生物统计学：
生物信息中的概率统计模型

2020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

- 课程网页
  - http://www.microbioinformatics.org/Biostatistics.html
  - QQ群：182996651

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…

MICROBIOME

A unified initiative to harness Earth's microbiomes

Transition from description to causality and engineering

By A. P. Alivisatos,* M. J. Blaser, E. L. Brodie, M. Chun, J. L. Dangl, T. J. Donohue, P. C. Dorrestein, J. A. Gilbert, J. L. Green, J. K. Jansson, R. Knight, M. E. Maxon, M. J. McFall-Ngai, J. F. Miller, K. S. Pollard, E. G. Ruby, S. A. Taha, Unified Microbiome Initiative Consortium

Despite their centrality to life on Earth, we know little about how microbes (1) interact with each other, their hosts, or their environment. Although DNA sequencing techniques have enabled a new view of the ubiquity and diversity of microorganisms, this has mainly yielded snapshots that shed limited light on microbial functions or community dynamics. Given that nearly every habitat and organism hosts a diverse constellation of microorganisms—its "microbiome"—such knowledge could transform our understanding of the world and launch innovations in agriculture, energy health, the environment, and more (see the photo). We propose an interdisciplinary Unified Microbiome Initiative (UMI) to discover and advance tools to understand and harness the capabilities of Earth's microbial ecosystems. The impacts of oceans and soil microbes on atmospheric CO_2 are critical for understanding climate change (2). By manipulating interactions at the root-soil-microbe interface, we may reduce agricultural pesticide, fertilizer, and water use enrich marginal land and rehabilitate degraded soils. Microbes can degrade plant cell walls (for biofuels), and synthesize novel small molecules for new bioproducts, including antibiotics (3). Restoring normal human microbial ecosystems can save lives (e.g., fecal microbiome transplantation for *Clostridium difficile* infections (4)). Rational management of microbial communities in and around us has implications for asthma, diabetes, obesity, infectious diseases, psychiatric illnesses, and other afflictions (5, 6). The human microbiome is a target and a source for new drugs (7) and an essential tool for precision medicine (8).

The National Science Foundation's Microbial Observatories, the U.S. Department of Energy's Genomic Sciences program, the National Institutes of Health's Human Microbiome Project, and other efforts in the United States and abroad have served as critical first steps in revealing the diversity of microbes and their communities. However, we lack many tools required to advance beyond descriptive approaches to studies that enable a mechanistic, predictive, and actionable understanding of global microbiome processes. Developing these tools requires new collaborations between physical, life, and biomedical sciences; engineering; and other disciplines.

AREAS OF EMPHASIS. A central purpose of the UMI is to develop cross-cutting platform technologies to accelerate basic discovery and translation to applications. We highlight key needs and opportunities.

Decrypting microbial genes and chemistries. Approaches for characterizing microbiomes increasingly rely on whole-community metagenomic sequencing, yet roughly half of the genes identified in these studies encode products of unknown function, and existing methods for assigning function are often incomplete or inaccurate (9). Technologies for resolving roles of uncharacterized genes with high throughput and high accuracy are needed. These approaches must integrate improved computational methods for in silico prediction of protein and RNA functions, rapid mutagenesis of model organisms or native strains under natural conditions, multiomics and high-resolution phenotyping platforms to test functional predictions in vitro and in situ, and improved capture of information in the literature.

> *"we envision...evidence-based, model-informed microbiome management..."*

Deciphering chemistries of microbiomes is essential. In untargeted metabolomics studies using mass spectrometry, less than 2% of data can be matched to known chemical compounds, and only a fraction of those map to recognized biochemical pathways (10). Advances have been made in predicting structures from mass spectra, but improvements are needed in both in silico and physical

Downloaded from www.sciencemag.org on November 9, 2015

COMMENT

UNIVERSITIES A call for better pedagogy in Muslim nations p.814 | PHOTO Wartime at Paris's natural history museum captured by Doisneau p.817 | EPISTEMOLOGY Guide to uncertainty from doyenne of European science policy p.818 | CANCER Will upcoming cuts be too small to tackle bovine tuberculosis? p.840

Even extreme environments such as Antarctic lakes host microbes.

Create a global microbiome effort

Understanding how microbes affect health and the biosphere requires an international initiative, argue Nicole Dubilier, Margaret McFall-Ngai and Liping Zhao.

Microbes have been discovered on Earth wherever anyone has looked for them, from the boiling waters of Yellowstone's hot springs in Wyoming to the depths of cold, dark Antarctic lakes under 800 metres of ice. A holistic understanding of the role of Earth's microbial community and its genome — its microbiome — in the biosphere and in human health is key to meeting many of the challenges that face humanity in the twenty-first century, from energy to infection to agriculture.

Recognizing this, a group of leading US scientists this week proposes[1] the creation of a Unified Microbiome Initiative (UMI). The UMI would bring together researchers and representatives from public and private agencies and foundations to study the activities of Earth's microbial ecosystems.

The UMI is conceived as a US initiative, springing from meetings sponsored by the White House Office of Science and Technology Policy and the Kavli Foundation of Oxnard, California. But Earth's biome is not defined by national borders, and efforts to unlock its secrets should go global.

We believe that to be successful, microbiome research will require a coordinated effort across the international community of biologists, chemists, geologists, mathematicians, physicists, computer scientists and clinical experts. As three scientists working in three countries — Germany, China and the United States — we call for an International Microbiome Initiative (IMI) supported by funding agencies and foundations around the world, in addition to the UMI. This would ensure the sharing of standards across borders and disciplines, and bring cohesion to the multitude of microbiome initiatives that exist.

MICROBIAL REVOLUTION

Science is only just realizing the full importance of the microbial world. This is thanks to developments such as low-cost high-throughput sequencing; advances in sample preparation that allow researchers to sequence genomes from individual cells as well as from microbial communities; improvements in computing power and imaging technologies; and the development of bioinformatics tools to help make sense of the data.

This biologists are gaining insight into the identity and function of microbes that cannot be grown in the laboratory — the vast majority of Earth's microbiome. Currently only 35 bacterial and archaeal phyla are recognized on the basis of classical approaches to microbial taxonomy. Sequencing efforts in the past few years have pushed the number closer to 1,000 (ref. 2).

Newfound groups of bacteria are throwing old assumptions about the tree of life into question, and revealing vast holes in our ▶

© 2015 Macmillan Publishers Limited. All rights reserved

Downloaded from www.sciencemag.org on November 9, 2015

Sciencexpress Reports

Anticancer immunotherapy by CTLA-4 blockade relies on the gut microbiota

Marie Vétizou,[1,2,3] Jonathan M. Pitt,[1,2,3] Romain Daillère,[1,2,3] Patricia Lepage,[4] Nadine Waldschmitt,[5] Caroline Flament,[1,2,6] Sylvie Rusakiewicz,[1,2,3] Bertrand Routy,[1,2,2,7] Maria P. Roberti,[1,2,3] Connie P. M. Duong,[1,2,3] Vichnou Poirier-Colame,[1,2,3] Antoine Roux,[1,2,7] Sonia Becharef,[1,2,3] Silvia Formenti,[8] Encouse Golden,[8] Sascha Cording,[9] Gerard Eberl,[9] Andreas Schlitzer,[10] Florent Ginhoux,[10] Sridhar Mani,[11] Takahiro Yamazaki,[1,2,3] Nicolas Jacquelot,[1,2,3] David P. Enot,[1,2,12] Marion Bérard,[13] Jérôme Nigou,[14,15] Paule Opolon,[1] Alexander Eggermont,[1,2,3] Paul-Louis Woerther,[16] Elisabeth Chachaty,[17] Nathalie Chaput,[1,18] Caroline Robert,[1,19,20] Christian Mateus,[1,19] Guido Kroemer,[7,21,22,23,24] Didier Raoult,[25] Ivo Gomperts Boneca,[26,27*] Franck Carbonnel,[5,28*] Mathias Chamaillard,[29*] Laurence Zitvogel[1,2,3,6†]

