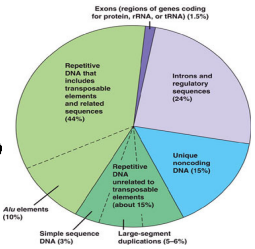
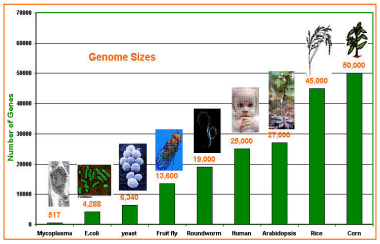
Introduction to Bioinformatics

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

What is Bioinformatics?





- The Human Genome Project (HGP)

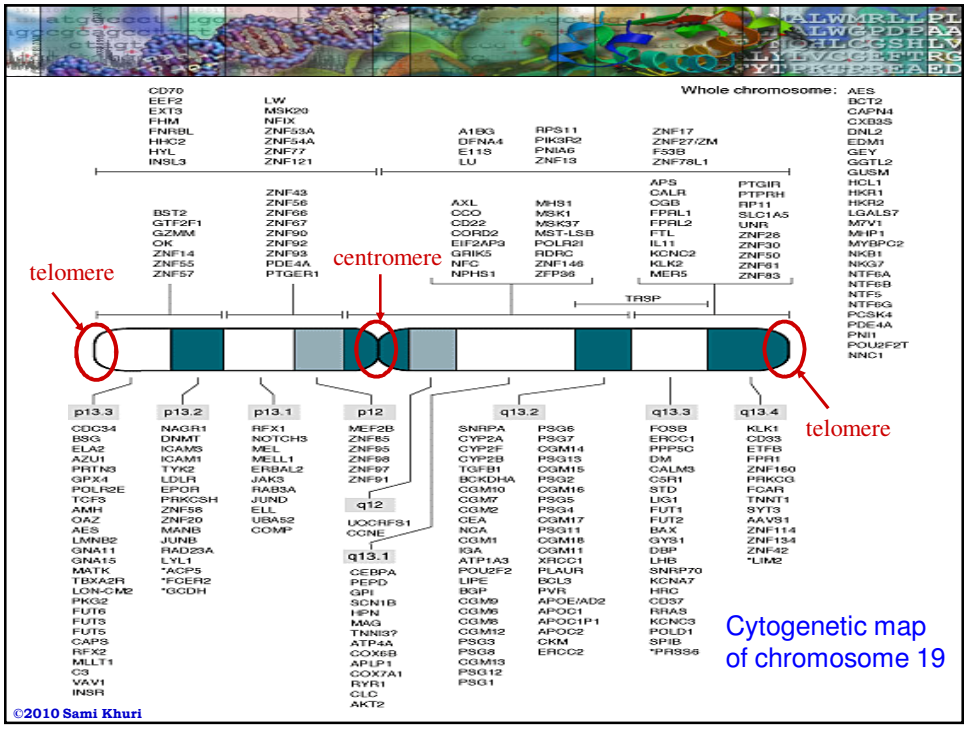
- Mapping
- Model Organisms
- Types of Databases
- Applications of Bioinformatics
- Genome Research

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.
  - As part of the project, genomes of other organisms such as bacteria, yeast, flies and mice are also being studied.
- The primary goal of the project is to make a series of descriptive diagrams (called **maps**) of each human chromosome at increasingly finer resolutions.

The HGP Goal

- 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**.
- **Mapping** involves:
  - dividing the chromosomes into smaller fragments that can be propagated and characterized
  - ordering (mapping) them to correspond to their respective locations on the chromosomes.



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.

