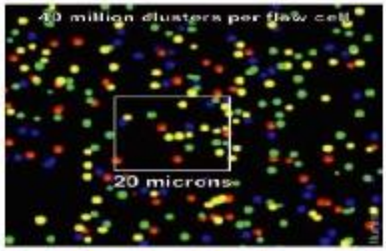
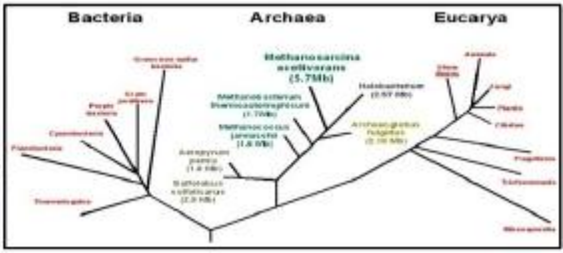
Bioinformatics: Introduction and Methods

Le Zhang

Computer Science Department, Southwest University

```
TAACCCTAACCCTAACCCTAACCCTAACCCTA
CCTAACCCTAACCCTAACCCTAACCCTAACCC
CCCTAACCCCTAACCCTAACCCTAACCCTAAC
AACCCTAACCCTAACCCTAACCCCTAACCCTA
ACCCTAACCCCAACCCCAACCCCAACCCCAAC
CTACCCTAACCCTAACCCTAACCCTAACCCTA
ACCCTAACCCTAACCCTAACCCCTAACCCTAA
```



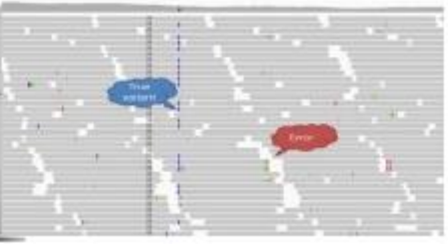


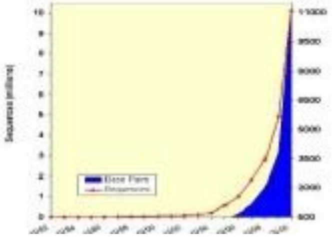
Introduction and History

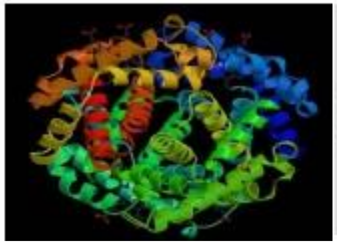
Le Zhang, Ph. D.

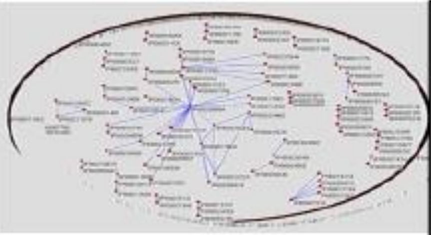
Computer Science Department

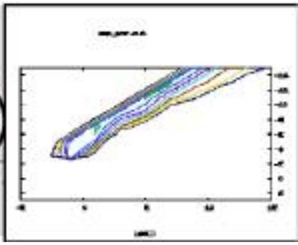
Southwest University



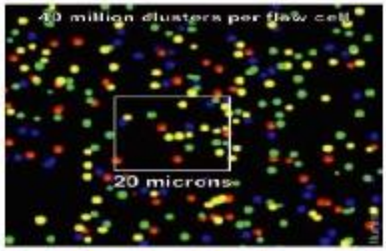


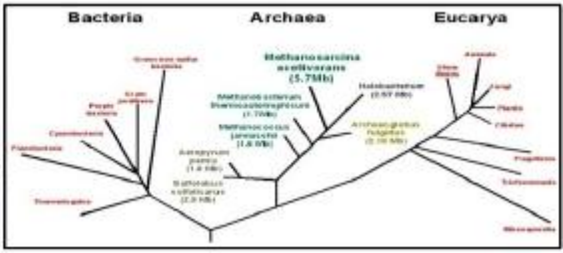






```
TAACCCTAACCCTAACCCTAACCCTAACCCTA
CCTAACCCTAACCCTAACCCTAACCCTAACCC
CCCTAACCCCTAACCCTAACCCTAACCCTAAC
AACCCTAACCCTAACCCTAACCCCTAACCCTA
ACCCTAACCCCAACCCCAACCCCAACCCCAAC
CTACCCTAACCCTAACCCTAACCCTAACCCTA
ACCCTAACCCTAACCCTAACCCCTAACCCTAA
```



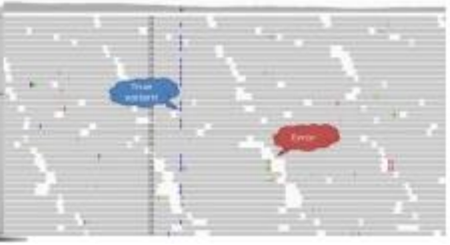


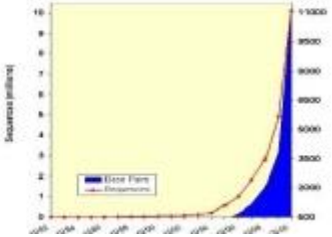
Unit 1: What is bioinformatics?

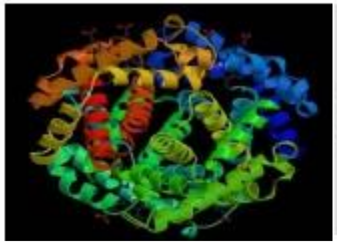
Le Zhang, Ph. D.

