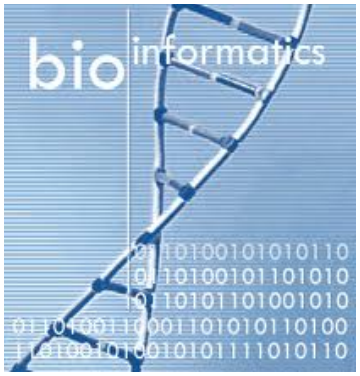
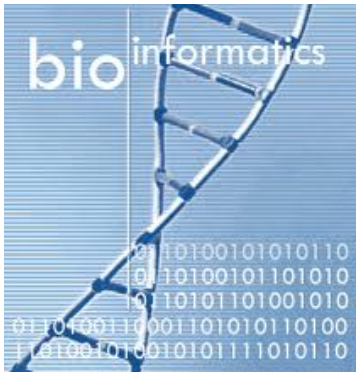
Bioinformatics

TWO

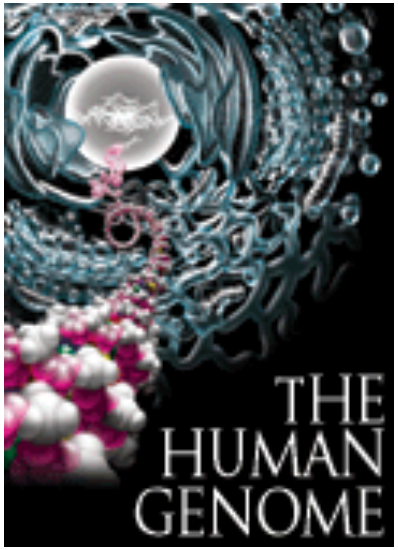
Introduction to Bioinformatics

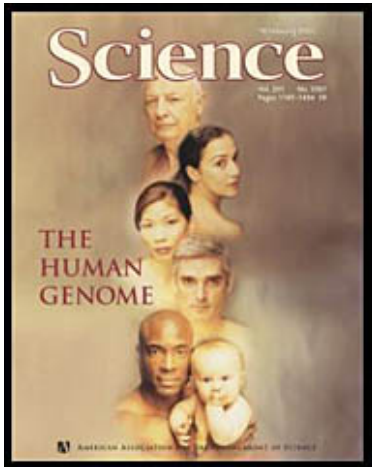
Wendy Lee
Dept of Computer Science
San José State University
Biology/CS/SE 123A
Fall 2014





What is Bioinformatics?





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

From the Preface

- We believe that to perform a proper analysis it is not sufficient to understand how to use a program and the kind of results (and errors!) it can produce.
- It is of also necessary to have some understanding of the technique used by the program and the science on which it is based.

Preface and Note to the Reader

- All research workers in the areas of biomolecular science and biomedicine are now expected to be competent in several areas of sequence analysis and often, additionally, in protein structure analysis and other more advanced bioinformatics techniques.
- The book is designed to be accessible both to students who wish to obtain a working knowledge of the bioinformatics applications, as well as to students who want to know how the applications work and maybe write their own.

The Human Genome Project

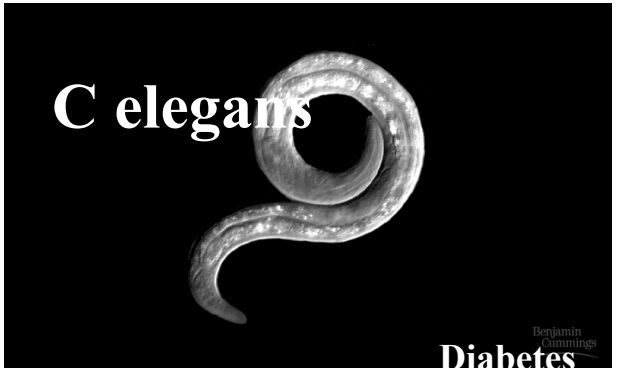
- The **HGP** is a multinational effort, begun by the USA in 1988, whose aim is to produce a complete physical map of all human chromosomes, as well as the entire human DNA sequence.
- The ultimate goal of genome research is to find all the **genes** in the **DNA sequence** and to develop tools for using 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.

