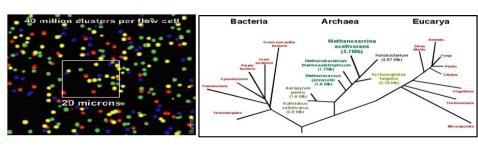
生物信息学:导论与方法 Bioinformatics: Introduction and Methods

Ge Gao 高歌 & Liping Wei 魏丽萍 Center for Bioinformatics, Peking University





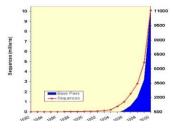
TAACCCTAACCCTAACCCTAACCCTA CCTAACCCTAACCCTAACCCTAACCC CCCTAACCCTAACCCTAACCCTAAC AACCCTAACCCTAACCCTAACCCTA ACCCTAACCCCAACCCCAACCCCAAC CTACCCTAACCCTAACCCTAACCCTA ACCCTAACCCTAACCCTAACCCTAA

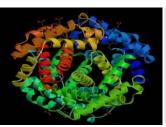


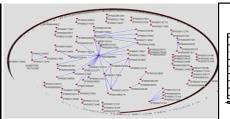
Introduction and History

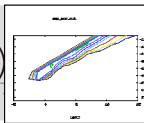
北京大学生物信息学中心 魏丽萍 Liping Wei, Ph.D. Center for Bioinformatics, Peking University





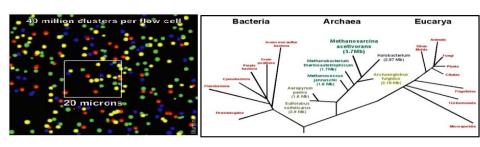








TAACCCTAACCCTAACCCTAACCCTA CCTAACCCTAACCCTAACCCTAACCC CCCTAACCCTAACCCTAACCCTAAC AACCCTAACCCTAACCCTAACCCTA ACCCTAACCCCAACCCCAACCCCAAC CTACCCTAACCCTAACCCTAACCCTA ACCCTAACCCTAACCCTAACCCTAACCCTAA

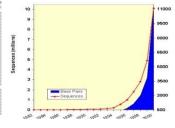


Unit 1: What is bioinformatics?

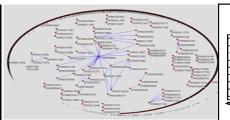
北京大学生物信息学中心 魏丽萍 Liping Wei, Ph.D.

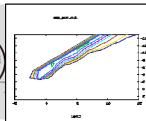
Center for Bioinformatics, Peking University



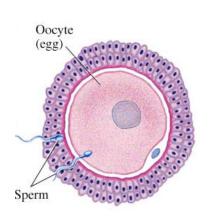








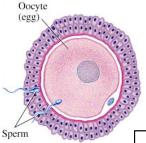
The miracle of life



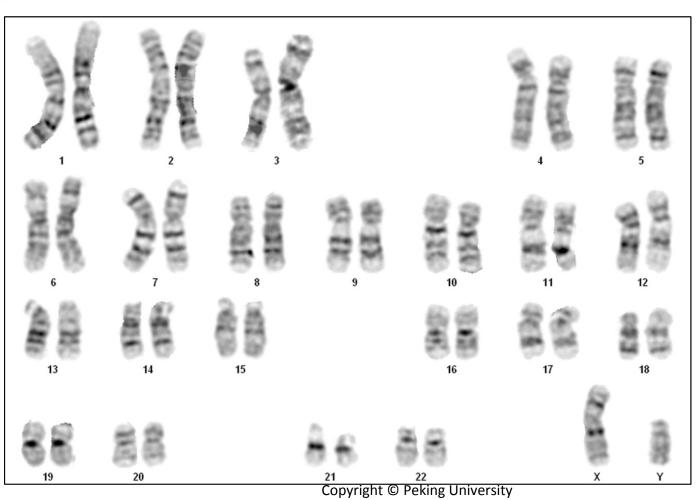




http://img.webmd.com/dtmcms/live/webmd/consumer_assets/site_images/media/medical/hw/n5551221.jpg
http://commons.wikimedia.org/wiki/File:Hospital_newborn_by_Bonnie_Gruenberg2.jpg
Copyright © Peking University



