Analysis of Biological Networks
Network Analysis
Metabolic Networks
Spring 2020
Sharif University of Technology
Metabolic Networks

Part I:

Basic Concepts
Part I: Basic Concepts

• **Amino acids** are organic compounds composed of nitrogen, carbon, hydrogen and oxygen, along with a variable side chain group.

![Amino acid structure](image)

• Your body needs **20 different amino acids** to grow and **function properly**. Though all 20 of these are important for your health, only 9 amino acids are classified as essential.
Part I: Basic Concepts

- **Glucose** is a simple sugar with the molecular formula \( C_6H_{12}O_6 \). Glucose is the most abundant carbohydrates in your cells. In energy metabolism, glucose is the most important source of energy in all organisms.

- **Carbohydrate** is a biomolecule consisting of carbon (C), hydrogen (H) and oxygen (O) atoms, usually with a hydrogen–oxygen atom ratio of 2:1 (as in water) and thus with the empirical formula \( C_m(H_2O)_n \).
Part I: Basic Concepts

- **Epigenomics**: The epigenome is a multitude of chemical compounds that can tell the genome what to do. When epigenomic compounds attach to DNA and modify its function, they are said to have "marked" the genome. These marks do not change the sequence of the DNA. Rather, they change the way cells use the DNA's instructions.

- **Phenotype**: The set of observable characteristics of an individual resulting from the interaction of its genotype with the environment.

- **Genotype**: is the part of the genetic makeup of a cell, and therefore of any individual, which determines one of its characteristics (phenotype).
Part I: Basic Concepts

- **Genome annotation** (DNA annotation): is the process of identifying the locations of genes and all of the coding regions in a genome and determining what those genes do. An annotation is a note added by way of explanation or commentary.
- **ADP**: Adenosine diphosphate is an important organic compound in metabolism and is essential to the flow of energy in living cells.
- **ATP**: Adenosine triphosphate is an organic compound that provides energy to drive many processes in living cells, e.g. muscle contraction, nerve impulse propagation, and chemical synthesis.
Metabolic Networks

Part II:
Why Metabolic Networks
What is a Metabolic Network
Part II: Metabolic Networks

• The most widely used approaches to analyze omics data mainly focus on genomics, transcriptomics, and proteomics, through differential expression or network-based coexpression analysis.

• However, genes and their expression alone do not always constitute a reliable indicator of cellular phenotype.

• When characterizing a phenotypic outcome, relying solely on gene or protein expression profiles will miss the highly nonlinear interaction between these biological layers.
Part II: Metabolic Networks

• Such approaches often overlook the metabolic level, the dense network of biochemical reactions occurring in a cell with the aim of converting nutrients into energy and cellular building blocks.

• Metabolism is the best indicator for the cell physiological state, because it is the best-characterized network in biological systems and also the closest to the phenotype.

• Metabolic network is the main contributor to cellular behavior.
Part II: Metabolic Networks

• Metabolic Network is a key player in a number of diseases such as diabetes, neurodegenerative diseases, and cancer, where altered metabolism is now accepted as a hallmark.

• The availability of high-throughput data regarding multiple layers of biological organization (omics) allows mapping of cellular processes at the levels of genes, mRNA, proteins, and metabolites.

• In a single experiment, these measurements are often at both the genotype level and at the phenotype level (the form and function of the cell).
Part II: Metabolic Networks

- A fundamental question in systems biology is the definition and understanding of the genotype-phenotype relationship.
Part II: Metabolic Networks

• A mechanistic link between genotype and phenotype is offered by genome-scale metabolic models that contain all known biochemical reactions occurring in a cell.

• Such models have been generated taking into account decades of studies in biochemistry and in most cases are able to predict the cellular phenotype with high accuracy.

• Constraint-based modelling is the most widely used approach to model the behavior of metabolism, often assuming that cells have to fulfill a given task or to optimize the production of a given compound.
Part II: Metabolic Networks

• These models have two main advantages:
  – They do not need dynamic or kinetic data as they are based on mass balance across the metabolic network.
  – They are suitable for integration of different omic layers at genome scale to improve their predictive performance.

• Multiomic **vertical integration** methods have been proposed to include omic layers (mainly transcriptomics and proteomics).

• Conversely, horizontal integration methods have focused on modelling different environments, cancers or growth conditions starting from the same model.
Part II: Metabolic Networks

• Multiomic integration in genome-scale models has provided a mechanistic link between the genotype and their phenotypic observables.

