Algorithms in Bioinformatics

生物信息学算法原理

Chaochun Wei（韦朝春）

ccwei@sjtu.edu.cn

http://cgm.sjtu.edu.cn/

Yue Zhang（张岳）

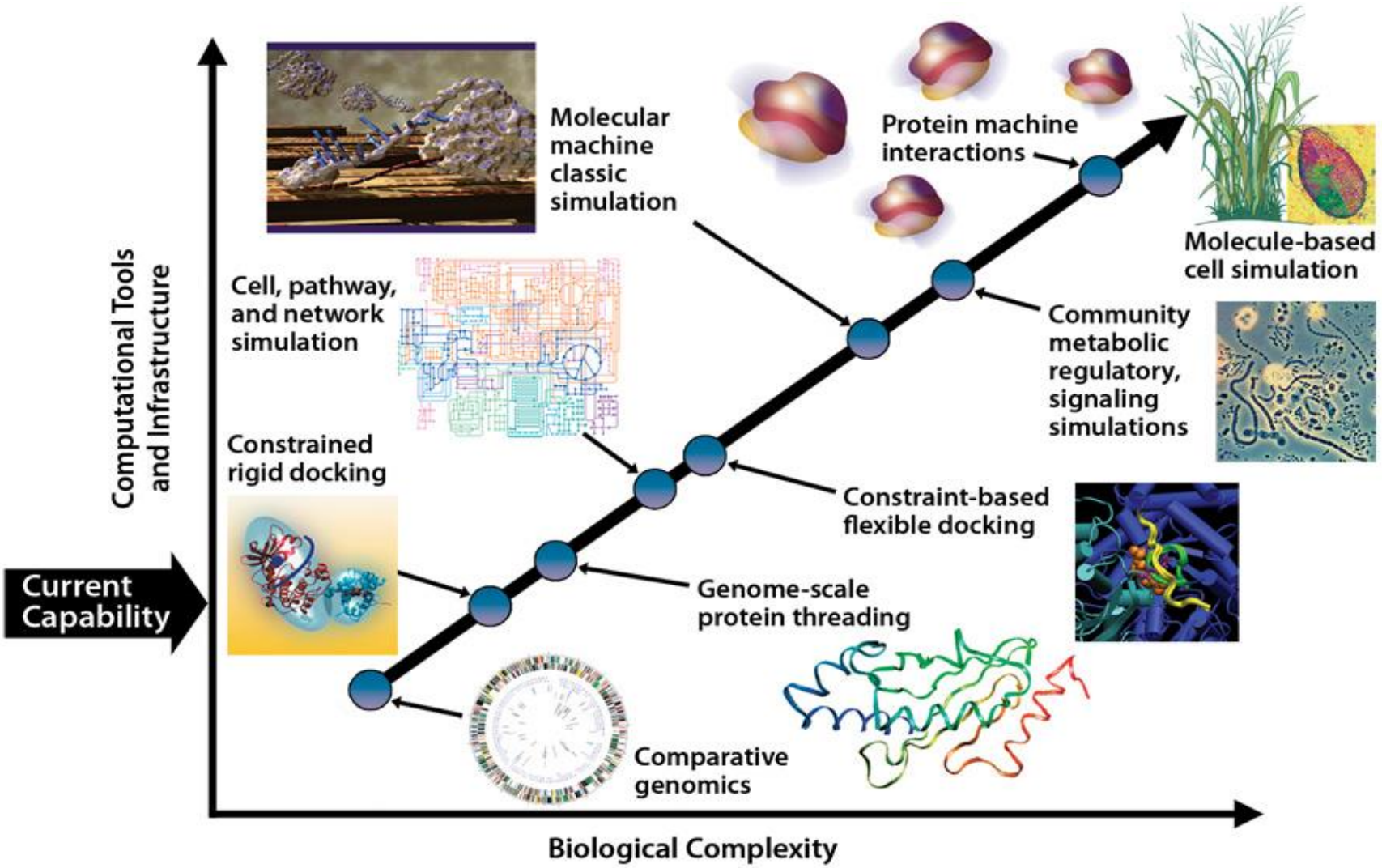
Spring 2018

Contents

- What is Bioinformatics
- What is an algorithm
- Why we need algorithm
- Course information
  - Goal
  - Contents
  - Organization
  - Grading

A Big Picture of Biology

"Biology is an information science" *-- Leroy Hood*



http://genomicscience.energy.gov/biofuels/

Bioinformatics

- **The science of collecting and analyzing complex biological data such as genetic codes.**

  -- Oxford Dictionary

- **Major research areas**
  - **Sequence analysis**
  - **Genome annotation**
  - **Computational evolutionary biology**
  - **Analysis of gene expression, regulation**
  - **Comparative genomics**
  - **Literature analysis**
  - **Biological systems modeling**
  - **Structural Biology**

Algorithm

- **A process or set of rules to be followed in calculations or other problem-solving operations, especially by a computer*: a basic algorithm for division***

  --- Oxford Dictionary

- **Examples:**
  - **Sorting**
  - **Calculation of Pi**
  - **Task arrangement**
  - **Printing**

Why do we need algorithms?

