# Algorithms in Bioinformatics

# 生物信息学算法原理

Chaochun Wei (韦朝春)

ccwei@sjtu.edu.cn

http://cbb.sjtu.edu.cn/~ccwei

Jing Li （李婧）

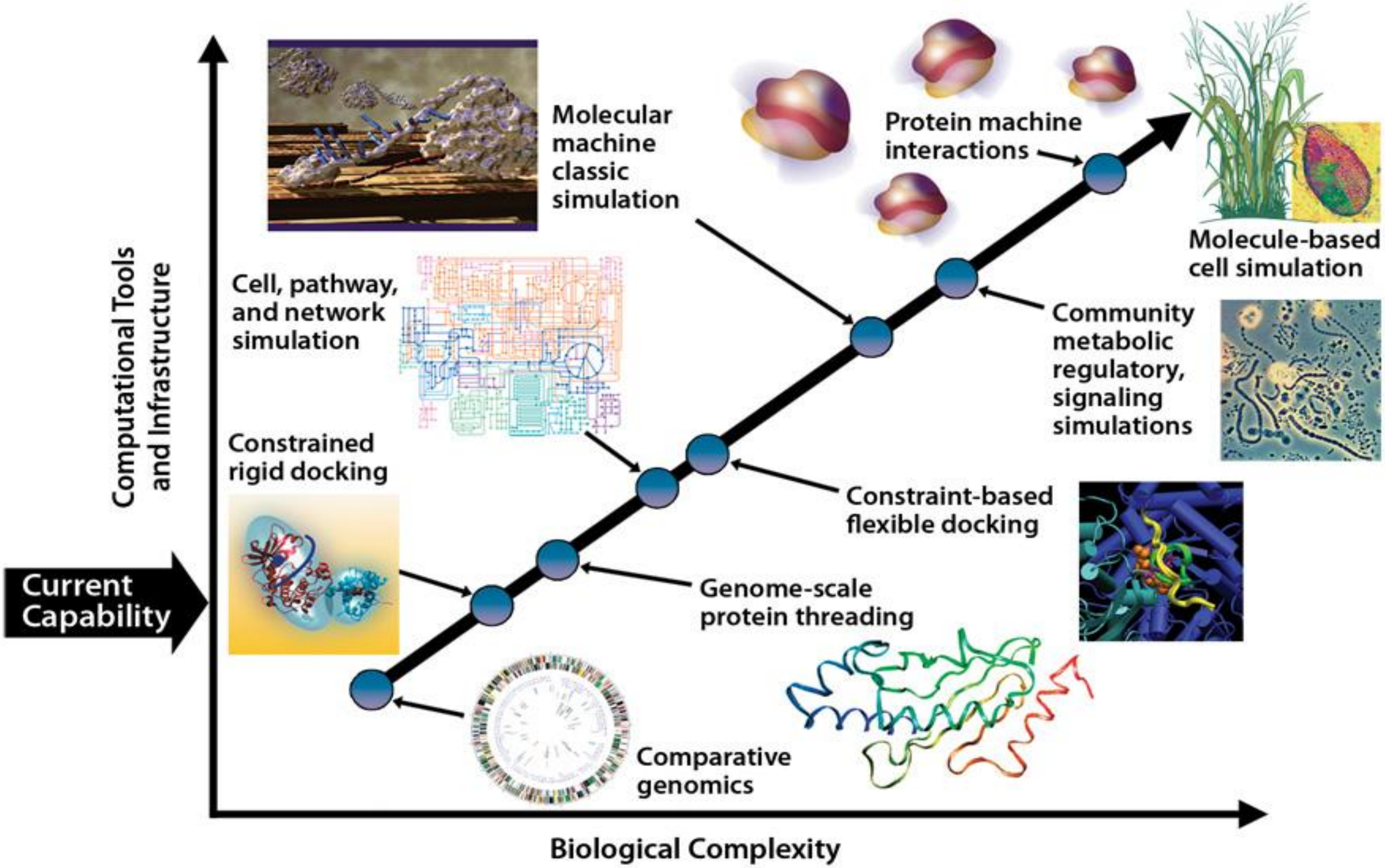
Fall 2014

# 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*



Molecular machine classic simulation
Protein machine interactions
Cell, pathway, and network simulation
Molecule-based cell simulation
Community metabolic regulatory, signaling simulations
Computational Tools and Infrastructure
Constrained rigid docking
Constraint-based flexible docking
Current Capability
Genome-scale protein threading
Comparative genomics
Biological Complexity