[1]Institut de Cancérologie Gustave Roussy Cancer Campus (GRCC), 114 rue Edouard Vaillant, 94805 Villejuif, France. [2]INSERM U1015, GRCC, Villejuif, France. [3]University of Paris Sud XI, Kremlin Bicêtre, France. [4]Institut National de la Recherche Agronomique (INRA), Micalis UMR1319, 78350 Jouy-en-Josas, France. [5]University of Lille, CNRS, INSERM, Centre Hospitalier Régional Universitaire de Lille, Institut Pasteur de Lille, U1019, UMR 8204, Centre d'Infection et Immunité de Lille (CIIL) F-59000 Lille, France. [6]Center of Clinical Investigations in Biotherapies of Cancer 1428, Villejuif, France. [7]Université Paris Descartes, Sorbonne Paris Cité, Paris, France. [8]Department of Radiation Oncology, New York University, New York, NY, USA. [9]Microenvironment and Immunity Unit, Institut Pasteur, Paris, France. [10]Singapore Immunology Network (SIgN), Agency for Science, Technology and Research (A*STAR), Singapore, Singapore. [11]Department of Genetics and Department of Medicine, Albert Einstein College of Medicine, Bronx, NY 10461, USA. [12]Metabolomics Platform, GRCC, Villejuif, France. [13]Animalerie Centrale, Institut Pasteur, Paris, France. [14]Centre National de la Recherche Scientifique, Institut de Pharmacologie et de Biologie Structurale (IPBS), Toulouse, France. [15]Université de Toulouse, Université Paul Sabatier, IPBS, F-31077 Toulouse, France. [16]Department of Medical Oncology, Institut Gustave Roussy, Villejuif, France. [17]Service de microbiologie, GRCC, Villejuif, France. [18]Laboratory of Immunomonitoring in Oncology, UMS 3655 CNRS/US 23 INSERM, GRCC, Villejuif, France. [19]INSERM U981, GRCC, Villejuif, France. [20]INSERM U848, Villejuif, France. [21]Equipe 11 Labellisée—Ligue Nationale contre le Cancer, Centre de Recherche des Cordeliers, INSERM U1138, Paris, France. [22]Pôle de Biologie, Hôpital Européen Georges Pompidou, Assistance Publique–Hôpitaux de Paris, Paris, France. [23]Orbit des Ribosomes, Faculté de Médecine, Université de la Méditerranée, Marseille, France. [24]Institut Pasteur, Unit of Biology and Genetics of the Bacterial Cell Wall, Paris, France. [25]INSERM, Equipe Avenir, Paris, France. [26]Gastroenterology Department, Hôpital Bichât, Assistance Publique–Hôpitaux de Paris, Paris, France.

*These authors contributed equally to this work.
†Corresponding author. E-mail: laurence.zitvogel@gustaveroussy.fr

Antibodies targeting CTLA-4 have been successfully used as cancer immunotherapy. We find that the antitumor effects of CTLA-4 blockade depend on distinct *Bacteroides* species. In mice and patients, T cell responses specific for *B. thetaiotaomicron* or *B. fragilis* were associated with the efficacy of CTLA-4 blockade. Tumors in antibiotic-treated or germ-free mice did not respond to CTLA blockade. This defect was overcome by gavage with *B. fragilis*, or by immunization with *B. fragilis* polysaccharides, or by adoptive transfer of *B. fragilis*-specific T cells. Fecal microbial transplantation from humans to mice confirmed that anti–CTLA-4 treatment of melanoma patients favored the outgrowth of *B. fragilis* with anticancer properties. This study reveals a key role for *Bacteroidales* in the immunostimulatory effects of CTLA-4 blockade.

Ipilimumab is a fully human monoclonal antibody directed against CTLA-4, a major negative regulator of T cell activation (*1*), approved in 2011 for improving the overall survival of patients with metastatic melanoma (MM) (*2*). However, blockade of CTLA-4 by ipilimumab often results in immune-related adverse events at sites that are exposed to commensal microorganisms, mostly the gut (*3*). Patients treated with ipilimumab develop antibodies to components of the enteric flora (*4*). Therefore, given our previous findings for other cancer therapies (*5*), addressing the role of gut microbiota in the immunomodulatory effects of CTLA-4 blockade is crucial for the future development of immune checkpoint blockers in oncology.

We compared the relative therapeutic efficacy of the anti–CTLA-4 9D9 Ab against established MCA205 sarcomas in mice housed in specific pathogen-free (SPF) versus germ-free (GF) conditions. Anti–CTLA-4 controlled tumor progression in SPF but not in GF mice (Fig. 1, A and B). Moreover, a combination of broad-spectrum antibiotics (ampicillin + colistin + streptomycin, ACS) (Fig. 1C), as well as imipenem alone, but not colistin (Fig. 1C), compromised the antitumor effects of anti–CTLA-4. These results, which suggest that the gut microbiota is required for the anticancer effects of CTLA-4 blockade, were confirmed in the Ret melanoma and the MC38 colon cancer models (fig. S1, A and B). In addition, in GF or ACS-treated mice, the anti–CTLA-4–induced activation of splenic effector CD4+ T cells and TILs was significantly decreased (Fig. 1, D and E, and fig. S1, C to E).

We next addressed the impact of the gut microbiota on the incidence and severity of intestinal lesions induced by anti–CTLA-4 treatment. A 'subclinical colitis' dependent on the gut microbiota was observed at late time points (figs. S2 to

Downloaded from www.sciencemag.org on November 9, 2015

NEWS | IN DEPTH

Academy of Sciences with colleagues at the University of Chicago in Illinois, Sverjensky extended knowledge about the dielectric constant to 60 kilobars, or about 200 kilometers—well into the mantle. It was a theoretical result, driven by a branch of computational chemistry called ab initio molecular dynamics, which predicts chemistry based on quantum mechanical first principles. "It's a hard calculation, and it requires a huge amount of computer time," says Craig Manning, an experimental geochemist at the University of California, Los Angeles.

The result suggested that water was a better solvent than expected at high pressures—and that it could hold in solution many more ions, such as carbonate. Previously, many geoscientists thought that the subduction of carbonate-bearing limestone would bury the carbon permanently in the mantle. But Sverjensky's calculations implied that carbonate dissolved in deep water could percolate up through the crust and back to the surface—an important new way for carbon to get back in circulation, where it can ultimately affect (among other things) atmospheric chemistry and climate. "It really tunes the deep carbon cycle," Daniel says.

To check his results experimentally, Sverjensky turned to Daniel, who creates deep-Earth pressures in her laboratory in France. In one key experiment, Daniel and colleagues put water and a single crystal of calcite—limestone—inside a gold-lined metal cylinder, smaller than a human hair. Then they squeezed the cylinder to temperatures up to 1000°C. He plans to offer workshops to help other researchers familiarize themselves with the model. In the meantime, he is skimming off some of the creamiest science. In a paper published last year, Sverjensky used his model to show that under the conditions in Earth's mantle, nitrogen can exist in water as N_2, which can easily spew back into the atmosphere through volcanoes. That helps explain why Earth's atmosphere is so much more nitrogen-rich than those of Venus and Mars, where the deep nitrogen is more likely to stay put in mantle rocks as ammonium (NH_4^+).

The new diamond-forming mechanism, meanwhile, could help explain why some diamonds have layered textures: Changes in the acidity of the water could make crystals grow and dissolve in fits and starts. The result might also force researchers to rethink a long-standing assumption that a high proportion of light carbon isotopes in diamond indicates that the carbon originally hailed from organic matter on the sea floor. Changes in acidity can create similar mixes of isotopes, Sverjensky says. "So diamonds with these isotopic signals may have nothing to do with ancient life."

Even more startling may be the

Layers in a diamond may be the result of changes in acidity of a carbon-bearing fluid.

they ramped up pressures and temperatures to understand. The field of theoretical ecology has a long history of using network models that are specifically intended to deal with large and

says, because "it's easier to change your gut microbiota than your genome."

