CS313 Computational Biology

Course Information

Instructor: Brian Tjaden
Pronouns: He, Him, His
Course Materials: http://cs.wellesley.edu/~cs313

Computational Biology is Multidisciplinary

Computational algorithms  Computational Biology  Biological data

DNA the Molecule of Life

chromosomes  gene  cell

DNA  gene

A-1  A-2  A-3  A-4
DNA: simplified

DNA: “program” for cell processes
Proteins: execute cell processes

DNA Structure

- Double helix
- Deoxyribose (sugar) - phosphate backbone
- Four bases - A, T, G, C
- Base pairing

DNA Structure

- Information polarity (anti-parallel strands)
- Either strand can function as a template (complementary strands)

Information Flow

Nucleic acids

Amino acids

DNA → RNA → Protein

transcription
translation
DNA → RNA → Protein

The Genetic Code

- 61 amino acid codons
- 1 start codon (Met)
- 3 stop codons

The Genetic Code

- Triplet code
- Non-overlapping codons
- Start and stop codons
- Degeneracy

4 nucleotides, 20 amino acids
Mutations

- Changes in DNA occur, despite cell’s best efforts

- Spontaneous events, copying errors, enviromental factors

- Mutations might change gene function

- Can be harmful, neutral, or beneficial

Normal RBCs

Sickle cell anemia

Ultroviolent (UV) Light Causes Sunburn and DNA Damage

Sample Genomes

A genome is the total DNA in a cell

<table>
<thead>
<tr>
<th>Species</th>
<th>Genome Size</th>
<th># of Genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>SARS-CoV-2</td>
<td>30 thousand</td>
<td></td>
</tr>
<tr>
<td>Epstein-Barr virus</td>
<td>172 thousand</td>
<td></td>
</tr>
<tr>
<td>Escherichia coli</td>
<td>4.6 million</td>
<td></td>
</tr>
<tr>
<td>Drosophila melanogaster</td>
<td>122 million</td>
<td></td>
</tr>
<tr>
<td>Homo sapiens</td>
<td>3.3 billion</td>
<td></td>
</tr>
<tr>
<td>Paris japonica</td>
<td>150 billion</td>
<td></td>
</tr>
</tbody>
</table>
Moore's Law: The number of transistors on microchips doubles every two years.

Cost of Sequencing a Human Genome
Sequencing a Genome

Is sequencing genome the end or beginning?

Open Questions

- Which regions of DNA have biologic function?
  (What are the genes?)
- What are their functions?
- When and how are genes turned on and off?
- How do genes and their products (proteins) interact with each other?
- What are the implications to health and medicine?

in other words...
How does the cell’s DNA “program” work?

Comparative Genomics

Comparative genomics involves understanding the relationships between the genomes of different species.

- Many genes have unknown functions
- Which genes are present (conserved vs unique)?
- How are genes arranged in the genome
- Infer function of genes by sequence similarity - homology to known genes

  Human and human: 99.9%
  Human and chimpanzee: 96%
  Human and mouse: 85%
  Human and banana: 55%

  Human and bacteria: ??%
Recurring Themes

• Bioinformatic tools are often hypotheses-generating

• Determining statistical significance of results generated by bioinformatic tools is useful

• Properties of data guide choice of algorithm

• Some problems are solved exactly or optimally. Other problems are addressed using a heuristic approach.

• Many computational approaches are improved by incorporating additional biological insights into their underlying method or model.

• Recent advances have allowed scientists to gather large amounts of, often heterogeneous, data. One of the roles of bioinformatic tools is efficient analysis of large data sets with the aim of extracting new biological insights.