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

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.

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.

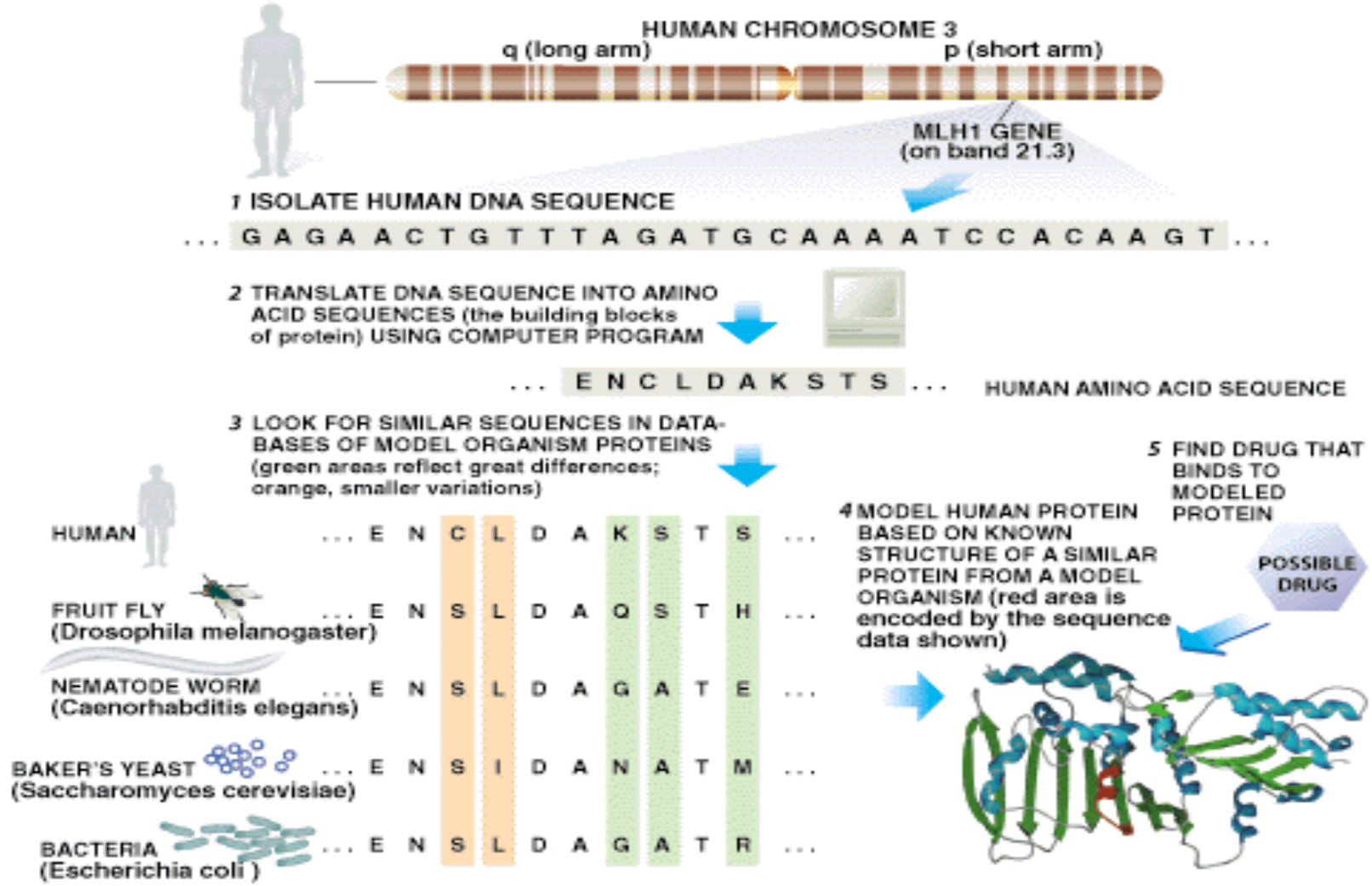
Definition of NCBI (National Center for Biotechnology Information)