Genome: the "manual of life" -- Almost



mitochondrial DNA

epigenetics

environments/nurture

chance

TAACCCTAACCCTA ACCCTA ACCCTA ACCCTA ACCCTA ACCCTA ACCCTA ACCCTAACCCT CCTAACCCTAACCC

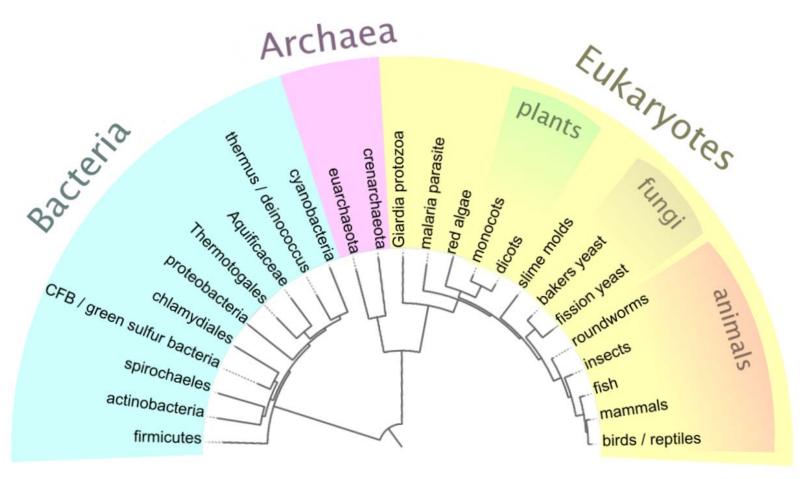
So simple, yet so mysterious... TAACCCTAACCC

CCCTAACCCCTAACCCCTAACCCCTAACCCCTAACCCTAACCCCTAACCCCTAACCCCTAACCCCTAACCCCTAACCCCTAACCCCTAACCCCTAACCCCTAACCCCTAACCCCTAACCCCTAACCCCCTAACCCCTAACCCCTAACCCCTAACCCCTAACCCCTAACCCCCTAACCCCTAACCCCTAACCCCTAACCCCTAACCCCTAACCCCCTAACCCCCTAACCCCTAACCCCTAACCCCTAACCCCTAACCCCTAACCCCCTAACCCCCTAACCCCTAACCCCTAACCC AACCCTAACCCTAACCCCTAACCCTAACCCTAAACCCTAAACCCTAACCCTAACCCTAACC ACCCTAACCCCAA CTACCCTAACCCCAA

CTACCCTAACCCT

Human genome has 3.1 billion base pairs. CTAACC ACCCTAACCCTAACCCCTAACCCTAACCCTAACCCTAACCCTCGCGGTACCCTCAGCCGGC CCCGCCCGGGTCTGACCACACAACTCTCCCCCCTTCACACTACCACCGAAATCTGTGCAGAG AACGCAGCTCCGCCC ~2.9% of the bases encode genes. ACTCCGCCGGCGCAG CGCAGAGAGGCGCGCCGCCGCCGCCAGGCGCAGAGACACATGCTACCGCGTCCAGGGGTGGAGGCG CGCAGGG ~97% of the genome was previously called "junk". AGGAGCAA They contain the regulatory elements that encode instructions on when, where, which, and how much proteins to make.

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

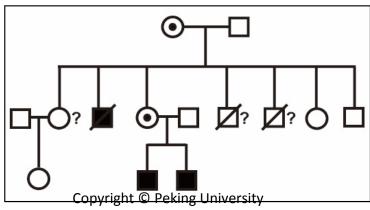


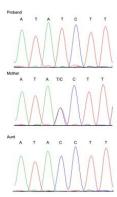
http://commons.wikimedia.org/wiki/File:Simplified_tree.png

