

# Motifs, Modules and Pathways

Proteomes Interactomes and Biological Networks

Emidio Capriotti

<http://biofold.org/>



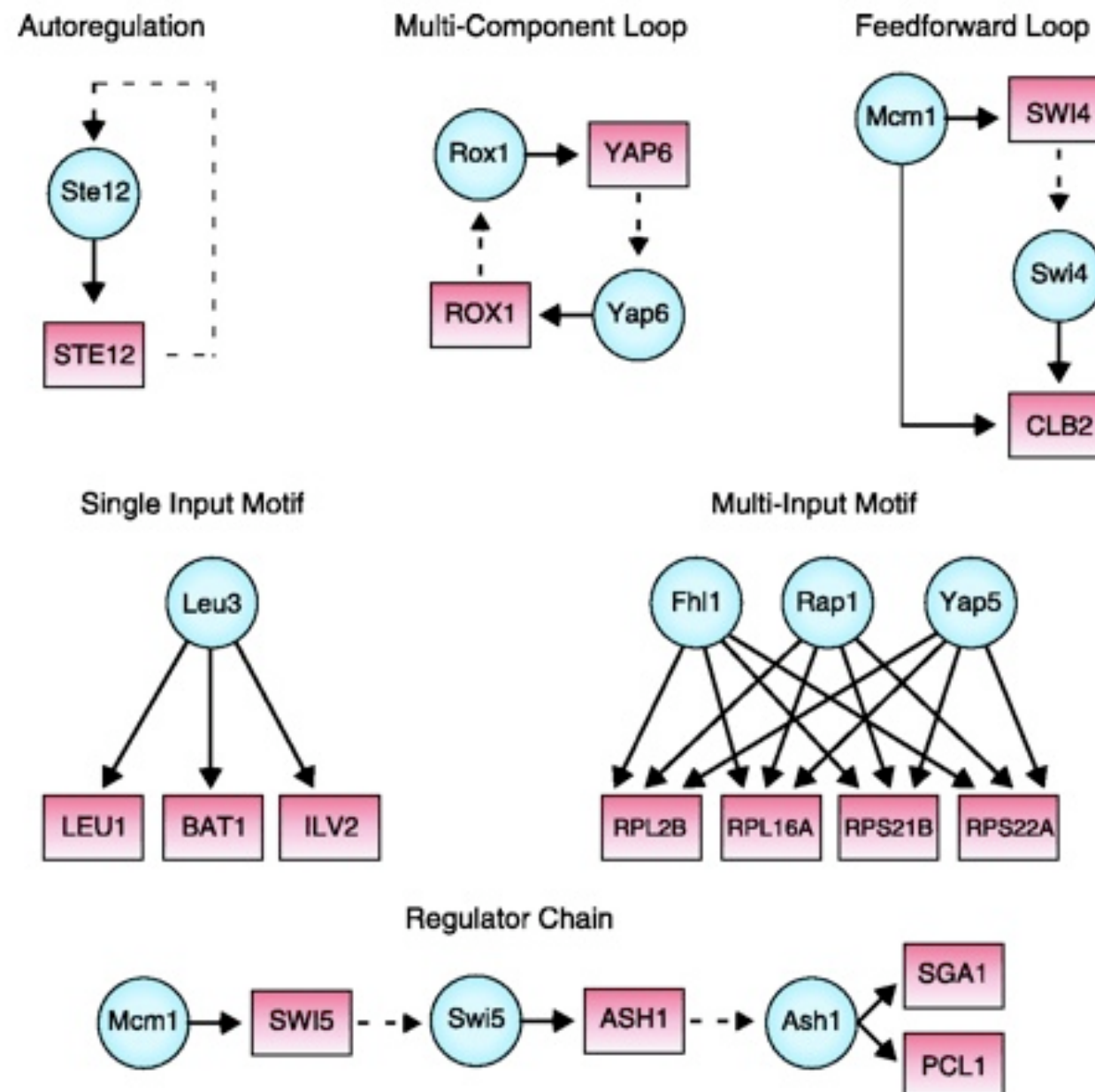
**Biomolecules**  
**Folding and**  
**Disease**

