Bioinf

Introduction to Bioinformatics

纪志梁，PhD
生物II馆114
Email: appo@bioinf.xmu.edu.cn

Bioinf The Definition

- *Computer-assisted data management discipline that helps us gather, analyze, and represent biological information in order to understand life's processes**
- Analysis of biological data with computing & statistical tools.
- A science combining multiple disciplines such as Biology, Informatics, Math, Chemistry, Physics,…

*Persidis A. Bioinformatics. *Nat Biotechnol* 1999 Aug; 17(8): 828-830

Bioinf The definition cont…

- Bioinformatics is conceptualizing **biology in terms of molecules** (in the sense of physical-chemistry) and then applying **“informatics” techniques** (derived from disciplines such as applied math, CS, and statistics) to **understand and organize the information associated** with these molecules, **on a large-scale.**

Brief History

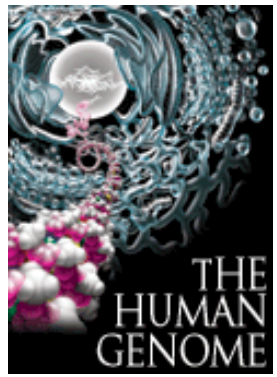
- 1956年，美国田纳西州召开的“生物学中的信息理论讨论会”
- 1980s，重新发展（得益于IT的发展）
- 1980s年代末，Dr林华安提出了结合生物及信息学的新词“Bioinformatics”
- 1990s，蓬勃发展至今
- 将来，空间无限

The Need for Bioinformatics

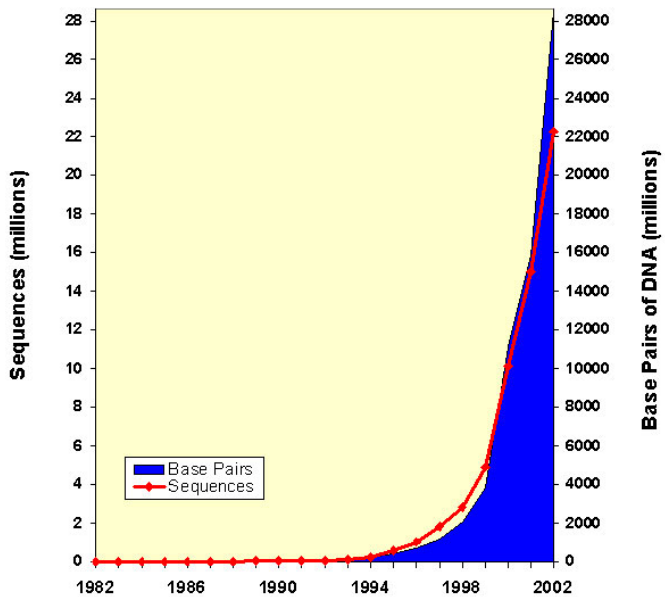
- Reveal essence of life
  - Genetic decoding, evolution (进化) understanding, molecule characterization…
- Support the development of related science
  - Providing useful information, clue, and strategy of solution…
- Possess huge market value
  - CADD，medical management…

The Need for Bioinformatics cont…

- Molecular biology / Cellular Biology
- (Functional) Genomics / Proteomics
- Systems biology
- Protein design and engineering
- Pharmaceutical development / Drug Design
- Medicine
- Ecology / population genetics
- Evolution



