Construction and Analysis of Enzyme Centric Network of *A. thaliana* using Graph Theory

Kasthuribai Viswanathan[1], Nita Parekh[1]
[1]Center for Computational Natural Science and Bioinformatics,
International Institute of Information Technology, Hyderabad - 500032, India
kasthuribai.vpg08@research.iiit.ac.in, nita@iiit.ac.in

Abstract. Graph comparisons, quantitative characterizations, computation of topological indices, clustering and partitioning are some of the major computations of graph that have yielded valuable results in various disciplines. Motivated by the potential benefits of graph theory application on biological data, we discuss the reconstruction and analysis of enzyme centric network of *Arabidopsis thaliana* using graph theory concepts. We had earlier constructed the metabolite network of *Arabidopsis thaliana* and witnessed the scale free and small world nature of the network. Compared to metabolites, the enzymes are more conserved in and across many pathways. So the aim of constructing the enzyme centric network is to see if the network follows similar network properties of the metabolite network and to look for additional details that a metabolite network cannot reveal. The enzyme flat file from KEGG FTP is used as the data set for the reconstruction of the enzyme centric network. We examined the network to find the relationship between topological connections among enzymes and their functions during evolution. The enzyme sequences of high degree and high betweenness enzymes belonged to ancient fold class and ancestry value showing evidence that they evolved very slowly.

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

1 Introduction

Metabolism is one of the most complex cellular processes. Connections between the biochemical reactions are represented as series of metabolic reactions which constitute the metabolic pathways. These pathways are used by researchers for the molecular evolution studies. Most studies of molecular evolution are focused on individual genes and proteins. The most interesting challenge in systems biology is constructing such biological networks and trying to interpret the hidden evolutionary details. The correlations that the researchers draw out in linear pathways are not highly visible and can emerge when analyzed as a whole network. We had previously developed a metabolite network from *Arabidopsis thaliana* pathway data using reaction files from the KEGG FTP. The metabolite network exhibited the small world, scale free nature and showed hierarchical organization. But the metabolite network is not very suitable for evolutionary analysis due to reasons like the metabolites are less conserved compared to the enzymes in pathways. Also, the number of enzymes is smaller compared to the number of metabolites enabling a closer look of the interactions.

The enzyme centric network is constructed with enzymes catalyzing the reactions as the vertices of the graph and the edge drawn between them if they one or more metabolites. Thus, an edge is drawn from an enzyme E1 to an enzyme E2 if E1 catalyzes a reaction in which compound A is the product and E2 consumes A, that is, it is a reactant in the second reaction [1]. For simplification of the network analysis, we assume that reactions are reversible and therefore each link or enzyme-enzyme relation in the network is undirected (bidirectional). For reconstruction of the enzyme centric network, the Enzyme flat file in the KEGG FTP is used as raw data. In order to include only functional relationships in the calculation of the enzyme connectivity, we excluded the 18 highly connected metabolites and co-factors. The enzyme network is then analyzed using Pajek, the network analysis tool, and various network properties such as degree, betweenness, etc. have been computed using this tool [2].

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* enzyme network construction shows that a few nodes have high degree and most of the nodes have low degree revealing the scale free nature. It would be interesting to investigate role of high-degree nodes in the evolution of the organism. The enzyme centric network exhibits modular architecture suggesting clusters of interacting enzymes. The enzymes having high betweenness values are observed to be relating different pathways and an analysis of the corresponding gene/protein sequences can help in assessing its age and conservation across species.

2 Materials and Methods

2.1 Dataset

The KEGG database is designed in a way to facilitate understanding of higher-order protein and cellular functions using genomic and molecular information. The main reason for choosing the pathways information from KEGG is because it is open source and free to academic users and the enzyme data is available as a single flat file format for easy processing. The enzyme flat file used for the analysis has been downloaded from KEGG FTP [3]. Each enzyme has details on the genes, organisms, reactants, products, references, etc.

