Pursuing Graduate Studies
(in Bioinformatics)

Chau-Wen Tseng

Department of Computer Science
University of Maryland, College Park
What You Learn

🔹 Bachelor (B.A.)
  - Basic principles & problem-solving techniques
  - Apply techniques to solve classroom problems

🔹 Master (M.S.)
  - State-of-the-art principles & techniques
  - Apply techniques to solve semi-real problems

🔹 Doctorate (Ph.D.)
  - Independently discover new principles & techniques
  - Apply techniques to solve cutting-edge problems
What Type of Job

- Bachelor (B.A.)
  - Greatest variety / number of jobs
  - May become boring over time
  - Less control over responsibilities / goals

- Master (M.S.)
  - Large variety / number of jobs
  - Greater potential for management

- Doctorate (Ph.D.)
  - Limited variety / number of jobs
    - Research & teaching positions
  - More ability to set your own responsibilities & goals
Graduate Studies Tradeoffs

- **Advantages**
  - More interesting job opportunities
  - More flexibility & control over job / career
  - Possibly higher income
  - Prestige

- **Disadvantages**
  - Many additional years of school (5-8+ for Ph.D.)
  - May not recover lost income
  - Limited career options (especially with Ph.D.)
  - Stressful on family & relationships
  - Not suitable for everyone
    - Only ~50% graduate students complete Ph.D.
What To Study?

- **Bioinformatics**
  - The creation and development of advanced information and computational techniques for solving problems in biology

- **Motivation**
  - Techniques for inexpensively measuring large amounts of biological data
Graduate Studies in Bioinformatics

- **Exciting area**
  - Biologists & doctors taking advantage of new technology & computing power
  - Area with lots of potential

- **Requires wide range of skills**
  - Molecular biology
  - Statistics
  - Algorithms
  - Artificial intelligence (machine learning)
  - Databases
Some Warnings

- **Bioinformatics jobs**
  - Highly specialized position
  - Area has been (over) hyped by media
  - Not many actual positions available in industry

<table>
<thead>
<tr>
<th>Job Listing Keyword</th>
<th>Listings in US</th>
<th>Listings in MD (Montgomery County)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Computer Security</td>
<td>5000+</td>
<td>199</td>
</tr>
<tr>
<td>Software Engineering</td>
<td>5000+</td>
<td>196</td>
</tr>
<tr>
<td>Bioinformatics</td>
<td>119</td>
<td>9 (4 biostatistics)</td>
</tr>
</tbody>
</table>
Bioinformatics Topics

- Sequence alignments
  - Find similarity between DNA / protein (amino acid) sequences
- Genome assembly
  - Combining genomic fragments to form whole genome
- Gene identification & annotation
  - Identify and classify genes on the genome
- Microarrays & gene expression analysis
  - Use DNA microarray (gene chip) to measure mRNA
- Protein folding
  - Compute 3D protein structure ↔ protein sequence
- Phylogenetic analysis
  - Find genetic relationships between sequences / species
Sequence Alignment

- **Pairwise sequence alignment**
  - Mutual arrangement of sequences
    
    \[
    \begin{align*}
    \text{C A T C A} \\
    \text{G A T} \\
    \text{G G T}
    \end{align*}
    \]
  - Similar sequence $\rightarrow$ similar function?
    
    \[
    \begin{align*}
    \text{C – T C A} \\
    \text{G G T}
    \end{align*}
    \]

- **Look for**

![Color Key for Alignment Scores]

Sequences producing significant alignments:

<table>
<thead>
<tr>
<th>Score</th>
<th>E Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>133</td>
<td>8e-31</td>
</tr>
<tr>
<td>133</td>
<td>8e-31</td>
</tr>
<tr>
<td>131</td>
<td>3e-30</td>
</tr>
<tr>
<td>122</td>
<td>2e-27</td>
</tr>
</tbody>
</table>
Sequence Assembly

- Combining short mRNA / DNA sequences (reads)
- Produce full transcript / genome
Gene Prediction

- Predicting expressed DNA (genes)
Measuring Gene Expression & Regulation

- Detecting DNA (genes) expressed as mRNA

Affymetrix DNA microarray

500,000 oligomers in 1.28 cm²
Protein Structure Prediction

- Predict protein 3D structure from (amino acid) sequence
- Predict interactions between proteins and ligands

5' atgcccaagctgaat ... 3'

atg ccc aag ctg aat ...

M P K L N ...

![Diagram of protein structure and DNA sequence]
Phylogenetic Analysis

- Study of evolutionary relationships
- Infer evolutionary relationship from shared features