Motifs, Modules and Pathways

Proteomes Interactomes and Biological Networks

November 26, 27 and 29, 2019

Emidio Capriotti
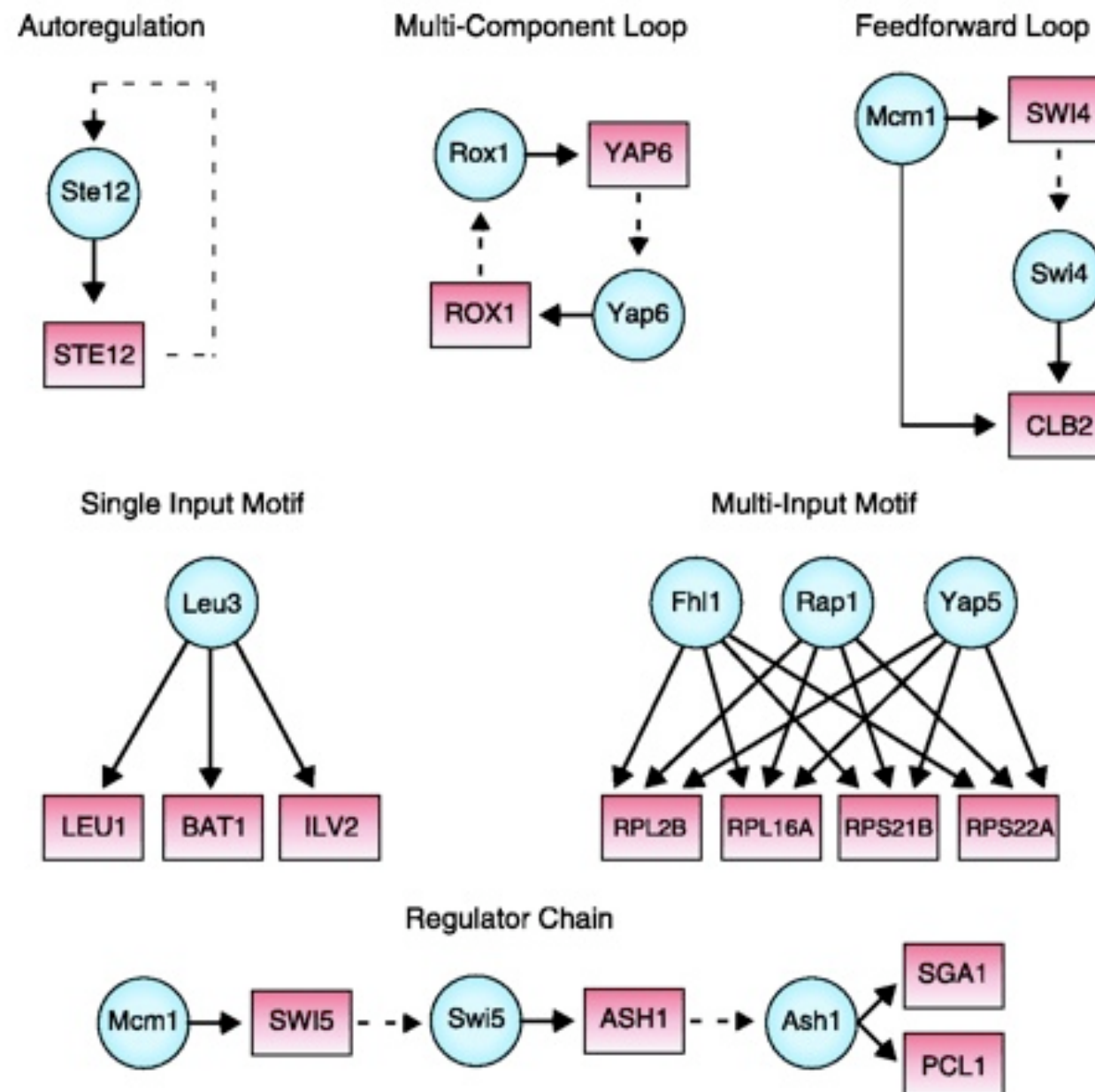http://biofold.org/

Department of Pharmacy and
Biotechnology (FaBiT)
University of Bologna

Biomolecules
Folding and
Disease

Network Motifs