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

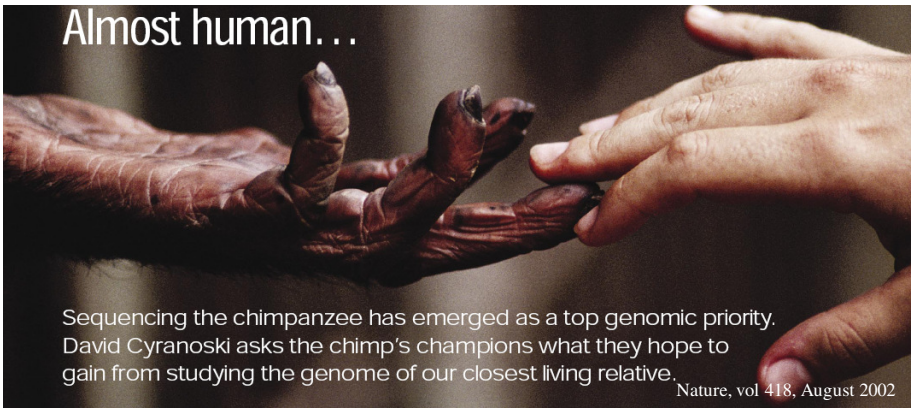
Model Organisms

- 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.

Name	Genome BP	Genes	Chromosomes
HSV1 (Herpes virus) 	1.5×10^5	70	1
Escherichia Coli 	4.6×10^6	4,300	1
Saccharomyces cerevisiae 	1.2×10^7	5,900	16
Caenorhabditis Elegans 	1.0×10^8	19,100	6
Drosophila melanogaster 	1.8×10^8	13,600	6
Arabidopsis Thalania 	1.2×10^8	25,500	5
Mus Musculus 	2.5×10^9	?30,000	20+X/Y
Homo sapiens 	2.9×10^9	?30,000	22+X/Y

David Gilbert

Sequencing the Chimpanzee



Studying Diseases

Comparison of disease susceptibility between chimps and humans

	Condition	*Human*	*Chimp*
Definite differences			
	HIV progression to AIDS	Common	Very rare
	Influenza A symptoms	Moderate/severe	Mild
	Hepatitis B/C late complications	Moderate/severe	Mild
	Plasmodium falciparum malaria	Susceptible	Resistant
	Menopause	Universal	Rare
Probable differences			
	Escherichia coli K99 gastroenteritis	Resistant	Sensitive?
	Alzheimer's disease pathology	Complete	Incomplete
	Coronary atherosclerosis	Common	Uncommon
	Epithelial cancers	Common	Rare

Source: Olson, M. V. *et al. A white paper advocating complete sequencing of the genome of the common chimpanzee,* Pan troglodytes. (2002).

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

F	W	Y	Neurological
+			Alzheimer-PS1
+			Alzheimer-APP
–			Creutzfeldt-Jakob-PRNP
+			Deafness, Hereditary-MYO15
+			Dementia, Multi-Infarct-NOTCH3
+			Duchenne MD^{+}-DMD
+			Fragile-X-FRAXA
+			Huntington-HD
+			Limb Girdle MD^{+}2A-CAPN3
+			Limb Girdle MD^{+}2B-YSF
–			Limb Girdle MD^{+}2E-BSG
+			Myotonic Dystrophy-DM1
+			Myotubular Myopathy 1-MTM1
–			Parkinson-SNCA
+			Parkinson-PARK2
+			Parkinson-UCHIL1
+			Tay-Sachs-HEXA

F	W	Y	Immune
+			Bruton Agammaglobulin.-BTK
+			Chronic Granulom.-CYBB
+			Immunodeficiency-DNA Ligase 1
–			Immunodeficiency-CD3G
+			SCID**-JAK3
–			SCID**-RAG1
–			SCID**-RAG2
+			SCID**-ZAP70

F	W	Y	Cardiovascular
+			Fam. Cardiac Myopathy-MYH7
+			HDL Deficiency 1-ABCA1

F	W	Y	Birth Defects
+			Holoprosencephaly 3-SHH
+			Holoprosencephaly-SIX3
+			Zellweger-PEX1

F	W	Y	Renal
+			Diabetes Insipidus 2-AQP2
–			Polycystic Kidney 1-PKD1
+			Polycystic Kidney 2-PKD2

F	W	Y	Endocrine
+			Diabetes-INS
+			Diabetes-INSR
+			Hyperinsulinism-ABCC8
+			Hyperinsulinism-KCNJ11
–			Obesity-LEP
–			Obesity-LEPR
+			Vitamin-D Resis. Rickets-VDR

