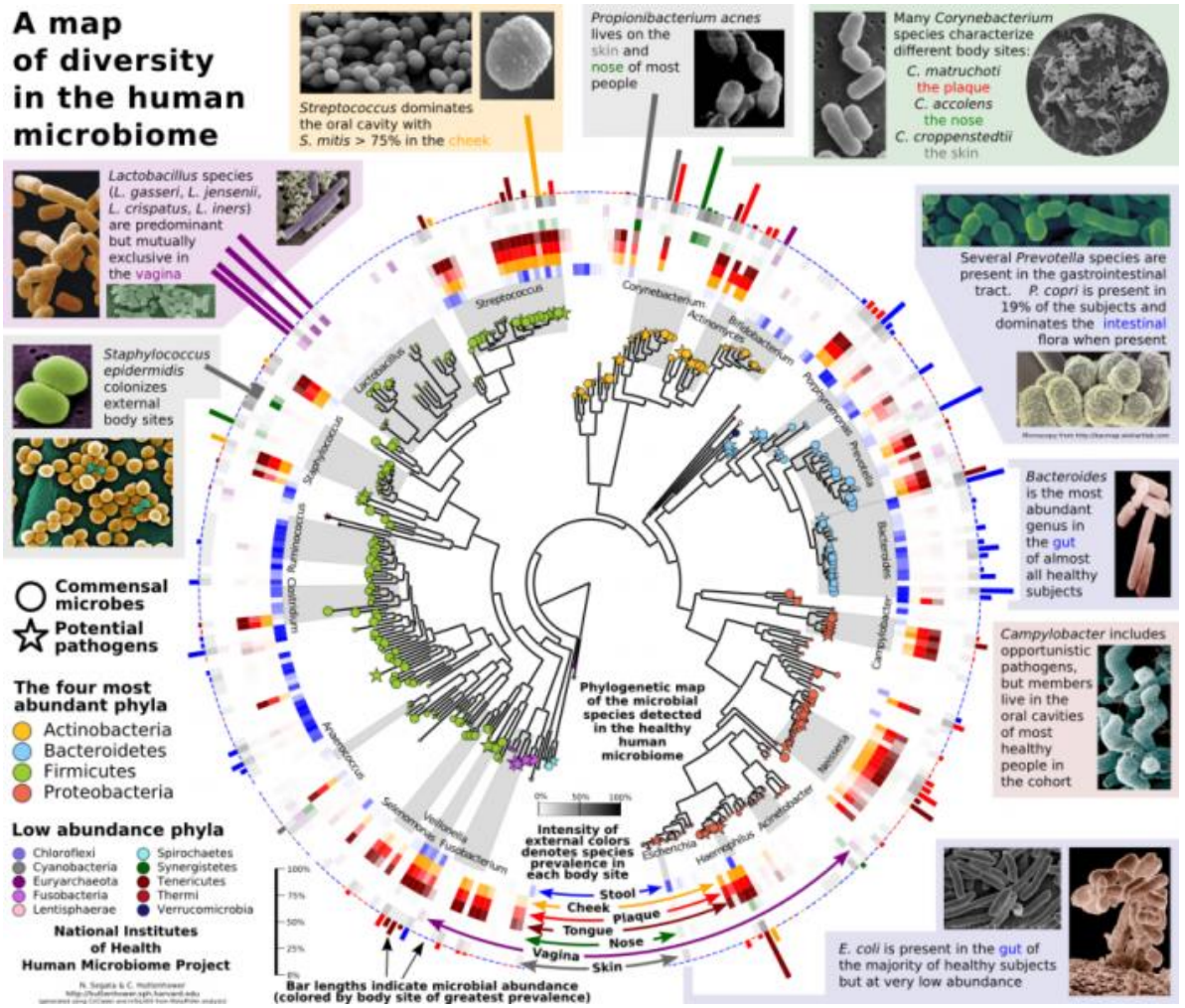
- A cell is already very complex
- 一个细胞已经非常复杂了

- A microbial community is much more complex than a cell
- 一个微生物群落就更为复杂了

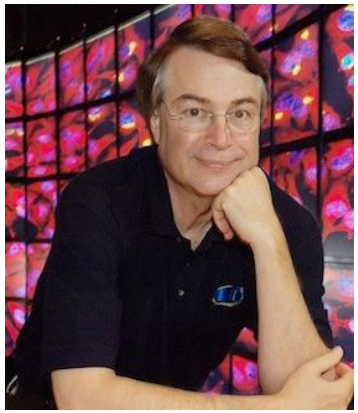
- But much more big-data
- 但是也代表了更多的数据

Microbiome and big-data…

在生物信息眼里，这全是大数据。。。



Microbiome and big-data…



Larry Smarr

Founding Director of the California Institute for Telecommunications and Information Technology (Calit2)

PUBLICATIONS

LARRY'S LATEST PAPERS

Large Memory High Performance Computing Enables Comparison Across Human Gut Microbiome Of Patients With Autoimmune Diseases And Healthy Subjects
Published in the XSEDE 2013 Proceedings of the Conference on Extreme Science and Engineering Discovery Environment: Gateway to Discovery. Article No. 25 (http://dl.acm.org/citation.cfm?doid=2484762.2484828)

Quantifying Your Body: A How-to Guide From A Systems Biology Perspective
Larry Smarr, Biotechnol. J. 2012, 7, 980-991

Supporting Information For Quantifying Your Body: A How-to Guide From A Systems Biology Perspective
Supporting Information for DOI 10.1002/biot.201100495

Essay: An Evolution Toward A Programmable Universe
Larry Smarr, Dec 5, 2011, The New York Times

Quantified Health: A 10-year Detective Story Of Digitally Enabled Genomic Medicine
Larry Smarr, with commentary by Mark Anderson, published as a Special Letter in the Strategic News Service Newsletter, September 30, 2011.

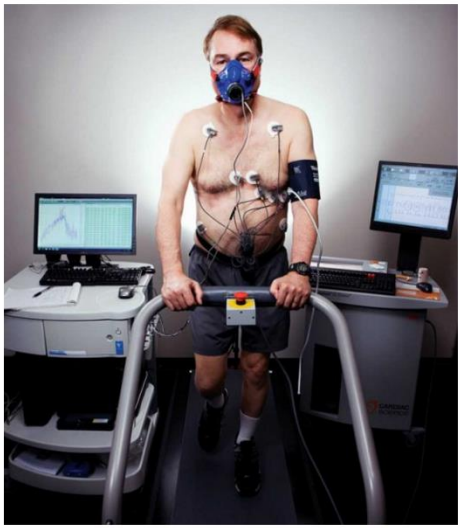
How I Improved My Health By Changing My Eating, Exercise, And Stress Management Habits: An Annotated Reading List
Larry Smarr, Requested by Mark Anderson, CEO Strategic News Service For Distribution to the Future in Review 2011 Attendees

Biomedicine

The Patient of the Future

Internet pioneer Larry Smarr's quest to quantify everything about his health led him to a startling discovery, an unusual partnership with his doctor, and more control over his life.

by Jon Cohen February 21, 2012

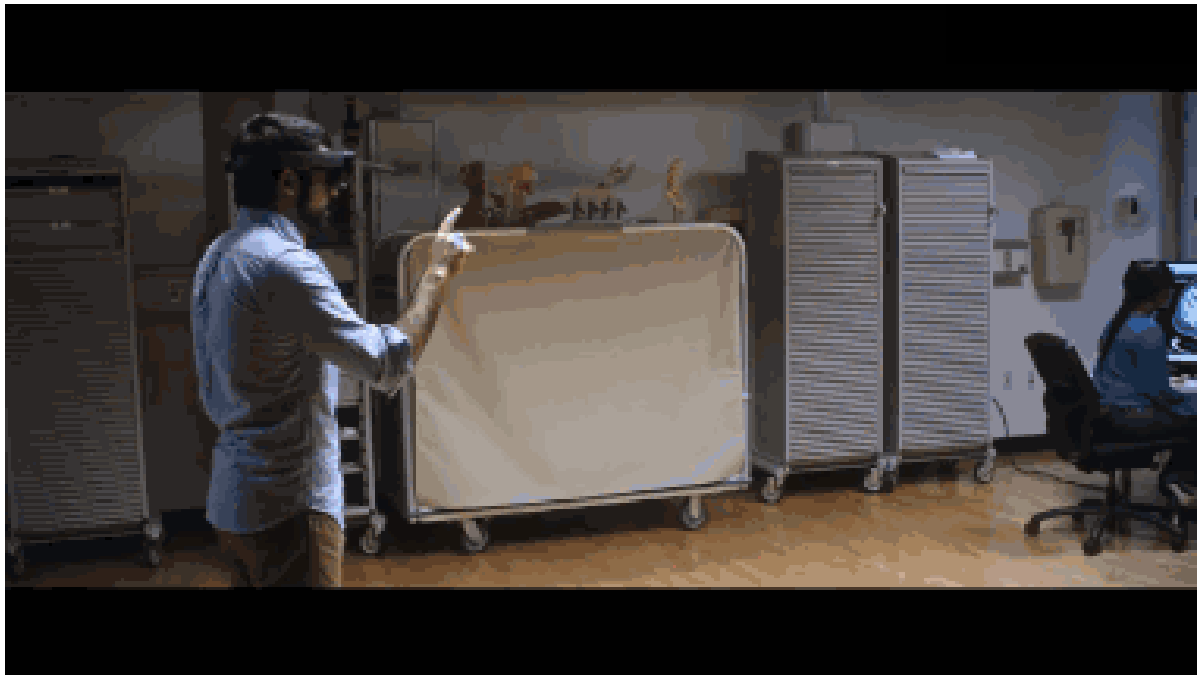


TEDMED Attend Speakers TEDMED Live Talks The Hive Partnerships About Blog

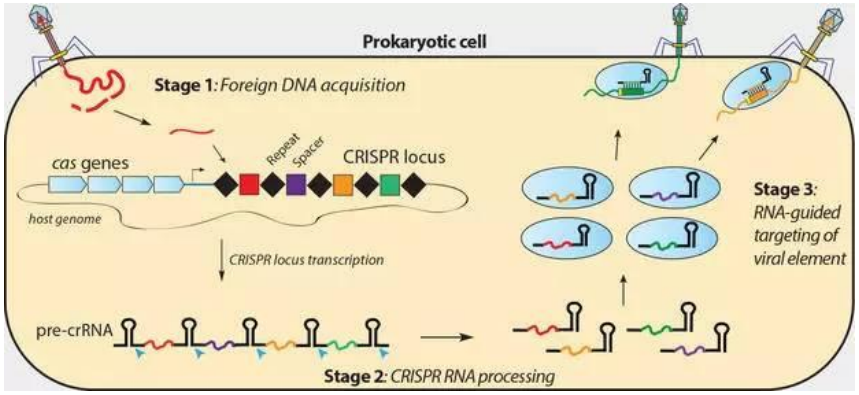
Biomedical big-data…

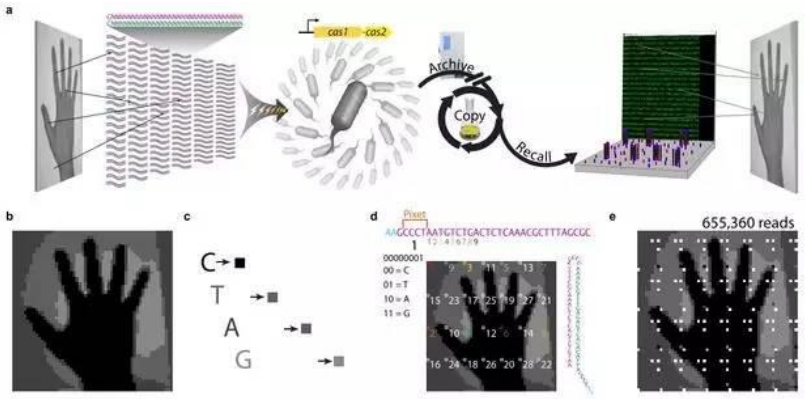
"Have you ever figured how information-rich your stool is?," Larry asks me with a wide smile, his gray-green eyes intent behind rimless glasses. "There are about 100 billion bacteria per gram. Each bacterium has DNA whose length is typically one to 10 megabases—call it 1 million bytes of information. This means human stool has a data capacity of 100,000 terabytes of information stored per gram. That's many orders of magnitude more information density than, say, in a chip in your smartphone or your personal computer. So your stool is far more interesting than a computer."

-- Larry Smarr



Understand it, create it!



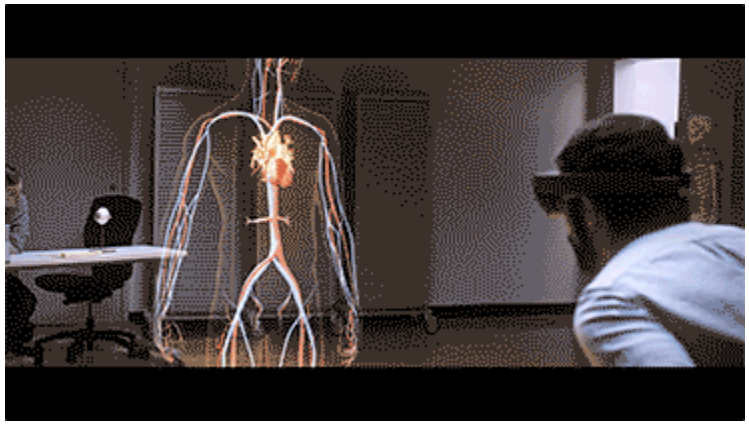


Original Image

原始图像

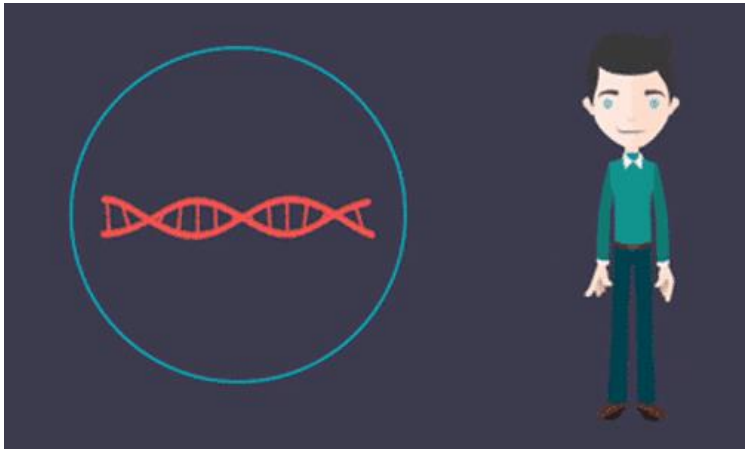
Image Reconstructed From Bacteria

从细菌DNA还原的图像



See it!

Understand it!



Create it!

生物统计学：计算科学视角

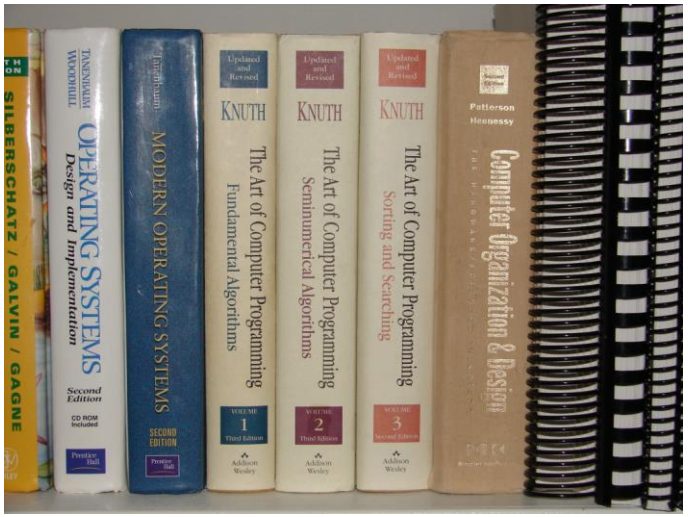
Donald Knuth（高德纳）

Donald Knuth, the "father of the analysis of algorithms.”

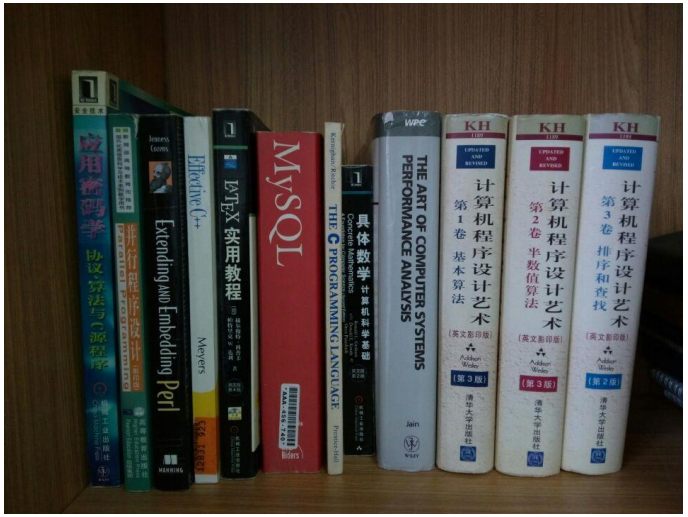
The Art of Computer Programming （计算机程序设计艺术）

Markup	Renders as
`The quadratic formula is $-b \pm \sqrt{b^2 - 4ac} \over 2a$ \bye`	The quadratic formula is $\frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$

“生物信息学为算法研究提供了500年的问题”– Don Knuth

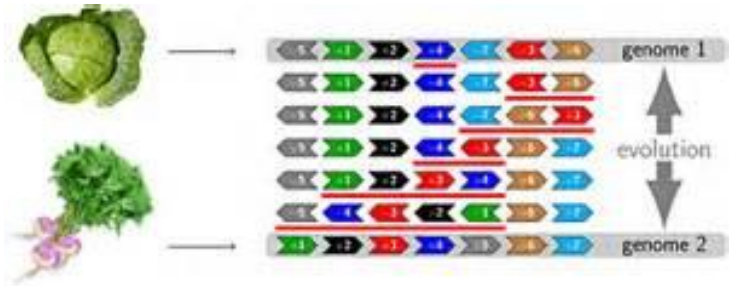


"definitely send me a résumé if you finish this fiendishly difficult book" – Bill Gates



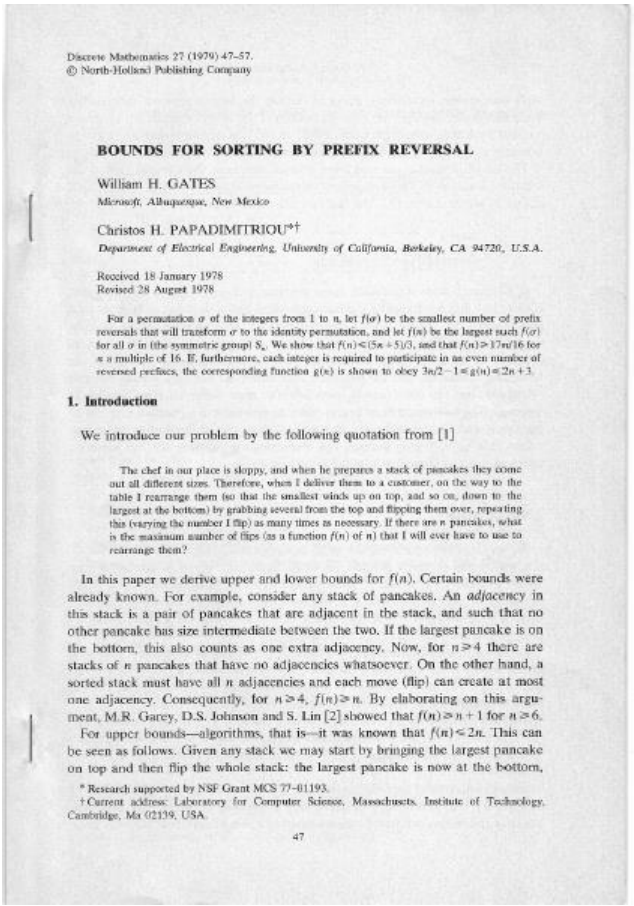
"definitely come to talk about algorithm if you read half of this book" – Kang Ning

Bill Gates（比尔盖茨）



比尔盖茨:下个世界首富出自基因检测领域

Sorting by reversal problem



Discrete Mathematics 27 (1979) 47–57.
© North-Holland Publishing Company

BOUNDS FOR SORTING BY PREFIX REVERSAL

William H. GATES
Microsoft, Albuquerque, New Mexico

Christos H. PAPADIMITRIOU*†
Department of Electrical Engineering, University of California, Berkeley, CA 94720, U.S.A.

Received 18 January 1978
Revised 28 August 1978

For a permutation σ of the integers from 1 to n, let $f(\sigma)$ be the smallest number of prefix reversals that will transform σ to the identity permutation, and let $f(n)$ be the largest such $f(\sigma)$ for all σ in (the symmetric group) S_n. We show that $f(n) \leqslant (5n+5)/3$, and that $f(n) \geqslant 17n/16$ for n a multiple of 16. If, furthermore, each integer is required to participate in an even number of reversed prefixes, the corresponding function $g(n)$ is shown to obey $3n/2 - 1 \leqslant g(n) \leqslant 2n+3$.

1. Introduction

We introduce our problem by the following quotation from [1]

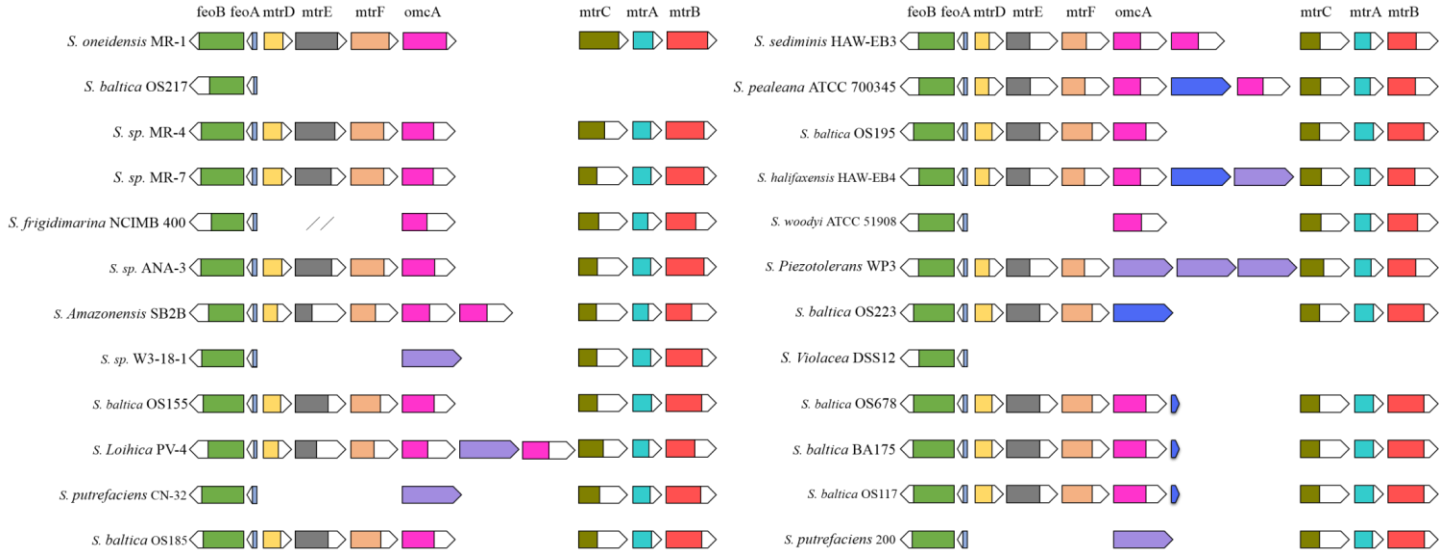
> The chef in our place is sloppy, and when he prepares a stack of pancakes they come out all different sizes. Therefore, when I deliver them to a customer, on the way to the table I rearrange them (so that the smallest winds up on top, and so on, down to the largest at the bottom) by grabbing several from the top and flipping them over, repeating this (varying the number I flip) as many times as necessary. If there are n pancakes, what is the maximum number of flips (as a function $f(n)$ of n) that I will ever have to use to rearrange them?

In this paper we derive upper and lower bounds for $f(n)$. Certain bounds were already known. For example, consider any stack of pancakes. An *adjacency* in this stack is a pair of pancakes that are adjacent in the stack, and such that no other pancake has size intermediate between the two. If the largest pancake is on the bottom, this also counts as one extra adjacency. Now, for $n \geqslant 4$ there are stacks of n pancakes that have no adjacencies whatsoever. On the other hand, a sorted stack must have all n adjacencies and each move (flip) can create at most one adjacency. Consequently, for $n \geqslant 4$, $f(n) \geqslant n$. By elaborating on this argument, M.R. Garey, D.S. Johnson and S. Lin [2] showed that $f(n) \geqslant n+1$ for $n \geqslant 6$.

For upper bounds—algorithms, that is—it was known that $f(n) \leqslant 2n$. This can be seen as follows. Given any stack we may start by bringing the largest pancake on top and then flip the whole stack: the largest pancake is now at the bottom,

* Research supported by NSF Grant MCS 77-01193.
† Current address: Laboratory for Computer Science, Massachusetts Institute of Technology, Cambridge, Ma 02139, USA.

How many reversal steps for this REAL case?

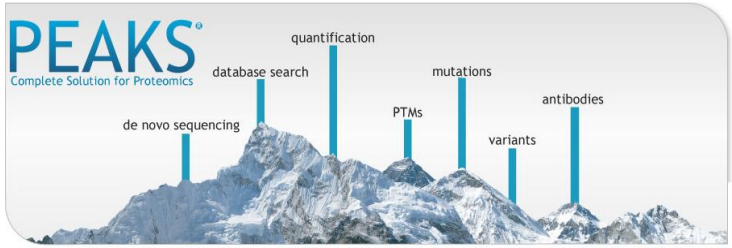


Ming Li （李明）

滑铁卢大学是微软招聘毕业生最多的学校之一

University of Waterloo

PatternHunter, ExonHunter, …



理论、生物信息、应用都重要！

Ming Li （李明）& Tao Jiang （姜涛）

SIAM J. COMPUT.
Vol. 24, No. 5, pp. 1122–1139, October 1995

012

ON THE APPROXIMATION OF SHORTEST COMMON SUPERSEQUENCES AND LONGEST COMMON SUBSEQUENCES*

TAO JIANG† AND MING LI‡

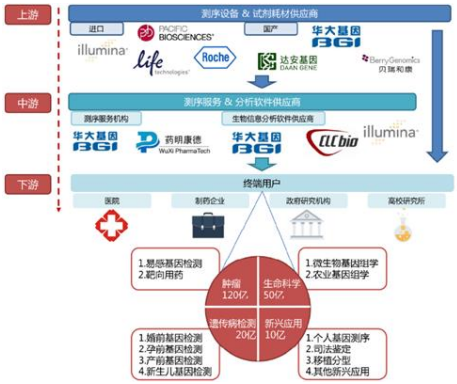
Abstract. The problems of finding shortest common supersequences (SCS) and longest common subsequences (LCS) are two well-known **NP**-hard problems that have applications in many areas, including computational molecular biology, data compression, robot motion planning, and scheduling, text editing, etc. A lot of fruitless effort has been spent in searching for good approximation algorithms for these problems. In this paper, we show that these problems are inherently hard to approximate in the worst case. In particular, we prove that (i) SCS does not have a polynomial-time linear approximation algorithm unless $\mathbf{P} = \mathbf{NP}$; (ii) There exists a constant $\delta > 0$ such that, if SCS has a polynomial-time approximation algorithm with ratio $\log^{\delta} n$, where n is the number of input sequences, then **NP** is contained in $\mathbf{DTIME}(2^{\text{polylog } n})$; (iii) There exists a constant $\delta > 0$ such that, if LCS has a polynomial-time approximation algorithm with performance ratio n^{δ}, then $\mathbf{P} = \mathbf{NP}$. The proofs utilize the recent results of Arora et al. [*Proc. 23rd IEEE Symposium on Foundations of Computer Science*, 1992, pp. 14–23] on the complexity of approximation problems.

In the second part of the paper, we introduce a new method for analyzing the average-case performance of algorithms for sequences, based on Kolmogorov complexity. Despite the above nonapproximability results, we show that near optimal solutions for both SCS and LCS can be found on the average. More precisely, consider a fixed alphabet Σ and suppose that the input sequences are generated randomly according to the uniform probability distribution and are of the same length n. Moreover, assume that the number of input sequences is polynomial in n. Then, there are simple greedy algorithms which approximate SCS and LCS with expected additive errors $O(n^{0.707})$ and $O(n^{1/2+\epsilon})$ for any $\epsilon > 0$, respectively.

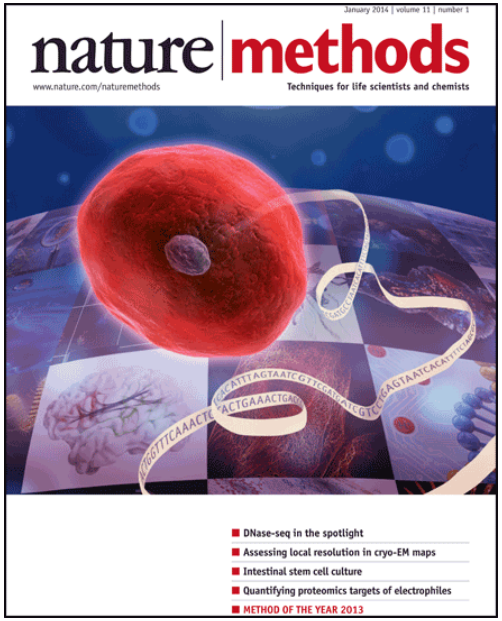
Incidentally, our analyses also provide tight upper and lower bounds on the expected LCS and SCS lengths for a set of random sequences solving a generalization of another well-known open question on the expected LCS length for two random sequences [K. Alexander, *The rate of convergence of the mean length of the longest common subsequence*, 1992, manuscript], [V. Chvatal and D. Sankoff, *J. Appl. Probab.*, 12 (1975), pp. 306–315], [D. Sankoff and J. Kruskall, eds., *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison*, Addison–Wesley, Reading, MA, 1983].

Key words. shortest common supersequence, longest common subsequence, approximation algorithm, **NP**-hardness, average-case analysis, random sequence

Current status（现今态势）

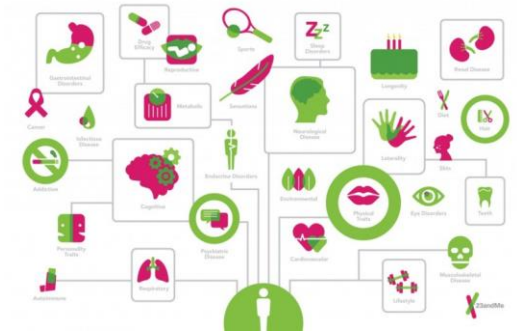


很难找到
与生物信息学和生物统计学
没有关系的
生物学与生物工程
研究和应用领域了。。。



Alphabet （谷歌）

Google 的基因组学梦想



The NEW ENGLAND JOURNAL of MEDICINE

HOME | ARTICLES & MULTIMEDIA | ISSUES | SPECIALTIES & TOPICS | FOR AUTHORS

CORRESPONDENCE

23andMe and the FDA

N Engl J Med 2014; 370:2248-2249 | June 5, 2014 | DOI: 10.1056/NEJMc1404692

Share:

Article | Citing Articles (3) | Metrics

To the Editor:

In their Perspective article (March 13 issue),[1] Annas and Elias state that the conflict between the genetic-testing company 23andMe and the Food and Drug Administration (FDA) concerns analytic and clinical validity, clinical utility, and ethical, legal, and social issues. However, their discussion is limited to a domestic U.S. perspective. After a person's raw genetic data have been determined from a DNA sample, the data are stored remotely and can be accessed easily anywhere in the world. For example, in Japan, maternal blood samples from Japanese mothers undergoing

nature biotechnology

Home | Current issue | News & comment | Research | Archive ▾ | Authors & referees ▾ | About the journal

home ▸ archive ▸ issue ▸ news ▸ full text

NATURE BIOTECHNOLOGY | NEWS

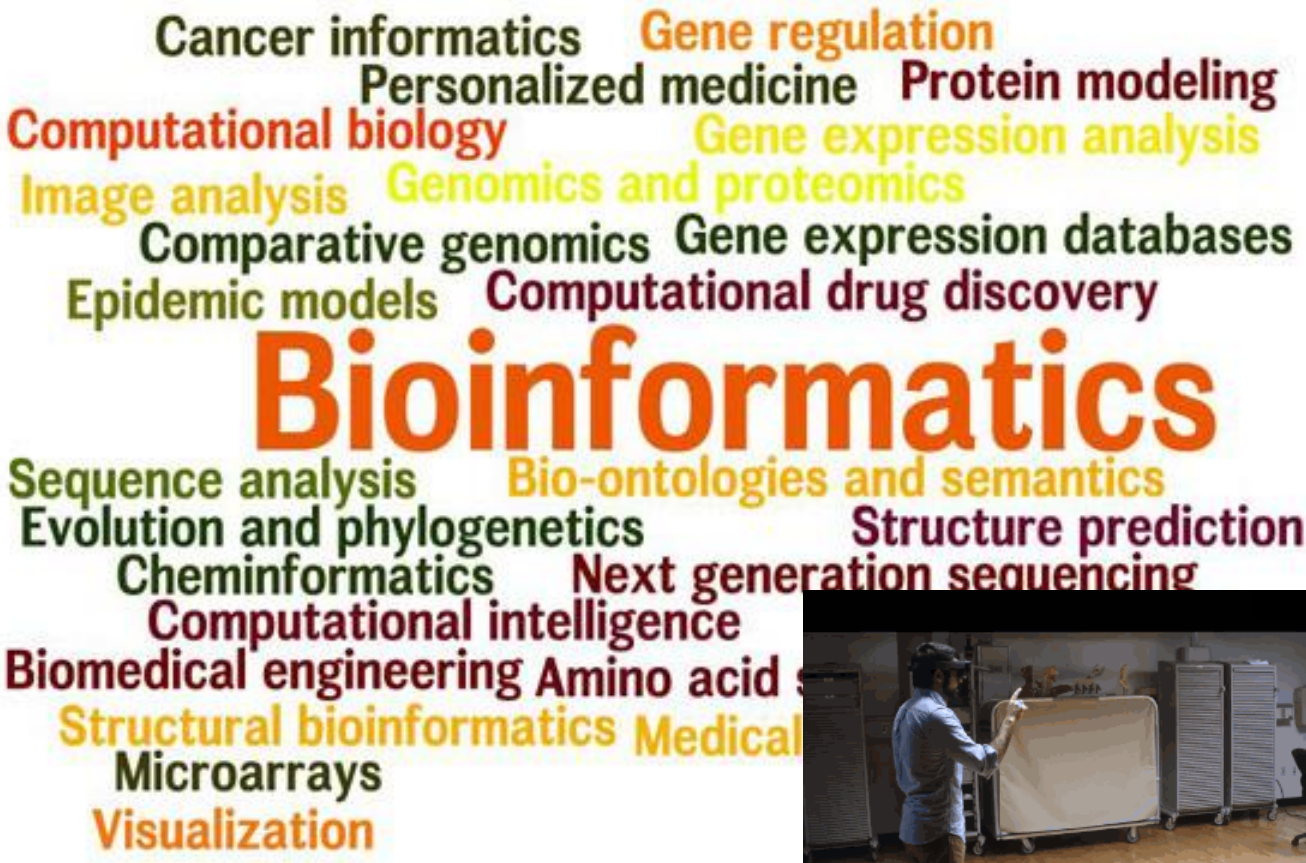
FDA approves 23andMe gene carrier test

Nature Biotechnology **33**, 435 (2015) | doi:10.1038/nbt0515-435a
Published online 12 May 2015

PDF | Citation | Reprints | Rights & permissions | Article metrics

23andMe, based in Mountain View, California, has received word from the US Food and Drug Administration (FDA) that their Bloom syndrome carrier screening test was approved as a class II device. The approval came in February, 15 months after the personal genomics company received a cease and desist letter from the regulator for its genetic tests because the company was dispensing health-related information to consumers without having obtained marketing clearance. The FDA website lists class II devices as moderate risk, requiring some regulatory controls, putting carrier screening tests in the same category as condoms. This turnaround follows the company's