Network analysis is important for detecting network motifs, which are recurrent and statistically significant sub-graphs or patterns.
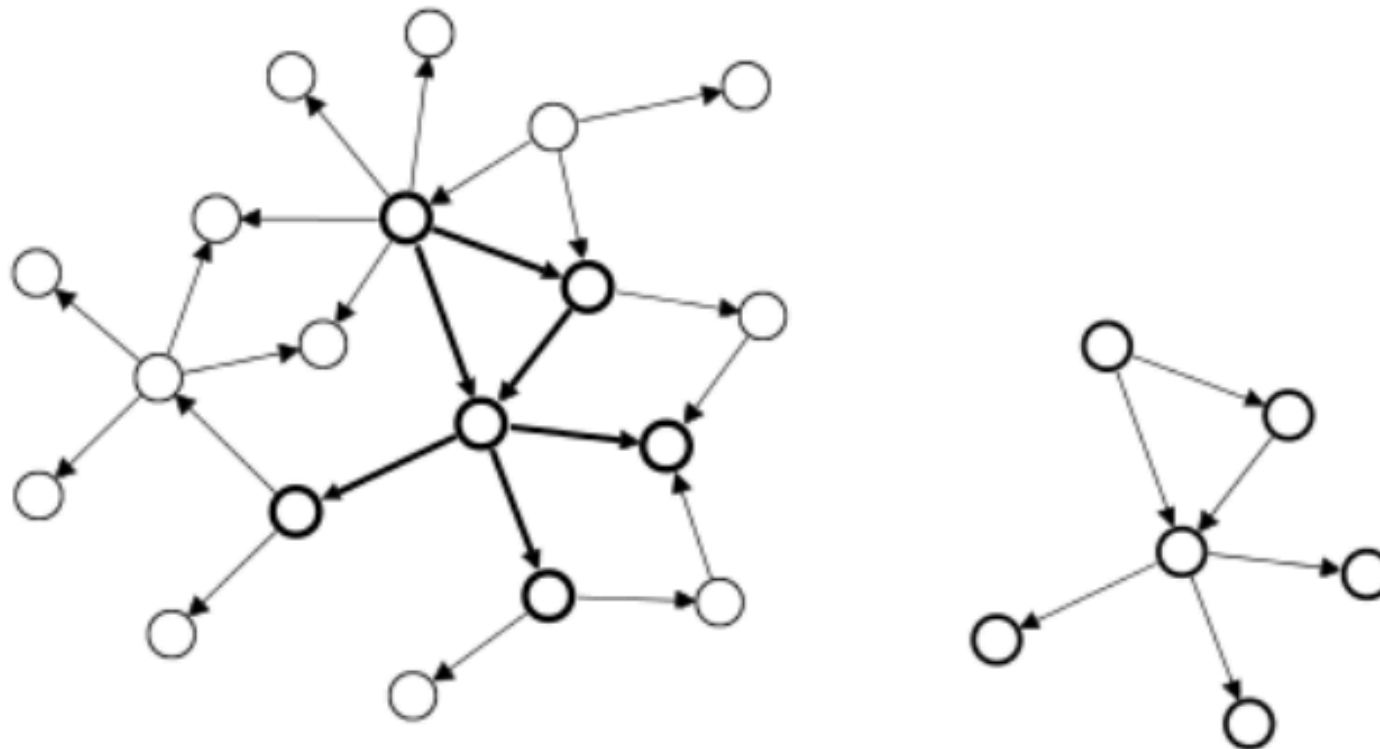
Motif Matching

A match of a motif G' in the target graph G = (V, E) is a subgraph G" = (V", E") which is isomorphic to motif G'