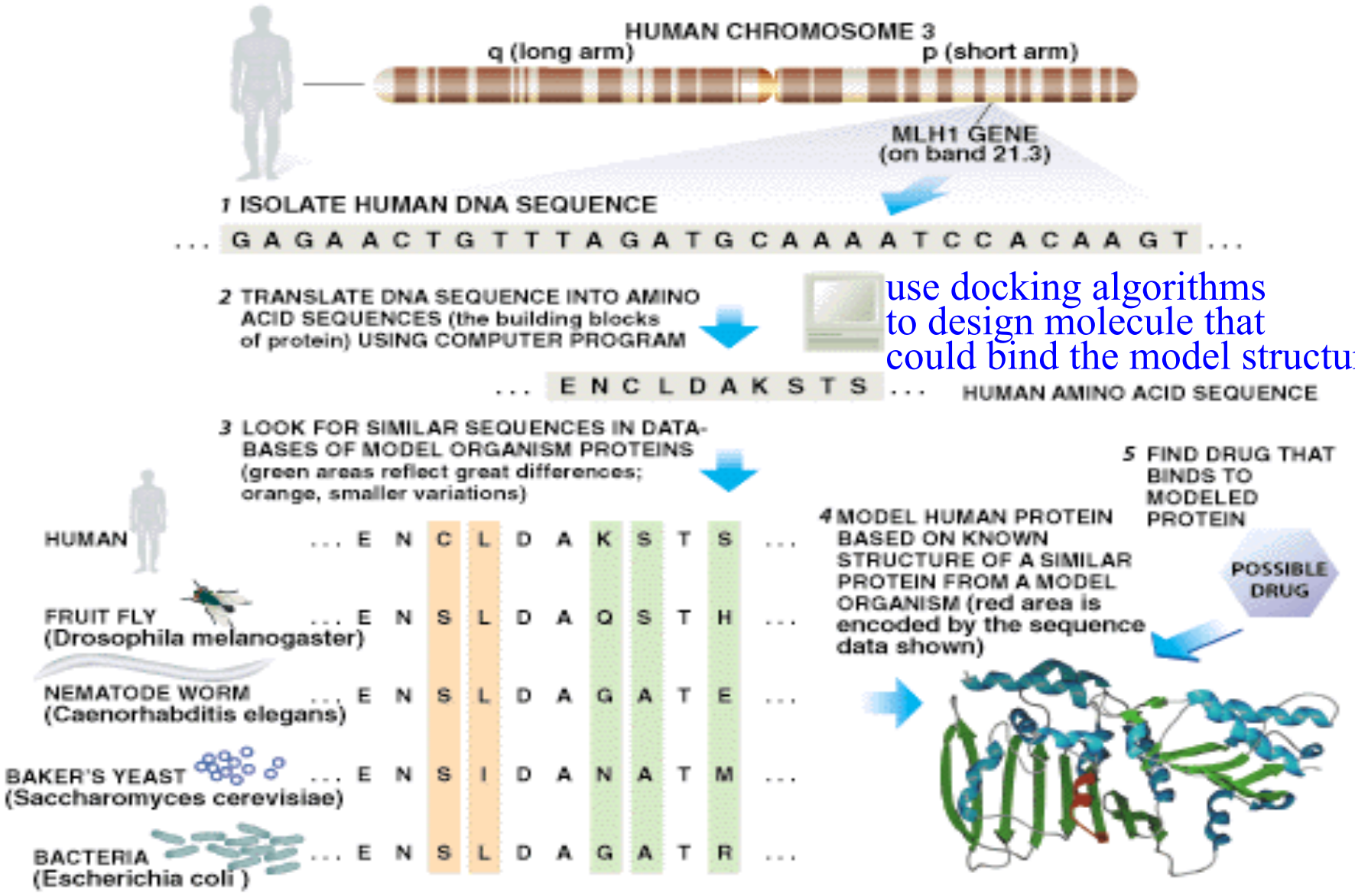
- The ultimate goal of **bioinformatics** is to enable the discovery of new biological insights and to create a global perspective from which unifying principles in biology can be discerned.

