Bioinformatics

What is Bioinformatics?

Sami Khuri
Department of Computer Science
San José State University
San José, California, USA
sami.khuri@sjsu.edu
www.cs.sjsu.edu/faculty/khuri

What is Bioinformatics?

- The Human Genome Project (HGP)
- Mapping
- Model Organisms
- Types of Databases
- Applications of Bioinformatics
- Genome Research

Pathway to Genomic Medicine

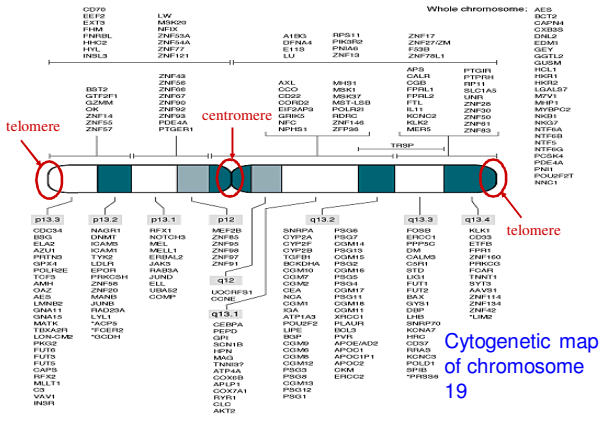
Human Genome Project	→ ENCODE Project	→ HapMap Project	→ Genomic Medicine
Sequencing of the human DNA	Interpreting the human genome sequence	Implicating genetic variants with human disease	Personalized medicine Cure for diseases

The Human Genome Project

- The **HGP** is a multinational effort, begun by the USA in 1988, whose aim is to produce a complete physical map of all human chromosomes, as well as the entire human DNA sequence.
- The ultimate goal of genome research is to find all the **genes** in the **DNA sequence** and to develop tools for using this information in the study of **human biology** and **medicine**.
- The primary goal of the project is to make a series of descriptive diagrams (called **maps**) of each human chromosome at increasingly finer resolutions.

Bioinformatics and the Internet

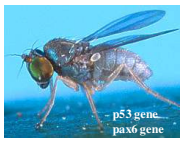
- The recent enormous increase in biological data has made it necessary to use computer information technology to collect, organize, maintain, access, and analyze the data.
- Computer speed, memory, exchange of information over the Internet has greatly facilitated bioinformatics.
- The bioinformatics tools available over the Internet are accessible, generally well developed, fairly comprehensive, and relatively easy to use.



Cytogenetic map of chromosome 19

Other Species

As part of the HGP, genomes of other organisms, such as bacteria, yeast, flies and mice are also being studied.



Baker's yeast

DNA repair
Cell division

Chimps are infected with SIV
Very rarely progress to AIDS

Other Sequenced Genomes

Model Organisms

- A **model organism** is an organism that is extensively studied to understand particular biological phenomena.
- **Why have model organisms?** The hope is that discoveries made in model organisms will provide insight into the workings of other organisms.
- **Why is this possible?** This works because evolution reuses fundamental biological principles and conserves metabolic, regulatory, and developmental pathways.

Studying Human Diseases

Organism	Human Diseases
E. coli	DNA repair; colon cancer and other cancers
Yeast	Cell cycle; cancer, Werner syndrome
Drosophila	Cell signaling; cancer
C. elegans	Cell signaling; diabetes
Zebrafish	Developmental pathways; cardiovascular disease
Mouse	Gene expression; Lesch-Nyhan disease, cystic fibrosis, fragile-X syndrome, and many other diseases

Goals of the HGP

- To ***identify*** all the approximately 20,000-25,000 genes in human DNA,
- To ***determine*** the sequences of the 3.2 billion chemical base pairs that make up human DNA,
- To ***store*** this information in databases,
- To ***improve*** tools for data analysis,
- To ***address*** the ethical, legal, and social issues (ELSI) that may arise from the project.

HGP Finished Before Deadline

- In 1991, the USA Congress was told that the HGP could be done by 2005 for $3 billion.
- It ended in 2003 for $2.7 billion, because of efficient computational methods.

What is Bioinformatics? Set of Tools

- The use of computers to collect, analyze, and interpret biological information at the molecular level.
- A set of software tools for molecular sequence analysis

What is Bioinformatics? A Discipline

- The field of science, in which **biology**, **computer science**, and **information technology** merge into a single discipline.

Definition of NCBI (National Center for Biotechnology Information)

- The ultimate goal of **bioinformatics** is to enable the discovery of new biological insights and to create a global perspective from which unifying principles in biology can be discerned.

What are Bioinformatics Tools Developed For?