F	W	Y
–		
–		
–		
–		
+		
+		

α-Thalassemia-HBA1
β-Thalassemia-HBB
δ-Thalassemia-HBD
ε-Thalassemia-HBE
Thrombophilia-PLG
Wiskott-Aldrich-WAS

F	W	Y	Metabolic
+			Cystinuria, Type 1-SLC3A1
–			Hypercalcemia-CASR
+			Niemann-Pick C-NPC1
+			SCID**-ADA

F	W	Y	Other
+			Cystic Fibrosis-ABCC7
+			Hereditary Pancreatitis-PRSS1
+			Juvenile Glaucoma-GLC1A
+			Wolfram-WFS1

- E-values $<1\times10^{-100}$
- E-values of 1×10^{-40} to 1×10^{-100}
- E-values of 1×10^{-6} to 1×10^{-40}
- E-values $>1\times10^{-6}$

Flies have **orthologs** to humans disease-causing genes in categories such as:

- neurological
- renal
- immunological
- endocrine
- cardiovascular
- metabolic
- blood-vessel and
- cancerous disorders

Flies can provide insights into human disease at the **systems level**, revealing how different genes interact in vivo

Discovering Genomics, Campbell, 2007

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.

Bioinformatics and the Internet

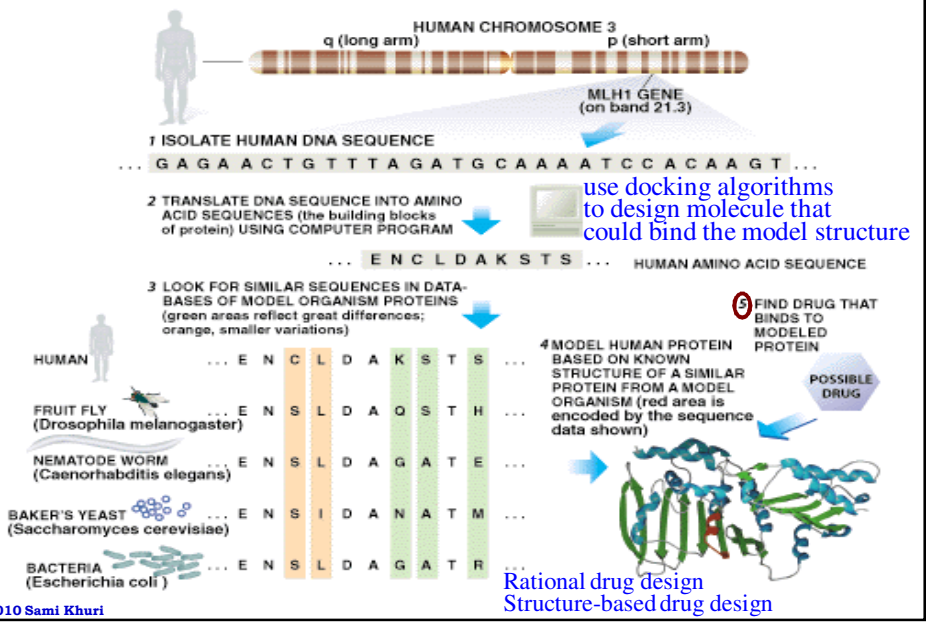
- The enormous increase in biological data has made it necessary to use computer information technology to collect, organize, maintain, access, and analyze the data.
- Computer speed, memory, and exchange of information over the Internet has greatly facilitated bioinformatics.
- The bioinformatics tools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

What do Bioinformaticians do?

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

Why Study Bioinformatics?

- 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.



