# Computational Structural Bioinformatics

**ECS129**

*Instructor: Patrice Koehl*


koehl@cs.ucdavis.edu

## Learning curve

<table>
<thead>
<tr>
<th>Math / CS</th>
<th>Biology / Chemistry</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre-requisite</td>
<td>Need to be able to access the web, to read and print PDF files</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>What you will learn</th>
<th>Not necessary</th>
</tr>
</thead>
<tbody>
<tr>
<td>Optimal alignment of two strings</td>
<td>How to solve the Poisson Boltzmann equation</td>
</tr>
<tr>
<td>Shape descriptors</td>
<td>Design a hashing function</td>
</tr>
<tr>
<td>Visualize and manipulate protein structures</td>
<td>Taxonomy, E.coli is a gram negative bacterium, organisation of protein kinases</td>
</tr>
<tr>
<td>Interactions between molecules</td>
<td>Protein families, structure prediction</td>
</tr>
<tr>
<td>Protein families</td>
<td>Use of bioinformatics databases and resources</td>
</tr>
<tr>
<td>Structure prediction</td>
<td></td>
</tr>
</tbody>
</table>

## 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

“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

- Biomedicine
- Mathematics
- Statistics
- Computer Science
- Molecular Biology
- Structural Biology
- Biophysics
- Ethical, legal and social implications
- Evolution

What data?

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

The Tree of Life

The affinities of all living things are classed by some botanists under a grand division, the Monera, which includes all single-celled beings; then a division of plants, and again of animals. But this is not the method I have adopted, for I consider it unnecessary to distinguish artificial groups which have no natural value. I am convinced that botanists will one day class all plants into one or two great groups, and zoologists will have the same result with the great classes of animals. The classification of the Tree of Life with its ever branching and beautiful ramifications.

Charles Darwin, 1859

http://tolweb.org
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 ~20,000 genes
<table>
<thead>
<tr>
<th>Organism</th>
<th>Estimated size</th>
<th>Estimated gene #</th>
<th>Number of chromosomes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>3,000 million bases</td>
<td>20,000</td>
<td>46</td>
</tr>
<tr>
<td>Rabbit (Oryctolagus cuniculus)</td>
<td>2,750 million bases</td>
<td>30,000</td>
<td>42</td>
</tr>
<tr>
<td>Macaque (Macaca fascicularis)</td>
<td>2,500 million bases</td>
<td>30,000</td>
<td>40</td>
</tr>
<tr>
<td>Oryza sativa L. (rice)</td>
<td>125 million bases</td>
<td>25,500</td>
<td>5</td>
</tr>
<tr>
<td>Drosophila melanogaster (fruit fly)</td>
<td>180 million bases</td>
<td>13,600</td>
<td>8</td>
</tr>
<tr>
<td>C. elegans (nematode)</td>
<td>97 million bases</td>
<td>19,100</td>
<td>6</td>
</tr>
<tr>
<td>Yeast (Saccharomyces cerevisiae)</td>
<td>12 million bases</td>
<td>6,300</td>
<td>16</td>
</tr>
<tr>
<td>E. coli</td>
<td>4.7 million bases</td>
<td>3,200</td>
<td>1</td>
</tr>
<tr>
<td>H. influenzae (bacteria)</td>
<td>1.8 million bases</td>
<td>1,700</td>
<td>1</td>
</tr>
<tr>
<td>Arabidopsis thaliana (plant)</td>
<td>125 million bases</td>
<td>25,500</td>
<td>5</td>
</tr>
</tbody>
</table>

### Genome sizes

- Homo sapiens (human): 2,900 million bases, ~20,000 genes, 46 chromosomes
- Rattus norvegicus (rat): 2,750 million bases, ~30,000 genes, 42 chromosomes
- Mus musculus (mouse): 2,500 million bases, ~30,000 genes, 40 chromosomes
- Oryza sativa L. (rice): 125 million bases, 25,500 genes, 12 chromosomes
- Drosophila melanogaster (fruit fly): 180 million bases, 13,600 genes, 8 chromosomes
- Caenorhabditis elegans (nematode): 97 million bases, 19,100 genes, 6 chromosomes
- Saccharomyces cerevisiae (yeast): 12 million bases, 6,300 genes, 16 chromosomes
- Escherichia coli (bacteria): 4.7 million bases, 3,200 genes, 1 chromosome
- Haemophilus influenzae (bacteria): 1.8 million bases, 1,700 genes, 1 chromosome
The genomics projects

http://www.genomesonline.org (GOLD)

Gene Databases

Growth of the GenBank Database

Base pairs

Entries

Av. length

Year

Genes (2)

- The ~20,000 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 highly conserved (functionally important)
Proteins: The Molecules of Life

- Ubiquitous molecules that are involved in all cellular functions
- Communication agents between cells
- Failure of a protein (missing, inactive, ...) can lead to serious health problems (prions, ...)

Why Proteins?

Function and Role of Proteins

Architecture:
- Structural proteins
- Cytoskeletal proteins
- Coiled proteins

Metabolism:
- Energy and Synthesis
- Catalytic enzymes

Sensory and Response:
- Regulation

Transport and Storage:
- Porins, transporters, Hemoglobin, transferrin

Regulation and Signaling:
- Transcription factors

Growth, Development and Reproduction:
- Flagella, cilia, Myosin, actin

Defence and Immunity:
- T-cell receptors, antibodies

Sensory and Response:
- Vision, hearing

Interactomics

Which proteins (biomolecules) interact with which proteins (biomolecules)?

Stanyon et al. Genome Biology 2004 & R96
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 patterns 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

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

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
<table>
<thead>
<tr>
<th></th>
<th>Rough Outline of the Course</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Overview of DNA, RNA and proteins</td>
</tr>
<tr>
<td>2</td>
<td>Sequence analysis</td>
</tr>
<tr>
<td>3</td>
<td>Structure analysis</td>
</tr>
<tr>
<td>4</td>
<td>Structure prediction</td>
</tr>
<tr>
<td>5</td>
<td>Molecular interactions</td>
</tr>
<tr>
<td>6</td>
<td>Drug design</td>
</tr>
<tr>
<td>7</td>
<td>Simulations</td>
</tr>
<tr>
<td>8</td>
<td>Available resources in bioinformatics</td>
</tr>
</tbody>
</table>