

# Construction and Analysis of Metabolite Network from *Arabidopsis thaliana* pathways

Kasthuribai viswanathan<sup>1</sup>, Nita parekh<sup>1</sup>

<sup>1</sup> Center for Computational Natural Science and Bioinformatics,  
IIIT- Hyderabad - 500032, India,  
[kasthuribai.v@research.iiit.ac.in](mailto:kasthuribai.v@research.iiit.ac.in), [nita@iiit.ac.in](mailto:nita@iiit.ac.in)

**Abstract.** The recent large scale advances in science and technology has resulted in accumulation of large amount of biological pathways data. Any metabolic pathway contains large number of enzymes, metabolites and reactions. To make sense of diverse data available on a system, one needs to correlate and analyze them as a whole. Motivated by the potential benefits of graph theory and its application in biological data, we discuss the automated reconstruction and analysis of metabolite network of *Arabidopsis thaliana* using concepts of graph theory. *A.thaliana* metabolite network was reconstructed and analysis of the global properties of its metabolite-centric graph shows that the network is small-world and scale-free in nature. The investigation of nodes with high centrality values like high degree and high betweenness in this network help in identifying important metabolites, reactions, etc. Newman's modularity-based approach has been used in the analysis of the metabolite network of *A. thaliana* to identify pathway clusters, isolated pathways, and orphan metabolites or products. Our analysis on network representations helps in understanding the relationship between the metabolites, enzymes and reactions of metabolic pathways in *A. thaliana*.

**Keywords:** metabolic pathways; graph theory; metabolic network; modularity; centrality measures.

## 1 Introduction

In recent decades, a large number of complete and draft genomes have been sequenced very rapidly. In spite of enormous metabolic reaction data, the accurate prediction of metabolite phenotypes remains difficult. Pathway reconstruction is an approach to corroborate the experimental data and to widen its utilities. Oldest and dynamic method of pathway reconstruction is the kinetic metabolic modeling[1]. It is based on rate laws of participating reactions and corresponding kinetic parameters. Despite the utilities, kinetic approach is not handy because, the determination and interpretation of concentrations and rate reactions are much difficult. On the other hand, pathway reconstruction using graph theory becomes advantageous since only very less information is required to construct the metabolic network or graph of the entire pathways in the organism. In depth functional analysis of metabolic pathways is succeeded by decomposition of this network.

A complete graph can be constructed using the existing knowledge of metabolites, enzymes and reactions from the metabolic pathway databases. The undirected metabolite network was constructed by considering each substrate as a node and an edge drawn between two substrates sharing the same reaction [2]. Even though the utility of pathway reconstruction is very high in plants, only a very few plant metabolic pathways have been reconstructed. We have chosen *Arabidopsis thaliana* for the study since it is a model organism which has significant metabolic pathways with remarkable functionalities like defense against pathogens and herbivores, UV protection, resistance against oxidative stress and Auxin transport.

---

Here, we have used an automated and efficient metabolite pathway reconstruction of *A.thaliana* using data set extracted from Kyoto Encyclopedia of Genes and Genomes (KEGG) Release 50.0, April 1, 2009 [3]. The XML file in the KEGG FTP contains reactions grouped under pathways of a specific organism. The XML file does not contain information about currency metabolites like ATP, H<sub>2</sub>O etc. List of edges and arcs that capture the biological relationship was computed. This file was visualized using open source visualization tools, such as Pajek and Centibin that help in plotting distributions, navigation within the network and calculating centralities of the biological networks.