- Lots and lots of data
- Huge computation
- Limited time and space

Milestone of modern biology：the human genome project

- Feb. 15, 2001 *Nature*

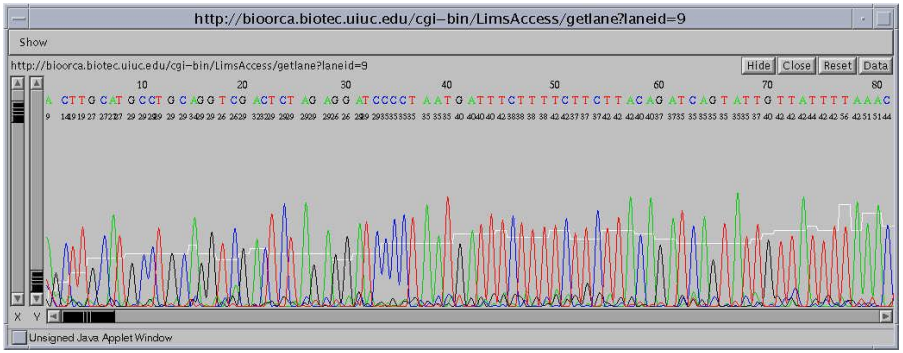
Feb. 16, 2001 *Science*

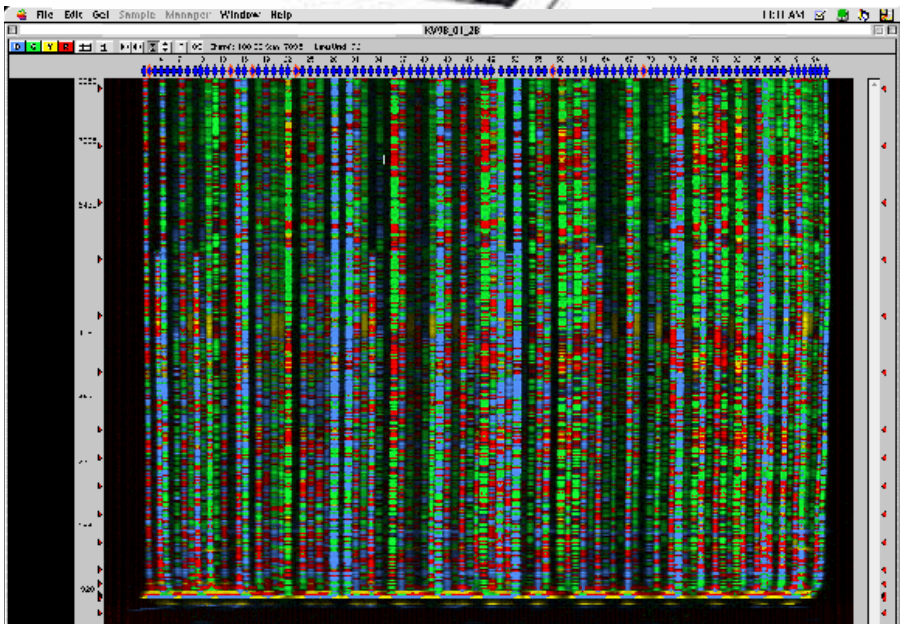
Human Gene
3 billion dollars, 3

- "This is ...
  - a WASTE of ta
    money….
  - We can do a lo

George Church
Professor of Genetics
Harvard Medical School

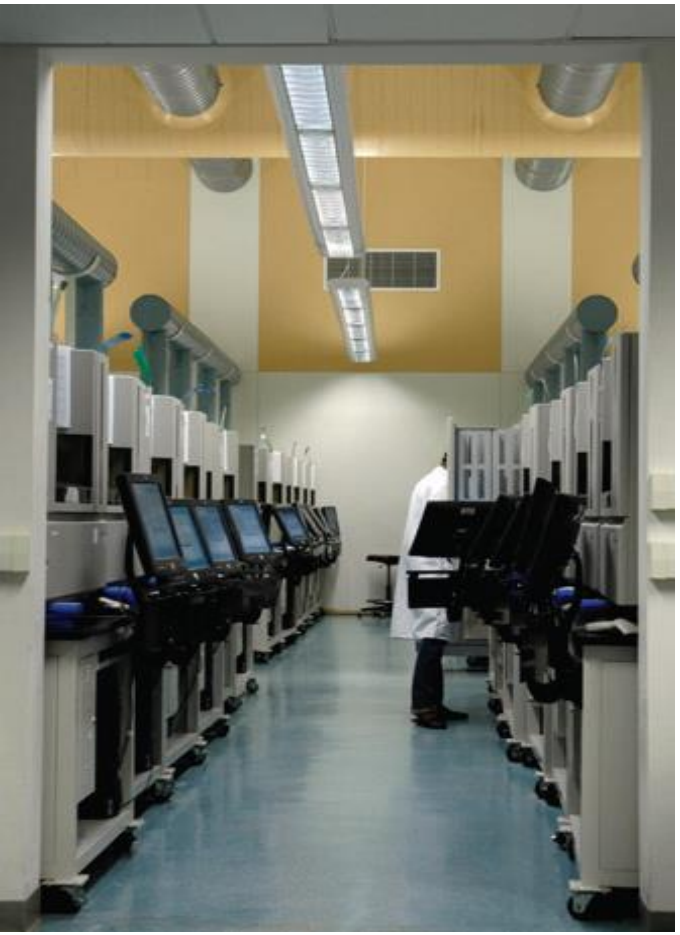
Automated high throughput sequencing





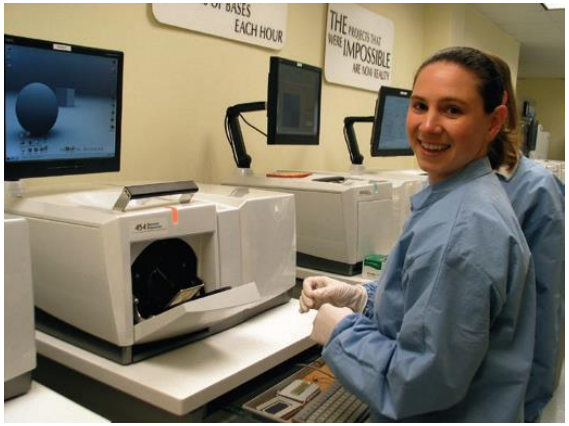
AGAACGACCATCAACTAAATCAAAATGCCTTTCAAACCAGCA
GACAACCCAAAATGCCCAAAATGCGGCAAATCCGTATACGCC
GCNGAAGAAAAAGTAGCTGGAGGATACAAATACCACAAATCC
TGCTTCAAATGCGGTATGTGCAATAAAATGCTCGACTCCACC
AACGTAACTGAACACGAAGCTGAATTGTACTGCAAAAATTGC
CATGGACGTAAATACGGACCTAAAGGATACGGATTCGGTGGT
GGAGCTGGGTGCTTAAGTATGGACGATGGAGCCCAATTCAA
GGGAACACAATAATTTTAAGAAGGAATCAATGTGAAGATGGC
GGCCAAAACCACACCAACTGTCAGCGGTCGTCAGTTCTACCC
TTTTCCATCCCCCACTATACACTAATGTAATATTTTTAGATCTT
AAATTACAGACTTAGTTTTAATTTATAAATTTTCGTATGACACG
TTATAAATAAGAATTCGGTTATTTTGTAATAATTGAATTAAATA
AATCTTATTTAAGACCAAAAAA

