56:198:582 Biological Networks
Lecture 1
Course goals

- Introduction to biological networks
  - Methods for modeling intracellular phenomena as networks of simple, interacting biochemical components
  - An understanding of current research directions
- Ability and confidence to tackle research
  - Problem sets
  - Project
Course organization

- Two complementary approaches to modeling and understanding biological networks
- Constraint-based modeling (Palsson)
  - System-wide
  - Metabolism
  - Steady-state
- Network motifs (Alon)
  - Small number of components at a time
  - Gene regulation
  - Dynamic
Why biological networks?

- 20th century biology was largely about understanding and characterizing components
- Now, we want to understand higher-level organization
- This permits deep understanding, simulation, design, and engineering

Introduction to Molecular Biology
“Central dogma” of molecular biology

DNA \rightarrow RNA \rightarrow Protein
What is DNA?

- DNA: Deoxyribonucleic acid
- DNA sequence: Bases from 5’ to 3’, e.g., ACTG or CAGT
- A is always paired with T, and C is always paired with G
- Knowing one strand tells you the other
DNA structure

Bases

A

G

C

T

U
DNA structure

Base pairing
DNA structure

Nucleoside

Nucleotide
DNA structure
DNA structure
DNA structure
RNA structure

- **Ribonucleic acid**, as opposed to **deoxyribonucleic acid**
Transcription
DNA structure

“Sense” strand

“Anti-Sense” strand

Reverse Complement

Convention:
Sense strand
Read 5′ → 3′
Transcription

- RNA polymerase (an enzyme) binds to DNA upstream of transcription start site
- Transcription factors (other enzymes) often also bind to the DNA and mediate binding of RNA polymerase
- RNA polymerase traverses the template (anti-sense) strand, producing a copy of the coding (sense) strand (with Us instead of Ts)
- RNA polymerase terminates by detaching from the template and releasing the RNA molecule
Protein structure

- Composed of amino acids
- Different amino acids have different side chains (R group)

Polymer is formed by peptide bonding
Protein structure

- Twenty amino acids (ignore selenocysteine)

  Arginine  Arg  R
  Histidine  His  H
  Lysine     Lys  K
  Aspartic acid  Asp  D
  Glutamic acid  Glu  E
  Serine    Ser  S
  Threonine    Thr  T
  Asparagine  Asn  N
  Glutamine  Gln  Q
  Cysteine    Cys  C
  Glycine     Gly  G
  Proline    Pro  P
  Alanine    Ala  A
  Isoleucine  Ile  I
  Leucine    Leu  L
  Methionine  Met  M
  Phenylalanine  Phe  F
  Tryptophan  Trp  W
  Tyrosine    Tyr  Y
  Valine      Val  V
Three RNA bases (a codon) are used to code for one amino acid according to the universal genetic code.

There are slight variations in the genetic code from organism to organism (and even within the same organism).
Translation

- Translation is performed by ribosomes (cellular components composed of RNA and proteins)
- tRNA (transfer RNA) carry amino acids to the ribosome
Metabolic Networks
What is metabolism?

- Chemical “engine” driving a cell
- Metabolism converts available raw materials into energy and into chemicals for building and maintaining cells
- Rough dichotomy of metabolic pathways:
  - *Catabolic pathways*: Used for breaking down chemical substrates
  - *Anabolic pathways*: Used for constructing cellular components, e.g., amino acids, nucleic acids
- A *metabolic network* or *metabolic reconstruction* is comprised of the full complement of metabolic reactions in a cell
An example: Glucokinase reaction

\[ \text{ATP} + \beta\text{-D-glucose} \rightleftharpoons \text{ADP} + \beta\text{-D-glucose-6-phosphate} \]

- ATP, β-D-glucose, ADP, etc. are *metabolites* (intermediates of metabolism)
A metabolic network

- Nodes: Metabolites
- Arcs/links: Reactions joining together two or more metabolites

www.expasy.ch/cgi-bin/show_thumbnails.pl
Reconstructing metabolic networks

► Generate a list of chemical reactions occurring in the cell
► Most reactions are enzymatically catalyzed (i.e. require a protein called an enzyme to make the reaction happen at an appreciable rate)
► Reconstructing a metabolic network mostly comes down to identifying the enzymes an organism makes, and the reactions catalyzed by those enzymes
Reconstructing metabolic networks

Four major sources of identifying reactions:

1. Biochemistry—An enzyme has been isolated from the organism and its function has been demonstrated.
2. Genomics—The DNA sequence of the organism indicates that it produces an enzyme of a particular function. Often, this inference is made by the similarity of the sequence to sequences of known function in other organisms.
3. Physiology and indirect information—Physiological information, e.g., the observed ability of the organism to produce a particular amino acid, requires that reactions be added to the network to make this possible.
4. In silico modelling data—Reactions are added to make the organism work in silico, e.g., reactions are added to ensure that all biomass components can be produced.
Enzyme commission (E.C.) numbers

- Enzyme commission (E.C.) numbers are used to systematically characterize enzymatic reactions.
- Sequenced genomes are typically annotated with predicted genes and associated E.C. numbers.
- The NCBI (National Center for Biotechnology Information) is the main repository for genomic information.
Enzyme commission (E.C.) numbers

Summary

Gene symbol: glik
Gene description: glucokinase
Primary source: EcoGene:EG12957
Locus tag: b2388
Seq related: ECOCYC:EG12957
Gene type: protein coding
RefSeq status: PROVISIONAL
Organism: Escherichia coli str. K-12 substr. MG1655 (strain: K-12 substrain: MG1655; origin: Escherichia coli K12) Lineage: Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacteriales; Enterobacteriaceae; Escherichia
Also known as: ECK2384; J2385
Summary: The cytoplasmic glucokinase is not required for growth on glucose as the carbon source, because glucose is phosphorylated during transport by the PTS system. [More information is available at EcoCyc: EG12957].
Enzyme commission (E.C.) numbers
**Enzyme commission (E.C.) numbers**