Human Genetic Variations









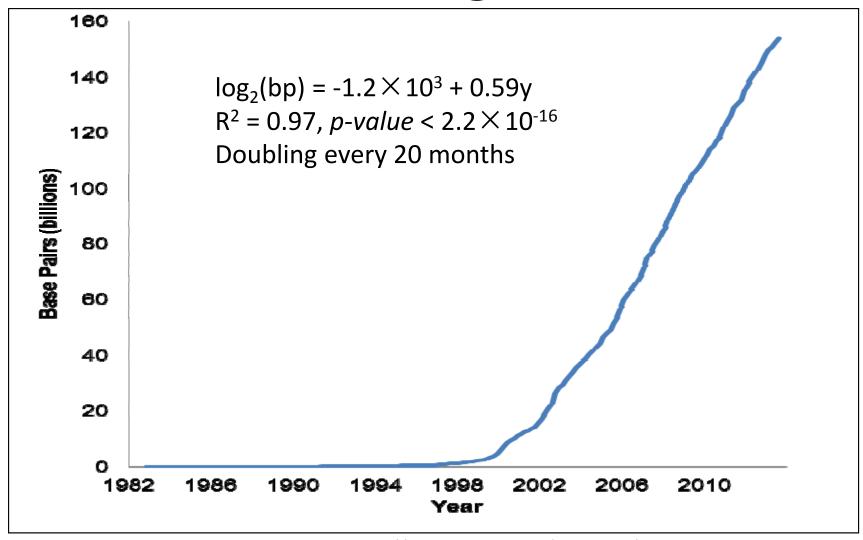
TAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTA	
$CC^{\mathtt{m}}$ א א פרויים א א פרויים א פרויים א	
cc How do you decode the instructions in this "manual"	
AACCCTAACCCTAACCCTAACCCCTAACCCTAAACCCTAAACCCTAAACC	CTAACCCTAACC
ACCCTAACCCCAACCCCAACCCCAACCCCAACCCCAACCCTAACCCCTAA	ACCCTAACCCT
If you print 100 characters per line and 50 lines per	page, Acc
cc you'll fill 600,000 pages, stacked 60 meter	
AACGCAGCTCCGCCCTCGCGGTGCTCTCCGGGTCTGTGCTGAGGAGAACGCAACTC	
CAGAGAGGCGCGCCGCCGCGCGCAGGCGCAGACACATGCTAGCGCGTCGGGGTGG	JAGGCGTGGCG(
CGCAGAGAGGG If you read one base per second, nonstop,	GGGTGGAGGC
CGCAGGCGCAC	CGTCCAGGGGT
gcgtggcgcac it will take you 100 years.	rtgcaggagca <i>i</i>
CGCACGGCGCCGGGGGGGGGGGGGGGGGGGGGGGGGGG	CACGTCACGGT
CGGCGCAGAGACGGGTAGAACCTCAGTAATCCGAAAAGCCGGGGATCGACCGCCCCT	rtgcttgcagc(

10¹⁵ basepairs in >165,000 organisms! GCTC TTGCT It will take you 30 million years to finish reading!