Department of Pharmacy and  
Biotechnology (FaBiT)  
University of Bologna



# Network Motifs

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

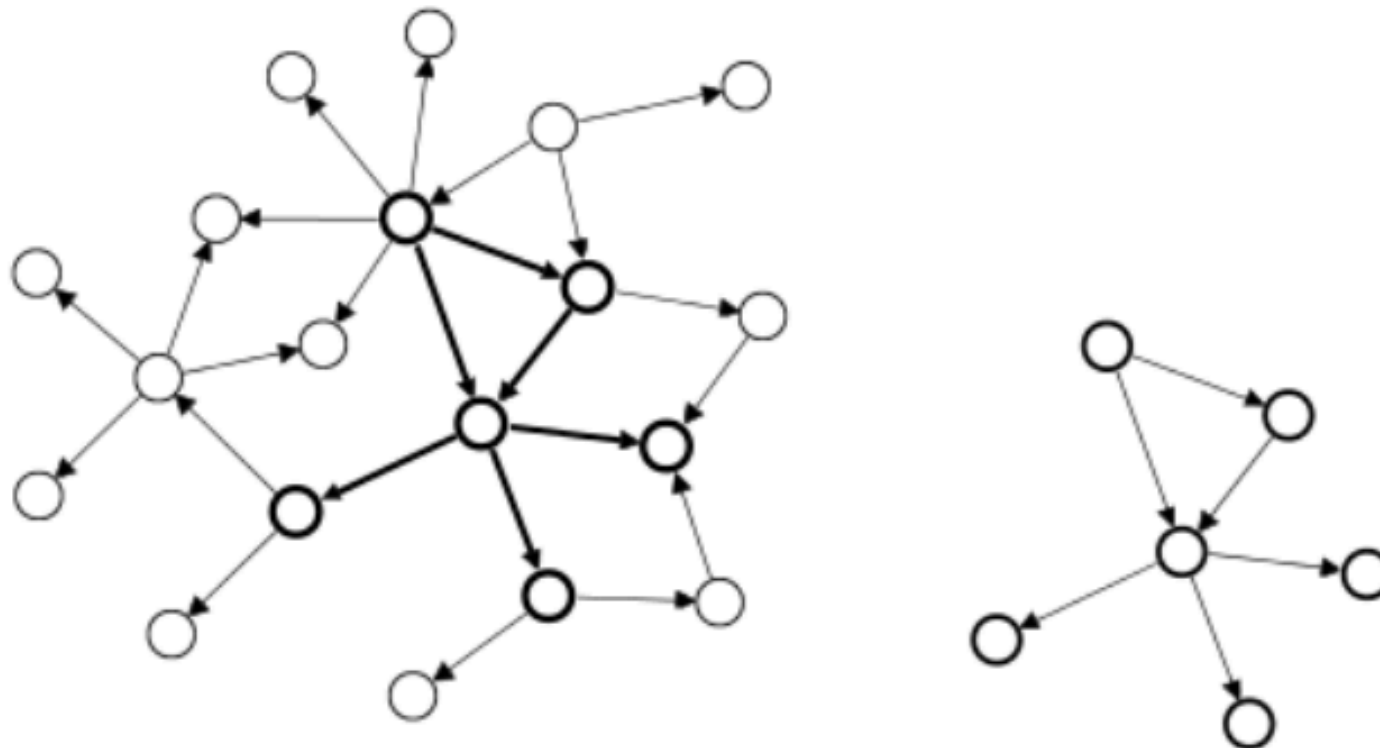


# Motif Matching

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)
- if the correspondence is known, the graph isomorphism belongs to P class of problems (problems that can be solved efficiently)
- if the correspondence is not known, the graph isomorphism is NP-complete (problems that are believed to be hard to 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