Future（未来）

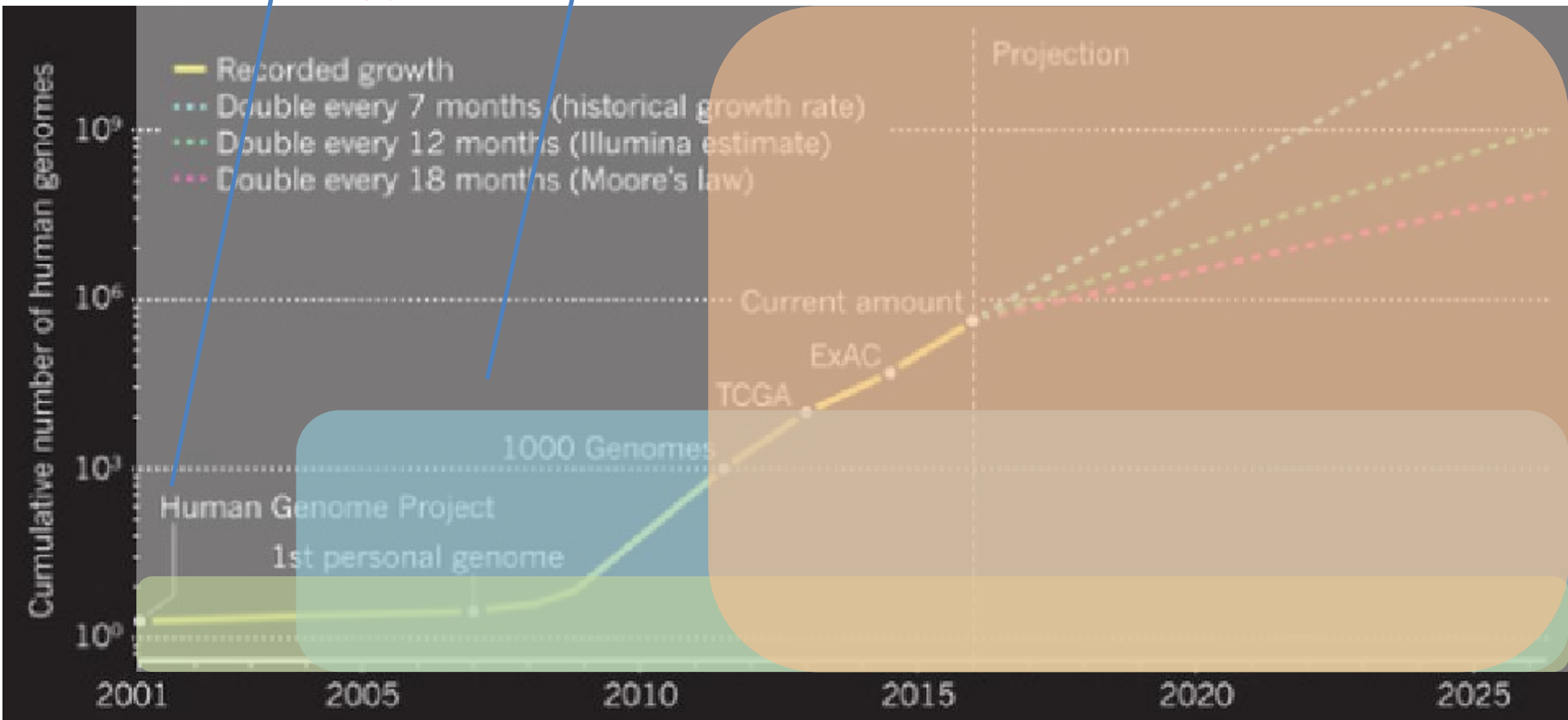

Cancer informatics
Gene regulation
Personalized medicine
Protein modeling
Computational biology
Gene expression analysis
Image analysis
Genomics and proteomics
Comparative genomics
Gene expression databases
Epidemic models
Computational drug discovery
Bioinformatics
Sequence analysis
Bio-ontologies and semantics
Evolution and phylogenetics
Structure prediction
Cheminformatics
Next generation sequencing
Computational intelligence
Biomedical engineering
Amino acid
Structural bioinformatics
Medical
Microarrays
Visualization

Biostatistics

湿实验
可验证范围

传统生物信息
处理范围

生物统计和深度学习
处理范围

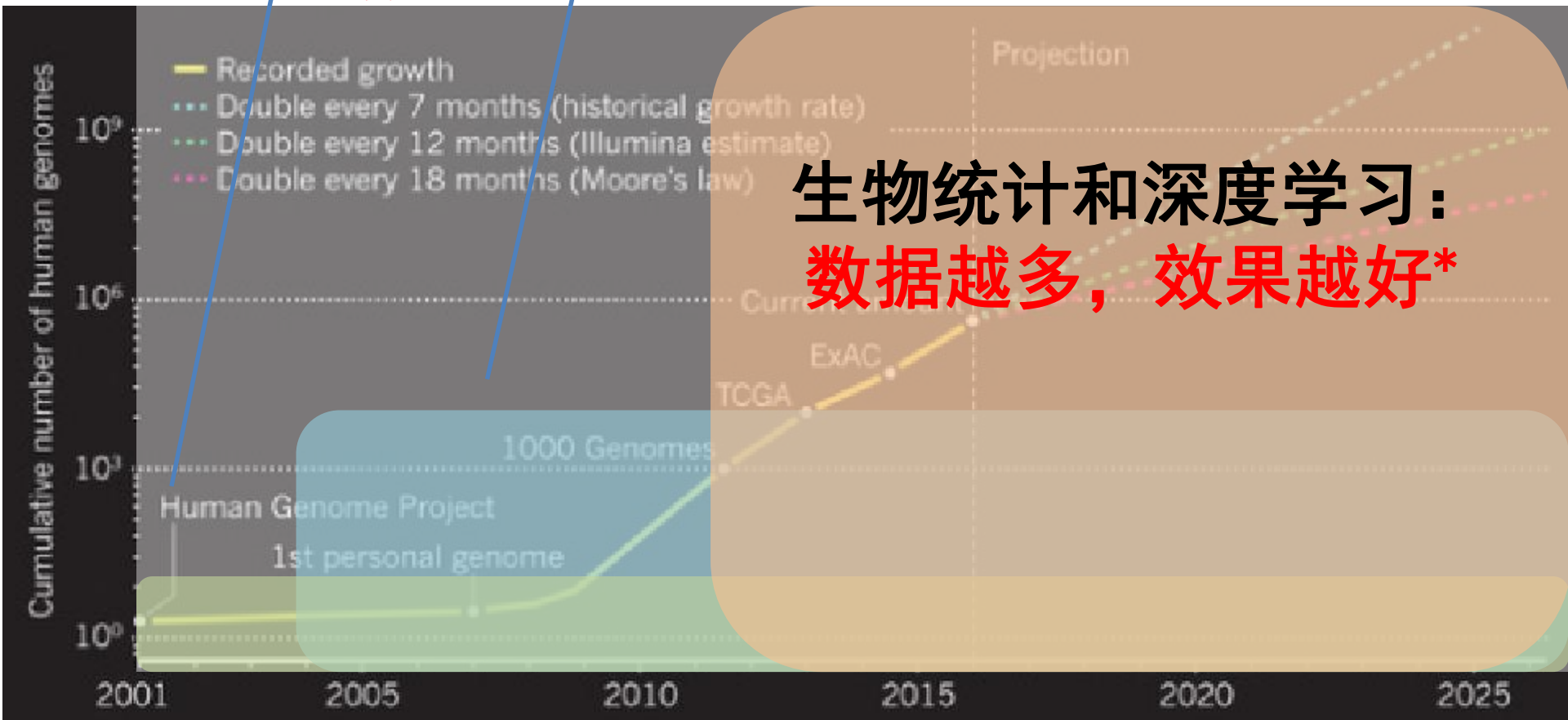


Biostatistics

湿实验
可验证范围

传统生物信息
处理范围

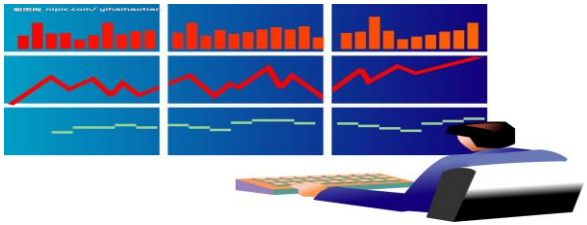
生物统计和深度学习
处理范围



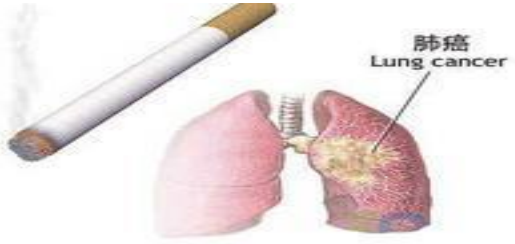
生物统计和深度学习：
数据越多，效果越好*

统计学：让数字说话！

人口数量和结构可以预测吗？

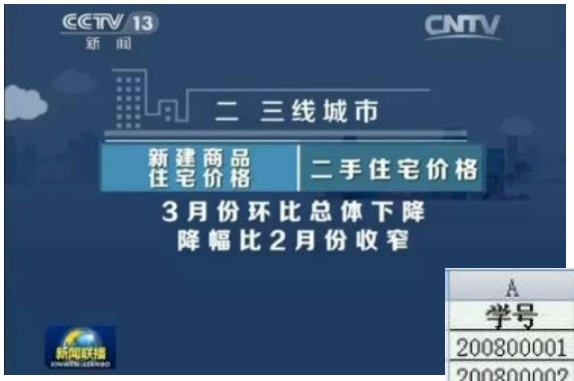


股市可以预测吗？



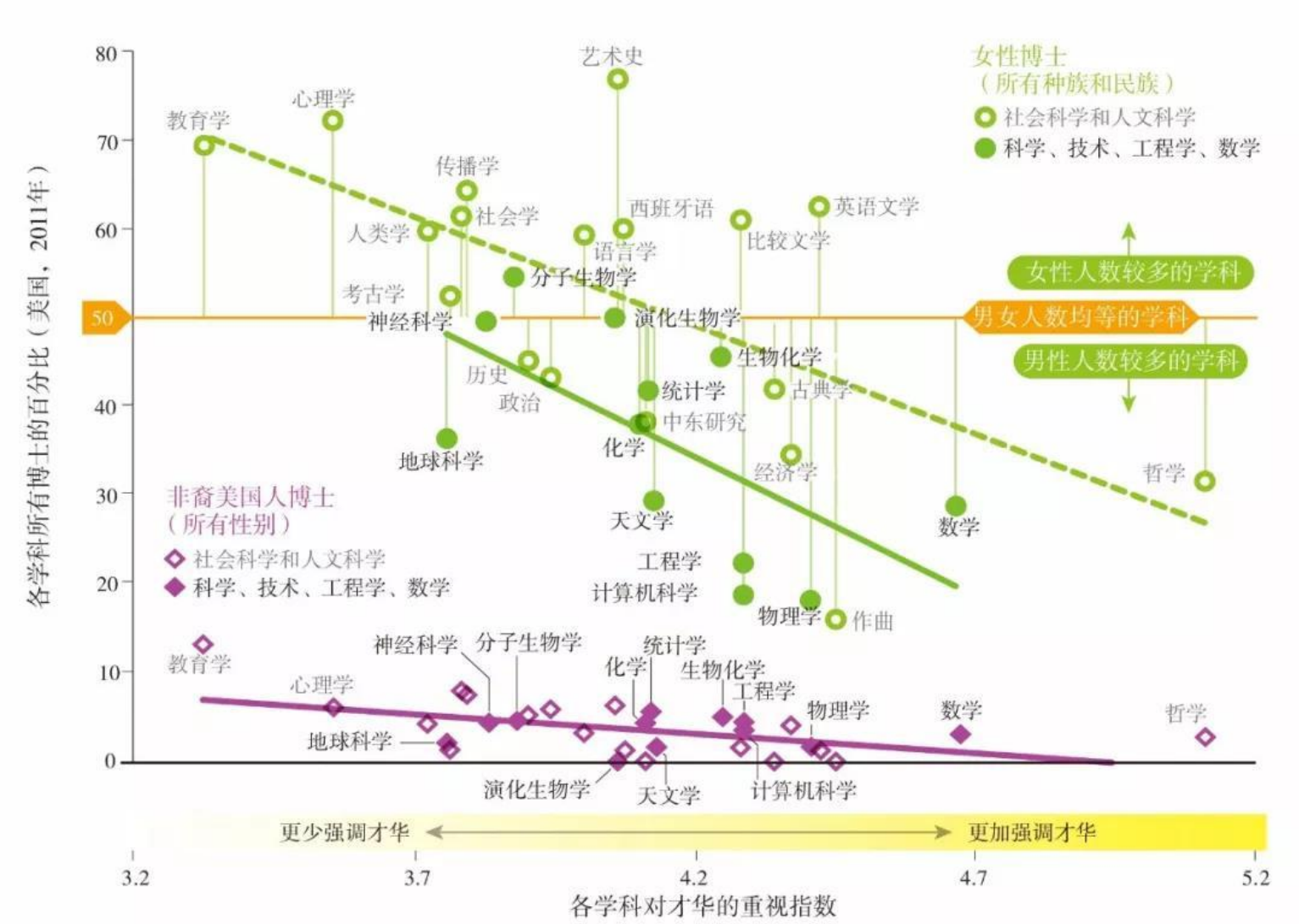
吸烟可以致癌？新药测试？

统计学：无处不在！



A	B	C	D	E	F	G
学号	姓名	语文	数学	英语	平均分	名次
200800001	狮子王	89	65	56	70.00	13
200800002	米老鼠	96	80	98	91.33	1
200800003	花木兰	83	82	83	82.67	9
200800004	唐老鸭	91	90	87	89.33	3
200800005	阿童木	93	92	88	91.00	2
200800006	灰姑娘	90	86	80	85.33	6
200800007	大力士	94	79	91	88.00	4
200800008	跳跳虎	83	43	94	73.33	10
200800009	小蚂蚁	89	91	82	87.33	5
200800010	皮诺曹	48	95	76	73.00	11
200800011	爱丽丝	87	83	85	85.00	7
200800012	睡美人	82	79.5	90	83.83	8
200800013	加菲猫	81	54	80	71.67	12

统计学：无处不在！


各学科所有博士的百分比（美国，2011年）
80
70
60
50
40
30
20
10
0
女性博士
（所有种族和民族）
社会科学和人文科学
科学、技术、工程学、数学
女性人数较多的学科
男女人数均等的学科
男性人数较多的学科
艺术史
心理学
教育学
传播学
社会学
人类学
西班牙语
英语文学
比较文学
语言学
分子生物学
考古学
神经科学
演化生物学
生物化学
历史
政治
统计学
古典学
中东研究
化学
地球科学
经济学
哲学
天文学
数学
工程学
计算机科学
物理学
作曲
非裔美国人博士
（所有性别）
社会科学和人文科学
科学、技术、工程学、数学
教育学
神经科学
分子生物学
统计学
心理学
化学
生物化学
工程学
物理学
数学
哲学
地球科学
演化生物学
天文学
计算机科学
更少强调才华
更加强调才华
3.2
3.7
4.2
4.7
5.2
各学科对才华的重视指数

统计学：无处不在！

习近平说：中国是世界第二大经济体，有13亿多人口的大市场，有960多万平方公里的国土，中国经济是一片大海，而不是一个小池塘。大海有风平浪静之时，也有风狂雨骤之时。没有风狂雨骤，那就不是大海了。狂风骤雨可以掀翻小池塘，但不能掀翻大海。经历了无数次狂风骤雨，大海依旧在那儿！经历了5000多年的艰难困苦，中国依旧在这儿！面向未来，中国将永远在这儿！

概率：

P(掀翻∩狂风骤雨|小池塘)=high
P(掀翻∩小雨|小池塘)=low
P(掀翻∩狂风骤雨|大海)=low
P(掀翻∩小雨|大海)=low

贝叶斯推断：

P(掀翻|大海)=P(大海|掀翻)*P(掀翻)/P(大海)=low
vs.
P(掀翻|小池塘)=P(小池塘|掀翻)*P(掀翻)/P(小池塘)=high

P(掀翻|大海)=P(掀翻∩狂风骤雨|大海)+P(掀翻∩小雨|大海)

统计学：产生价值！

基于大数据统计分析的防控平台
（社会效益）

基于大数据统计分析的决策平台
（经济效益）

统计学：概率不等于事实！

盖洛普民意测验与美国总统大选关联度一览表（1936—2000）

年代	候选人	盖洛普最后民意测验结果（%）	总统选举真实结果（%）	盖洛普误差（%）
2000	布什	48.0	47.9	+0.1
1996	克林顿	52.0	49.2	+2.8
1992	克林顿	49.0	43.3	+5.7
1988	老布什	56.0	53.9	+2.1
1984	里根	59.0	59.2	-0.2
1980	里根	47.0	50.8	-3.8
1976	卡特	48.0	50.1	-2.1
1972	尼克松	62.0	61.8	+0.2
1968	尼克松	43.0	43.5	-0.5
1964	约翰逊	64.0	61.3	+2.7
1960	肯尼迪	51.0	50.1	+0.9
1956	艾森豪威尔	59.5	57.8	+1.7
1952	艾森豪威尔	51.0	55.4	-4.4
1948	杜鲁门	44.5	49.5	-5.0
1944	罗斯福	51.5	53.8	-2.3
1940	罗斯福	52.0	55.0	-3.0
1936	罗斯福	55.7	62.5	-6.8

盖洛普民意测验创始人
乔治·盖洛普

统计学：历史和地位！

人口普查

F检验

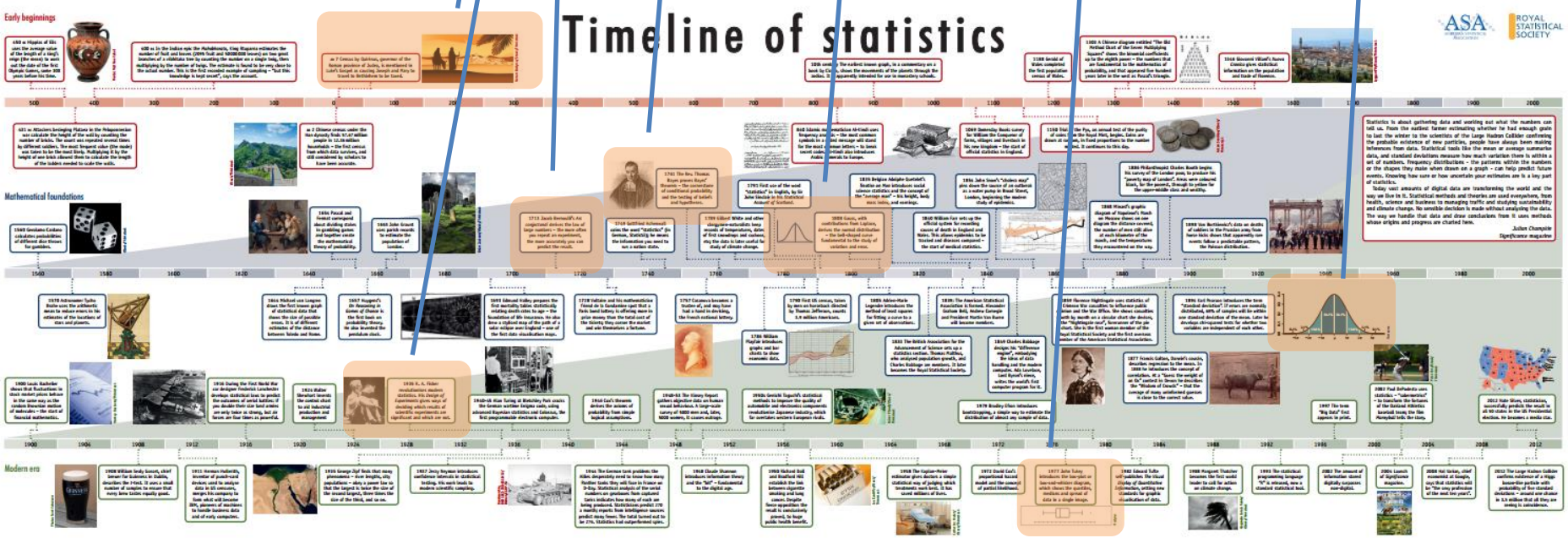
大数定理

贝叶斯推断

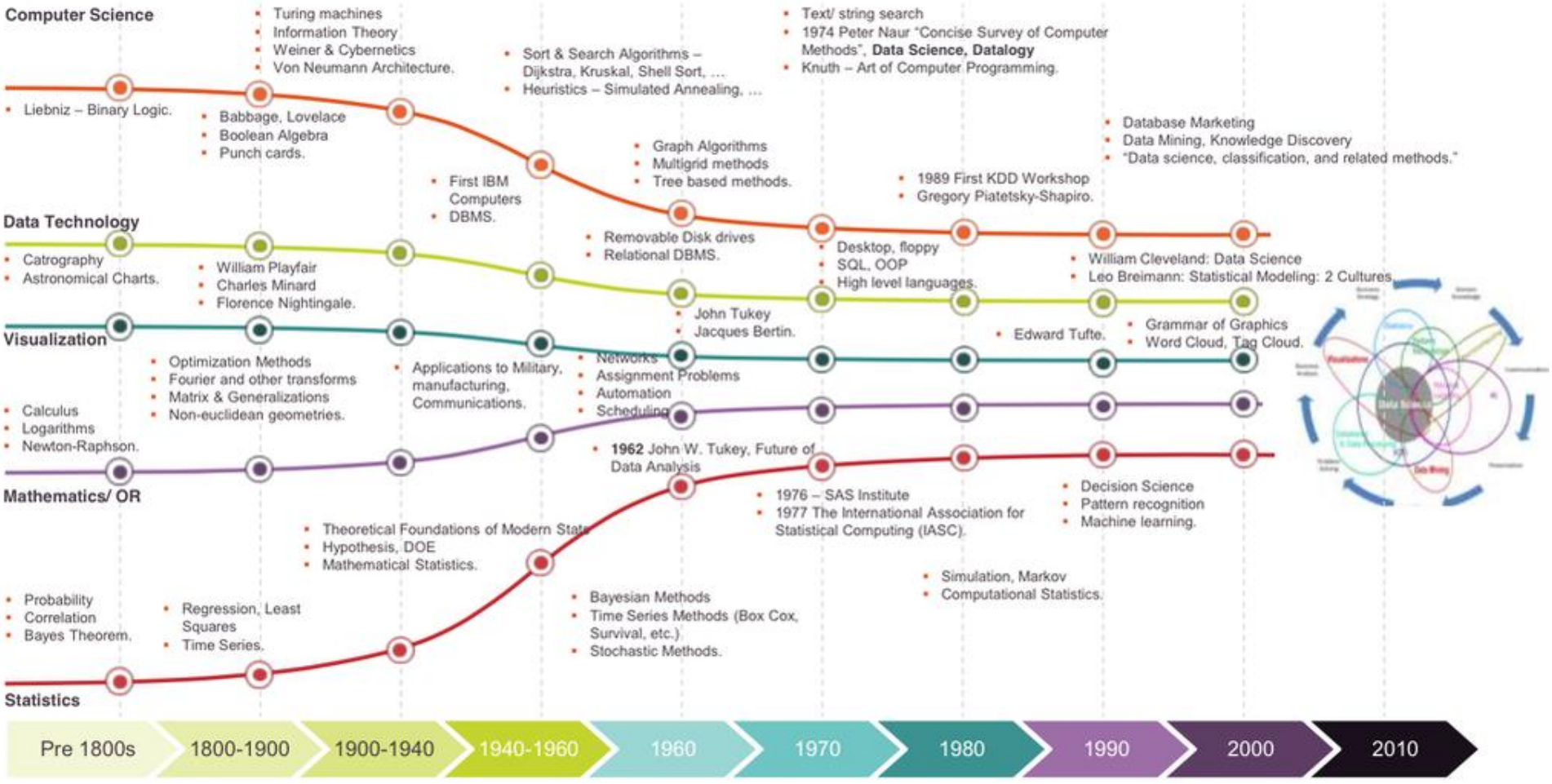
正态分布

箱式图

方差

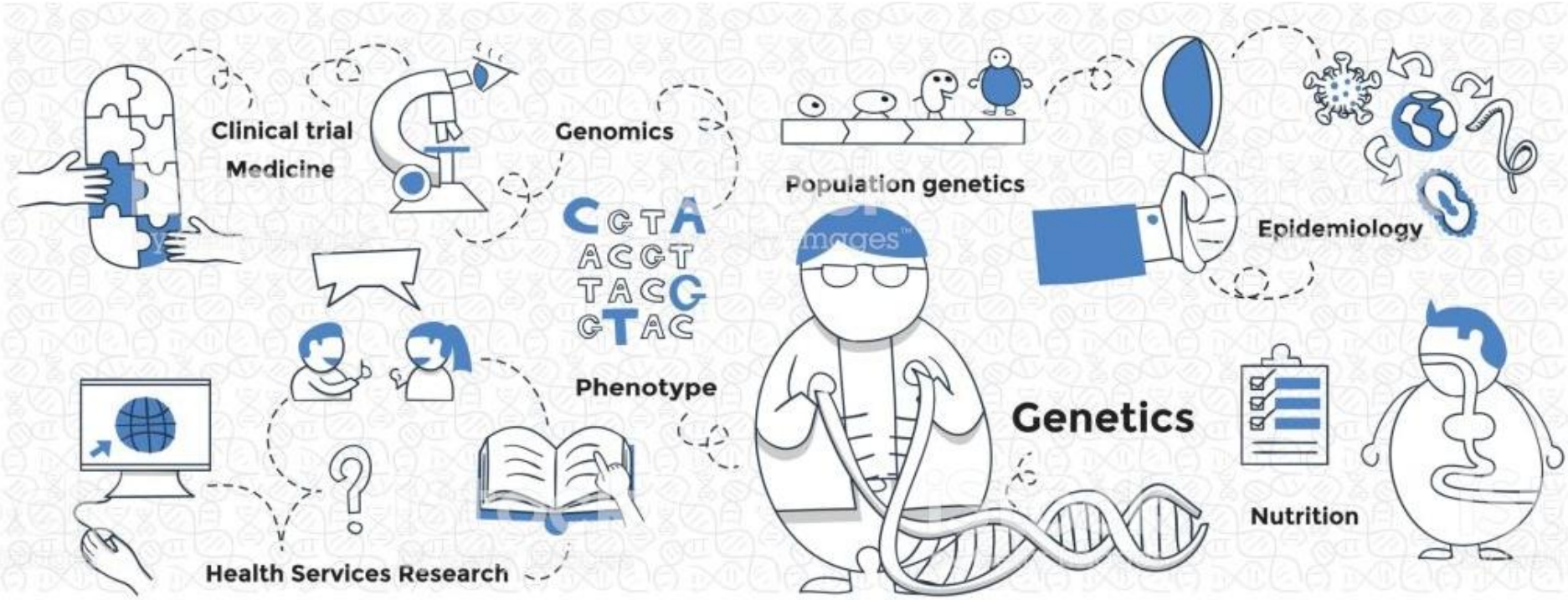


统计学：历史和地位！


Computer Science
Turing machines
Information Theory
Weiner & Cybernetics
Von Neumann Architecture.
Liebniz – Binary Logic.
Babbage, Lovelace
Boolean Algebra
Punch cards.
Sort & Search Algorithms – Dijkstra, Kruskal, Shell Sort, …
Heuristics – Simulated Annealing, …
Text/ string search
1974 Peter Naur "Concise Survey of Computer Methods", Data Science, Datalogy
Knuth – Art of Computer Programming.
Graph Algorithms
Multigrid methods
Tree based methods.
First IBM Computers
DBMS.
1989 First KDD Workshop
Gregory Piatetsky-Shapiro.
Database Marketing
Data Mining, Knowledge Discovery
"Data science, classification, and related methods."
Data Technology
Catography
Astronomical Charts.
William Playfair
Charles Minard
Florence Nightingale.
Removable Disk drives
Relational DBMS.
Desktop, floppy
SQL, OOP
High level languages.
William Cleveland: Data Science
Leo Breimann: Statistical Modeling: 2 Cultures
John Tukey
Jacques Bertin.
Visualization
Edward Tufte.
Grammar of Graphics
Word Cloud, Tag Cloud.
Optimization Methods
Fourier and other transforms
Matrix & Generalizations
Non-euclidean geometries.
Applications to Military, manufacturing, Communications.
Networks
Assignment Problems
Automation
Scheduling
Calculus
Logarithms
Newton-Raphson.
1962 John W. Tukey, Future of Data Analysis
Mathematics/ OR
1976 – SAS Institute
1977 The International Association for Statistical Computing (IASC).
Decision Science
Pattern recognition
Machine learning.
Theoretical Foundations of Modern Stats
Hypothesis, DOE
Mathematical Statistics.
Simulation, Markov
Computational Statistics.
Probability
Correlation
Bayes Theorem.
Regression, Least Squares
Time Series.
Bayesian Methods
Time Series Methods (Box Cox, Survival, etc.)
Stochastic Methods.
Statistics
Pre 1800s
1800-1900
1900-1940
1940-1960
1960
1970
1980
1990
2000
2010

统 计 学：生物统计学！


Bridging the sciences
BIOSTATISTICS


Clinical trial
Medicine
Genomics
Population genetics
Epidemiology
Phenotype
Genetics
Nutrition
Health Services Research