MICROBIOME

Microbes aid cancer drugs

Gut bacteria boost immunotherapies

By Mitch Leslie

A new class of cancer treatments that unleash the power of the immune system on tumors may depend on some unlikely allies. Two studies of mice in this week's issue of *Science* demon-

RESEARCH | REPORTS

MICROBIOME

The ecology of the microbiome: Networks, competition, and stability

Katharine Z. Coyte,[1,2†] Jonas Schluter,[1,3,4†] Kevin R. Foster[1,2‡]

The human gut harbors a large and complex community of beneficial microbes that remain stable over long periods. This stability is considered critical for good health but is poorly understood. Here we develop a body of ecological theory to help us understand microbiome stability. Although cooperating networks of microbes can be efficient, we find that they are often unstable. Counterintuitively, this finding indicates that hosts can benefit from microbial competition when this competition dampens cooperative networks and increases stability. More generally, stability is promoted by limiting positive feedbacks and weakening ecological interactions. We have analyzed host mechanisms for maintaining stability, including immune suppression, spatial structuring, and feeding of community members—and support our key predictions with recent data.

The human microbiome contains hundreds of species and trillions of cells that reside predominantly in the gastrointestinal tract (*1*, *2*). These microbes provide many health benefits, including the breakdown of complex molecules in food, protection from pathogens, and healthy immune development (*3*–*6*). The gut microbiome is often noted for its ecological stability. Different people may carry different microbial species, but any one individual tends to carry the same key set of species for long periods (*6*–*8*). This stability is considered critical for host health and well-being, because it ensures that beneficial ecological and evolutionary functions and provide health benefits for the host (*9*, *10*–*14*). There is a clear rationale for this assumption. Competition between microbes—captured by the number and magnitude of mutually negative interactions in our models—is associated with both antibiotic warfare (*15*, *16*) and a reduction in the cooperative secretions that promote community productivity (*17*). Both have the potential to severely reduce the efficiency of any cooperative metabolism that benefits the host (*1*, *3*, *11*–*14*, *18*). However, although intuitive, this argument neglects the potential effects of cooperation or competition on the ecological stability of microbiome communities.

To understand ecological stability in the microbiome, therefore, we must consider the specific effects of cooperation and how cooperation interacts with other community characteristics, such as microbiome diversity, interference competition dynamics. To do this, we have developed a dedicated body of theory based on Wigner's semicircle law and subsequent analyses (*19*, *20*, *30*). First, we derive a general analytic stability criterion that captures all potential community types, covering the full ranges of possible interactions and species diversities (Fig. 1 and supplementary method 1 (*S1*)). We develop our theory for unstructured ecological networks because, unlike in plant-pollinator communities or food webs (*31*, *32*), there is no evidence of strong structuring within microbial communities (*16*). However, although we lack such strong structure features in these communities, our mathematics

© 2015 AAAS

Downloaded from www.sciencemag.org on November 9, 2015

REVIEW SUMMARY

MICROBIOME

Microbiomes in light of traits: A phylogenetic perspective

Jennifer B. H. Martiny,* Stuart E. Jones, Jay T. Lennon, Adam C. Martiny

BACKGROUND: Microbial communities—microbiomes—are intricately linked to human health and critical ecosystem services. New technologies allow the rapid characterization of hundreds of samples at a time and provide a sweeping perspective on microbiome patterns. However, a systematic understanding of what determines microbiome diversity and composition and its implications for system functioning is still lacking. A focus on the phenotypic characteristics of microorganisms—their traits—offers a path for interpreting the growing amount of microbiome data. Indeed, a variety of trait-based approaches have been proposed for plants and animal communities, and this approach has helped to clarify the mechanisms underlying community assembly, diversity-process relationships, and ecosystem responses to environmental change.

Although there is a growing emphasis on microbial traits, the concept has not been fully appreciated in microbiology. However, a trait focus for microorganisms may present an even larger research opportunity than for macro organisms. Not only do microorganisms play a central role in nutrient and energy cycling in most systems, but the techniques used to characterize microbiomes usually provide extensive molecular and phylogenetic information.

ADVANCES: One major difference between macro- and microorganisms is the potential for horizontal gene transfer (HGT) in microbes. Higher rates of HGT mean that many microbial traits might be more

lated to the history of the vertically descended parts of the genome. If true, then the taxonomic composition of a microbiome might reveal little about the health or functioning of a system. We first review key aspects of microbial traits and then recent studies that document the distribution of microbial traits over the tree of life. A synthesis of these studies reveals that, despite the promiscuity of HGT, microbial traits appear to be phylogenetically conserved, or not distributed randomly across the tree of life. Further, microbial traits appear to be conserved in a hierarchical fashion, possibly linked to their biochemical and genetic complexity. For instance, traits such as pH and salinity preference are relatively deeply conserved, such that taxa within deep clades tend to share the trait. In contrast, other traits like the ability to use simple carbon substrates or to take up organic phosphorus are shallow conserved, and taxa share these traits only within small, shallow clades.

OUTLOOK: The phylogenetic, trait-based framework that emerges offers a path to interpret microbiome variation and its connection to the health and functioning of en-

that means maintaining human health or mitigating climate change impacts. Future

ON OUR WEB SITE

Read the full article at http://dx.doi.org/10.1126/science.aac8480

© 2015 AAAS

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 $f(n) \leq (5n+5)/3$, and that $f(n) \geq 17n/16$ 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 $3n/2-1 \leq g(n) \leq 2n+3$.

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 $n \geq 4$ 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 $n \geq 4$, $f(n) \geq n$. By elaborating on this argument, M.R. Garey, D.S. Johnson and S. Lin [2] showed that $f(n) \geq n+1$ for $n \geq 6$.

For upper bounds—algorithms, that is—it was known that $f(n) \leq 2n$. 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

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

© 1995 Society for Industrial and Applied Mathematics
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

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！

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
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
各学科所有博士的百分比（美国，2011年）
各学科对才华的重视指数
更少强调才华
更加强调才华
80
70
60
50
40
30
20
10
0
3.2
3.7
4.2
4.7
5.2
教育学
心理学
艺术史
传播学
社会学
人类学
考古学
西班牙语
语言学
比较文学
英语文学
分子生物学
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
神经科学
分子生物学
统计学
化学
生物化学
工程学
物理学
数学
哲学
演化生物学
天文学
计算机科学

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：产生价值！

[公司文件] 加强与国内大学合作，吸纳全球优秀人才，共同推动中国基础研究
——任总与中国科学技术大学包信和校长座谈的讲话 荐

2018-12-13 17:59 6799 113 只看楼主

总 裁 办 电 子 邮 件

电邮讲话【2018】128号　　签发人：任正非

加强与国内大学合作，吸纳全球优秀人才，共同推动中国基础研究

——任总与中国科学技术大学包信和校长座谈的讲话

2018年11月19日

在高校学科设置上，我特别支持你们重视统计学。计算机科学不仅仅是技术，还应该以统计学为基础。大数据需要统计学，信息科学需要统计学，生命科学也需要统计学。国家要搞人工智能，更要重视统计学。统计学不是一个纯粹的学科，而是每一个学科都要以统计学为基础。

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！

it's a long long story

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
Liebniz – Binary Logic.
Turing machines
Information Theory
Weiner & Cybernetics
Von Neumann Architecture.
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.
Database Marketing
Data Mining, Knowledge Discovery
"Data science, classification, and related methods."
First IBM Computers
DBMS.
1989 First KDD Workshop
Gregory Piatetsky-Shapiro.
Data Technology
Catrography
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。

生物统计学：生物学背景知识

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

生物统计学：统计建模和人工智能背景知识

人工智能与机器学习、深度学习的关系

云栖社区 yq.aliyun.com

Introduction

- Artificial Intelligence
- Machine learning
- “Deep” learning

Artificial Intelligence

Artificial intelligence (AI, also machine intelligence, MI) is intelligence displayed by machines, in contrast with the natural intelligence (NI) displayed by humans and other animals.

