# Computational Structural Bioinformatics

ECS129

Instructor: Patrice Koehl

http://koehllab.genomecenter.ucdavis.edu/teaching/ecs129
koehl@cs.ucdavis.edu

# Learning curve

| | Math / CS | Biology/ Chemistry |
|---|---|---|
| Pre-requisite | - Need to be able to access the web, to read and print PDF files<br>- Basic knowledge of statistics, probability | Molecules, basic cell biology |
| What you will learn | -Optimal alignment of two strings<br>-Shape descriptors<br>- Visualize and manipulate protein structures | -Interactions between molecules<br>-Protein families<br>-Structure prediction<br>-Use of bioinformatics databases and resources |
| Not necessary | -How to solve the Poisson Boltzmann equation<br>-Design a hashing function | Taxonomy, E.coli is a gram negative bacterium, organisation of protein kinases |

# Science, then, and now…

At the beginning,
there were
thoughts,
and
observation….

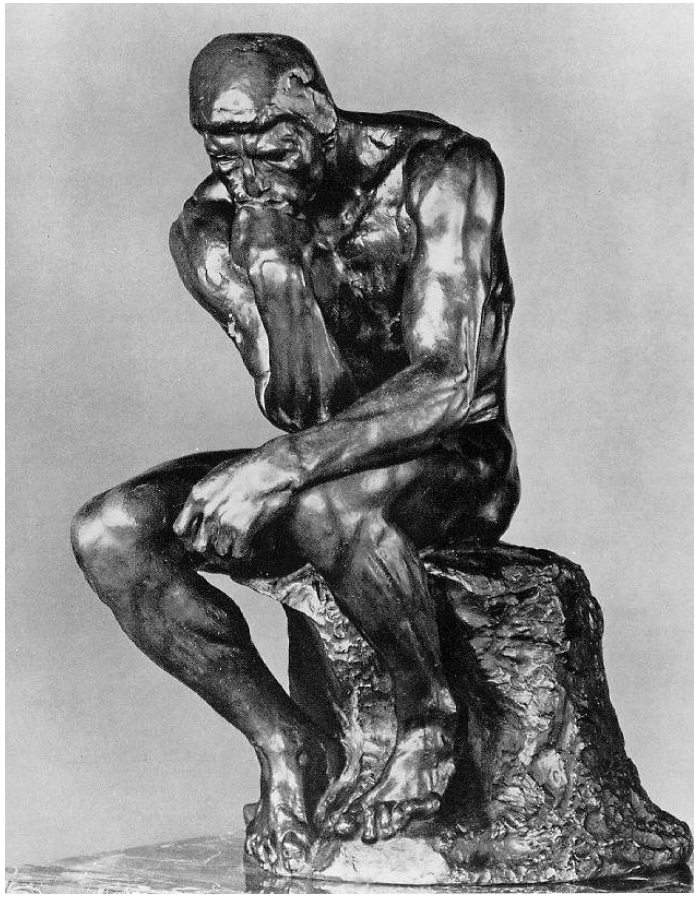
# Science, then, and now…

- For a long time, people thought that it would be enough to reason about the existing knowledge to explore everything there is to know.

- One single person could possess all knowledge in her cultural context.

  (encyclopedia of Diderot and D'Alembert)

- Reasoning, and mostly passive observation were the main techniques in scientific research



# Science, then, and now…

*"All science is either physics, or stamp collecting"*

Rutherford, chemist and physicist, 1876-1937

# Science, then and now

- Today's experiment yields massive amounts of data
- From hypothesis-driven to exploratory data analysis:
  - data are used to formulate new hypotheses
  - computers help formulate hypotheses
- No single person, no group has an overview of what is known

# Context: Biology

- "Life sciences" have their origins in ancient Greece

  Aristotle wrote influential treatises on zoology, anatomy and botany, that remained influential till the Renaissance

- "Life sciences" have always relied both on observation and discovery

  taxonomy, classifications, theory of evolution,…

- Biology is changing with the arrival of massive amount of data from the different genomics experiments