• This is a key added feature that such models possess if compared with genome-wide association studies (GWAS), which are able to associate gene variations to phenotypic traits, but not to provide a mechanistic explanation of the associations observed.

• In the next section we consider Genome-Scale Metabolic Networks, Constraint-Based Modelling and Flux Balance Analysis, and Multiomic Flux Balance Analysis.
Metabolic Networks

Part III:
Metabolic Network Modeling
Part III: Metabolic Network Models

Reconstruction of Genome-Scale Metabolic Networks

- The human knowledge on metabolism was condensed on the famous Boehringer Mannheim wall chart of metabolic pathways.
Part III: Metabolic Network Models

Reconstruction of Genome-Scale Metabolic Networks

• The human knowledge on metabolism was condensed on the famous Boehringer Mannheim wall chart of metabolic pathways.
Part III: Metabolic Network Models

Reconstruction of Genome-Scale Metabolic Networks

• Until the emergence of the first complete bacterial genome in 1995, it was impossible to access the complete and species-specific metabolic network (also see https://www.ncbi.nlm.nih.gov/).

• Many of these networks are available online: Kyoto Encyclopedia of Genes and Genomes (KEGG), EcoCyc is an extensively human-curated database, and KEGG and BioCyc maintain databases for many organisms but with little human curation effort.
Part III: Metabolic Network Models

Reconstruction of Genome-Scale Metabolic Networks

- To reconstruct a genome-scale metabolic (GSM) network we need to follow a workflow:
  - Given a genome annotation extract the Enzyme Commission (EC) numbers.
  - Use a reaction database to find the corresponding reactions for the genome and associated EC numbers.
  - Convert reactions into a connection matrix.
  - Extract connections via currency metabolites to obtain the metabolic network.
Part III: Metabolic Network Models

Reconstruction of Genome-Scale Metabolic Networks

1. Genome annotation
2. Enzymes, e.g., EC: 2.7.1.2; Enzyme name: glucokinase
3. Reaction database
4. Reactions, e.g., \( \text{ATP} + \text{d-glucose} \rightarrow \text{ADP} + \text{d-glucose 6-phosphate} \)
5. Reversibility information
6. Connection matrix, e.g., C00031-C00092
7. Remove connections via currency metabolites
8. Metabolic network
Part III: Metabolic Network Models

Reconstruction of Genome-Scale Metabolic Networks

- Currency metabolites are the metabolites that are mainly used as carriers for transferring electrons and certain functional groups (hydrogen, phosphate, amino group, one carbon unit, methyl group, etc.).
- For example, in the reaction Glucose + ATP = G6P + ADP, ADP and ATP are currency metabolites for transferring phosphate to glucose.
- The connection matrix should be further treated by removing connections via currency metabolites such as H2O, CO2, and ATP.
- The currency metabolites are often not shown in the metabolic pathway maps in KEGG.
Part III: Metabolic Network Models

Reconstruction of Genome-Scale Metabolic Networks

• When considering the connections through currency metabolites, structure analysis often produces biologically meaningless results.

• For example, in the glycolysis pathway, the path length from glucose to pyruvate should be nine in terms of biochemistry (number of reaction steps in the pathway).

• However, if ATP and ADP are considered as vertices in the network there would be only two steps from glucose to pyruvate (the first reaction uses glucose and produces ADP, while the last reaction consumes ADP and produces pyruvate).
Part III: Metabolic Network Models

Reconstruction of Genome-Scale Metabolic Networks

glycolysis pathway: conventional (left), Network based on currency metabolites (right)
Part III: Metabolic Network Models

Reconstruction of Genome-Scale Metabolic Networks

- A simple way is to exclude the top-ranked metabolites based on their connection degree (number of edges connected with a metabolite).
- The problem is that certain primary metabolites such as pyruvate may also have high connection degrees.
- Moreover, currency metabolites cannot be defined by compounds but should be defined according to the reaction.
- It is necessary to check reactions manually in order to remove the biologically meaningless connections.
Part III: Metabolic Network Models

Reconstruction of Genome-Scale Metabolic Networks

• Therefore, the removal of connections through currency metabolites is an essential step to draw biologically meaningful conclusions from graph analysis of metabolic networks.
Part III: Metabolic Network Models
Connectivity and Centrality in Metabolic Networks

• Metabolic networks exhibit typical characteristics of small-world network; power law connection degree distribution, high cluster coefficients and a short network diameter.