The revolution caused by Next-Generation Sequencing (NGS)



Sanger sequencing requires industrialized lab and many staffs

<



NGS: 2nd generation
One staff, one machine

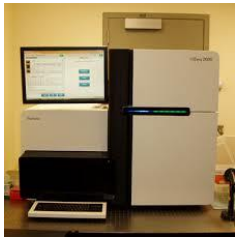
<?

3rd generation:
Oxford Nanopore

NGS platforms

Applied Biosystems
ABI 3730XL

Roche / 454
Genome Sequencer
FLX



Illumina / Solexa HiSeq

Illumina / Solexa MiSeq

HeliScope™
Single Molecule
Sequencer

Applied Biosystems
SOLiD

3rd generation:
Oxford Nanopore
MinION

Pacific Biosciences RS II

Comparison of sequencing platforms (2018.2)

Platforms	Sanger	454	HiSeq X Ten *	MiSeq *	NovaSeq *	PacBio RS II**	Nanopore
Read length	650-1100	150-1000	150	36-300	2x150	Up to 60k	Very long
# of reads/run	96	0.4-2M	5.3-6 B	12M – 50M	1.6-20B	~55,000	Up to 500
Error rate	10^-3	<10^-2	~10^-3	~10^-3	~10^-3	~10%	Varies
Cost ($/Mbp)	5000	~5	<0.01	~0.5	<0.001	~1.5	~1
Time/run	~3 hours	~7 hours	<3 days	4-56 hours	19-40hr	0.5-4 hours	No fixed run tim
Throughput	100Kb	~1Gb	1.6-1.8Tb	540Mb-15Gb	167Gb-6Tb	500Mb-1Gb	Up to 1 Gb

* http://www.illumina.com/systems
** http://files.pacb.com/pdf/PacBio_RS_II_Brochure.pdf