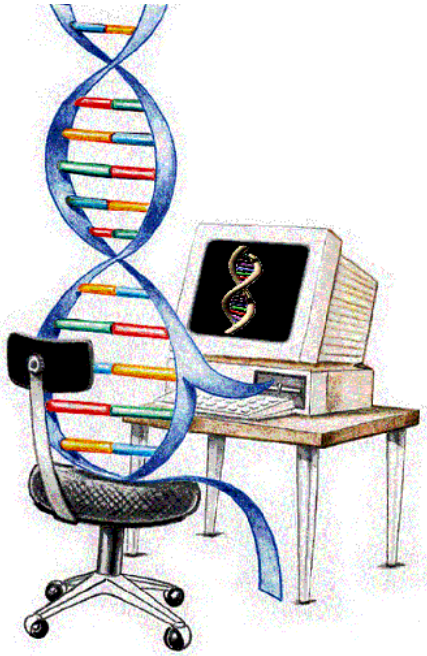
# What is ‘bioinformatics’?

- The term was originally proposed in 1988 by Dr. Hwa Lim

- The original definition was :

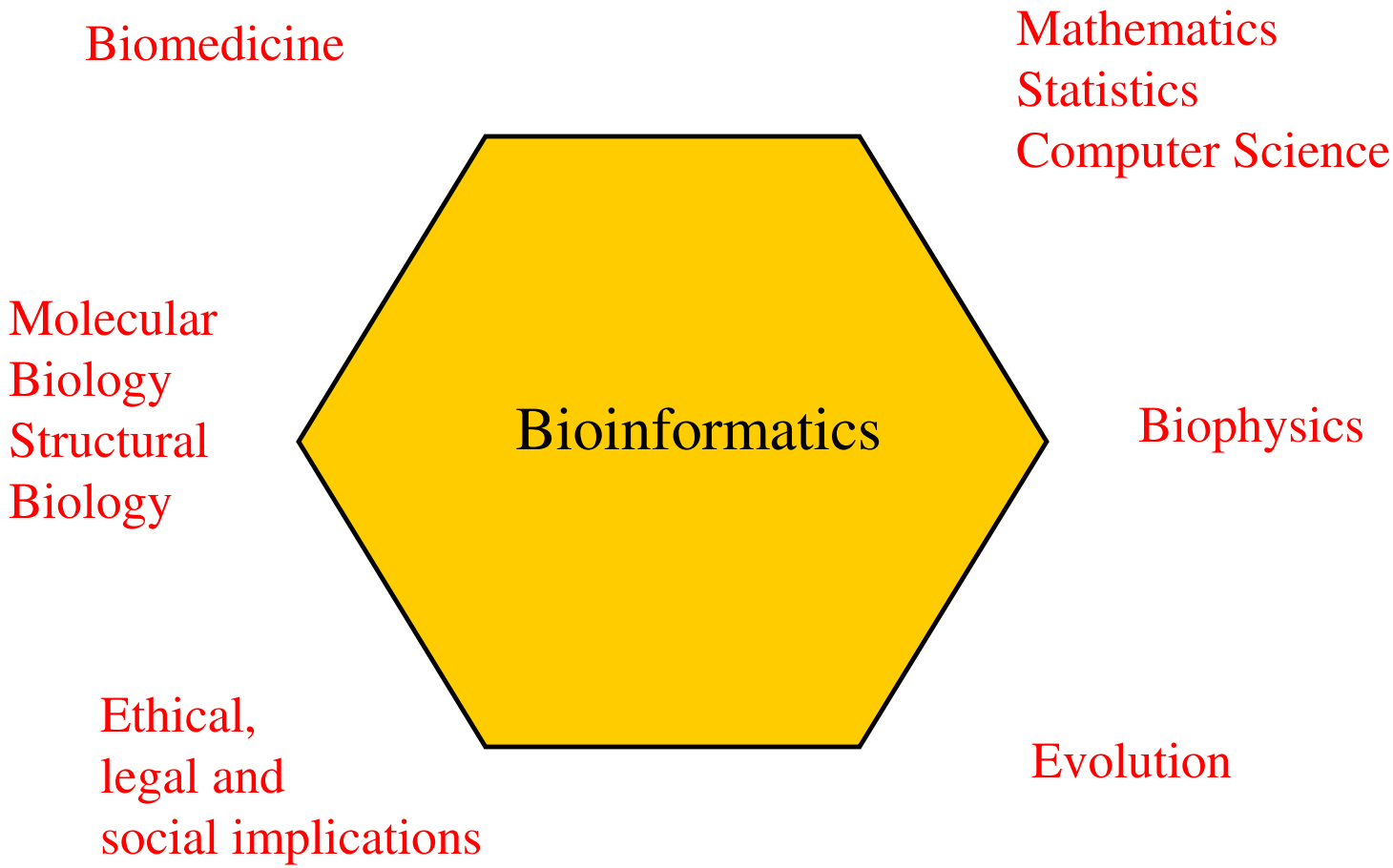
*“a collective term for data compilation, organisation, analysis and dissemination”*