Databases for Storage and Analysis

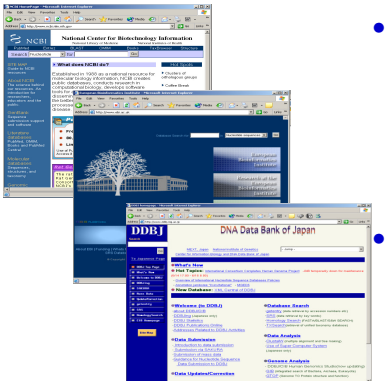
- Databases store data that need to be analyzed
- By comparing sequences, we discover:
  - How organisms are related to one another
  - How proteins function
  - How populations vary
  - How diseases occur
- The improvement of sequencing methods generated a lot of data that need to be:

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

Types of Databases

- **Sequence**
  - Genbank, SwissProt, 3D structure, carbohydrates, organism specific, phylogenetic, sequence patterns
- **Literature**
  - Medline, OMIM, Patents, eJournals
- **Graphical**
  - Swiss2D-Page
- **Expression Analysis Databases**
  - Microarrays
- **Protein Interaction Databases**
  - Pathways

Three Major Databases

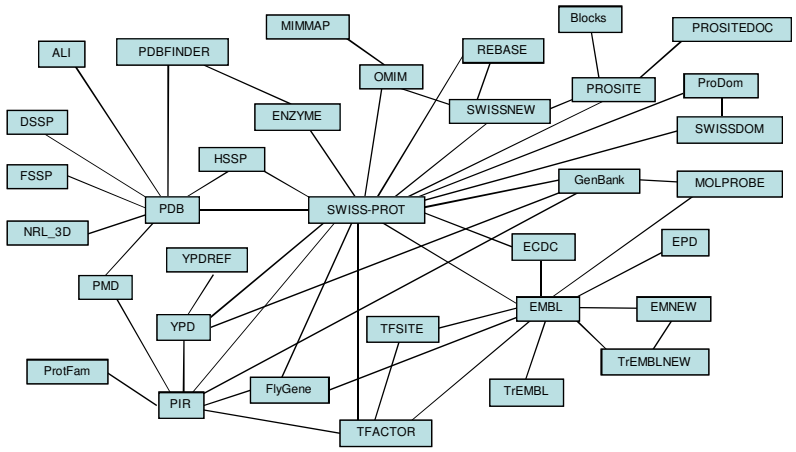


- GenBank from the NCBI (National Center of Biotechnology Information), National Library of Medicine http://www.ncbi.nlm.nih.gov
- EBI (European Bioinformatics Institute) from the European Molecular Biology Library http://www.ebi.ac.uk
- DDBJ (DNA DataBank of Japan) http://www.ddbj.nig.ac.jp

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.

The Annotation Challenge

- One of the biggest challenges for maintainers of biological databases is the **annotation**:
  - putting sufficient **information** in the database such that there is no question of what the gene is,
  - creating the proper **links** between that information and the gene sequence and serial number.
- Correct **annotation** of genomic data is an active area of research.

Data Formats

- Most databases are relational DB. The format can be changed to export the data:
  - GenBank data is stored by NCBI in a Sybase database, but made public in a flatfile format.
- Each database and each sequence analysis program store/accept data in a different format.
  - The most commonly used data format is FASTA

>gi|37222328|gb|AY350722.1| Pan troglodytes masticatory myosin heavy chain (MYH16)

GAGCAGCTGAACAAGCTGATGACCACCCTCCATAGCACCGCACCCCATTTTG
TCCGCTGTATTATCCCCAATGAGTTTAAGCAATCGG

Reasons for Searching Databases

Searching a database can answer the following questions:

- A researcher has just sequenced a gene.
  Has someone already found it?
- A researcher has a sequence of unknown function.
  Is there a homology with another sequence that has a known function?
- A researcher has found a new protein in a lower organism.
  Is there a homology in a higher species?

Protein Databases

- GenPept from NCBI.
- ExPASy (Expert Protein Analysis System)
  http://www.expasy.ch
- SwissProt, TrEMBL
  http://www.ebi.ac.uk
- PIR (Protein Identification Resource)
  http://www-nbrf.georgetown.edu/pirwww
- DISC - DNA Information and Stock Center, Japan
  http://www.dna.affrc.go.jp

SwissProt and SRS

- The SwissProt protein sequence database is at ISREC (Swiss Institute for Experimental Cancer Research) in Epalinges, Lausanne.
- The Sequence Retrieval System (SRS) at the European Bioinformatics Institute:
  - allows both simple and complex concurrent searches of one or more sequence databases,
  - may also be used on a local machine to assist in the preparation of local sequence databases.