Why Study Bioinformatics (I)

- Bioinformatics is intrinsically interesting.
- Bioinformatics offers the prospect of finding better drug targets earlier in the drug development process.
  - By looking for genes in model organisms that are similar to a given human gene, researchers can learn about protein the human gene encodes and search for drugs to block it.

How can Bioinformatics Help?





Rational drug design
Structure-based drug design

Scientific American July 2000

Why Study Bioinformatics (II)

- Molecular biology is the new frontier of 21st century science.
  - DNA, RNA, genes, stem cells, etc.. are everywhere in the news.
- Science Magazine celebrated its 125th anniversary by issuing twenty five big questions facing science over the next quarter-century.

www.sciencemag.org/sciext/125th

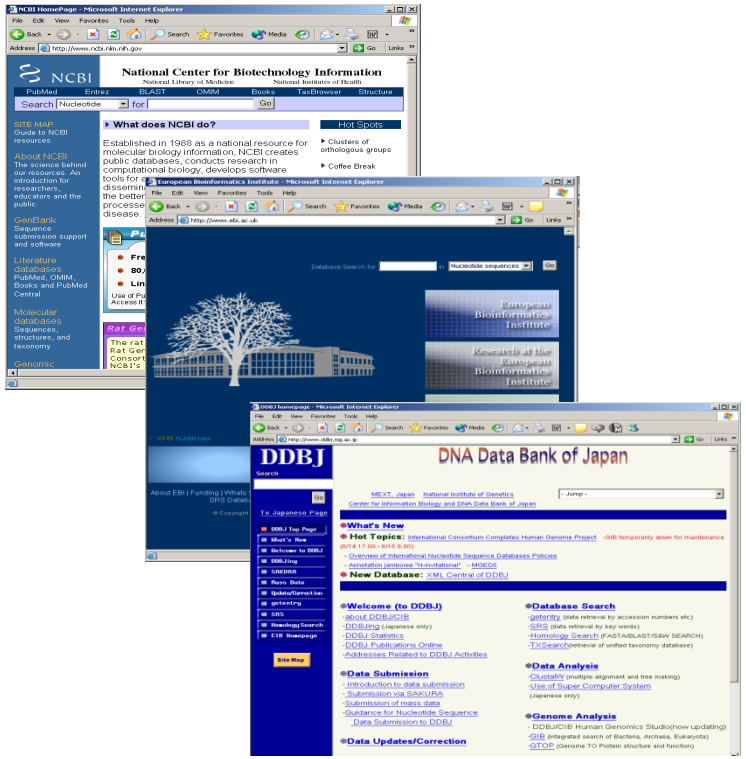
What do Bioinformaticians do?

- 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
  - annotated
  - accessed
  - organized
  - managed
  - assessed
  - curated
  - networked