Growth of GenBank



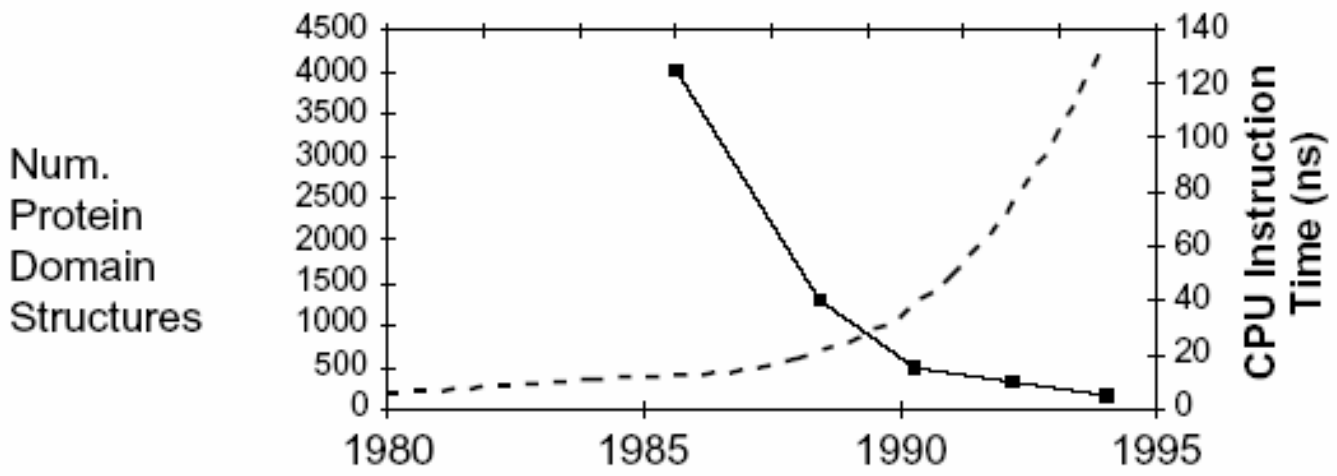
Driving Force

- **CPU vs Disk & Net**

  As important as the increase in computer speed has been, the ability to store large amounts of information on computers is even more crucial

- **New Biotechnologies**

  The exponential increase in biological data



Picture adapted from D Brutlag, Stanford

Bioinf True & False

- Digital Libraries
  - Automated Bibliographic Search and Textual Comparison
  - Knowledge bases for biological literature

True

- Pattern Determination
  - Motifs/domains identification
  - EST identification

True

Bioinf True & False

- Methods for Structure Determination
  - Computational Crystallography
    - Refinement

True / False?

- NMR Structure Determination
  - Distance Geometry

True

- Biochip 生物芯片
  - Gene chip — **True**
  - Protein chip — **True**
  - DNA computer — **False**

Bioinf True & False

- Genomic Sequencing Methods
  - Sequencing — **False**
  - Assembling — **True**
  - Annotation — **True**
  - Genomic mapping — **True**
- DNA/RNA/Protein structure prediction — **True**
- Gene identification by algorithms — **True**

Bioinf True & False

- Phylogenic study
  - Based on mutation of molecules (DNA/RNA/Protein) — **True**
  - Based on whole genomes — **True**
  - Based on non-molecular organisms — **False**
- Radiological image processing — **False**
- Artificial life simulation — **True**
- Metabolic pathway simulation — **True**

Bioinf True & False

- Drug Design
  - Lead identification — **False**
  - Lead optimization — **False**
  - Lead synthesis — **False**
  - Target identification — **True**
  - Ligand-Receptor docking — **True**
  - High throughput screen based on ligand — **False**
  - High throughput screen based on protein — **True**

Biology Aspects of Bioinformatics

- **Macromolecules**
  - DNA/RNA, protein, complex
  - Characteristics, annotation, structure, function
- **Physiological Process**
  - Cellular process, molecular machine,...
  - Biological pathway, signal transduction,...
- **Whole genomes**
  - Assembling, annotation, mapping,...
  - Comparison, evolution,...
  - Genome, proteome, interactome, polymorphism,...
- **More...**

Informatics Aspects of Bioinformatics

- **Databases**
  - Building, Querying
  - Object DB, DBMS
- **Text String Comparison**
  - Text Search
  - 1D Alignment
  - Significance Statistics
  - Alta Vista, grep
- **Patterns Search**
  - AI / Machine Learning
  - Clustering, HMM
- **Geometry**
  - Robotics
  - Graphics (Surfaces, Volumes)
  - Comparison and 3D Matching (Visision, recognition, docking)
- **Physical Simulation**
  - Newtonian Mechanics
  - Electrostatics
  - Numerical Algorithms
  - Simulation
  - Motion, Energy

Biology Vs. Computer Science

Bioinf

Biologists

- Nothing COMPLETELY True or False
- Understand complex and messy NATURAL world
- Data driven
- Experimentally-focused
- Comfortable to data with error
- Mainly science driven

Computer Scientists

- Everything is True or False
- Build a clean and organized virtual world
- Algorithm driven
- Problem-solving
- Can NOT stand error
- Engineering/marketing driven