ACTGC

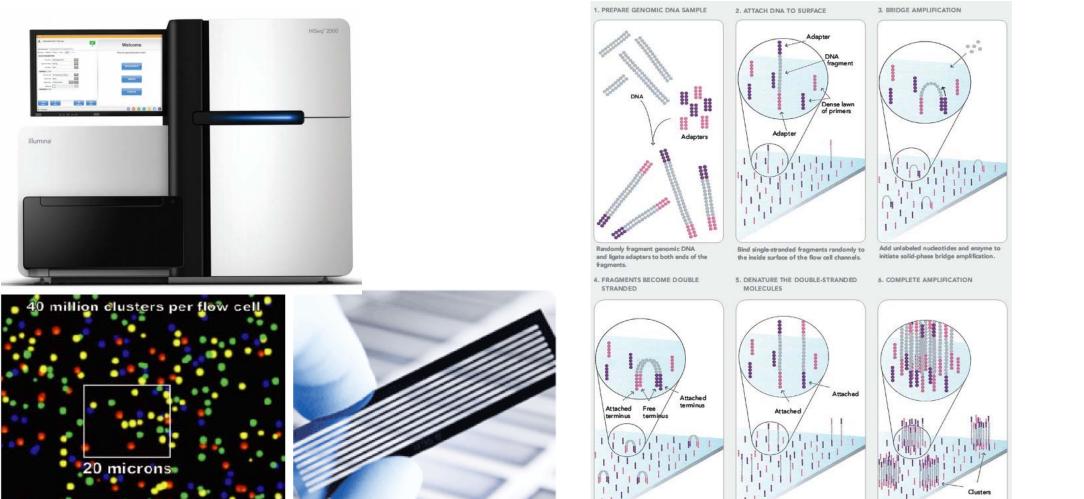
CACI

Genbank growth



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

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



Copyright © Peking University

The enzyme incorporates nucleotides to

build double-stranded bridges on the solid-

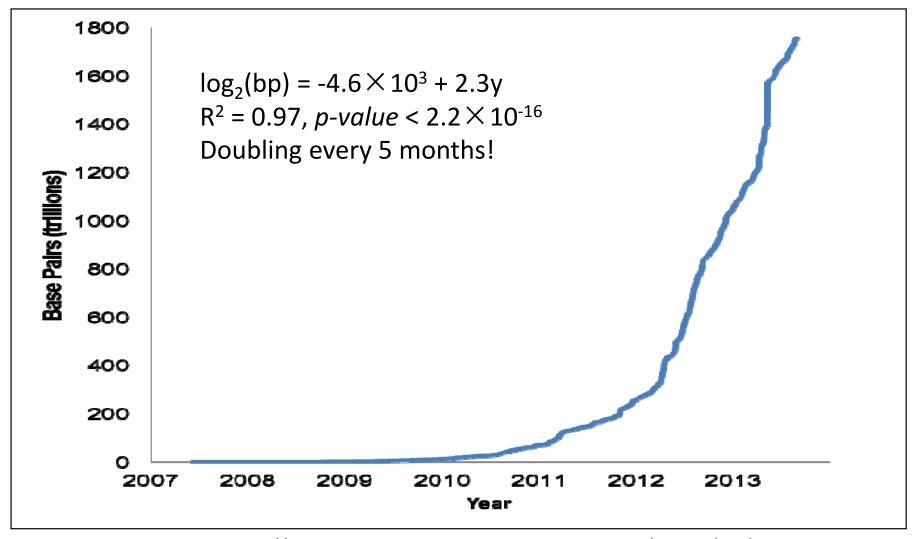
Several million dense dustr

stranded DNA are generate

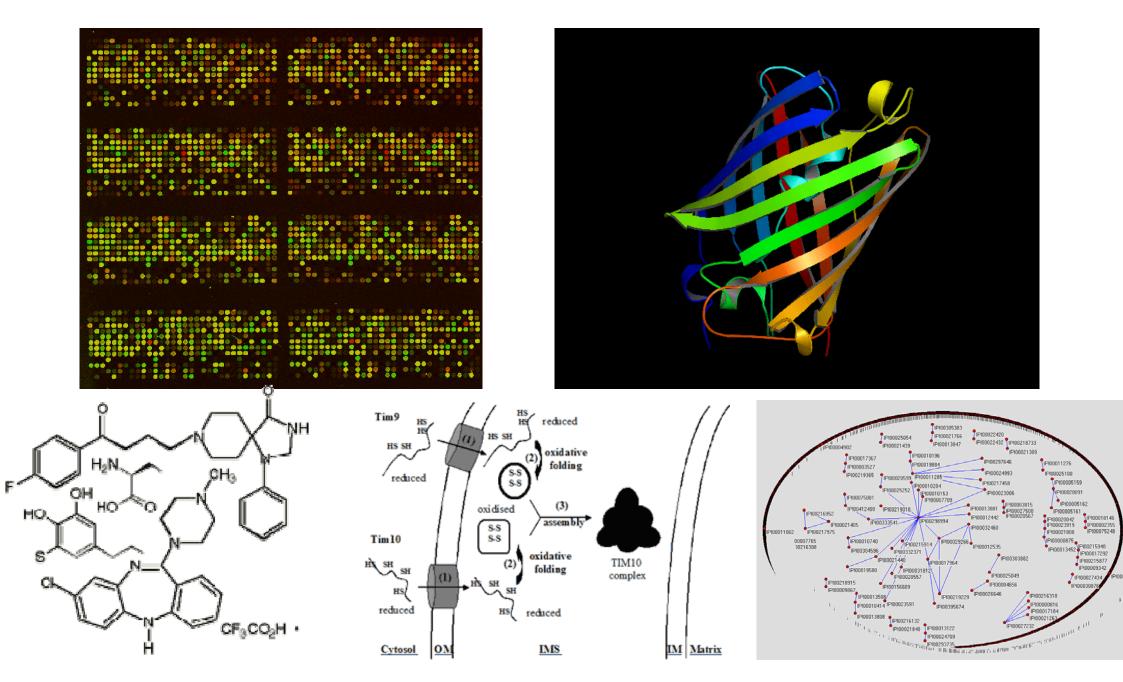