Hatley, L. (2016). Presentation, New Designs for Learning: Games and Gamification
https://en.wikipedia.org/wiki/Applications_of_artificial_intelligence

Big-data

Source: The Connectivist

Source: Cisco Global Cloud Index

Big-data

facebook

1B+ USERS
30+ PETABYTES

WIKIPEDIA
The Free Encyclopedia

32 million
pages

YouTube

100+ hours video
uploaded every minute

twitter

645 million users
500 million tweets / day

Big-model

Convolution
Fully connected
Source: University of Bonn
$\Delta\theta_{(\mathcal{D})}$
L0 (Input)
L1
L2
L3
L4
F5
F6

Big-model

Google Brain
Deep Learning
for images:
1~10 Billion
model parameters

Multi-task Regression
for simplest whole-genome analysis:
100 million ~ 1 Billion
model parameters

Topic Models
for news article analysis:
Up to 1 Trillion
model parameters

Collaborative filtering
for Video recommendation:
1~10 Billion
model parameters

NETFLIX

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

[Simonyan arxiv 2014]

Year 2015

MSRA

[He ICCV 2015]

Big-data + Deep learning

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

Figures copyright IEEE, 2015. Reproduced for educational purposes

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

Machine Learning

Machine learning is the science of getting computers to act without being explicitly programmed.

General workflow of Machine Learning

Pat Alvarado et al,2016,Building the Machine Learning Infrastructure

Machine Learning

Machine Learning

拥抱人工智能从机器学习开始

- 机器学习最大的特点是利用数据而不是指令来进行各种工作，其学习过程主要包括：数据的特征提取、数据预处理、训练模型、测试模型、模型评估改进等几部分。

机器学习算法是使计算机具有智能的关键

- 算法是通过使用已知的输入和输出以某种方式“训练”以对特定输入进行响应。代表着用系统的方法描述解决问题的策略机制。人工智能的发展离不开机器学习算法的不断进步。

机器学习算法分类

云栖社区 yq.aliyun.com

1. 线性回归：找到一条直线来预测目标值

- 一个简单的场景：已知房屋价格与尺寸的历史数据，问面积为2000时，售价为多少？

House Size sq.ft (X)	1400	1600	1700	1875	1100	1550	2350	2450	1425	17(
House Price$ (Y)	245,000	312,000	279,000	308,000	199,000	219,000	405,000	324,000	319,000	255,

云栖社区 yq.aliyun.com

云栖社区 yq.aliyun.com

线性回归的应用

- **预测客户终生价值：** 基于老客户历史数据与客户生命周期的关联关系，建立线性回归模型，预测新客户的终生价值，进而开展针对性的活动。
- **机场客流量分布预测：** 以海量机场WiFi数据及安检登机值机数据，通过数据算法实现机场航站楼客流分析与预测。
- **货币基金资金流入流出预测：** 通过用户基本信息数据、用户申购赎回数据、收益率表和银行间拆借利率等信息，对用户的申购赎回数据的把握，精准预测未来每日的资金流入流出情况。
- **电影票房预测：** 依据历史票房数据、影评数据、舆情数据等互联网公众数据，对电影票房进行预测。

2. 逻辑回归：找到一条直线来分类数据

- 逻辑回归虽然名字叫回归，却是属于分类算法，是通过Sigmoid函数将线性函数的结果映射到Sigmoid函数中，预估事件出现的概率并分类。

目标：预测肿瘤的性质

输入：肿瘤的体积，
患者的年龄

输出：良性或恶性

云栖社区 yq.aliyun.com

逻辑回归从直观上来说是画出了一条分类线。位于分类线一侧的数据，概率>0.5,属于分类A；位于分类线另一侧的数据，概率<0.5,属于分类B。

3. K-近邻：用距离度量最相邻的分类标签

- 一个简单的场景：已知一个电影中的打斗和接吻镜头数，判断它是属于爱情片还是动作片。当接吻镜头数较多时，根据经验我们判断它为爱情片。那么计算机如何进行判别呢？

4. 朴素贝叶斯：选择后验概率最大的类为分类标签

- 一个简单的场景：一号碗(C1)有30颗水果糖和10颗巧克力糖，二号碗(C2)有水果糖和巧克力糖各20颗。现在随机选择一个碗，从中摸出一颗糖，发现是水果糖。
- 问这颗水果糖(X)最有可能来自哪个碗？

似然函数

先验概率

$$P(\omega_j \mid x) = \frac{p(x \mid \omega_j) P(\omega_j)}{p(x)}$$

后验概率

证据因子
（evidence）

$$p(x) = \sum_{j=1}^{2} p(x \mid \omega_j) P(\omega_j)$$

例如上面的例子中： P(X): 水果糖的概率为5/8

P(X|C1): 一号碗中水果糖的概率为3/4

P(X|C2): 二号碗中水果糖的概率为2/4

P(C1)=P(C2): 两个碗被选中的概率相同，为1/2

则水果糖来自一号碗的概率为:

$P(C1|X)=P(X|C1)P(C1)/P(X)=(3/4)(1/2)/(5/8)=3/5

水果糖来自二号碗的概率为:

P(C2|X)=P(X|C2)P(C2)/P(X)=(2/4)(1/2)/(5/8)=2/5

P(C1|X)＞P(C2|X)

因此这颗糖最有可能来自一号碗。

朴素贝叶斯的主要应用有文本分类、垃圾文本过滤，情感判别，多分类实时预测等。

5. 决策树：构造熵值下降最快的分类树

- 一个简单的场景：
  相亲时，可能首先检测相亲对方是否有时间。如果有，则考虑进一步接触，再观察其是否有上进心，如果没有，直接Say Goodbye。如果有，则在看帅不帅，帅的可以列入候选名单。

决策树可以应于：用户分级评估、贷款风险评估、选股、投标决策等。

6. 支持向量机（SVM）：构造超平面，分类非线性数据

一个简单的场景：

要求用一根线将不同颜色的球分开，要求尽量在放更多球之后，仍然适用。 A、B两条线都可以满足条件。再继续增加球，线A仍可以将球很好的分开，而线B则不可以。

进一步增加难度，当球没有明确的分界线，用一条直线已经无法将球分开，该怎么解决？

SVM 可应用于垃圾邮件识别、手写识别、文本分类、选股等。

7. K-means：计算质心，聚类无标签数据

- 在上面介绍的分类算法中，需要被分类的数据集已经有标记，例如数据集已经标记为○或者×，通过学习出假设函数对这两类数据进行划分。而对于没有标记的数据集，希望能有一种算法能够自动的将相同元素分为紧密关系的子集或簇，这就是聚类算法。

云栖社区 yq.aliyun.com

• 8. 关联分析：挖掘啤酒与尿布（频繁项集）的关联规则

- 算法中几个相关的概念：

➢ **频繁项集：** 在数据库中大量频繁出现的数据集合。例如购物单数据中{'啤酒'}、{'尿布'}、{'啤酒', '尿布'}出现的次数都比较多。

➢ **关联规则：** 由集合 A，可以在某置信度下推出集合 B。即如果 A 发生了，那么 B 也很有可能会发生。例如购买了{'尿布'}的人很可能会购买{'啤酒'}。

➢ **支持度：** 指某频繁项集在整个数据集中的比例。假设数据集有 10 条记录，包含{'啤酒', '尿布'}的有 5 条记录，那么{'啤酒', '尿布'}的支持度就是 5/10 = 0.5。

➢ **置信度：** 有关联规则如{'尿布'} -> {'啤酒'}，它的置信度为 {'尿布'} -> {'啤酒'}

假设{'尿布', '啤酒'}的支持度为 0.45，{'尿布'}的支持度为 0.5，则{'尿布'} -> {'啤酒'}的置信度为 0.45 / 0.5 = 0.9。

9. PCA降维：减少数据维度，降低数据复杂度

- 降维是指将原高维空间中的数据点映射到低维度的空间中。因为高维特征的数目巨大，距离计算困难，分类器的性能会随着特征数的增加而下降；减少高维的冗余信息所造成的误差,可以提高识别的精度。

10. 人工神经网络：逐层抽象，逼近任意函数

- 前面介绍了九种传统的机器学习算法，现在介绍一下深度学习的基础：人工神经网络。 它是模拟人脑神经网络而设计的模型，由多个节点（人工神经元）相互联结而成，可以用来对数据之间的复杂关系进行建模。