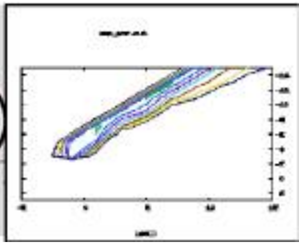
Computer Science Department

Southwest University

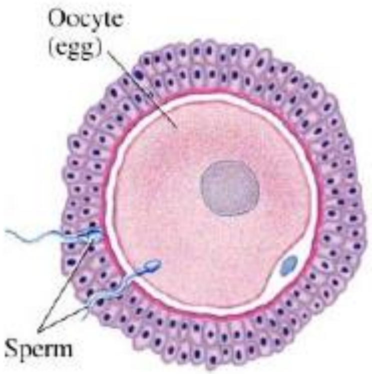






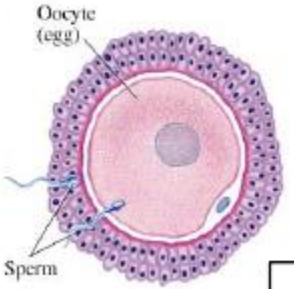


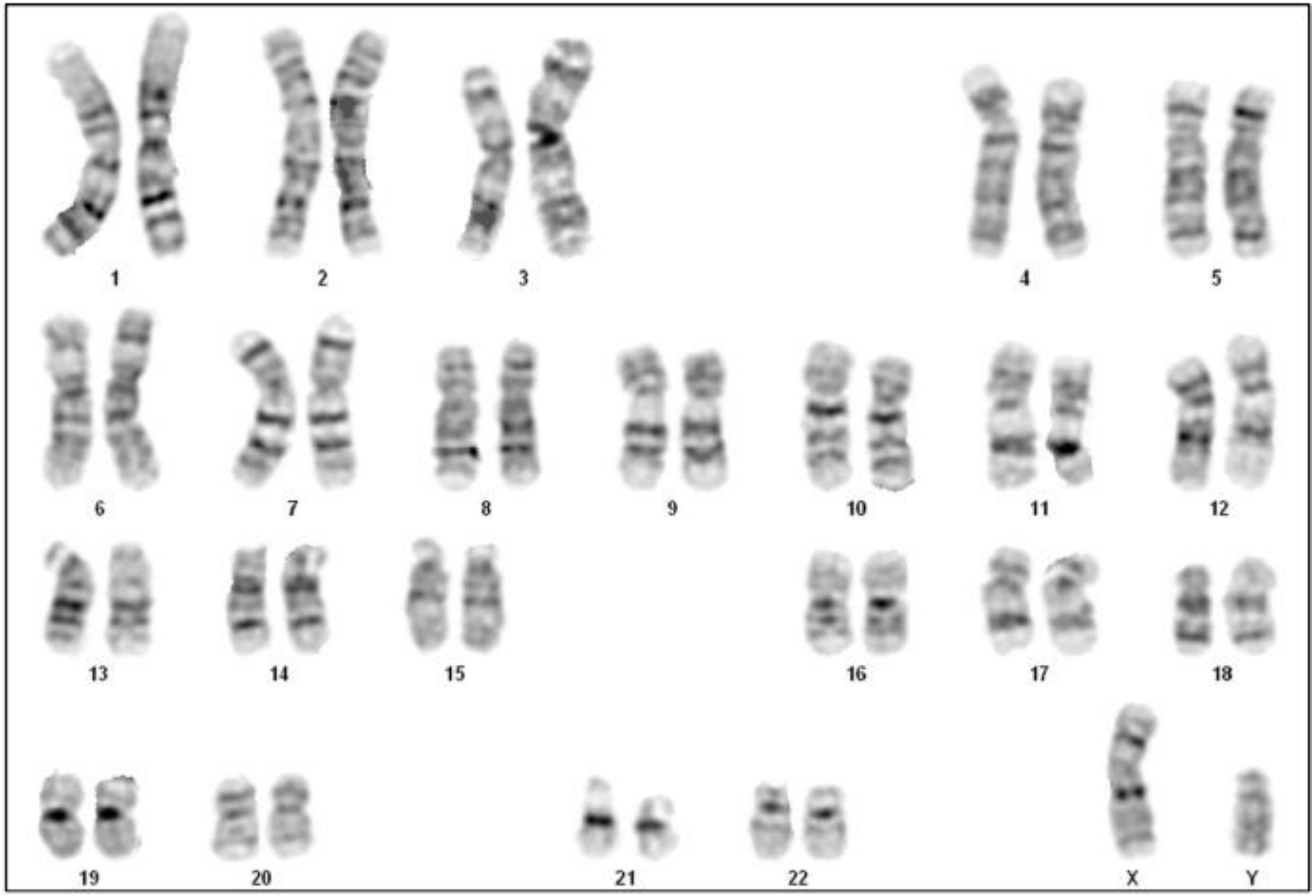
The miracle of life



How does the oocyte become the baby and how does the baby become a girl

Genome: the “manual of life” ***-- Almost***





mitochondrial DNA

epigenetics

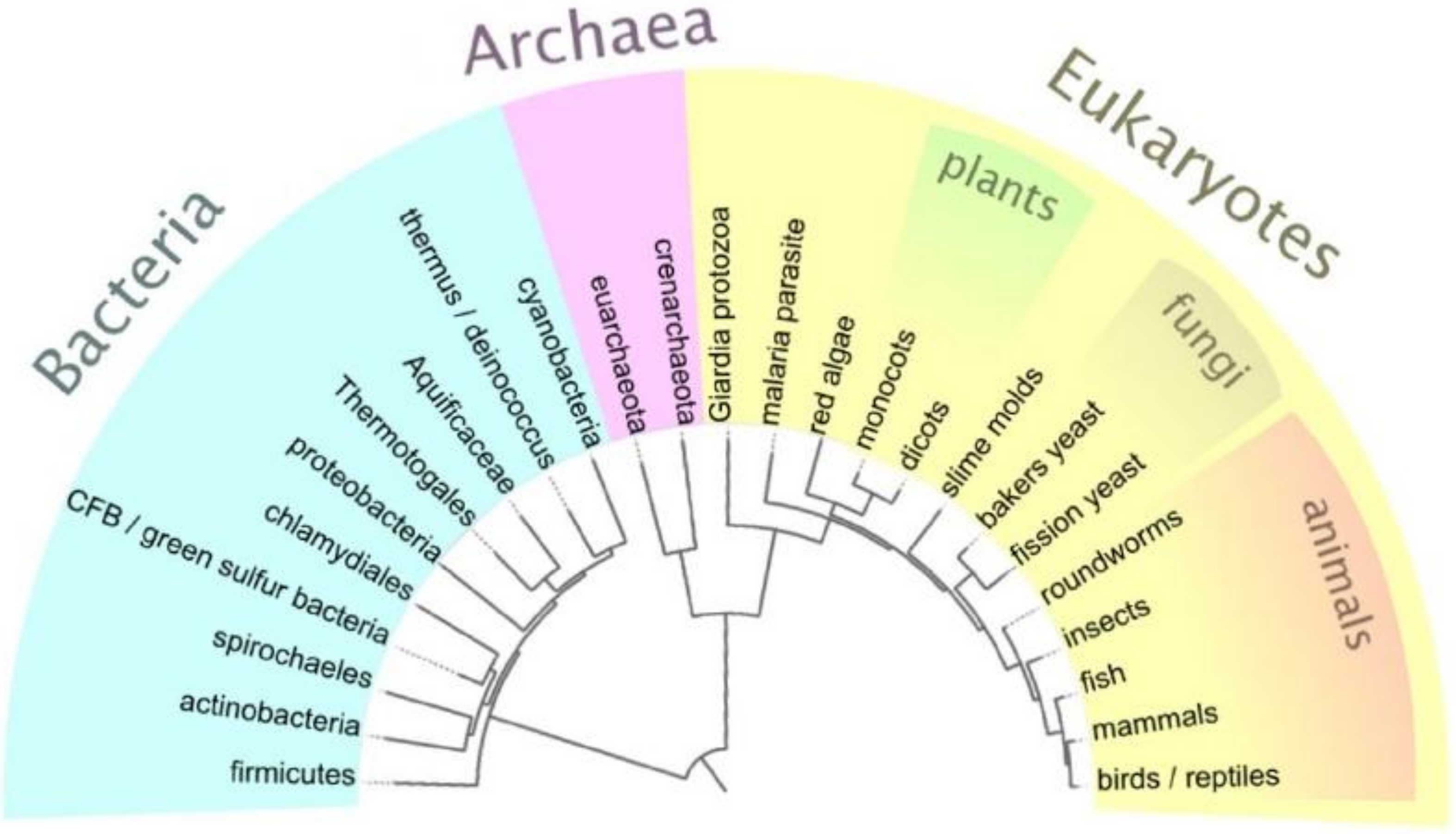
environments/nurture

chance

```
GCAGCACGCCCACCTGCTGGCAGCTGGGGACACTGCCGGGCCCTCTTGCTCCAACA
GTACTGGCGGATAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAA
CCCTAACCCTAACCCTAACCCTCCTAACCCTAACCCTAACCCTAACCCTAACCCTA
ACCCAACCCTAACCCTAACCCTAACCCTAACCCCCCTAACCCCTAACCCTAACCCT
AACCCTAACCCTAACCTAACCCTAACCCTAACCCTAACCCTAACAACCCTAACCCT
AACCCTAACCCCTAACCCTAACCCTAAACCCTAAACCCTAACCCTAACCCTAACC
```

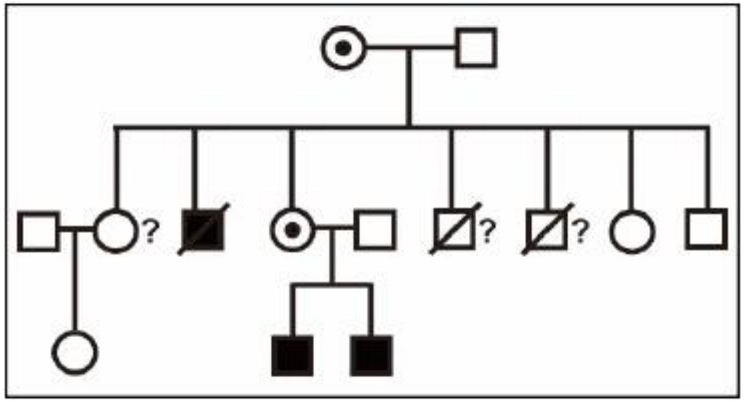
- Life is so simple but so mysterious
- Human genome has 3.1 billion bases
- 97% of the genome were called junks
- ~2.9% of the bases encode genes
- They contain the regulatory elements that encode instruction on when, where, which, and how much proteins to make

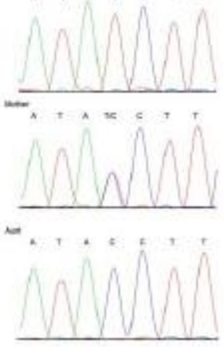
The universal code: Other species' genomes
The Tree of Life



http://commons.wikimedia.org/wiki/File:Simplified_tree.png

Human Genetic Variations





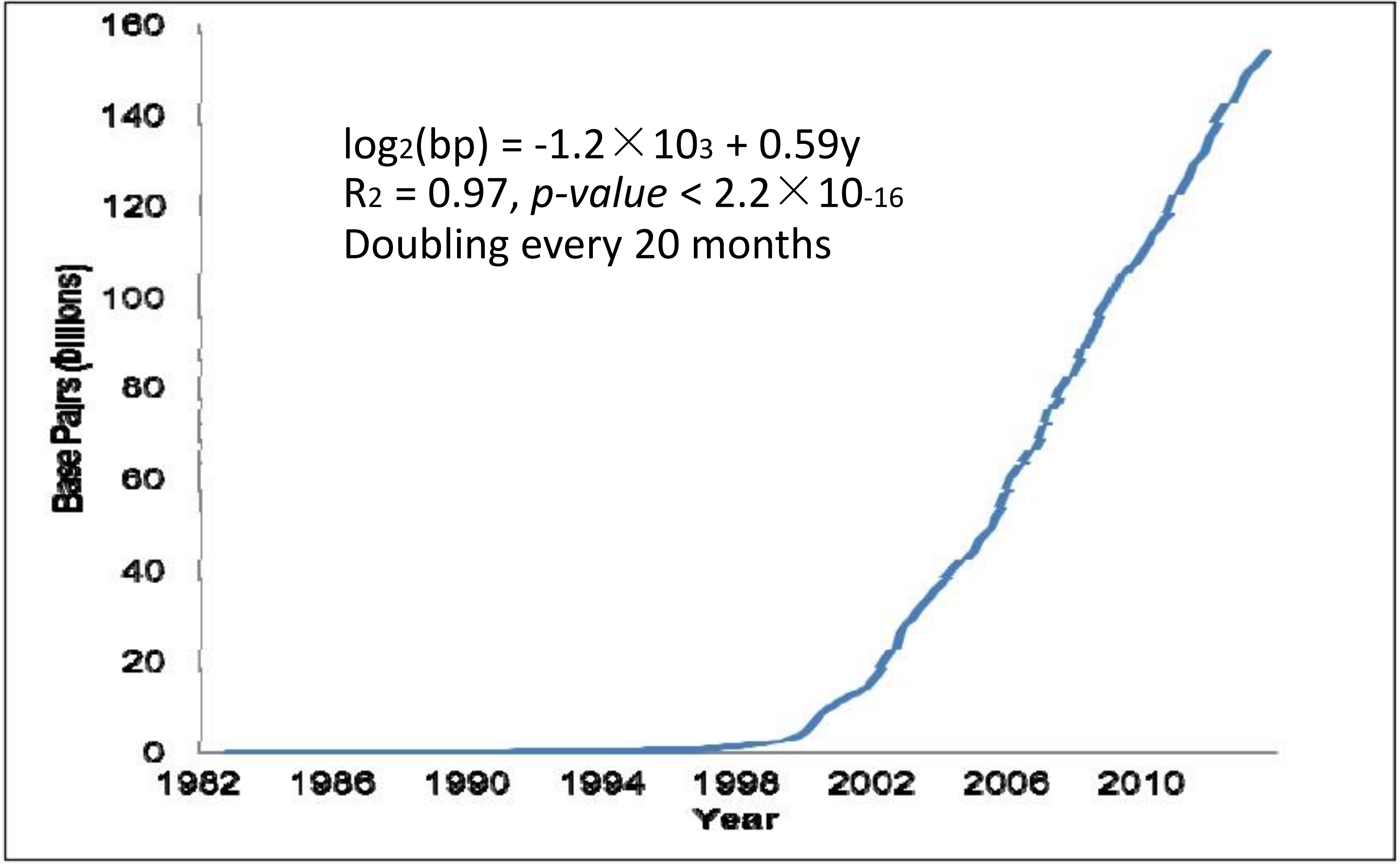
```
GCAGCACGCCCACCTGCTGGCAGCTGGGGACACTGCCGGGCCCTCTTGCTCCAACA
GTACTGGCGGATAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAA
CCCTAACCCTAACCCTAACCCTCCTAACCCTAACCCTAACCCTAACCCTAACCCTA
ACCCAACCCTAACCCTAACCCTAACCCTAACCCCCCTAACCCCTAACCCTAACCCT
AACCCTAACCCTAACCTAACCCTAACCCTAACCCTAACCCTAACAACCCTAACCCT
AACCCTAACCCCTAACCCTAACCCTAAACCCTAAACCCTAACCCTAACCCTAACC
```

- How do you **decode** the instructions in this “manual of life”?