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.

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.

GenBank DNA Sequence Entry I

The main important fields of an entry are:

- **Locus:** name of locus, length and type of sequence, classification of organism, date of entry. Not maintained among other databases
- **Definition:** description of entry
- **Accession:** accession number of original source. A citable entity; does not change when record is updated.

GenBank DNA Sequence Entry II

- **Keywords:** keywords for cross referencing this entry
- **Source:** source organism of DNA
- **Organism:** description of organism
- **Comment:** biological function or database information
- **Features:** information about sequence by base position or range of positions

GenBank DNA Sequence Entry III

- **Base count:** count of A, C, G, T and other symbols
- **Origin:** text indicating the start of the sequence.

  The sequence entry is assumed by all computer programs to lie between the identifiers **ORIGIN** and **//**.

Division of Organisms

BCT	Bacterial	PLN	Plant
FUN	Fungal	PRI	Primate
HUM	Homo sapiens	PRO	Prokaryotic
INV	Invertebrate	ROD	Rodent
MAM	Other mammalian	SYN	Synthetic & chimeric
ORG	Organelle	VRL	Viral
PHG	Phage	VRT	Other vertebrate

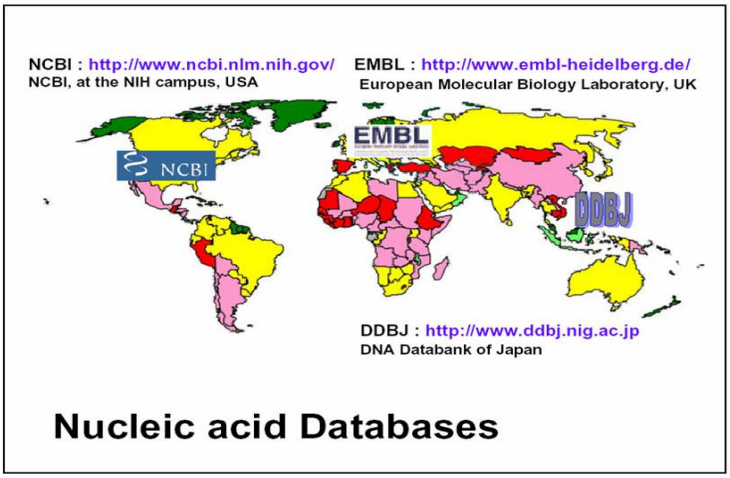
Interesting Databases

- UCSC Human Genome Browser
  - http://genome.ucsc.edu/
- Organism specific information:
  - Yeast: http://genome-www.stanford.edu/Saccharomyces/
  - Arabidopis: http://www.tair.org/
  - Mouse: http://www.jax.org/
  - Fruit fly: http://www.fruitfly.org/
  - Nematode: http://www.wormbase.org/

European Molecular Biology Laboratory

- The **European Molecular Biology Laboratory** (**EMBL**) was established in 1974.
- It is supported by sixteen countries.
- EMBL consists of five facilities:
  - The main Laboratory in Heidelberg (Germany),
  - Outstations in Hamburg (Germany), Grenoble (France) and Hinxton (the U.K.), and an external Research Programme in Monterotondo (Italy).

NCBI – EMBL - DDJB



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.