# That means….



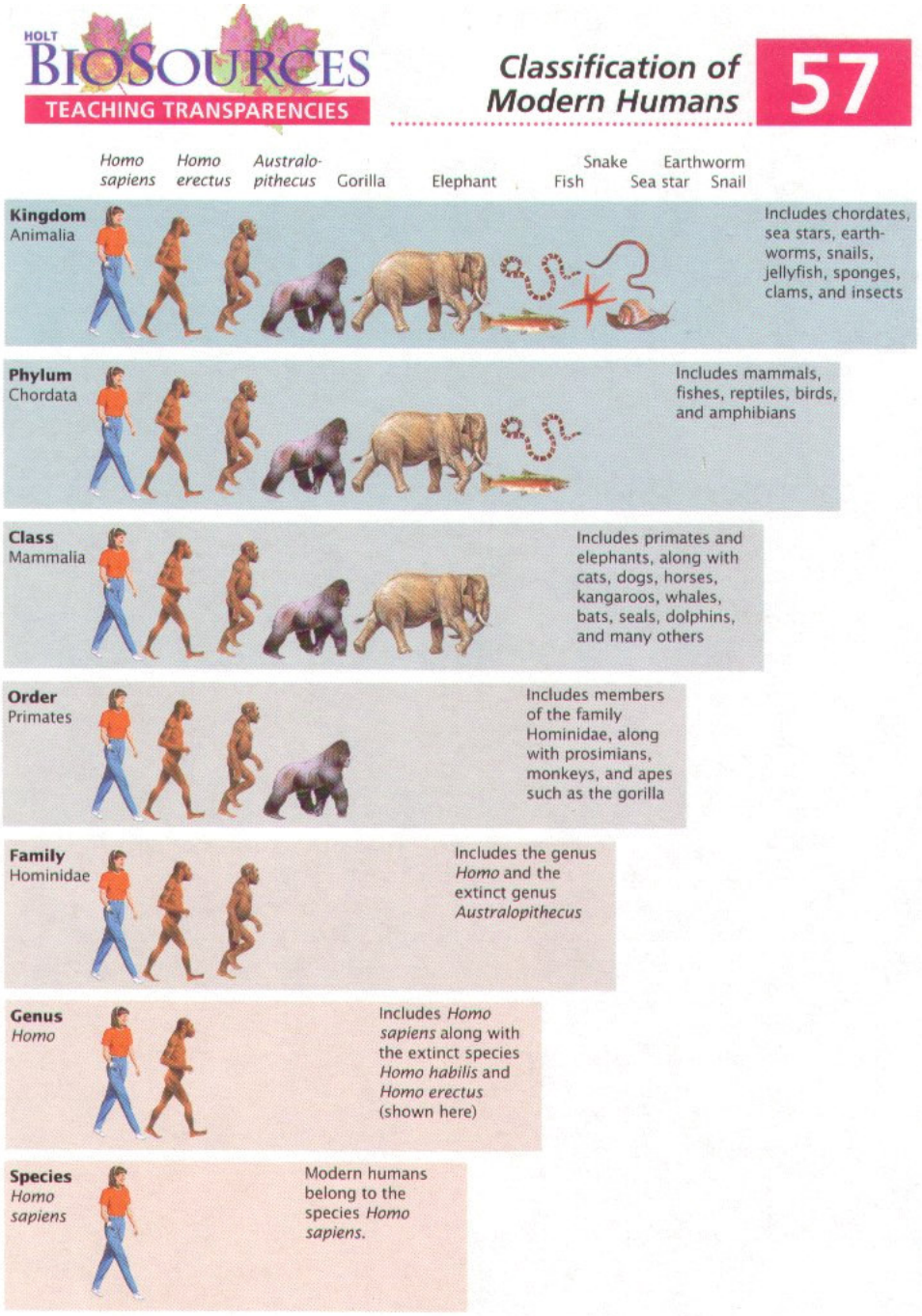
- Using information technology to help solve biological problems by designing novel algorithms and methods of analyses *(computational biology)*
- It also serves to establish innovative software and create new or maintain existing databases of information, allowing open access to the records held within them *(bioinformatics)*