<table>
<thead>
<tr>
<th>Accepted Name</th>
<th>Alternative Name(s)</th>
<th>Reaction catalysed</th>
<th>Comment(s)</th>
<th>Cross-references</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glucokinase</td>
<td></td>
<td>ATP + D-glucose ⇌ ADP + D-glucose 6-phosphate</td>
<td>A group of enzymes found in invertebrates and microorganisms highly specific for glucose.</td>
<td></td>
</tr>
</tbody>
</table>

- **Biochemical Pathways; map number(s)**: B6 : C8
- **PROSITE**: PDOC00370 : PDOC00866
- **BRENDA**: 2.7.1.2
- **EC2PDB**: 2.7.1.2
- **PRIAM enzyme-specific profiles**: 2.7.1.2
- **KEGG Ligand Database for Enzyme Nomenclature**: 2.7.1.2
- **IUBMB Enzyme Nomenclature**: 2.7.1.2
Gene-protein-reaction (GPR) associations

- Besides listing chemical reactions, we wish to know the genes that give rise to particular enzymes (which catalyze particular reactions)
- These associations can be complicated

![Gene-protein-reaction (GPR) associations diagram](image-url)
E. coli reconstructions

Transcriptional Regulatory Networks
What is transcriptional regulation?

- Not all genes are transcribed (turned on / expressed) at all times
- Transcriptional regulation is the process that governs the expression of genes
- Metabolism: conversion of “small” molecules
- Transcriptional regulation: interaction of “large” molecules (primarily protein-protein or protein-DNA interactions)
An example: *lac* operon of *Escherichia coli*

- The operon (a set of genes transcribed together) consists of three genes: *lacZ*, *lacY*, and *lacA*, which are used by *E. coli* to break down lactose

- Two DNA-binding regulatory proteins:
  - LacI repressor binds to DNA only in the absence of lactose
  - CAP (cAMP Activator Protein) binds to DNA only in the absence of glucose
Transcriptional regulatory network motifs

- The proteins that bind to DNA and regulate gene expression are called transcription factors.
- Transcription factor genes and their regulatory targets (which can be other transcription factors) form an interaction network.
- In this network, some motifs occur regularly.
- We will discuss such network motifs more later in the course.

Shen-Orr et al., *Nat Genet*, 2002
Three fundamental data types

1. Component data. Data on regulatory components, e.g., transcription factors, binding sites
2. Interaction data. Data on interactions between components, e.g., protein-DNA and protein-protein interactions
3. Network state data. Gene expression and phenotyping data, especially under perturbations, e.g., genetic, environmental
Top-down data types

- High-throughput data that allow a large number of variables or states to be measured, e.g.,
  1. Experimental measurement of gene expression.
  2. Computational identification of promoter sites from genome sequence.
  3. Experimental identification of protein binding sites on DNA.
ChIP-seq

- ChIP: chromatin immunoprecipitation
- seq: massively parallel DNA sequencing

Mardis, Nat Methods, 2007
Bottom-up data types

- Data derived from classical biochemistry or genetics, focused on a small number of interactions
- Some transcription factors and their targets are well-studied (e.g., lac operon in *E. coli*), others are not
- RegulonDB: Database of interactions from biochemical data for *E. coli*. Contains data on 186 transcription factors.
- Bottom-data is laborious and slow to acquire, but is high-confidence
Regulation of the *Caulobacter crescentus* cell cycle

- Activated CtrA represses DNA replication initiation in swarmer cells

Laub et al., *Science*, 2000
Early development of the sea urchin *Strongylocentrotus purpuratus*

- Gene expression plays an important role in differentiating specialized cells in the embryo
Integrated regulatory-metabolic reconstructions

- Regulatory and metabolic reconstructions can be combined and used for predictive modeling
- *E. coli* reconstruction
- 104 regulatory genes
- 102 stimuli
- 906 metabolic genes
- 479 (of the 906) metabolic genes are regulated through boolean relationships, e.g.,

\[
GdhA = \text{NOT}((\text{Nac}) \text{ OR } (\text{glu-L(e)} > 0))
\]

Covert et al., *Nature*, 2004
Signaling Networks
What is cell signaling?

- Cell signaling is the process by which environmental signals are communicated from outside the cell to inside.
- Signal transduction often involves:
  1. Binding of a ligand (the signaling molecule) to an extracellular receptor
  2. Phosphorylation (adding of a phosphate group) of an intracellular enzyme
  3. Amplification and passage of the signal
  4. Resultant change in cellular function (e.g., increase in expression of a gene)
An example: G-protein signaling
An example: JAK-STAT signaling
Reconstruction of signaling networks

- Very few signaling networks have been reconstructed
- Reconstruction methods are often similar to those for regulatory networks (e.g., measuring expression after a genetic perturbation)
Yeast two-hybrid

- Yeast two-hybrid (Y2H) is an assay for determining protein-protein interactions
- Transcription factor (TF) consisting of binding domain (BD) and activation domain (AD)
- TF functions only when AD and BD are present (in close proximity, not necessarily binding)
- Reporter gene essential for survival of organism (yeast or E. coli)
- Bait protein fused with BD
- Prey protein fused with AD
- Organism survives only if bait and prey bind
- Noisy small degree of congruence between data sets
- False positives may result because proteins are spatially or temporally segregated in vivo
Human B-cell signaling network

- Human B-cells produce antibodies (immunoglobins, Ig) that recognize foreign antigens and are an important component of the adaptive immune response.