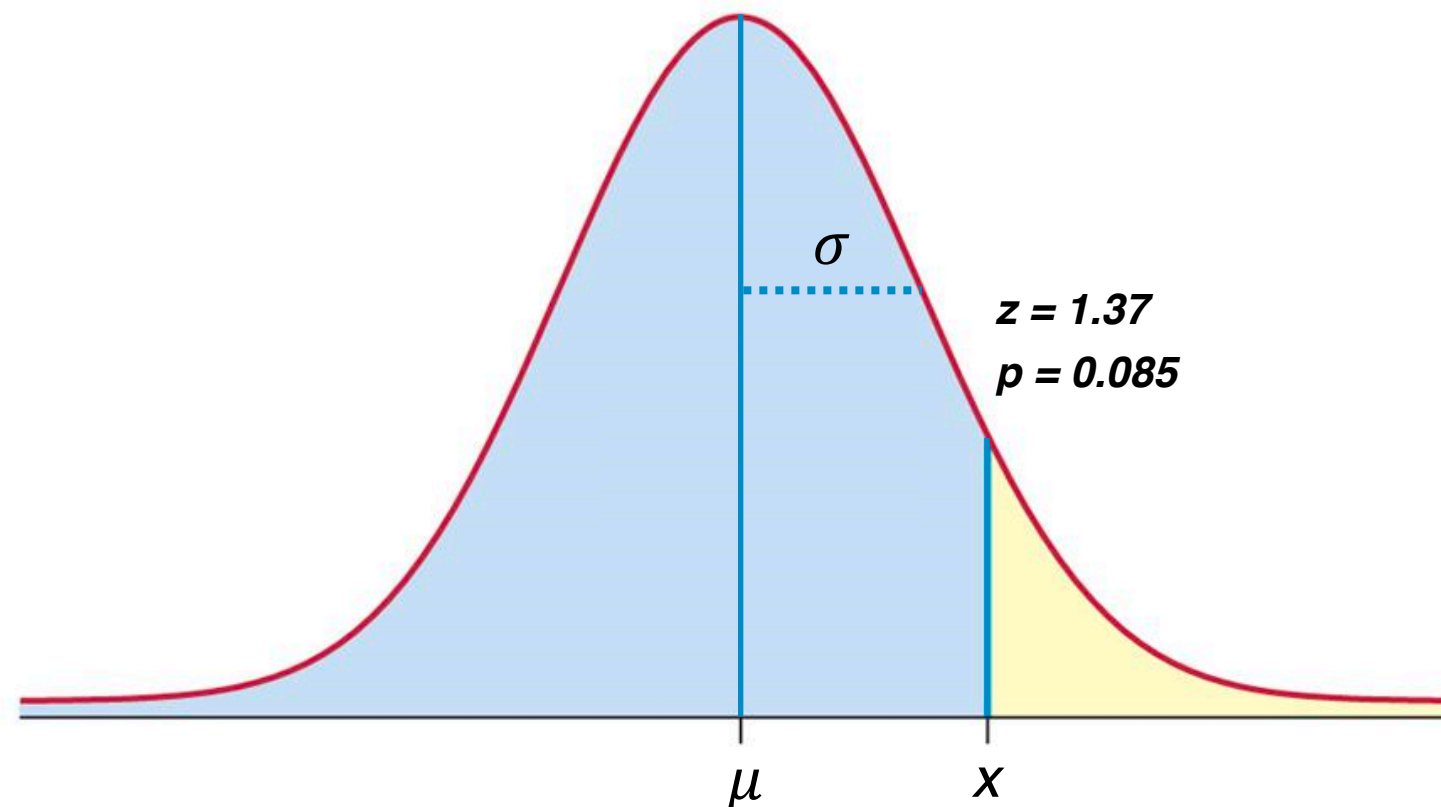
- Generation of random networks with the Erdos-Renyi algorithm
- Random shuffling of the edges

# Z-score

The z-score represents the distance of a number  $x$  from the mean value ( $\mu$ ) of a distribution in terms of standard deviations ( $\sigma$ )

$$Z = \frac{X - \mu}{\sigma}$$

Assuming a normal distribution of the data the probability  $p(t > x)$  can be calculated as a function of the z-score



# Empirical probability

If the distribution of data can not be fitted to any known distribution, the **empirical probability** is estimated based on the fraction of the events.

Given a sorted list of measures

$$M = \{ t_1, t_2, \dots, t_i, \dots, t_N \}$$

being  $t_i$  the lowest measure with  $t_i > x$  the empirical probability  $f(t > x)$  can be calculates as

$$f(t > x) = \frac{N - i + 1}{N} \quad \text{with } N \rightarrow \infty \quad f(t > x) \rightarrow p(t > x)$$



# Detection of Motifs

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

```
>>> g = nx.Graph()  
>>> g.add_edges_from([(1,2),(1,3)])  
  
>>> mot = nx.Graph()  
>>> mot.add_edges_from([("A","B")])  
  
>>> g1 = g.subgraph([1,2])  
>>> nx.is_isomorphic(g1,mot)
```