Latest sequencing platforms

Pacific Biosciences RS II
Read length: Maximum > 30 Kb

Ion Torrent

- **Complete Genomics (obtained by BGI, 2013)**
  - **Finish 10,000 genomes/year (2010)**
- **Oxford Nanopore**
  - **Human genome：$3000, 6 hours（2012）**
- **Visigen**
- **BIGIS (China)**
- **more…**

Latest sequencing platforms

- **HiSeq X series**
  - **Ten: > 18,000 human genomes/year**
  - **<$1000 per 30X genome**
- **NovaSeq**
  - **<$100/human genome**

Personal Genomics

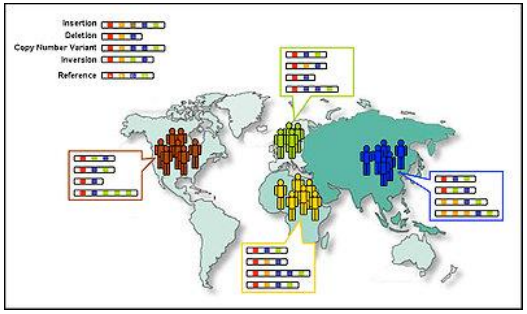
- Craig Venter genome
- James Watson genome
- 2 Koran genomes
- 1 Chinese genome
- 2 cancer genomes
- 1 African genome
- Stephen Quake genome
- Family of Four by Institute of System Biology
- ……

From Dawei Lin, http://bioinformatics.ucdavis.edu, 2011

Personal genomics projects

- 1000 Genome Projects (UK, China, US)
- ClinSeq (NHGRI)
- 23andMe Research Revolution (US)
- International Cancer Genome Consortium (Canada)

International Cancer Genome Consortium



From Dawei Lin http://bioinformatics.ucdavis.edu

Latest personal genome projects

- 23andMe Research Revolution (US)
- 10,000 Human Genome Project (USA)
- 300,000 Human Genome Project (Iceland)
- 500,000 Genome Projects (UK *)
- 1,000,000 Human Genome Project (USA)
- Cancer genomes

* http://www.ukbiobank.ac.uk/

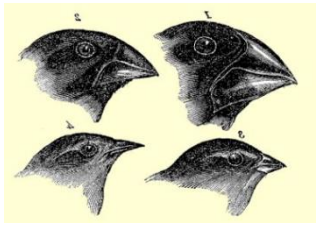
New ideas, new projects

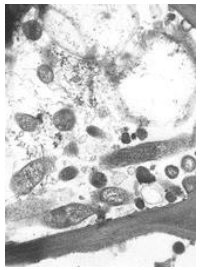
- **De novo sequencing**
  - targeted sequencing
  - a large number of small genomes
- **SNP discovery**
  - without reference genomes
- **Transcriptom study**
  - Unknown Transcriptom
- **Metagenome study**
  - Microbial genomes in nature
- **Epigentics study**
- **Regulatory element**
  - Chip-seq, RNA-seq
- **Other new projects**
  - High throughput sequence alignment









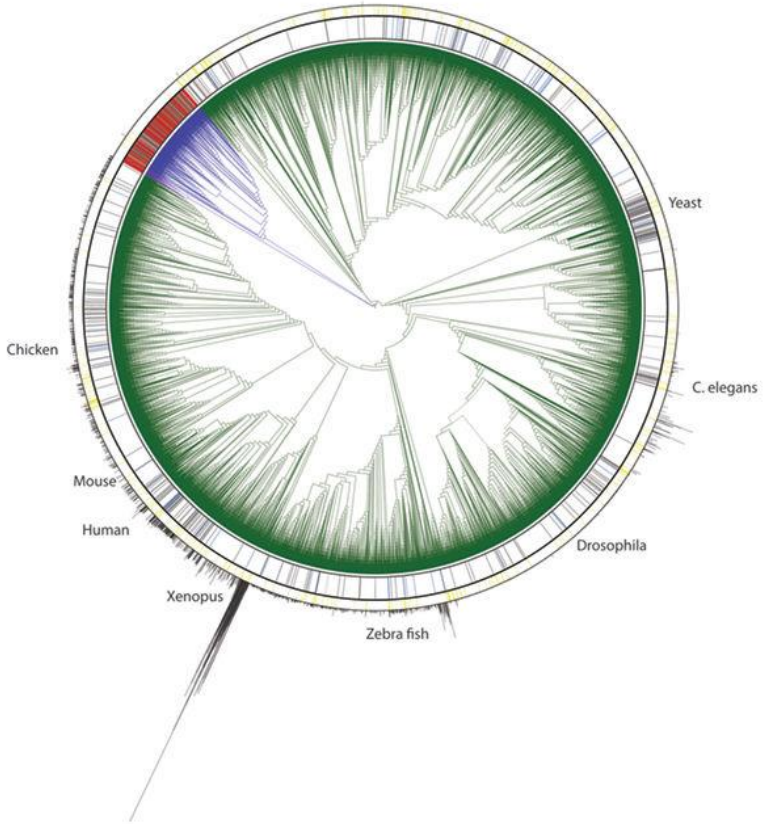



From Dawei Lin, http://bioinformatics.ucdavis.edu

Metagenomes, pan-genomes, single-cell sequencing …

- **Metagenomes**
  - **HMG**
  - **Marine metagenomes**
- **Pan-genomes**
  - **3,000 rice genomes (50TB raw reads)**
  - **1,000 genomes of *Arabdopsis thaliana***
  - **15,000 tomato genomes?**
- **Single-cell sequencing**
  - **Different tissues, development stages, conditions**
  - **…**

Sequencing the world!