2.2 Construction of Enzyme Centric Network

The flowchart in Figure 1 explains the steps in the reconstruction of enzyme network starting with the enzyme flat file obtained from KEGG FTP. The enzyme flat file is a complete listing of all the known enzymes with an enzyme commission name (EC) and is delimited with "///" (Fig. 1). A total of 5391 enzymes are listed in this file. It is not an organism specific file, hence *A. thaliana* specific enzymes need to be extracted by parsing this file. This was done by first splitting this single file into 5391 separate files with enzyme name as the file name. Each of these files consisted of all the relevant information such as the enzyme name, synonyms, class, reactants, products, organisms and gene for each enzyme and references. This format was chosen to simplify computational processing and to minimize data duplication and extract only *A. thaliana*

specific enzymes. Next only those enzyme files which have “ath” in the organism names list were extracted; a total of 3277 enzymes were identified to be *A. thaliana* specific and used for constructing the network.

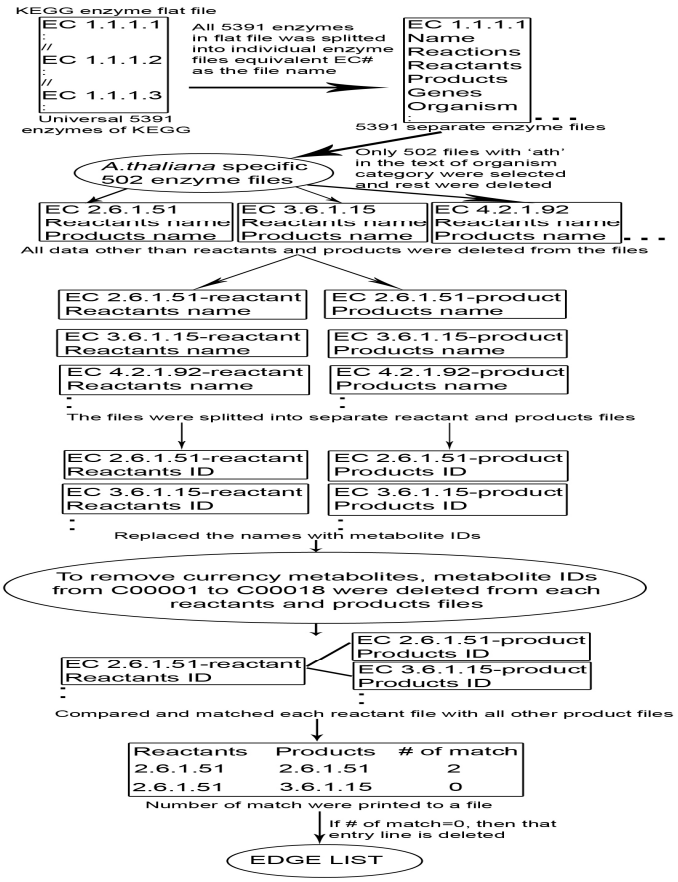

Fig. 1. Flow chart for the construction of enzyme-centric network.

The next step was to extract reactant and product information corresponding to every reaction catalyzed by an enzyme. Scripts were developed to automate the extraction process from each file, and a number of problems were addressed, such as removing incomplete and redundant reactants and products. The metabolite name had many synonyms, for example, NADH-glyoxylate reductase, glyoxylic acid reductase, and NADH-dependent glyoxylate reductase are all synonyms of glyoxylate reductase. So we replaced all the metabolite names with the KEGG metabolite ID and so the synonyms of the metabolite are not considered as a new metabolite to the list Fig. 1). Removing the most abundant substrates, called “currency metabolites” was the second processing step in the resonstruction of the network. Differentiating the currency metabolite from the primary metabolite is troublesome. For example, Glutamate (GLU) is a current metabolite for transferring amino groups in many reactions, but in the following reaction

it is defined as a primary metabolite:

$$AKG + NH3 + NADPH = GLU + NADP^{+} + H2O$$

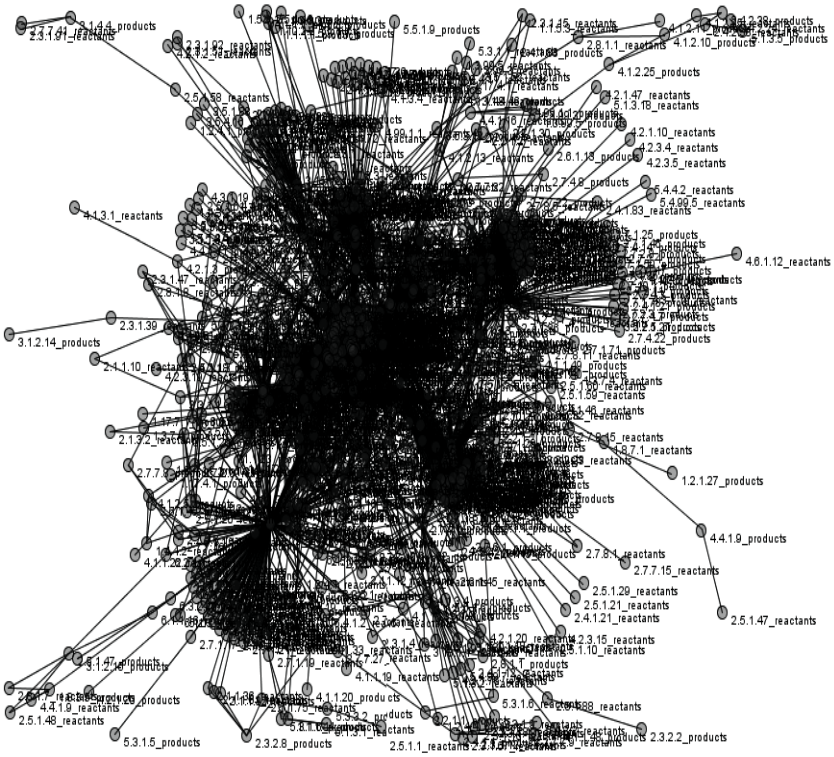

Fig. 2. The enzyme centric network of *Arabidopsis thaliana* visualized in Pajek. The circular balls represent the nodes (enzymes) and the lines represent the edges (metabolites catalyzed by the enzyme). The distribution of the connection is heterogeneous with few nodes very densely connected and few with only a single connection.

Jeong *et al.* (2000) ranked the metabolites according to their connection degree. Based on their analysis, the metabolites C00001 to C00018 (ATP, H, H20, ADP, H202, pyrophosphate, orthophosphate, CO2, NAD, glutamate, NADP, NADH, NADPH, AMP, NH3, and CoA) were labelled as currency metabolites that occurred multiple times in most of the reactions [4]. We removed these metabolites during our construction of the enzyme network. To compare enzymes that catalyze the reactants with the enzymes that catalyze products they share, we wrote perl scripts that split each enzyme file into two files, listing reactants and products respectively. and enzyme product. For example, 1.1.1.1_reactant file contains all the reactants that take part in the reaction catalyzed by enzyme 1.1.1.1. The file 1.1.1.1_product contains the products of the reaction catalyzed by 1.1.1.1. Now each enzyme product file is compared with all the enzyme reactant files to find the number of reactants and products that exactly matched.

A file containing the pair of enzymes and the number of shared reactants and products between them is listed and those that never shared any reactants or products were removed because product in the enzyme product file, say, EPi, is present in the enzyme reactant file as product, ERj, then a link is formed between the enzymes catalyzing the two reactions, Ei and Ej [5]. The edge list of the interacting enzymes is thus constructed. Each node is labeled with an EC number. The enzyme network of *Arabidopsis thaliana* thus constructed is shown in Fig. 2. From the 5391 enzyme files, the number of files drops to 507 Arabidopsis *thaliana* specific enzyme files. The enzyme network has 502 enzyme nodes. There are 4950 metabolites that utilize these enzymes in the network.

3 Results and Discussion

The analysis of the enzyme centric network constructed as discussed in the above section was carried out. It is clear from Fig. 2 that the enzyme network of *A. thaliana* is not a random network and a clear clustering of nodes is observed. In Table 1 is summarized the global properties of the network. The diameter of the network, which is the largest distance between two nodes, is 8. The low diameter reveals that the information flow between the enzymes is faster which means that the reactions are very closely connected. The average path length is 2.9. The low value of average path length means every node is close to all of the others in the network, they can reach others quickly without going through too many intermediaries.