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

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

DNA Identification II

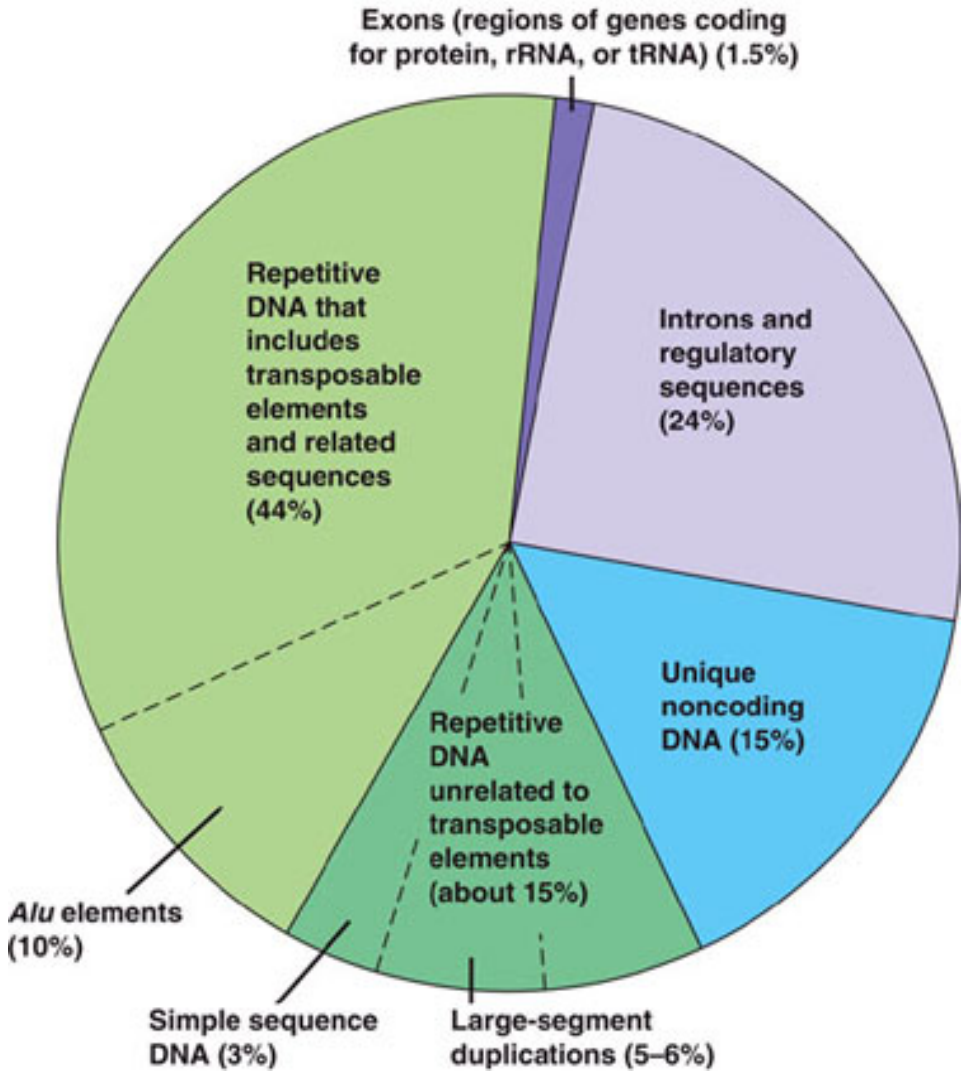
- Identify endangered and protected species as an aid to wildlife officials and also to prosecute poachers
- Detect bacteria and other organisms that may pollute air, water, soil, and food
- Determine pedigree for seed or livestock breeds
- Authenticate consumables such as wine and caviar