- **Sequencing all Eukaryotes**

http://www.sciencemag.org/news/2017/02/biologists-propose-sequence-dna-all-life-earth

Sequencing the Earth!

- **Earth Microbiome Project**
  - **Constructing the Microbial Map for Planet Earth**
  - 200,000 samples
  - 500,000 reconstructed microbial genomes

http://www.earthmicrobiome.org/

Caner Genomes

NATIONAL CANCER INSTITUTE
THE CANCER GENOME ATLAS

TCGA BY THE NUMBERS

TCGA produced over **2.5** PETABYTES of data

To put this into perspective, **1 petabyte** of data is equal to **212,000** DVDs

TCGA data describes **33** DIFFERENT TUMOR TYPES ...including **10** RARE CANCERS

...based on paired tumor and normal tissue sets collected from **11,000** PATIENTS

...using **7** DIFFERENT DATA TYPES

TCGA RESULTS & FINDINGS

MOLECULAR BASIS OF CANCER	Improved our understanding of the genomic underpinnings of cancer	For example, a TCGA study found the basal-like subtype of breast cancer to be similar to the serous subtype of ovarian cancer on a molecular level, suggesting that despite arising from different tissues in the body, these subtypes may share a common path of development and respond to similar therapeutic strategies.
TUMOR SUBTYPES	Revolutionized how cancer is classified	TCGA revolutionized how cancer is classified by identifying tumor subtypes with distinct sets of genomic alterations.*
THERAPEUTIC TARGETS	Identified genomic characteristics of tumors that can be targeted with currently available therapies or used to help with drug development	TCGA's identification of targetable genomic alterations in lung squamous cell carcinoma led to NCI's Lung-MAP Trial, which will treat patients based on the specific genomic changes in their tumor.

THE TEAM

20 COLLABORATING INSTITUTIONS across the United States and Canada

WHAT'S NEXT?

The Genomic Data Commons (GDC) houses TCGA and other NCI-generated data sets for scientists to access from anywhere. The GDC also has many expanded capabilities that will allow researchers to answer more clinically relevant questions with increased ease.

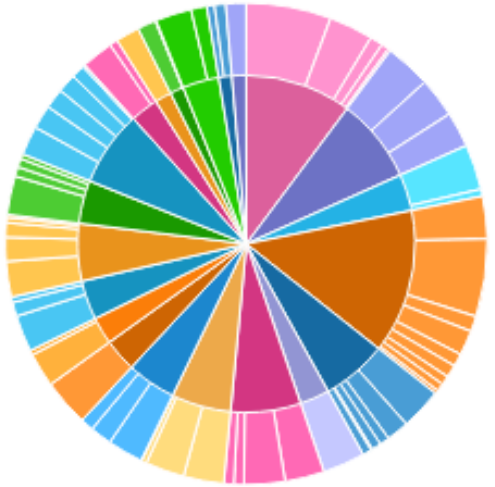
*TCGA's analysis of stomach cancer revealed that it is not a single disease, but a disease composed of four subtypes, including a new subtype characterized by infection with Epstein-Barr virus.

