Sequence and Omics Data

Cydney Nielsen

University of British Columbia
British Columbia Cancer Agency
Vancouver, British Columbia, Canada
Big Picture

Genes (DNA sequences) are transcribed into RNA molecules (expressed)

RNA molecules transcribed from a gene often called *transcripts*

*RNA transcripts* are translated into *proteins*
Big Picture

eukaryotic cell

DNA  Nucleotides (A, C, G, T)

RNA  Nucleotides (A, C, G, U)

PROTEIN  Amino acids (21 letters)
Big Picture

eukaryotic cell

prokaryotic cell
High-Throughput Technologies

http://www.bcgsc.ca/platform
Omics Data – High-Throughput Measurements

DNA → genome → entire DNA sequence of an organism
RNA → transcriptome → set of all RNA molecules in one or a population of cells
PROTEIN → proteome → set of expressed proteins in one or a population of cells
To make sense of DNA, RNA, protein sequence data, we first label all of the functional elements we know about. We call these labels annotations.
For well studied genomes (e.g. *E. coli*, mouse, human) can *annotate* the genome sequence with known genes.
Annotations

For well studied genomes (e.g. *E. coli*, mouse, human) can *annotate* the genome sequence with known genes and with known functions.
Annotations

DNA
RNA
PROTEIN

genes

genome 1

genome 2

genome 3
Presence/absence patterns of genes in different genomes can inspire hypotheses about the underlying biology:

Does gene $X$ drive the pathogenicity of bacteria with genome 1?
Annotations

Rich online sources of annotations for DNA, RNA, and proteins

Examples
Annotations

Rich online sources of annotations for DNA, RNA, and proteins

Challenges

How to integrate diverse resources?
How to facilitate query and exploration of diverse annotations?

**Paper 3: COMBat: Visualizing Co-Occurrence of Annotation Terms**
Multiple Sequence Alignments

DNA → RNA → PROTEIN

genes

Paper 2:
Large-Scale Multiple Sequence Alignment Visualization through Gradient Vector Analysis
Multiple Sequence Alignments

- DNA
- RNA
- PROTEIN

Genomes:
- Genome 1
- Genome 2
- Genome 3
- Genome 4

Newly sequenced genome
Computational gene predictions but no known functions
Multiple Sequence Alignments

Alignments allow you to use sequence conservation to infer function.
Sequence similarity suggests evolutionarily relationships

Villarreal and DeFilippis. J. Virology 2000
Multiple Sequence Alignments

Sequence similarity suggests evolutionarily relationships

Phylogenetic tree inferred from sequence alignments

Villarreal and DeFilippis J. Virology 2000
Multiple Sequence Alignments

**Challenges**

How to represent many thousand aligned sequences?

How to indicate potential errors introduced by the alignment algorithms?

**DNA**

**RNA**

**PROTEIN**

---

**Walker loop phosphate-binding motif**

---

**Paper 2:**

Large-Scale Multiple Sequence Alignment Visualization through Gradient Vector Analysis

Roßbach et al. BMC Structural Biology 2005
In vitro selection

DNA

RNA

PROTEIN

Paper 1:
invis: Exploring High-Dimensional RNA Sequences from In Vitro Selection
In vitro selection

DNA

RNA

PROTEIN

Input RNA sequence pool
In vitro selection

DNA

RNA

PROTEIN

Input RNA sequence pool

SELECTION

Select for a function of interest (e.g. binding to a particular compound)
**In vitro selection**

DNA → RNA → PROTEIN

- **Input RNA sequence pool**
- **Enriched RNA sequence pool**

**SELECTION**
**In vitro selection**

- **DNA**
- **RNA**
- **PROTEIN**

Need to make more material for subsequent selection

**AMPLIFICATION**

**SELECTION**

**Input RNA sequence pool**

**Enriched RNA sequence pool**

Need to make more material for subsequent selection.
**In vitro selection**

1. **DNA** → **RNA**
2. **RNA** → **PROTEIN**
3. **AMPLIFICATION**
4. **REPEAT**
5. **SELECTION**

- Input RNA sequence pool
- Enriched RNA sequence pool
**In vitro selection**

### DNA

### RNA

**Random sequence pool**
- Allows for discovery of previously unknown functional RNAs

**Randomly mutated versions of a known functional sequence**
- Allows you to explore the diversity of sequences that maintain the target functionality

### PROTEIN
**In vitro selection**

**Challenges**

How to visualize and explore high dimensional sequence space? How to reveal its overall structure while also showing details at the individual sequence level?

**Paper 1:**

invis: Exploring High-Dimensional RNA Sequences from In Vitro Selection
Sequence and Omics Data Papers

**Paper 1:**

invis: Exploring High-Dimensional RNA Sequences from In Vitro Selection

**Paper 2:**

Large-Scale Multiple Sequence Alignment Visualization through Gradient Vector Analysis

**Paper 3:**

COMBat: Visualizing Co-Occurrence of Annotation Terms