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

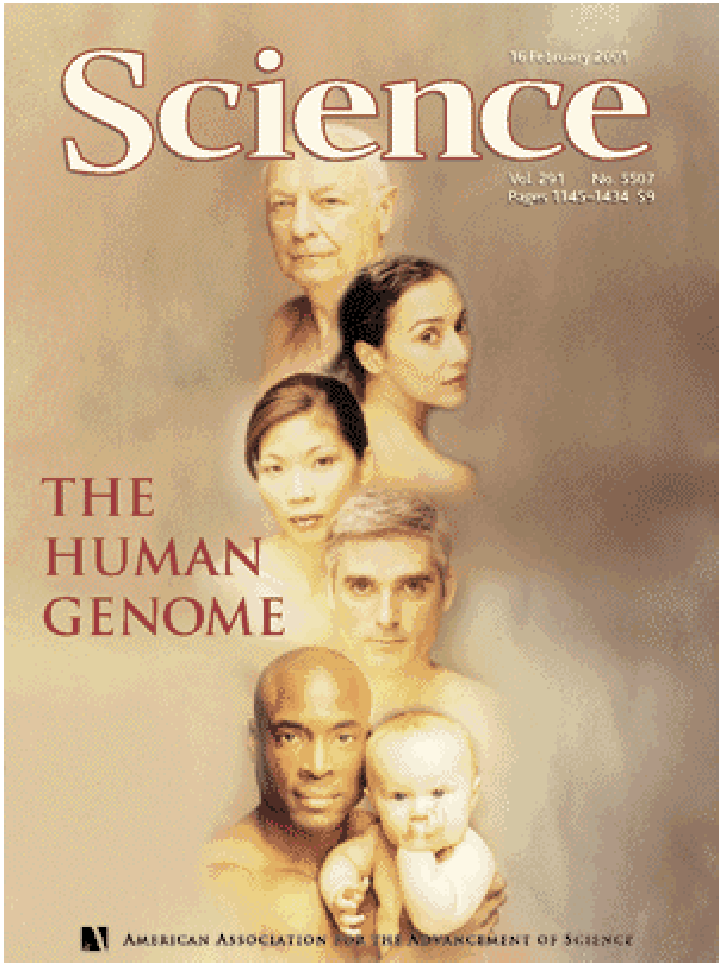
# Why do we need algorithms?