# Bioinformatics is interdisciplinary



# What data?

*Biologists have been classifying data on plants and animals since the Greeks*



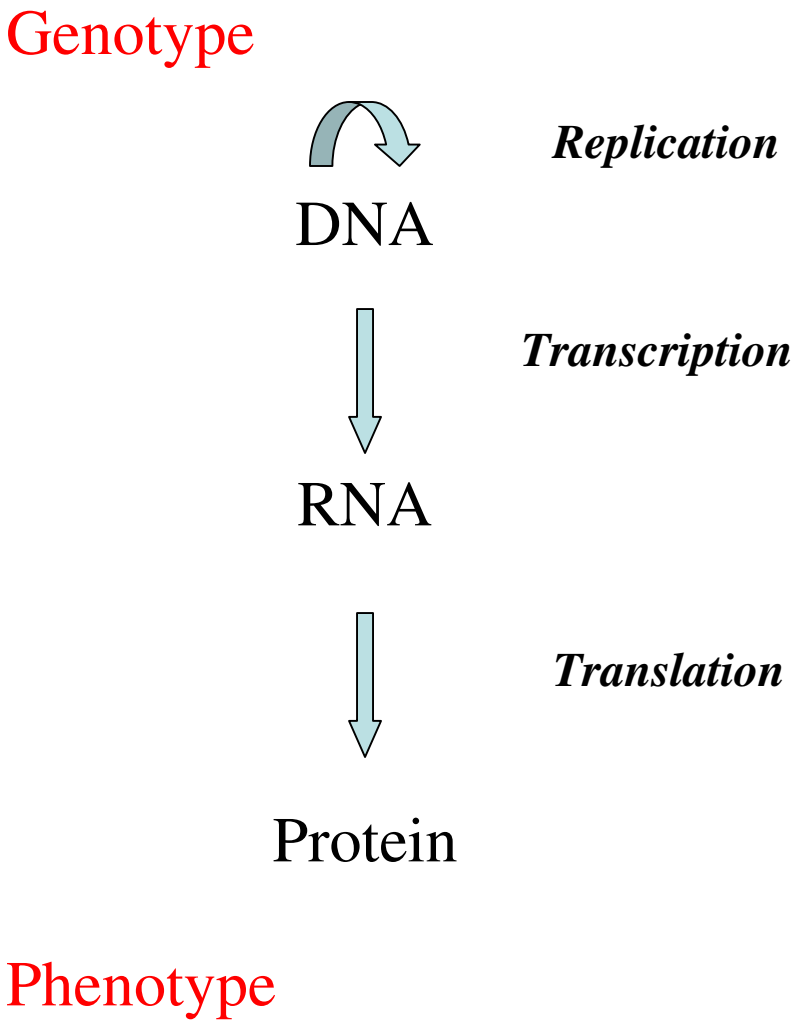
# The Tree of Life

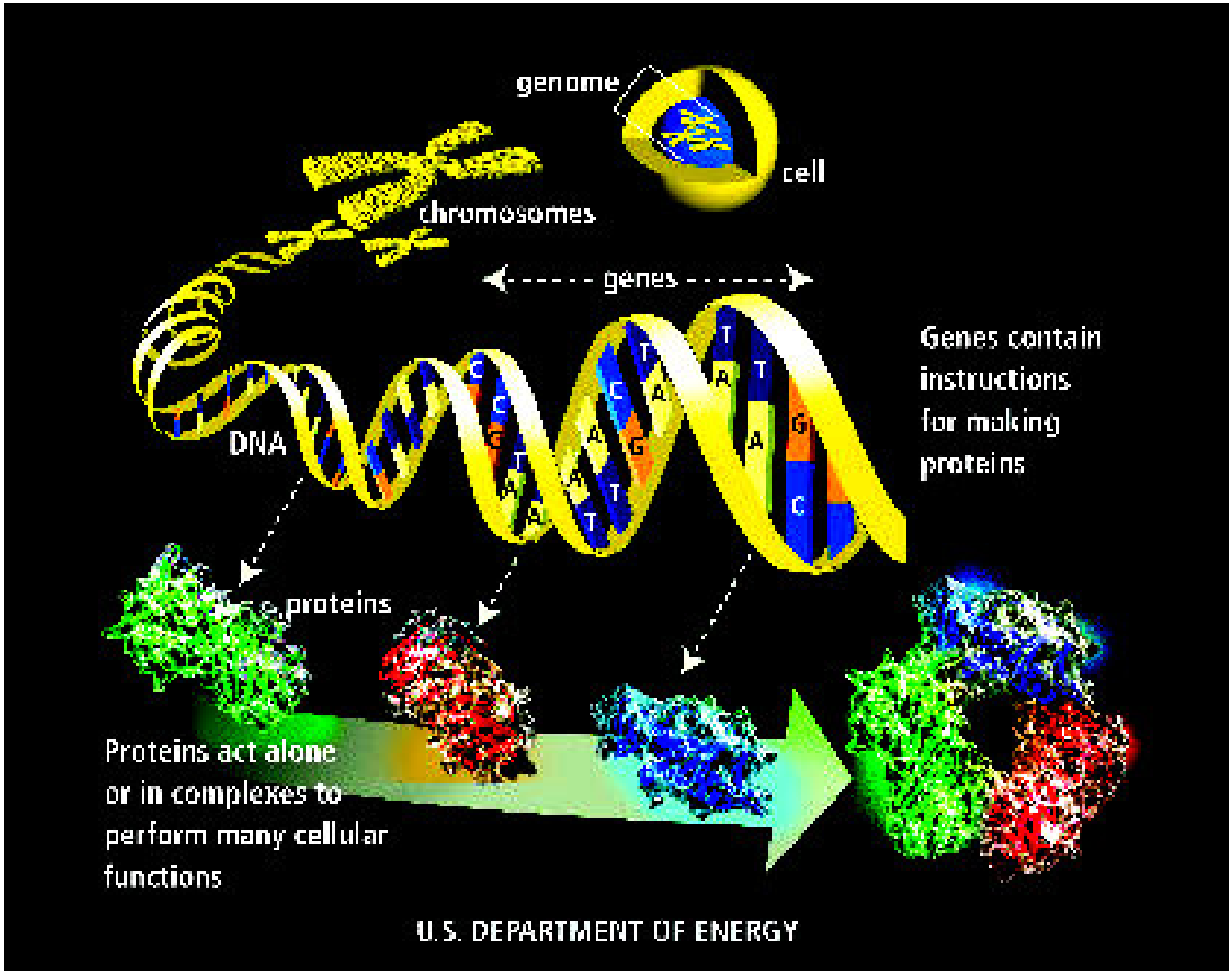
*"The affinities of all beings of the same class have sometimes been represented by a great tree… As buds give rise by growth to fresh buds, and these if vigorous, branch out and overtop on all sides many a feebler branch, so by generation I believe it has been with the great Tree of Life, which fills with its dead and broken branches the crust of the earth, and covers the surface with its ever branching and beautiful ramifications."*

*Charles Darwin, 1859*

***http://tolweb.org***

# Central Dogma of Molecular Biology




genome
cell
chromosomes
genes
DNA
Genes contain instructions for making proteins
proteins
Proteins act alone or in complexes to perform many cellular functions
U.S. DEPARTMENT OF ENERGY