为什么要学习生物统计学

两品种小麦的对比

转基因小麦

非转基因小麦

一株转基因小麦各项指标都优于另一株非转基因小麦，是否可以确定转基因小麦产量提高？

实际情况可能是：

所选转基因小麦是实验田中长得最好的，而所选非转基因小麦是麦田中非常普通的一株。

怎样才能确认转基因小麦更优呢？

选取所有的小麦进行比较？

如何选取

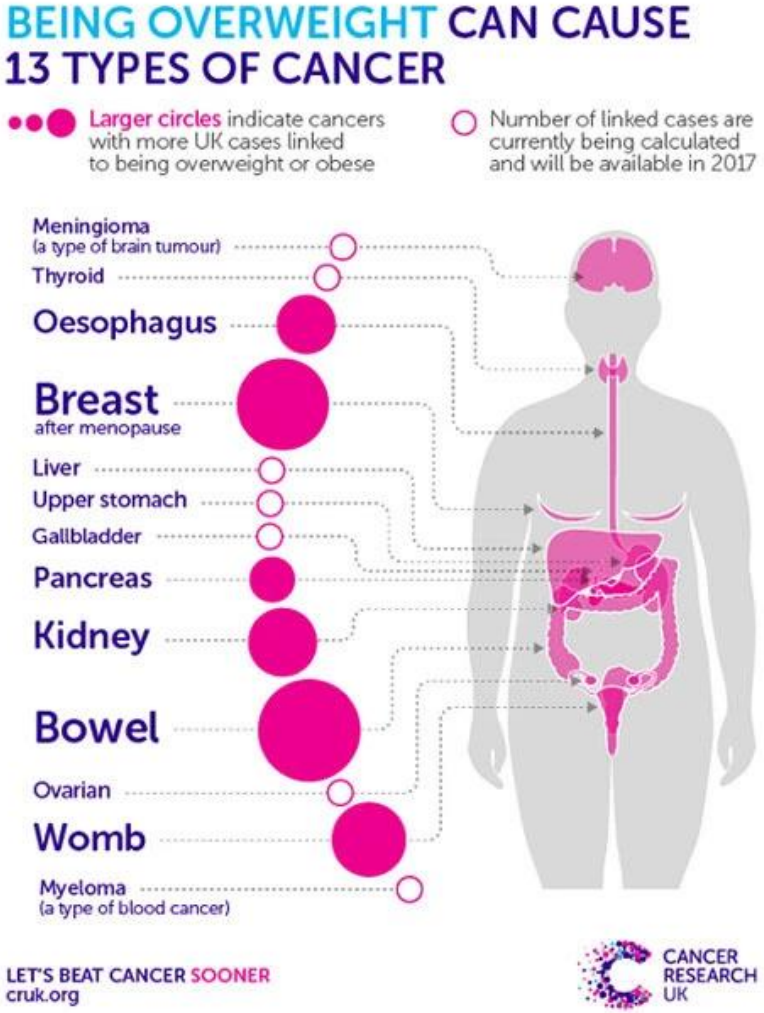
摘取多少小麦才能更好地保证对比结果的准确性呢？

为什么要学习生物统计学

阿司匹林对抗癌的疗效

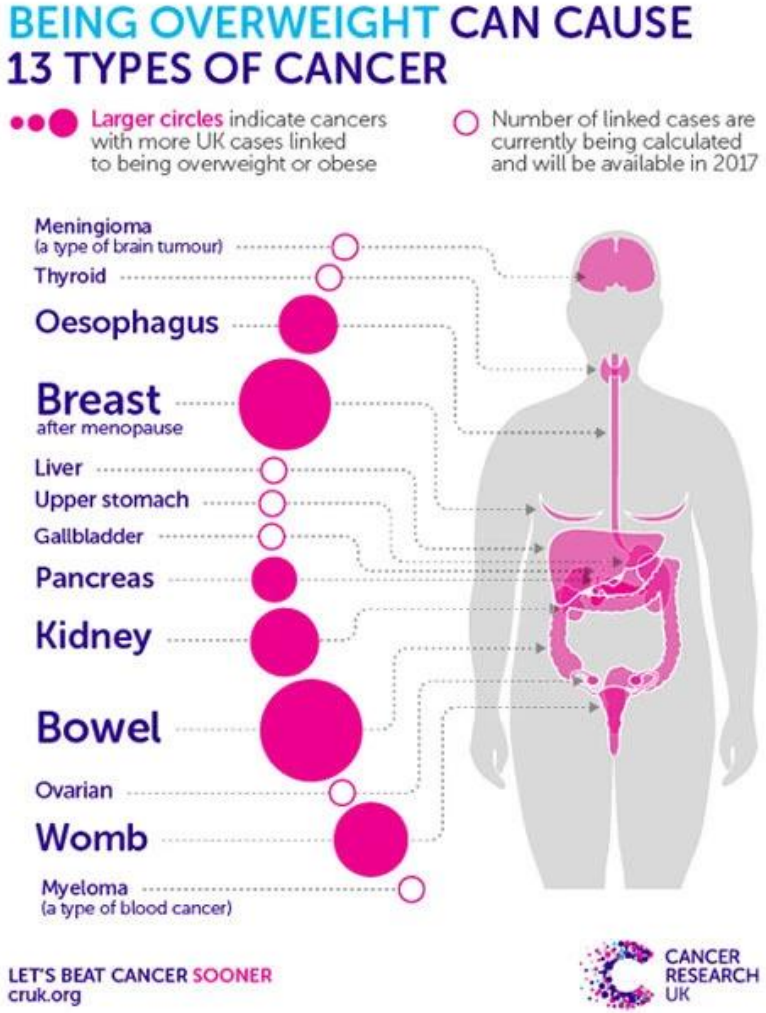
口服阿司匹林，可能有助于治疗好多种癌症。

是真的吗？



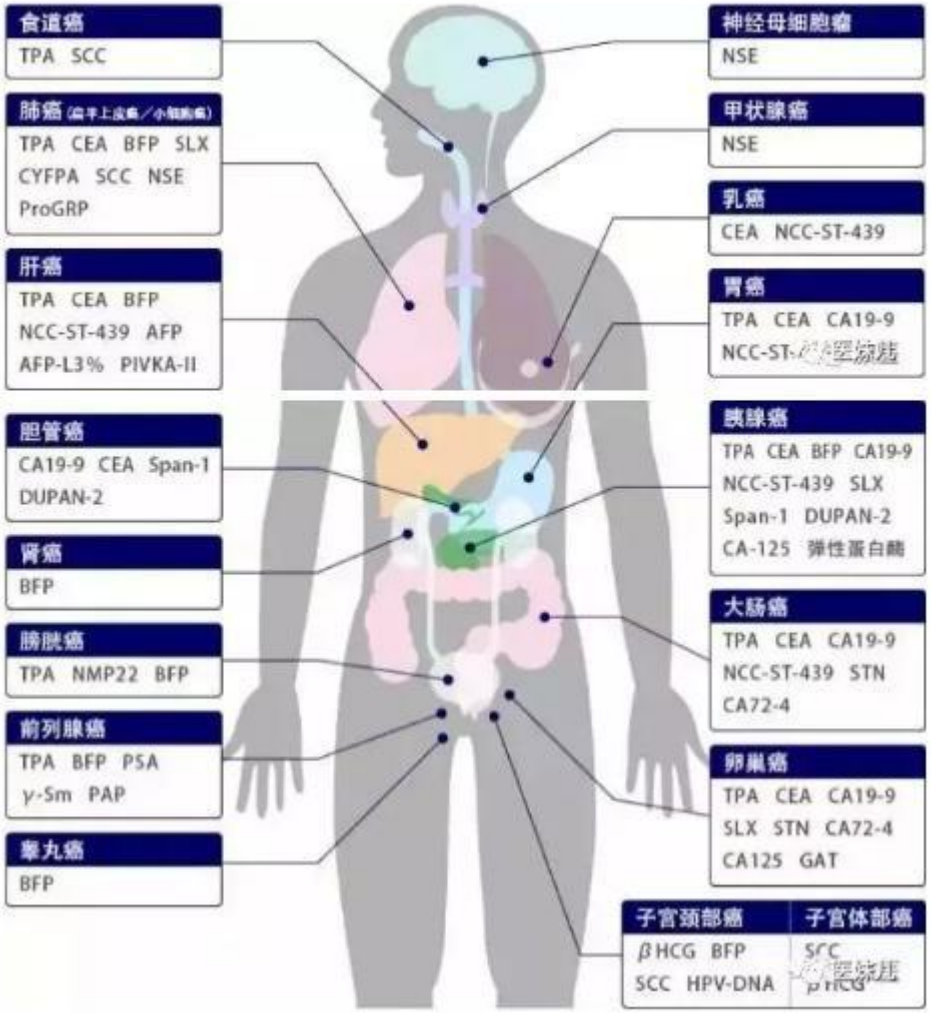
实际情况可能是：

口服阿司匹林的同时，可能也口服了其它抗癌药物。。。



如何选取

要评估口服阿司匹林的效果，必须做好实验设计，控制单变量，评估其影响。



如何确立可靠的关联性

为什么要学习生物统计学

- 生物统计学是生物科学研究的基本工具

➢ 生物现象的特点：

变异性：个体之间存在差异

不确定性（随机性）：变异不能准确推算

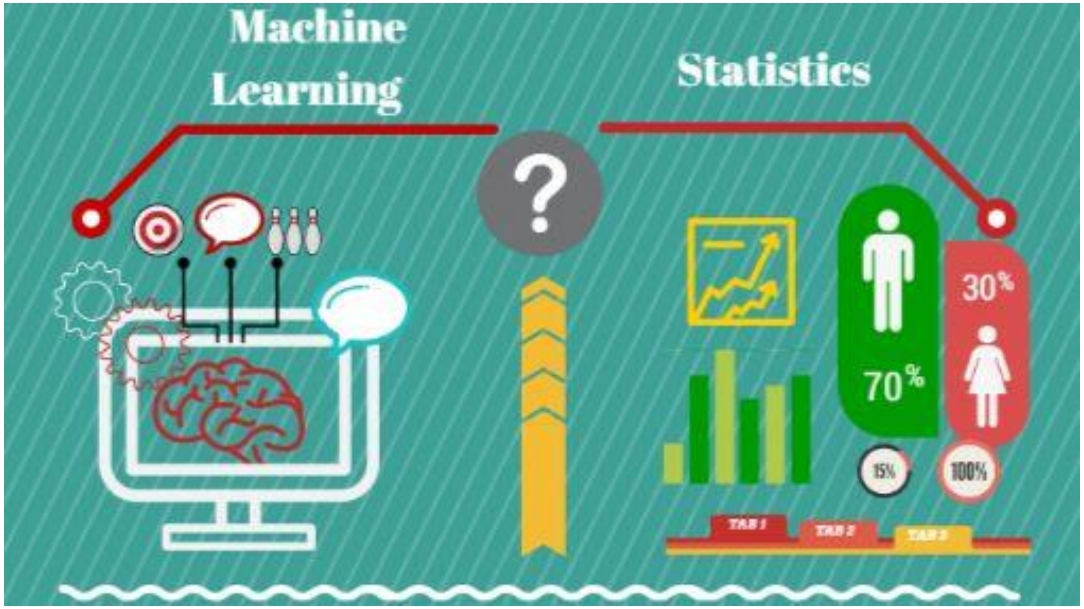
复杂性：影响因素众多，有些是未知的

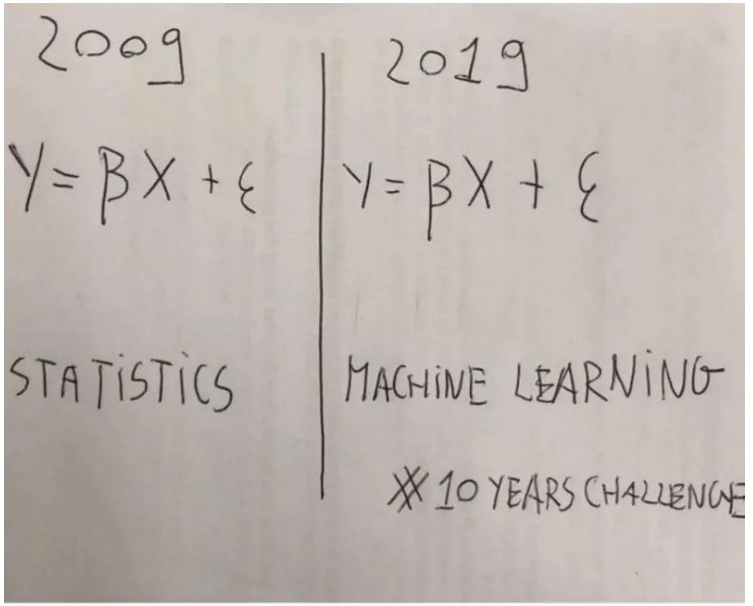
常规的数学方法不能解决问题

为什么要学习生物统计学

- 必须利用生物统计学才能回答的问题
  - 疾病已经进入哪个阶段了？
  - 哪些基因在疾病发生发展中起到关键作用？
  - 基因和环境是否有关？
  - 新药物是否更有效？
  - 遗传与环境哪个更重要？
  - ……

为什么要学习生物统计学





https://mp.weixin.qq.com/s/xCJBowXS89UlHA07R8WNuw

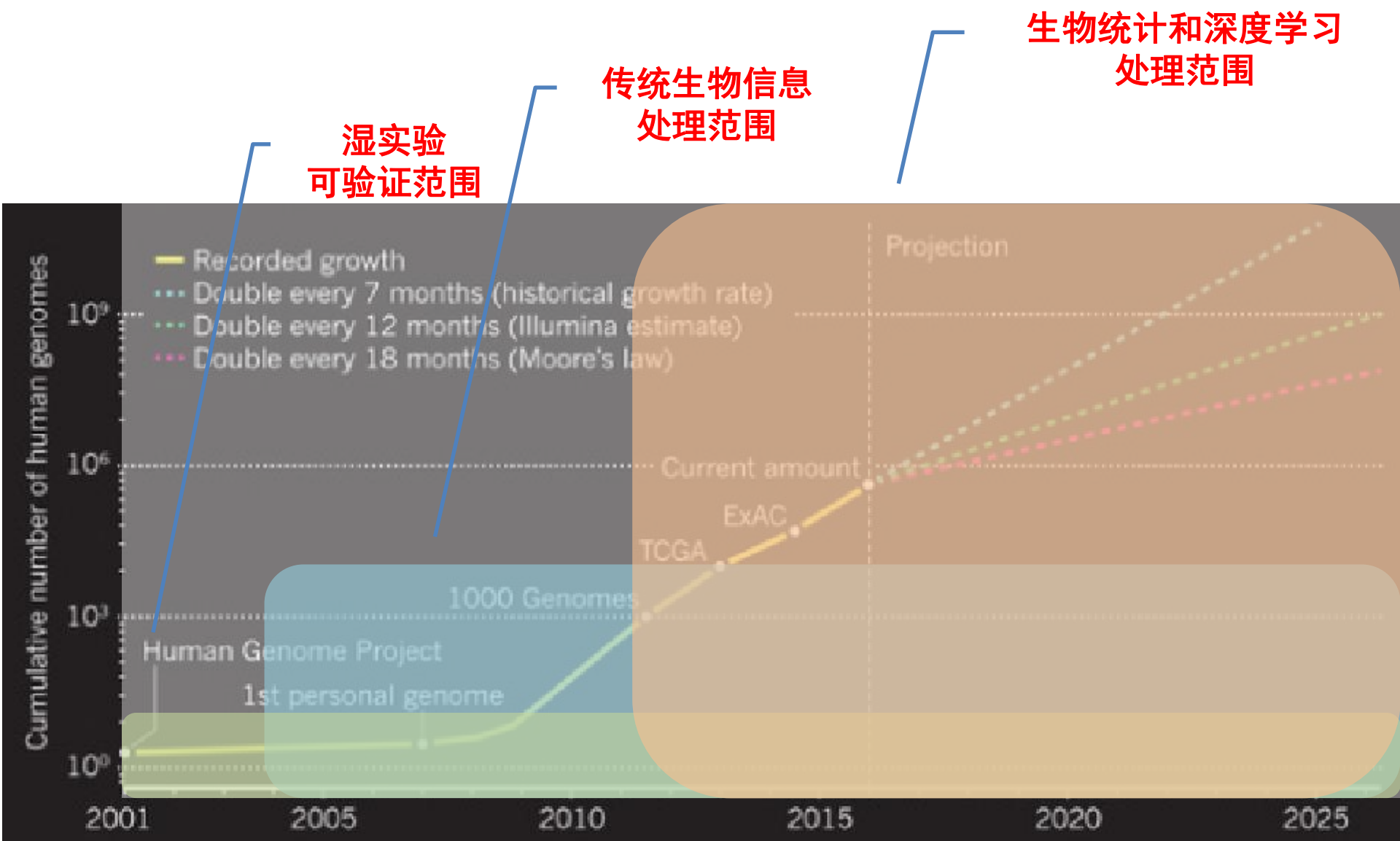
https://towardsdatascience.com/the-actual-difference-between-statistics-and-machine-learning-64b49f07ea3?gi=412e8f93e22e

为什么要学习生物统计学

- 如果你只是想从数据中找出哪类人更容易得某种疾病，机器学习可能是更好的选择。
- 如果你希望找出变量之间的关系或从数据中得出推论，选择统计模型会更好。

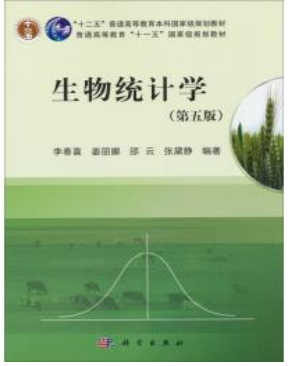
https://towardsdatascience.com/the-actual-difference-between-statistics-and-machine-learning-64b49f07ea3?gi=412e8f93e22e

为什么要学习生物统计学



学习方法与要求

- 要弄懂统计的基本原理和基本公式；
- 要认真做好习题作业，加深对公式及统计步骤的理解，达到能熟练地应用统计方法；
- 注意培养科学的统计思维方法，理论联系实际，结合专业，了解统计方法的实际应用。



教材及参考书目

- **教学参考书：**
- 《生物统计学》（第5版），普通高等教育十二五国家级规划教材). 科学出版社. 2013年6月出版. 李春喜，姜丽娜，邵云，张黛静编著.
- 《生物序列分析》（第1版）.科学出版社. 2010年8月出版. R. Durbin等编著，王俊等主译.

- **课外文献阅读：**
- 《生物统计学》(第4版). 高等教育出版社. 2013年12月出版. 杜荣骞主编.
- 《生物统计学》(普通高等院校生命科学类十二五规划教材). 华中科技大学出版社. 2015年3月出版. 彭明春，马纪主编.

课程范围

- **生物统计学的范围**
  - 一切和生物相关数据的分析有关的统计
- **面向生物信息和大数据挖掘的生物统计学特点**
  - 兼容并包、同时注重方法和应用
- **生物统计学的应用**
  - 精准医学的应用

课程结构

- 生物统计学基础；
- **生物信息中的算法设计与概率统计模型**；
- 生物大数据和深度学习。

Biostatistics

Biostatistics is the application of statistics to a wide range of topics in biology.

The science of biostatistics encompasses the design of biological experiments, especially in medicine, pharmacy, agriculture and fishery; the collection, summarization, and analysis of data from those experiments; and the interpretation of, and inference from, the results.

A major branch of this is medical biostatistics, which is exclusively concerned with medicine and health.

Bioinformatics

Bioinformatics is the research, development, or application of computational tools and approaches for expanding the use of biological, medical, behavioral or health data, including those to acquire, store, organize, archive, analyze, or visualize such data;

Computational biology is the development and application of data-analytical and theoretical methods, mathematical modeling and computational simulation techniques to the study of biological, behavioral, and social systems.

Bioinformatics

- Overlay of **Biology**, **Computer science** and **Statistics**.
- Topics:

  Sequence alignment

  Protein folding

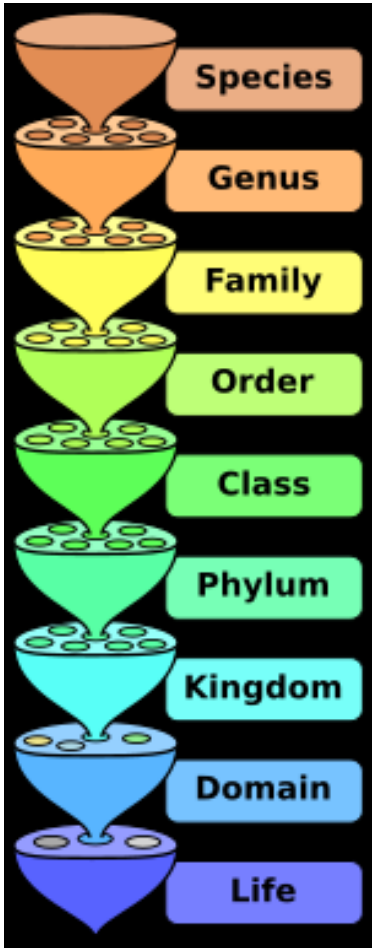
  Gene finding

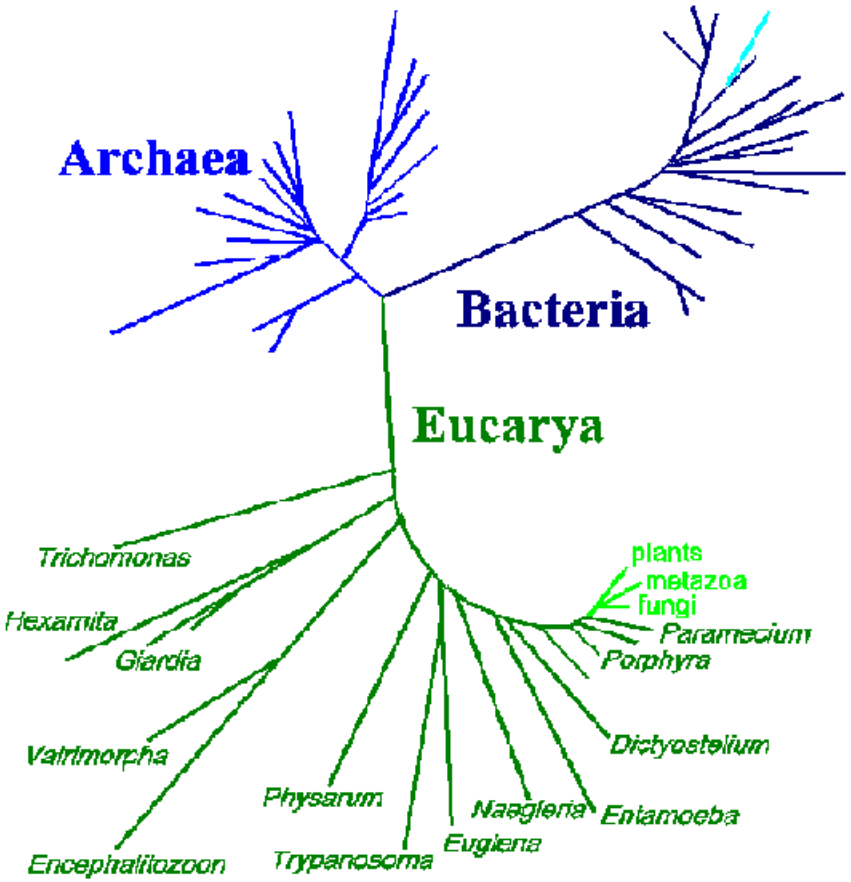
  Functional annotation

  Network inference

研究对象: 生物序列, 进化树, 生物网络, 基因表达...

Tree of Life





modified from N.R. Pace, ASM News 62:464, 1996

Molecules of Life

- DNA
- RNA
- Protein

DNA

- Deoxyribonucleic acid(脱氧核糖核酸)
- Consist of four nucleotides
  - A Adenine(腺嘌呤)
  - C Cytosine(胞嘧啶)
  - G Guanine(鸟嘌呤)
  - T Thymine(胸腺嘧啶)

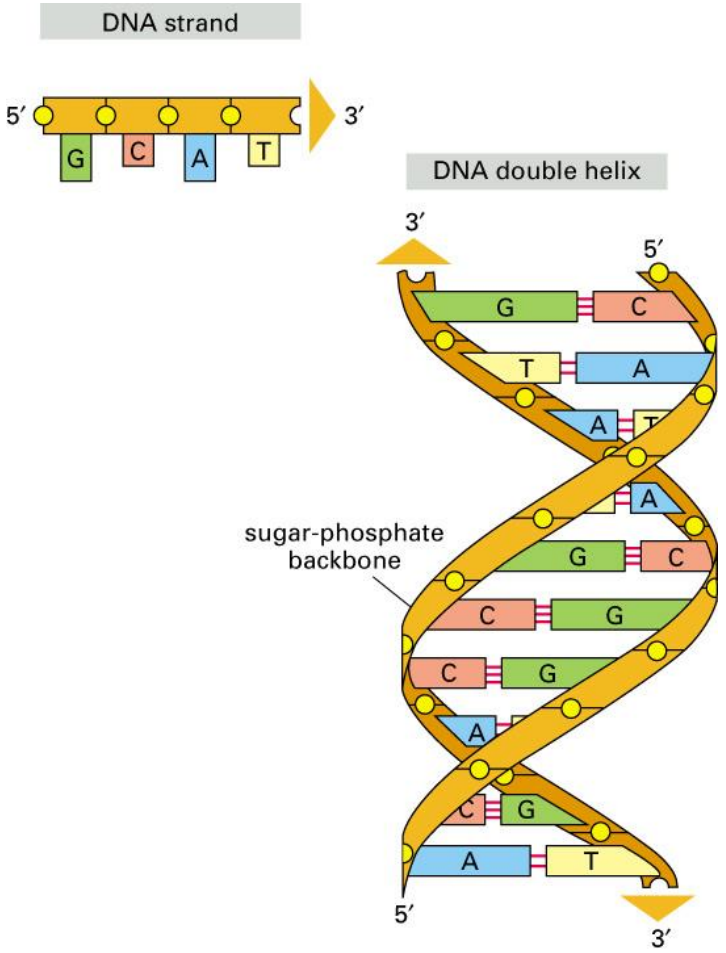


Figure 4–3 part 2 of 2. Molecular Biology of the Cell, 4th Edition.

RNA

- Ribonucleic acid (核糖核酸)
  - mRNA : Messenger RNAs, code for proteins
  - rRNA: Ribosomal RNAs, form the basic structure of the ribosome and catalyze protein synthesis.
  - tRNA: Transfer RNAs, central to protein synthesis as adaptors between mRNA and amino acids.

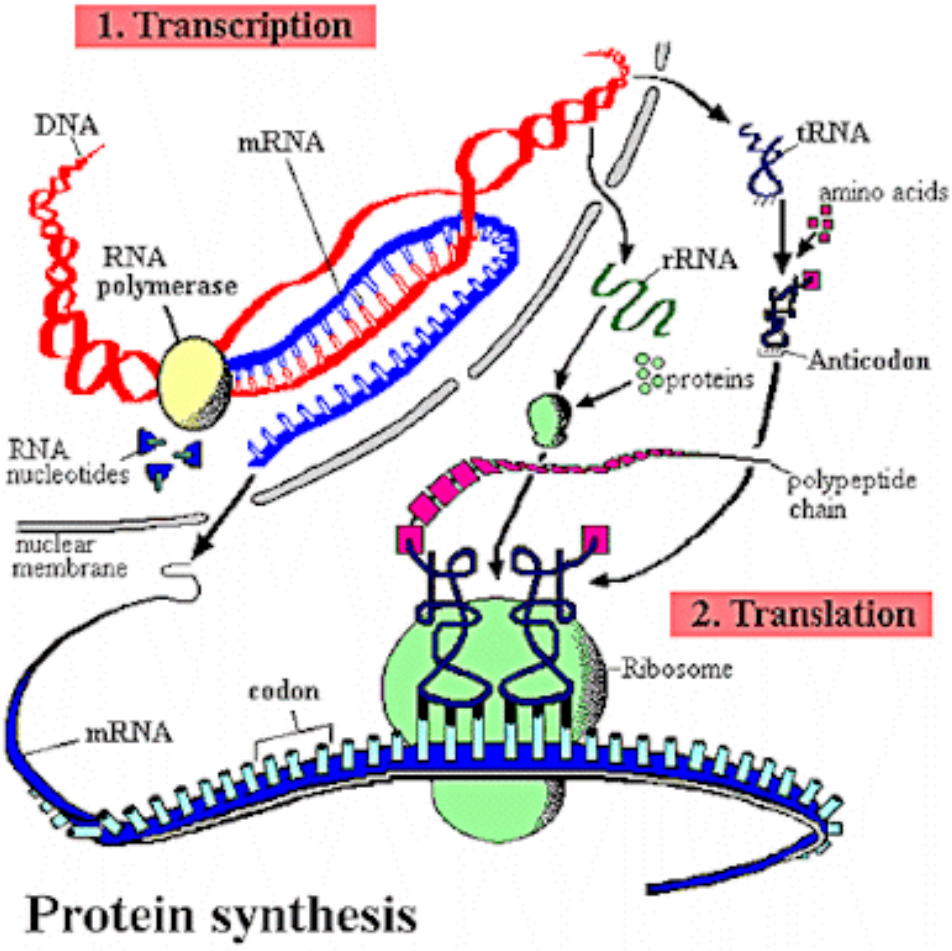
Proteins

Main building blocks and functional molecules, take up ~20% of eukaryotic cell's weight.

- Structural proteins
- Enzymes
- Antibodies
- Transmembrane proteins

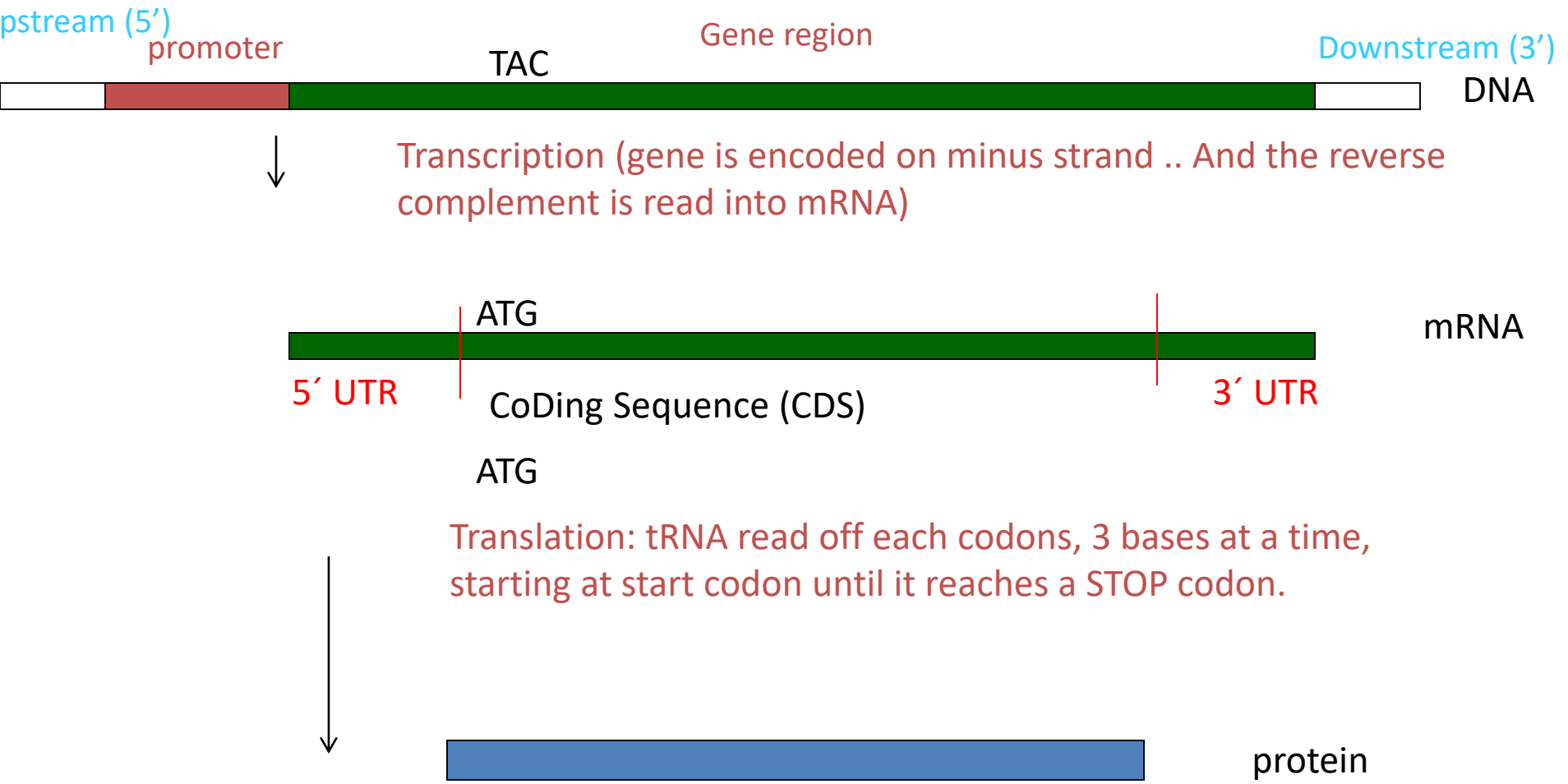
The Central Dogma

DNA→RNA→Protein

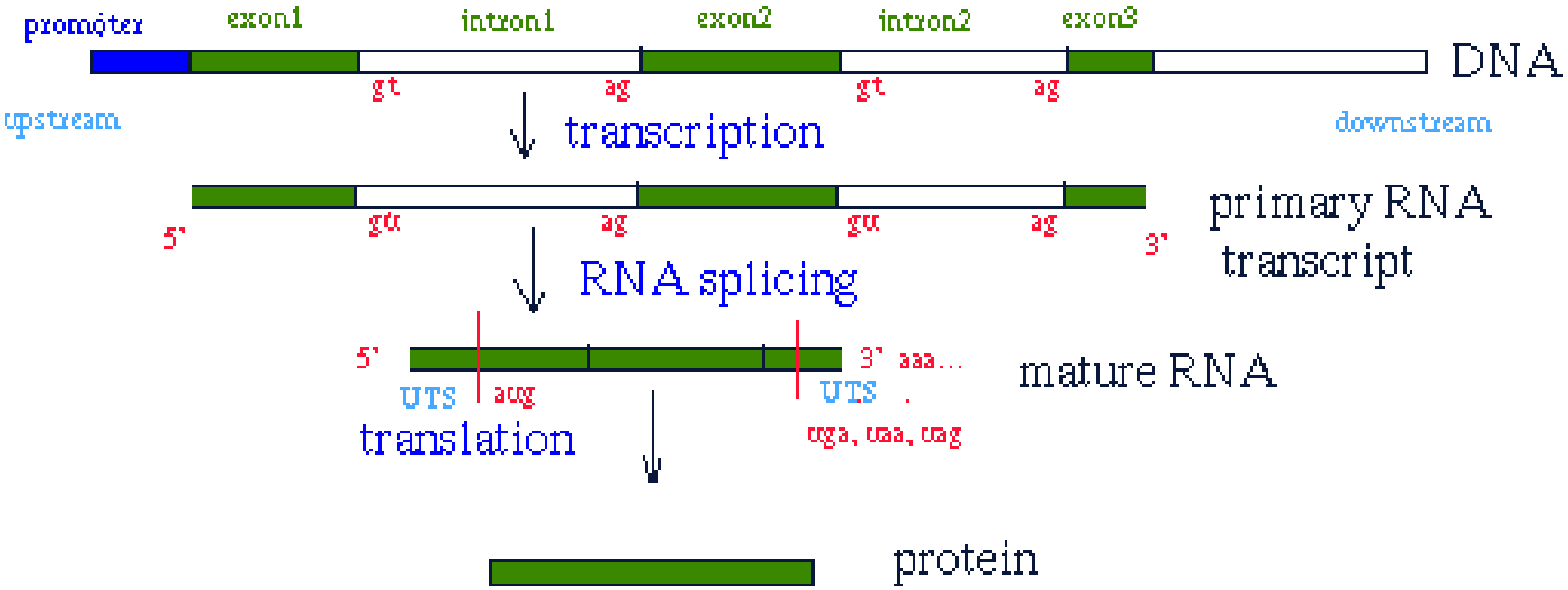


Prokaryotic Genes

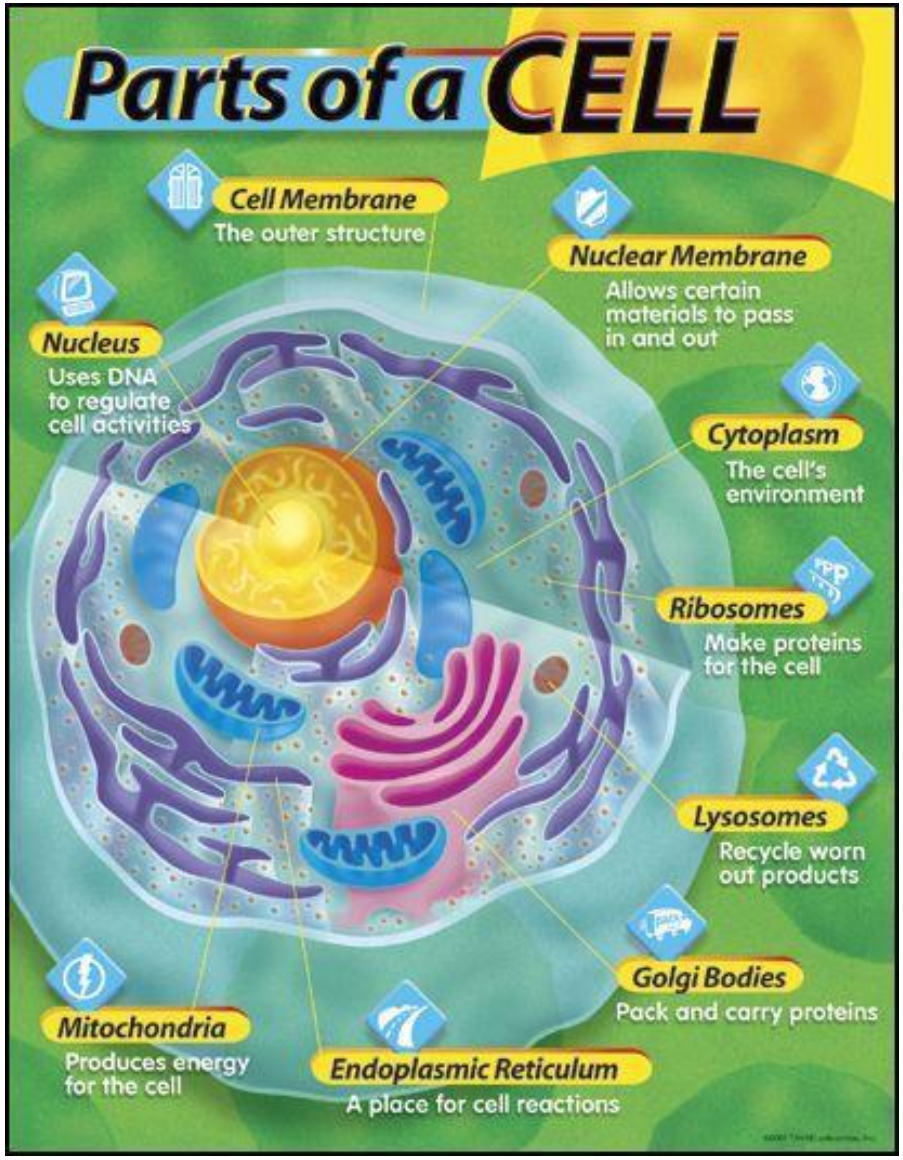
Prokaryotes (intronless protein coding genes)



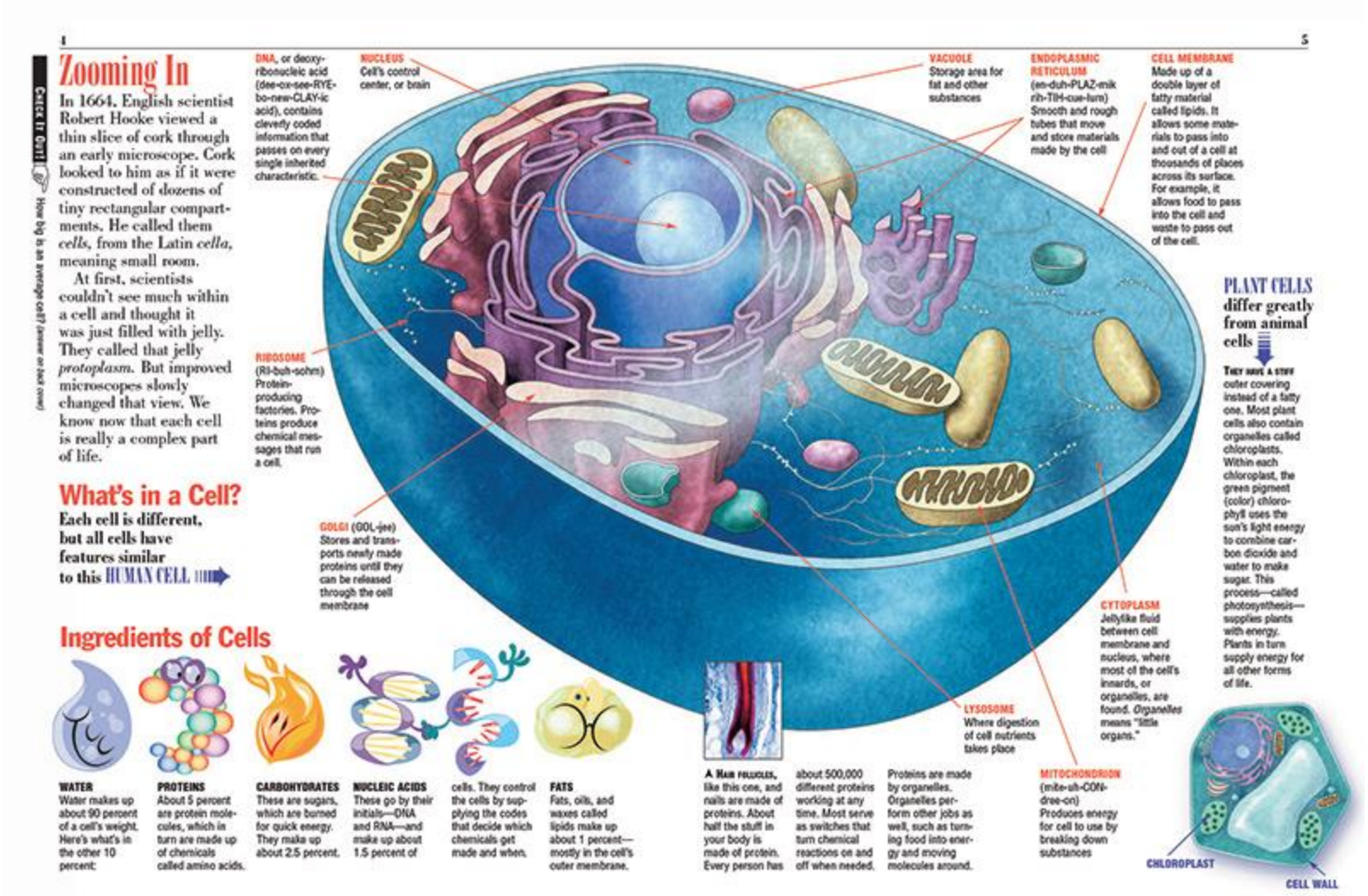
Eukaryotic Genes



Cell structure



Cell structure



Biostatistics

Case 1：基因表达数据分析

- 差异表达基因分析
- 基因共表达分析
- 基因表达数据的聚类和分类
- 基因集分析
- 基因调控网络

Biostatistics

Case 2：基因集合分析

- 通过测序，找到了一批“interesting” 的基因
  - 差异表达基因
  - 不做差异基因鉴定，直接做基因集分析
- 生物学功能上是否存在关联？
  - 某种功能是否显著？
- 计算分析方法
  - 基因本体 (Gene Ontology)
  - KEGG (Kyoto Encyclopedia of Genes and Genomes)
  - 超几何分布

Biostatistics

Case 3：微生物组生态数据分析

- 物种的富集分析
- 功能的富集分析
- **MWAS分析**

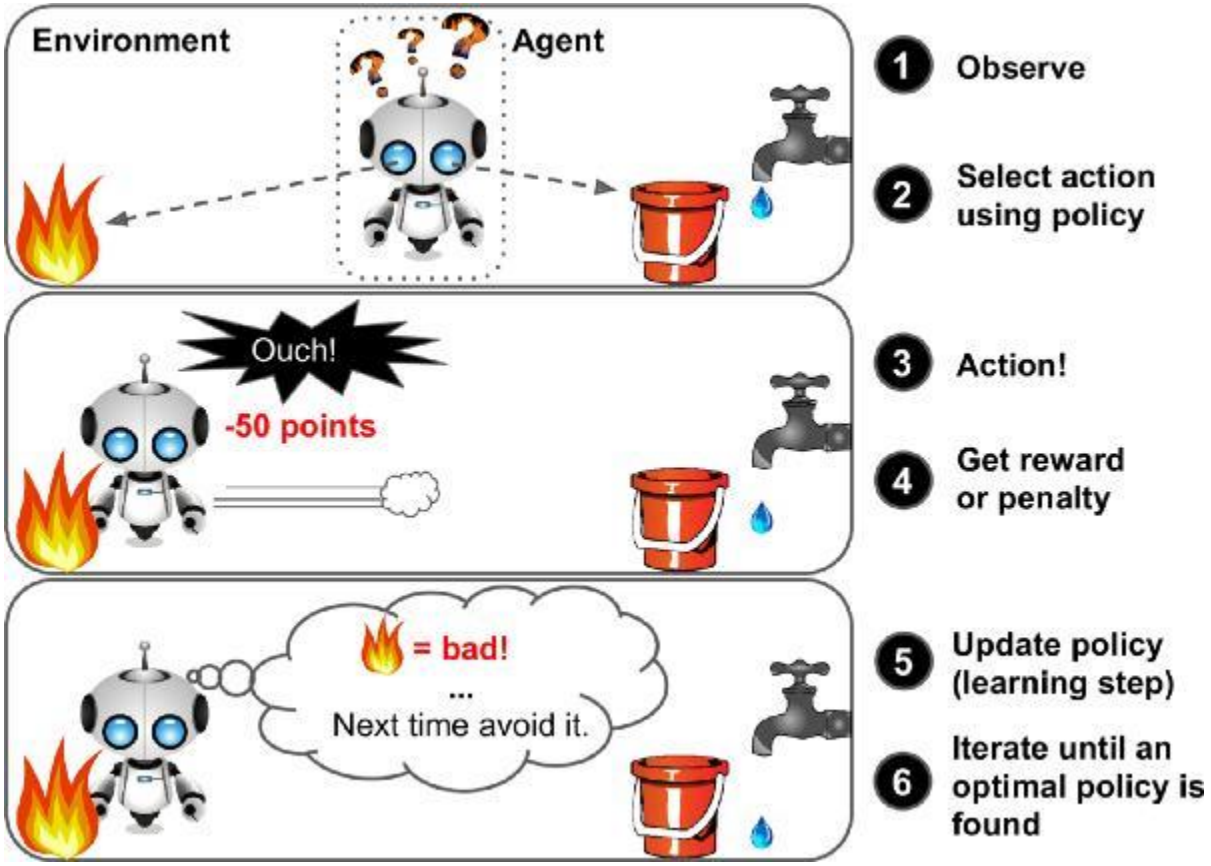
Big-data + Deep learning

Big-data is a term for data sets that are so large or complex that traditional data processing application software is inadequate to deal with them. Challenges include capture, storage, analysis, data curation, search, sharing, transfer, visualization, querying, updating and information privacy.