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

# Milestone of modern biology：the human genome project

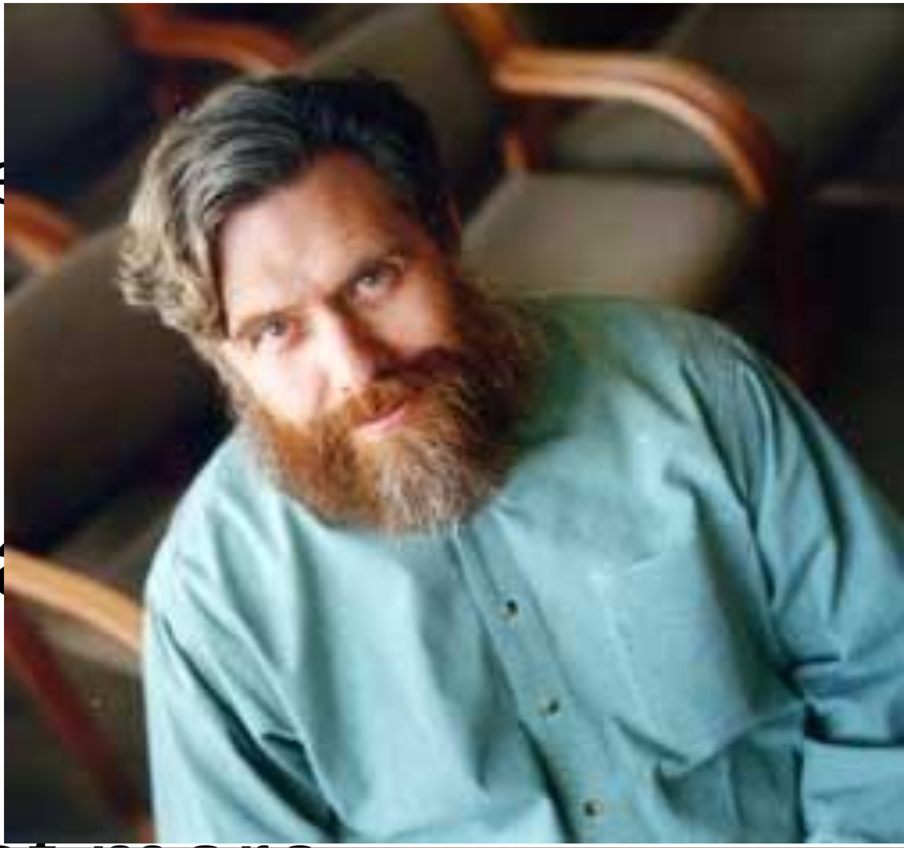
Feb. 15, 2001 *Nature*

Feb. 16, 2001 *Science*



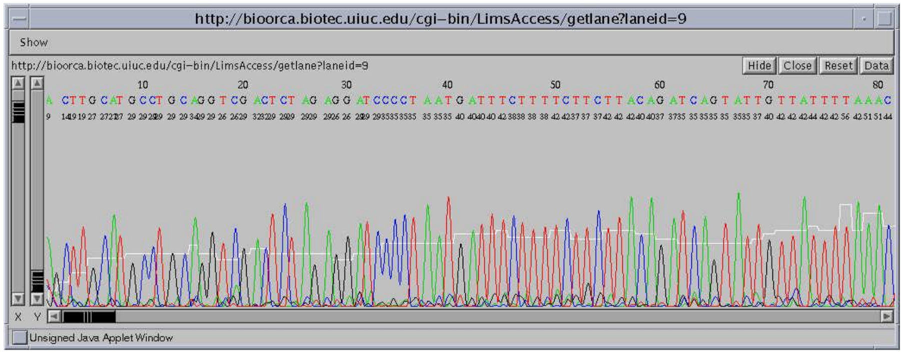
Human Genome Project
3 billion dollars，3 billion re

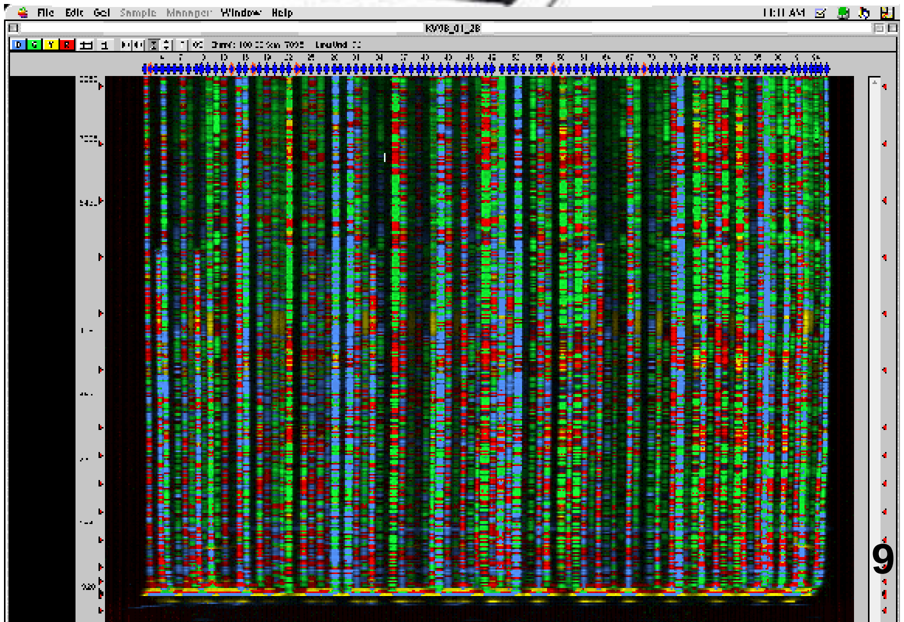
- “This is ...
  - a WASTE of ta money….
  - We can do a lo



George Church
Professor of Genetics
Harvard Medical School

# Automated high throughput sequencing


http://bioorca.biotec.uiuc.edu/cgi–bin/LimsAccess/getlane?laneid=9
Show
http://bioorca.biotec.uiuc.edu/cgi–bin/LimsAccess/getlane?laneid=9
Hide
Close
Reset
Data
10 20 30 40 50 60 70 80
A CTTGCATGCCTGCAGGTCGACTCTAGAGGATCCCCTAATGATTTCTTTTCTTCTTACAGATCAGTATTGTTATTTTAAAC
X Y
Unsigned Java Applet Window