云栖社区 yq.aliyun.com

- 例如利用单层神经网络实现逻辑与门和同或门

与门AND

$$h_\Theta(x) = g(-30 + 20x_1 + 20x_2)$$

x1	x2	hΘ(x)
0	0	g(−30)≈0
0	1	g(−10)≈0
1	0	g(−10)≈0
1	1	g(10)≈1

同或门XNOR

x1	x2	a1(2)	a2(2)	hΘ(x)
0	0	0	1	**1**
0	1	0	0	**0**
1	0	0	0	**0**
1	1	1	0	**1**

云栖社区yq.aliyun.com

- 多层神经网络的每一层神经元学习到的是前一层神经元值的更抽象的表示，通过抽取更抽象的特征来对事物进行区分，从而获得更好的区分与分类能力。

11. 深度学习：赋予人工智能以璀璨的未来

- 深度学习就是一种基于对数据进行表征学习的方法，使用多层网络，能够学习抽象概念，同时融入自我学习，逐步从大量的样本中逐层抽象出相关的概念，然后做出理解，最终做出判断和决策。通过构建具有一定“深度”的模型，可以让模型来自动学习好的特征表示（从底层特征，到中层特征，再到高层特征），从而最终提升预测或识别的准确性。

传统机器学习的数据处理流程。

深度学习的数据处理流程。

云栖社区 yq.aliyun.com

深度学习的历史变迁：

深度学习经历了三次浪潮：

- 20世纪40年代～60年年代，深度学习的雏形出现在控制论中；
- 20世界80年代～90年代，深度学习表现为联结主义；
- 2006年以后，正式以深度学习之名复兴。

- **第一次浪潮：以感知机和线性模型为代表**

  不能解决与或问题

- **第二次浪潮：以多层感知机和BP模型为代表**

  以统计学为基础，应用核函数和图模型的支持向量机算法（SVM算法）等各种浅层有监督的机器学习模型广泛应用，且深度神经网络不可训练

- **第三次浪潮：以无监督学习为代表。**

  解决了深层神经网络的计算能力问题；解决了深度神经网络后向误差反馈梯度消失的问题。

与日俱增的每个神经元的连接数

- 最初，人工神经网络中神经元之间的连接数受限于硬件能力。而现在，神经元之间的连接数大多是出于设计考虑。一些人工神经网络中每个神经元的连接数与猫一样多，并且对于其他神经网络来说，每个神经元的连接数与较小哺乳动物（如小鼠）一样多，这种情况是非常普遍的。甚至人类大脑每个神经元的连接数也没有过高的数量。
- 1. 自适应线性单元（Widrow and Hoff, 1960）；2. 神经认知机（Fukushima, 1980）；3. GPU-加速卷积网络（Chellapilla et al., 2006）；4. 深度玻尔兹曼机（Salakhutdinov and Hinton, 2009a）；5. 无监督卷积网络（Jarrett et al., 2009b）；6. GPU- 加速多层感知机（Ciresan et al., 2010）；7. 分布式自编码器（Le et al., 2012）；8. Multi-GPU 卷积网络（Krizhevsky et al., 2012a）；9. COTS HPC 无监督卷积网络（Coates et al., 2013）；10. GoogLeNet（Szegedy et al., 2014a

与日俱增的神经网络规模

- 自从引入隐藏单元，人工神经网络的规模大约每 2.4 年翻一倍。
- 1. 感知机 （Rosenblatt, 1958, 1962）； 2. 自适应线性单元 （Widrow and Hoff,1960）； 3. 神经认知机 （Fukushima, 1980）； 4. 早期后向传播网络 （Rumelhart et al., 1986b）； 5. 用于语音识别的循环神经网络 （Robinson and Fallside, 1991）； 6. 用于语音识别的多层感知机 （Bengio et al., 1991）； 7. 均匀场 sigmoid 信念网络 （Saul et al., 1996）； 8. LeNet5 （LeCun et al., 1998c）； 9. 回声状态网络 （Jaeger and Haas, 2004）； 10. 深度信念网络 （Hinton et al., 2006a）； 11. GPU- 加速卷积网络 （Chellapilla et al., 2006）； 12. 深度玻尔兹曼机 （Salakhutdinov and Hinton, 2009a）； 13. GPU加速深度信念网络 （Raina et al., 2009a）； 14. 无监督卷积网络 （Jarrett et al., 2009b）； 15. GPU- 加速多层感知机 （Ciresan et al., 2010）； 16. OMP-1 网络 （Coates and Ng, 2011）； 17. 分布式自编码器 （Le et al., 2012）； 18. MultiGPU 卷积网络 （Krizhevsky et al., 2012a）； 19. COTS HPC 无监督卷积网络 （Coates et al., 2013）； 20. GoogLeNet （Szegedy et al., 2014a）

AI in a nutshell

- 目前深度学习的应用十分广泛，例如图像识别、语音识别、机器翻译、自动驾驶、金融风控、智能机器人等。

Finance

Radiology

云栖社区 yq.aliyun.com

Machine Learning in biology

Machine learning algorithm: supervised and unsupervised

An exemplification of Machine learning in biology : classification model

Angermueller C, et al, 2016, Deep learning for computational biology

Machine Learning in biology

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

生物统计学：案例分析

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

Gene coding region

Promoter region

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
Exon 1
Intron1
Exon N
3’ UTR
5’ UTR
GT
AG
Splice Sites
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

Bioinformatics @ math.pku

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

我们有讲义啦

- 《生物统计学：生物大数据的概率统计模型与机器学习方法》，宁康，2020。

还有配套习题集：

- 《生物统计学习题集》，宁康，2020。

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
〔英〕 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)

Slides credits

——生物信息学研究方法概述：北京大学生物信息中心

——生物统计学：卜东波@中国科学院计算技术研究所，邓明华@北京大学

——神经网络与深度学习: 邱锡鹏@复旦大学

——Introduction to Computational Biology and Bioinformatics: Xiaole Shirley Liu Lab@Havard University

——Combinatorial Methods in Computation Biology: Ken Sung Lab@NUS

——Deep Learning in the Life Sciences: MIT

——Probabilistic Graphical Models: Eric Xing@CMU

——Numerous other leading researchers and leading labs.....

补充知识

生物信息学与生物统计学：

- 两种视角
- 四种技术
- 两类方法

生物信息学与生物统计学：两种视角，四种技术，两类方法

- 两种视角
  - 生物学视角
  - 计算视角
- 四种技术
  - 算法
  - 数据
  - 超算
  - 深度学习
- 两类方法
  - 经典软件
  - 核心工具

生物信息学：生物学视角

我们是谁？我们从哪里来？我们到哪里去？

先看视屏！

- 开讲啦：如何健康地活到100岁？（汪健@CCTV, http://www.biodiscover.com/talk/research/103880.html）
- 基因组学（Barry Schuler@TED, https://www.ted.com/talks/barry_schuler_genomics_101）
- 微生物组学（Rob Knight@TED, https://www.ted.com/talks/rob_knight_how_our_microbes_make_us_who_we_are）
- 一个IT人的故事（Larry Smarr@TED, http://www.tedmed.com/talks/show?id=18018）
- 大数据（Kenneth Cukier@TED, https://www.ted.com/talks/kenneth_cukier_big_data_is_better_data）

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）
- 。。。

Understand it, create it!

DNA数据存储的现在和未来

Understand it, create it!

Original Image

原始图像

Image Reconstructed From Bacteria

从细菌DNA还原的图像

Understand it, create it!

Our test on DNA storage: how can it tolerate mutations?

5% SNPs

2% SNPs

1% SNPs

0.1% SNPs

0.05% SNPs

Understand it, create it!

CRISPR-Cas9

See it!

Understand it!

Create it!

Understand it, create it!

读物推荐

Understand it, create it!

Understand it, create it!

CRISPR和基因编辑技术

OR

CRISPR和基因编辑技术

生物信息学：计算科学视角

生物信息学@HUST

我们是谁？我们从哪里来？我们到哪里去？

生命科学的前沿 – 生物大数据

生物大数据 – 机遇与挑战

Next generation sequencing…

大数据的挑战 – 需要高通量的解决方案

1962

高通量测序

2003

2014

大数据的挑战 – 需要高通量的解决方案

1951

>100候选，>10000实验

1971

2015

2003

>>60亿研究对象，
>>EB(1000,000TB)数据量

????