Two graphs G' and G" are isomorphic if there is a bijective mapping between the edge and vertex identities

i.e. G' is transformed to G" by changing the vertex and edge identities

Problem Complexity

The complexity of graph isomorphism is in the 'grey area' of complexity:

- It belongs to NP class of problems (problems where solution is easy to verify once found)

- It is not known if graph isomorphism belongs to P class of problems (problems that can be solved efficiently)

- It is not known if graph isomorphism is NP-complete (problems that are believed to be hard ot solve but easy to verify)

- Subgraph isomorphism, checking if a subgraph G'' that is isomorphic to given graph G' exists in a larger graph G, is known to be NP-complete

- No hope for really fast algorithms for finding motifs.

Statistical Significance

A motif is a statistically overrepresented pattern of local interactions in the network

- Overrepresentation = occurring more frequently than expected by chance

- The motif has emerged several times therefore it has been conserved in the evolution of the network

- The rationale is that overrepresentation may denote possible conservation of the function

Significance tests

The statistical significance can be tested calculating the z-score of the presence of the motif on a set of randomly generated graphs obtained with

- Generation of random networks with the Erdos-Renyi algorithm

- Random shuffling of the edges

Detection of Motifs

Networkx allow to select a subgraph of a the whole graph and verify if two graphs are isomorphic

```
>>> g = nx.Graph().add_nodes_from([(1,2),(1,3)])


>>> mot = nx.Graph().add_nodes_from([("A","B")])


>>> g1 = g.subgraph([1,2])


>>> nx.is_isomorphic(g1,mot)
```

Exercise

Given the Feed Forward Loop (FFL) and 3-Cycle write the code to detect the motif in the graph with 6 nodes and 8 edges.

Calculate occurrence on random network and the z-score.

Feed Forward Loop
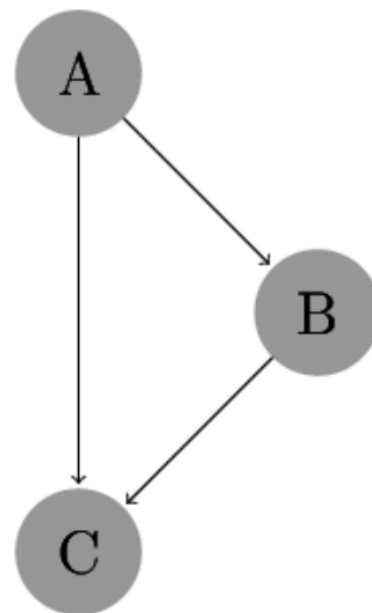
3-Cycle
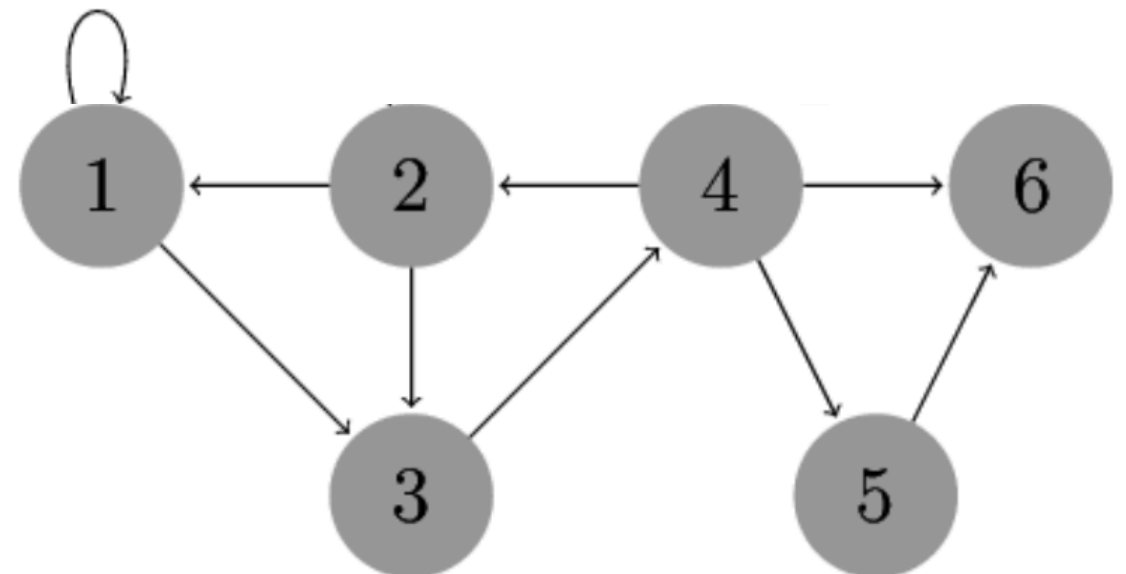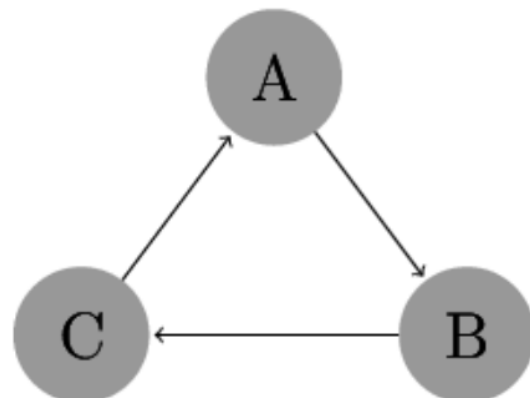
RegulonDB