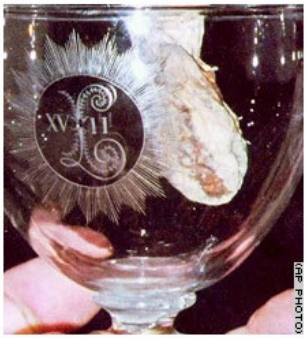
Microbial Genomics

- 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 Identification II

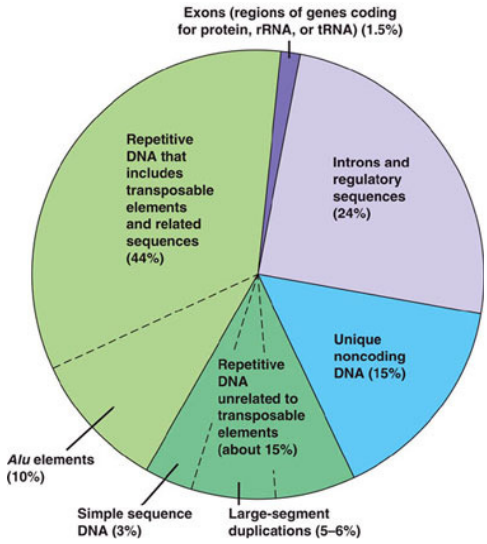
- Identify endangered and protected species as an aid to wildlife officials and also to prosecute poachers
- Detect bacteria and other organisms that may pollute air, water, soil, and food
- Determine pedigree for seed or livestock breeds
- Authenticate consumables such as wine and caviar

Agriculture, Livestock Breeding and Bioprocessing

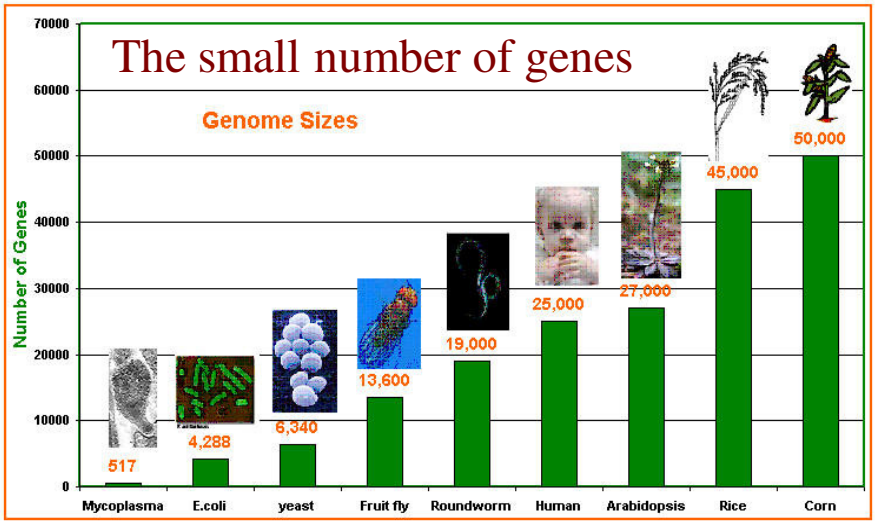
- Grow disease-resistant, insect-resistant, and drought-resistant crops
- Breed healthier, more productive, disease-resistant farm animals
- Grow more nutritious produce
- Develop biopesticides
- Incorporate edible vaccines into food products

What have we learned from HGP?

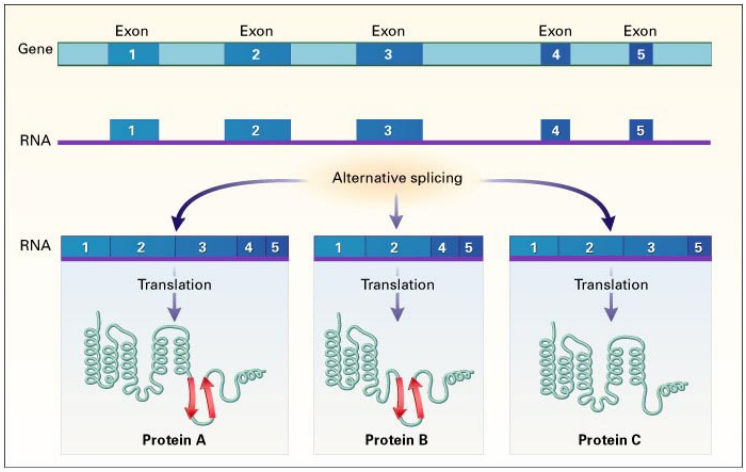
A small portion of the genome codes for proteins, tRNAs and rRNAs