Deep learning is part of a broader family of machine learning methods based on learning representations of data. Research in this area attempts to make better representations and create models to learn these representations from large-scale unlabeled data.

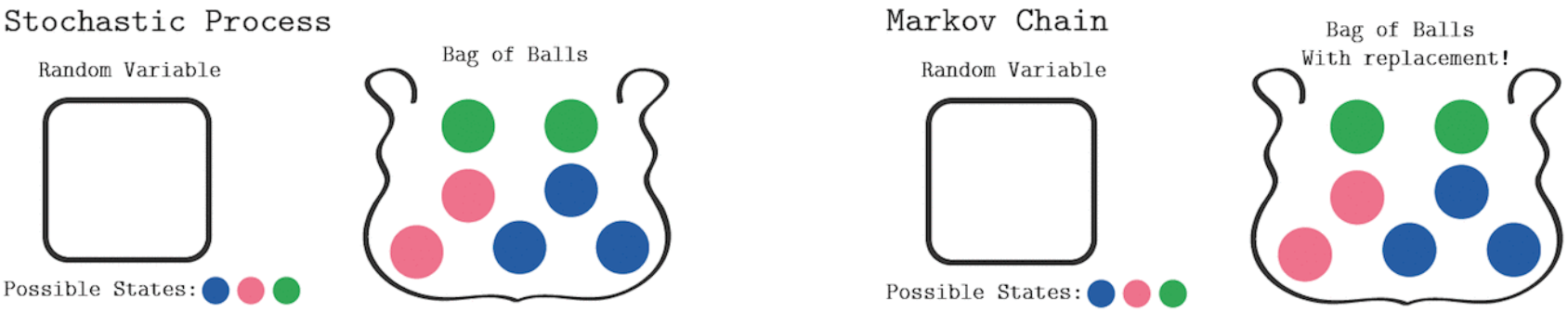
Big-data + Deep learning

Reinforcement Learning



Big-data + Deep learning

Markov chain



“The future is independent of the past given the present!”

Big-data + Deep learning

IMAGENET Large Scale Visual Recognition Challenge

Year 2010

NEC-UIUC

Dense descriptor grid: HOG, LBP

Coding: local coordinate, super-vector

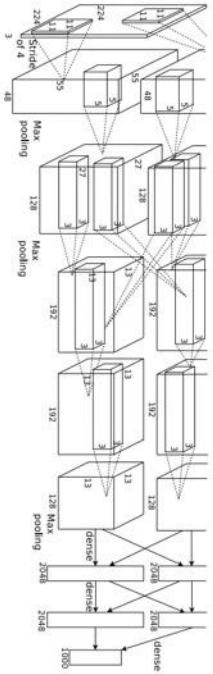
Pooling, SPM

Linear SVM

[Lin CVPR 2011]

Year 2012

SuperVision

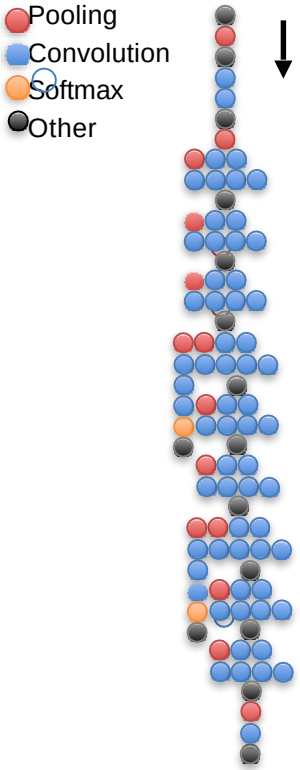


[Krizhevsky NIPS 2012]

Figure copyright Alex Krizhevsky, Ilya

Year 2014

GoogLeNet



[Szegedy arxiv 2014]

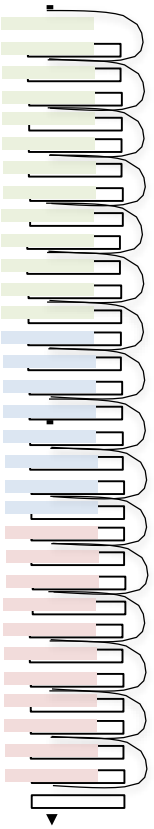
VGG

Image
conv-64
conv-64
maxpool
conv-128
conv-128
maxpool
conv-256
conv-256
maxpool
conv-512
conv-512
maxpool
conv-512
conv-512
maxpool
fc-4096
fc-4096
fc-1000
softmax

[Simonyan arxiv 2014]

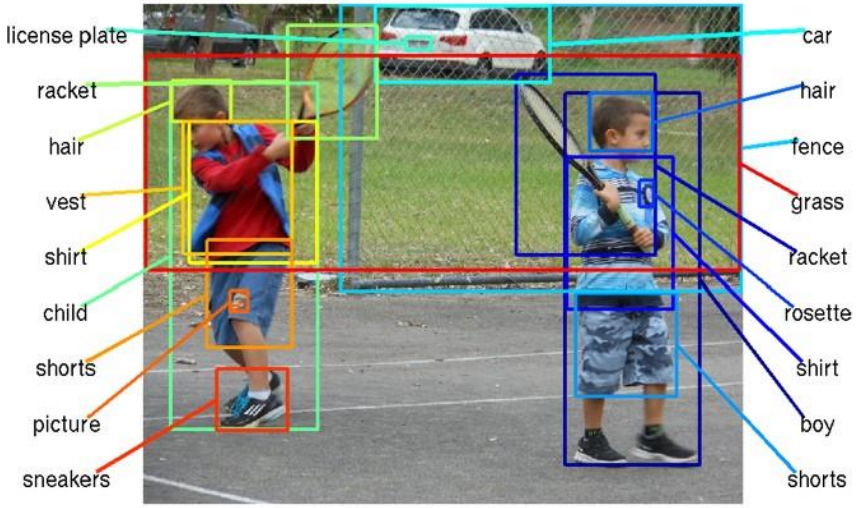
Year 2015

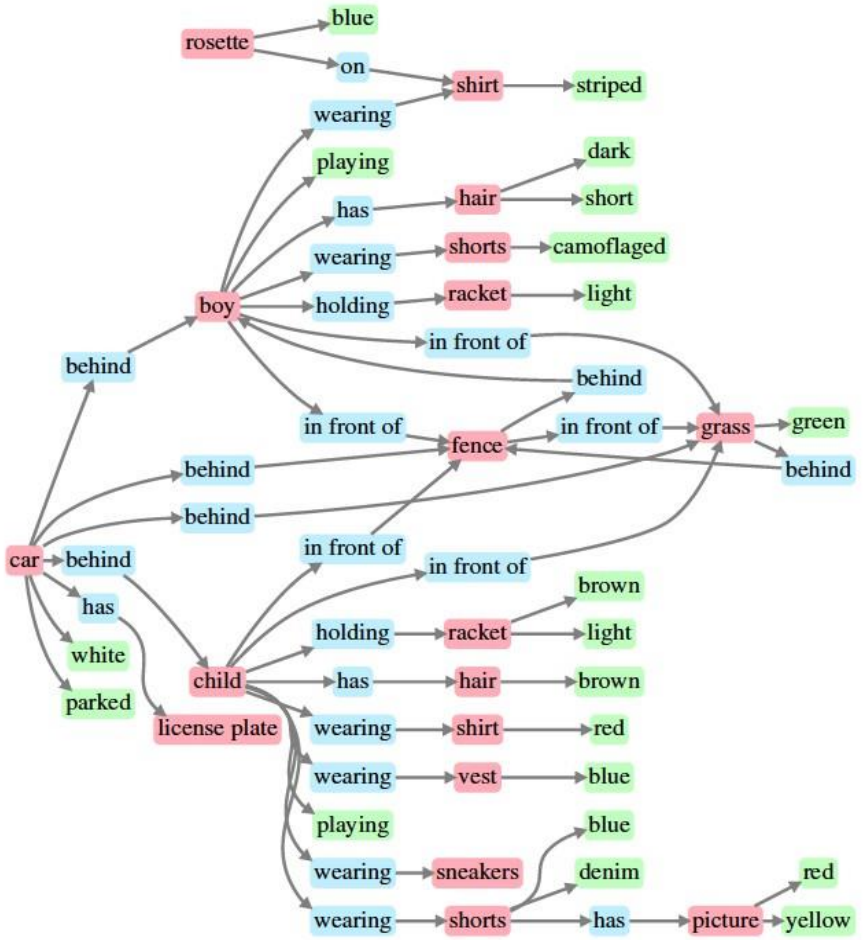
MSRA



[He ICCV 2015]

Big-data + Deep learning





Johnson *et al.*, "Image Retrieval using Scene Graphs", CVPR 2015

Big-data + Deep learning

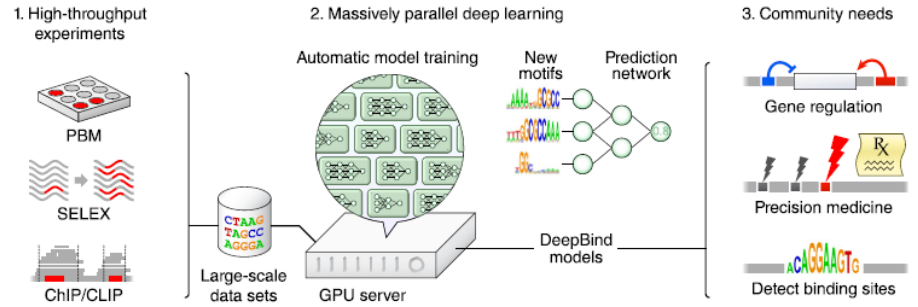


Figure 1 DeepBind's input data, training procedure and applications. 1. The sequence specificities of DNA- and RNA-binding proteins can now be measured by several types of high-throughput assay, including PBM, SELEX, and ChIP- and CLIP-seq techniques. 2. DeepBind captures these binding specificities from raw sequence data by jointly discovering new sequence motifs along with rules for combining them into a predictive binding score. Graphics processing units (GPUs) are used to automatically train high-quality models, with expert tuning allowed but not required. 3. The resulting DeepBind models can then be used to identify binding sites in test sequences and to score the effects of novel mutations.

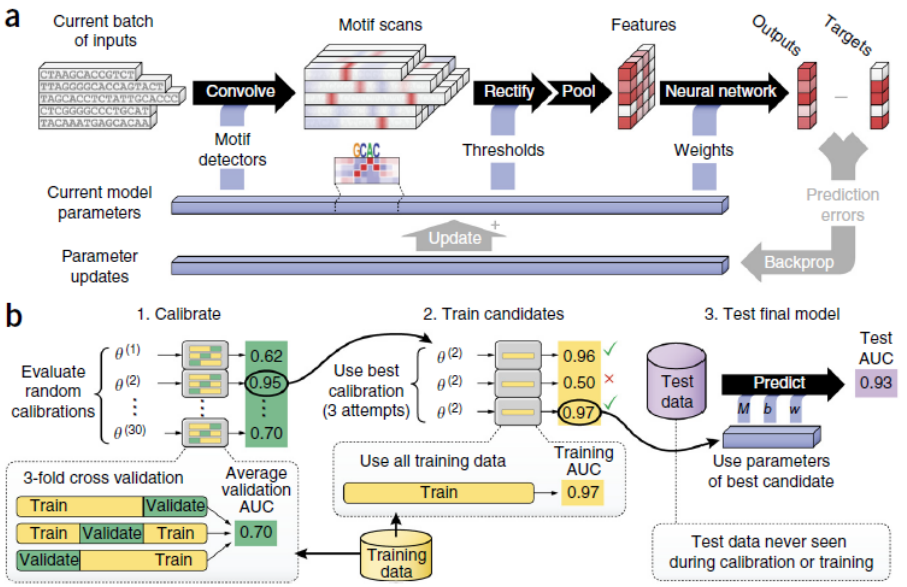
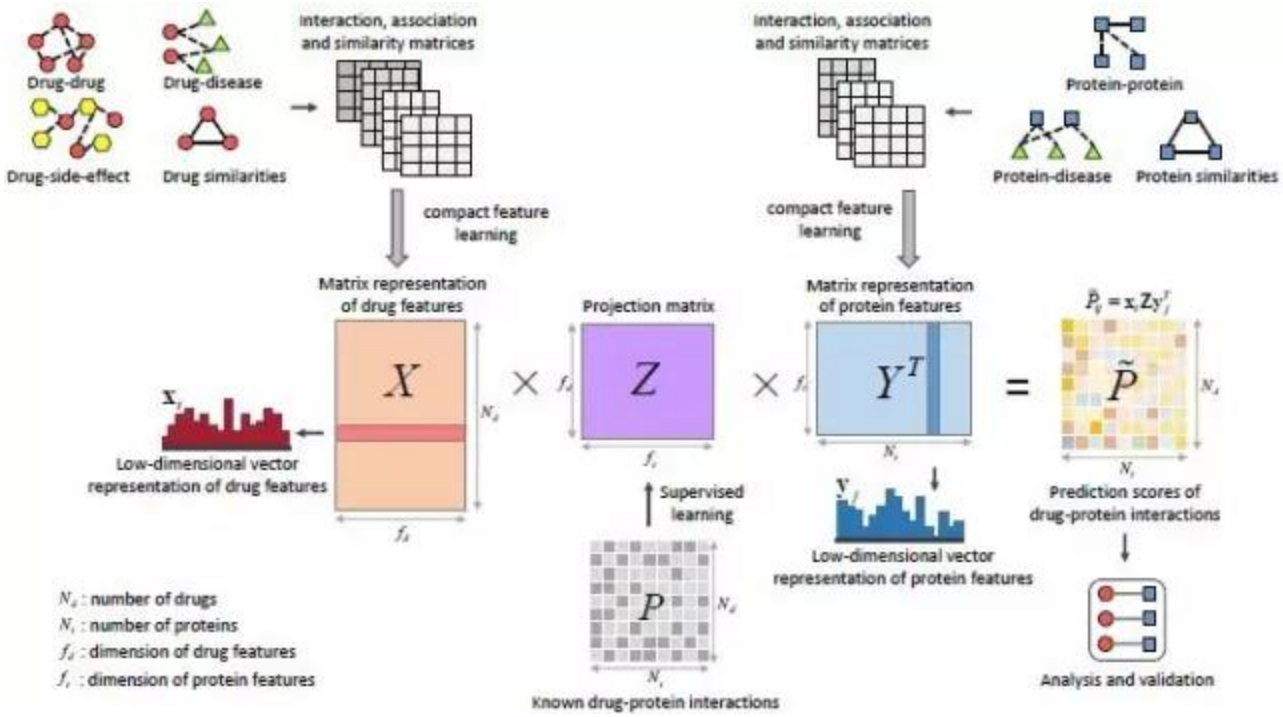


Figure 2 Details of inner workings of DeepBind and its training procedure. (**a**) Five independent sequences being processed in parallel by a single DeepBind model. The convolve, rectify, pool and neural network stages predict a separate score for each sequence using the current model parameters (**Supplementary Notes**, sec. 1). During the training phase, the backprop and update stages simultaneously update all motifs, thresholds and network weights of the model to improve prediction accuracy. (**b**) The calibration, training and testing procedure used throughout (**Supplementary Notes**, sec. 2).

Brendan Frey *et al.*, Nature Biotechnology, 2015

Big-data + Deep learning



Jianyang Zeng *et al.*, Nature Communications, 2017

课程安排

- 生物背景和课程简介
- 传统生物统计学及其应用
- 生物统计学和生物大数据挖掘
  - Hidden Markov Model (HMM)及其应用
    - Markov Chain
    - HMM理论
    - HMM和基因识别 (Topic I)
    - HMM和序列比对 (Topic II)
  - 进化树的概率模型 (Topic III)
  - Motif finding中的概率模型 (Topic IV)
    - EM algorithm
    - Markov Chain Monte Carlo (MCMC)
  - 基因表达数据分析 (Topic V)
    - 聚类分析-Mixture model
    - Classification-Lasso Based variable selection
  - 基因网络推断 (Topic VI)
    - Bayesian网络
    - Gaussian Graphical Model
  - 基因网络分析 (Topic VII)
    - Network clustering
    - Network Motif
    - Markov random field (MRF)
  - Dimension reduction及其应用 (Topic VIII)
- 面向生物大数据挖掘的深度学习

研究对象:
生物序列,
进化树,
生物网络,
基因表达
...

方法:
生物计算与生物统计

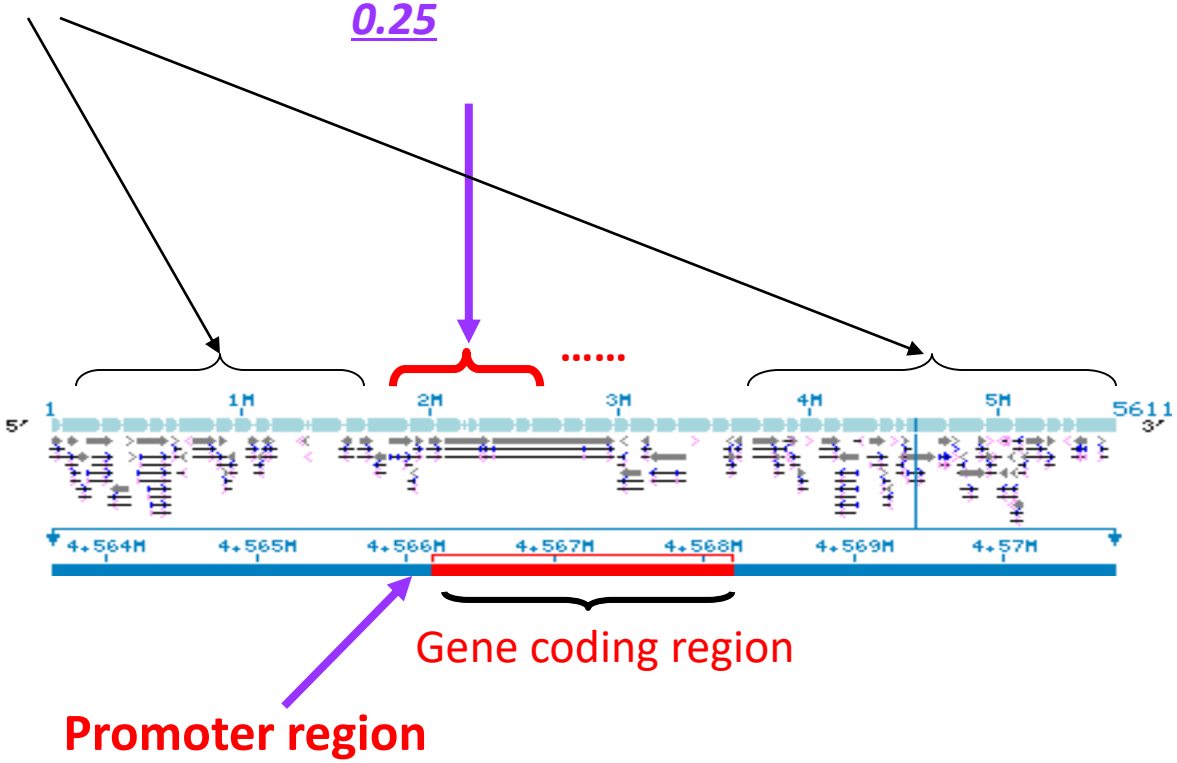
Topic I: Sequence's Feature Detection

- Problem I: CpG island finding
- Problem II: Gene finding (promoter prediction, Splicing site prediction, Translation Initial Site Prediction etc.)
- Hidden Markov Model is a powerful method for these problem

什么是CpG岛？

CG-poor regions: P(CG) ~ 0.07!

CG-rich region: P(CG) ~ 0.25



CpG岛的生物学意义

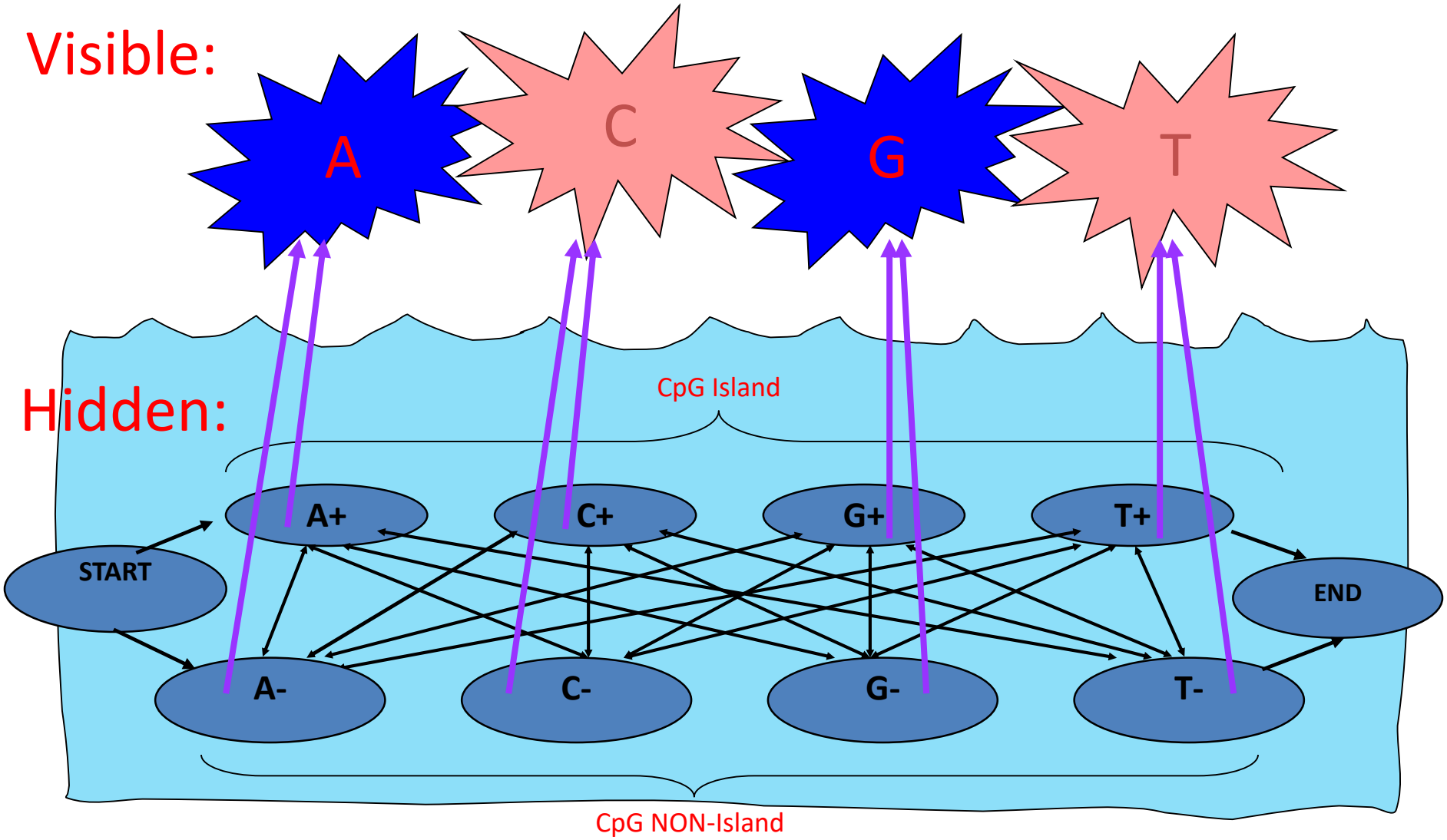
- CpG-rich regions are associated with genes which are *frequently transcribed*.
- Helps to understand gene expression related to *location* in genome.

HMM对于CpG岛识别的意义

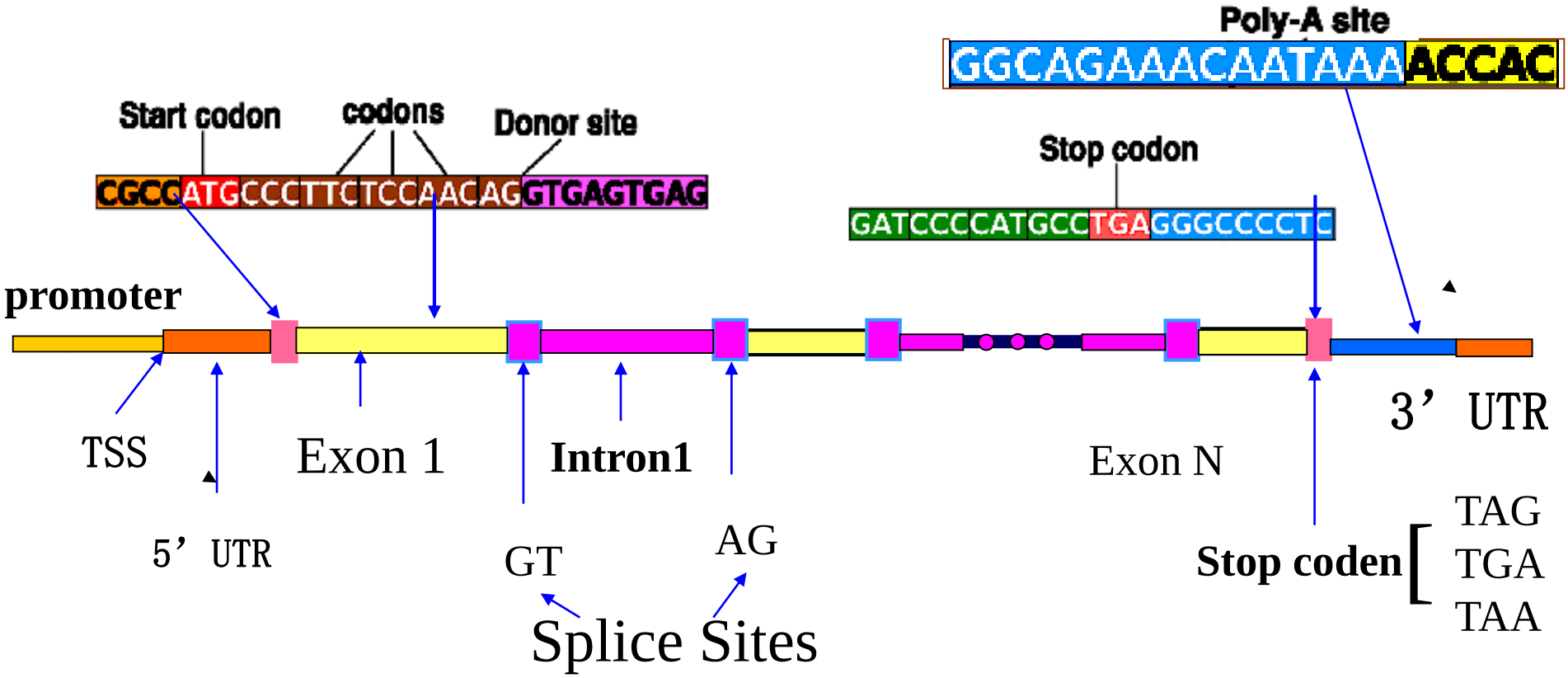
- Q: Why an HMM?
- It can answer the questions:
  - Short sequence: *does it come from a CpG island or not?*
  - Long sequence: *where are the CpG islands?*
- So, what's a good model?
  - Well, we need states for ISLAND bases and

    NON-ISLAND bases …

HMM示意


Visible:
A
C
G
T
Hidden:
CpG Island
A+
C+
G+
T+
START
END
A-
C-
G-
T-
CpG NON-Island

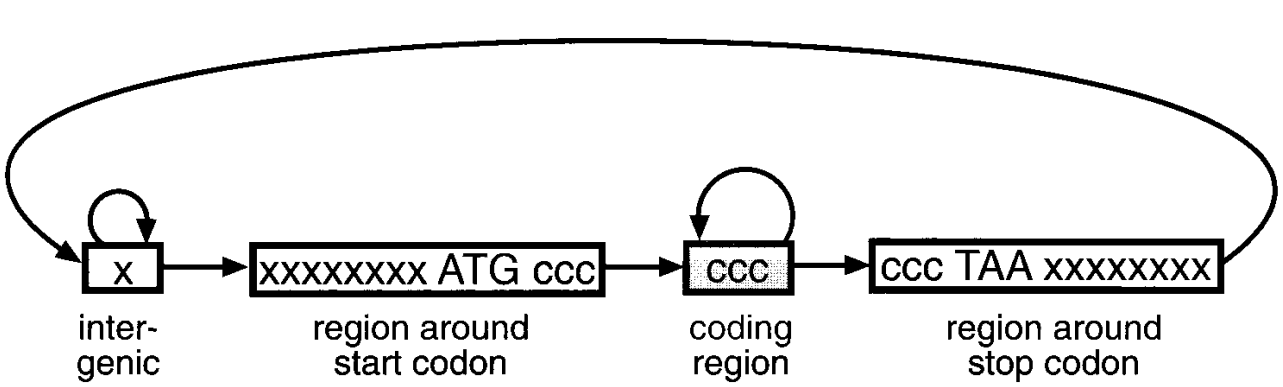
基因的结构


Poly-A site
GGCAGAAACAATAAAACCAC
Start codon
codons
Donor site
CGCCATGCCCTTCTCCAACAGGTGAGTGAG
Stop codon
GATCCCCATGCCTGAGGGCCCCTC
promoter
TSS
5’ UTR
Exon 1
Intron1
GT
AG
Splice Sites
Exon N
3’ UTR
Stop coden
TAG
TGA
TAA

HMMs and Gene Structure

- Nucleotides *{A,C,G,T}* are the observables
- Different states generates generate nucleotides at different frequencies

  A simple HMM for unspliced genes:

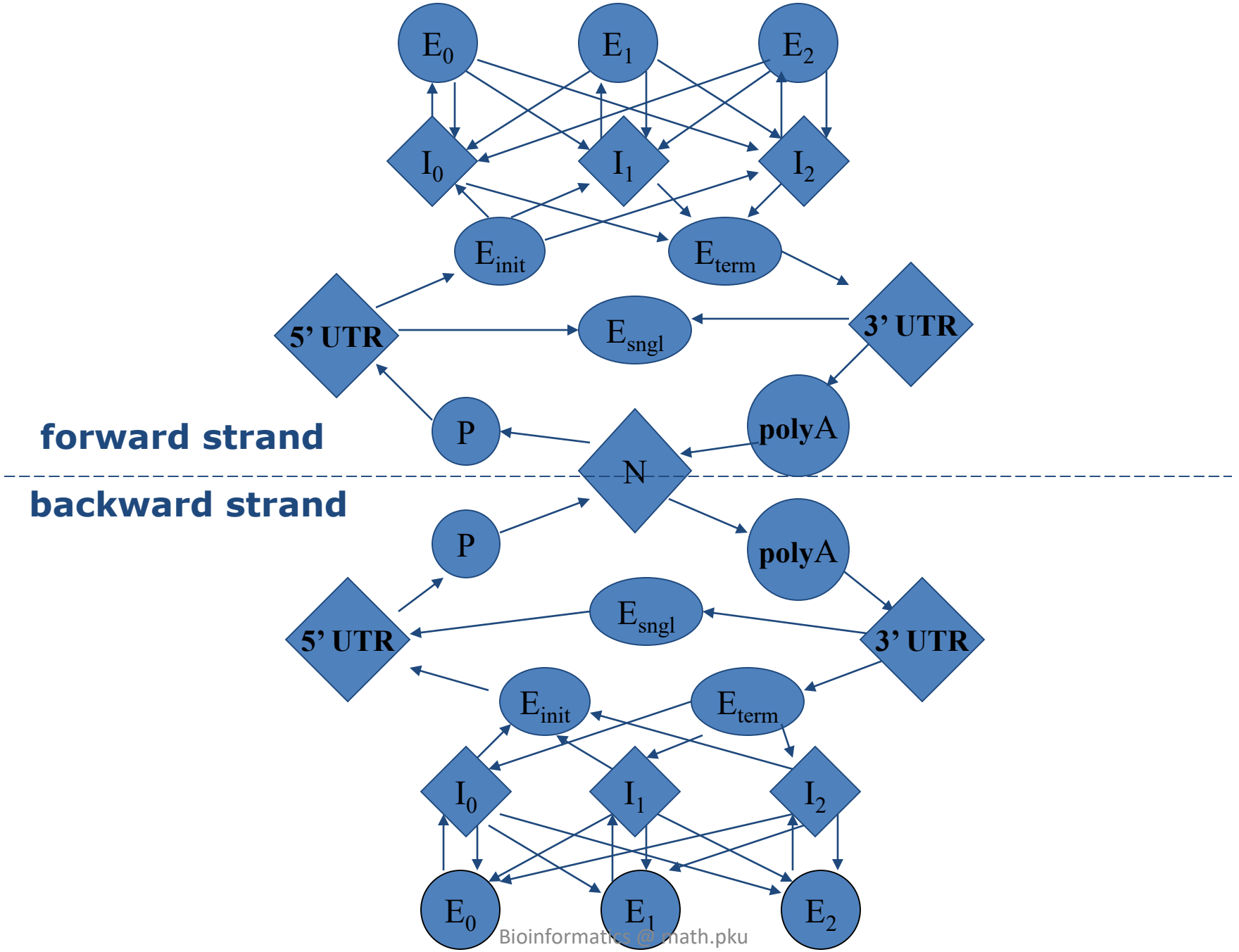


  AAAGC ATG CAT TTA ACG AGA GCA CAA GGG CTC TAA TGCCG

- The sequence of states is an annotation of the generated string – each nucleotide is generated in intergenic, start/stop, coding state

Genscan

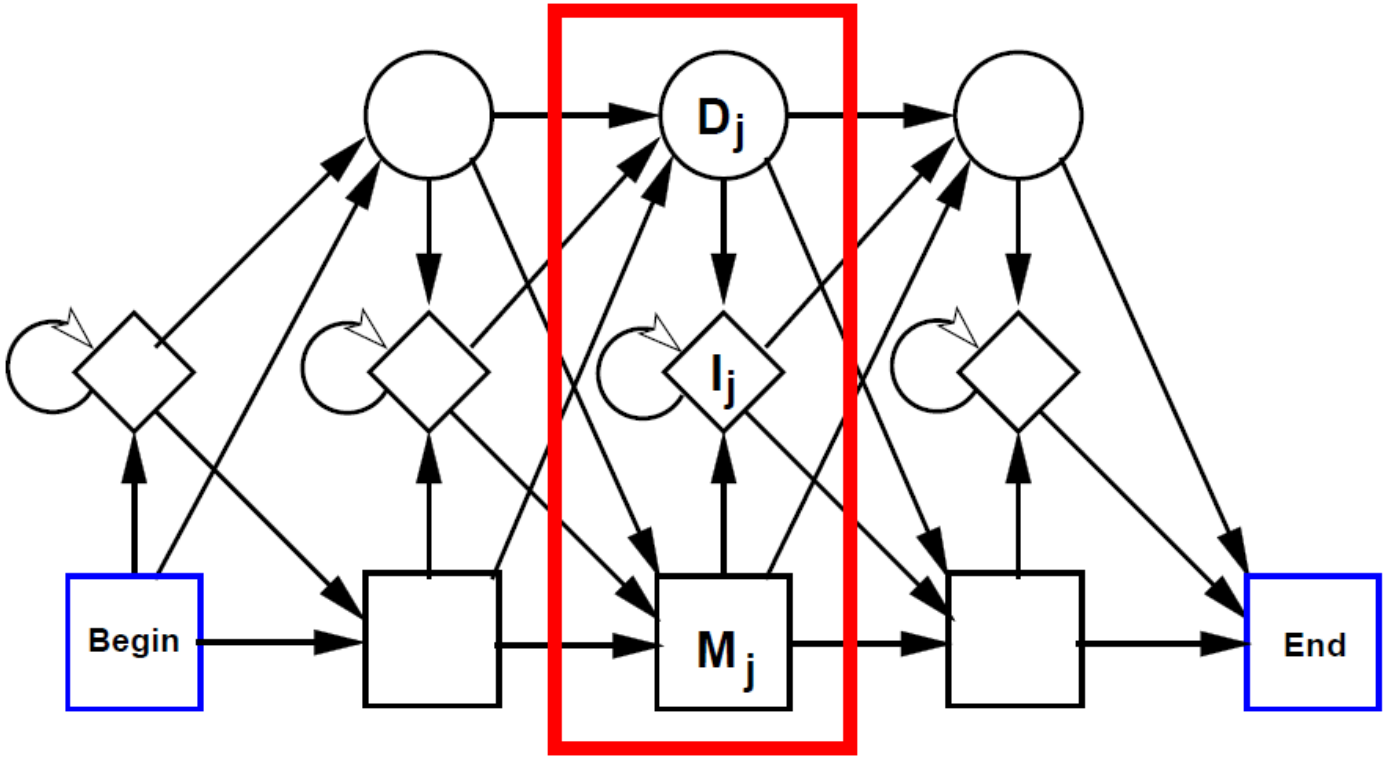
- Developed by Chris Burge 1997
- One of the most accurate *ab initio* programs
- Uses explicit state duration HMM to model gene structure (different length distributions for exons)
- Different model parameters for regions with different GC content



Topic II: Multiple Alignment

	X	X	.	.	.	X
bat	A	G	-	-	-	C
rat	A	-	A	G	-	C
cat	A	G	-	A	A	-
gnat	-	-	A	A	A	C
goat	A	G	-	-	-	C
	1	2	.	.	.	3

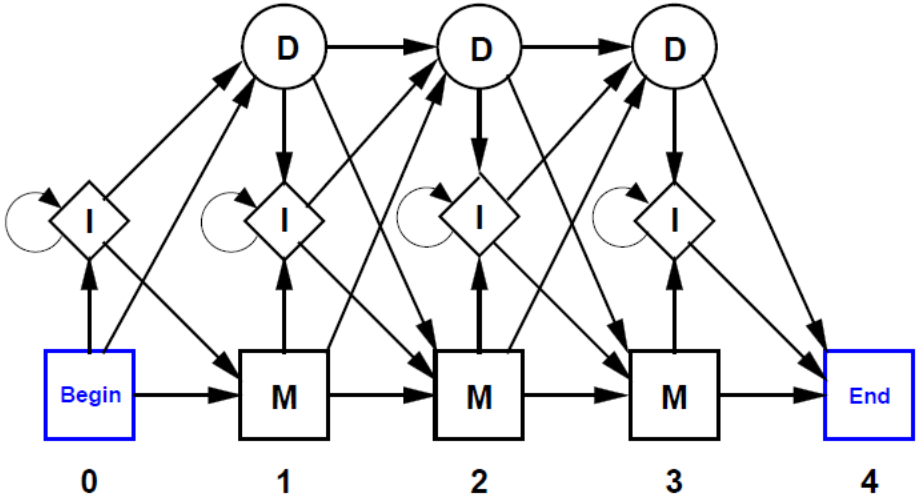
Profiled HMM



Transition structure of a profile HMM

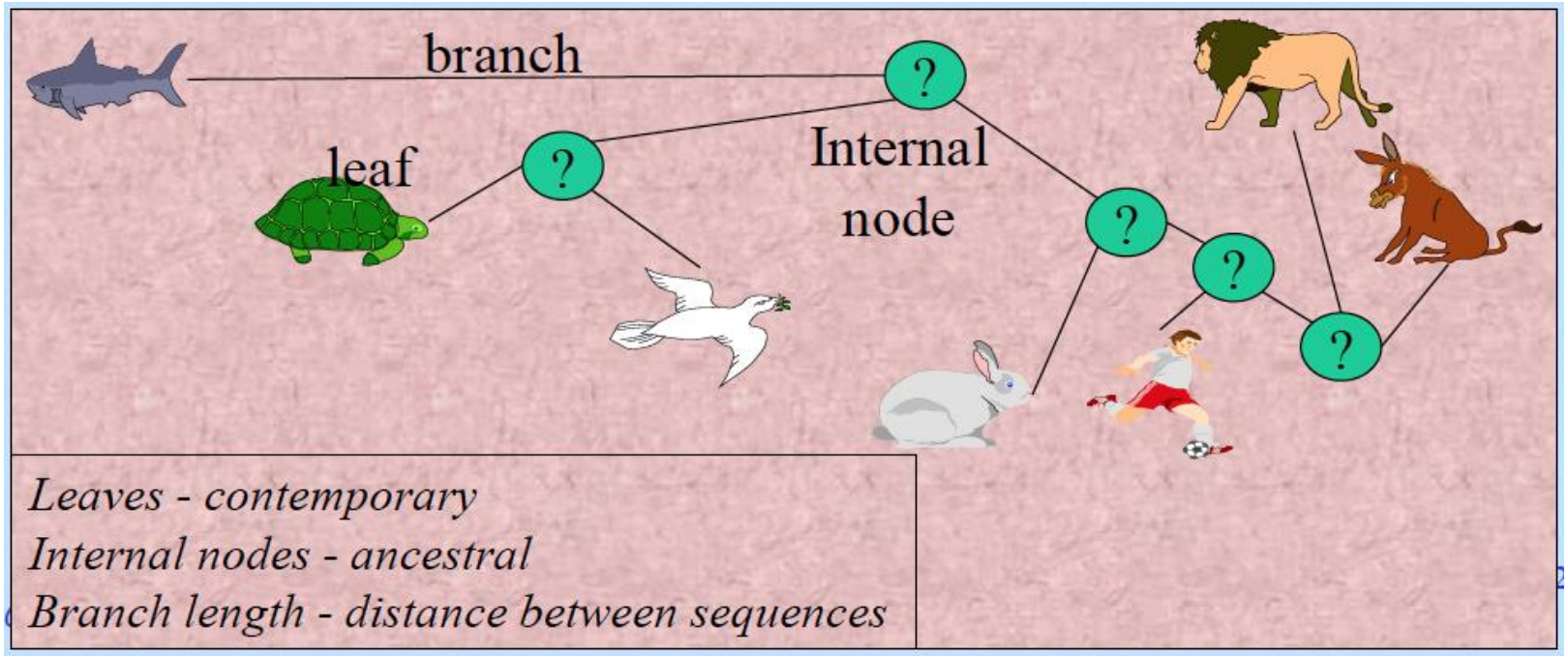
Example of Profile HMM

	X	X	.	.	.	X
bat	A	G	-	-	-	C
rat	A	-	A	G	-	C
cat	A	G	-	A	A	-
gnat	-	-	A	A	A	C
goat	A	G	-	-	-	C
	1	2	.	.	.	3



Topic III: Tree of life

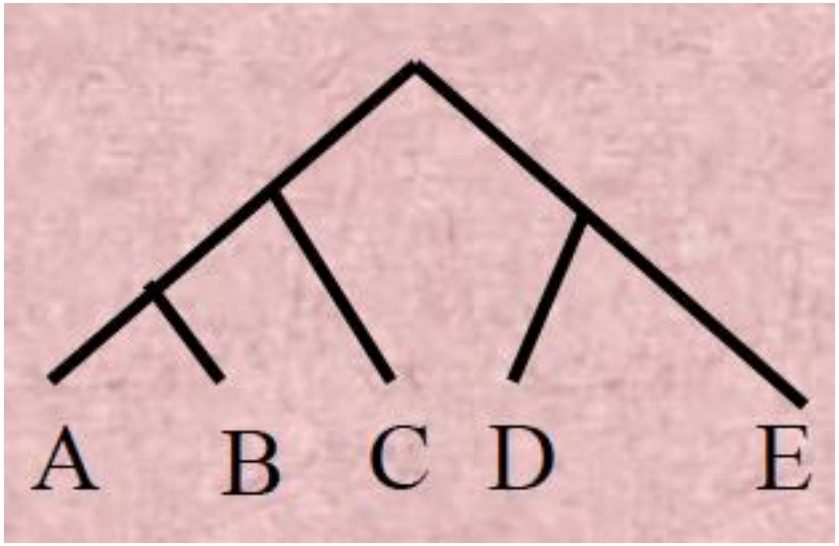
- Phylogeny: the ancestral relationship of a set of species
- Represented by a phylogenetic tree



Inferring a phylogenetic tree

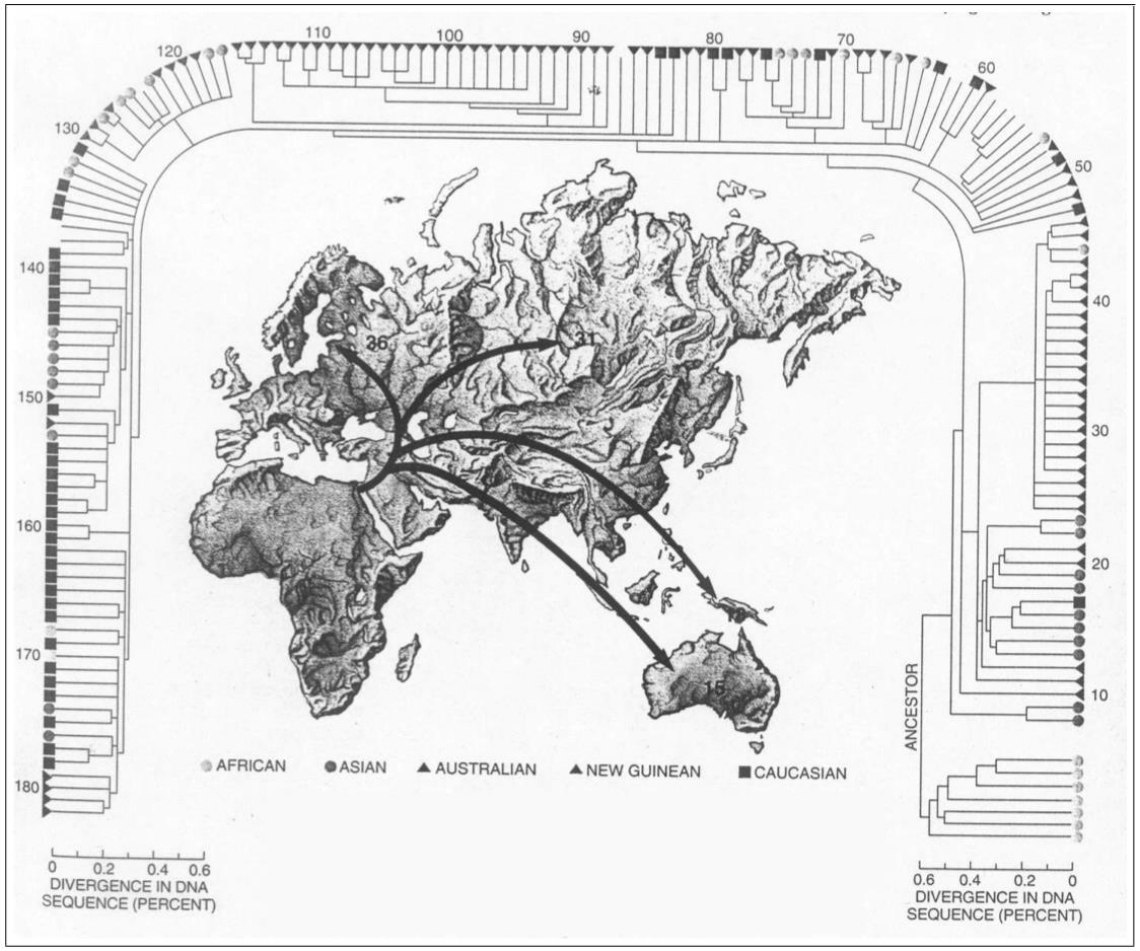
- Classical: morphological characters
- Modern: molecular sequences

```
A: CAGGTA
B: CAGACA
C: CGGGTA
D: TGCACT
E: TGCGTA
```



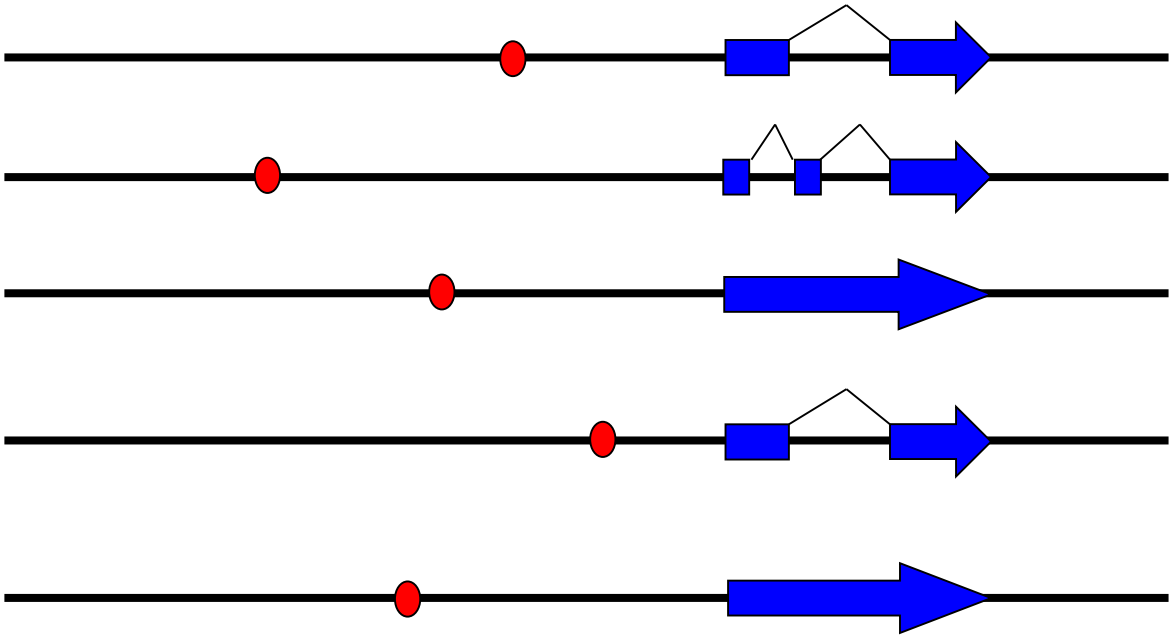
- Approaches: probabilistic model, bootstrap

An example: Out of Africa



Topic IV: Motif Finding

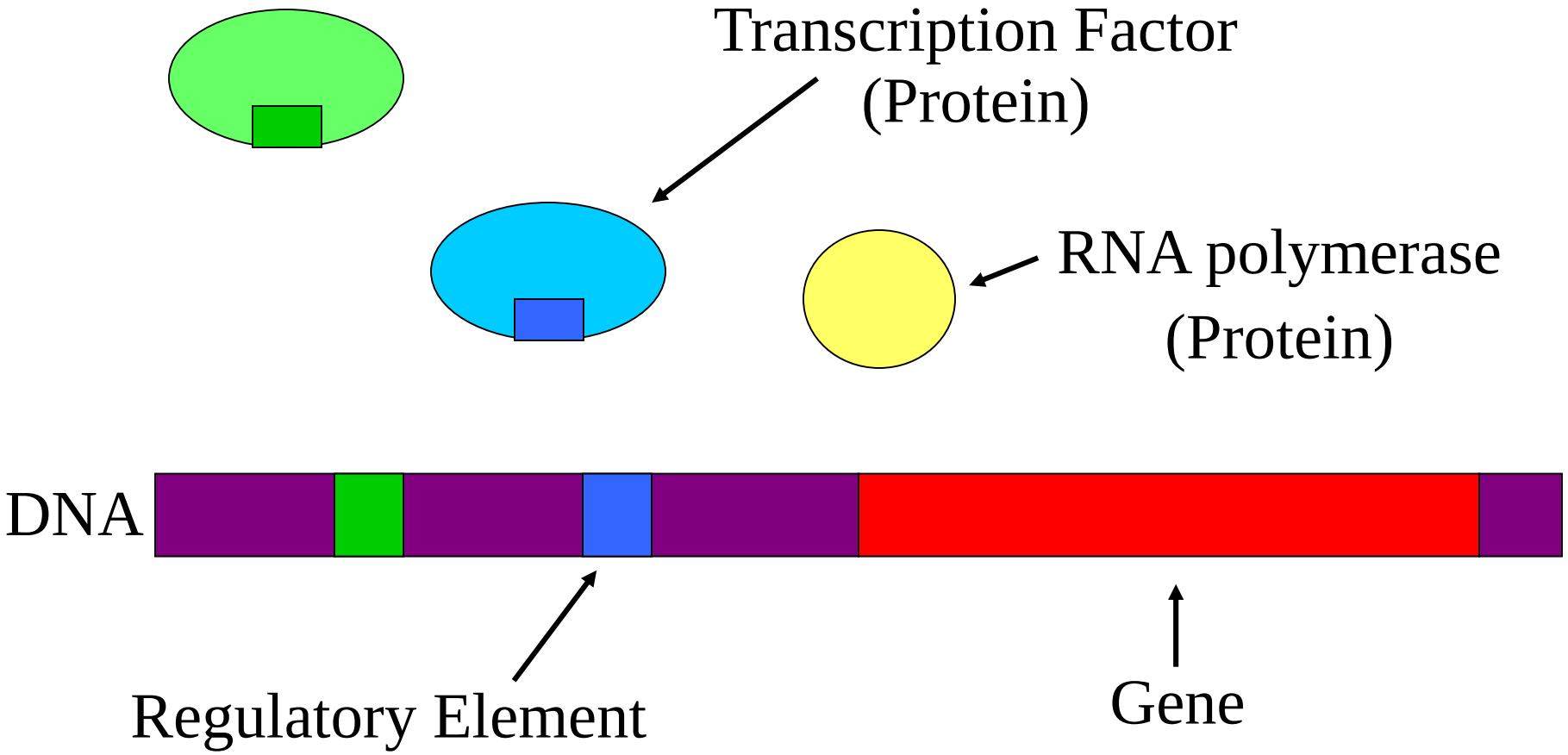
- Find promoter motifs associated with **co-regulated** or **functionally related** genes



Transcriptional Regulation

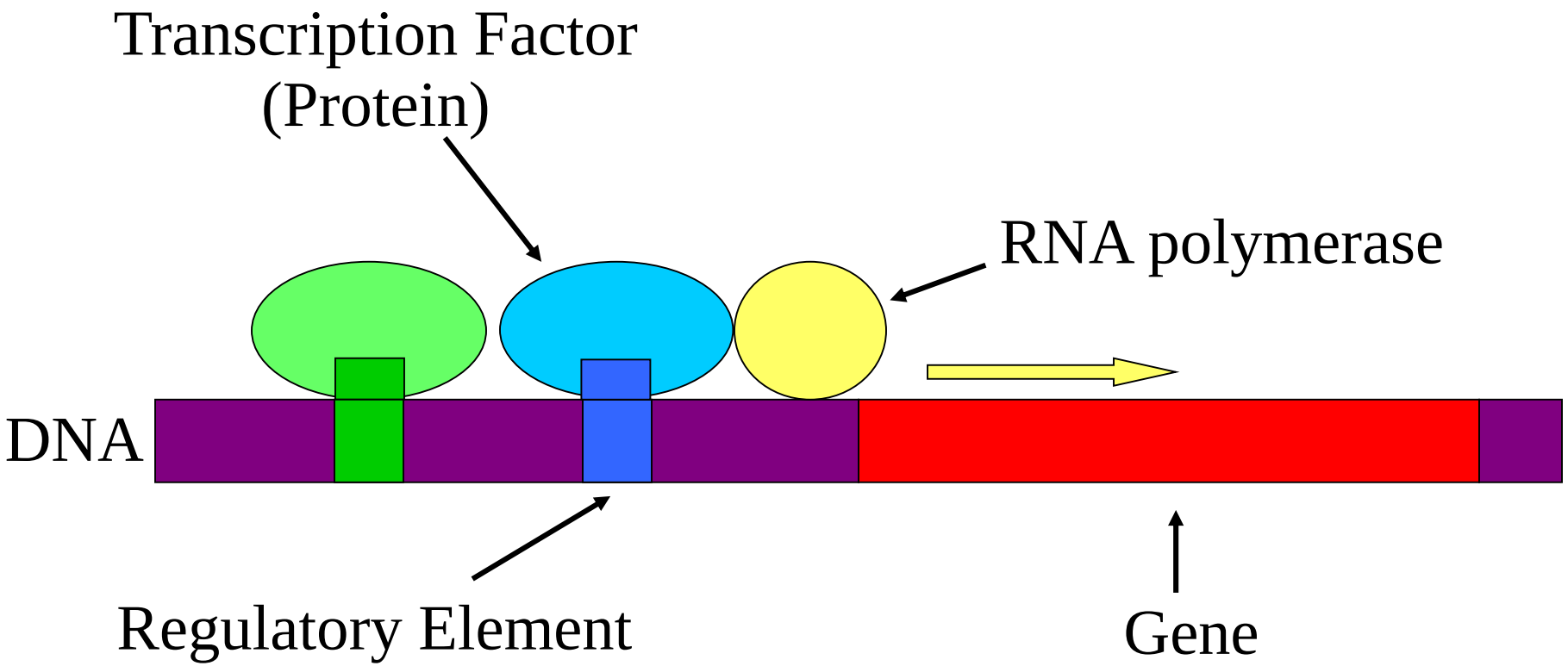
- The transcription of each gene is controlled by a regulatory region of DNA relatively near the transcription start site (TSS).
- two types of fundamental components
  - short DNA regulatory elements
  - *gene regulatory proteins* that recognize and bind to them.

Regulation of Genes



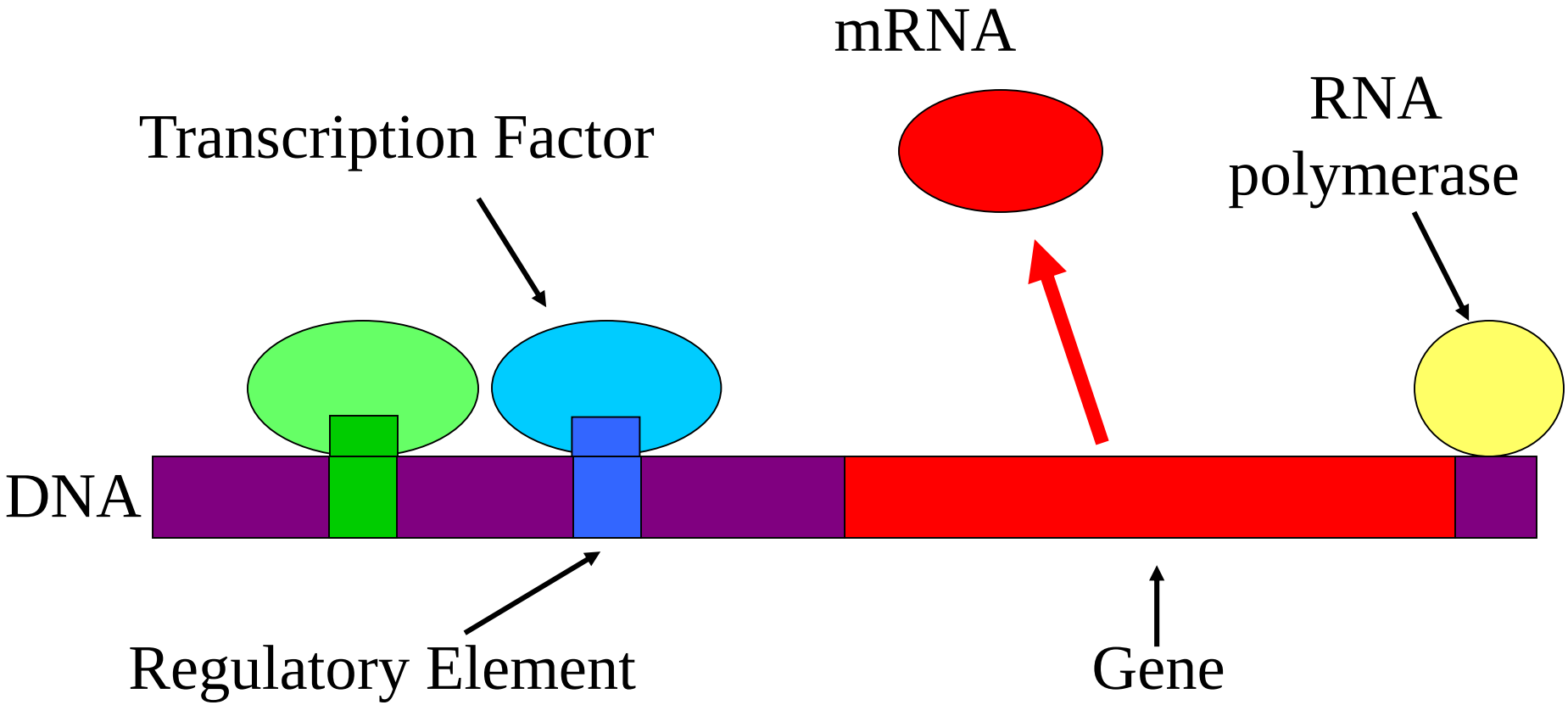
source: M. Tompa, U. of Washington

Regulation of Genes



source: M. Tompa, U. of Washington

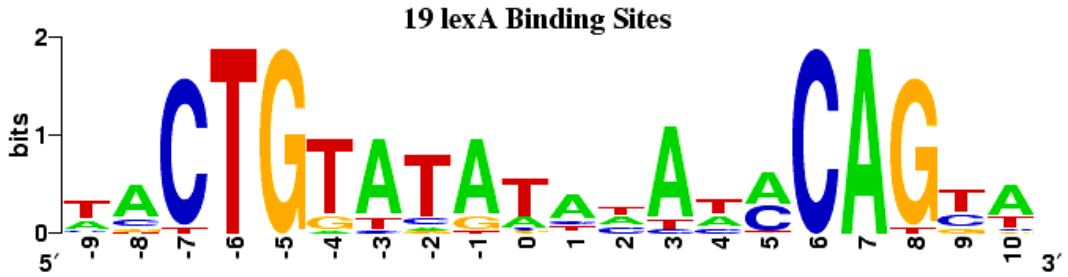
Regulation of Genes



source: M. Tompa, U. of Washington

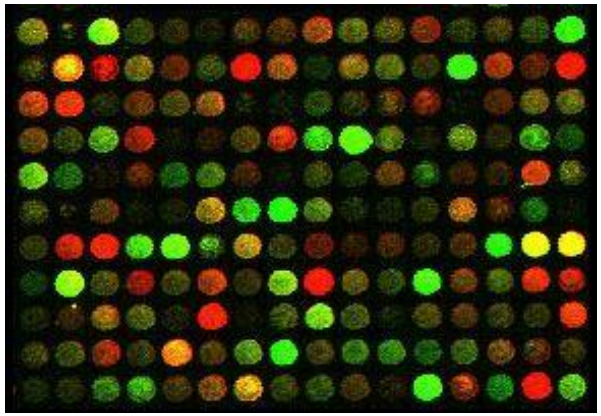
Motif Finding Problem

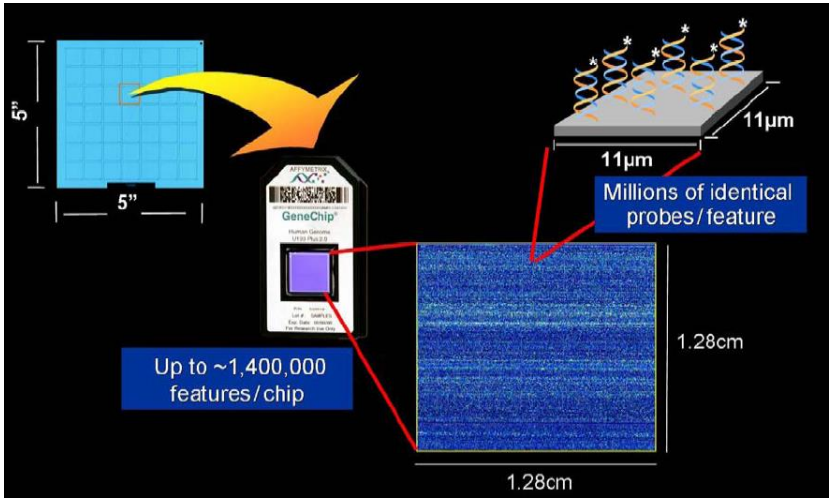
- Characterizing the motif: Positional weight matrix



- Finding the motif
  - Gibbs Sampler (AlignACE)
  - EM algorithm (MEME)

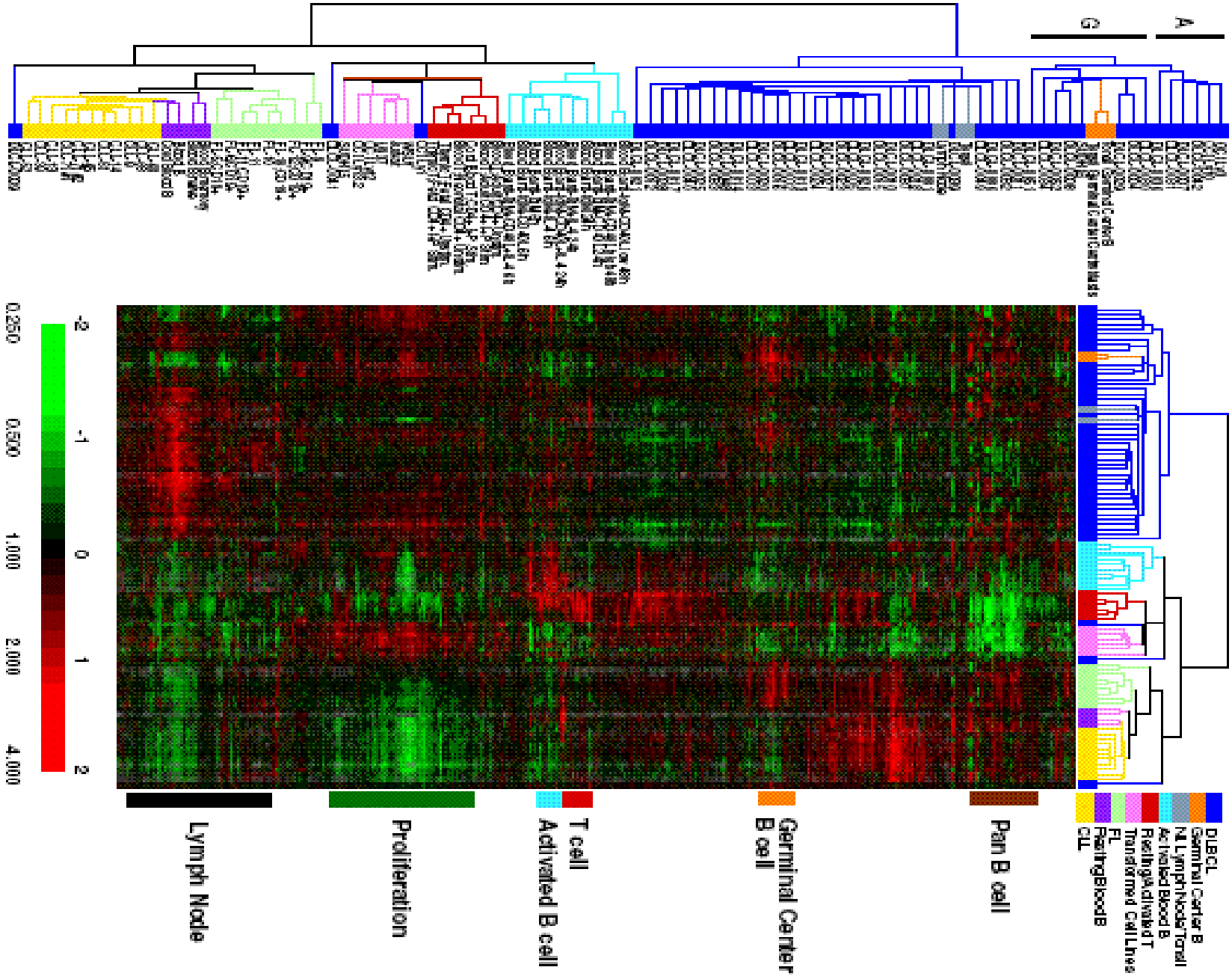
Topic V: Gene Expression Data Clustering and Biomarker Discovery





Microarrays

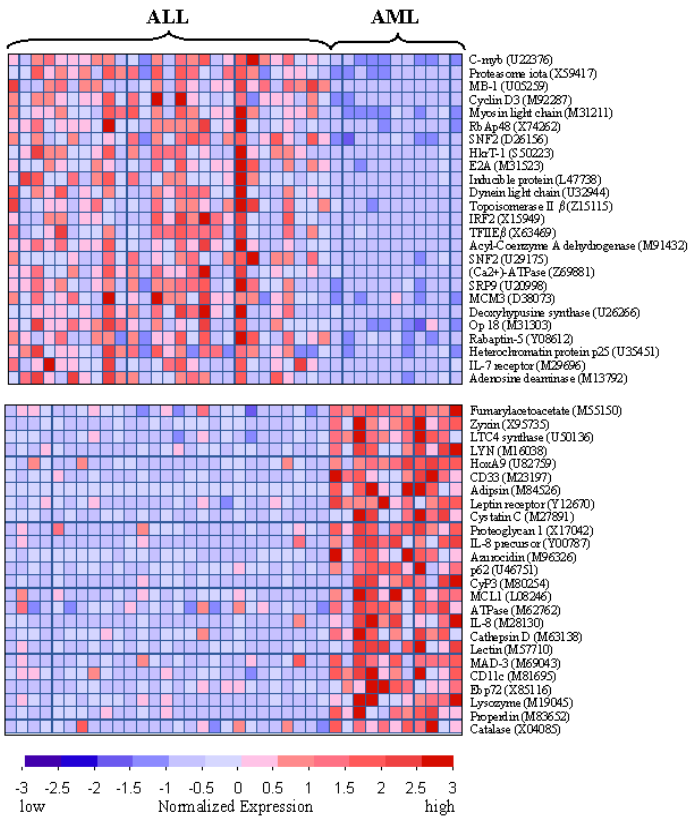
- ***DNA microarray*** technology rely on the hybridization properties of nucleic acids to monitor DNA or RNA abundance on a genomic scale in different types of cells.