Database of Escherichia coli K-12 Transcriptional Regulatory Network

http://regulondb.ccg.unam.mx/

Regulation Data

The regulation data includes information about the transcription factors (TF) that activate or repress the expression of the genes with associated supporting evidences.

```
# Release: 10.6.2 Date: 10-04-2019
# _____#
# Columns:
#  (1) Transcription Factor (TF) name
#  (2) Gene regulated by the TF (regulated gene)
#  (3) Regulatory effect of the TF on the regulated gene (+ activator, - repressor, +- dual, ? unknown)
#  (4) Evidence that supports the existence of the regulatory interaction
#
AcrR    acrA    -       [BCE, BPP, GEA, HIBSCS] Strong
AcrR    acrB    -       [BCE, BPP, GEA, HIBSCS] Weak
AcrR    acrR    -       [AIBSCS, BCE, BPP, GEA, HIBSCS] Weak
AcrR    flhC    -       [GEA, HIBSCS]   Weak
AcrR    flhD    -       [GEA, HIBSCS]   Weak
AcrR    marA    -       [BPP, GEA, HIBSCS]      Strong
AcrR    marB    -       [BPP, GEA, HIBSCS]      Strong
AcrR    marR    -       [BPP, GEA, HIBSCS]      Strong
AcrR    micF    -       [AIBSCS]        Weak
AcrR    soxR    -       [BPP, GEA, HIBSCS]      Strong
```

Nodes ad Edges

With networkx we can assign attributes to nodes and edges

```
>>> G=nx.DiGraph()

>>> G.add_node(1, color='blue')

>>> G.add_node(2, color='red')

>>> G.add_edge(1, 2, sign='+')

>>> G.node[1]

>>> G.edge[1][2]
```

Matches Node and Edges

Matches can be performed based on node and edges attributes

```
>>> import networkx.algorithms.isomorphism as iso

>>> em=so.categorical_edge_match('sign'='+')

>>> nm=iso.categorical_node_match('color'='red')

>>> nx.is_isomorphic(G1,G2,edge_match=em, node_match=nm)
```
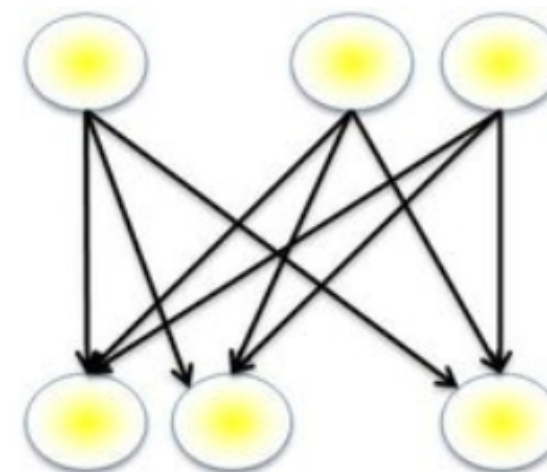
Exercise

Write a program to analyze the RegulonDB network considering only data with strong supporting information.