- The three central biological processes around which bioinformatics tools are generally being developed:
  - **DNA sequence** determines **protein sequence**
  - **Protein sequence** determines **protein structure**
  - **Protein structure** determines **protein function**

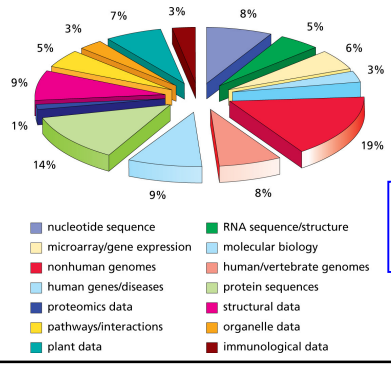
What do Bioinformaticians do?

- They analyze and interpret data
- Develop and implement algorithms
- Design user interface
- Design database
- Automate genome analysis
- They assist molecular biologists in data analysis and experimental design.

Databases for Storage and Analysis

- Databases store data that need to be analyzed
- By comparing sequences, we discover:
  - How organisms are related to one another
  - How proteins function
  - How populations vary
  - How diseases occur
- The improvement of sequencing methods generated a lot of data that need to be:
  - stored
  - organized
  - curated
  - annotated
  - managed
  - networked
  - accessed
  - assessed

Types of Databases



In 2006 there were 858 databases classified into 14 major categories

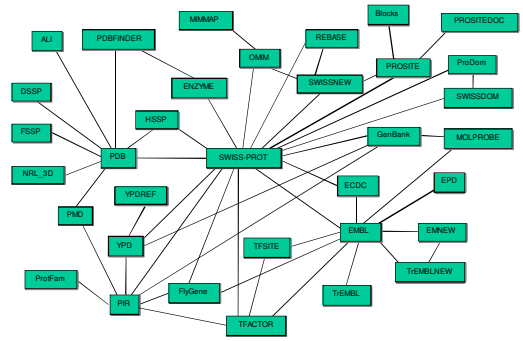
Three Major Databases

- GenBank from the NCBI (National Center of Biotechnology Information), National Library of Medicine http://www.ncbi.nlm.nih.gov
- EBI (European Bioinformatics Institute) from the European Molecular Biology Library http://www.ebi.ac.uk
- DDBJ (DNA DataBank of Japan) http://www.ddbj.nig.ac.jp

GenBank Taxonomic Sampling

Homo sapiens	62.1%
Mus musculus	7.7%
Drosophila melanogaster	6.1%
Caenorhabditis elegans	3.3%
Arabidopsis thaliana	2.9%
Oryza sativa	1.3%
Rattus norvegicus	0.8%
Danio rerio	0.6%
Saccharomyces cerevisiae	0.6%

Databanks Interconnection



The major DNA databases are updated and synchronized daily.

GenBank

GenBank is the NIH genetic sequence database of all publicly available DNA and derived protein sequences, with annotations describing the biological information these records contain.

What does NCBI do?

NCBI: established in 1988 as a national resource for molecular biology information.

- it creates public databases,
- it conducts research in computational biology,
- it develops software tools for analyzing genome data, and
- it disseminates biomedical information,

all for the better understanding of molecular processes affecting human health and disease.

Applications of Genome Research

Current and potential applications of Genome Research include:

- Molecular Medicine
- Microbial Genomics
- Risk Assessment
- Bioarcheology, Anthropology, Evolution and Human Migration
- DNA Identification
- Agriculture, Livestock Breeding and Bioprocessing

Molecular Medicine

- Improve the **diagnosis** of disease
- Detect genetic **predispositions** to disease
- Create drugs **based on molecular information**
- Use **gene therapy** and control systems as drugs
- Design **custom drugs** on individual genetic profiles.

Microbal Genomics

- Swift detection and treatment in clinics of disease-causing microbes: pathogens
- Development of new energy sources: biofuels
- Monitoring of the environment to detect chemical warfare
- Protection of citizens from biological and chemical warfare
- Efficient and safe clean up of toxic waste.

DNA Identification I

- Identify potential suspects whose DNA may match evidence left at crime scenes
- Exonerate persons wrongly accused of crimes
- Establish paternity and other family relationships
- Match organ donors with recipients in transplant programs

Louis XVII

Louis XVII: son of Louis XV1 and Marie-Antoinette who died from tuberculosis in 1795 at the age of 12

DNA and Human Trafficking

13 Haitian Children Returned To Their Families Thanks To DNA Analyses: DNA-Prokids Bolivia

Natural disasters frequently turn into human tragedies, such as family separations. The Haiti earthquake of January 12, was followed by emotive worldwide solidarity actions. But this can not outshine extremely serious incidents, like the fact that the human trafficking mafias could take advantage of the catastrophe to get children off the island.

Last January, more than seventy people from Haiti arrived at Santa Cruz de la Sierra (Bolivia), via Lima. Visa problems stopped them on their way to Brazil or Argentina. Bolivian Police suspicions opened a deep investigation and proved that the 25 Haitian children in the group were not accompanied by their relatives. In February, their families in Haiti started to look for them.