Denaturation leaves single-stranded

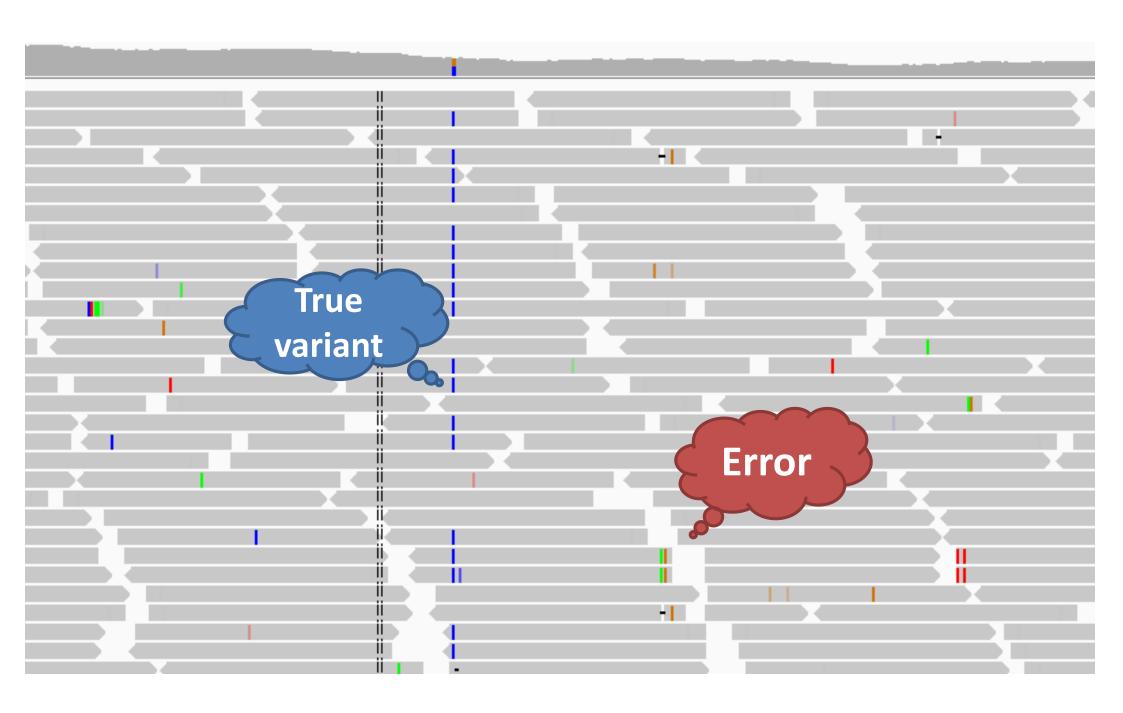
templates anchored to the substrate.

SRA Growth



Data source: http://0-www.ncbi.nlm.nih.gov.elis.tmu.edu.tw/Traces/sra/





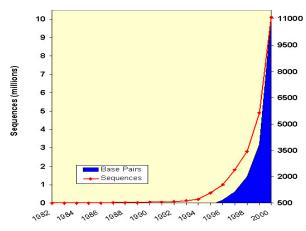
Opportunities and challenges hand-in-hand: the driving forces of bioinformatics

High-throughput data

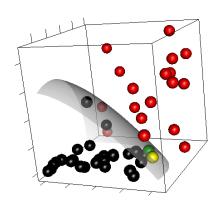
- Huge amount
- Explosive growth
- Low signal-to-noise ratio
- Multiple types