- If you print 100 character per line and 50 lines per page, it will have 600000 pages

- The biological data is a big data

Genbank growth



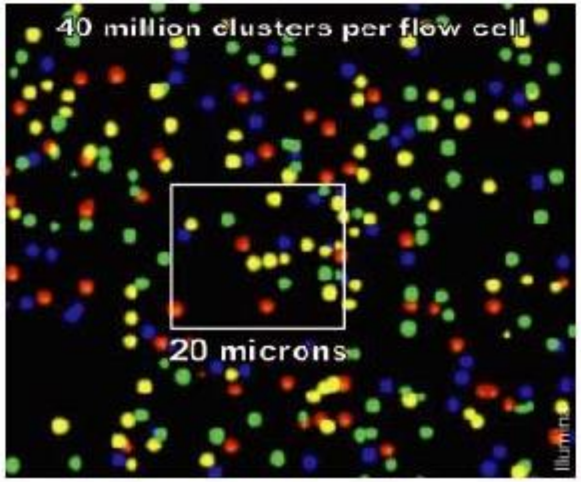
$\log_2(bp) = -1.2 \times 10_3 + 0.59y$

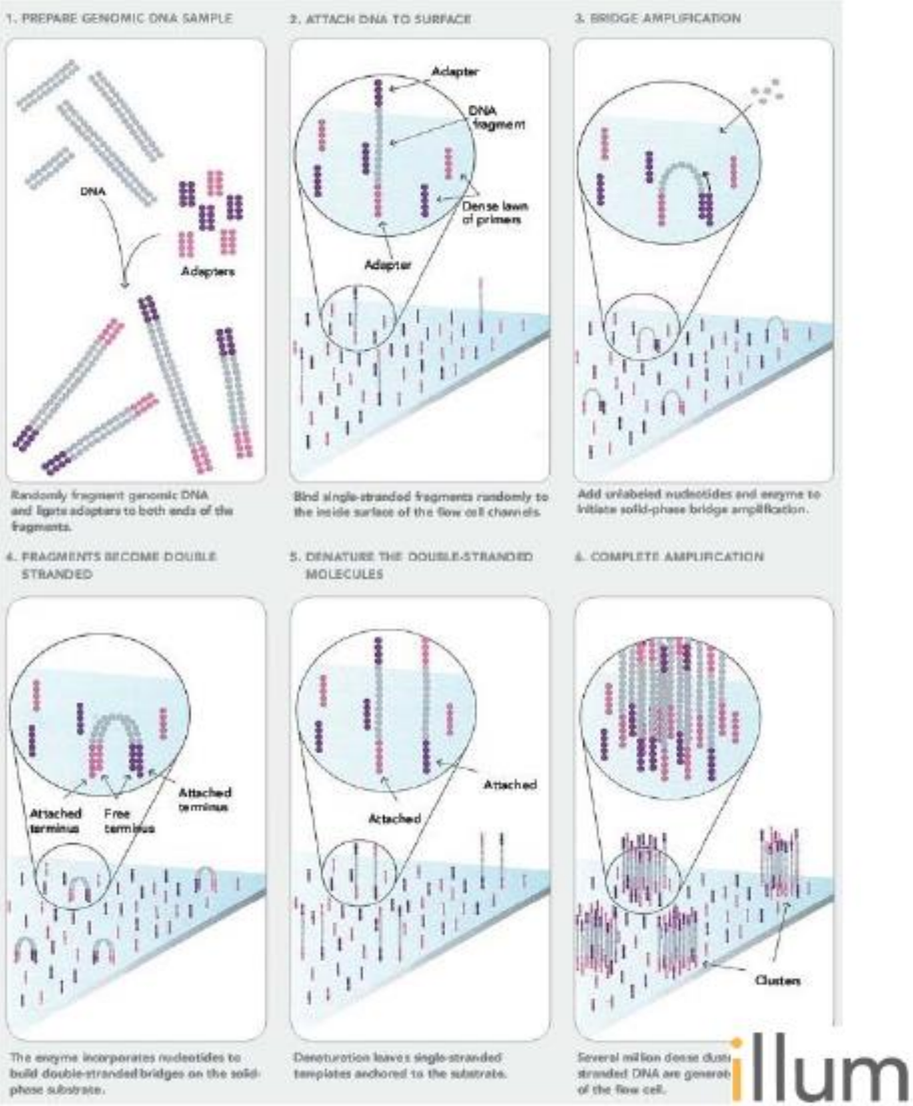
$R_2 = 0.97$, *p-value* $< 2.2 \times 10_{-16}$

Doubling every 20 months

Data Source: ftp://ftp.ncbi.nih.gov/genbank/gbrel.txt

Next-Generation Sequencing: Your genome, one day, $3000!




1. PREPARE GENOMIC DNA SAMPLE

Randomly fragment genomic DNA and ligate adapters to both ends of the fragments.

2. ATTACH DNA TO SURFACE

Bind single-stranded fragments randomly to the inside surface of the flow cell channels.

3. BRIDGE AMPLIFICATION

Add unlabeled nucleotides and enzyme to initiate solid-phase bridge amplification.

4. FRAGMENTS BECOME DOUBLE STRANDED

The enzyme incorporates nucleotides to build double-stranded bridges on the solid-phase substrate.

5. DENATURE THE DOUBLE-STRANDED MOLECULES

Denaturation leaves single-stranded templates anchored to the substrate.

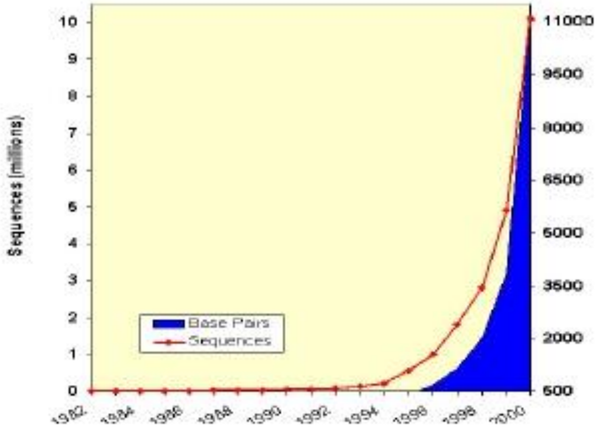
6. COMPLETE AMPLIFICATION

Several million dense clust stranded DNA are generat of the flow cell.

illumina

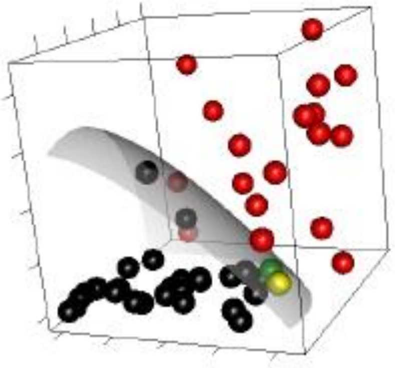
Opportunities and challenges hand-in-hand: the driving forces of bioinformatics

- High-throughput data
  - Huge amount
  - Explosive growth
  - Low signal-to-noise ratio
  - Multiple types



ftp://ftp.ncbi.nih.gov/genbank/gbrel.txt

- Requirements for the methods
  - Data needs to be stored in efficient ontology-based database systems
  - The huge amount of data requires efficient algorithms
  - Exponential growth requires scalable methods
  - The low signal-to-noise ratio requires accurate methods
  - Multiple types of data require data integrative methods



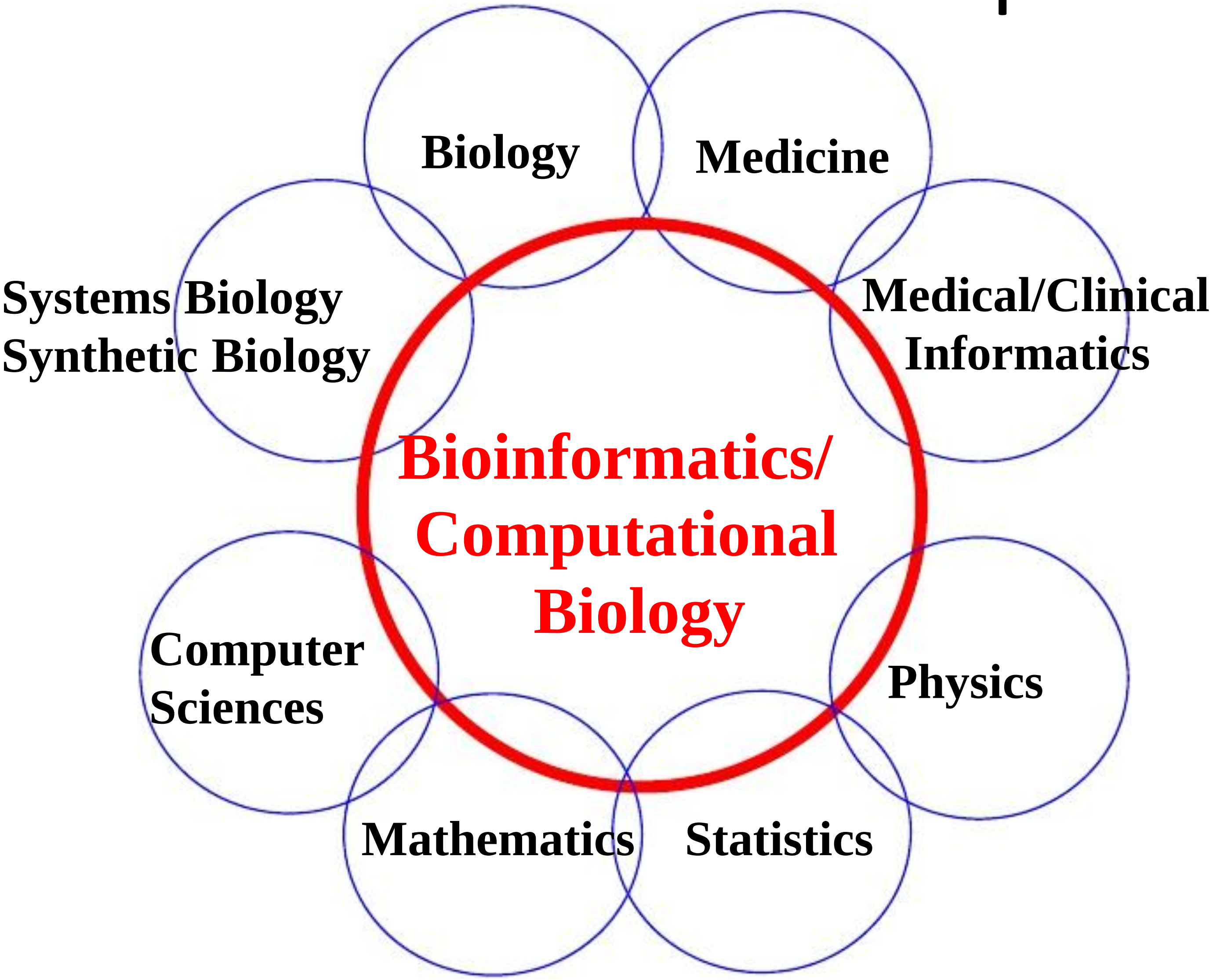
What is bioinformatics?