The Bolivian Attorney General's Office requested the collaboration of the Laboratory of Forensic Genetics of the Bolivia Forensic Research Institute, which applied the DNA-Prokids action protocol. The genetic research results were unquestionable: eight parents (seven mothers and a father) looking for their 13 children have recovered them, thanks to the DNA identification (two mothers looked for two children each, a mother looked for three children, four mothers looked for a child each, a father looked for two children).

From Haiti to Bolivia

DNA Identification II

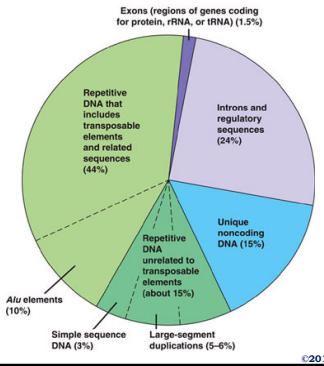
- Identify endangered and protected species as an aid to wildlife officials and also to prosecute poachers
- Detect bacteria and other organisms that may pollute air, water, soil, and food
- Determine pedigree for seed or livestock breeds
- Authenticate consumables such as wine and caviar

Agriculture, Livestock Breeding and Bioprocessing

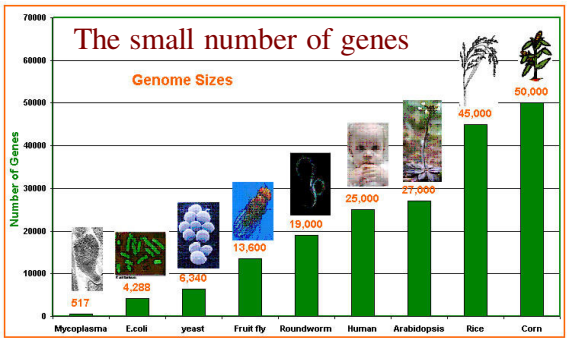
- Grow disease-resistant, insect-resistant, and drought-resistant crops
- Breed healthier, more productive, disease-resistant farm animals
- Grow more nutritious produce
- Develop biopesticides
- Incorporate edible vaccines into food products

What have we learned from the HGP?

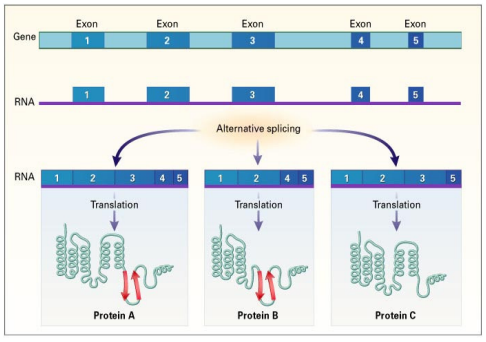
A small portion of the genome codes for proteins, tRNAs and rRNAs



What have we learned from the HGP?

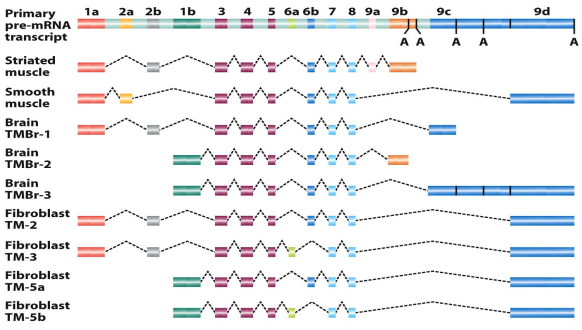


Alternative Splicing



Genomic Medicine by Guttmacher et al., NEJM, 2002

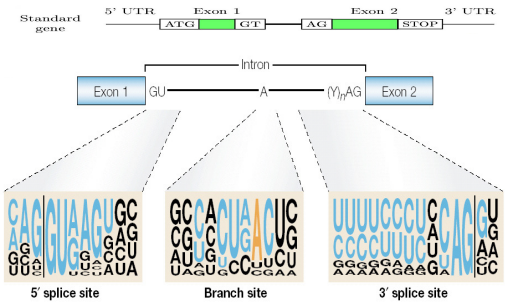
The Alpha-Tropomyosin Gene



Gene Prediction

- **Problem**: Given a genomic DNA sequence, identify where the **genes** are.
- **Input**: A genomic DNA sequence.
- **Output**: Location of **gene elements** in the raw, genomic DNA sequence, including (for eukaryotes):
  - **exons**
  - **introns**

Anatomy of an Intron



Convert all this progress into real riches for science, society, and patients

Concluding Remarks

- Biology is becoming an information science
- Progression: **in vivo** to **in vitro** to **in silico**
- Are natural languages adequate in predicting quantitative behavior of biological systems?
  - Need to produce biological knowledge and operations in ways that natural languages do not allow
- Today's biologists need to think quantitatively and from a multidisciplinary perspective.
- Today's biology courses need to cast a wide net to capture the imaginations of students representing many different interests, skills, and viewpoints.