• The structure of metabolic networks still have the characteristics of a scale-free network after deleting the connections through currency metabolites.

• The connection degree is defined as the number of connections linked with each metabolite (vertex). Considering the direction, the number of connections starting from the metabolite is called output degree, and the number of connections ending at the metabolite is called input degree.
Part III: Metabolic Network Models
Connectivity and Centrality in Metabolic Networks

• Let $P(k)$ denote the fraction of vertices that have a $k$-degree of outputs. It is calculated by dividing the number of metabolites, which had $k$ output connections with the total number of metabolites in the organism.

• Except for the first point with $k = 1$, a clear power law distribution (linear relations in the logarithmic scale coordinates) can be detected: $P(k) = \alpha k^{-\gamma}$

• For the log–log plot: $\log P(k) \approx \log \alpha - \gamma \log k$

• where $-\gamma$ is the slope of the linear approximation of the curve.
Part III: Metabolic Network Models

Connectivity and Centrality in Metabolic Networks

- The power law degree distribution exists in metabolic networks of all the organisms.

Output degree distribution in four typical organisms. \( P(k) \) is the fraction of vertices that have a k-degree of output connections. The original data have been logarithmically binned; hsa: Homo sapiens (eukaryote), eco: Escherichia coli (gram negative bacteria), bsu: Bacillus subtilis (gram positive bacteria), ape: Aeropyrum pernix (archaea).
Part III: Metabolic Network Models
Connectivity and Centrality in Metabolic Networks

• Degree centrality can show the vertices with highest number of connections (hubs) in the network. In a metabolic network, hubs are the vertices that can be converted into more metabolites.

• Vertices with the highest betweenness centrality are the ones with the highest number of shortest pathways going through them. For a metabolic network, this may mean vertices that participate in more metabolites conversions.

• This is not always true because in many cases the shortest path is not the one used in the reality.
The biosynthesis of arginine from glutamate: The shortest path from glutamate to arginine is the path B. But the metabolic pathway actually used for the biosynthesis of arginine is path A which is longer than B.
Part III: Metabolic Network Models

Connectivity and Centrality in Metabolic Networks

- Closeness centrality identifies vertices in the central part of a network and vertices in the periphery part.
- Vertices in the central part have a shorter distance to other vertices in the network.
- In a metabolic network, this means that these metabolites can be converted to others in fewer steps.

- Three Metabolites with Highest Betweenness Centrality for B. Subtilis Network:

<table>
<thead>
<tr>
<th>Rank</th>
<th>Betweenness centrality</th>
<th>KEGG</th>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.432132</td>
<td>C00022</td>
<td>Pyruvate</td>
</tr>
<tr>
<td>2</td>
<td>0.310035</td>
<td>C00117</td>
<td>d-Ribose 5-phosphate</td>
</tr>
<tr>
<td>3</td>
<td>0.297295</td>
<td>C00119</td>
<td>5-Phospho-alpha-d-ribose 1-diphosphate</td>
</tr>
</tbody>
</table>
Part III: Metabolic Network Models
Connectivity and Centrality in Metabolic Networks

- Three Metabolites with Highest Closeness Centrality for B. Subtilis Network.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Closeness centrality</th>
<th>KEGG&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.124260</td>
<td>C00022</td>
<td>Pyruvate</td>
</tr>
<tr>
<td>2</td>
<td>0.122129</td>
<td>C04442</td>
<td>2-Dehydro-3-deoxy-6-phospho-D-gluconate</td>
</tr>
<tr>
<td>3</td>
<td>0.122058</td>
<td>C11437</td>
<td>1-Deoxy-D-xylulose 5-phosphate</td>
</tr>
</tbody>
</table>

- Ten Metabolites with Highest Degree Centrality for B. Subtilis Network.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Degree centrality</th>
<th>KEGG&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.038095</td>
<td>C00022</td>
<td>Pyruvate</td>
</tr>
<tr>
<td>2</td>
<td>0.026190</td>
<td>C00024</td>
<td>Acetyl-CoA</td>
</tr>
<tr>
<td>3</td>
<td>0.023810</td>
<td>C00025</td>
<td>Glutamate</td>
</tr>
</tbody>
</table>
Part III: Metabolic Network Models

Modularity & Decomposition of Metabolic Networks

• In biochemistry, modules consisting of several interacting bioreactions or metabolic pathways that build discrete functional units of metabolism.