Bioinformatics: an interdisciplinary field that develops and applies computer and computational technologies to study biomedical questions

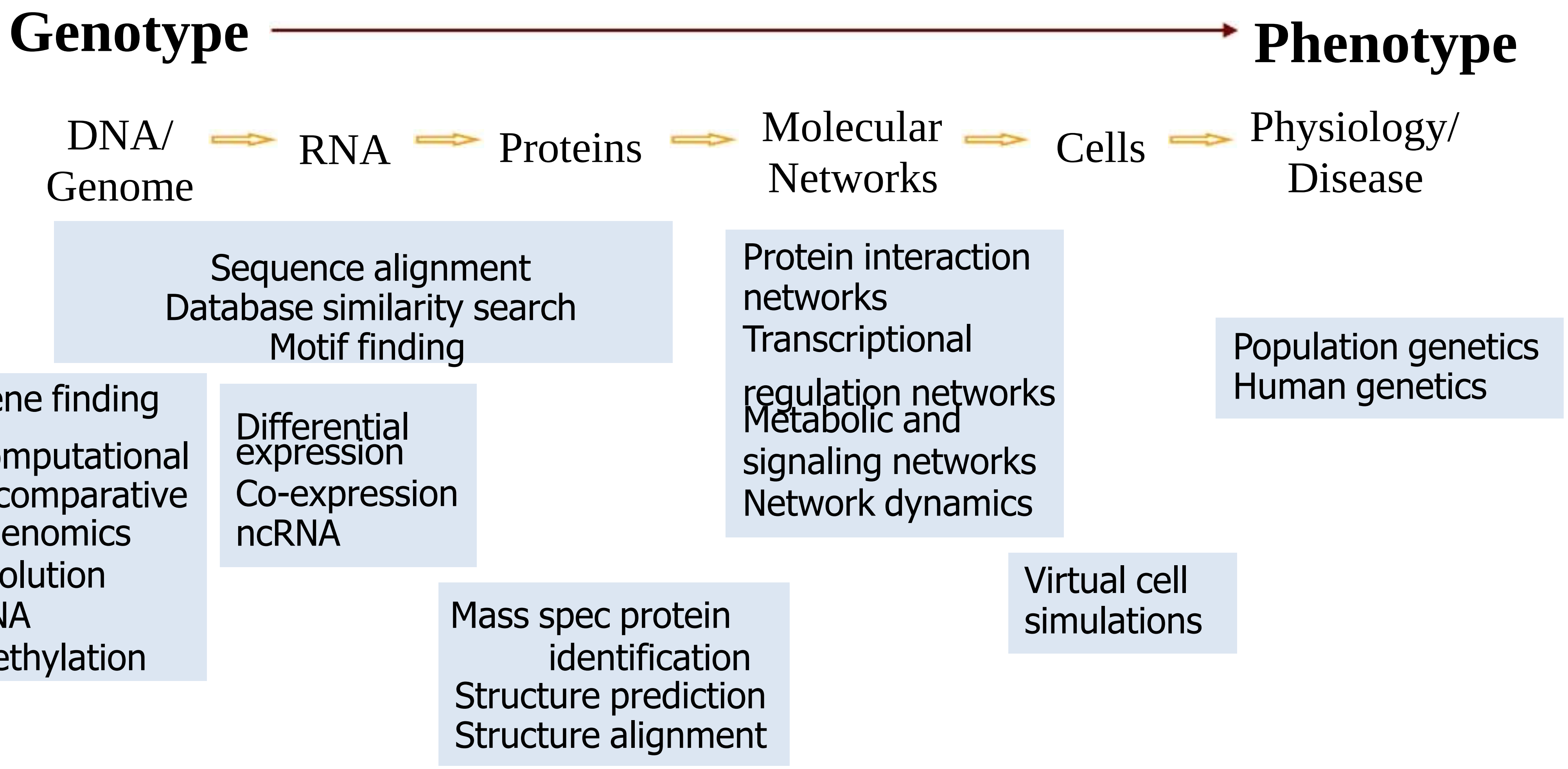
- As a technology, bioinformatics is a powerful technology to manage, query, and analyze big data in life sciences.

- As a methodology, bioinformatics is a top-down, holistic, data-driven, genome-wide, and systems approach that generates new hypotheses, finds new patterns, and discovers new functional elements.

Bioinformatics is an interdisciplinary field



The Bio- in Bioinformatics



Genotype → Phenotype

DNA/ Genome → RNA → Proteins → Molecular Networks → Cells → Physiology/ Disease

Sequence alignment
Database similarity search
Motif finding

Gene finding
Computational & comparative genomics
Evolution
DNA methylation

Differential expression
Co-expression
ncRNA

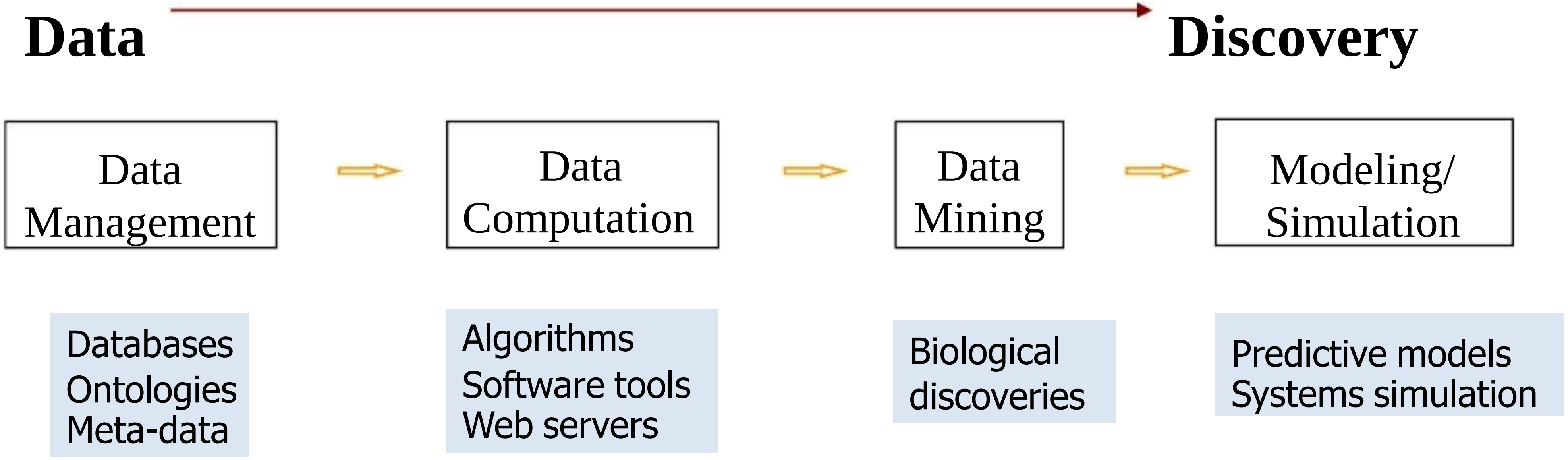
Mass spec protein identification
Structure prediction
Structure alignment

Protein interaction networks
Transcriptional regulation networks
Metabolic and signaling networks
Network dynamics

Virtual cell simulations

Population genetics
Human genetics

The –informatics in Bioinformatics



Data → **Discovery**

Data Management ⇨ Data Computation ⇨ Data Mining ⇨ Modeling/ Simulation

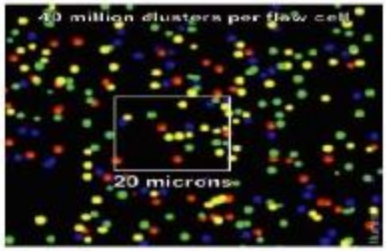
Databases
Ontologies
Meta-data

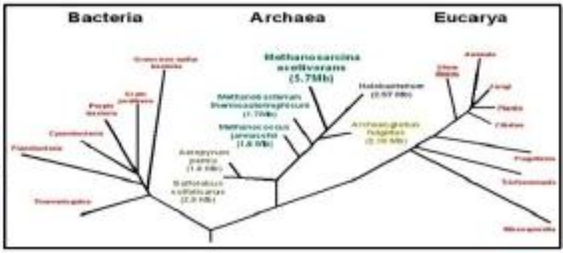
Algorithms
Software tools
Web servers

Biological discoveries

Predictive models
Systems simulation

```
TAACCCTAACCCTAACCCTAACCCTAACCCTA
CCTAACCCTAACCCTAACCCTAACCCTAACCC
CCCTAACCCCTAACCCTAACCCTAACCCTAAC
AACCCTAACCCTAACCCTAACCCCTAACCCTA
ACCCTAACCCCAACCCCAACCCCAACCCCAAC
CTACCCTAACCCTAACCCTAACCCTAACCCTA
ACCCTAACCCTAACCCTAACCCCTAACCCTAA
```



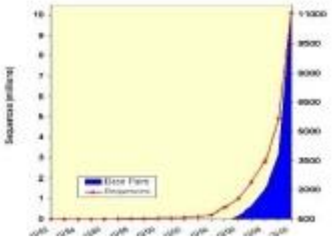


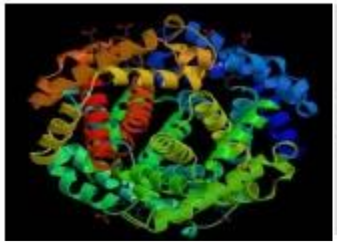
Unit 2: History of bioinformatics

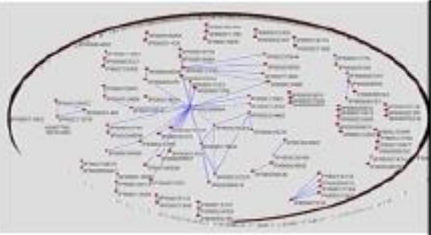
Le Zhang, Ph. D.