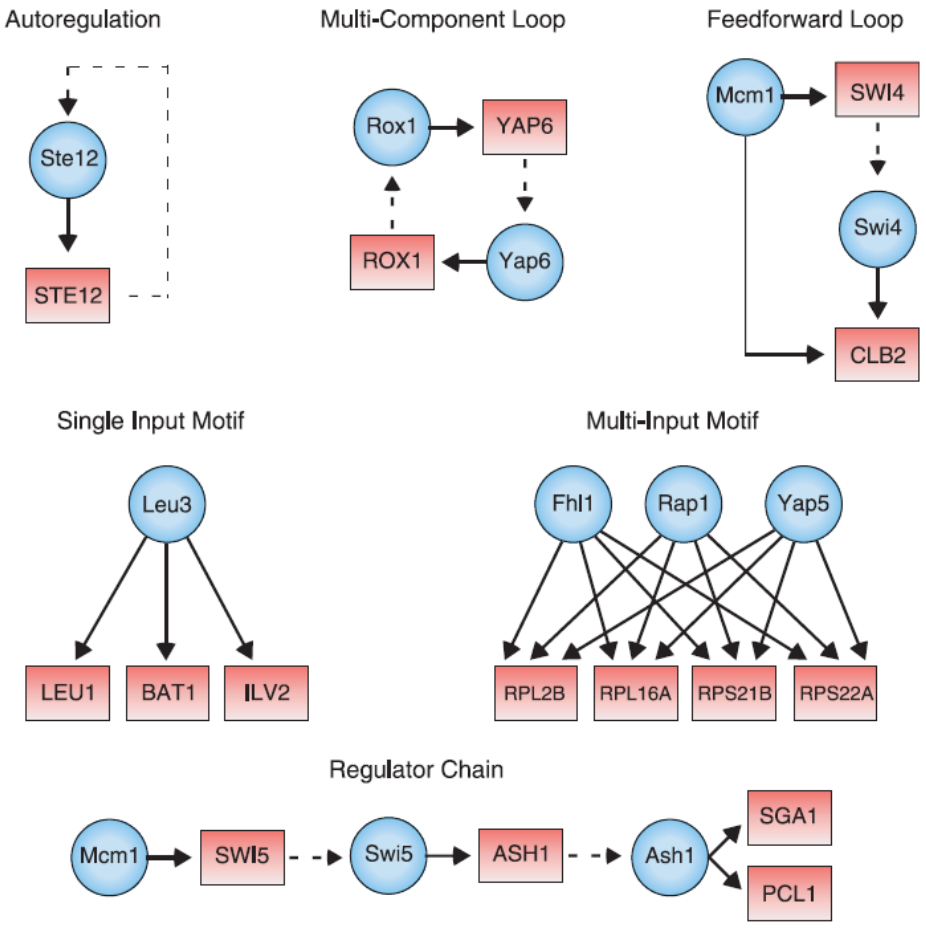
Alizadeh et al., Nature 403:503-11, 2000

Classification

- Given: the case, control gene expression data
- Find: a set of genes (biomarker) can discriminate two classes
- Method: variable selection



Topic VI: Regulatory Network Inference from Gene Expression Data



Lee et al. Science 298: 799(2002)

Network Inference: Reverse Engineering

- Given: a large set of gene expression observations
- Find:
  - Wiring diagram
  - Transition rules

  To fit the observation data
- Methods
  - Bayesian Network
  - Gaussian graphical model

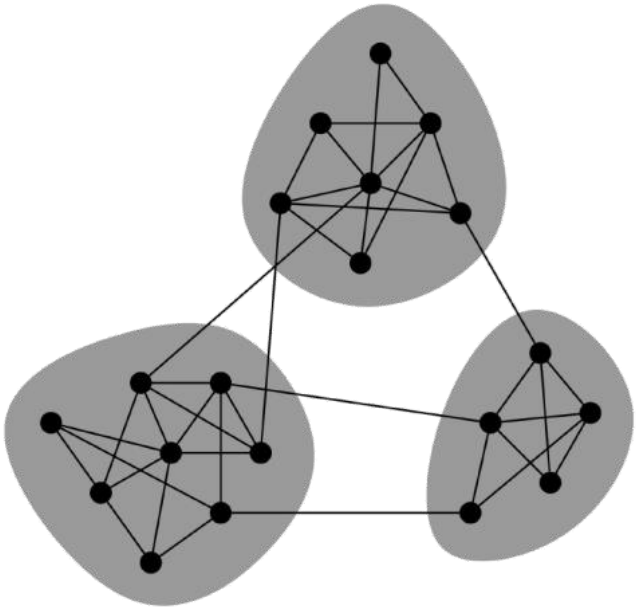
Dream Project

- DREAM: Dialogue for Reverse Engineering Assessments and Methods

- http://dreamchallenges.org/

Topic VII: Network Analysis

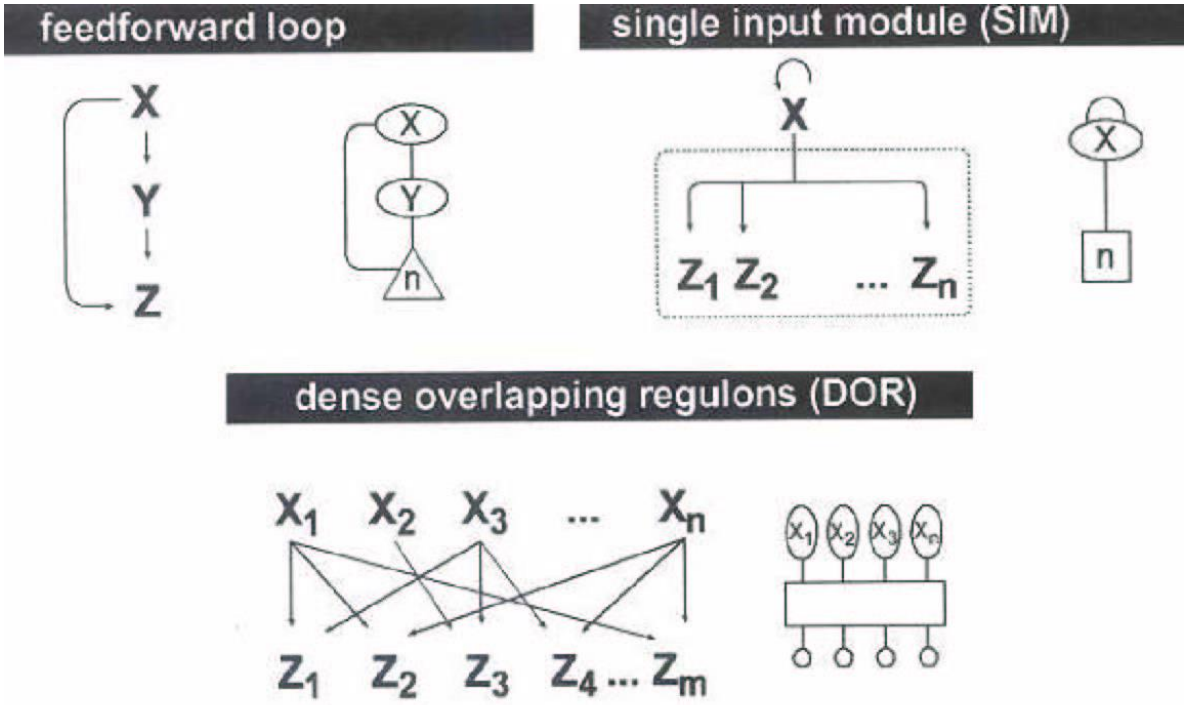
- Network modular (network clustering)



Network Motif

- Definition: Patterns of interconnections occurring in complex networks at numbers that are significantly higher than those in randomized networks (Milo, R., et. al. *Science* **298,** 824–827)

Network Motifs



Milo et al. *Science*, **298**, 2002

Topic VIII: Dimension Reduction

- Curse of dimensionality

- Visualization in low dimension

Curse of Dimensionality

- A major problem is *the curse of dimensionality.*
- If the data x lies in high dimensional space, then an enormous amount of data is required to learn distributions or decision rules.
- Example: 50 dimensions. Each dimension has 20 levels. This gives a total of 20^{50} cells. But the no. of data samples will be far less. There will not be enough data samples to learn.

Curse of Dimensionality

- One way to deal with dimensionality is to assume that we know the form of the probability distribution.
- For example, a Gaussian model in N dimensions has N + N(N-1)/2 parameters to estimate.
- Requires $O(N^2)$ data to learn reliably. This may be practical.

Dimension Reduction

- One way to avoid the curse of dimensionality is by projecting the data onto a lower-dimensional space.
- Techniques for dimension reduction:
  - Principal Component Analysis (PCA)
  - Singular value decomposition (SVD)
  - Multi-dimensional Scaling (MDS).

Recap

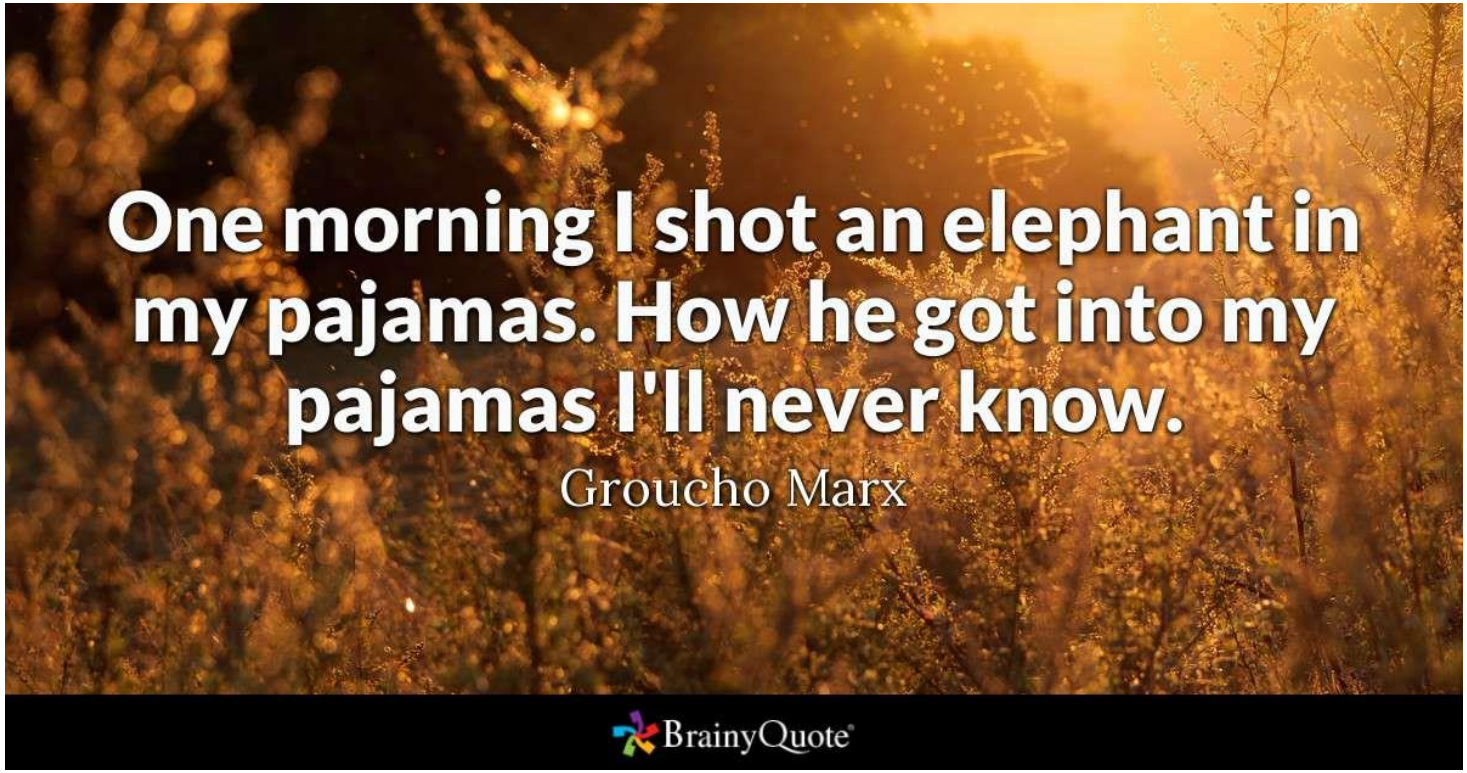
- a breif summary...

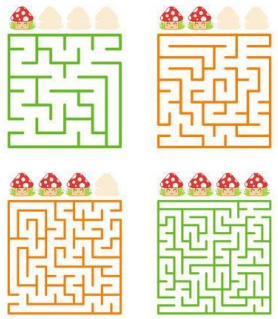
- Think again:
  - Why biostatistics
  - What is biostatistical modeling
  - Where is the applications

Statistical modeling

数据无处不在，复杂性无处不在。。。

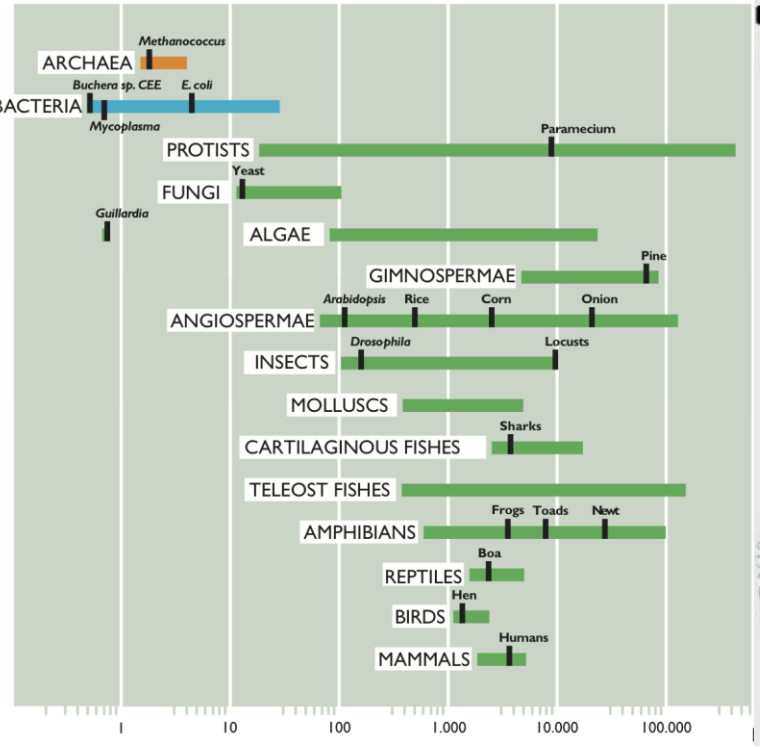
一天早上我在我的睡衣里打死一只大象，他怎么跑到我睡衣里来的，我就不知道了。

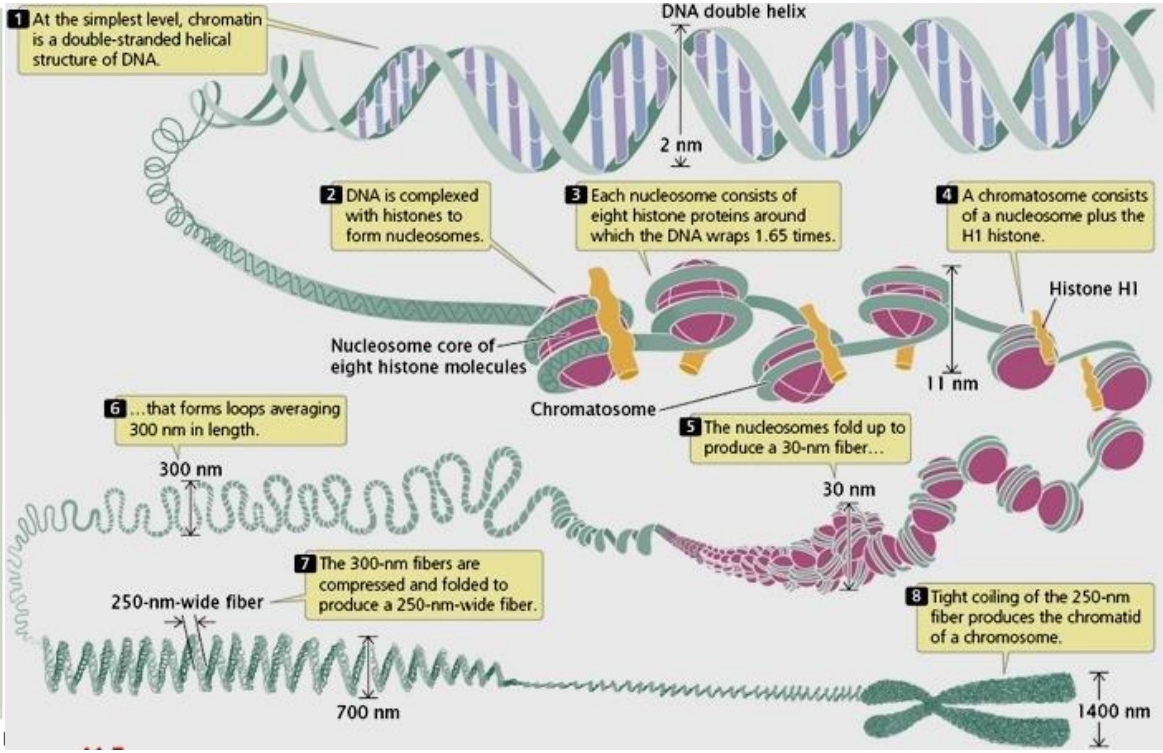




Statistical modeling

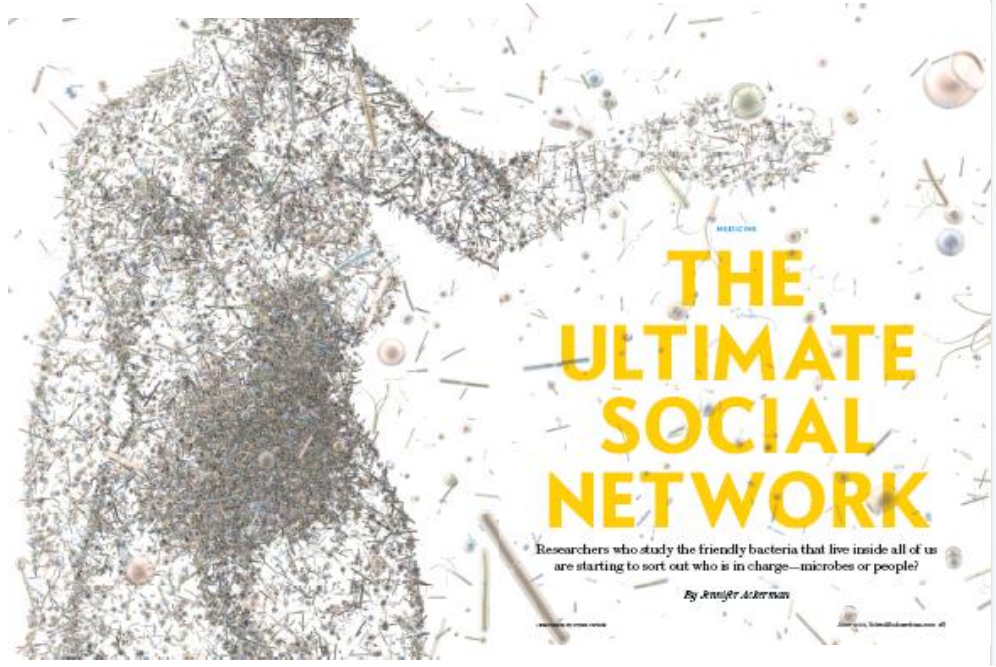
数据无处不在，复杂性无处不在。。。

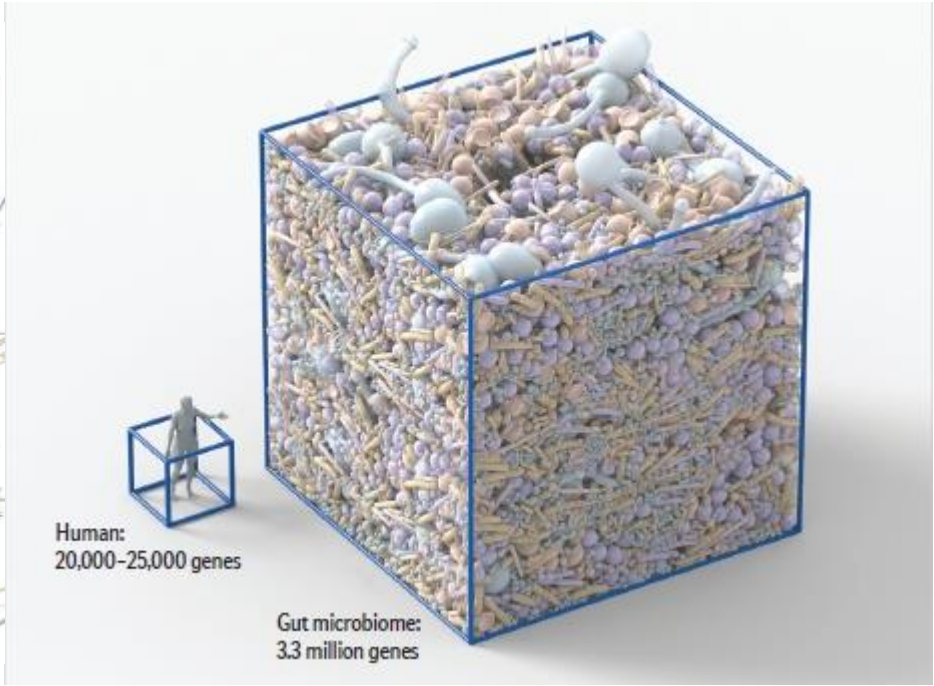




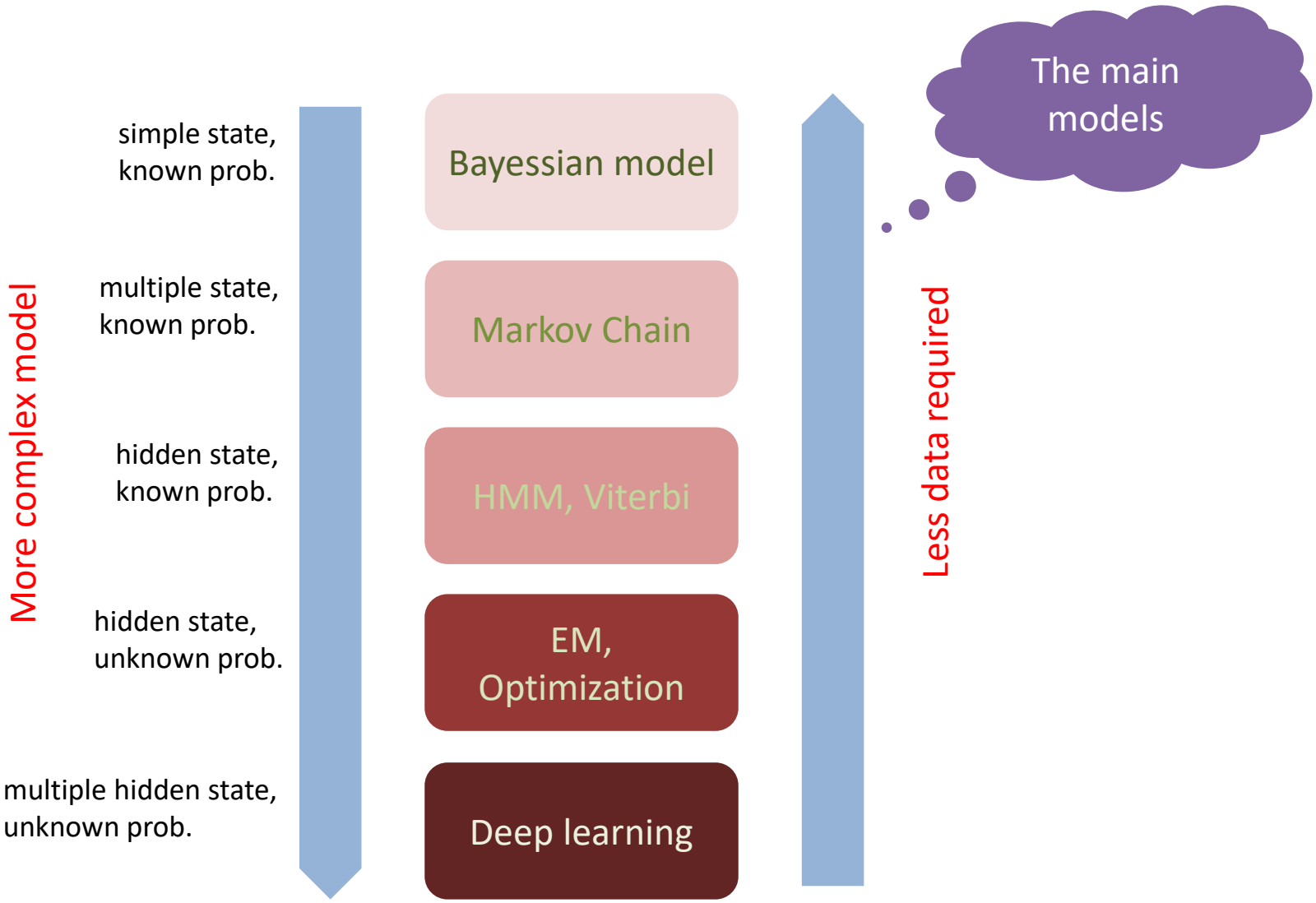
Statistical modeling

数据无处不在，复杂性无处不在。。。

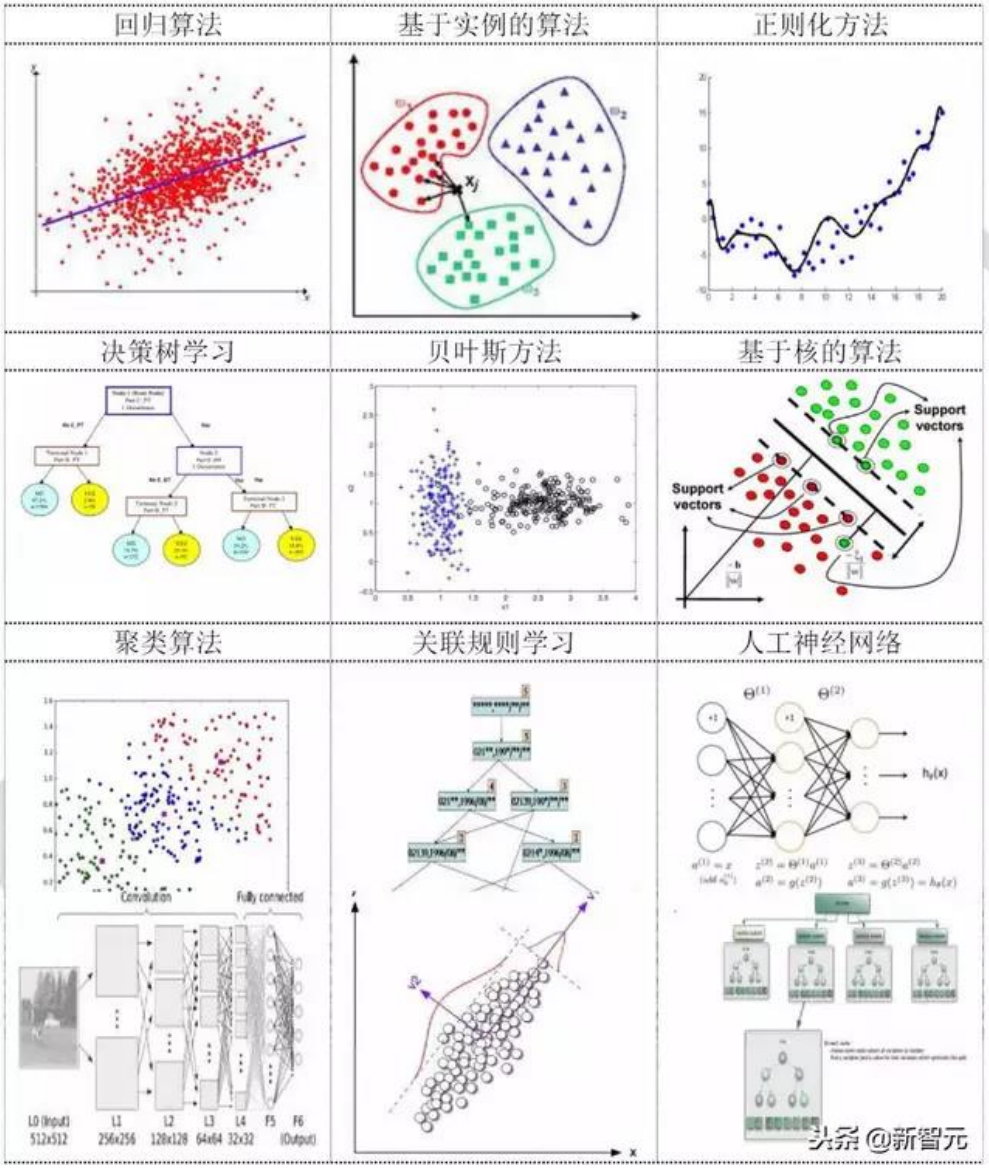




Statistical modeling



Statistical modeling



课程安排

- 生物背景和课程简介
- 传统生物统计学及其应用
- 生物统计学和生物大数据挖掘
  - Hidden Markov Model (HMM)及其应用
    - Markov Chain
    - HMM理论
    - HMM和基因识别 (Topic I)
    - HMM和序列比对 (Topic II)
  - 进化树的概率模型 (Topic III)
  - Motif finding中的概率模型 (Topic IV)
    - EM algorithm
    - Markov Chain Monte Carlo (MCMC)
  - 基因表达数据分析 (Topic V)
    - 聚类分析-Mixture model
    - Classification-Lasso Based variable selection
  - 基因网络推断 (Topic VI)
    - Bayesian网络
    - Gaussian Graphical Model
  - 基因网络分析 (Topic VII)
    - Network clustering
    - Network Motif
    - Markov random field (MRF)
  - Dimension reduction及其应用 (Topic VIII)
- 面向生物大数据挖掘的深度学习

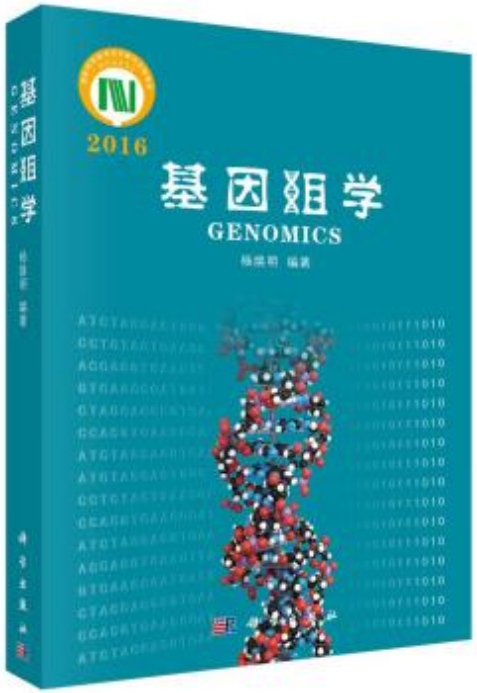
研究对象:
生物序列,
进化树,
生物网络,
基因表达
...

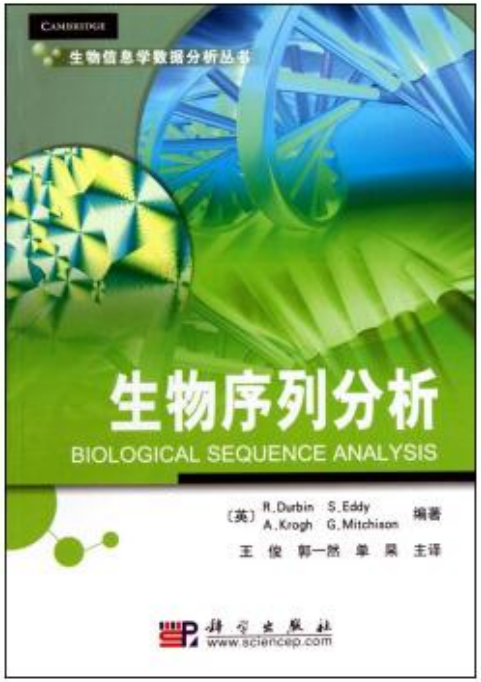
方法:
生物计算与生物统计

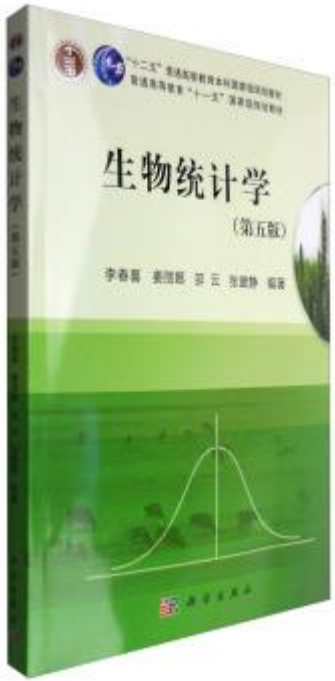
References

- James D. Watson, Tania A. Baker, Stephen P. Bell. Molecular Biology of the Gene. Benjamin-Cummings Publishing Company. 2008.
- Bruce Alberts. Molecular Biology of the Cell. Garland Publishing Inc. 2007.
- Jocelyn E. Krebs, Stephen T. Kilpatrick, Elliott S. Goldstein. Lewin‘s Genes XI. Jones and Bartlett Publishers, Inc. 2012.