如果没有
大数据分析技术...

大数据的挑战 – 需要高通量的解决方案

2003

>60亿研究对象，>>EB数据量

!!!!

光有机器也枉然...

亟需高性能大数据分析技术...

生物医学大数据3V特征

生物医学大数据涉及到广泛的领域

生物医学大数据涉及到完整研究链条

生物医学大数据依赖于大数据的分析方法

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

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*

生物信息学与生物统计学：两种视角，四种技术，两类方法

- 两种视角
  - 生物学视角
  - 计算视角
- 四种技术
  - 算法
  - 数据
  - 超算
  - 深度学习
- 两类方法
  - 经典软件
  - 核心工具

Binformatics

- Ontology and pathway (DAVID, GSEA)
- Batch removal (COMBAT)
- NGS for DNA-seq (FastQC, BWA, STAR)
- RNA-seq (RSeQC, Salmon, DESeq2)
- scRNA-seq (Seurat)
- ChIP/DNase-seq (MACS, Cistrome)
- scATAC-seq (Seurat / MAESTRO)
- Genome-resequencing (Mutect)
- CRISPR screens (MAGeCK)
- Immunology (Polysolver, NetMHC, TIMER, TRUST, TIDE)
- Do not be afraid to learn new algorithms

Bioinformatics Research

- Level 0: modeling for modeling's sake
- Published in low impact non-biological journals, read and cited only by non-biologists

Bioinformatics Research

- Level 1: analyze unpublished data from a collaborator's project
- Is it an important biological problem?
- Are there existing algorithms to analyze the data?
- Does the analyses and experiments inform each other?
- Does the study have real biological findings?
- Does the experimental group value the bioinformatics contribution?
- Familiarity of bioinformatics algorithms and understanding of the biological problem are important
- Can learn domain biological knowledge and motivate methodology development
- Avoid being a data janitor.

Bioinformatics Research

- Level 2: algorithm development for mining biological data
- Have a keen eye on cost-effective new techniques
- Use sound statistics and develop user-friendly software
- Continue to maintain and update algorithm with new data and new challenges
- Stay in the biological domain, and provide complete solution to the domain's problem
- User adoption is more important than publication journal IF
- Time is the ultimate measure: Is the algorithm and cited? Do people use it to make real findings?
- Do well by doing good.

Bioinformatics Research

- Level 3: make biological discoveries from public data integration and modeling
- True big data mining!
- Have unique biological angles, and not necessarily use sophisticated algorithms
- Make real novel biological discoveries rather than “putting numbers on theories”
- Is there experimental validation informatics group or their exper collaborators)?

Bioinformatics Research

- Level X: integrative papers from big science consortia
- Sometimes have numbers on the paper title
- Exhausting and time-consuming, political struggles of the consortium members
- Published in high profile journals and highly cited, but contributions often dubious
- Trainees might get some good papers / faculty positions, but may not do well later
- Informatics group should caution not to sacrifice Levels I, II, III for Level X-only research

Break

生物信息学研究的三个层面

初级层面
中级层面
高级层面

初级层面

基于现有的生物信息数据库和资源，利用成熟的生物信息学工具（专业网站、软件）解决生物信息学问题

——生物信息数据库（NCBI、EBI等）

——基因组序列分析、序列比对软件（GCG、BLAST、CLUSTAL等）

——系统发育树构造软件（PHYLIP、PALM、MEGA等）

——分子动力学模拟软件（GROMACS、NAMD等）

——搜集、整理有特色的生物信息学数据集

中级层面

利用数值计算方法、数理统计方法和相关的工具，研究生物信息学问题

——概率、数理统计基础

——科学计算基础

——现有的数理统计和科学计算工具（EXCEL、SPSS、SAS、MATLAB等）

——建立有特色的生物信息学数据库

高级层面

提出有重要意义的生物信息学问题；自主创新，发展新型方法，开发新型工具，引领生物信息学领域研究方向。

——面向生物学领域，解决生物学问题

——数学、物理、化学、计算科学等思想和方法

——建立模型，发展算法

——自行编程，开发软件，建立网页（Linux系统、C/C++、PERL、数据库技术）

从事生物信息学研究应具备多方面的科学基础：

（1）、一定的计算能力，包括相应的软、硬设备。要有各种数据库或者能与国际、国内的数据库系统进行有效的交流。要有发达、稳定的互联网络系统；
（2）、强有力的创新算法和软件。没有算法创新，生物信息学就无法获得持续的发展；
（3）、与实验科学，特别是与自动化的大规模高通量的生物学研究方法与平台技术建立广泛、紧密的联系。这些技术，既是产生生物信息数据的主要方法，又是验证生物信息学研究结果的关键手段。

从事生物信息学研究的人员必须具备多学科交叉的知识。

生物信息学的“降龙十八掌”

（1）

要掌握生物信息数据库及其查询搜索方法

(Database & searching)

——对分子生物信息数据库的种类以及某些具体数据库的掌握和了解

——从现有数据库中熟练获得需要的数据信息（尤其是二级数据库）

——能熟练地进行数据库查询和数据库搜索（数据库查询系统**Entrez**、**SRS**；搜索工具**BLAST**等）

——数据库技术、互联网技术

（2）

要学会生物信息学软件和工具的应用

(Software & application)

利用成熟的生物信息学工具（专业网站、软件）解决生物信息学问题

——基因组序列分析、序列比对软件（**GCG**、**BLAST**、**CLUSTAL**等）

——系统发育树构造软件（**PHYLIP**、**PALM**等……）

——基因芯片检测分析软件（商业软件**ScanArray**、**Array-Pro**等 ……）

——分子动力学模拟软件（**GROMACS**、**NAMD**等……）

（3）

概率论基础

(Probability theory)

——随机事件、概率

——随机变量、概率分布

——大数定律、中心极限定理

——几乎用于生物信息学的各个方面

"Most of the problems in computational sequence analysis are essentially statistical."

——"Biological sequence analysis"

（4）

数理统计基础

(Statistical methods)

——样本和统计量（方差、均值……）

——参数估计、假设检验

——基本的统计分析（方差分析、协方差分析、回归分析）

——常用统计软件的运用（SPSS、SAS）

——几乎用于生物信息学的各个方面

第五式　羚羊触藩

（5）

基于频率的组分分析方法和权重矩阵方法

(Composition analysis & weight matrix method)

——符号（如碱基）频率反映具有生物学意义的序列特征，如内含子剪接位点的发现，KOZAK规则的发现等

——核酸组分、氨基酸组分、密码子使用频率

——主要用于具有特定生物学意义的序列特征的分析

Figure 1. Logo for *E. coli* ribosome binding sites. Only −18 to +8 of the −20 to +13 site is shown. The first translated codon is just to the right of the 2 bit high vertical bar. 149 natural sites were used to create the logo (9).

权重矩阵分析方法举例

——针对序列信号（一段核酸、蛋白），计算每一位点所使用的词汇或叫符号（*碱基、氨基酸*）频率，频率的偏好性反映信号的序列特征（**sequence pattern**）。

例：人类基因内含子/外显子剪接位点的序列特征分析

Figure 6–28. Molecular Biology of the Cell, 4th Edition.