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

# Next-generation sequencing technology

<

Sanger method:
huge lab, numerous machines and staffs

Next-generation：
one staff, one machine

**NATURE METHODS，16 | VOL.5 NO.1 | JANUARY 2008**

# 新一代测序技术平台

Applied Biosystems
ABI 3730XL

Roche / 454
Genome Sequencer
FLX

HeliScope™
Single Molecule
Sequencer

Oxford
Nanopore
MinION

Illumina / Solexa
Genetic Analyzer

Applied Biosystems
SOLiD

# Comparison of NGSs vs. traditional technology

| Platforms | Sanger | 454 | Solexa | SOLiD |
|---|---|---|---|---|
| Read Length（bps） | 650-1100 | 150-250 | 35-150 | 25-50 |
| Capacity<br>（reads/run) | 96 | 400,000 | 200,000,000 | 2,000,000,000 |
| Error Rate | 10^-3 | <10^-2 | ~10^-2 | ~10^-2 |
| Cost（$/Mbp) | 5000 | ~5 | ~0.6 | ~0.2 |
| Time/run | ~3h | ~7h | 2-14d | 3-14d |
| Throughput | 100Kb | ~1Gb | ~600Gb | 100-300Gb |

# Shanghai NGS Ally（*SGA*）

| Platform | Number | Throughput |
|---|---|---|
| 454 GS FLX | 2 | 4 Gb/day |
| Solexa GA IIx | 3 | 1Gb/day |
| HiSeq 2000 | 4 | 100 Gb/day |
| Solid 4 | 7 | 7 Gb/day |
| HQ | （7） | （128 Gb/day） |
| 合计 | 16 | >110 （220）Gb/day |

SBC
上海生物芯片有限公司
生物芯片上海国家工程研究中心

CHGC
国家人类基因组南方研究中心
National Human Genome Center at Shanghai

Scbit
上海生物信息技术研究中心
SHANGHAI CENTER FOR BIOINFORMATION TECHNOLOGY

# Latest sequencing technologies

- Pacific Biosciences
  - Human genome： $100, 15 minutes （2013）
- Complete Genomics
  - 10,000 genomes /year (from 2010)
- Ion Torrent
- Oxford Nanopore (2012)
- Visigen
- more…

# 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
- ……….

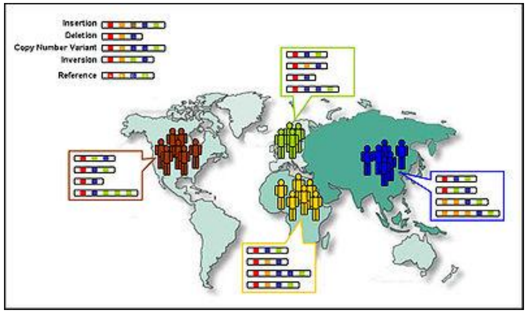
From Dawei Lin, http://bioinformatics.ucdavis.edu

# Latest personal genomics projects

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

International Cancer Genome Consortium

ClinSeq®



23andMe

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

# 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

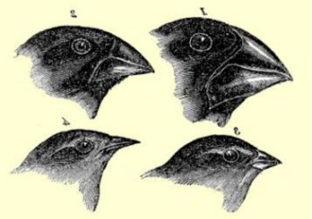
- Epigentics study

- Regulatory element

  Chip-seq, RNA-seq

- Other new projects

  High throughput sequence alignment



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

With so many bioinformation data, we need algorithms!

# Goals

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

# Course organization

- ➢Introduction ( Week 1)
- ➢ Part I: Algorithms for Sequence Analysis (Week 1 - 11)
  - – Chapter 1-3, Models and theories
    - » Probability theory and Statistics (Week 2)
    - » Algorithm complexity analysis (Week 3)
    - » Classic algorithms (Week 4)
    - » Lab: Linux and Perl
  - – Chapter 4, Sequence alignment (week 6)
  - – Chapter 5, Hidden Markov Models ( week 8)
  - – Chapter 6. Multiple sequence alignment (week 10)
  - – Chapter 7. Motif finding (week 11)
  - – Chapter 8. Sequence binning (week 11)
- ➢ Part II: Algorithms for Network Biology (Week 12 - 16)

# Course organization (2)

- Friday (Every Week)
  - Lectures (东中院2-403）

- Wednesday(Even weeks)
  - Lab (生物药楼4号楼-302，生信实验室)
    - Unix and Perl （Week 2, 4, 6）
    - HMM  (week 8, 10)