The degree of a node in a network is the number of connections or edges the node has with other nodes. The degree distribution of the *A.thaliana* metabolite network construction show that a few nodes have high degree and most of the nodes have low degree revealing the scale free nature [4]. Construction of a random network with the same number of nodes and edges as the *A. thaliana* metabolite network exhibited similar path length but smaller clustering coefficient compared to the *A. thaliana* metabolite network suggesting its small world nature. Correlation between high degree and high betweenness of the network construction shows that there are many nodes with high betweenness and low degree. These nodes typically connect pathways or two groups of reactions and are important to be analyzed. We analyzed the robustness of the network by random and targeted removal of nodes in both the metabolite network and its random counterpart. Targeted removal was performed on nodes exhibiting high centrality values (degree and betweenness). When under attack by nodes with high degrees, the random network does not show any difference whether the nodes are selected randomly or based on the decreasing values of degree whereas the metabolite network shows a drastic change in diameter when the nodes are targeted for removal. The community detection analysis of *A. thaliana* metabolite network suggests its modular nature. Modularity analysis (using Newman's algorithm) of the network showed hierarchical architecture and also helped in identifying isolated and orphan metabolites.

## 2 Materials and Methods

### 2.1 Dataset

In KEGG metabolic pathway database, the pathway maps are validated, manually drawn and updated frequently and the enzymes are cross-referenced to other relevant databases like GenBank, PDB, etc. [5]. Hence for reconstruction of metabolic network of *A.thaliana*, we have used the Kyoto Encyclopedia of Genes and Genomes database. KEGG FTP contains metabolic pathways as XML files for each listed organisms. A total of one hundred metabolic pathways listed under *A.thaliana* were downloaded as individual XML files forming the dataset for the analysis.  
(<ftp://ftp.genome.jp/pub/kegg/xml/kgml/metabolic/organisms/ath/>).

### 2.2 Substrate Centric Graph

The KEGG XML file has unique reaction id for each reaction in the pathway followed by the unique ids for the reactants and products. These files are incomplete without detailed information like secondary metabolites in reactants and products. Using perl scripts, the reaction id are matched with KEGG entire reaction list which has complete reaction information and the missing information's are made complete. Since the network we constructed is undirected and does not contain currency metabolite, the information on the direction of the reactions and the currency metabolites are neglected. Reactants and products of the same reaction are connected by edges. Each reactant and the product becomes each node in the network, the reaction id is assigned to the edge connecting two nodes. Edge list is

computed by listing the connected edges and their corresponding reaction id's. The Edge list captures the network property and this file is used for the network analysis.

### 3 Results and Discussion

The metabolite network constructed has metabolites as nodes and the reactions they take part as edges. The metabolite network we generated for *Arabidopsis thaliana* has 2801 unique metabolites. The network does not contain the common small molecules such as ATP, NADH, water etc. There are 3639 unique reactions in the network. The diameter of the network, which is the largest distance between two nodes, is 56. To know how our network differs from similar networks, we compare the properties of our metabolite network with the random network constructed with same number of nodes and edges and with the Radrich's *Arabidopsis* metabolite network model.(Table1)

**Table 1: Global properties of metabolite network, random network and the properties analyzed by Radrich in *A.thaliana* network construction**

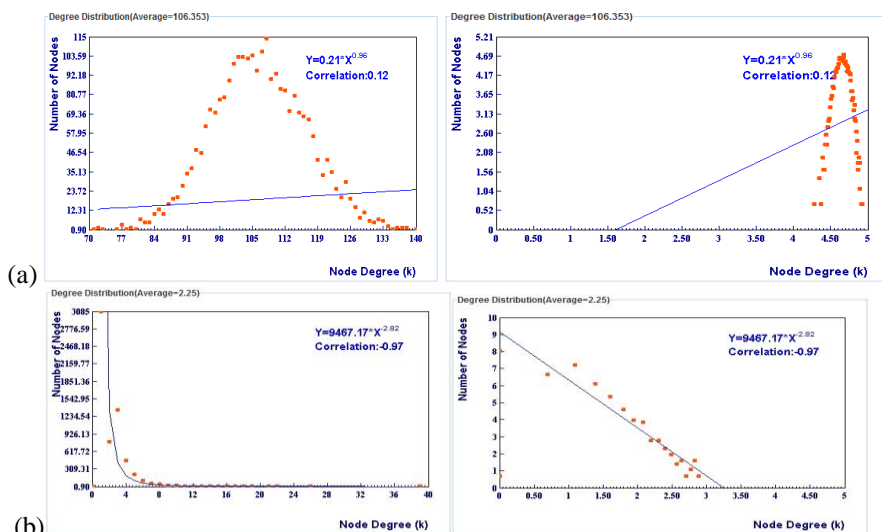
	Metabolite Network	Random Network	Radrich Network
Nodes	2801	2801	2288
Edges	3639	3639	6547
Diameter	56	8	10
Clustering Coeff.	0.215	0.001	0.186
Avg. path length	3.486	4.642	3.286

