Computational Biology
Fall 2017

Administrivia

Instructor: Brian Tjaden

Preferred Pronouns: He, Him, His

Meeting: Mondays & Thursdays, 1:30-2:40pm

Office Hours: Mondays 10:30am-12:30pm
Wednesdays 11:00am-12:30pm

Google Group: CS-313-01-FA17

Course Materials

http://cs.wellesley.edu/~cs313

Computational Biology is Multidisciplinary
**DNA: simplified**

DNA → Gene → Protein

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

transcription → translation

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
tRNAs (transfer RNAs)

Amino acid

codons

anti-codon

tRNA

GAACGCU AUGCUUGGUG CUCUAAGUA ACGCUAG mRNA

Mutations

- Changes in DNA occur, despite cell’s best efforts
- Spontaneous events, copying errors, environmental factors
- Mutations might change gene function
- Can be harmful, neutral, or beneficial

Normal RBCs

Sickle cell anemia

Types of Mutations

Single base substitutions

A → T

Insertions and Deletions

Amplifications

Translocations

9 22

CML

Inversions
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>Epstein-Barr virus</td>
<td>172 thousand</td>
<td></td>
</tr>
<tr>
<td><em>Escherichia coli</em></td>
<td>4.6 million</td>
<td></td>
</tr>
<tr>
<td><em>Drosophila melanogaster</em></td>
<td>122 million</td>
<td></td>
</tr>
<tr>
<td><em>Homo sapiens</em></td>
<td>3.3 billion</td>
<td></td>
</tr>
<tr>
<td><em>Paris japonica</em></td>
<td>150 billion</td>
<td></td>
</tr>
</tbody>
</table>
Sequencing a Genome

Comparative Genomics

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

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

State of the Art

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?
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.

The Next Big Thing

"The next big thing in mathematics? Biology. ... The mathematics involved in studying the genome and the folding of proteins is deep, elegant, and beautiful ... a spectacular new area of research that is certain to grow enormously in the next 10 years."

John Ewing - American Mathematical Society