The Change of Biologists

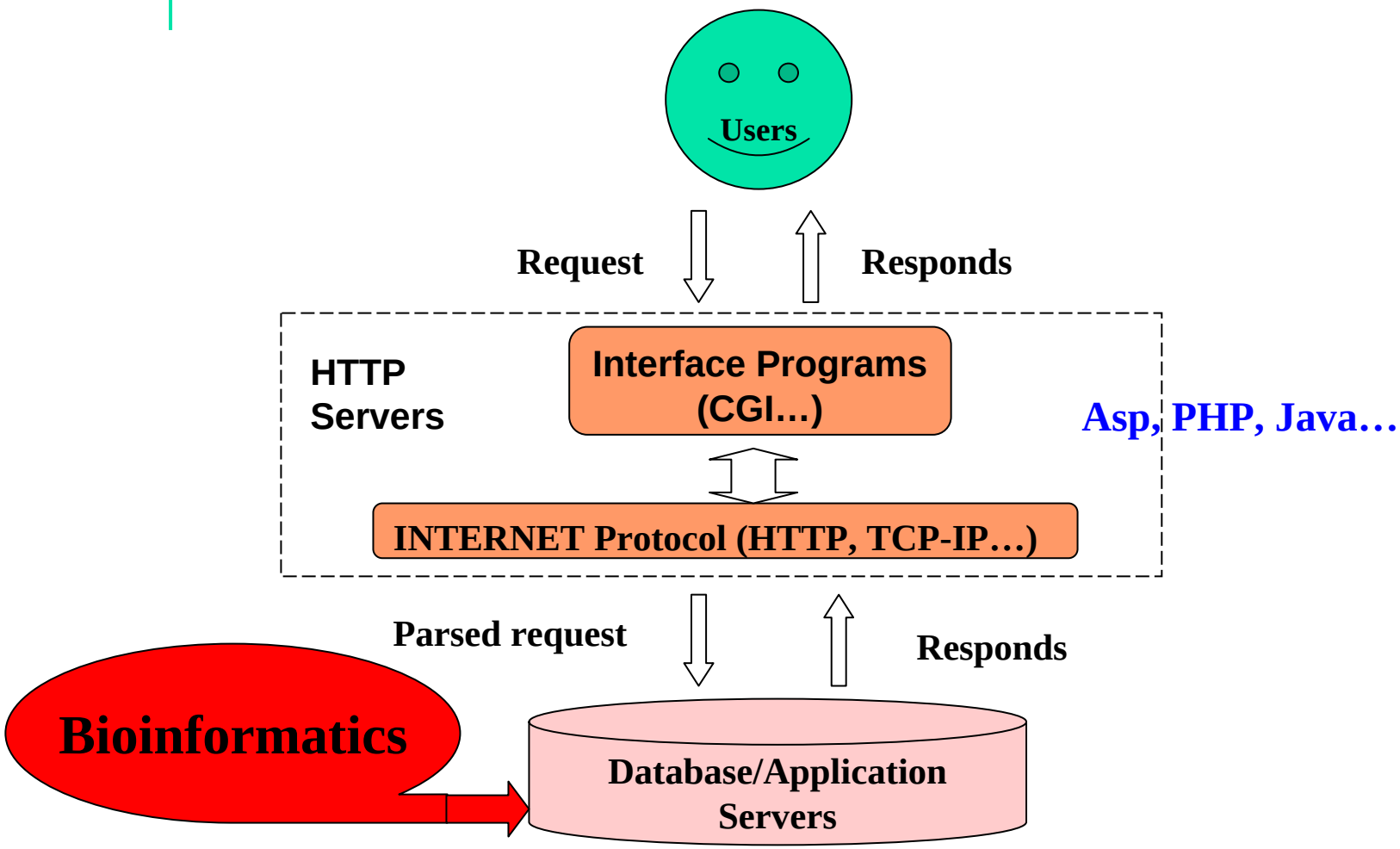
- Spend more time using computers
- spend more time on data analysis
- become more quantitative thinking
- Seek clue using computer before experiment

Bioinf To be Bioinformatician

Good Bioinformatics

- A bridge of Biological science and informatics' science
- Excellent communication and understanding between biologists and computer scientists is the key

Bioinformatics and The Internet



Basic Internet Knowledge

- Remote communication began in early 1960s, first internet ARPANET was created by ARPA, USA in 1969.
- In 1981, BINET was introduced for point-to-point communication.
- In 1982, Transmission Control Protocol (TCP) and the Internet Protocol (IP) was introduced to allow different networks to be connected to and communicate with one another until today.

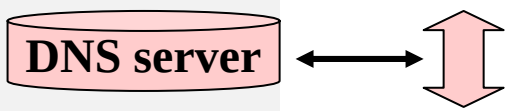
IP & Domain Names

TABLE 1.1. Top-Level Doman Names

TOP-LEVEL DOMAIN NAMES	
.com	Commercial site
.edu	Educational site
.gov	Government site
.mil	Military site
.net	Gateway or network host
.org	Private (usually not-for-profit) organizations
EXAMPLES OF TOP-LEVEL DOMAIN NAMES USED OUTSIDE THE UNITED STATES	
.ca	Canadian site
.ac.uk	Academic site in the United Kingdom
.co.uk	Commercial site in the United Kingdom
GENERIC TOP-LEVEL DOMAINS PROPOSED BY IAHC	
.firm	Firms or businesses
.shop	Businesses offering goods to purchase (stores)
.web	Entities emphasizing activities relating to the World Wide Web
.arts	Cultural and entertainment organizations
.rec	Recreational organizations
.info	Information sources
.nom	Personal names (e.g., *yourlastname.nom*)

A complete listing of domain suffixes, including country codes, can be found at http://www.currents.net/resources/directory/noframes/nf.domains.html.