The random network has low clustering coefficient compared to the metabolite network constructed by us. Radrich's semi automated genome-scale reconstruction network on *Arabidopsis* by integration of metabolic databases [6]uses current metabolites and pathway data that were common in both KEGG and AraCyc. There are more edges in Radrich model due to the currency metabolites. The diameter of the metabolite network is very high compared to that of random network with same number of nodes and edges and Radrich network. The larger diameter in our network reveals that the information flow is between metabolites of two completely unrelated pathways leads to larger path lengths between those nodes. The lower clustering coefficient of a random network compared to *Arabidopsis* metabolite network explains occurrence of meaningful clustering in biological network. In metabolite network, we see the average path length depends on the system size but does not change drastically with it.

#### 3.1 *Arabidopsis thaliana* metabolite network is scale free and small world

The degree of a node in a network is the number of connections or edges the node has to other nodes. The degree distribution  $P(k)$  gives the fraction of nodes that have degree  $k$  and is obtained by counting the number of nodes  $N(k)$  that have  $k = 1, 2, 3, \dots$  edges and dividing it by the total number of nodes  $N$ . From Fig.1, degree distribution graph, we see that it follows the power law which appears as a straight line on a logarithmic plot and hence proving metabolite network follows 'scale free nature'[2]. Using this function  $P(k)$  it is evident that there is a high diversity in the degree of the nodes (Fig.1). This nature becomes more evident by comparing it with a random graph with the same number of edges and arcs.

We constructed a random graph, using the Erdoes Renyi model that assumes each pair of nodes in the network is connected randomly with probability  $p$ . This graph reflects the expected properties of a network which is random with respect to the node's position and their interaction compared to a metabolite network of the same size[3]. Random network have a bell-shaped degree distribution, indicating that the majority of nodes have a degree close to the average degree  $\langle k \rangle$ . The average clustering coefficient of a random graph equals  $\langle k \rangle / N$  and thus is very small for large  $N$ [7]. We compare the degree distribution of the metabolite network with random network containing same number of nodes (Fig 1).



**Fig 1:** Comparison between the degree distribution of (a) Random graph having the same number of nodes and edges as the *Arabidopsis* metabolite network. For better understanding the same two distributions are plotted both on a linear (left) and logarithmic (right) scale for all the networks. The bell-shaped degree distribution of random graphs peaks at the average degree and decreases fast for both smaller and larger degrees, indicating that these graphs are statistically homogeneous. By contrast, the degree distribution of the scale-free network follows the power law  $P(k) = Ak^{-3}$ , which appears as a straight line on a logarithmic plot.

We compare the metabolite network with random network. Another observation by comparing the metabolite network with and random network is that the average clustering coefficient of the random network is much smaller than that of *A.thaliana* metabolite network and the average path length was closer in the random graph, justifying the small world nature of the metabolite network[7].

### 3.2 Error and attack tolerance nature

The nodes in *Arabidopsis* metabolite network are capable of staying interconnected and communicate even by unrealistically high failure rates. However, most networks become extremely vulnerable to attacks on selected nodes that bridge highly interconnected nodes in the network. We tested the error and attack tolerance nature of *Arabidopsis* network comparing random and scale free networks.

Attack vulnerability shows a decreased performance of a network due to the selected removal of nodes or edges[8]. Here, it means the prevention of a metabolic reaction to take place due to the removal of an enzyme or primary substrate. Studying the attack vulnerability of networks is very important for identifying the weak or strong ‘links’ in the network[1]. Subsequently, this knowledge can be used to protect the network from outside attacks. In order to study the attack tolerance, we removed a fraction of nodes from both random and *Arabidopsis* metabolite networks and studied the effect of this removal on the diameter and clustering coefficient.