Agriculture, Livestock Breeding and Bioprocessing

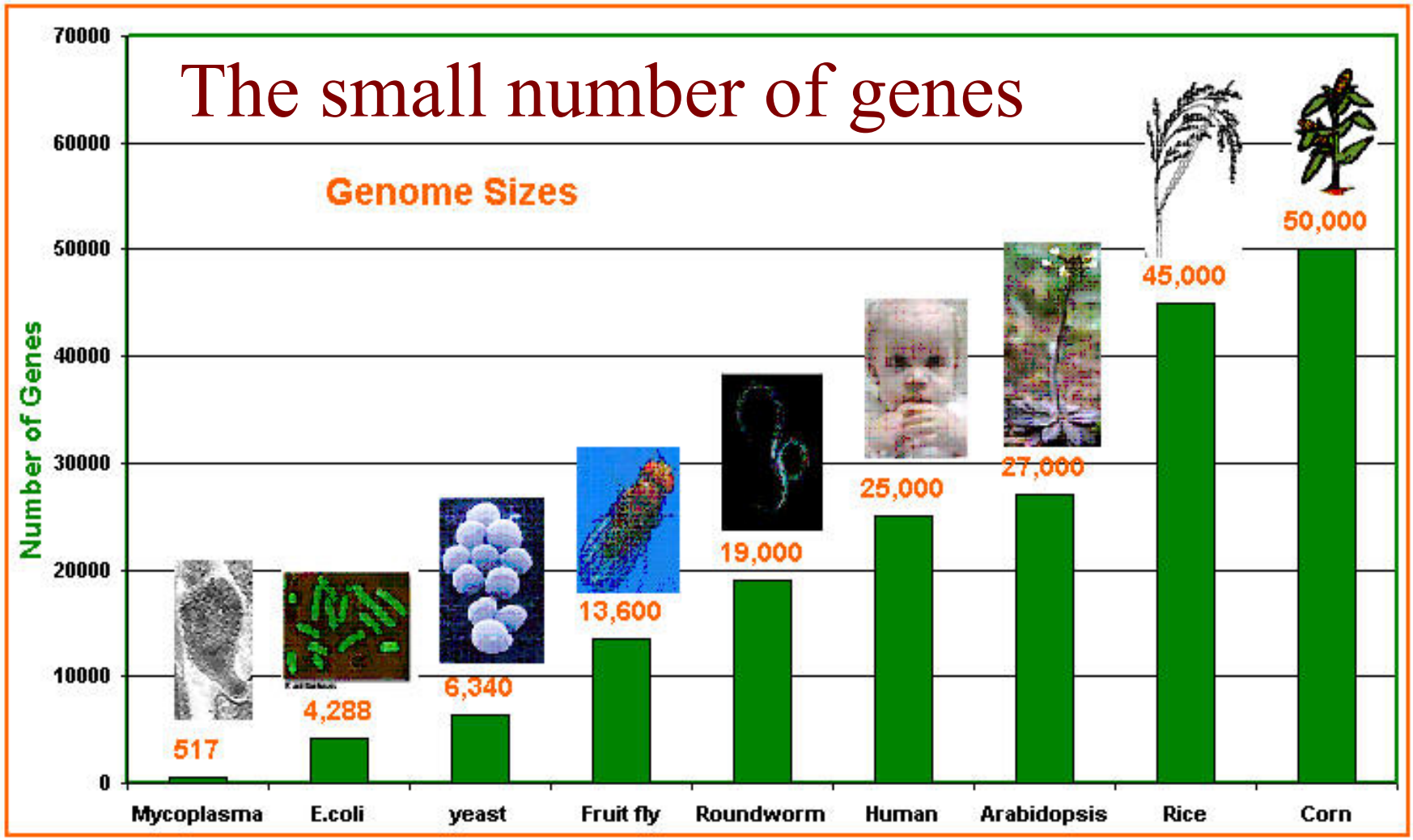
- Grow disease-resistant, insect-resistant, and drought-resistant crops
- Breed healthier, more productive, disease-resistant farm animals
- Grow more nutritious produce
- Develop biopesticides
- Incorporate edible vaccines into food products

What have we learned from the HGP?