# Course features

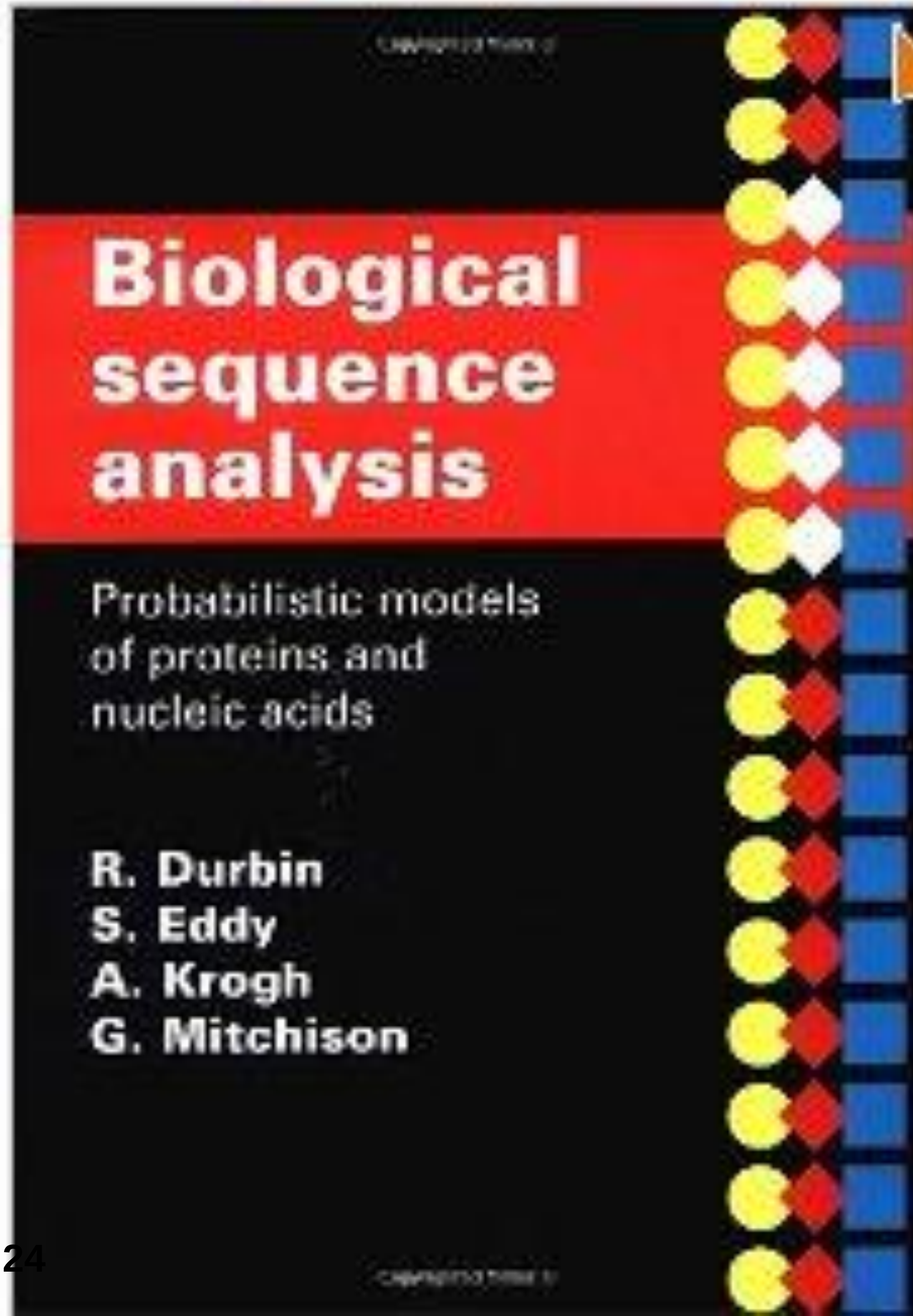
- Subjects
  - Biological sequences ( Genomic, and Proteomic seqs)
  - From reads to whole genomes (peptides to proteomes)
- Topics
  - Algorithms
  - Models
  - Biology
- Theory and Practice：
  - Probability Theory
  - Complexity analysis for algorithms
  - Design and implementation of an HMM-based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



Biological sequence analysis

Probabilistic models of proteins and nucleic acids

R. Durbin
S. Eddy
A. Krogh
G. Mitchison

# 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://www.cs.washington.edu/education/courses/527/
- 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://www-helix.stanford.edu/courses/bmi214/
- MIT(Foundations of Computational and Systems Biology)
  - http://www.core.org.cn/OcwWeb/Biology/7-91JSpring2004/LectureNotes/index.htm
- Imperial College (Introduction to Bioinformatics)
  - http://www.doc.ic.ac.uk/~sgc/teaching/341/

# Course website

- http://cbb.sjtu.edu.cn/~ccwei/pub/courses/2014/algorithms_in_bioinformatics/ab.php
- If you have any questions, send me an email at: ccwei@sjtu.edu.cn