• Find the TF that regulates more genes (activation and suppression)

• Find the gene that is regulated by more TFs

• Find the Double-Positive Feedback loop and Multi-Input module
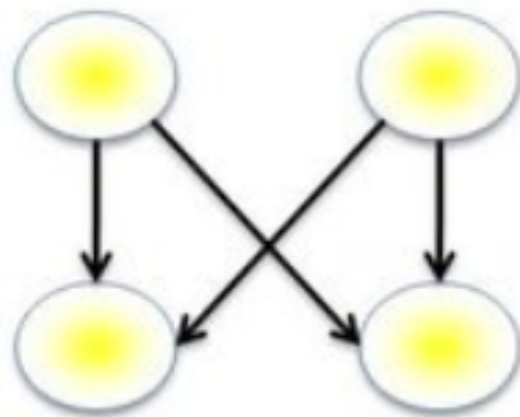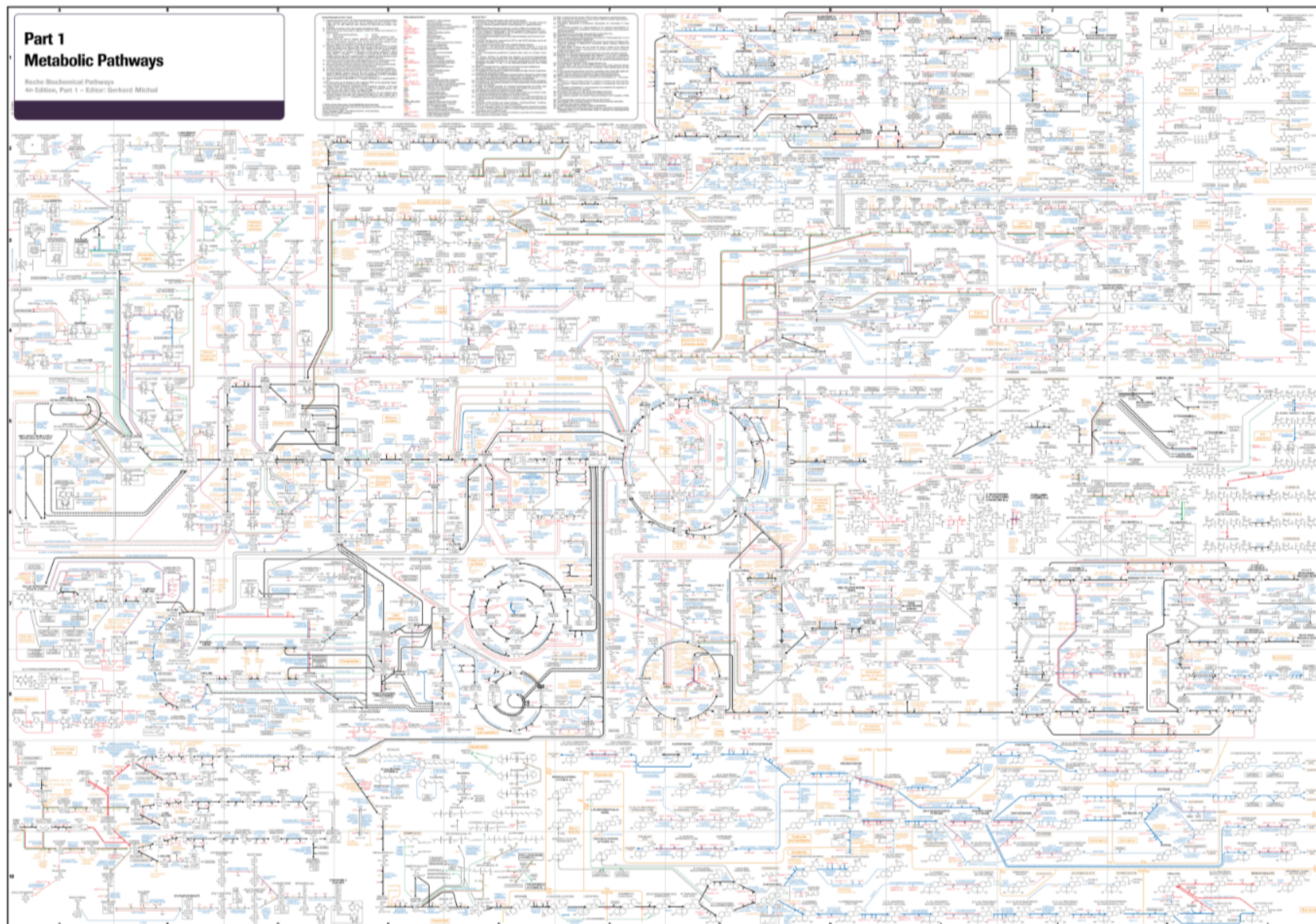

