Views of this Class

- Graduate Comp. Biol. 2
  - Covers much of comp. biol. that is not covered in CMSC 858E (which is more sequence-based)
  - CMSC 858E is **not** a pre-requisite however.

- Computational Systems Biology
  - Understanding large systems as a whole, and how their parts work together to perform some function.

  - Various uses of graphs, graph theory, and graph algorithms to solve biological problems.
Topics

- Background
- Annotating proteins with function
- Comparing & searching interaction networks
- Finding significant motifs & patterns
- Dynamics of regulation networks
- Other network analysis

- Genome rearrangements
- Phylogenetic trees & networks
- Sequence assembly
- Protein structure prediction
- Sequence motif-finding

Protein-protein interaction networks
*First 2/3rd of class*

Other graph models for bio applications
*Last 1/3 of class*
Class Work

- Midterm (25%)
- Final (30%)
- Class project. Will be fairly open-ended; details in a few weeks (~25%)
- Homeworks / quizzes (~15%)
- Participation (~5%)
Life

- **prokaryotes**: lack nucleus, includes bacteria, “simple”
- **eukaryotes**: have nucleus, includes humans, more “complex”
  - can be single-cellular (e.g. yeast)
  - or multi-cellular (e.g. human, or fruit fly)
- Humans have 30,000-40,000 genes
Recent history

- First genome sequenced in 1995
- Human genome (draft #1): 2001
- Now: over 800 bacterial genomes
- Hundreds of eukaryotic genomes done or in progress.
Central Dogma of Biology:

DNA =
- double-stranded, linear molecule
- each strand is string over \{A,C,G,T\}
- strands are complements of each other (A \leftrightarrow T; C \leftrightarrow G)
- substrings encode for genes, most of which encode for proteins

Transcription

mRNA

proteins

(T \rightarrow U)
Proteins

mRNA
\[ \Sigma = \{A,C,G,U\} \]

protein
\[ |\Sigma| = 20 \text{ amino acids} \]

Amino acids with flexible side chains strung together on a backbone

Proteins are the Building Blocks of Life

Their shape is instrumental in determining their function.
Polymers

- Proteins, DNA, RNA are examples of polymers:
- Macromolecules built from chains of "monomers," smaller subunits.
- Polysaccharides are another example (chains of sugars)
Examples of Proteins

- Alcohol dehydrogenase
- Trypsin: breaks down other proteins
- Antibodies
- TATA DNA binding protein
- Collagen: forms tendons, bones, etc.

Examples of “Molecules of the Month” from the Protein Data Bank
http://www.rcsb.org/pdb/
“Cartoon” drawing, showing major features such as alpha helices and beta sheets
The function of proteins

- Structural: the organelles of the cell
- Signaling: pass information from the environment and between different parts of the cell; turn genes on & off.
- Catalyze reactions (act as enzymes).
Protein “complex”

Group of proteins that interact to form a stable functioning unit.

Estimate:
- ~800 protein complexes in yeast
- ~3000 in human
Fas Signaling Pathway

Involved in programmed cell death (apoptosis)

Such cell death is often desirable in multicellular organisms.
Edge Types:
“-”: inhibitory
“+”: stimulatory
o: neutral
?: undefined

(darkness indicates confidence)
Types of Biological Networks

- Yeast transcription network
- Yeast Protein-Protein interaction network
- Yeast Phosphorylation network
- E. coli metabolic network
- Yeast SSL network

Transcription network, aka regulatory network:

**Transcription Factors** = proteins that bind to DNA to activate or repress the nearby, downstream genes.

the regulated gene might also be a transcription factor

leads to a directed graph
Protein-Protein interaction network:

Proteins physically interact

Assumption of binary interactions is imperfect.

Sometimes several proteins must bind simultaneously for there to be any "interaction" (could be modeled as a hyperedge)
Metabolic network

Proteins are enzymes.

They label the edges and their substrates are the nodes.

Valine, Leucine, and Isoleucine biosynthesis (from KEGG)
Yeast (aka *Saccharomyces cerevisiae*) interaction network (Two-hybrid)

GRID ([http://thebiogrid.org](http://thebiogrid.org))

8,742 edges

3,113 nodes (= proteins) (out of ~6,000 genes)
Human interaction network (Two-hybrid)

6,434 edges

4,083 nodes (out of ~25,000 genes)
How do we extract biological insight from these graphs?
1. What role does each protein play in the cell?

2. How do we uncover the true graph from noisy samples?

3. How do we compare interaction graphs across species?

4. What are the characteristics of interaction graphs?
Function Prediction

- Proteins with known function + network topology → function assignment for unknown proteins.
- Guilt by association
- Simple: Majority Rule:

Doesn’t take into account connections between neighbors
Or annotations at distance > one
Completing Defective Cliques (Yu, et al, 2006):

P, Q both adjacent to all nodes in clique (there are two \((n-1)\)-cliques that overlap by \((n-2)\) nodes) \(\Rightarrow\) likely that P,Q should be adjacent.
Aligning Networks

- Let $G_1 = (V_1, E_1)$, $G_2$, ... $G_k$ be graphs, each giving noisy experimental estimations of interactions between proteins in organisms $1,.., k$.

- If $G_i = (V_i, E_i)$, we also have a function:

$$\text{sim}(u, v) : V_i \times V_j \rightarrow \mathbb{R}$$

that gives the sequence similarity between $u$ and $v$. 

![Diagram of aligned networks](image-url)
Are there connection patterns that occur frequently?

“Frequently” = more often than you’d expect in a random graph with the same degree distribution.

Milo et al, 2002 found the feed forward motif over-represented in gene regulation networks.

Difficulty is larger motifs: subgraph isomorphism is NP-hard.
Other Uses of Graphs to Model Biological Problems:
Given sequences or distance matrix, find the tree that “best fits” that data.
Phylogenetic Networks

- Genetic material followed inconsistent evolutionary history
  - Horizontal gene transfer
  - Recombination

- Directed, acyclic graphs

- Given a set of sequences or distances build the DAG that best fits them (e.g.)
- Genome of *Mycoplasma genitalium*.
- Node = subsequence of length 75.
- Edge if two nodes overlap by 74 someplace in the genome.
- Paths of degree 1 compressed.
- Some Eulerian tour = the real genome.
Protein Design

- Position choice
- Interaction
- Side chain