Computational Structural Bioinformatics

ECS129
Instructor: Patrice Koehl

koehl@cs.ucdavis.edu

Learning curve

<table>
<thead>
<tr>
<th>Pre-requisite</th>
<th>Biology/Chemistry</th>
</tr>
</thead>
<tbody>
<tr>
<td>Need to be able to access the web, read and print PDF files</td>
<td>Molecules, basic cell biology</td>
</tr>
<tr>
<td>Basic knowledge of statistics, probability</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>What you will learn</th>
<th>Biology/Chemistry</th>
</tr>
</thead>
<tbody>
<tr>
<td>Optimal alignment of two strings</td>
<td>Interactions between molecules</td>
</tr>
<tr>
<td>Shape descriptors</td>
<td>Protein families</td>
</tr>
<tr>
<td>Visualize and manipulate protein structures</td>
<td>Structure prediction</td>
</tr>
<tr>
<td>Use of bioinformatics databases and resources</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Not necessary</th>
<th>Biology/Chemistry</th>
</tr>
</thead>
<tbody>
<tr>
<td>How to solve the Poisson Boltzmann equation</td>
<td>Taxonomy, E.coli, a gram negative bacterium</td>
</tr>
<tr>
<td>Design a hashing function</td>
<td>organisation of protein kinases</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

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

What data?

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

The affinities of all beings of the same class have sometimes been represented by means of a tree, in which the branches on one side stretch out over the surface of the earth, and of the other over the ocean. As buds give rise by growth to fresh buds, and these if vigorous, branch out and overtop one another, 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 and the “omics”
Integrative Systems Biology

Geneomics

Transcriptomics

Proteomics

Regulation

Interactomics

Metabolism

Metabolomics

DNA

RNA

Protein

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 chromosome</th>
</tr>
</thead>
<tbody>
<tr>
<td>Homo sapiens (human)</td>
<td>2000 million bases</td>
<td>~20,000</td>
<td>46</td>
</tr>
<tr>
<td>Rattus norvegicus (rat)</td>
<td>2,750 million bases</td>
<td>~30,000</td>
<td>42</td>
</tr>
<tr>
<td>Mus musculus (mouse)</td>
<td>2200 million bases</td>
<td>~10,000</td>
<td>40</td>
</tr>
<tr>
<td>Oryza sativa (rice)</td>
<td>450 million bases</td>
<td>~40,000</td>
<td>12</td>
</tr>
<tr>
<td>Drosophila melanogaster (fruit fly)</td>
<td>130 million bases</td>
<td>~13,000</td>
<td>8</td>
</tr>
<tr>
<td>Arabidopsis thaliana (plant)</td>
<td>125 million bases</td>
<td>12,500</td>
<td>5</td>
</tr>
<tr>
<td>Caenorhabditis elegans (roundworm)</td>
<td>97 million bases</td>
<td>16,100</td>
<td>6</td>
</tr>
<tr>
<td>Drosophila melanogaster (fruit fly)</td>
<td>5 million bases</td>
<td>~5,000</td>
<td>1</td>
</tr>
<tr>
<td>Escherichia coli (bacteria)</td>
<td>1.8 million bases</td>
<td>1,800</td>
<td>1</td>
</tr>
<tr>
<td>H. Influenzae (bacteria)</td>
<td>1.8 million bases</td>
<td>1,800</td>
<td>1</td>
</tr>
</tbody>
</table>
The genomics projects

[Graph or table showing genomics projects]

Gene Databases

[Graph showing growth of the GenBank database]

Projects by Domain

[Graph showing project growth by domain]
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 1/2 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, Cytoskeletal proteins
- Metabolism: Energy and Synthesis, Catalytic enzymes
- Sensory and response
- Locomotion: Flagella, cilia, Myosin, actin
- Transport and Storage: Porins, transporters, Hemoglobin, transferrin
- Defence and Immunity
- Regulation and Signaling: Transcription factors
- Growth, Development and Reproduction

Interactomics

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

Stanyon et al: Genome Biology 2004 5: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 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