References


2016
基因组学
GENOMICS


CAMBRIDGE
生物信息学数据分析丛书
生物序列分析
BIOLOGICAL SEQUENCE ANALYSIS
R.Durbin S.Eddy
A.Krogh G.Mitchison
编著
王俊 郭一然 单昊 主译
www.sciencep.com


生物统计学
(第五版)

补充知识

- 大数据；
- 高通量测序；
- 生物统计学经典软件；
- 生物统计学核心工具；
- 深度学习。

1. 大数据

生物数据很大

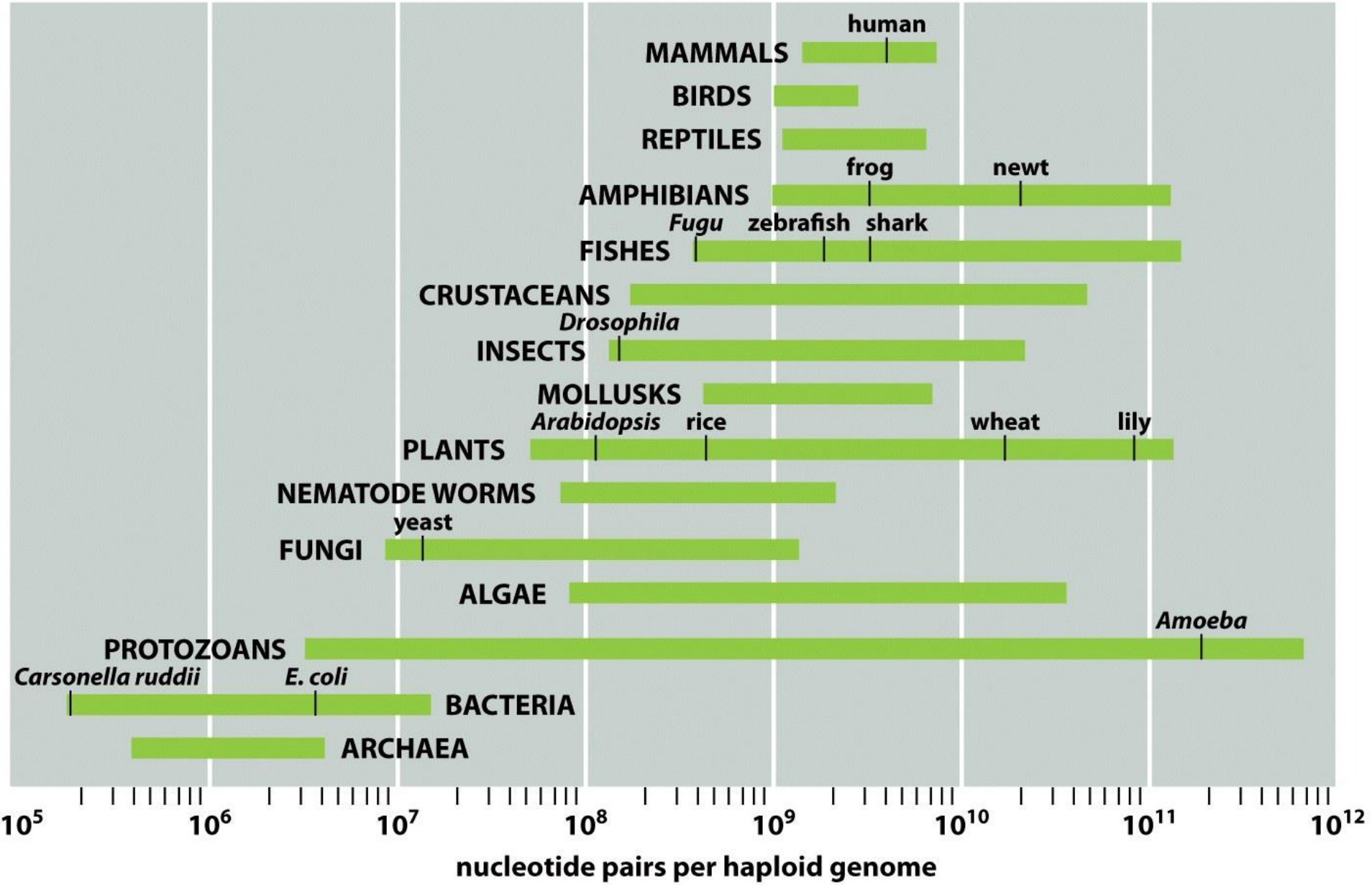


Figure 1-41 Essential Cell Biology 3/e (© Garland Science 2010)

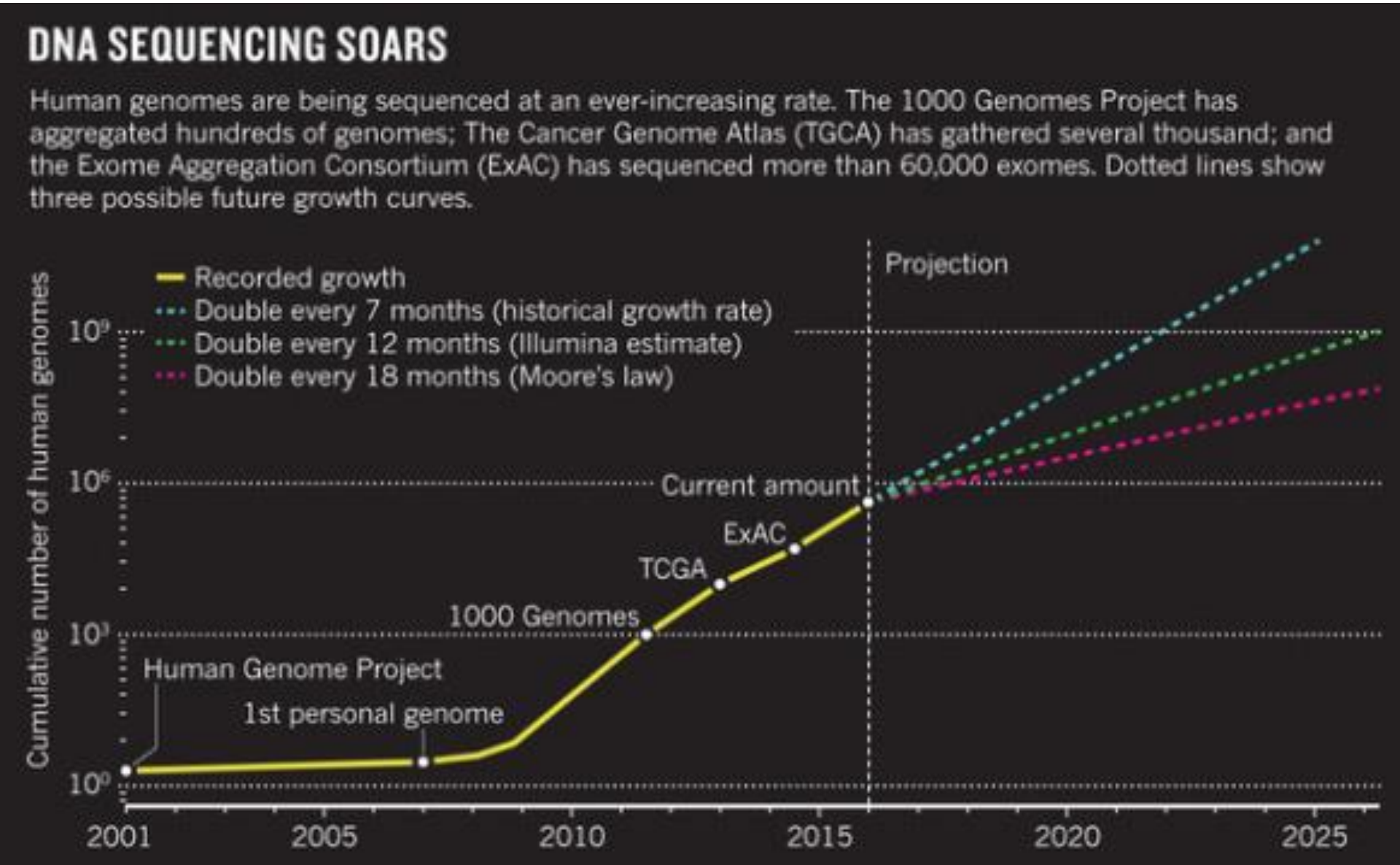
生物数据很大


4 September 2008 | www.nature.com/nature | £10
THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE
nature
THE BITER BIT
Viral infections for viruses
TROPICAL CYCLONES
The strong get stronger
BLACK HOLE PHYSICS
A new window on the Galactic Centre
BIG DATA
NATUREJOBS
Minnesota musings
SCIENCE IN THE PETABYTE ERA


nature
THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE
INSIDE STORY
Results from the Integrative Human Microbiome Project
PAGES 599, 623, 632, 641, 655 & 663
SCIENCE PUBLISHING
IMPACT FACTOR 2.0
The path to better journal metrics
PAGE 621
ATOMIC PHYSICS
GOING SUPERSONIC
Hawking radiation from analogue black holes
PAGES 634 & 688
ELECTRONICS
HUMAN TOUCH
A tactile glove that senses how a hand grasps
PAGES 630 & 698
NATURE.COM
30 May 2019
Vol 569 No 7758


Science Signaling
22 MAY 2018
AAAS

生物数据很大

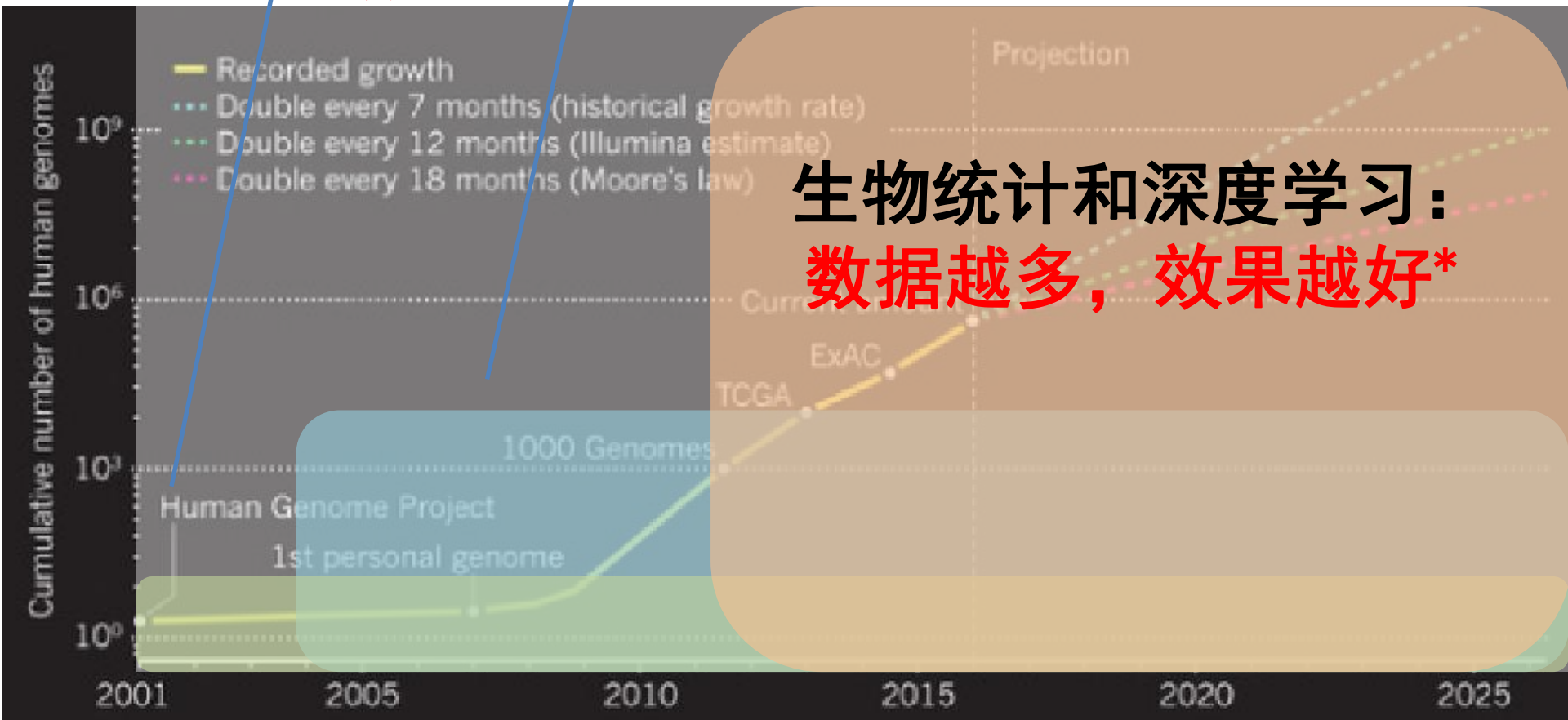

DNA SEQUENCING SOARS
Human genomes are being sequenced at an ever-increasing rate. The 1000 Genomes Project has aggregated hundreds of genomes; The Cancer Genome Atlas (TGCA) has gathered several thousand; and the Exome Aggregation Consortium (ExAC) has sequenced more than 60,000 exomes. Dotted lines show three possible future growth curves.
Recorded growth
Double every 7 months (historical growth rate)
Double every 12 months (Illumina estimate)
Double every 18 months (Moore's law)
Projection
Cumulative number of human genomes
10^9
10^6
10^3
10^0
Current amount
ExAC
TCGA
1000 Genomes
Human Genome Project
1st personal genome
2001
2005
2010
2015
2020
2025

Biostatistics

湿实验
可验证范围

传统生物信息
处理范围

生物统计和深度学习
处理范围



生物统计和深度学习：
数据越多，效果越好*

生物数据很大

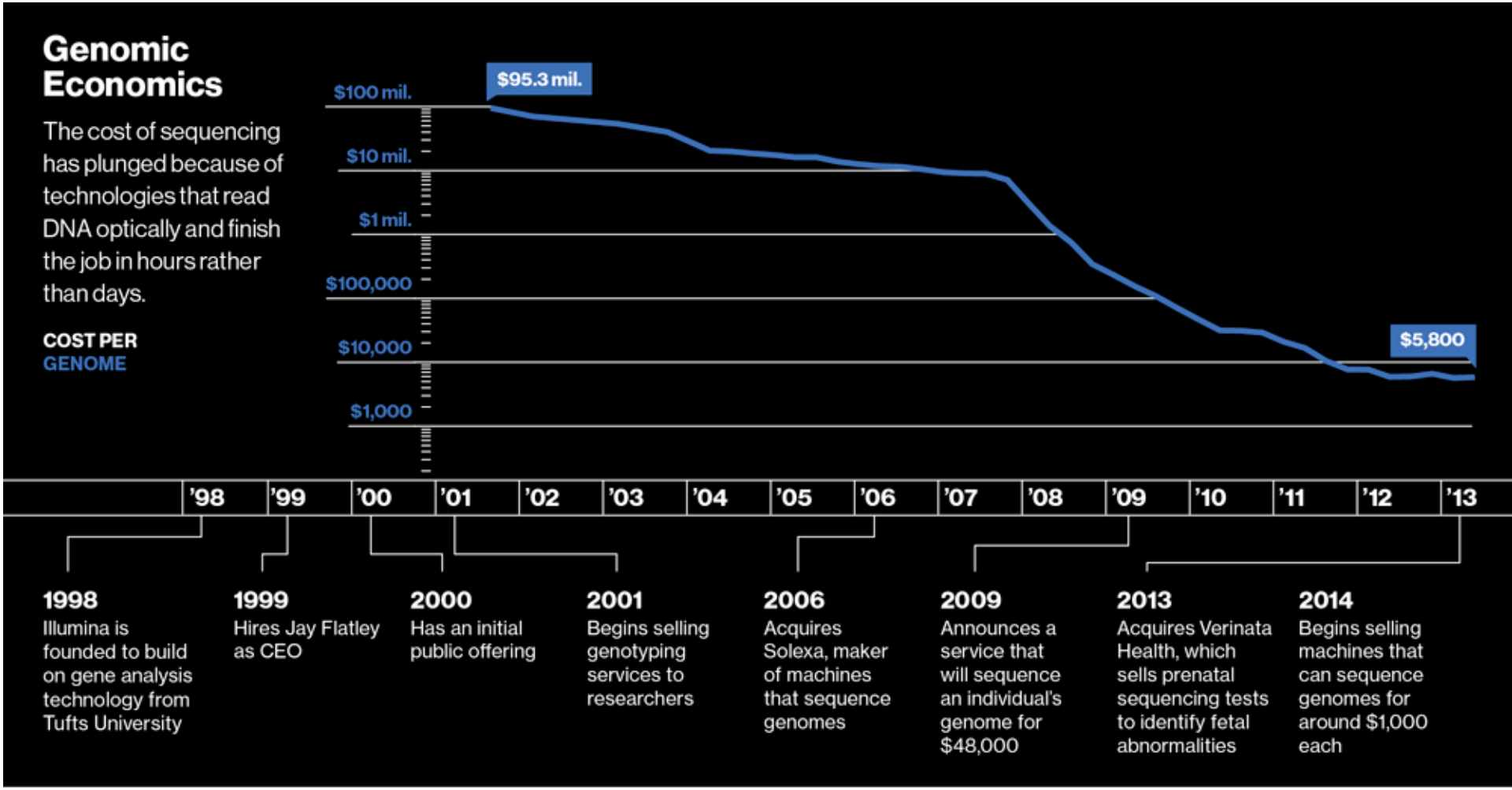
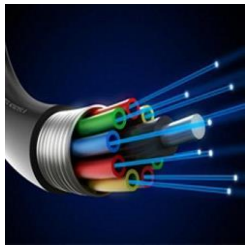


CHART BY INFOGRAPHICS.COM; DATA FROM NATIONAL HUMAN GENOME RESEARCH INSTITUTE

制约数据交换的实际是网络

光纤



北京-->武汉：1Gb/s，5000元/月

Infiniband

服务器之间：~50Gb/s，10万元

快递小哥

北京-->武汉：
(4TB*20)/(60*60*24*2)=231.5MB/s=3.7Gb/s，200元

他可以多装点，而且次日达可以更快的。。。

超级计算机平台

TOP 10 Sites for June 2017

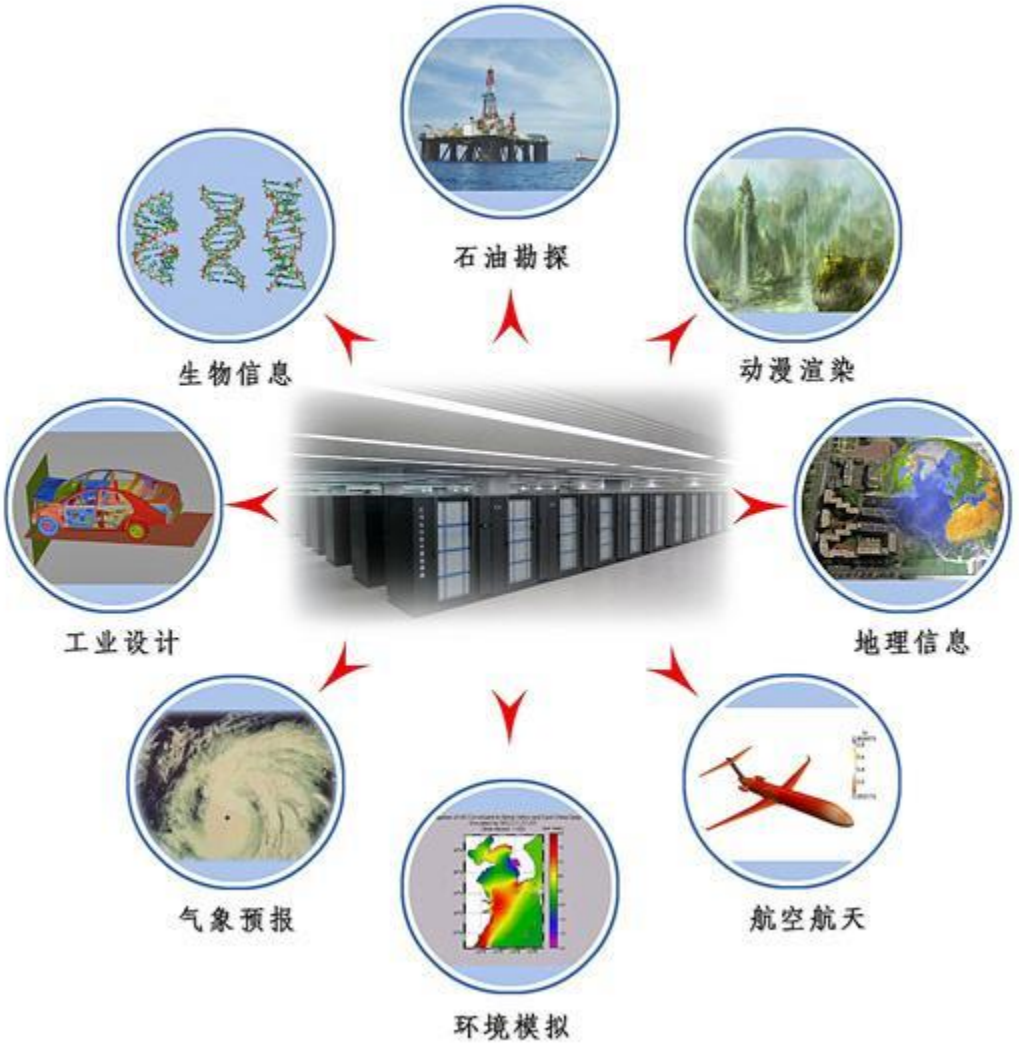
For more information about the sites and systems in the list, click on the links or view the complete list.

1-100 | 101-200 | 201-300 | 301-400 | 401-500

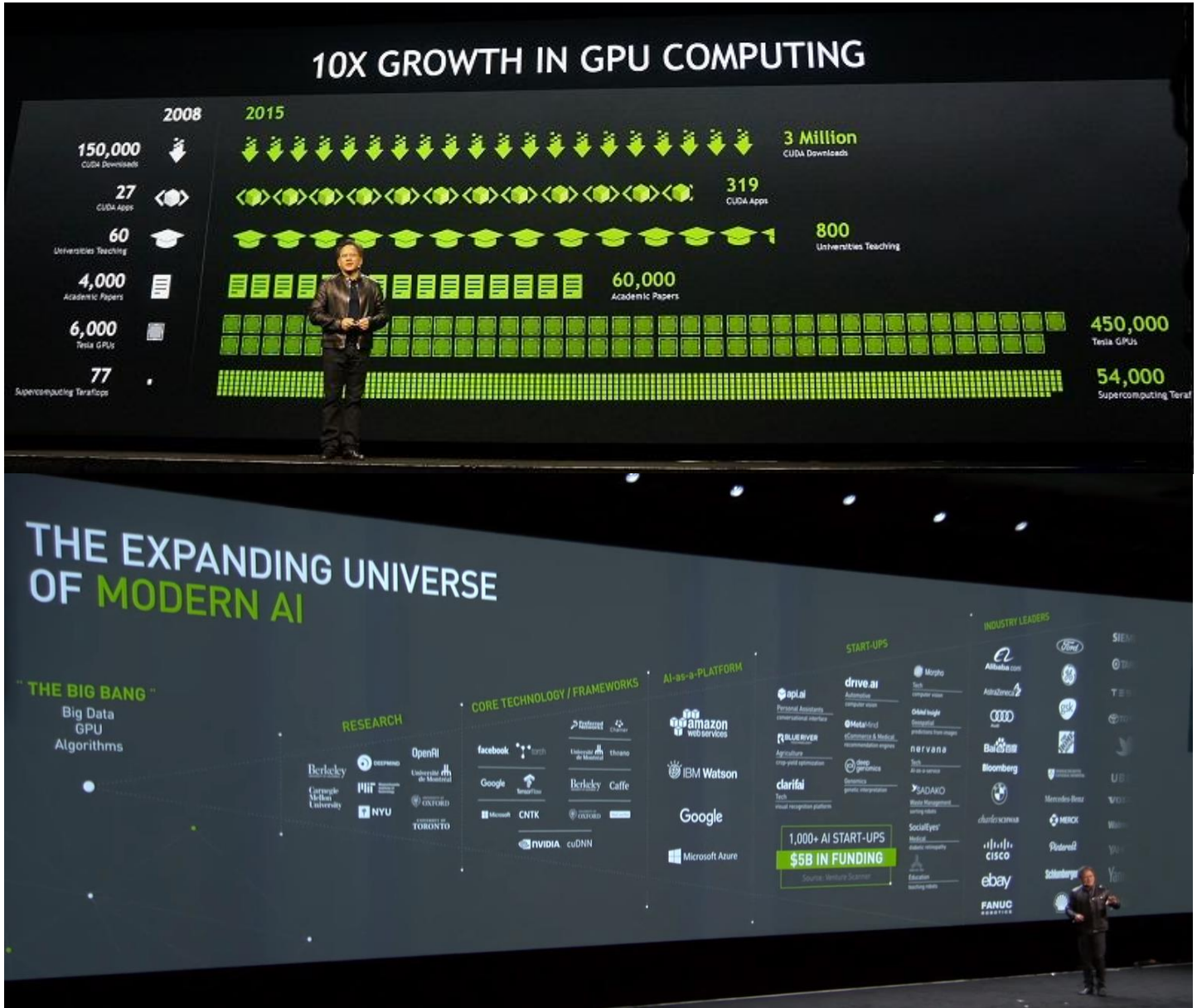
Rank	System	Cores	Rmax (TFlop/s)	Rpeak (TFlop/s)	Power (kW)
1	**Sunway TaihuLight** - Sunway MPP, Sunway SW26010 260C 1.45GHz, Sunway , NRCPC National Supercomputing Center in Wuxi China	10,649,600	93,014.6	125,435.9	15,371
2	**Tianhe-2 (MilkyWay-2)** - TH-IVB-FEP Cluster, Intel Xeon E5-2692 12C 2.200GHz, TH Express-2, Intel Xeon Phi 31S1P , NUDT National Super Computer Center in Guangzhou China	3,120,000	33,862.7	54,902.4	17,808
3	**Piz Daint** - Cray XC50, Xeon E5-2690v3 12C 2.6GHz, Aries interconnect , NVIDIA Tesla P100 , Cray Inc. Swiss National Supercomputing Centre (CSCS) Switzerland	361,760	19,590.0	25,326.3	2,272
4	**Titan** - Cray XK7, Opteron 6274 16C 2.200GHz, Cray Gemini interconnect, NVIDIA K20x , Cray Inc. DOE/SC/Oak Ridge National Laboratory United States	560,640	17,590.0	27,112.5	8,209
5	**Sequoia** - BlueGene/Q, Power BQC 16C 1.60 GHz, Custom , IBM DOE/NNSA/LLNL United States	1,572,864	17,173.2	20,132.7	7,890
6	**Cori** - Cray XC40, Intel Xeon Phi 7250 68C 1.4GHz, Aries interconnect , Cray Inc. DOE/SC/LBNL/NERSC United States	622,336	14,014.7	27,880.7	3,939
7	**Oakforest-PACS** - PRIMERGY CX1640 M1, Intel Xeon Phi 7250 68C 1.4GHz, Intel Omni-Path , Fujitsu Joint Center for Advanced High Performance Computing Japan	556,104	13,554.6	24,913.5	2,719
8	K computer, SPARC64 VIIIfx 2.0GHz, Tofu interconnect , Fujitsu RIKEN Advanced Institute for Computational Science (AICS) Japan	705,024	10,510.0	11,280.4	12,660

超级计算机平台

Tianhe-2



GPU计算


10X GROWTH IN GPU COMPUTING
2008
2015
150,000
CUDA Downloads
3 Million
CUDA Downloads
27
CUDA Apps
319
CUDA Apps
60
Universities Teaching
800
Universities Teaching
4,000
Academic Papers
60,000
Academic Papers
6,000
Tesla GPUs
450,000
Tesla GPUs
77
Supercomputing Teraflops
54,000
Supercomputing Teraf
THE EXPANDING UNIVERSE
OF MODERN AI
" THE BIG BANG "
Big Data
GPU
Algorithms
RESEARCH
CORE TECHNOLOGY / FRAMEWORKS
AI-as-a-PLATFORM
START-UPS
INDUSTRY LEADERS
amazon web services
IBM Watson
Google
Microsoft Azure
drive.ai
clarifai
nervana
1,000+ AI START-UPS
$5B IN FUNDING

GPU计算

NVIDIA

PLATFORMS ▸ DEVELOPERS ▸ COMMUNITY ▸ SHOP DRIVERS ▸ SUPPORT ABOUT NVIDIA ▸

TESLA

NVIDIA Home > Products > High Performance Computing > Industry Applications > Bioinformatics & Life Sciences

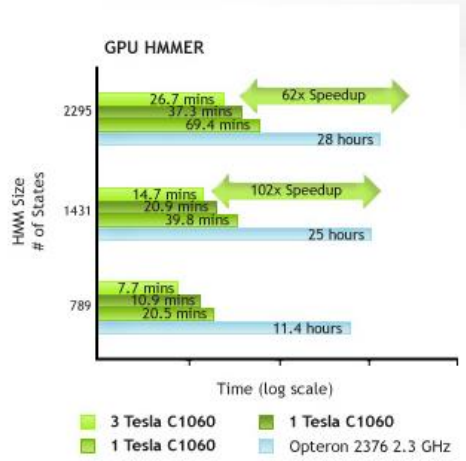
Subscribe

GPU APPLICATIONS

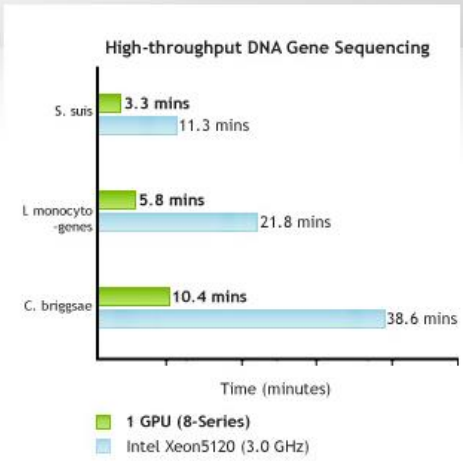
Transforming computational research and engineering

BIOINFORMATICS AND LIFE SCIENCES

Sequencing and protein docking are very compute-intensive tasks that see a large performance benefit by using a CUDA-enabled GPU. There is quite a bit of ongoing work on using GPUs for a range of Bioinformatics and life sciences codes.

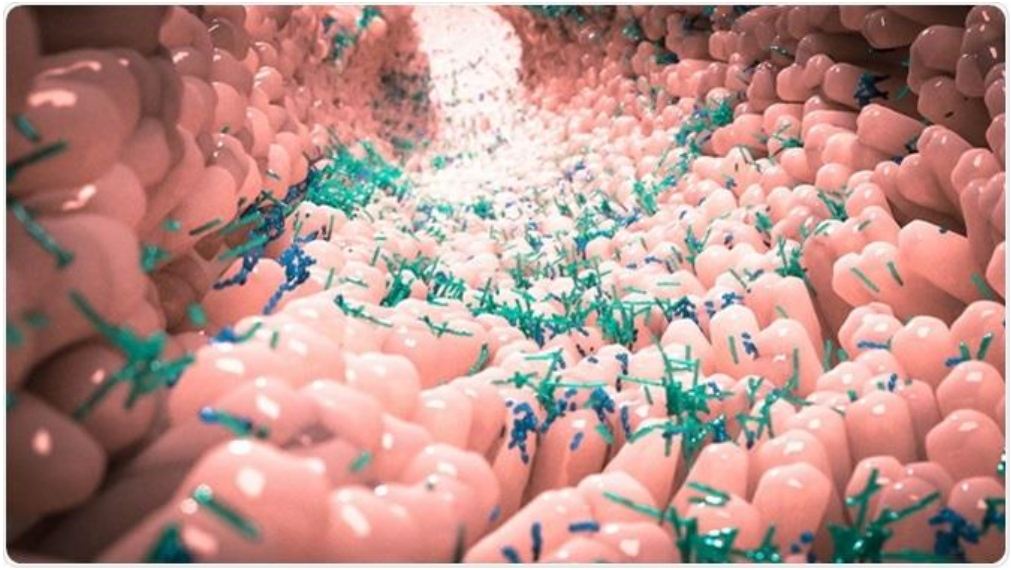


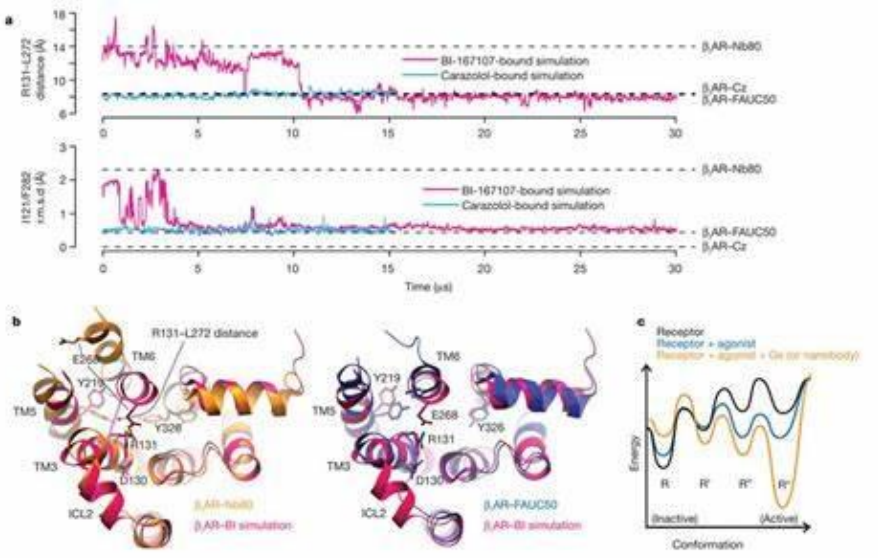
Accelerating HMMER using GPUs
Scalable Informatics