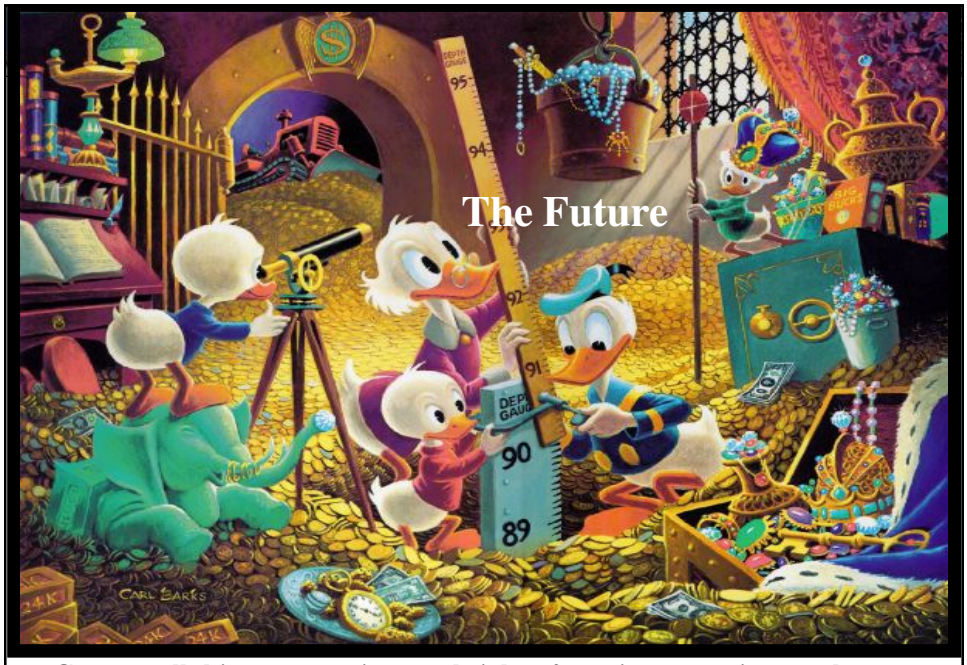
What have we learned from HGP?



Alternative Splicing



Genomic Medicine by Guttmacher et al., NEJM, 2002



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

Objectives of Molecular Biology

- Extract the information in the genomes.
- Understand the structure of the genome.
- Apply this understanding to the diagnosis and treatment of genetic diseases.
- Explain the process of evolution by comparing genomes of related species.

Goals of Modern Molecular Biology

- Read the entire genomes of living things
- Identify every gene
- Match each gene with the protein it encodes
- Determine the structure and function of each protein.

Objectives of Bioinformatics

Development and use of mathematical and computer science techniques to help solving the problems in molecular biology.

Bioinformatics Problems

- Reconstructing long DNA sequences from overlapping string fragments.
- Comparing two or more sequences for similarities.
- Storing, retrieving and comparing DNA sequences and subsequences in databases.
- Exploring frequently occurring patterns of nucleotides.
- Finding informative elements in protein and DNA sequences.
- Finding evolutionary relationships between organisms.

Main Aim of the Problems

- The aim of these problems is to learn about the functionality and/or the structure of protein without actually having to physically construct the protein itself.
- The research is based on the assumption that similar sequences produce similar proteins.

Functional: Coding v/s Noncoding

	Coding Sequence (Genes)	Non-Coding Sequence
Identifying Computational Tools	Relatively Easy Improving Tools	Very Hard Poor predictive tools
Signals What to look for	We Have a Good Understanding	Very little is known
Complementary data we can use	Available – Ex. ESTs and cDNAs	Unavailable

Post Human Genome Project

- Major role for comparative sequence analysis will be the identification of functionally important, non-coding sequences.
- Need to study the relation between Sequence Conservation and Sequence Function.
- Focus on the interpretation of the human genome.
- Learn the functional landscape of the human genome.
- **Challenge**: go from sequence to function
  - i.e., define the role of each gene and understand how the genome functions as a whole.

Impact on Other Fields of Science

- Mathematical Models
- Visualization/Animation of molecular structures
- Visualization/Animation of algorithms
- More sophisticated databases
- Data mining
- Statistical analysis of results
- Metabolic pathways
- Predictive algorithms