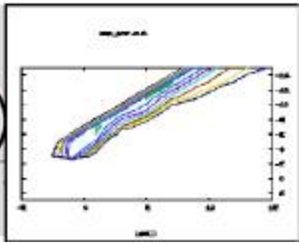
Computer Science Department

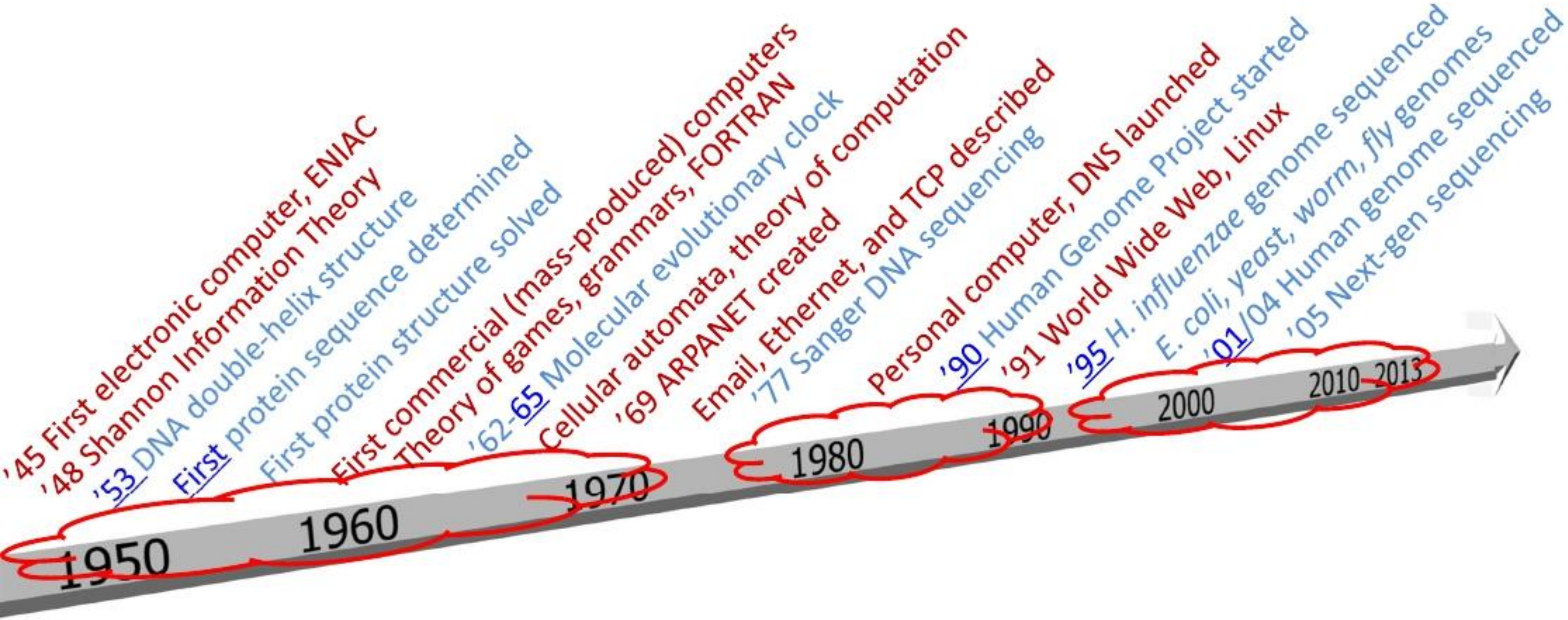
Southwest University











'45 First electronic computer, ENIAC

'48 Shannon Information Theory

'53 DNA double-helix structure

First protein sequence determined

First protein structure solved

First commercial (mass-produced) computers

Theory of games, grammars, FORTRAN

'62-65 Molecular evolutionary clock

Cellular automata, theory of computation

'69 ARPANET created

Email, Ethernet, and TCP described

'77 Sanger DNA sequencing

Personal computer, DNS launched

'90 Human Genome Project started

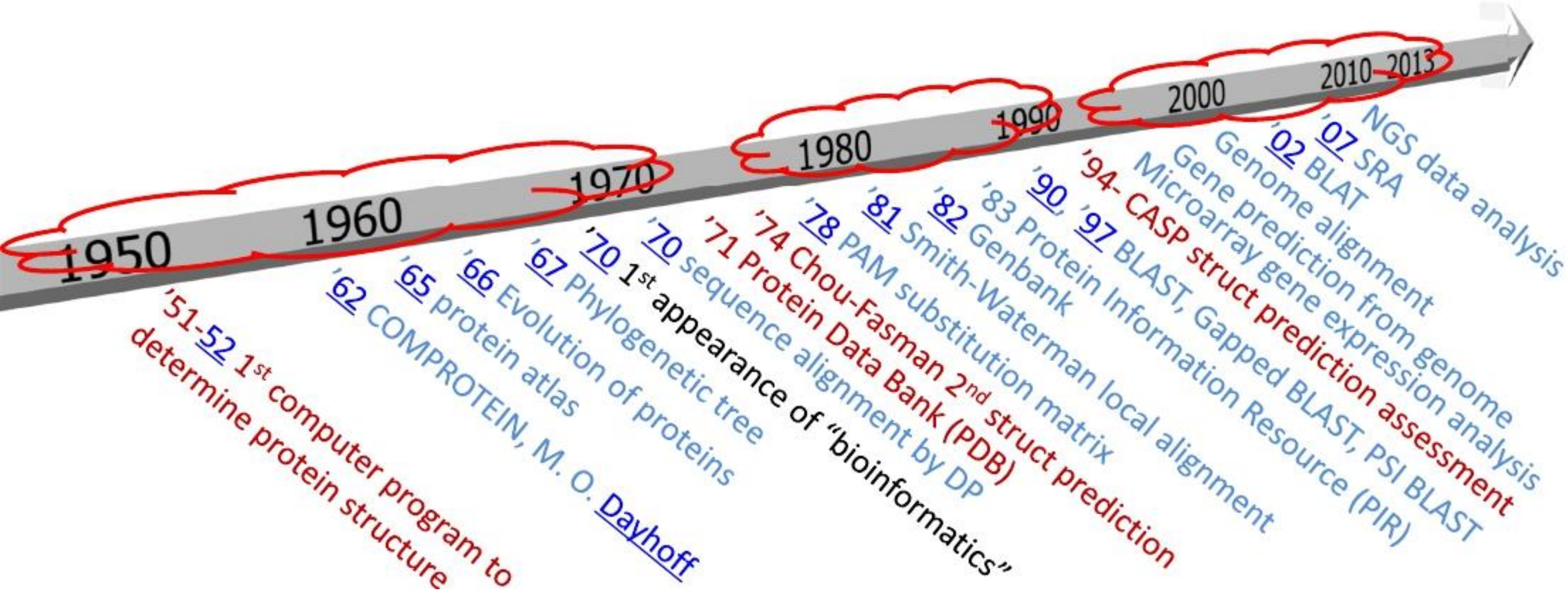
'91 World Wide Web, Linux

'95 *H. influenzae* genome sequenced

E. coli, yeast, worm, fly genomes

'01/04 Human genome sequenced

'05 Next-gen sequencing



'51-52 1st computer program to determine protein structure

'62 COMPROTEIN, M. O. Dayhoff

'65 protein atlas

'66 Evolution of proteins

'67 Phylogenetic tree

'70 1st appearance of "bioinformatics"

'70 sequence alignment by DP

'71 Protein Data Bank (PDB)

'74 Chou-Fasman 2nd struct prediction

'78 PAM substitution matrix

'81 Smith-Waterman local alignment

'82 Genbank

'83 Protein Information Resource (PIR)

'90, '97 BLAST, Gapped BLAST, PSI BLAST

'94- CASP struct prediction assessment

Microarray gene expression analysis

Gene prediction from genome

Genome alignment

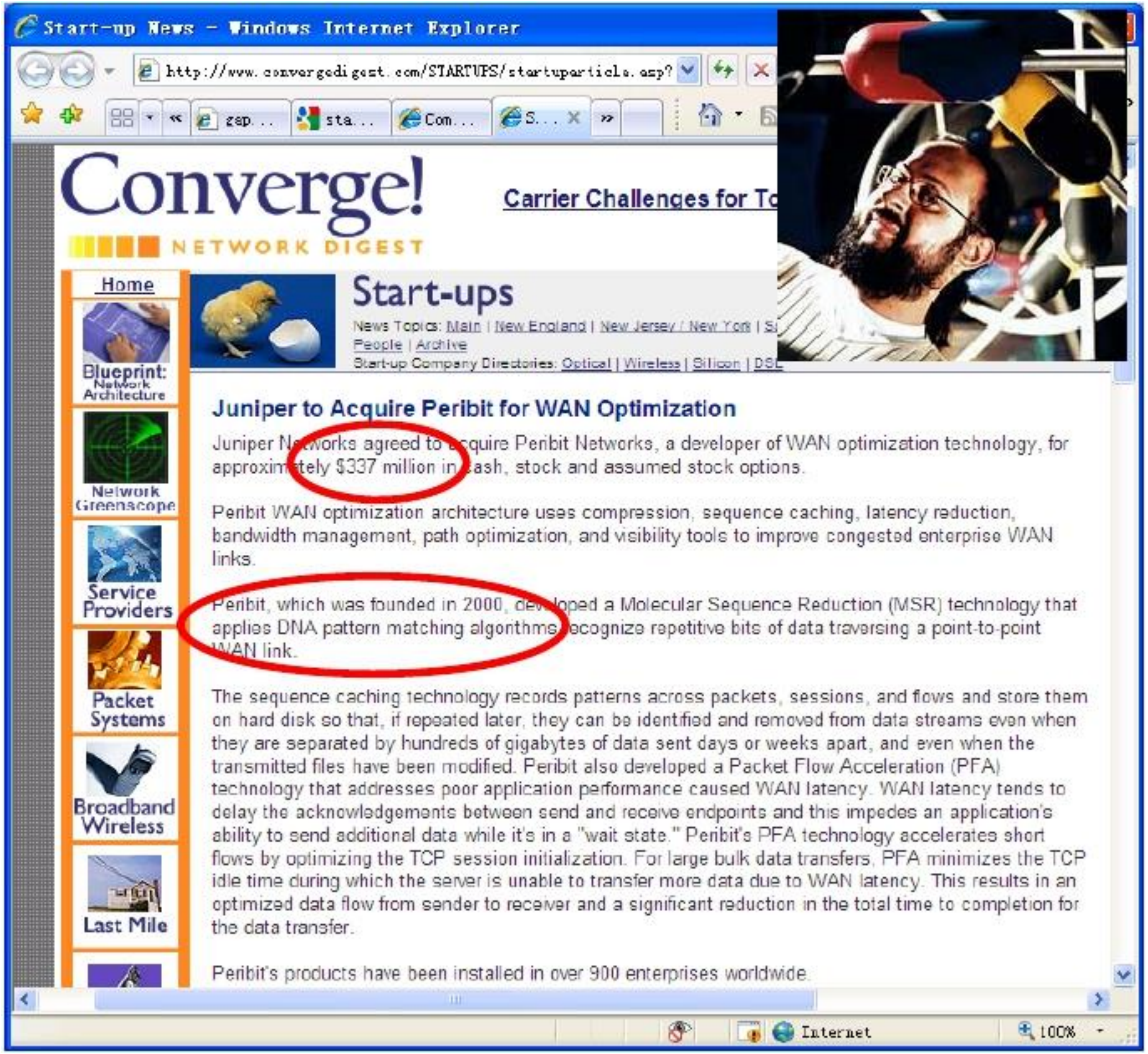
'02 BLAT

'07 SRA

NGS data analysis

Examples of contributions to computer sciences

- Neural networks
- Genetic programming
- WAN


Start-up News - Windows Internet Explorer
http://www.convergedigest.com/STARTUPS/startuparticle.asp?
gap...
sta...
Com...
S...
Converge!
NETWORK DIGEST
Carrier Challenges for To
Home
Blueprint: Network Architecture
Network Greenscope
Service Providers
Packet Systems
Broadband Wireless
Last Mile
Internet
100%

Start-ups

News Topics: Main | New England | New Jersey / New York | S
People | Archive
Start-up Company Directories: Optical | Wireless | Silicon | DSL

Juniper to Acquire Peribit for WAN Optimization

Juniper Networks agreed to acquire Peribit Networks, a developer of WAN optimization technology, for approximately $337 million in cash, stock and assumed stock options.

Peribit WAN optimization architecture uses compression, sequence caching, latency reduction, bandwidth management, path optimization, and visibility tools to improve congested enterprise WAN links.

Peribit, which was founded in 2000, developed a Molecular Sequence Reduction (MSR) technology that applies DNA pattern matching algorithms to recognize repetitive bits of data traversing a point-to-point WAN link.

The sequence caching technology records patterns across packets, sessions, and flows and store them on hard disk so that, if repeated later, they can be identified and removed from data streams even when they are separated by hundreds of gigabytes of data sent days or weeks apart, and even when the transmitted files have been modified. Peribit also developed a Packet Flow Acceleration (PFA) technology that addresses poor application performance caused WAN latency. WAN latency tends to delay the acknowledgements between send and receive endpoints and this impedes an application's ability to send additional data while it's in a "wait state." Peribit's PFA technology accelerates short flows by optimizing the TCP session initialization. For large bulk data transfers, PFA minimizes the TCP idle time during which the server is unable to transfer more data due to WAN latency. This results in an optimized data flow from sender to receiver and a significant reduction in the total time to completion for the data transfer.

Peribit's products have been installed in over 900 enterprises worldwide.