We randomly removed 5, 10,15,20,25 percentage of nodes from the *Arabidopsis thaliana* metabolite network. In random network, due to the homogeneity all nodes contribute equally to the diameter, so the removal of each node caused the same effect(Fig 2a,2b). But in case of metabolite network (scale free) due to the extremely inhomogeneous degree distribution, many nodes have only a few links. The nodes with small connectivity will be selected with a much higher probability and these removals changed the diameter in a small scale[4]. During attack on high degrees nodes, the random network does not show any difference irrespective of

selection with random or descending degree nodes[4]. In scale-free metabolite network, targeted removal (Fig 3a, 3b) show drastic change in diameter due to small number of nodes with very high connectivity. The diameter almost doubles when 5% of the nodes were removed.

**Table (2a) Random Removal of Nodes from the metabolite network**

Nodes	Clustering Coefficient	Diameter
90	0.045	56
180	0.045	51
270	0.045	39
360	0.046	37
450	0.048	34

**Table (2b) Random Removal of Nodes from the Random network**

Nodes	Clustering Coefficient	Diameter
90	0.001	9
180	0.001	10
270	0.001	10
360	0.001	10
450	0.001	11

**Table (3a) Targeted Removal of Nodes from the metabolite network**

Nodes	Clustering Coefficient	Diameter
90	0.01853	53
180	0.0123	30
270	0.00525	25
360	0.00225	25
450	0.002	25
540	0.002	11

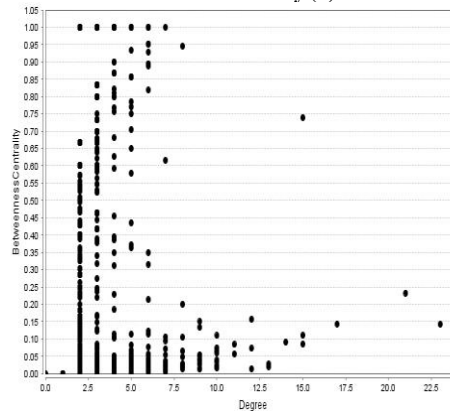
**Table (3b) Targeted Removal of Nodes from the Random network**

Nodes	Clustering Coefficient	Diameter
90	0.001	8
180	0.001	9
270	0.001	9
360	0.001	11
450	0.001	12

### 3.3 Betweenness Vs Degree distribution

In order to understand the relation between high degree and high betweenness, the betweenness is plotted as a function of connectivity (Fig2). The metabolite network showed that most metabolites have low neighborhood connectivity but very high betweenness. This shows that many metabolites typically connect pathways and are potentially important metabolites. These results suggest that the network has modular organization with the high-betweenness and low-connectivity nodes as important links between these modules. The selected nodes with *high degree* and *betweenness* centrality are hubs and they are important nodes that control the overall network interaction. Hub metabolites include Pyruvate, Gibberelin, Stemmadiene, Antracene Cis 1,2hidhdriole which have been investigated to important in *Arabidopsis*.

**Fig 2: Betweenness ( $B$ ) is plotted as a function of connectivity ( $k$ ) for metabolite network**



### 3.4 Modularity

Modularity means that cellular functionality can be seamlessly partitioned into a collection of modules. Each module is a discrete entity of several elementary components and performs an identifiable task, separable from the functions of other modules. We used the Newman and Girvan edge-betweenness method to calculate the number of clusters available in the network. This algorithm identifies edges in a network that lie between communities and then removes them, leaving behind just the communities themselves[6]. We have utilized the radatools [9] to apply the algorithm and the input files were the dot net files of the network.

Community detection using Newman's algorithm[10] detects 101 communities in metabolite network. The largest community had 506 metabolites that had the highest interaction within the group and lower interaction outside the group (Fig 2a). Metabolites taking part in similar functional type of reactions will share common properties. These metabolites were further traced back to the pathways (Fig 2b) that contained these metabolites and the list of pathways for the largest community was collected. They mainly constituted the amino acid metabolism pathways. There were 27 communities with only one metabolite called isolated metabolites (Fig 2c). They produce similar intermediate compounds and have hence interacted more closely. The pathways in the highest cluster were the amino acid metabolism pathways and the chlorophyll metabolism pathways.