# Genes (1)

- Genes are the basic units of heredity
- A gene is a sequence of bases that carries the information required for constructing a particular protein (gene “encode” the protein)
- The human genome comprises ~ 32,000 genes

# Genome sizes

| Organism | DNA length | Genes |
|---|---|---|
| *Mycoplasma genitalium* | 0.5 Mb | 470 |
| *Deinococcus radiodurans* | 3 Mb in 4-10 copies! | 3 200 |
| *Escherichia coli* | 4.5 Mb | 4 400 |
| *Saccharomyces cerevisiae* | 12 Mb | 6 200 |
| *Caenorhabditis elegans* | 97 Mb | 22 000 |
| *Drosophila melanogaster* | 120 Mb | 18 000 |
| *Homo sapiens* | 3200 Mb | 32 000 |

| Organism | Estimated size | Estimated gene number | Number of chromosome |
|---|---|---|---|
| *Homo sapiens* (human) | 2900 million bases | ~30,000 | 46 |
| *Rattus norvegicus* (rat) | 2,750 million bases | ~30,000 | 42 |
| *Mus musculus* (mouse) | 2500 million bases | ~30,000 | 40 |
| *Oryza sativa L.* (rice) | 450 million bases | ~40,000 | 12 |
| *Drosophila melanogaster* (fruit fly) | 180 million bases | 13,600 | 8 |
| *Arabidopsis thaliana* (plant) | 125 million bases | 25,500 | 5 |
| *Caenorhabditis Elegans* (roundworm) | 97 million bases | 19,100 | 6 |
| *Saccharomyces cerevisiae* (yeast) | 12 million bases | 6300 | 16 |
| *Escherichia coli* (bacteria) | 4.7 million bases | 3200 | 1 |
| *H. Influenzae* (bacteria) | 1.8 million bases | 1700 | 1 |