www.cancer.gov/ccg

Data Release 22

August 23rd, 2016

Donor Distribution by Primary Site



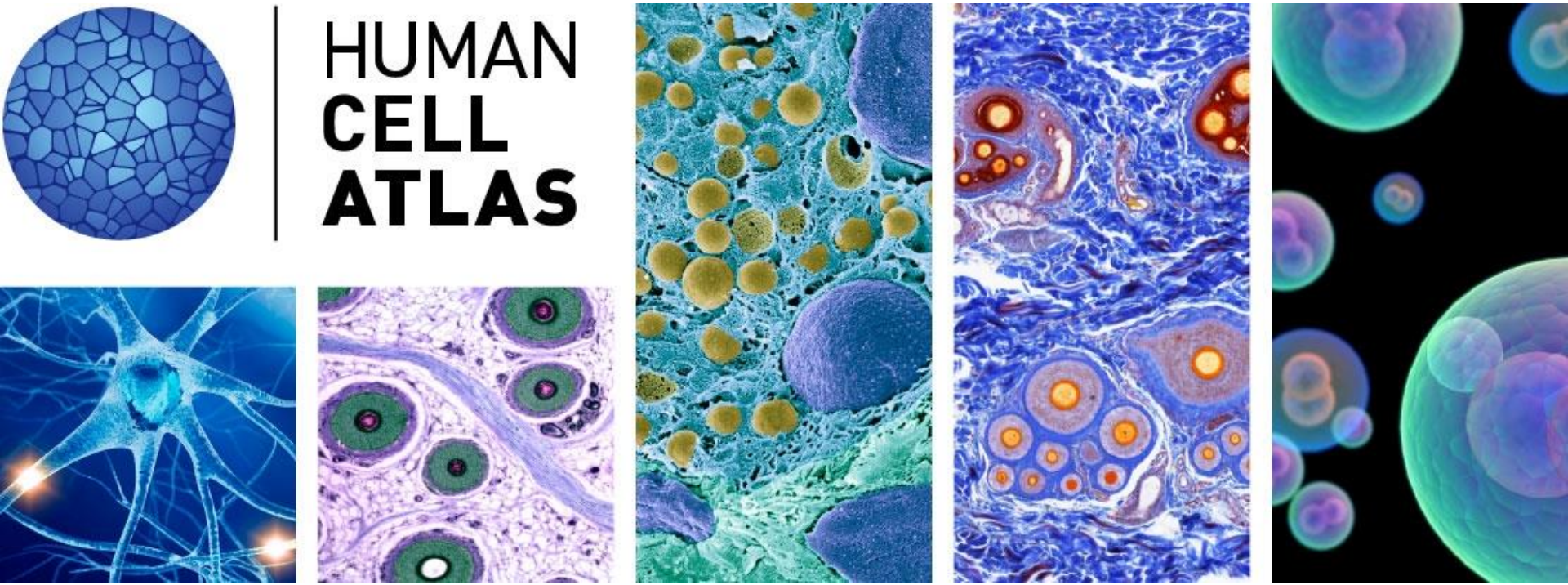
Cancer projects	70
Cancer primary sites	21
Donors with molecular data in DCC	16,236
Total Donors	19,290
Simple somatic mutations	46,429,997
Mutated Genes	57,658