IP address
130.14.25.1



Domain name
ncbi.nlm.nih.gov

Performance of Internet

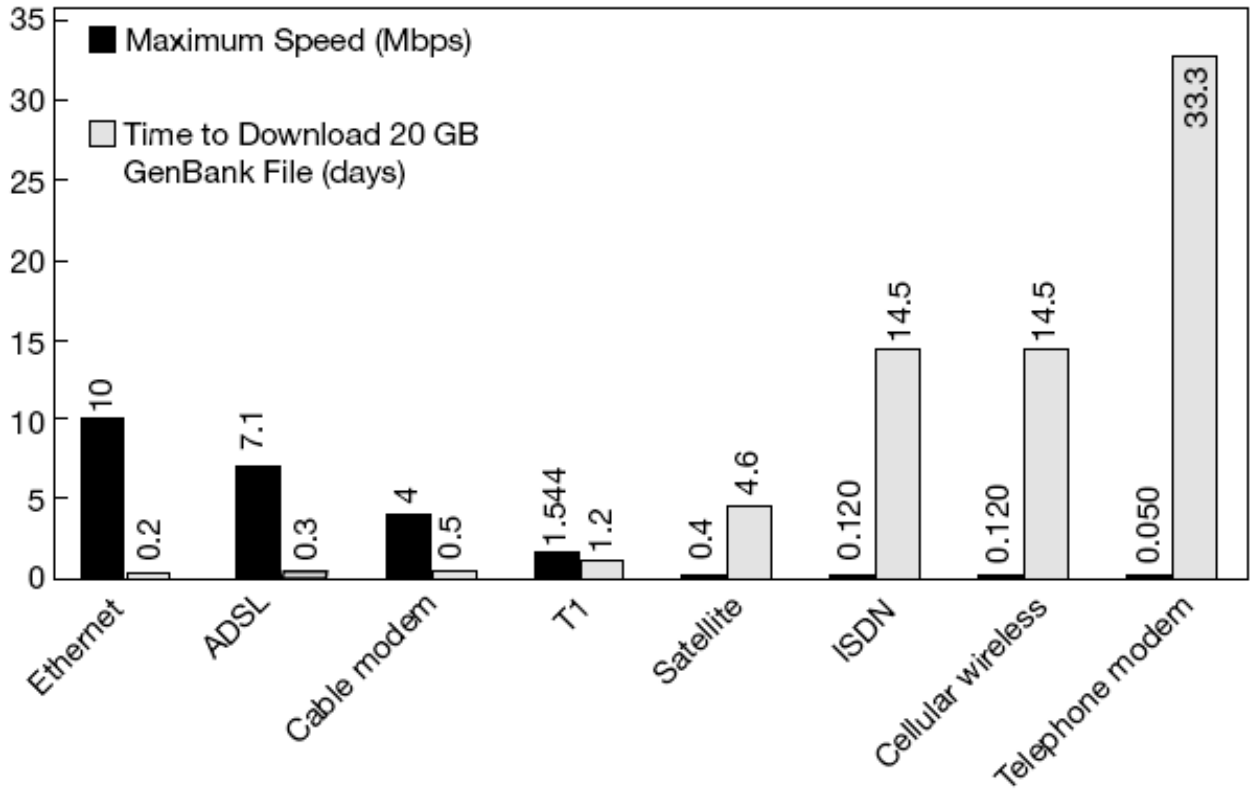


Figure 1.1. Performance of various types of Internet connections, by maximum throughput. The numbers indicated in the graph refer to peak performance; often times, the actual performance of any given method may be on the order of one-half slower, depending on configurations and system conditions.

Some Terminology

- World Wide Web (WWW)
  - Hyper Text Transfer Protocol (HTTP)
  - Hypertext Markup Language (HTML)
- File Transfer Protocol (FTP)
- Uniform Resource Locator (URL)

General form	*protocol://computer.domain*
FTP site	*ftp://ftp.ncbi.nlm.nih.gov*
Gopher site	*gopher://gopher.iubio.indiana.edu*
Web site	*http://www.nhgri.nih.gov*

Bioinf

Next Topic

From Gene to Protein

- Genetic code & ORF
- Physic-chemical analysis of macromolecules