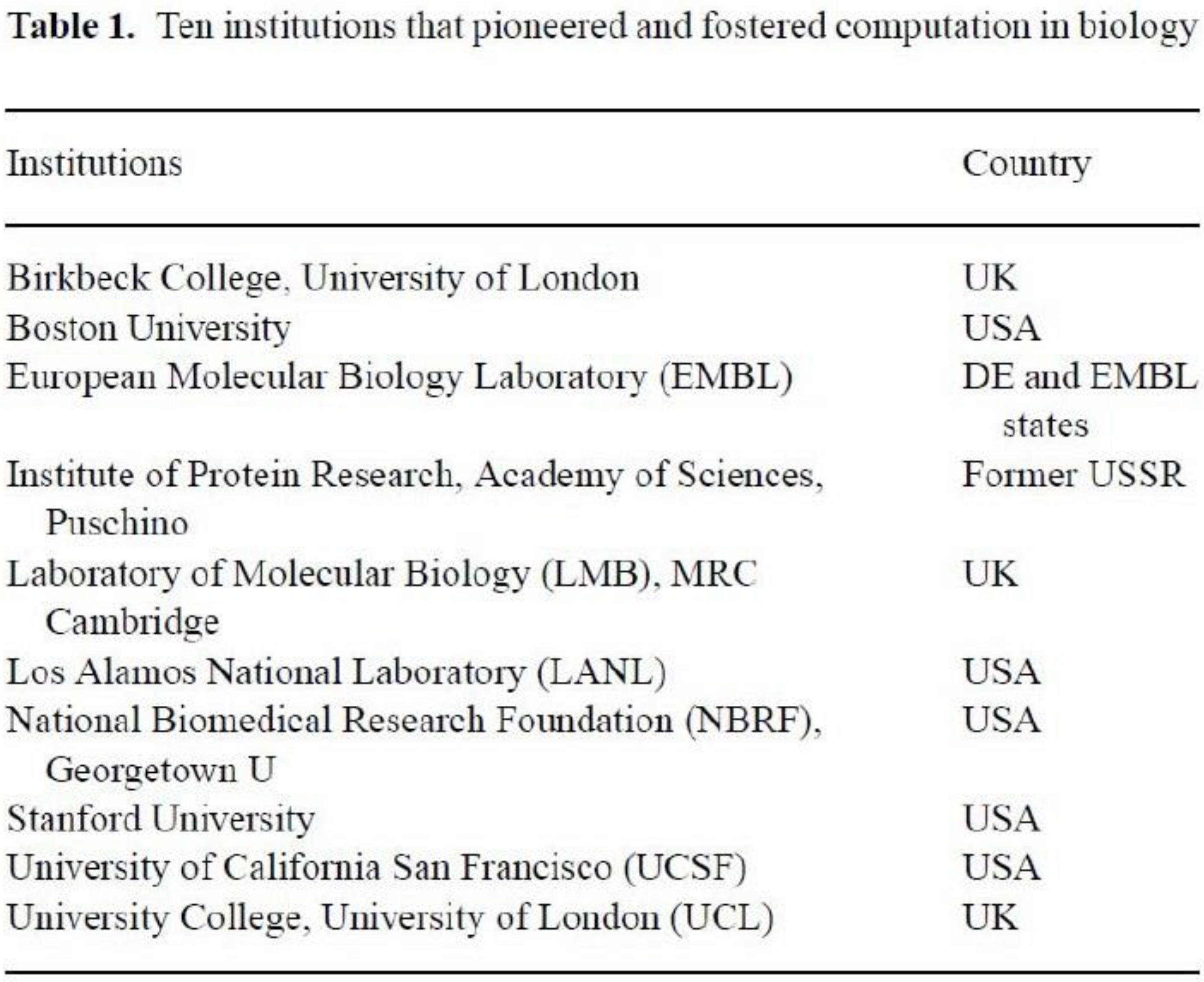


Table 1. Ten institutions that pioneered and fostered computation in biology

Institutions	Country
Birkbeck College, University of London	UK
Boston University	USA
European Molecular Biology Laboratory (EMBL)	DE and EMBL states
Institute of Protein Research, Academy of Sciences, Puschino	Former USSR
Laboratory of Molecular Biology (LMB), MRC Cambridge	UK
Los Alamos National Laboratory (LANL)	USA
National Biomedical Research Foundation (NBRF), Georgetown U	USA
Stanford University	USA
University of California San Francisco (UCSF)	USA
University College, University of London (UCL)	UK

Ouzounis & Valencia, *Bioinformatics*, ‘03

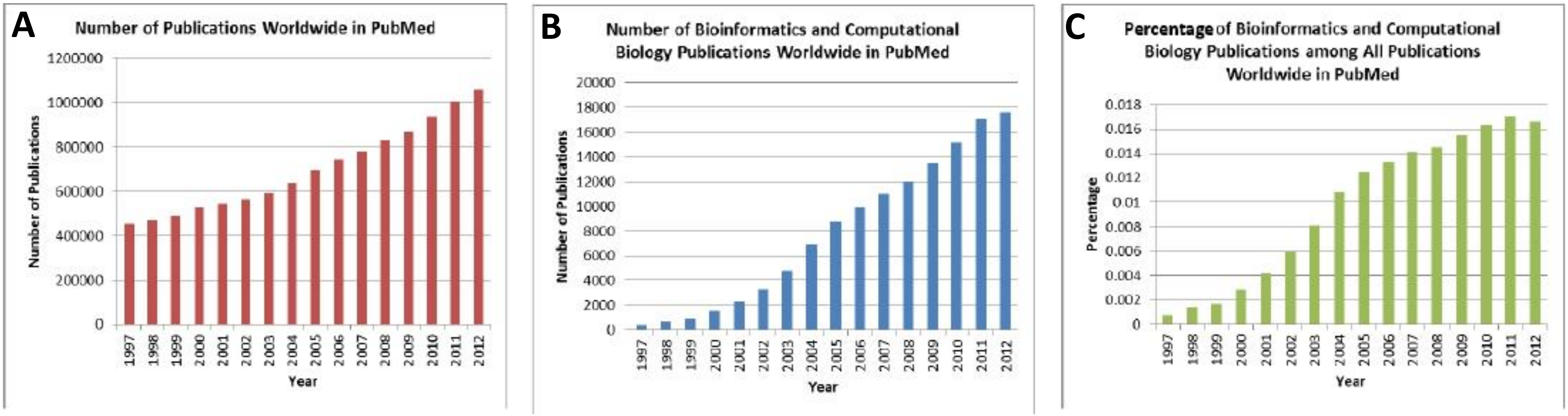
Current Bioinformatics Journals

- Bioinformatics
- BMC Bioinformatics
- BMC Systems Biology
- Briefings in Bioinformatics
- Bulletin of Mathematical Biology
- Cancer Informatics

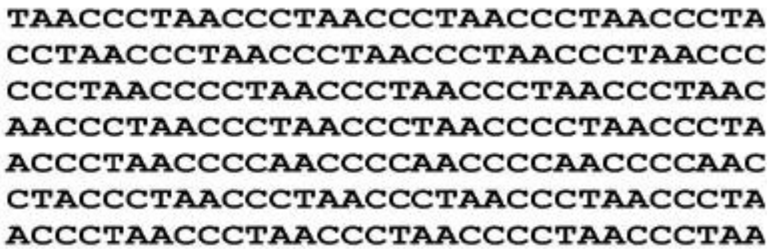
- Computational Biology and Chemistry
- Computers in Biology and Medicine
- Database: The Journal of Biological Databases and Curation
- IEEE/ACM Transactions on Computational Biology and Bioinformatics
- Journal of Bioinformatics and Computational Biology
- Journal of Biomedical Informatics
- Journal of Computational Biology
- Journal of Integrative Bioinformatics

- Journal of Mathematical Biology
- Journal of Theoretical Biology
- PLoS Computational Biology
- Source Code for Biology and Medicine
- Statistical Applications in Genetics and Molecular Biology

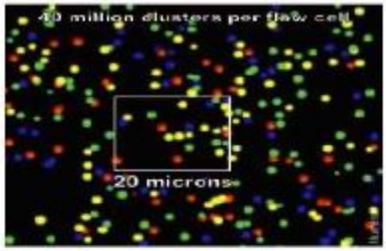
- Nucleic Acids Research
- Genome Research
- Nature Methods
- Nature Biotechnology

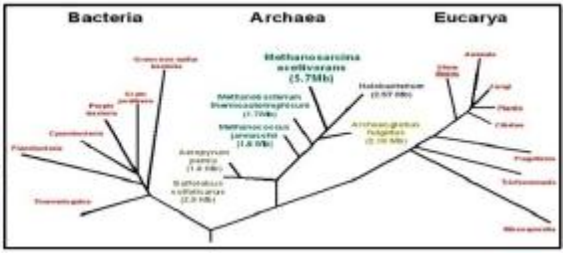

A
Number of Publications Worldwide in PubMed
Number of Publications
1200000
1000000
800000
600000
400000
200000
0
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
Year
B
Number of Bioinformatics and Computational Biology Publications Worldwide in PubMed
Number of Publications
20000
18000
16000
14000
12000
10000
8000
6000
4000
2000
0
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
Year
C
Percentage of Bioinformatics and Computational Biology Publications among All Publications Worldwide in PubMed
Percentage
0.018
0.016
0.014
0.012
0.01
0.008
0.006
0.004
0.002
0
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
Year

China?



TAACCCTAACCCTAACCCTAACCCTAACCCTA
CCTAACCCTAACCCTAACCCTAACCCTAACCC
CCCTAACCCCTAACCCTAACCCTAACCCTAAC
AACCCTAACCCTAACCCTAACCCCTAACCCTA
ACCCTAACCCCAACCCCAACCCCAACCCCAAC
CTACCCTAACCCTAACCCTAACCCTAACCCTA
ACCCTAACCCTAACCCTAACCCCTAACCCTAA



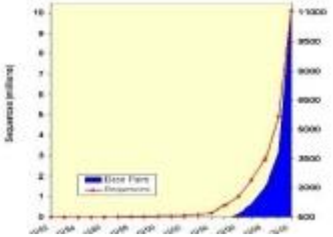


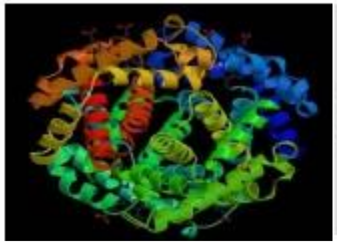
Unit 3:
Bioinformatics in Mainland China

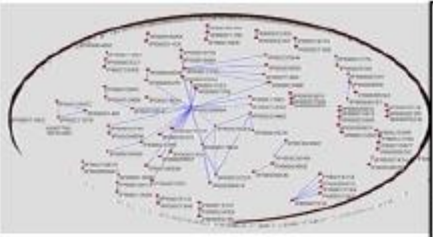
Le Zhang, Ph. D.