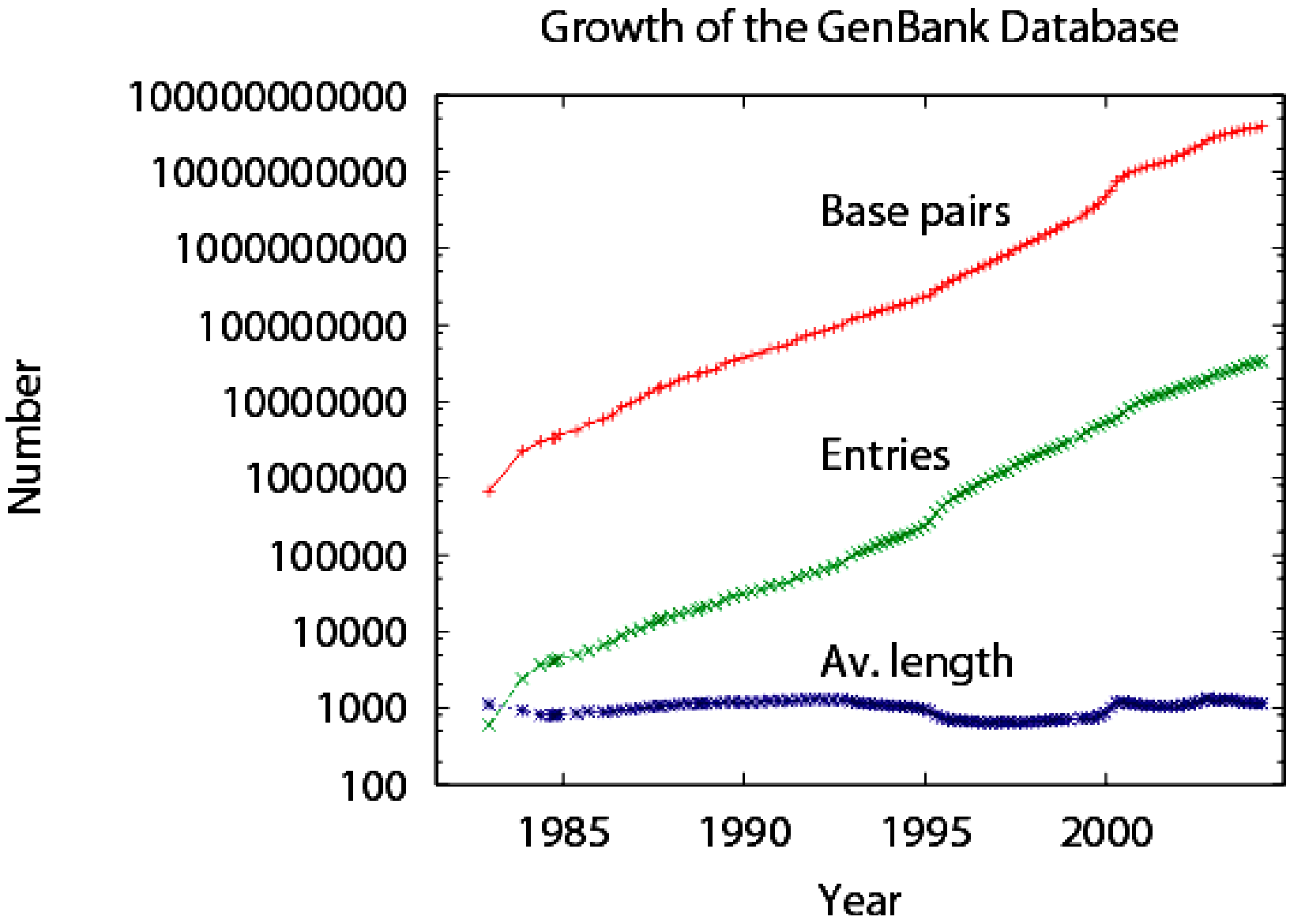
# The genomics projects


## GOLD[TM]
## Genomes OnLine Database v.2


| Contact: Genomesonline | Last Update: September 11, 2005 | Location www.genomesonline.org |
|---|---|---|
| | Search GOLD: 1567 genome projects | |
| 294 Published Complete Genomes | 740 Prokaryotic Ongoing Genomes | 532 Eukaryotic Ongoing Genomes including 8 chromosomes |

| Contact: Genomesonline | Last Update: September 19, 2008 | Location www.genomesonline.org |
|---|---|---|
| 857 Published Complete Genomes | Search GOLD: 4029 genome projects | 131 Metagenomes |
| 98 Archaeal Ongoing Genomes | 1953 Bacterial Ongoing Genomes | 990 Eukaryotic Ongoing Genomes |