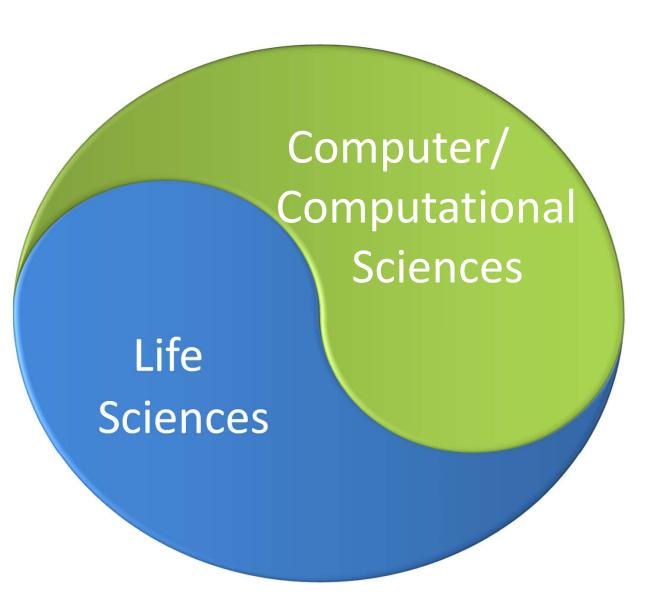
Requirements for the methods

- 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



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





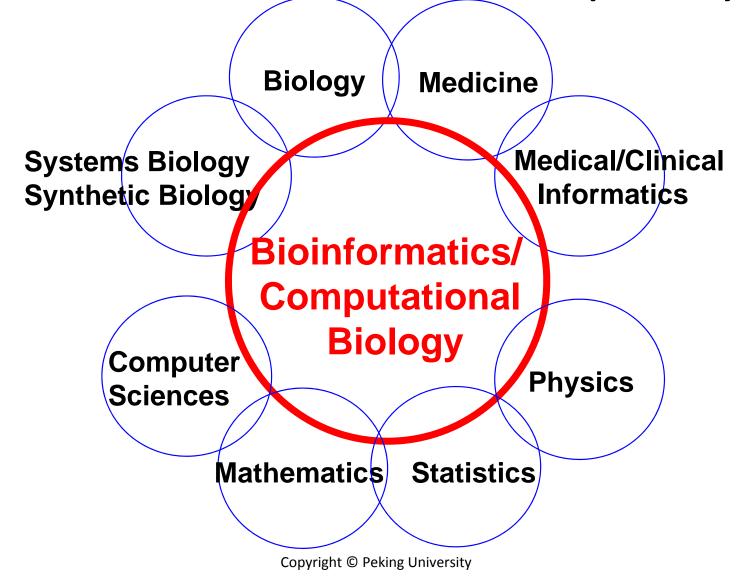
Bioinformatics

What is bioinformatics?

<u>Bioinformatics</u>: 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, datadriven, 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/ Proteins Molecular Cells Physiology/
Networks Disease

Sequence alignment
Database similarity search
Motif finding

Gene finding
Computational
& comparative
genomics
Evolution
DNA
methylation

Differential expression Co-expression ncRNA

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

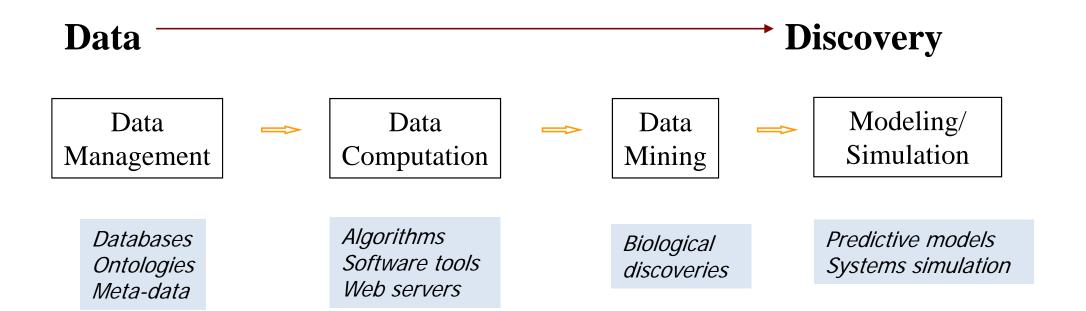
Population genetics

Human genetics

Mass spec protein identification
Structure prediction
Structure alignment

Virtual cell simulations

The -informatics in Bioinformatics



Summary Questions

- What is bioinformatics?
- Can you name some of the biomedical questions studied in bioinformatics?
- Can you name some of the technical questions studied in bioinformatics?
- Why are you interested in bioinformatics?

生物信息学:导论与方法 Bioinformatics: Introduction and Methods