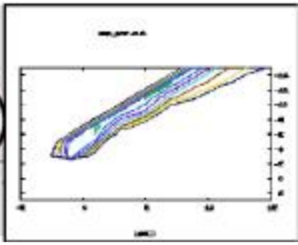
Computer Science Department

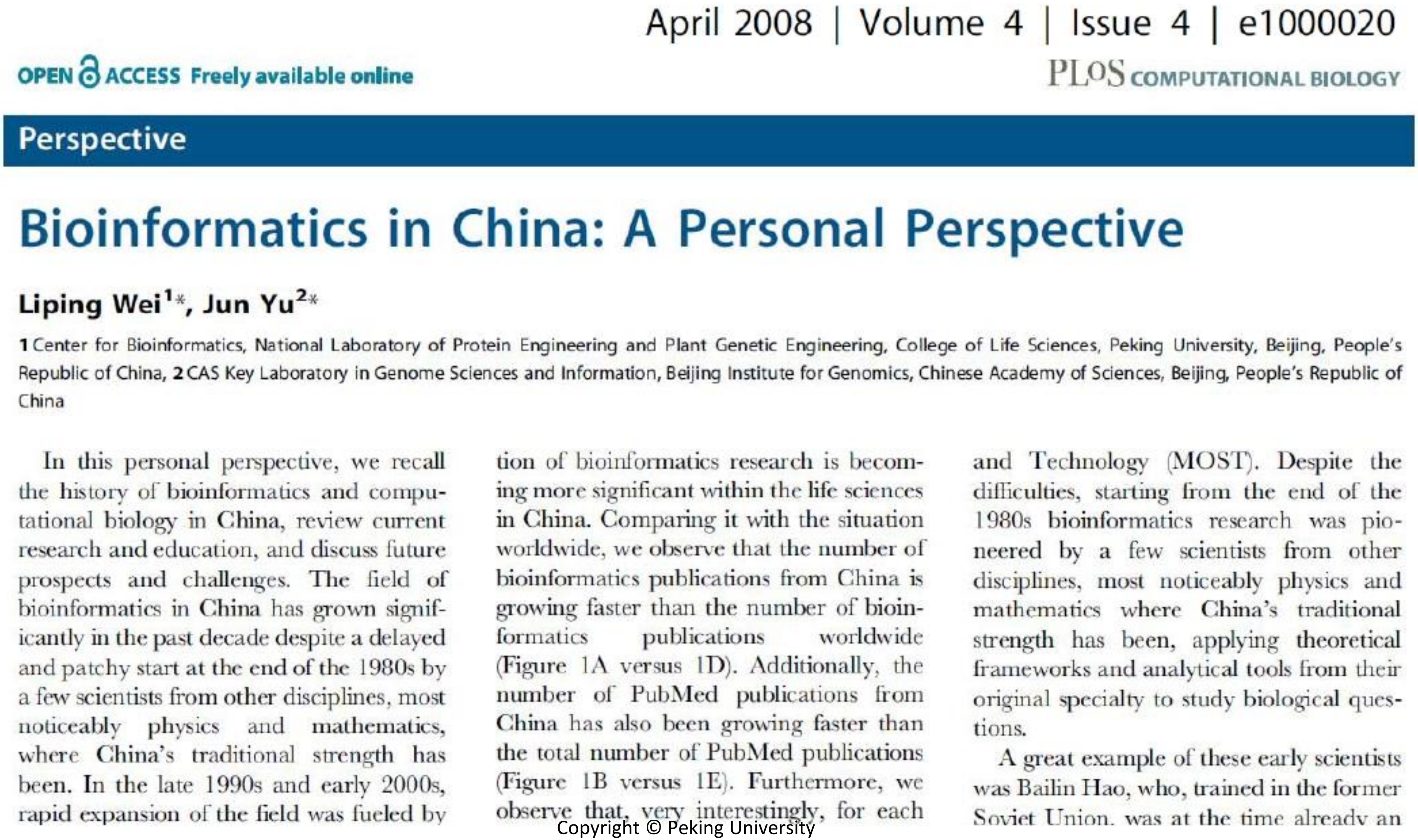
Southwest University











April 2008 | Volume 4 | Issue 4 | e1000020

OPEN ACCESS Freely available online

PLoS COMPUTATIONAL BIOLOGY

Perspective

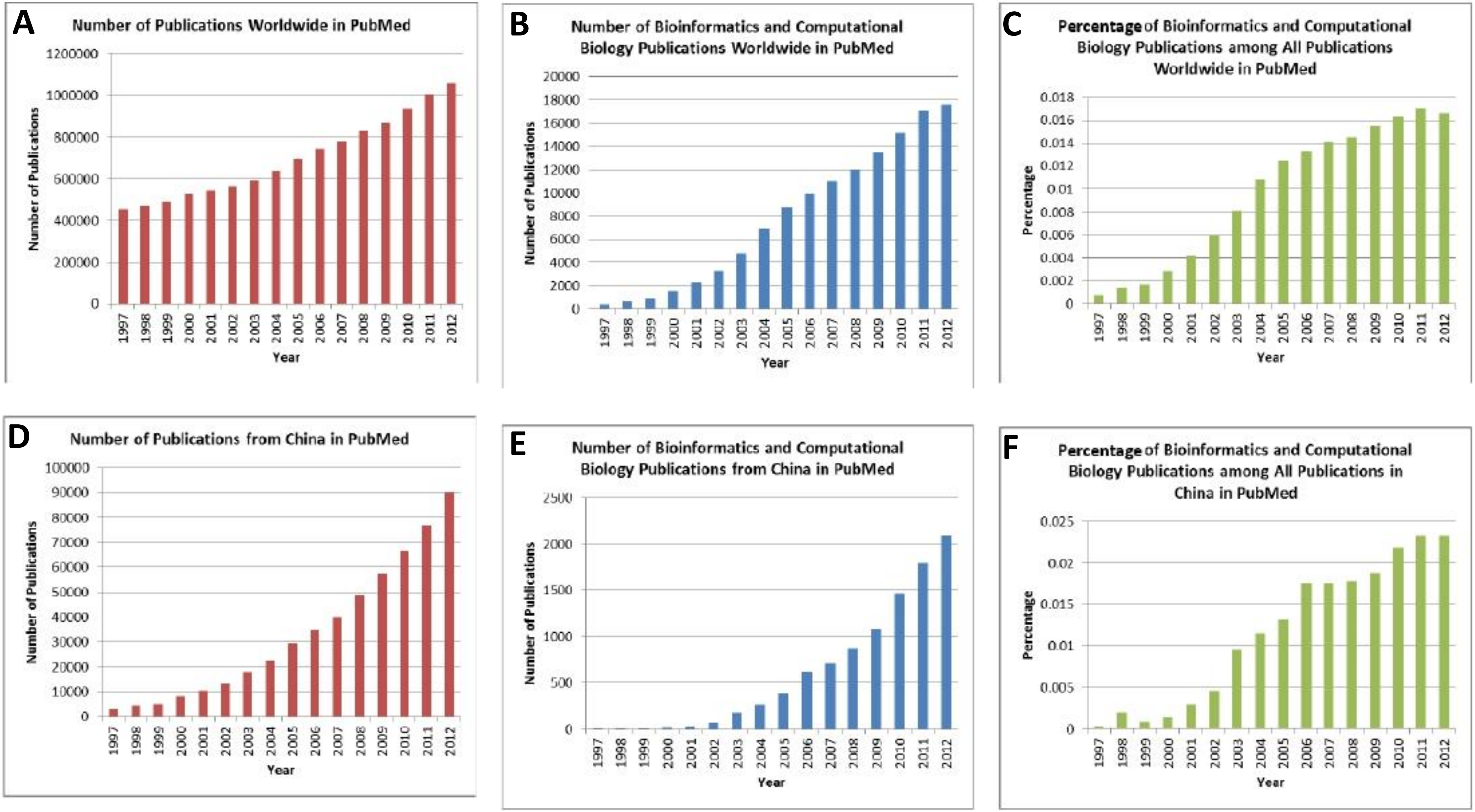
Bioinformatics in China: A Personal Perspective

Liping Wei[1]*, Jun Yu[2]*

[1] Center for Bioinformatics, National Laboratory of Protein Engineering and Plant Genetic Engineering, College of Life Sciences, Peking University, Beijing, People's Republic of China, [2] CAS Key Laboratory in Genome Sciences and Information, Beijing Institute for Genomics, Chinese Academy of Sciences, Beijing, People's Republic of China

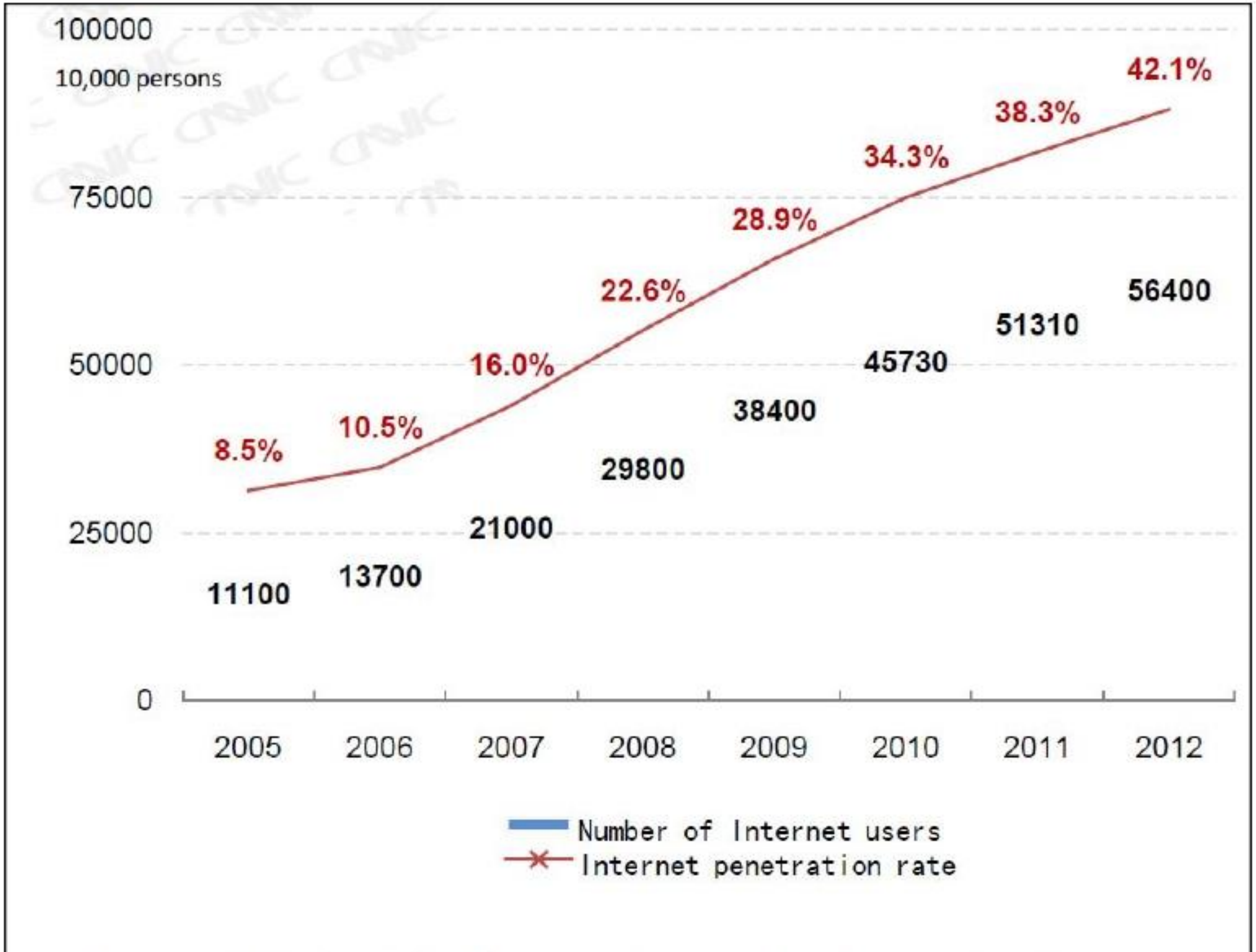
In this personal perspective, we recall the history of bioinformatics and computational biology in China, review current research and education, and discuss future prospects and challenges. The field of bioinformatics in China has grown significantly in the past decade despite a delayed and patchy start at the end of the 1980s by a few scientists from other disciplines, most noticeably physics and mathematics, where China's traditional strength has been. In the late 1990s and early 2000s, rapid expansion of the field was fueled by tion of bioinformatics research is becoming more significant within the life sciences in China. Comparing it with the situation worldwide, we observe that the number of bioinformatics publications from China is growing faster than the number of bioinformatics publications worldwide (Figure 1A versus 1D). Additionally, the number of PubMed publications from China has also been growing faster than the total number of PubMed publications (Figure 1B versus 1E). Furthermore, we observe that, very interestingly, for each and Technology (MOST). Despite the difficulties, starting from the end of the 1980s bioinformatics research was pioneered by a few scientists from other disciplines, most noticeably physics and mathematics where China's traditional strength has been, applying theoretical frameworks and analytical tools from their original specialty to study biological questions.