Double-Positive Feedback loop

Multi-Input module

Metabolic Pathway

Metabolic pathway is a linked series of chemical reactions occurring within a cell.

Part 1
Metabolic Pathways

Roche Biochemical Pathways
4th Edition, Part 1 – Editor: Gerhard Michal

http://biochemical-pathways.com/

KEGG Database

It is the Kyoto Encyclopedia of Genes and Genomes. It collects many databases the most important one is KEGG Pathway which contains maps divided in 7 groups

KEGG PATHWAY Database

Wiring diagrams of molecular interactions, reactions and relations

| Menu | **PATHWAY** | **BRITE** | **MODULE** | **KO** | **GENES** | **LIGAND** | **NETWORK** | **DISEASE** | **DRUG** | **DBGET** |

Select prefix   Enter keywords

[map] [Organism]   [   ] [Go]   Help

[New pathway maps | Update history]

Pathway Maps

KEGG PATHWAY is a collection of manually drawn pathway maps representing our knowledge on the molecular interaction, reaction and relation networks for:
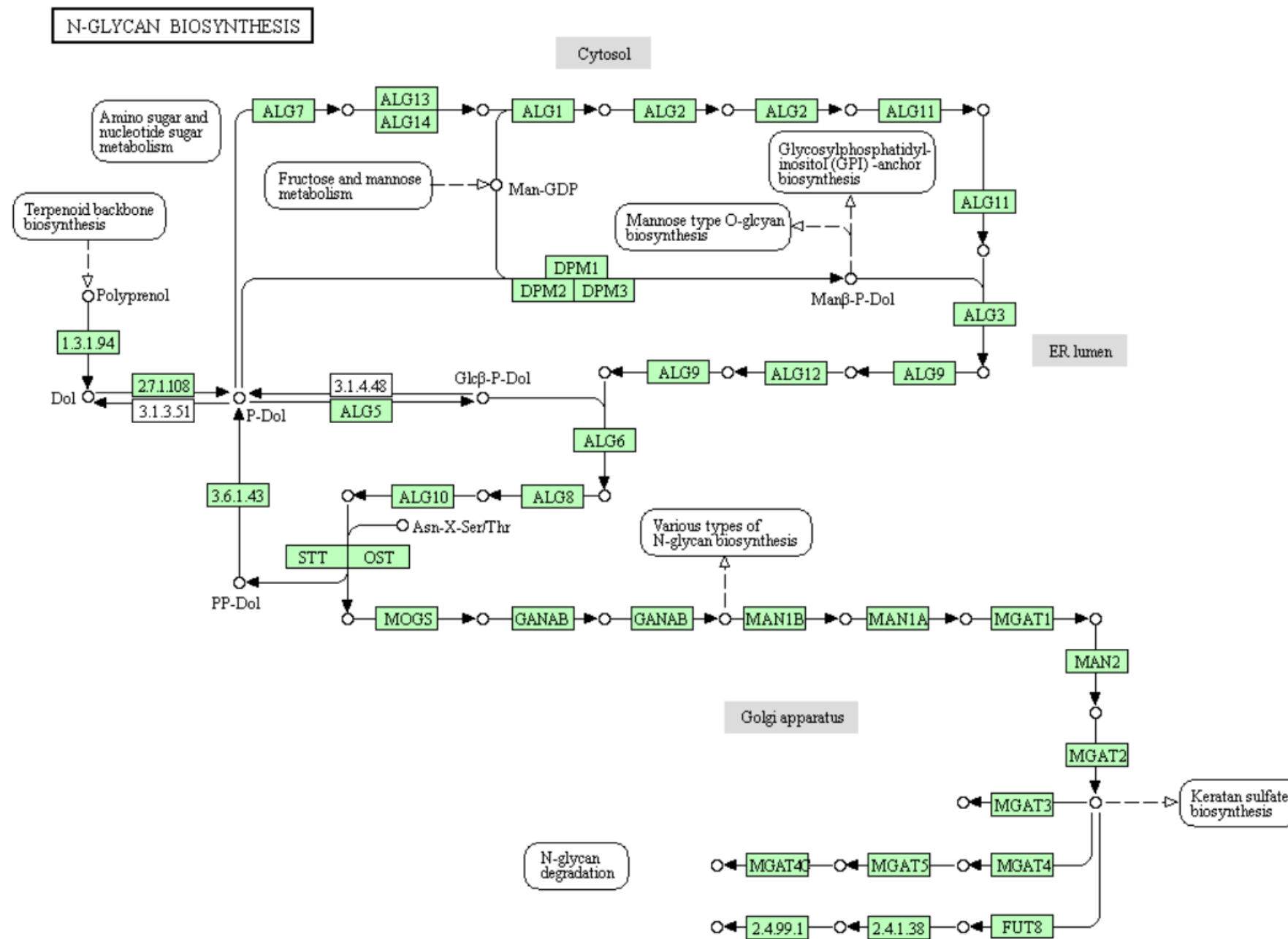
1. **Metabolism**
   Global/overview Carbohydrate Energy Lipid Nucleotide Amino acid Other amino Glycan
   Cofactor/vitamin Terpenoid/PK Other secondary metabolite Xenobiotics Chemical structure
2. **Genetic Information Processing**
3. **Environmental Information Processing**
4. **Cellular Processes**
5. **Organismal Systems**
6. **Human Diseases**
7. **Drug Development**

KEGG PATHWAY is the reference database for pathway mapping in **KEGG Mapper**.

https://www.genome.jp/kegg/pathway.html

Pathway Map

It is a representation of a set of reactions in which the reactants, products, and intermediates of an enzymatic reaction, known as metabolites, are modified by a sequence of reactions catalyzed by enzymes.

KEGG Data

Given a pathway KEGG provide several information about the genes, the metabolites and the enzymes involved in the series of reactions

KGML Format

The KEGG Markup Language (KGML) is an exchange format of the KEGG pathway maps.

The KGML files for metabolic pathway maps contain two types of graph object patterns:

- boxes (enzymes) are linked by "relations"
- circles (chemical compounds) are linked by "reactions".

The information are provided in xml format. Enzymes are always indicated with Enzyme Commission number (EC number).

The EC number is composed by four numbers separated by periods. Those numbers represent a progressively finer classification of the enzyme.