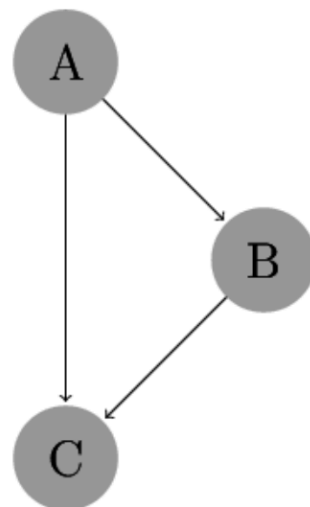
# Exercise

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

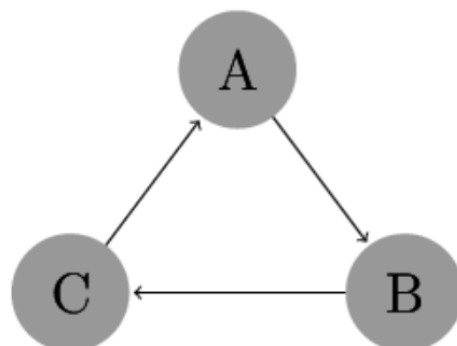
Calculate occurrences on random networks and the z-score.

What is the difference the the FFL is matched on G'?

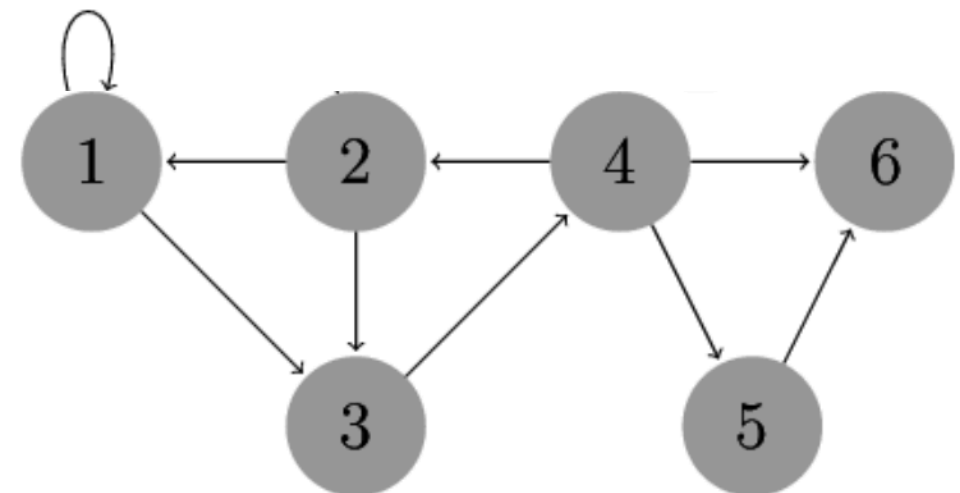
**Feed Forward Loop**



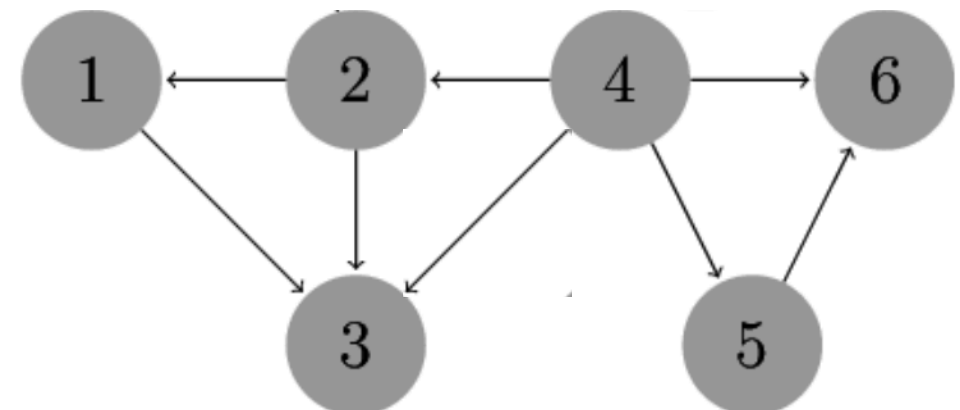
**3-Cycle**



**G**



**G'**






# RegulonDB


Database of Escherichia coli K-12 Transcriptional Regulatory Network





[Home](#) [Features](#) [Integrated Views & Tools](#) [Downloads](#) [Doc & Help](#)

 **Search in RegulonDB**  
  
**Search**   
Example: "araC AND arabinose", "araC transcriptional regulator"  
[Regulon list](#)

 **Downloads**

**Experimental Datasets**  
Data files of manually curated biological objects with experimental evidence (confirmed, strong or weak). 