TCGA：http://cancergenome.nih.gov/

ICGC：http://icgc.org/

Sequencing every cells in a human body

- 10^14 cells



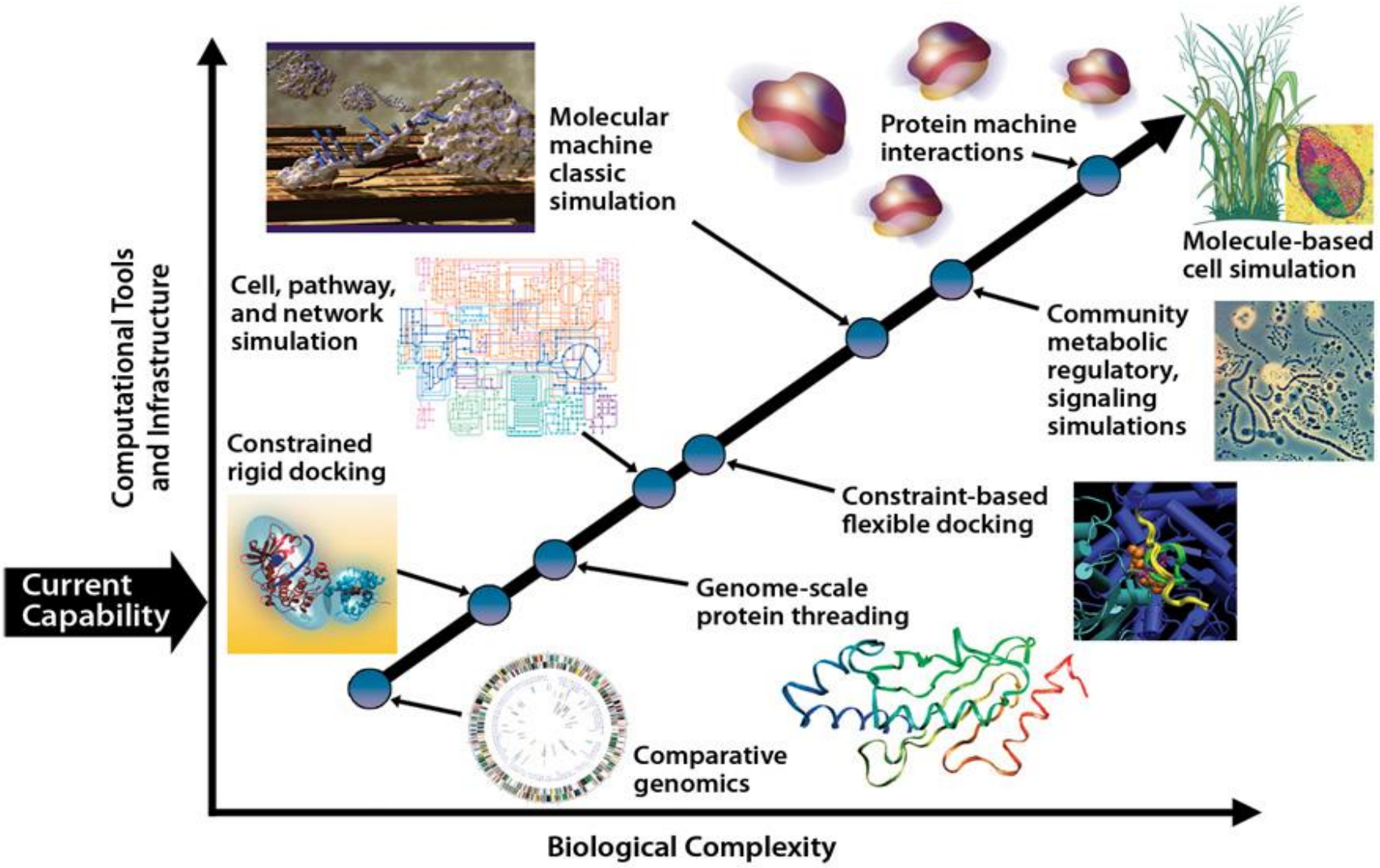
https://www.humancellatlas.org/

With so many bioinformation data,

we need algorithms!

A Big Picture of Biology

"Biology is an information science" *-- Leroy Hood*



 http://genomicscience.energy.gov/biofuels/

Goals

- General introduction about algorithms
- Basic knowledge about algorithms in Bioinformatics
- Some practice about Bioinformatics analysis

Course organization

- **Introduction (Week 1-2)**
  - Course introduction
  - A brief introduction to molecular biology
  - A brief introduction to sequence comparison
- **Part I: Algorithms for Sequence Analysis (Week 1 - 8)**
  - Chapter 1-3, Models and theories
    - » Probability theory and Statistics (Week 2)
    - » Algorithm complexity analysis (Week 3)
    - » Classic algorithms (Week 4)
  - Chapter 4. Sequence alignment (week 6)
  - Chapter 5. Hidden Markov Models (week 7）
  - Chapter 6. Multiple sequence alignment (week 8)
- **Part II: Algorithms for Network Biology (Week 9 - 16)**
  - Chapter 7. Omics landscape (week 9)
  - Chapter 8. Microarrays, Clustering and Classification (week 10)
  - Chapter 9. Computational Interpretation of Proteomics (week 11)
  - Chapter 10. Network and Pathways (week 12,13)
  - Chapter 11. Introduction to Bayesian Analysis (week 14,15)
  - Chapter 12. Bayesian networks (week 16)

Course organization (2)

- **Monday (Every Week)**
  - Lectures (东下院403）
- **Thursday(Even weeks)**
  - Lab (生物药楼4号楼-302，生信实验室)
    - BLAST（Week 2）
    - UCSC Genome Browser (Week 4)
    - RNA-seq analysis (Week 6)
    - HMM (week 8)