Bayesian打分函数用于剪接位点预测的公式

The likelihood that a property value *v* (of a new structure) is drawn from the splicing site is:

$$P(site \mid \upsilon) = \frac{P(\upsilon \mid site)P(site)}{P(\upsilon \mid site)P(site) + P(\upsilon \mid nonsite)P(nonsite)'}$$

Score for the overall likelihood of the query sequence being a site is:

$$\sum_{\substack{\text{properties at} \\ \text{associated volumes}}} \log\left(\frac{P(site \mid \upsilon)}{P(site)}\right)$$

Say we have a sequence S= $S_1S_2\ldots S_n$. Then one need to calculate

$$\frac{P(S \mid \text{splice site})}{P(S \mid \text{background})}$$

So to look for a donor site in the sequence, we might calculate

第六式　潜龙勿用

（6）

信息论方法

(Information method)

——信息的度量：是信息符号出现何种状态的一种不确定性程度，信息的获得要对不确定性进行否定。

——生物信息的符号如ACGT四种符号，状态空间即其所有可能的排列

——用于结构预测

——信息熵 $$H = -\sum_i p_i \log p_i$$

——信息熵H刻画了由$\{p_i\}$表示的随机试验结果的先验不确定性，或观察到输出时所获得的信息量。

（7）

期望最大化（EM）方法

(Expectation Maximization)

——EM算法是一种在不完全数据情况下计算极大似然估计或者后验分布的迭代算法。

——适用于具有隐变量的模型和问题，

——用于结构的识别，如Motif识别的MEME方法、HMM中的Baum-Welch算法

（8）

动态规划方法

(Dynamic Programming)

——一种常用的多阶段决策的寻优算法

——动态规划用得最多的方面是**DNA**序列或者蛋白质序列比对

（9）

迭代方法

(Iteration)

——迭代的目的通常是在状态空间找到目标函数收敛的稳定解

——在运用模式识别方法时，对系统参数的学习通常要经过迭代来实现

——迭代必须能够不断逼近稳定解

——用于上述某些方法的方法

（10）

回归、拟合、相关性分析、关联分析

(Regression, fitting, correlation & association)

——经典的统计分析方法
——主要目的：描述和预测自变量与因变量间的关系
——用于上述某些方法的方法

（11）

判别分析方法

(Discriminant analysis)

——用于判别样品所属类型的统计分析方法

条件：已知研究对象总体的类别数目及其特征（如：分布规律，或各类的训练样本）

目的：判断未知类别的样本的归属类别

——用于基因识别、医学诊断、人类考古学

第十二式　鱼跃于渊

（12）

聚类分析方法

(Clustering method)

——聚类分析（群分析）是实用多元统计分析的一个新分支，正处于发展阶段。理论上尚未完善，但应用十分广泛。实质上是一种分类问题，目的是建立一种分类方法，将一批数据按照特征的亲疏、相似程度进行分类。

——条件：研究对象总体的类别数目未知，也不知总体样本的具体分类情况

——目的：通过分析，选定描述个体相似程度的统计量、确定总体分类数目、建立分类方法；对研究对象给出合理的分类。(“物以类聚”是聚类分析的基本出发点)

——定性、经验的分类的局限

分类较粗、数据量小、凭借经验

——谱系聚类法（系统聚类法）、动态聚类法、模糊聚类法

——生物信息学中的聚类分析问题：

根据DNA芯片获得的基因表达数据进行基因聚类（数据量庞大）

蛋白质相互作用网络的分类

根据不同物种的大分子序列进行相似性比较并构建系统发育树

（13）

Markov模型的应用

(Markov model)

——Markov过程：从一种状态转移到另一种状态时，过程仅取决于前面n种状态，是一种有序n模型。 n是影响下一个状态选择的状态数。

——最简单的Markov过程是一阶过程，状态的选择完全取决于前一状态，这种选择是依照概率来选择的。

——状态的选择是概率的，而非确定的。故Markov过程本质上是一种随机过程。

第十四式　损则有孚

（14）

隐Markov模型方法

(HMM method)

——将核苷酸序列看成一个随机序列，DNA序列的编码部分与非编码部分在核苷酸的选用频率上对应着不同的Markov模型。由于这些Markov模型的统计规律是未知的，而HMM能够自动寻找出它们隐藏的统计规律。对于高等生物这样复杂的DNA序列，HMM必须学习不同的基因结构的信号。

（1）序列比较与搜寻（尤其是多序列比对）

（2）基因及信号的识别、预测（包括DNA编码与非编码区的识别、真核基因剪接位点信号识别、非编码区的转录调控信号识别、信号肽识别……）

（3）蛋白质二级结构、家族、超家族预测、分类等……

（15）

感知器与人工神经网络方法

(Perceptron & ANN method)

——计算机人工神经网络是对大脑神经网络的模拟，在生物信息学研究中，无论是基因识别还是蛋白质结构预测，神经网络都取得了比其它方法更为准确的结果。

Axon (Carries signals away)
Nucleus
Cell
Dendrites (Carries signals in)
Synapse size changes in response to learning

X1
W1
X2
W2
Wn
Xn
Y = F(X)
Y

Axon
Synapse
Dendrites

IN1
IN2
IN3
IN4
OUT

（16）

决策树、支持向量机及其它模式识别方法

(Decision tree & SVM method)

——模式识别是在输入样本中寻找特征并识别对象的一种方法。

——模式识别主要有两种方法，一种是根据统计特征进行识别，另一种是根据对象的结构特征进行识别，而后者常用的方法为句法识别。

——在基因识别中，对于DNA序列上的功能位点和特征信号的识别都需要用到模式识别。

（17）

微分方程的数值方法

(Numerical methods)

——分子动力学模拟：研究生物大分子的构象，主要还是用基于半经验势函数的分子动力学方法，而量子力学则在确定势函数的参数和研究局部性质时起作用。对蛋白质进行动力学研究是利用计算机进行模拟实验的基础。

——分子动力学得到一组动力学微分方程，要求得到初值问题的解。

——微分方程的数值求解：有限差分法、有限元法

（18）

最终要诀：各类方法综合运用

All in one!

——综合运用不同的研究方法

——始终面向生物学问题

——知识和技能的学习方法

——文献的查阅和阅读方法

——中、英文论文的写作方法

生物信息学的“东邪西毒南帝北丐中顽童”

——东邪：**de Bruijn**图算法（基因组拼接）

——西毒：近似算法和似然估计（进化树分析）

——南帝：核函数（基因和物种分类问题）

——北丐：统计建模（序列分析）

——中顽童：深度学习和生成对抗神经网络**GAN**（判断问题）

Slides credits

——生物信息学研究方法概述：北京大学生物信息中心

——生物统计学：卜东波@中国科学院计算技术研究所，邓明华@北京大学

——神经网络与深度学习: 邱锡鹏@复旦大学

——Introduction to Computational Biology and Bioinformatics: Xiaole Shirley Liu Lab@Havard University

——Combinatorial Methods in Computation Biology: Ken Sung Lab@NUS

——Deep Learning in the Life Sciences: MIT

——Probabilistic Graphical Models: Eric Xing@CMU

——Numerous other leading researchers and leading labs.....

生物信息学与生物统计学：两种视角，四种技术，两类方法

- 两种视角
  - 生物学视角
  - 计算视角

- 四种技术
  - 算法
  - 数据
  - 超算
  - 深度学习

- 两类方法
  - 经典软件
  - 核心工具

1. 算法

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 $f(n) \leq (5n+5)/3$, and that $f(n) \geq 17n/16$ 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 $3n/2-1 \leq g(n) \leq 2n+3$.

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 $n \geq 4$ 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 $n \geq 4$, $f(n) \geq n$. By elaborating on this argument, M.R. Garey, D.S. Johnson and S. Lin [2] showed that $f(n) \geq n+1$ for $n \geq 6$.

For upper bounds—algorithms, that is—it was known that $f(n) \leq 2n$. 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.

实例

How many reversal steps for this REAL case?

Ming Li （李明）& Tao Jiang （姜涛）

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

© 1995 Society for Industrial and Applied Mathematics
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

姚期智（Andrew Yao）

也许觉得算法在今后不会碰到那么多。。。

终身学习。。。

终身成长
卡罗尔·德韦克
坐镇美国亚马逊心理类畅销榜10年
累计印量180万册
MINDSET
The New Psychology of Success
CAROL S.DWECK, Ph.D.

TWO MINDSETS
CAROL S. DWECK, Ph.D.
Graphic by Nigel Holmes
Fixed Mindset
Intelligence is static
Growth Mindset
Intelligence can be developed
Leads to a desire to look smart and therefore a tendency to...
Leads to a desire to learn and therefore a tendency to...
CHALLENGES
...avoid challenges
...embrace challenges
OBSTACLES
...give up easily
...persist in the face of setbacks
EFFORT
...see effort as fruitless or worse
...see effort as the path to mastery
CRITICISM
...ignore useful negative feedback
...learn from criticism
SUCCESS OF OTHERS
...feel threatened by the success of others
...find lessons and inspiration in the success of others
As a result, they may plateau early and achieve less than their full potential.
All this confirms a deterministic view of the world.
As a result, they reach ever-higher levels of achievement.
All this gives them a greater sense of free will.