**Computational Predictions Datasets**  
Data files of genome-wide computationally predicted biological objects. 

**RegulonDB Full Version**  
Get the latest version of the complete RegulonDB database in different formats: TXT, XMLS, DMP file and BioPAX Level 3 format (Registration required). 

**Escherichia coli K-12 Transcriptional Regulatory Network**

Currently the best electronically-encoded regulatory network of any free-living organism. [Read more](#)

**RegulonDB Features**

- RegulonDB is the primary database on transcriptional regulation in *Escherichia coli* K-12 containing knowledge manually curated from original scientific publications, complemented with high throughput datasets and comprehensive computational predictions.
- Graphic and text-integrated environment with friendly navigation where regulatory information is always at hand.
- We strive for facilitating integrated views for users to understand as well as organized knowledge in computable form.

[Read our latest release notes](#)

<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

[http://regulondb.ccg.unam.mx/menu/download/datasets/files/network\\_tf\\_gene.txt](http://regulondb.ccg.unam.mx/menu/download/datasets/files/network_tf_gene.txt)

# 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

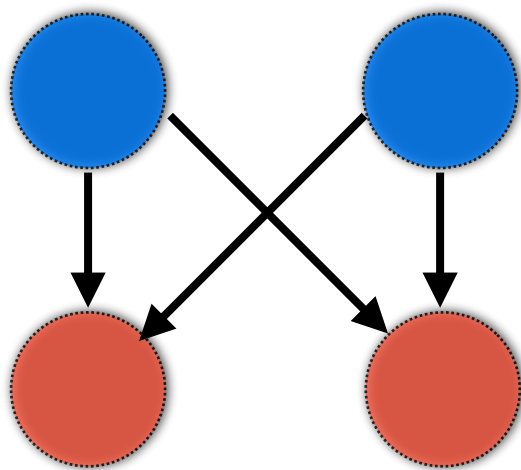
Matches can be performed based on node and edges attributes

```
>>> import networkx.algorithms.isomorphism as iso  
>>> em=iso.categorical_edge_match('sign'='+')  
>>> nm=iso.categorical_node_match('color'='red')  
>>> nx.is_isomorphic(G1,G2,edge_match=em, node_match=nm)
```

