Introduction of Computational Biology

By Rohayanti Hassan (PhD)
Why **Computational Biology**?

- Lots of data (* lots of data)
- There are rules
- Pattern finding
- It’s *all* about data
- Ability to visualize
- Simulations
- Guess + verify (generate hypotheses for testing)
- Propose mechanisms / theory to explain observations
- Networks / combinations of variables
- Efficiency (reduce experimental space to cover)
- Informatics infrastructure (ability to combine datasets)
- Correlations
- *Life itself is digital.* Understand cellular instruction set
Computational biology is particularly exciting today because:

- the problems are large enough to motivate efficient algorithms,
- the problems are accessible, fresh and interesting,
- biology is increasing becoming a computational science.

Computational biology is increasing of interest in both life science and computational science departments.

Source of complex questions and real-life data.

- Many problem ideas go from biology to CS: e.g. fragment assembly, sequence analysis, algorithms for phylogenetic trees.
- Many problem ideas go from CS to biology: e.g. sequencing by hybridization, DNA computing.
Do you know, these data belong to?
Do you know, these data belong to?
Computational Biology?

- **Computational biology** involves the development and application of data-analytical and theoretical methods, mathematical modeling and computational simulation techniques to the study of biological, behavioral, and social systems.

**Bioinformatics/Computational Biology**

(the application of a core technology of computer science, e.g. algorithms, artificial intelligence, data bases, to problems arising from biology.)

- **Computational Biology**
  - Molecular dynamics (MD) simulation
  - Simulation of population
  - Theory of Protein folding
  - Biological image (2D/3D) processing

- **Bioinformatics**
  (massive data analysis database)

- **Systems Biology**
  (looking at biology as a system often using network structures)

- **Translational Bioinformatics**
  (additional uses of clinical info to improve medicine)

In this course, Bioinformatics ≈ Computational Biology

- The field is broadly defined and includes foundations in computer science, applied mathematics, animation, statistics, biochemistry, chemistry, biophysics, molecular biology, genetics, genomics, ecology, evolution, anatomy, neuroscience, and visualization.
The Challenges

Challenges in Computational Biology

1. Gene Finding
2. Sequence alignment
3. Database lookup
4. Gene expression analysis
5. Comparative Genomics
6. Evolutionary Theory
7. Regulatory motif discovery
8. Genome Assembly
9. Cluster discovery
10. Gibbs sampling
11. Protein network analysis
12. Metabolic modelling
13. Emerging network properties
Proteins carry out the cell’s chemistry

- More complex polymer
  - Nucleic Acids have 4 building blocks
  - Proteins have 20. Greater versatility
  - Each amino acid has specific properties

Sequence → Structure → Function

- The amino acid sequence determines the three-dimensional fold of protein
- The protein’s function largely depends on the features of the 3D structure

Proteins play diverse roles

- Catalysis, binding, cell structure, signaling, transport, metabolism

Enjoy this video:
Cellular Visions the Inner Life of a Cell

Nuclei, proteins and lipids move with bug-like authority, slithering, gliding and twisting through 3D space. "All of those things that you see in the animation are going on in every one of your cells in your body all the time," says XVIVO lead animator John Liebler
Computational Solution

1. Pre-Analytics
   - Discretization
   - Binning
   - Normalization
   - Standard Deviation
   - Feature Selection
   - Feature Ranking
   - etc

2. Post-Analytics
   - Supervised Learning
   - Unsupervised Learning
   - Probabilistic Method
   - etc

Output

Biological Data

Big Data

Different dimensions:
  - Volume
  - Velocity
  - Variety
  - Veracity

www.utm.my
Current Trend? BIG DATA? DATA SCIENCE

**Annual Sequence Data Generation**

- **Original Data Generation from Annual Sequencing**
- **Data Expansion from Interpretation & Analysis**

<table>
<thead>
<tr>
<th>Year</th>
<th>Data Generation (PetaBytes)</th>
<th>Data Expansion (ExaBytes)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2014</td>
<td>105</td>
<td>3.3</td>
</tr>
<tr>
<td>2015</td>
<td>308</td>
<td>1.6</td>
</tr>
<tr>
<td>2016</td>
<td>776</td>
<td></td>
</tr>
<tr>
<td>2017</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2018</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2019</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Main categories of high-performance computing platforms**

<table>
<thead>
<tr>
<th>Large-scale computing platform</th>
<th>computing architectures</th>
<th>Advantages</th>
<th>Disadvantages</th>
<th>example applications</th>
</tr>
</thead>
</table>
| Cluster computing              | Multiple computers linked together, typically through a fast local area network, that effectively function as a single computer | Cost-effective way to realize supercomputer performance | Requires a dedicated, specialized facility, hardware, system administrators and IT support | • BLAST  
• Bayesian network reconstruction  
• Computing genetic associations in large-scale GWAS studies |
| Cloud computing                | Computing capability that abstracts the underlying hardware architectures (for example, servers, storage and networking), enabling convenient, on-demand network access to a shared pool of computing resources that can be readily provisioned and released (NIST Technical Report) | The virtualization technology used results in extreme flexibility; good for one-off HPC tasks, for which persistent resources are not necessary | Privacy concerns; less control over processes; bandwidth is limited as large data sets need to be moved to the cloud before processing | • Searching sequence databases  
• Aligning raw sequencing reads to genomes  
• General purpose genomics tools (for example, GeneSifter from Geospiza)  
• Most applications running on a cluster can be transferred to a cloud |
Prospect as a Data Scientist

8 Data Skills to Get You Hired

Basic Tools

Basic Statistics

Machine Learning

Multivariable Calculus and Linear Algebra

Data Munging

Data Visualization & Communication

Thinking Like A Data Scientist

Software Engineering

Thinking Like A Data Scientist

www.utm.my