A great example of these early scientists was Bailin Hao, who, trained in the former Soviet Union, was at the time already an


A
Number of Publications Worldwide in PubMed
Number of Publications
0
200000
400000
600000
800000
1000000
1200000
1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012
Year
B
Number of Bioinformatics and Computational Biology Publications Worldwide in PubMed
Number of Publications
0
2000
4000
6000
8000
10000
12000
14000
16000
18000
20000
1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012
Year
C
Percentage of Bioinformatics and Computational Biology Publications among All Publications Worldwide in PubMed
Percentage
0
0.002
0.004
0.006
0.008
0.01
0.012
0.014
0.016
0.018
1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012
Year
D
Number of Publications from China in PubMed
Number of Publications
0
10000
20000
30000
40000
50000
60000
70000
80000
90000
100000
1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012
Year
E
Number of Bioinformatics and Computational Biology Publications from China in PubMed
Number of Publications
0
500
1000
1500
2000
2500
1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012
Year
F
Percentage of Bioinformatics and Computational Biology Publications among All Publications in China in PubMed
Percentage
0
0.005
0.01
0.015
0.02
0.025
1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012
Year

Rapid growth since 1990s

- **Driving force #1: internet**
  - 1994: full TCP/IP internet connection between China and the rest of the world was established.
    - A dial-up X.25 connection between the Institute of High-Energy Physics (IHEP) in Beijing and the Stanford Linear Accelerator Center (SLAC).
  - On May 17th 1994, the official connection to FIX-West was announced and the U.S.-based Energy Sciences Network (ESnet) agreed to carry China IP traffic.



Source: CNNIC Statistical Survey on Internet Development in China

- **Driving force #2: genomics**
  - China National Human Genome Center (Shanghai and Beijing), Beijing Genomics Institute (BGI), Beijing Institute of Genomics, CAS
  - 1% of human genome sequencing
  - Sequencing of rice and many other genomes

CHGC

国家人类基因组南方研究中心

国家人类基因组北方研究中心

Chinese National Human Genome Center, Beijing

- **Driving force #3: increased research funding**
  - Ministry of Science and Technology (MOST): “863” and “973” grants
  - Natural Science Foundation of China (NSFC): young investigator awards and research grants

中华人民共和国科学技术部

Ministry of Science and Technology of the People´s Republic of China

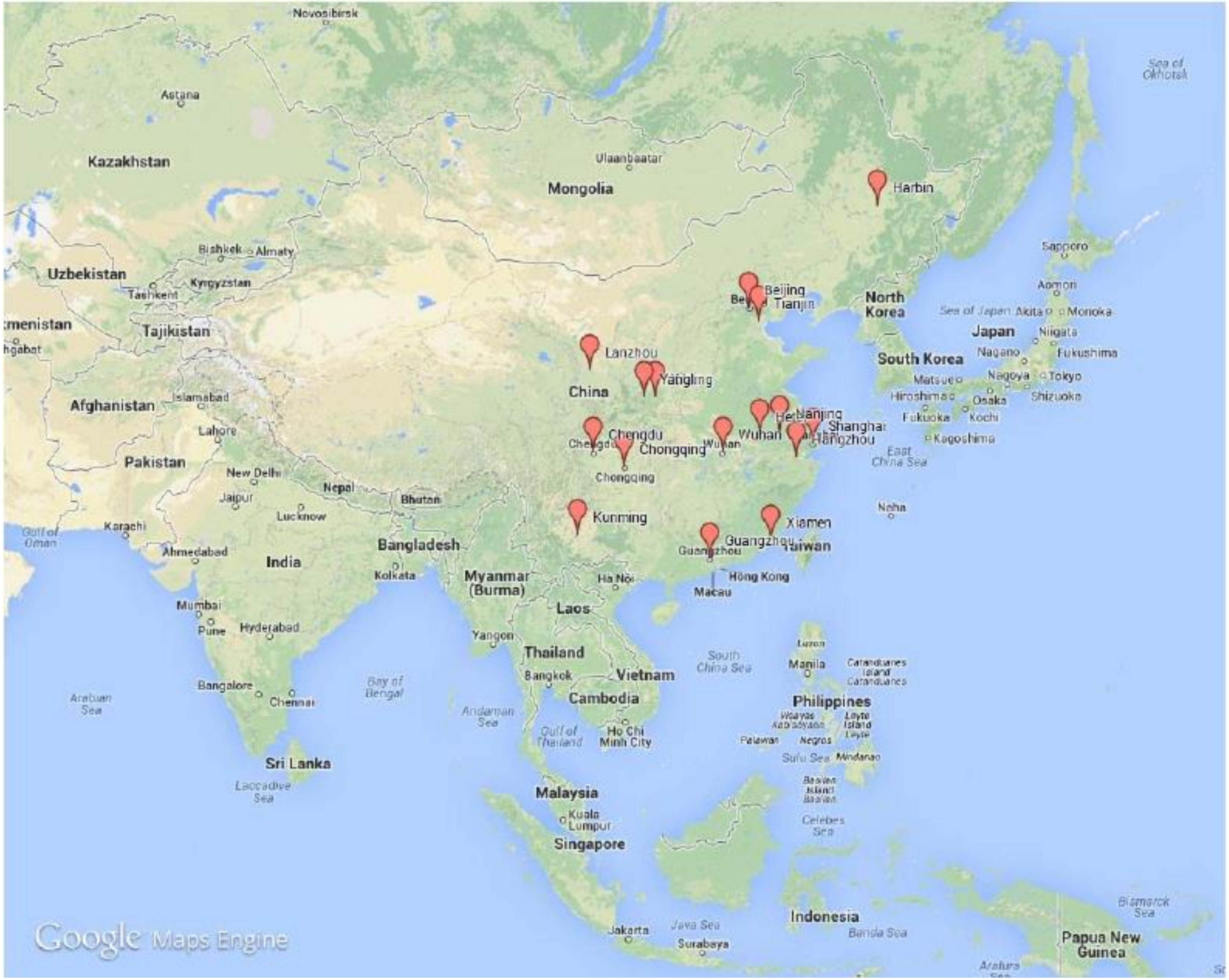
国家自然科学基金委员会

National Natural Science Foundation of China

- **Driving force #4: critical mass of researchers & students**

Pioneers in the late 80s and early 90s

Name	Original background	Research in bioinformatics
Bai-Lin Hao	Theoretical physics	Phylogeneticanalysis with k-strings
Run-Sheng Chen	Biophysics	Small RNAs and noncoding RNAs
Chun-Ting Zhang	Theoretical physics	Z-score forDNA sequence analysis
YandaLi	Automation control	Gene expression regulation
Yunyu Shi	Biophysics	Structural biology and bioinformatics
Liaofu Luo	Theoreticalphysics	Genome evolution
Dafu Ding	Mathematics	Protein structure modeling and design
ZhirongSun	Automation control	Molecular network and pathway analysis
LuhuaLai	Chemistry	Docking and drug design
XiaochengGu	Biophysics	Protein structureanalysis
JingchuLuo	Biology	Plant transcriptionregulation and evolution



Based on Google Map

Bioinformatics degree programs in Mainland China

City	University/Institute		School/Center	Degrees
Beijing	Peking University	北京大学	Center for Bioinformatics, School of Life Sciences; Center for Quantatitive Biology	PhD
Beijing	Tsinghua University	清华大学	Department of Biological Sciences and Biotechnology; Institute of Bioinformatics, Department of Automation	PhD
Beijing	Chinese Academy of Sciences	中科院	Beijing Institute of Genomics; Center of Systems Biology, Institute of Biophysics; Center of Molecular Systems Biology, Institute of Genetics and Developmental Biology	PhD
Beijing	China Agricultural University	中国农业大学	College of Biological Sciences	PhD, Master
Beijing	Beijing Normal University	北京师范大学	College of Life Sciences, Laboratory of Computational Molecular Biology	Master
Chengdu	Sichuan University	四川大学	School of Life Sciences	PhD, Master
Chongqing	Chongqing University of Posts and Telecommunications	重庆邮电大学	College of Bio-information	Bachelor
Guangzhou	Sun Yat-sen University	中山大学	Center for Bioinformatics, College of Life Sciences	PhD, Master
Hangzhou	Zhejiang University	浙江大学	School of Life Science, Institute of Bioinformatics	Bachelor
Harbin	Harbin Medical University	哈尔滨医科大学	College of Bioinformatics Science and Technology	Master,Bachelor
Hefei	University of Science and Technolog	中国科学技术大学	School of Life Sciences	PhD, Master

Nanjing	Nanjing University	南京大学	School of Life Science	PhD, Master
Nanjing	Nanjing Agricultural University	南京农业大学	Center for Bioinformatics, College of Life Sciences	Master
Nanjing	Southeast University	东南大学	State Key Laboratory of Bioelectronics, School of Biological Science & Medical Engineering	PhD, Master
Nanjing	China Pharmaceutical University	中国药科大学	School of Life Science and Technology	PhD, Master
Shanghai	Fudan University	复旦大学	School of Life Sciences	PhD, Master
Shanghai	Shanghai Institute for Biological Sciences	上海生命科学研究所	Key Laboratory of Systems Biology	PhD
Shanghai	Tongji University	同济大学	School of Life Science	PhD, Master, BS
Shanghai	Shanghai Jiao Tong University	上海交通大学	Department of Biomedical Engineering, College of Life Science and Biotechnology	Master
Shanghai	East China Normal University	华东师范大学	School of Life Sciences	PhD, Master
Shanghai	Shanghai Jiao Tong University	上海大学	School of Life Sciences	Master
Tianjin	Nankai University	南开大学	College of Life Sciences	PhD, Master
Tianjin	Tianjin University	天津大学	Tianjin University BioInformatics Centre	Master
Wuhan	Huazhong Agricultural University	华中农业大学	School of Life Sciences	Master, Bachelor
Wuhan	Huazhong University of Science and Technology	华中科技大学	School of Life Science and Technology	PhD, Master, BS
Xiamen	Xiamen University	厦门大学	Department of Chemistry	PhD, Master
Xi'an	Xi'an Jiaotong University	西安交通大学	School of Life Science, Institute of Bioinformatics	PhD, Master
Yangling	Northwest Agriculture and Forestry University	西北农林科技大学	Center for Bioinformatics, College of Life Sciences	PhD, Master

Bioinformatics: Introduction and Methods

Computer Science Department, Southwest University

Thank you