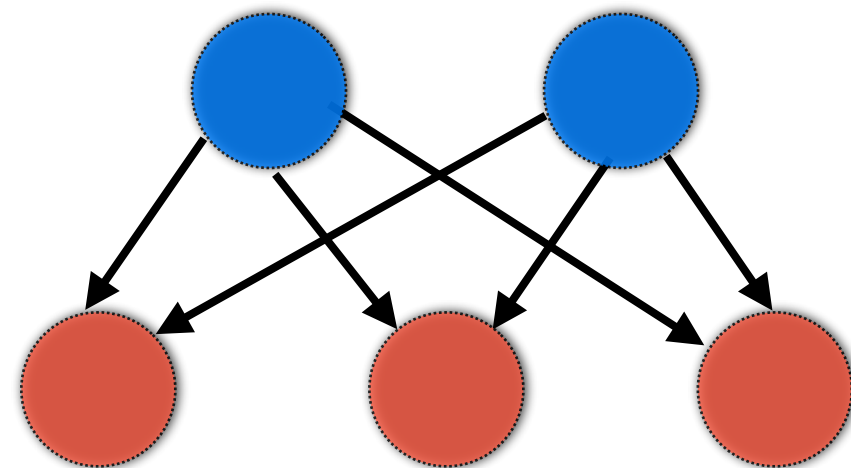
# Exercise

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**

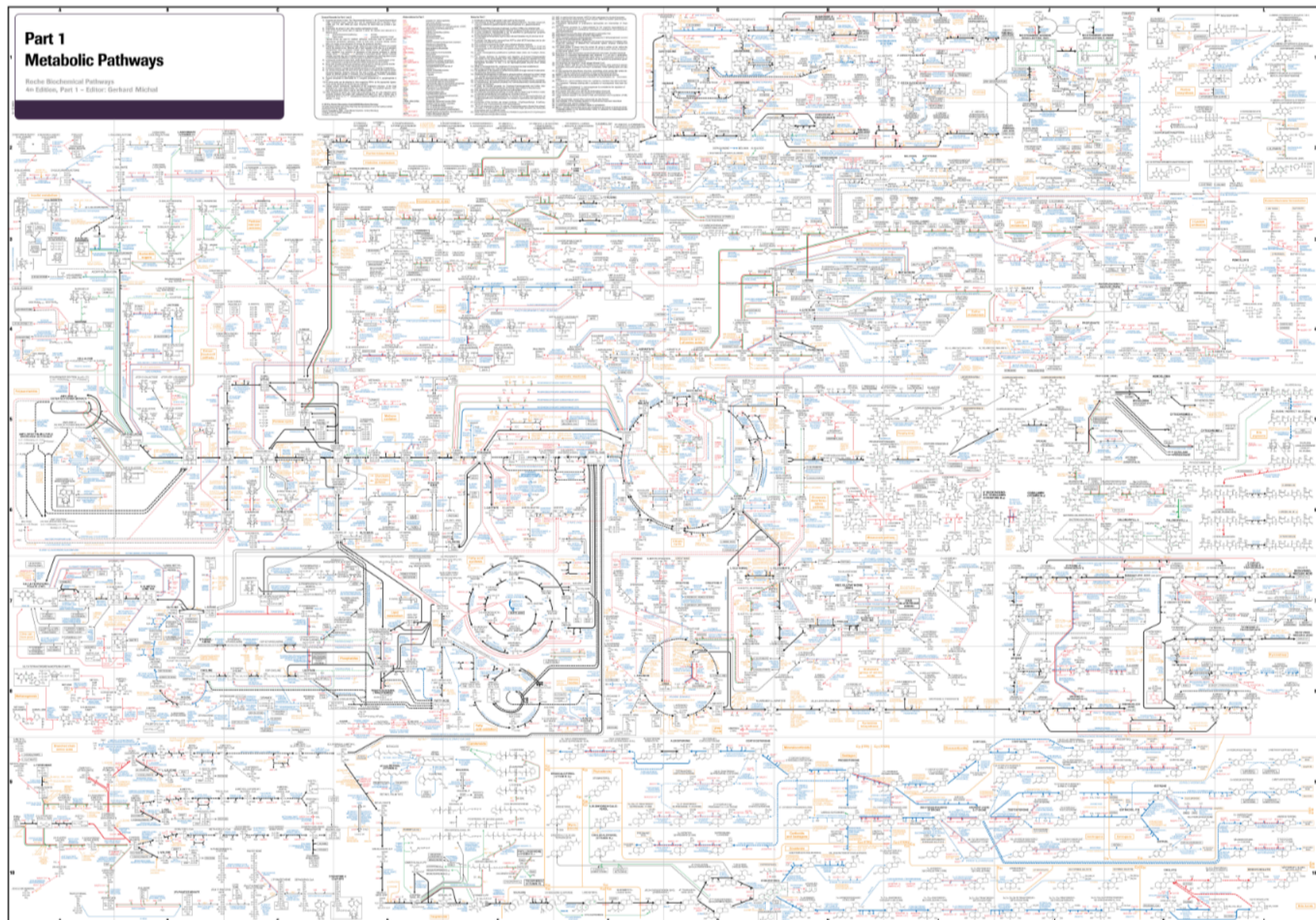


**Multi-Input module**



# Metabolic Pathway

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





# 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

map

Organism

Enter keywords

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

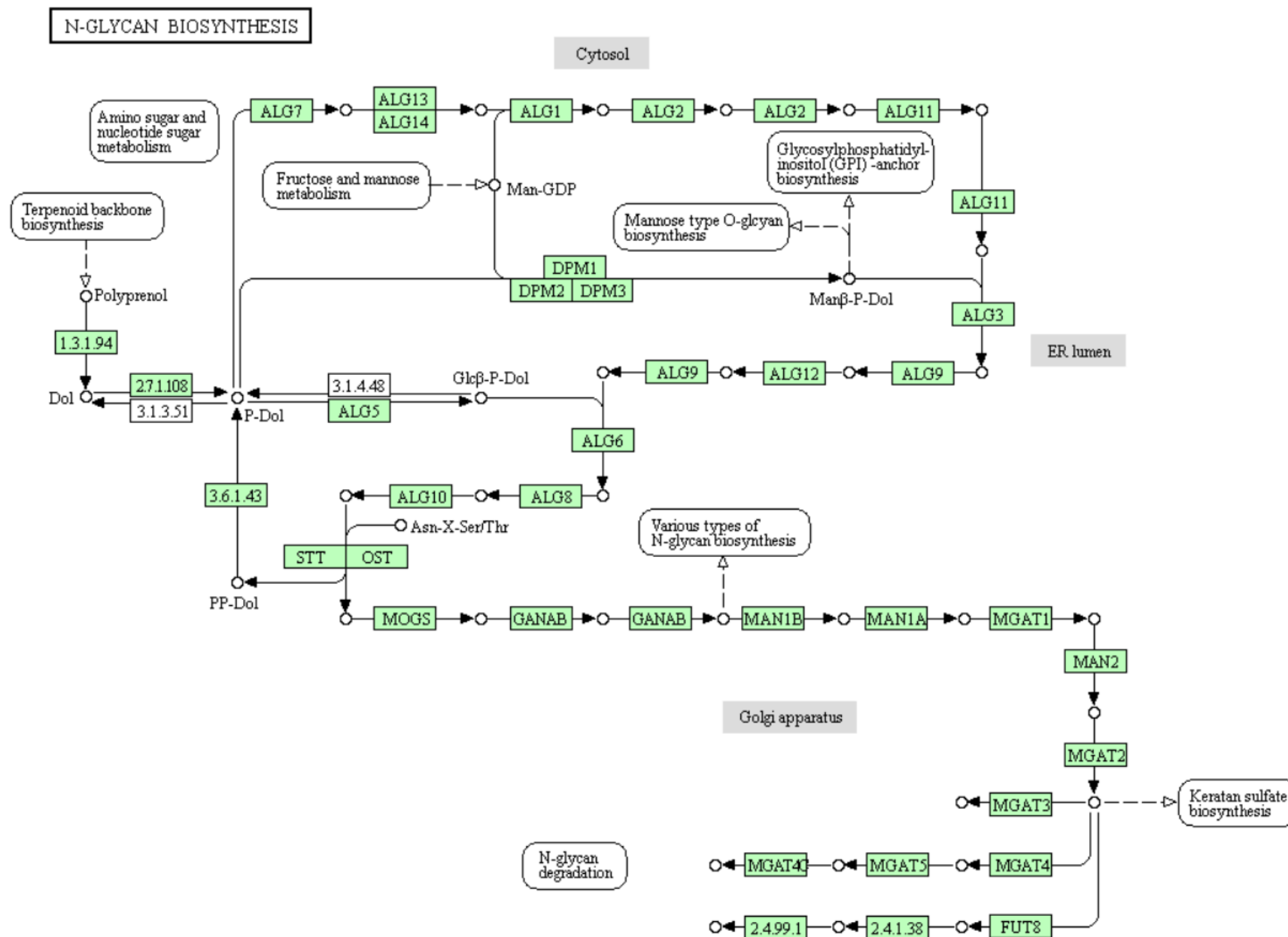
#### 7. Drug Development

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



PATHWAY: map00510

Help

Entry	map00510	Pathway
Name	N-Glycan biosynthesis	
Description	<p>N-glycans or asparagine-linked glycans are major constituents of glycoproteins in eukaryotes. N-glycans are covalently attached to asparagine with the consensus sequence of Asn-X-Ser/Thr by an N-glycosidic bond, GlcNAc b1- Asn. Biosynthesis of N-glycans begins on the cytoplasmic face of the ER membrane with the transferase reaction of UDP-GlcNAc and the lipid-like precursor P-Dol (dolichol phosphate) to generate GlcNAc a1- PP-Dol. After sequential addition of monosaccharides by ALG glycosyltransferases [MD:M00055], the N-glycan precursor is