• These modules are further nested to form a complex metabolic network.

• On the other hand, the structural analysis of metabolic networks indicates a small-world property for metabolic networks.

• Therefore, all the vertices in the whole network are linked through a short path and modularity seems to be missing in this small-world network.
Part III: Metabolic Network Models
Modularity & Decomposition of Metabolic Networks

• Solution: Utilize a hierarchical modularity model for metabolic networks.

• According to this model, metabolic networks of organisms are organized as many small, but highly connected modules that combine in a hierarchical manner to larger, less cohesive units.

• Several studies using concepts such as the reaction betweenness centrality distribution and the dependency of metabolites have further verified that metabolic networks are organized in a hierarchical way.
Part III: Metabolic Network Models
Modularity & Decomposition of Metabolic Networks

• Methods for a rational decomposition of metabolic network into relatively independent functional subsets are essential to better understand the modularity and organization principle of large-scale, genome-wide networks.

• Several methods such as elementary flux mode analysis and extreme pathway analysis have been developed for analyzing the pathway structure of metabolic networks.

• For large-scale networks reconstructed from genome information, decomposition methods should be first used to divide the whole network into small subsystems.

• The pathway structure of these subsystems may then be properly analyzed by these methods.
Part III: Metabolic Network Models
Modularity & Decomposition of Metabolic Networks

- Some methods make use of a hierarchical clustering by using different criteria for clustering such as distance, connection degree, and modularity coefficient.

- Decompose a network based on distance is shown below. One can then build a hierarchical tree (dendrogram) for finding clusters with similar reaction properties. To select the best number of modules the modularity coefficient can be used.
Part III: Metabolic Network Models

Modularity & Decomposition of Metabolic Networks

• A good partition of a network into modules must comprise many within-module edges and as few as possible between-module edges.

• A better measure to achieve optimal modularity is the Modularity measure:

\[
M = \sum_{s=1}^{N_M} \left[ \frac{l_s}{L} - \left( \frac{d_s}{2L} \right)^2 \right]
\]

• where \(N_M\) is the number of modules, \(L\) is the number of edges in the network, \(l_s\) is the number of edges between vertices in module \(s\), and \(d_s\) is the sum of the degrees (number of edges) of the vertices in module \(s\).
Part III: Metabolic Network Models

Modularity & Decomposition of Metabolic Networks

- If the number of within-communities is no better than random, we will get $M = 0$.
- Values approaching $M = 1$, which is the maximum, indicate a strong community structure.
- For metabolic networks, this value is around 0.8, which shows the strong community structure in this kind of networks.
- In the modularity-based decomposition algorithm each vertex in the network is first considered to be a module itself.
Part III: Metabolic Network Models
Modularity & Decomposition of Metabolic Networks

• At the beginning of modularity-based algorithm, the number of modules is the same as the number of vertices.

• Then, the modules are joined in pairs to form new modules. The criterion to choose the modules to join is based on changes in modularity.

• The change in the modularity caused by the union of modules $i$ and $j$:

$$
\Delta M = \frac{l_{i,j}}{L} - \left(\frac{d_{i,j}}{2L}\right)^2
$$
Part III: Metabolic Network Models

Modularity & Decomposition of Metabolic Networks

- where $l_{i,j}$ is the number of edges between the two modules, $L$ is the total number of edges in the network, and $d_{i,j}$ is the total degree of the vertices in the two modules.

- New modules that results in the greatest increase (or smallest decrease) in modularity will be created first.

- The progress of the algorithm can be represented as a dendrogram like in the distance-based method.

- Decomposition of E. coli network using this modularity method illustrated in the next page.
Part III: Metabolic Network Models

Modularity & Decomposition of Metabolic Networks

• In this figure, vertices represent the modules while vertex size represents the number of metabolites included in the module. Edge labels show the number of connections between modules.
Part III: Metabolic Network Models
Elementary Flux Modes & Extreme Pathways

• In a metabolic network analysis how do we identify and analyze metabolic pathways at a genome scale?
• Two related concepts address this issue:
  – Elementary Flux Modes (EFMs)
  – Extreme Pathways (EPs)
• EFM is defined as a minimal set of enzymes that can operate at a steady state with all irreversible reactions proceeding in the appropriate direction.
• Minimal means that if only the enzymes belonging to this set were operating, complete inhibition of one of these would lead to termination of any steady-state flux in the system.
Part III: Metabolic Network Models

Elementary Flux Modes & Extreme Pathways

• Extreme pathways are a unique and minimal set of vectors that completely characterize the steady-state capabilities of genome-scale metabolic networks.