Public key system

RSA algorithm

Adi Shamir, Ron Rivest and Len Adleman
1977

Collision attack

Xiaoyun Wang @ China
2004

Extension: from statistics back to DNA

Hamilton path problem

DNA computing

Solution

Leonard Adleman

SCIENCE CLASSICS

BY LARRY GONICK

Discover magazine published an article in comic strip format about Leonard Adleman's discovery of DNA computation. Not only entertaining, but also the most understandable explanation of molecular computation I have Ever seen.

Understand it, create it!

CRISPR–Cas encoding of a digital movie into the genomes of a population of living bacteria, Nature, 2017

2. 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

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

生物数据很大

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

MIT Technology Review

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)=463MB/s=3.7Gb/s，200元

他可以多装点，而且次日达可以更快的。。。

3. 超算

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

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
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

Meta-Storms

高性能的生物计算（蓝色）
HPC

利用先进的数据库和索引技术处理群落比较和搜索

Bioinformatics, 2012
NAR, in revision

Meta-Storms

Meta-Storms

高性能的生物计算（蓝色）
HPC

利用并行计算技术加速大规模群落比较和搜索

NVIDIA CUDA RESEARCH CENTER

Data source: MG-RAST database

Bioinformatics, 2014
Scientific Reports, 2017

Meta-Prism

高性能的生物计算（蓝色）
HPC

利用双引擎计算，同时提高**准确性和运行速度**

Bioinformatics, in press

4. 深度学习

Deep Learning

数据很重要！

算法更重要！

TWO TYPES OF ARTICLES ABOUT MACHINE LEARNING

MACHINE LEARNING
CLASSICAL LEARNING
UNSUPERVISED
SUPERVISED
Clustering
DBSCAN
K-Means
Agglomerative
Mean-Shift
Fuzzy C-Means
Classification
Naive Bayes
K-NN
SVM
Decision Trees
Logistic Regression
Pattern search
Euclat
Apriori
FP-Growth
Regression
Linear Regression
Polynomial Regression
Ridge/Lasso Regression
DIMENSION REDUCTION (generalization)
t-SNE
PCA
LSA
SVD
LDA
REINFORCEMENT LEARNING
Genetic Algorithm
Q-Learning
SARSA
Deep Q-Network (DQN)
A3C
ENSEMBLE METHODS
Stacking
Bagging
Random Forest
Boosting
XGBoost
AdaBoost
LightGBM
CatBoost
NEURAL NETS AND DEEP LEARNING
Convolutional Neural Networks (CNN)
DCNN
Perceptrons (MLP)
Recurrent Neural Networks (RNN)
LSM
LSTM
GRU
Generative Adversarial Networks (GAN)
Autoencoders
seq2seq

THE MAIN TYPES OF MACHINE LEARNING

Simple data
Clear features

CLASSICAL ML

No data,
but we have
an environment
to interact with

REINFORCEMENT LEARNING

When quality is
a real problem

ENSEMBLES

eternal competitors

Complicated data
Unclear features
Belief in a miracle

NEURAL NETWORKS AND DEEP LEARNING

CLASSICAL MACHINE LEARNING

Data is pre-categorized or numerical

Data is not labeled in any way

SUPERVISED

Predict a category

Predict a number

CLASSIFICATION

«Divide the socks by color»

REGRESSION

«Divide the ties by length»

UNSUPERVISED

Divide by similarity

Find hidden dependencies

Identify sequences

CLUSTERING

«Split up similar clothing into stacks»

ASSOCIATION

«Find what clothes I often wear together»

DIMENSION REDUCTION (generalization)

«Make the best outfits from the given clothes»

THE SIMPLEST SPAM-FILTER

(used until 2010)

hey	... 1829
I'm	...1710
no	... 1191
where	... 1012
you	...985
speak	... 873
learn	...747
one	... 739

good letters

672 times

«KITTY»

13 times

viagra	... 1552
casino	... 1492
100%	... 1320
credit	... 1184
sale	... 985
press	... 873
free	... 747
enlarge	... 739

spam letters

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

BAYES' THEOREM

NOT SPAM

NAIVE BAYES

SEPARATE TYPES OF ALCOHOL

SUPPORT VECTOR MACHINE

PREDICT TRAFFIC JAMS

REGRESSION

PUT KEBAB KIOSKS IN THE OPTIMAL WAY

(also illustrating the K-means method)

1. Put kebab kiosks in random places in city

2. Watch how buyers choose the nearest one

3. Move kiosks closer to the centers of their popularity

4. Watch and move again

5. Repeat a million times

6. Done!
You're god of kebabs!

SEPARATE DOCUMENTS BY TOPIC

1. Build a matrix of how often each word can be found in each document
(black - more often)

3. Get visual topic clusters.
Even if the words haven't met together

LATENT SEMANTIC ANALYSIS (LSA)

THAT MEATBAG BOUGHT A SOFA
PROBABLY, HE LOVE SOFAS!!!
RECOMMEND HIM 148 MORE SOFAS
ebay
AliExpress
amazon

DIFFERENT ALGORITHMS
SAME DATA
K-NN
DECISION TREE
SVM
FINAL DECISION ALGORITHM
ANSWER
STACKING

SAME
ALGORITHMS
BAGGING ON TREES
//
RANDOM FOREST
MAKE
DIFFERENT SETS
OF DATA FROM
INITIAL SET
!!!
JUST AVERAGING
ALL THE RESULTS
ANSWER
BAGGING

INITIAL DATASET
TREE
NEW SET OF DATA
(priority for cases on which the previous algorithm was mistaken)
WOW, ANOTHER TREE!
COLLECT BAD CASES AGAIN
GUESS WHO IS HERE AGAIN?
REPEAT UNTIL IT FEELS GOOD
BOOSTING

Neural Networks

NEURON
10
0.5
7
1.0
3
0.1
sum > 10
?
1 (YES)
12.3
10x0.5 + 7x1.0 + 3x0.1

MULTILAYER PERCEPTRON (MLP)

The neural network itself learns how to build important features from the simple lines

Perceptron finds important features specific to a cat

CONVOLUTIONAL NEURAL NETWORK (CNN

RECURRENT NEURAL NETWORK (RNN)

A mostly complete chart of

Neural Networks

©2016 Fjodor van Veen - asimovinstitute.org

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

生物信息学与生物统计学：两种视角，四种技术，两类方法

- 两种视角
  - 生物学视角
  - 计算视角

- 四种技术
  - 算法
  - 数据
  - 超算
  - 深度学习

- 两类方法
  - 经典软件
  - 核心工具

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/

生物统计经典软件

基因组可视化：Genome Browser，(http://genome.ucsc.edu/), (tracks, annotations, etc.)

序列保守性：WebLogo，(http://weblogo.berkeley.edu/logo.cgi),

基因预测：MEME, (http://meme-suite.org/).

进化树：iTOL, (https://itol.embl.de/),

基因调控网络： GeneNetwork, (http://gn2.genenetwork.org/), Cytoscape, (https://cytoscape.org/),

代谢通路：KEGG, (https://www.kegg.jp/); iPATH, (https://pathways.embl.de/),

蛋白结构与功能：PDB，(http://www.rcsb.org); pFAM, (http://pfam.xfam.org/),

微生物组: EBI Magnify. (https://www.ebi.ac.uk/metagenomics/),

蛋白和小分子互作数据：STITCH，(http://stitch.embl.de/); STRING, (http://string-db.org),

药物数据库：DrugBank, (https://www.drugbank.ca/),

生物数据分析平台：Galaxy, (https://usegalaxy.org/),

生物数据可视化：Echart, (https://www.echartsjs.com/examples/zh/index.html),

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

Current status（现今态势）

很难找到
与生物信息学没有关系的
生物学和计算科学
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

Future（未来）

我们是谁？我们从哪里来？我们到哪里去？