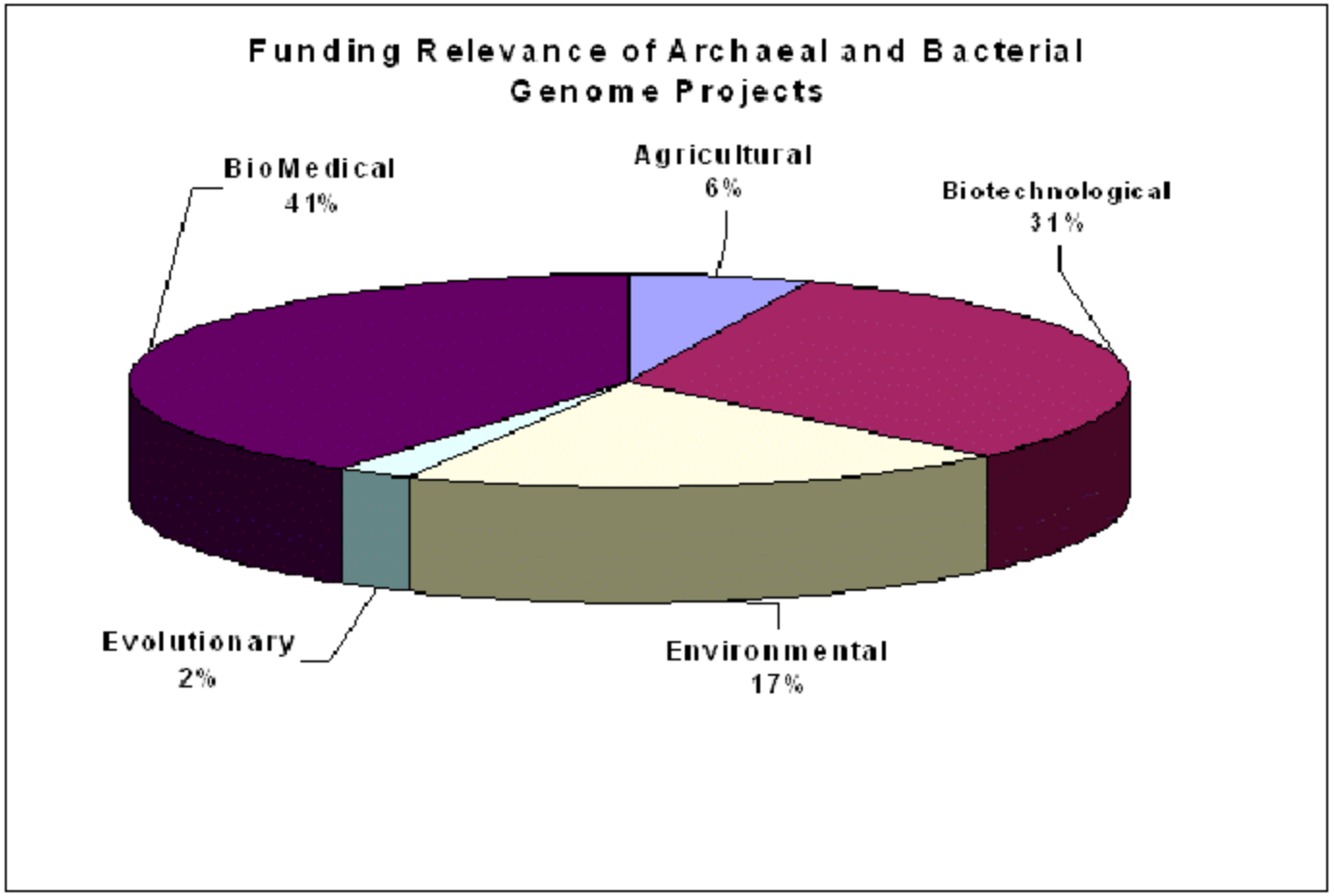
# Gene Databases



# Statistics on Genome Databases (2005)

QuickTime™ and a
TIFF (LZW) decompressor
are needed to see this picture.

# Funding Relevance of Archaeal and Bacterial Genome Projects



# Genes (2)

- The ~35,000 genes of the human genome encode > 100,000 polypeptides
- Not all of the DNA in a genome encodes protein

  microbes: 90% coding gene

  human: 3% coding gene
- About ½ of the non-coding DNA in humans is conserved (functionally important)

*Is there a danger, in molecular biology, that the accumulation of data will get so far ahead of its assimilation into a conceptual framework that the data will eventually prove an encumbrance ?*

John Maddox, 1988

# Top ten challenges for bioinformatics

1) Precise models of where and when transcription will occur in a genome (initiation and termination) ability to predict where and when transcription will occur in genome
2) Precise, predictive models of alternative RNA splicing: ability to predict the splicing pattern of any primary transcript in any tissue
3) Precise models of signal transduction pathways; ability to predict cellular responses to external stimuli
4) Determining protein:DNA, protein:RNA, protein:protein recognition codes
5) Accurate ab-initio protein structure prediction

# Top ten challenges for bioinformatics

6) Rational design of small molecule inhibitors of proteins
7) Mechanistic understanding of protein evolution: understanding exactly how new protein functions evolve
8) Mechanistic understanding of speciation: molecular details of how speciation occurs
9) Development of effective gene ontologies: systematic ways to describe gene and protein function
10) Education: development of bioinformatics curricula

*Source: Birney (EBI), Burge (MIT), Fickett (Glaxo)*

# Rough Outline of the Course

1) Overview of DNA, RNA and proteins
2) Sequence analysis
3) Structure analysis
4) Structure prediction
5) Molecular interactions
6) Drug design
7) Simulations
8) Available resources in bioinformatics