MUMmerGPU: High-through DNA sequence alignment using GPUs
Schatz, et al

GPU计算

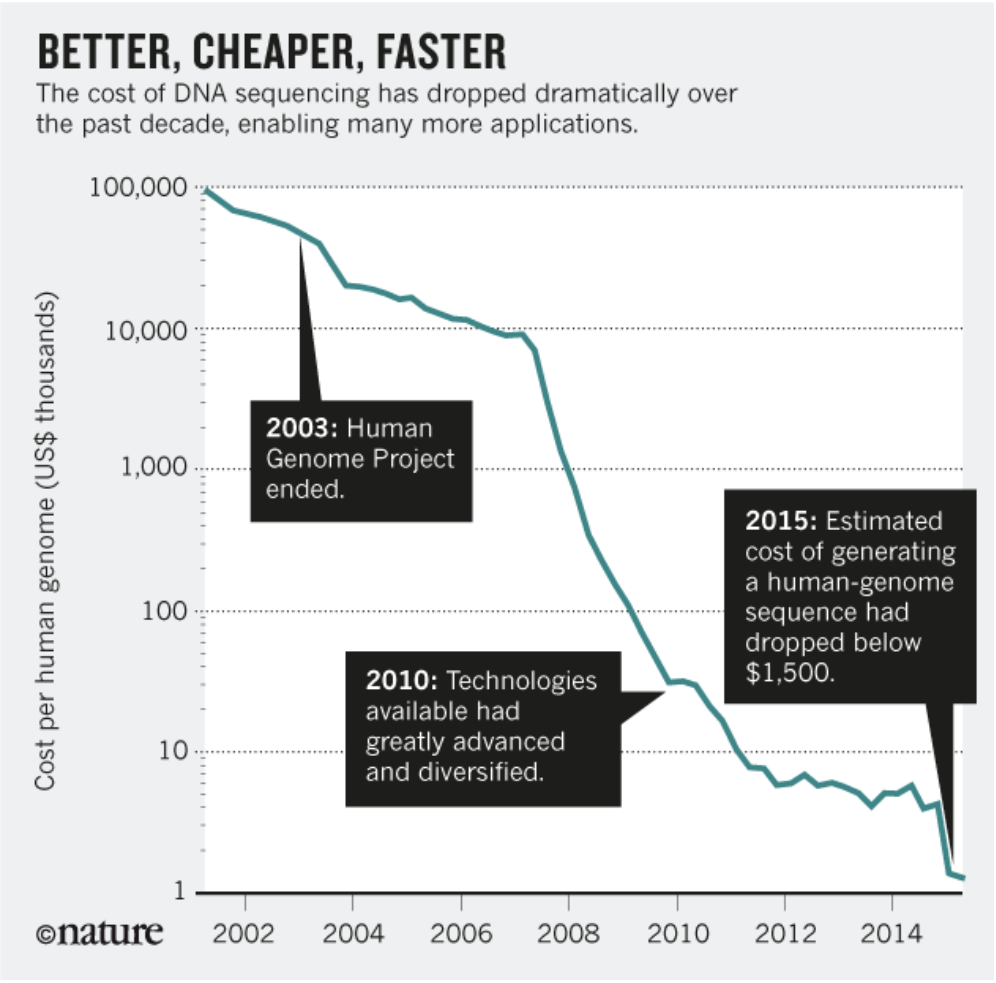


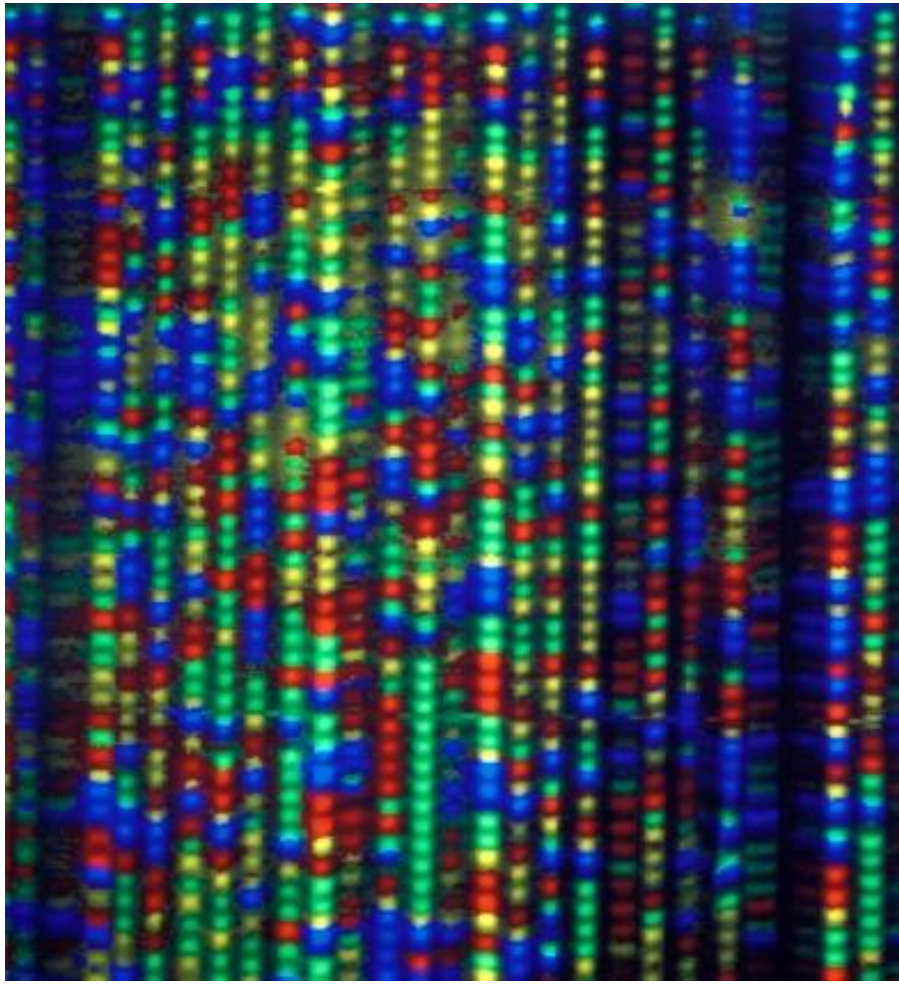


- 高通量测序数据挖掘（深度学习）
- 蛋白和分子对接（药物设计）
- 分子动态网络分析
- 。 。 。

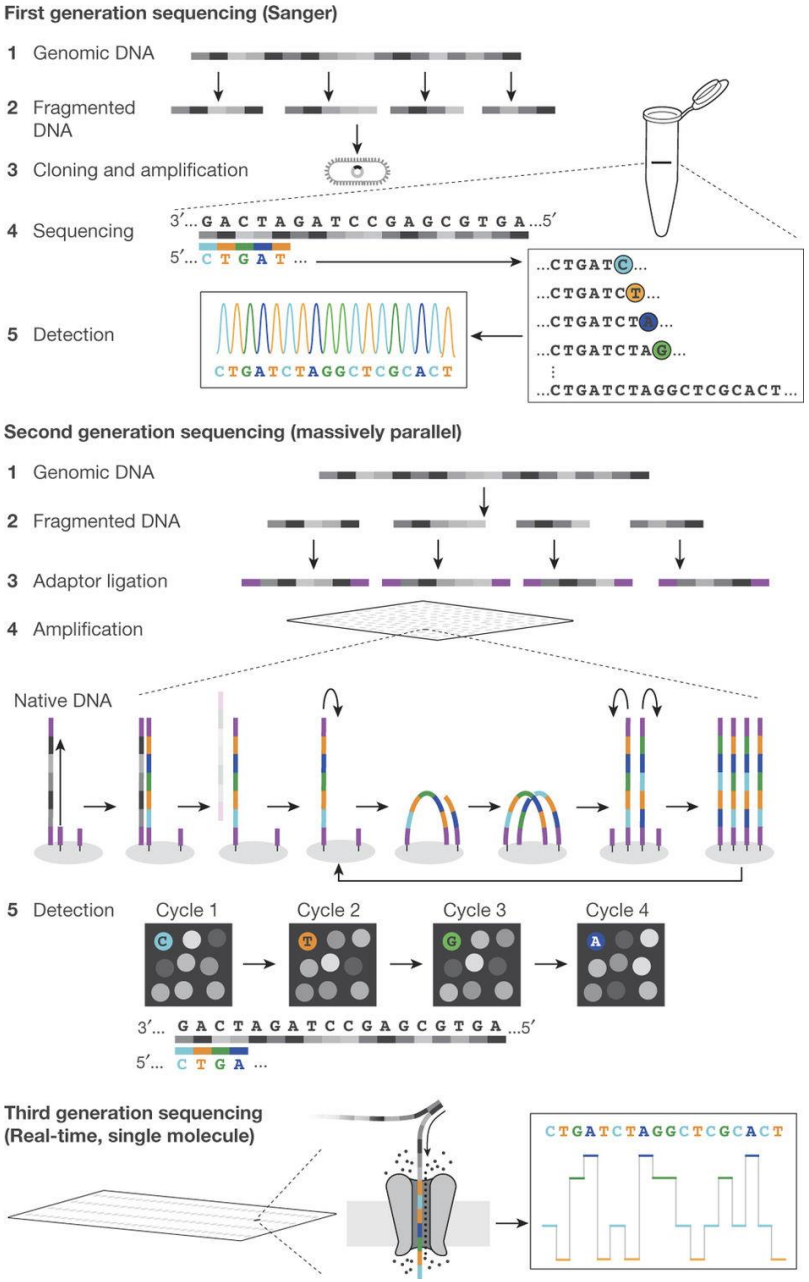
2. 高通量测序

DNA sequencing and bioinforamtics

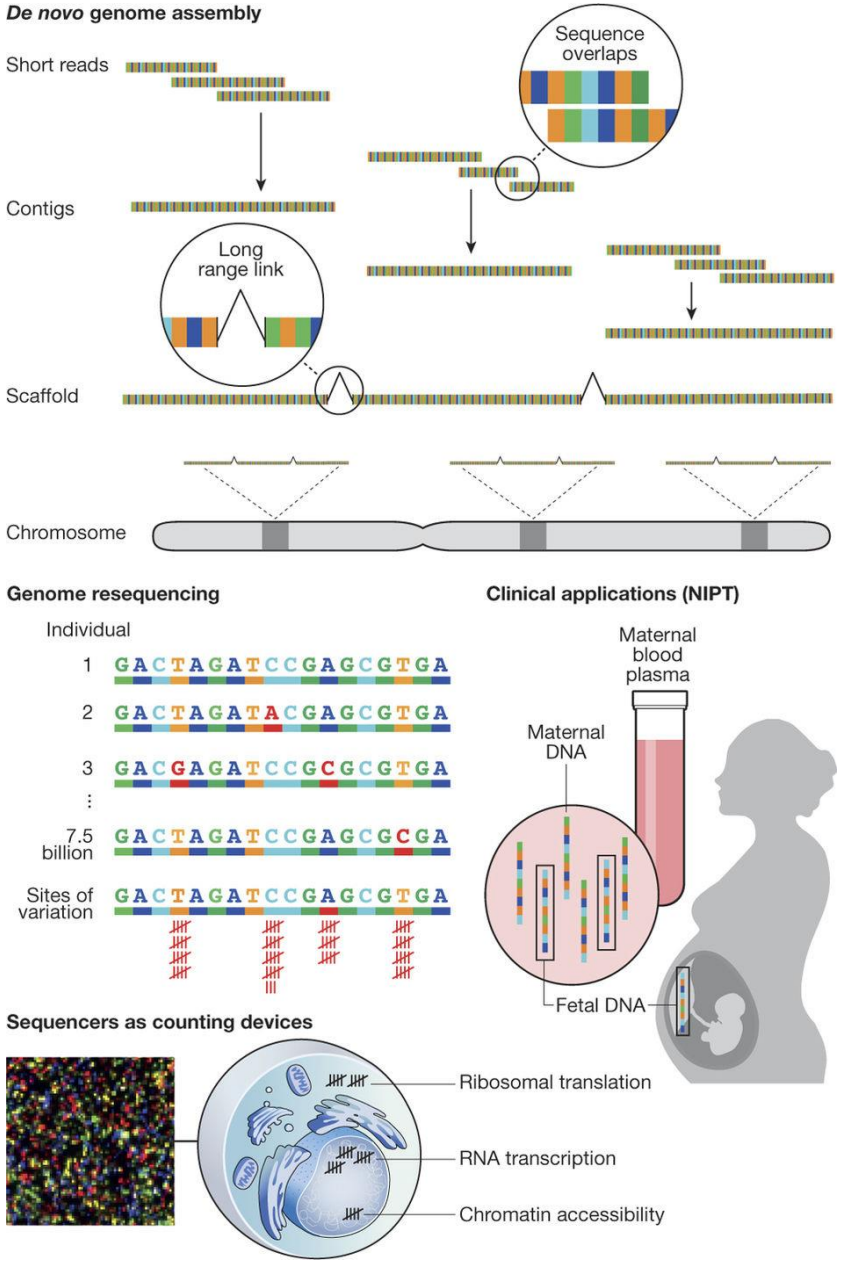




DNA Sequencing



Sequencing applications



在今天，DNA测序技术已经在诸多方面达到了临床应用的具体要求。

科学家估计，全世界每年大约有四百万到六百万孕妇正在通过外周血游离DNA对胎儿的21三体综合征进行诊断，而十年之内，这个数字将超过1500万。

在高收入国家，基因组测序已经广泛用于多种疾病的产前诊断，可以揭示大约30%的出生缺陷，同时，这一数字也正随着数据解读能力的成熟而逐渐上升。

在肿瘤学领域，液体活检在最近几年已经成为了肿瘤相关学术，产业以及投资界的新宠。基于DNA测序的液体活检被认为正逐步发展为癌症诊断与预后评估的标准方法，能够在可知的时间内逐步补充甚至取代传统的创伤性癌症诊断技术。

同样，手持DNA测序仪等设备的开发也使得流行病学家甚至能够在最为偏远的地区高效完成对人类样本，动物以及昆虫病原载体甚至是空气，水，食物的基因检测。

流行病学家和公共卫生专家也开始讨论如何通过对城市垃圾中微生物的DNA测序辅助传染性疾病的预防与控制。

海洋生物学家也正在通过宏基因组学技术来对海洋的生态健康进行监测与研究。

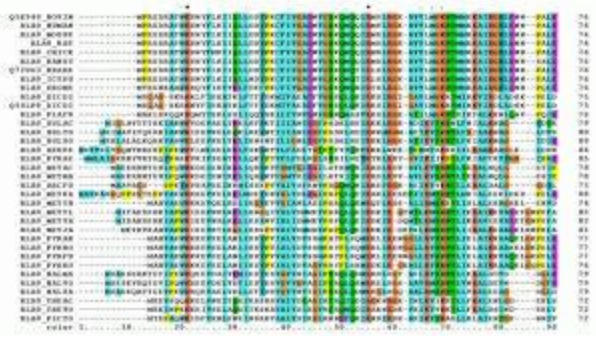
在法医领域，便携式DNA测序仪可以将DNA分析带出法医实验室，使其成为一线警务工作人员的随身工具。帮助警方即使通过DNA监测确定嫌疑人，发展成为诸如酒精探测器一类的便捷工具。

在人们的家里，DNA测序设备或许也可以成为下一个“智能”或“连接”设备，一些评论者甚至认定厕所是通过实时DNA测序监测家庭成员健康的理想场所。

3. 生物统计学经典软件

生物统计经典软件

高通量测序数据分析

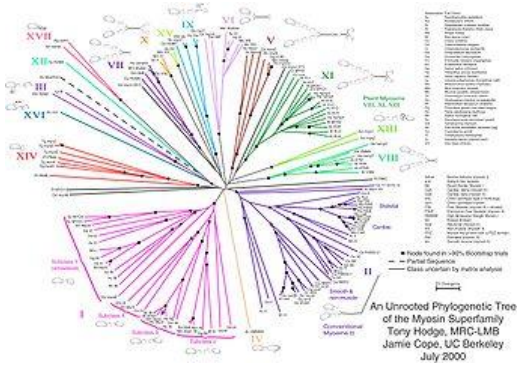


- MEME： http://meme-suite.org/
- GenScan： http://genes.mit.edu/GENSCAN.html
- HMMalign：

http://www.biology.wustl.edu/gcg/hmmeralign.html

生物统计经典软件

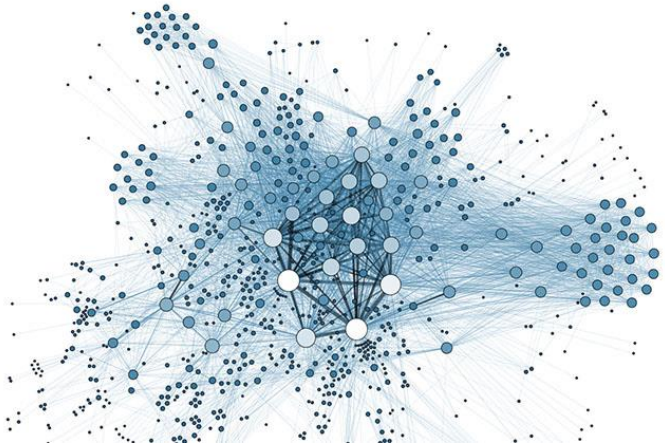
物种和基因进化分析



- iTOL： https://itol.embl.de/
- MEGA： http://www.megasoftware.net/

生物统计经典软件

生物分子网络分析



- Cytoscape： http://www.cytoscape.org/

生物统计经典软件

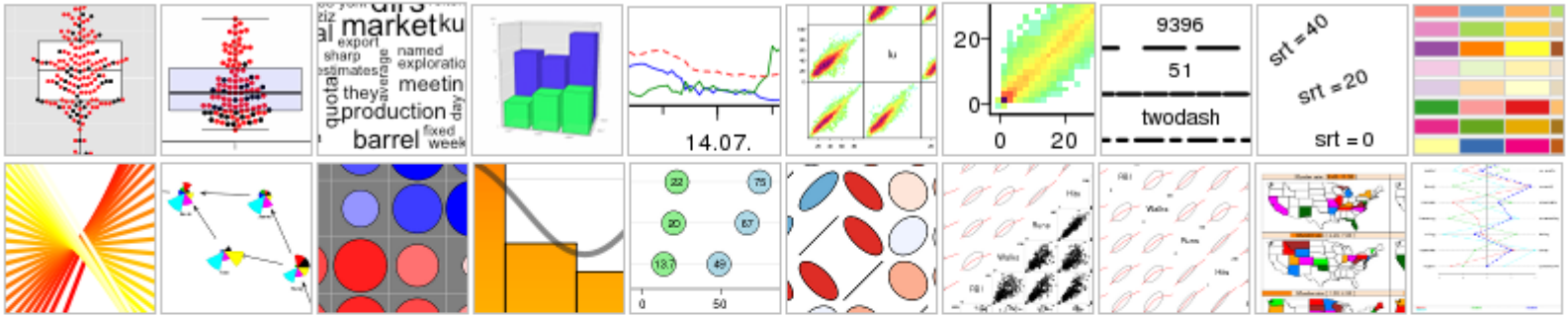
降维、分子结构等分析

- PCA analysis： http://biit.cs.ut.ee/clustvis/
- DREAM Challenge： http://dreamchallenges.org/

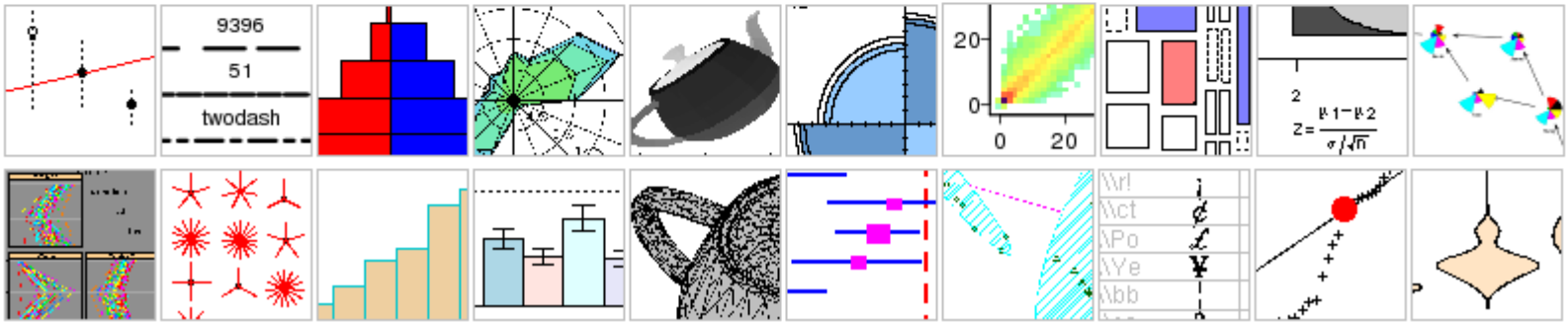
3. 生物统计学核心工具

R： https://www.r-project.org

» Last entries ...



» Random entries



http://blog.revolutionanalytics.com/2011/09/the-r-graph-gallery-goes-social.html

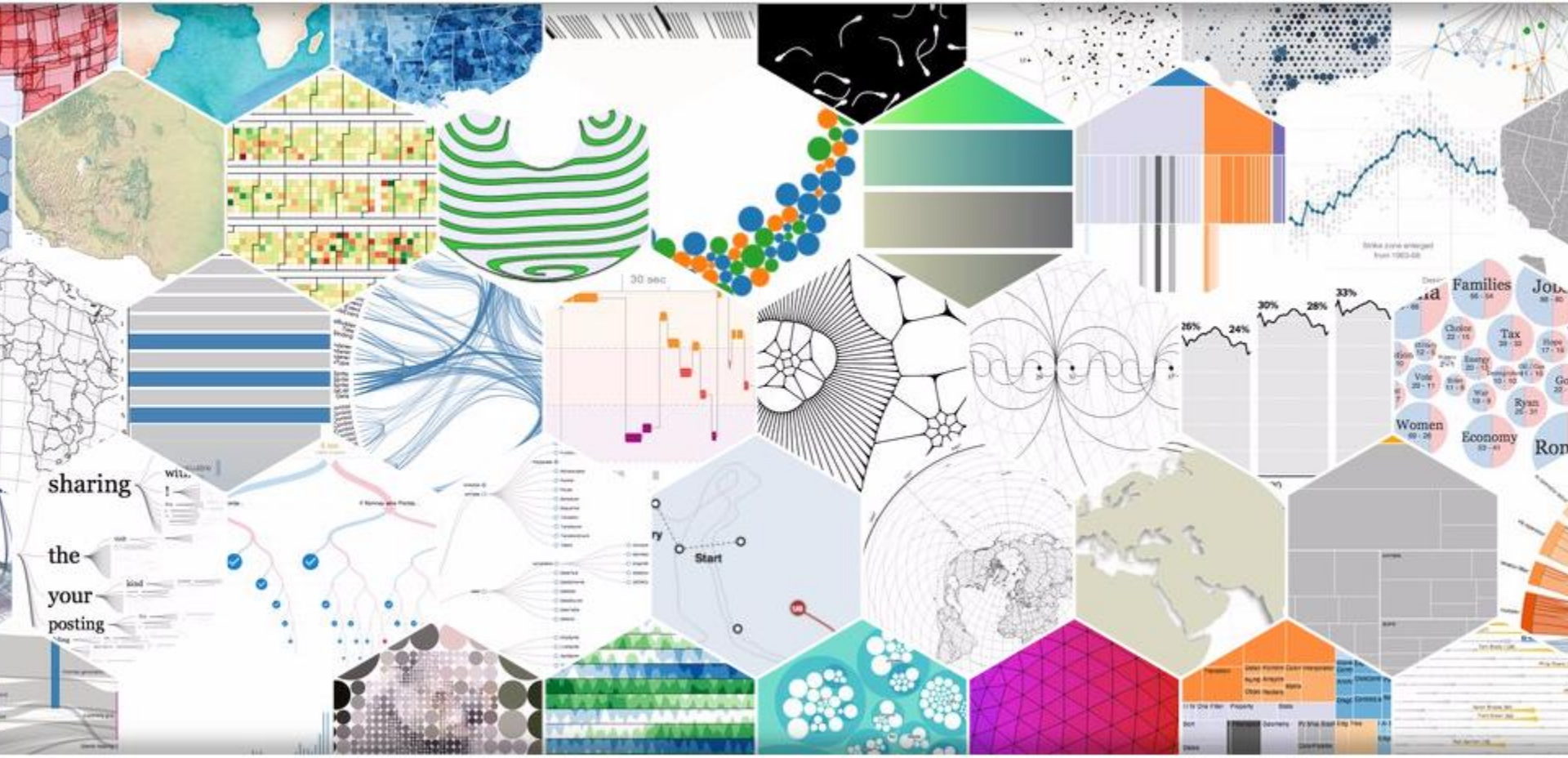
Python and Biopython：

https://www.python.org/

http://biopython.org/

D3.js for visualization：

https://d3js.org/



Echart for visualization：

http://echarts.baidu.com

4. 深度学习

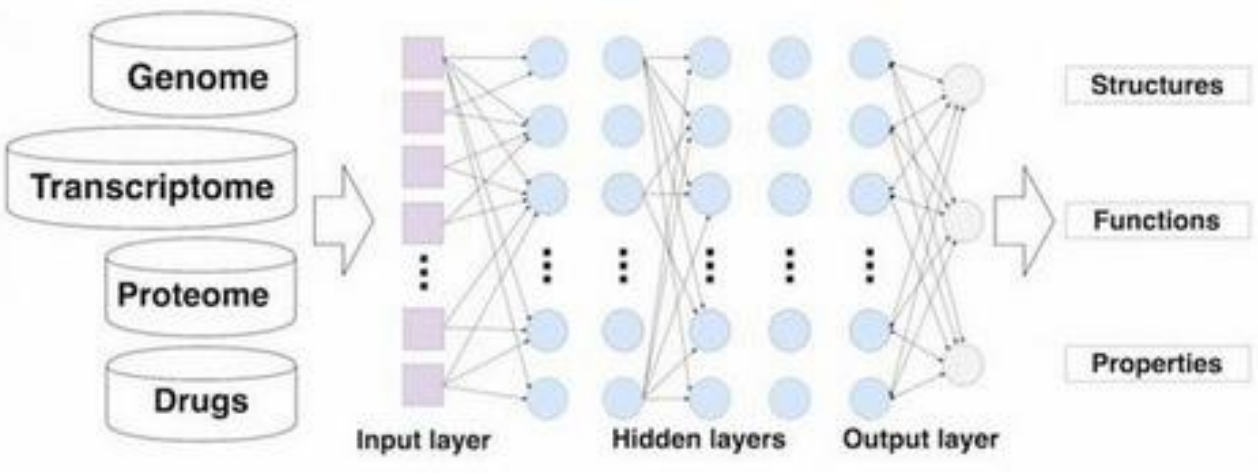
Deep Learning

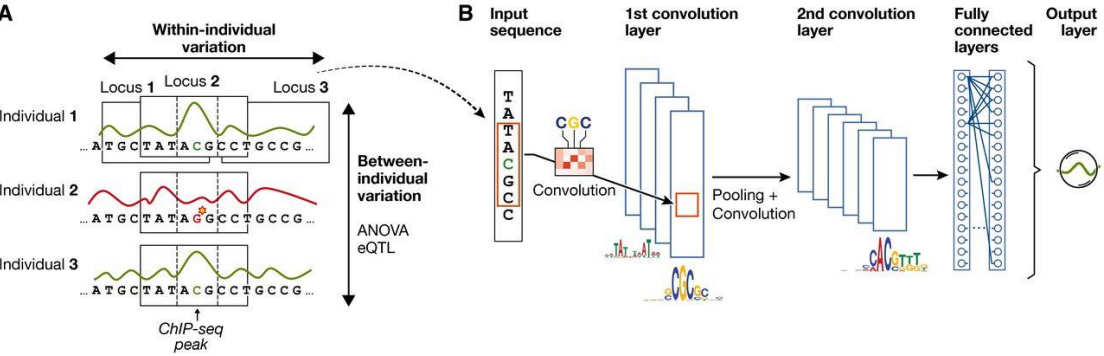
数据很重要！

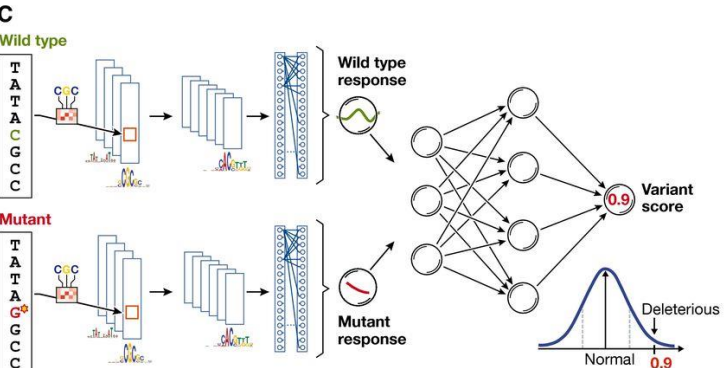
算法更重要！

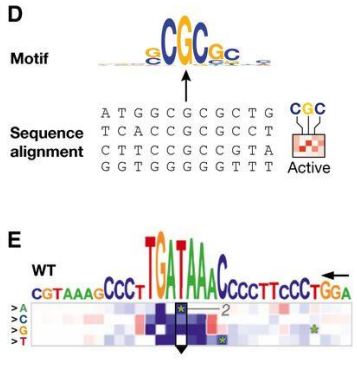
Deep Learning

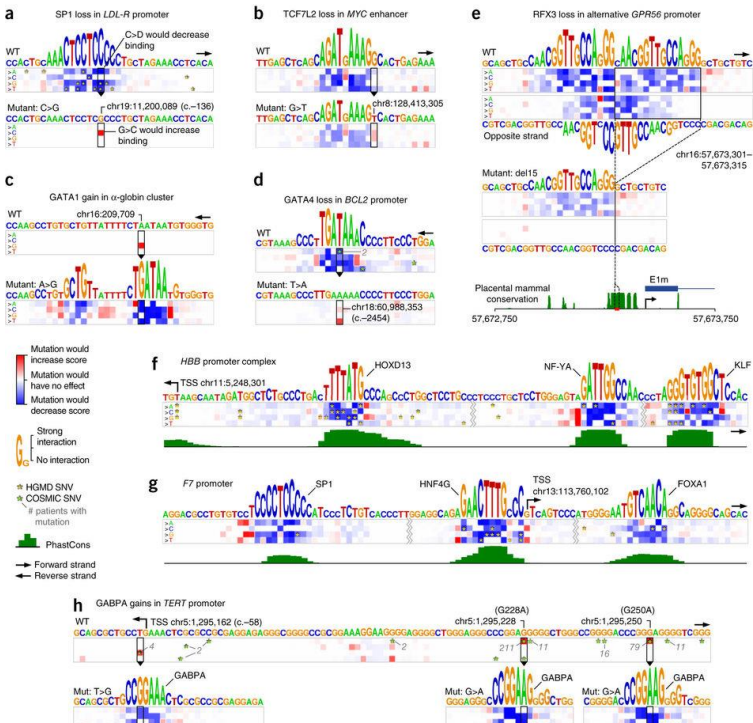
Deep Learning for Bioinformatics



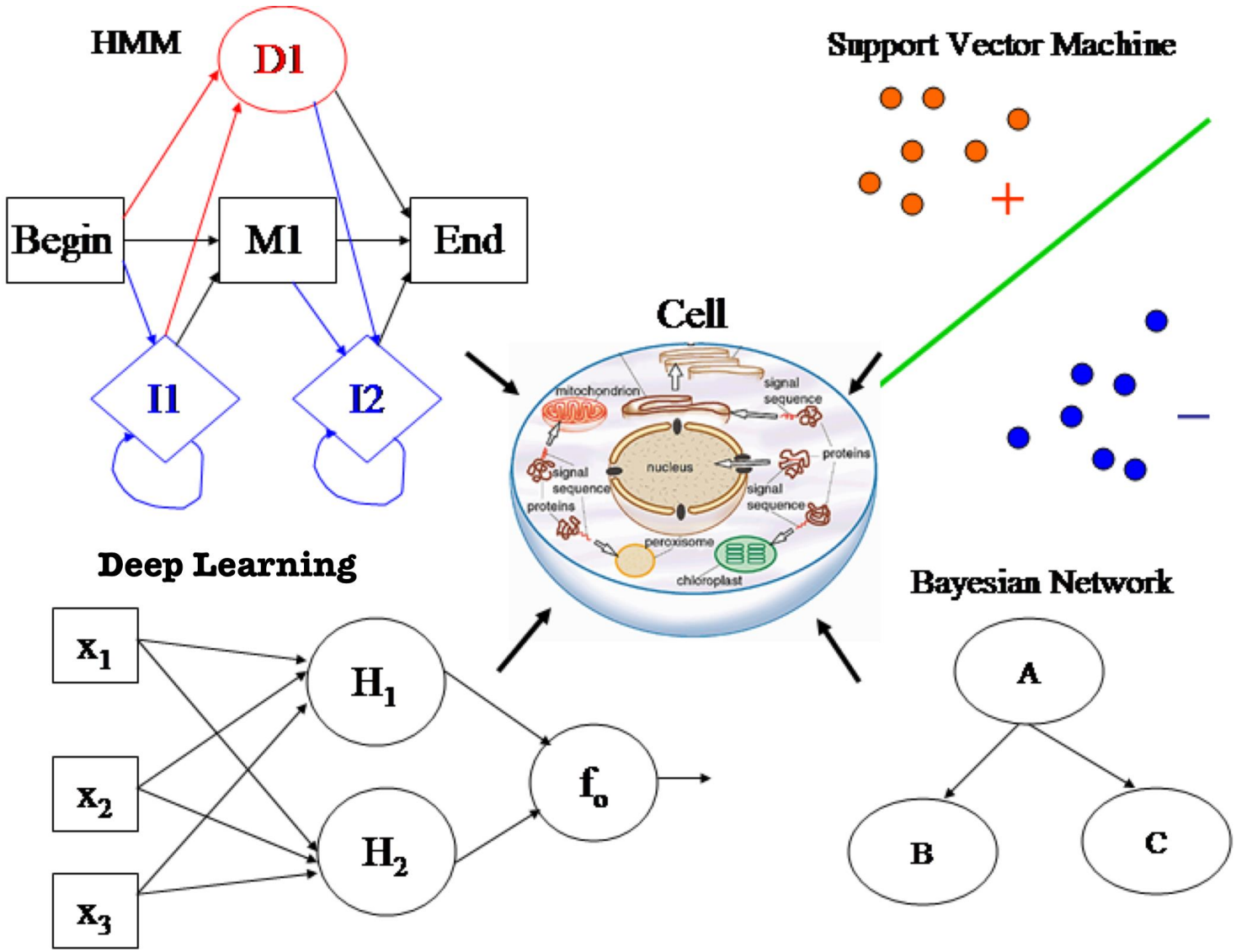






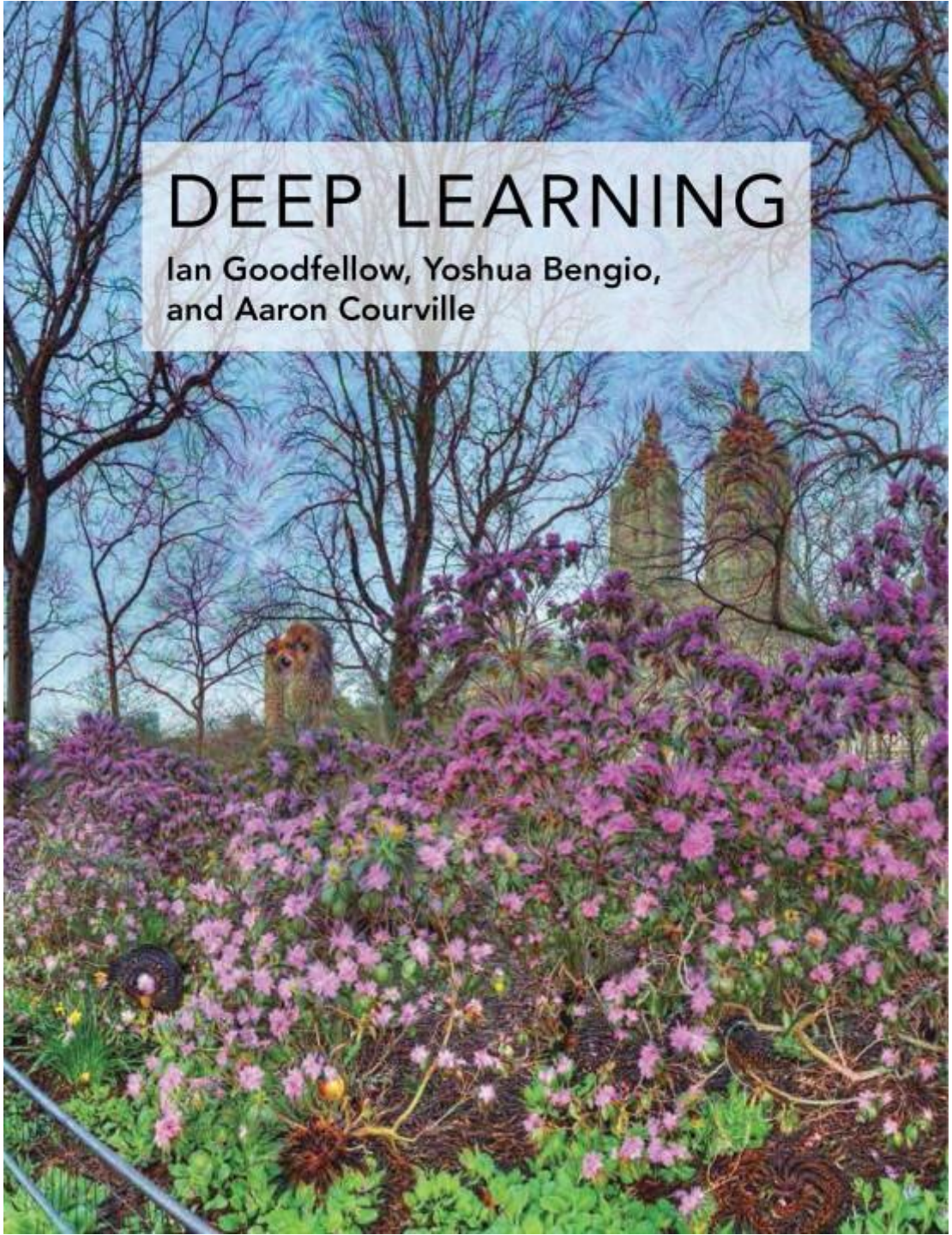


Deep Learning for Bioinformatics



Deep Learning for Bioinformatics

- 高通量测序数据挖掘
- 蛋白和分子对接（药物设计）
- 生物影像分析
- 。 。 。


DEEP LEARNING
Ian Goodfellow, Yoshua Bengio,
and Aaron Courville