Course features

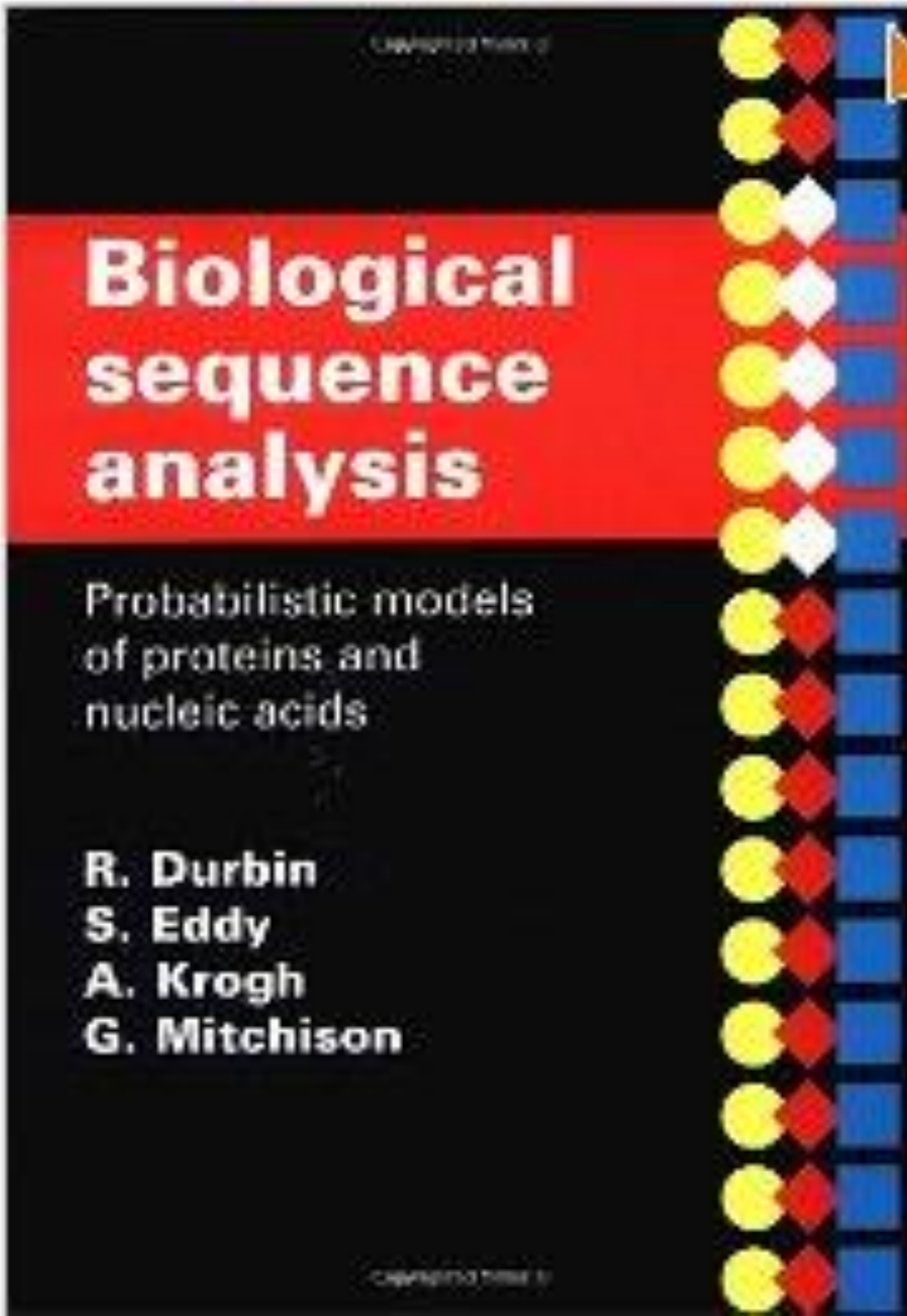
- Subjects
  - Biological sequences (Genomic, and Proteomic seqs)
  - From reads to whole genomes (peptides to proteomes)
- Topics
  - Algorithms
  - Models
  - Biology
- Theory and Practice：
  - Probability Theory
  - Complexity analysis for algorithms
  - Design and implementation of an HMM-based gene prediction system

Prerequisites

- Mathematics (a little bit)
  - Calculus
  - Probability Theory
  - Statistics
  - Advanced Algebra
- Computer Science (a little bit)
  - Programming
- Biology (a little bit)
  - Molecular Biology

Text Book

- Biological sequence analysis：Probabilistic Models for Proteins and Nucleic Acids, R. Durbin, S. Eddy, A. Krogh, G.Mitchison, Cambridge University Press, 1999



References

- 生物信息学基础，孙啸， 陆祖宏， 谢建明清华大学出版社，2004
- Introduction to Algorithms, Thomas Cormen, Charles Leiserson, and Ronald Rivest, The MIT Press.
- Unix and Perl （V.2.3.4）, K. Bradnam & I. Korf, 2009
- An Introduction to Bioinformatics Algorithms
  Neil C. Jones and Pavel A. Pevzner

  中译本： 生物信息学算法导论，【美】 N.C琼斯 P.A.帕夫纳 著 王翼飞 等译， 化学工业出版社 （生物.医药出版分社）

Grading

- Homework 30%
- Projects(1+1) 20%
- Exam 50%

作业规定

- 作业允许合作，但是必须注明各人的贡献
- 作业报告必须用自己的语言独立完成
- 期末考试需要独立完成
- 严禁抄袭
  - 抄袭者：不及格(F)
  - 被抄袭者：成绩降一级（A→B, B→C, C→D, D→F）

Similar courses in other universities

- Washington University (Algorithms for Computational Biology)
  - http://bio5495.wustl.edu/syllabus.html
- University of Washington （Computational Biology)
  - http://courses.cs.washington.edu/courses/cse527/
- Tel Aviv University School of Computer Science (Algorithms in Molecular Biology)
  - http://www.cs.tau.ac.il/~rshamir/algmb/01/algmb01.html
- Stanford (Representations and Algorithms for Computational Molecular Biology)
  - http://scpd.stanford.edu/search/publicCourseSearchDetails.do?method=load&courseId=1167871
- MIT(Foundations of Computational and Systems Biology)
  - https://ocw.mit.edu/courses/biology/7-91j-foundations-of-computational-and-systems-biology-spring-2014/

Course website

- http://cgm.sjtu.edu.cn/pub/courses/2018/pab/ab.php

- If you have any questions, send me an email at: ccwei@sjtu.edu.cn