• The length of an extreme pathway is the number of reactions that comprise it.

• Reaction participation is the percentage of extreme pathways that utilize a given reaction.

• The concept of EPs is related to EFMs. However, for EFMs reversible reactions are unnecessary to be split into two opposite irreversible reactions (see this clip).
Metabolic Networks:
Part IV:
Case Study
with Cytoscape
Case Study

D. Melanogaster Metabolic pathways

• Drosophila melanogaster is a species of fly (the taxonomic order Diptera) in the family Drosophilidae.

• D. melanogaster (fruit fly) is typically used in research owing to its rapid life cycle, relatively simple genetics with only four pairs of chromosomes, and large number of offspring per generation.
Database

• KEGG

KEGG (Kyoto Encyclopedia of Genes and Genomes) is a database resource that integrates genomic, chemical and systemic functional information. In particular, gene catalogs from completely sequenced genomes are linked to higher-level systemic functions of the cell, the organism and the ecosystem.
Data

• You can access the KEGG website at: https://www.genome.jp/
Data

- Search for Drosophila melanogaster
Data

- Find *Drosophila melanogaster* Genome (dme) and click it!
Data

- You can find any information about *Drosophila melanogaster* (fruit fly) genome, click on KEGG pathway.
Data

- In this screen you can find all pathways of *Drosophila melanogaster* (fruit fly), we choose one of the metabolic pathways for example Galactose metabolism
Galactose Metabolism

- Galactose metabolism network for the fruit fly.
Galactose Metabolism

- Galactose, which is metabolized from the milk sugar, lactose (a disaccharide of glucose and galactose), enters glycolysis by its conversion to glucose-1-phosphate (G1P).
Galactose Metabolism

- Galactose can exist in two different stereoisomeric forms; α-D-galactose and β-D-galactose. The α-form is that which is metabolized in the Leloir pathway. Conversion of the β-form of galactose to the α-form requires the enzyme galactose mutarotase encoded by the GALM gene (also known as aldose 1-epimerase). The first reaction of the Leloir pathway is the phosphorylation of α-D-galactose by galactokinase to yield galactose-1-phosphate.
Download XML Data

- To use and analyze this data you should download it as KGML (XML) format.
Cytoscape

• We use Cytoscape version 3.8.0 for analysis.
Import KGML file into Cytoscape

• Install KEGGParser app from Cytoscape and import your KGML file into Cytoscape.
Network

• The network is ready!
MetDisease

- Install MetDisease app from Cytoscape.
- The MetScape plugin for Cytoscape provides a bioinformatics framework for the visualization and interpretation of metabolomic and expression profiling data in the context of human metabolism.
- It allows users to build and analyze networks of genes and compounds, identify enriched pathways from expression profiling data, and visualize changes in metabolite data.
MetDisease

- MetDisease uses Medical Subject Headings (MeSH) disease terms mapped to PubChem and KEGG compounds through literature to annotate compound networks.
- Use Labels as attribute. Click OK!
Disease Association

- Now you can see diseases and their corresponding nodes associated to this pathway.
Other Analysis

You may use other apps such as KEGGscape and the apps we used in previous case study to analyze your network.

Try to perform some analysis such as:

• Module Detection
• Comparison with random network
• Node degree distribution and other parameters like betweenness centrality.
Metabolic Networks

Part V:
Summary & Conclusions
Summary of PPI Networks

• Metabolic network composed of metabolites and their interconversions (biochemical reactions) in an organism.

• The sequencing of genomes and functional genomics help us to reconstruct and understand the structure and function of metabolic networks.

• Computational tools and biological concepts are being utilized for reconstruction, visualization, and graph representation of genome-scale metabolic networks for structural analysis such as connectivity and centrality.
Summary of PPI Networks

• Present methods primarily address static properties and functions of metabolic networks.
• New tools and concepts are needed to capture the dynamic structure and function of metabolic networks.