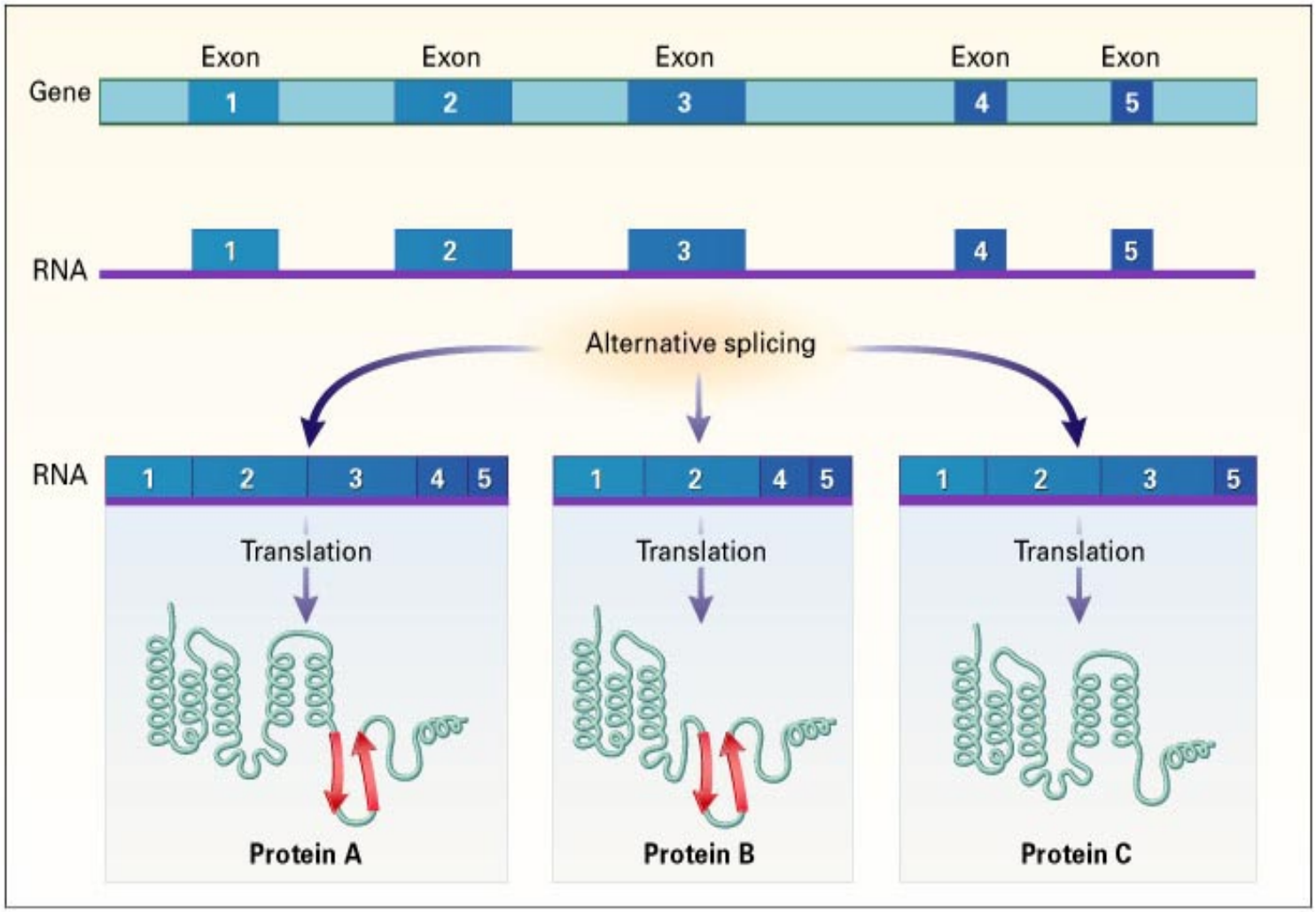
A small portion of the genome codes for proteins, tRNAs and rRNAs



What have we learned from the HGP?



Alternative Splicing



Genomic Medicine by Guttmacher et al., NEJM, 2002

Concluding Remarks (I)

- 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
- "Biology easily has 500 years of exciting problems to work on". Donald Knuth
- We are here to add what we can *to*, not to get what we can *from*, Life. William Osler

Concluding Remarks (II)

- Today's biology courses need to cast a wide net to capture the imaginations of students representing many different interests, skills, and viewpoints.
- Today's biologists need to think quantitatively and from a multidisciplinary perspective.
- The role that mathematics and computer science are playing in biology is still in an embryonic stage.