Network Modules

A module is a set of genes/proteins performing a distinct biological function are characterized by coherent behavior with respect to certain biological property.
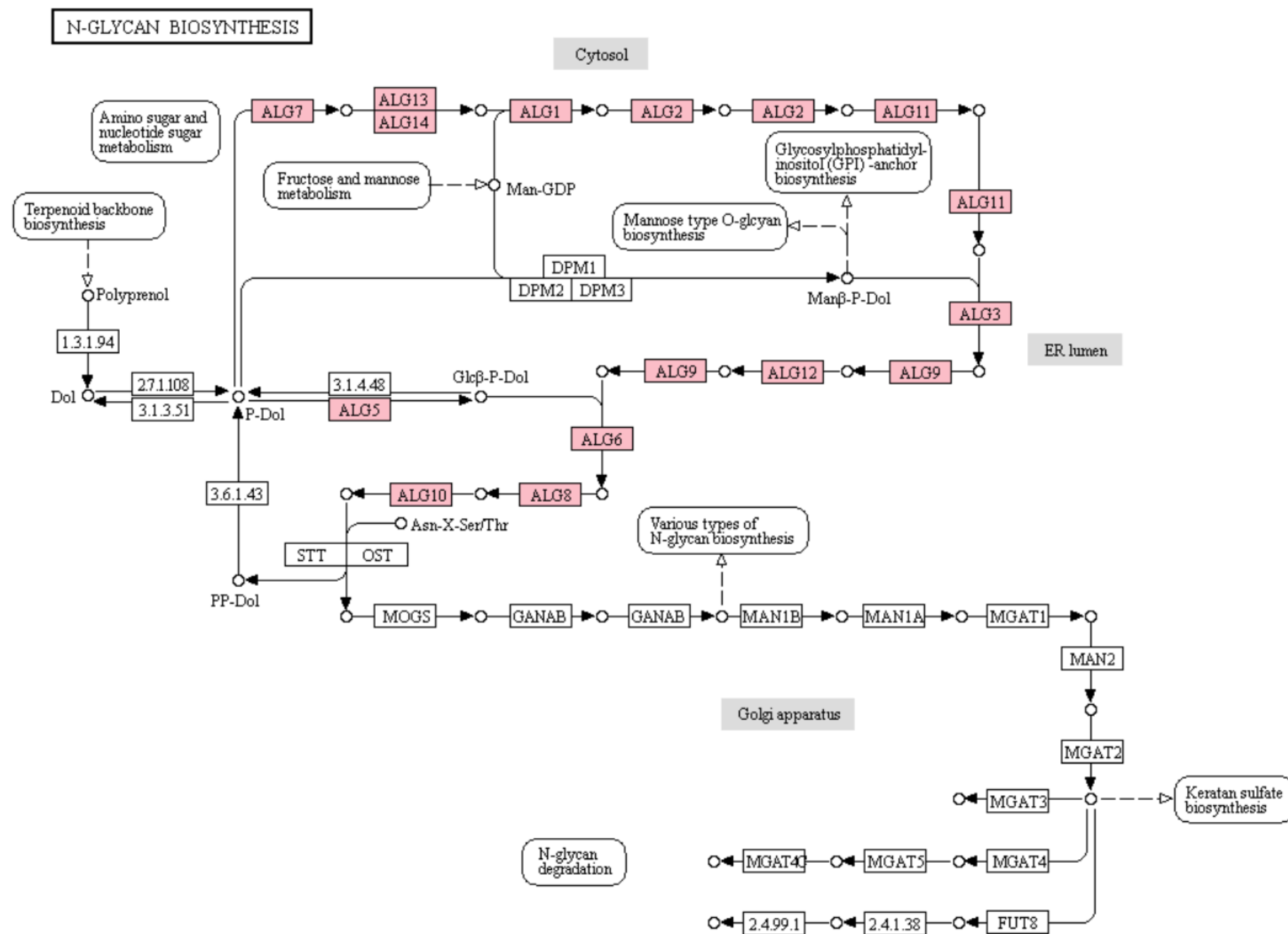
Exercise

Download kegg-kgml-parser-python from GitHub and install on your machine.

- Use the program to parse a KGML file and generate a network with networks.

- Identify the module N-glycan precursor biosynthesis (M00055) and visualize it in the graph assigning different color to its nodes.