attached by the OST (oligosaccharyltransferase) complex to the polypeptide chain that is being synthesized and translocated through the ER membrane. The protein-bound N-glycan precursor is subsequently trimmed, extended, and modified in the ER and Golgi by a complex series of reactions catalyzed by membrane-bound glycosidases and glycosyltransferases. N-glycans thus synthesized are classified into three types: high-mannose type, complex type, and hybrid type. Defects in N-glycan biosynthesis lead to a variety of human diseases known as congenital disorders of glycosylation [DS:H00118 H00119].</p>	
Class	Metabolism; Glycan biosynthesis and metabolism	
	<a href="#">BRITE hierarchy</a>	

## All links

[Pathway \(5\)](#)  
    [KEGG MODULE \(5\)](#)  
[Disease \(2\)](#)  
    [KEGG DISEASE \(2\)](#)  
[Chemical substance \(38\)](#)  
    [KEGG COMPOUND \(8\)](#)  
    [KEGG GLYCAN \(30\)](#)  
[Chemical reaction \(72\)](#)  
    [KEGG ENZYME \(35\)](#)  
    [KEGG REACTION \(37\)](#)  
[Gene \(117824\)](#)  
    [KEGG ORTHOLOGY \(45\)](#)  
    [RefGene \(117779\)](#)  
[Literature \(6\)](#)  
    [PubMed \(6\)](#)  
[All databases \(117947\)](#)