**Table 2 (a) The Communities in *A.thaliana* substrate graph and the number of nodes in each community (b): The pathways corresponding to nodes in Community 1 with 506 nodes (c) Pathways corresponding to the single node communities in the metabolite centric graph**

Number of Community	Number of nodes
1	506
2	399
3	375
4	337
5	295
6	250
7	37
8	36
9	36
10	18
11	16
12	15
12	14
14	14
15	13
16	12
17	12
18	12
19	11
20	10
21	10
22-24	9
25-28	7
29-32	6
33-37	5
38-46	4
47-51	3
52-72	2

High Cluster Pathways	
1	Biosynthesis of alkaloids derived from shikimate pathway
2	Porphyrin and chlorophyll metabolism
3	Naphthalene and anthracene degradation
4	Glycerolipid metabolism
5	Cyanoamino acid metabolism
6	ABC transporters
7	Pentose and glucuronate interconversions
8	Ascorbate and aldarate metabolism
9	Selenoamino acid metabolism
10	Biosynthesis of Terpenoids and steroids etc.,

Isolated Pathways	
1	Steroid hormone biosynthesis
2	Tyrosine metabolism
3	Monoterpenoid biosynthesis
4	Arachidonic acid metabolism
5	Indole alkaloid biosynthesis
6	Glycine, serine and threonine metabolism
7	Porphyrin and chlorophyll metabolism
8	Fructose and mannose metabolism
9	Tryptophan metabolism
10	Methane metabolism
11	Biosynthesis of phenylpropanoids
12	Butanoate metabolism etc.,

## 4 Conclusion

By using a method proposed earlier by others we come up with automated way of constructing the metabolite network with KEGG metabolic pathway data. This gives us a complete idea of interaction between enzymes, reactions, and metabolites. The substrate centric graph helps in finding the conserved metabolites and reactions. This construction and analysis procedures can be further applied to an enzyme network and the enzyme evolution in *Arabidopsis thaliana* can be studied.

## References

- [1] Z. N. Oltvai and A. L. Barabasi, "Systems biology. Life's complexity pyramid," *Science*, vol. 298, pp. 763-4, Oct 25 2002.
- [2] A. L. Barabasi and Z. N. Oltvai, "Network biology: understanding the cell's functional organization," *Nat Rev Genet*, vol. 5, pp. 101-13, Feb 2004.
- [3] L. H. Hartwell, *et al.*, "From molecular to modular cell biology," *Nature*, vol. 402, pp. C47-52, Dec 2 1999.
- [4] A. L. Barabasi, "Scale-free networks: a decade and beyond," *Science*, vol. 325, pp. 412-3, Jul 24 2009.
- [5] M. Kanehisa, *et al.*, "From genomics to chemical genomics: new developments in KEGG," *Nucleic Acids Res*, vol. 34, pp. D354-7, Jan 1 2006.
- [6] K. Radrich, *et al.*, "Integration of metabolic databases for the reconstruction of genome-scale metabolic networks," *BMC Syst Biol*, vol. 4, p. 114, 2010.
- [7] A. L. Barabasi and E. Bonabeau, "Scale-free networks," *Sci Am*, vol. 288, pp. 60-9, May 2003.
- [8] H. W. Ma and A. P. Zeng, "The connectivity structure, giant strong component and centrality of metabolic networks," *Bioinformatics*, vol. 19, pp. 1423-30, Jul 22 2003.
- [9] B. H. Junker, *et al.*, "Exploration of biological network centralities with CentiBiN," *BMC Bioinformatics*, vol. 7, p. 219, 2006.
- [10] E. Ravasz, *et al.*, "Hierarchical organization of modularity in metabolic networks," *Science*, vol. 297, pp. 1551-5, Aug 30 2002.