[Download RDF](#)

# 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:

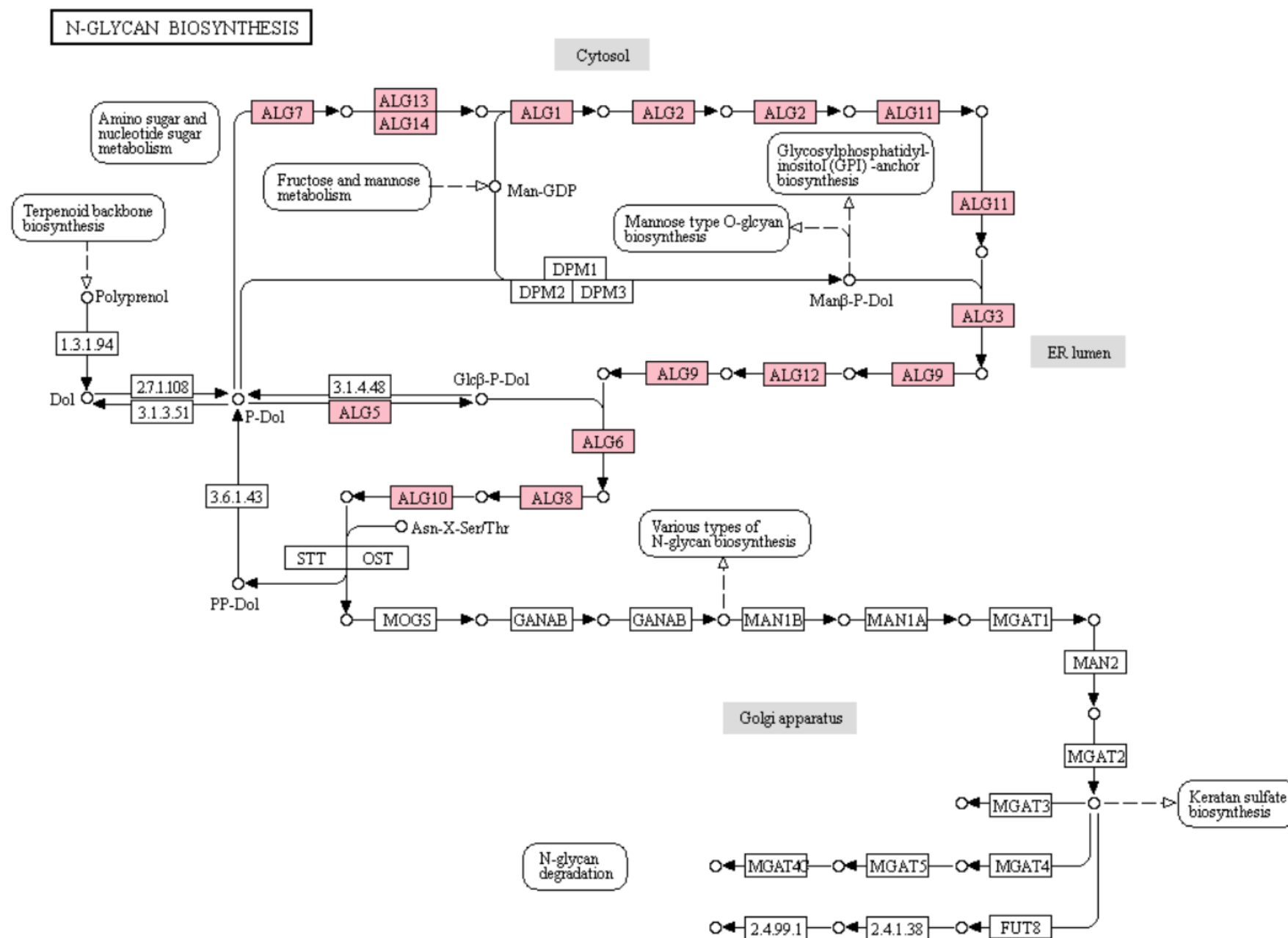
- boxes (enzymes) are linked by "relations"
- circles (chemical compounds) are linked by "reactions".

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