Table 1. Global Properties of Enzyme Network

Number of Nodes	502
Number of Edges	4950
Clustering Coefficient	0.47
Diameter	8
Average path length	2.9

3.1 Scale free Nature of *Arabidopsis thaliana* Enzyme Centric Network

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 $k = 1, 2, 3,\ldots$ edges and dividing it by the total number of nodes N. From the degree distribution graph in Fig. 3, it is clear that the *A. thaliana* enzyme network exhibits power law behavior (Fig. 3(a)), revealed by the straight line on a logarithmic plot (Fig. 3(b)). This indicates the 'scale free nature' of the enzyme network [6]. That is, a high heterogeneity in the degree of the nodes is observed, critically influencing the topological properties of the network. Following Barabasi-Albert algorithm of preferential growth model, the power-law behavior of enzyme network proposes that, during evolution, new nodes tend to attach preferentially to well connect network nodes. As a consequence, the nodes having high degree are likely to be ancient enzymes.

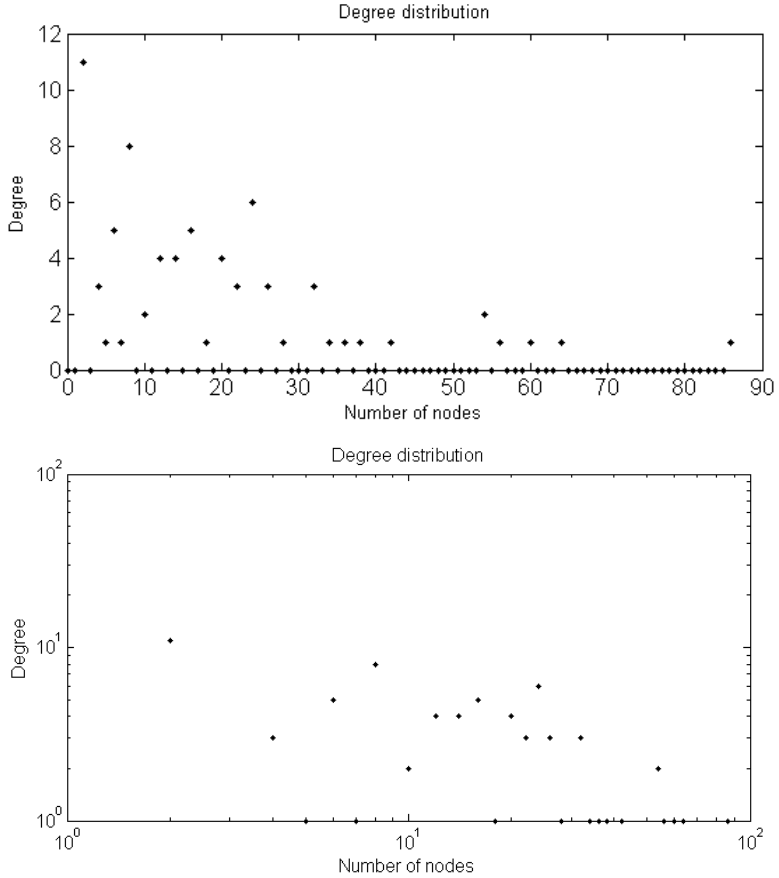

Fig. 3. Degree distribution of the enzyme network plotted both on linear and logarithmic scale. The degree distribution of the scale-free network follows the power law P (k) = $Ak^{-1.26}$, which appears as a straight line on a logarithmic plot.

3.2 Analysis of High Betweenness Enzymes

The betweenness of a node *v* is defined as the number of shortest paths going through that node:

$$C_B(v) = \frac{1}{(|V|-1)(|V|-2)} \sum_{s \neq v \neq t \in V} g_{st}(v) / g_{st}$$

where *V* is the set of nodes and |*V*| represents the number of nodes in *V*; g_{st} is the number of shortest paths from node *s* to node *t*; $g_{st}(v)$ is the number of shortest paths from node *s* to node *t* lying on node *v*. The nodes connecting pathways are critical and their removal can have a deleterious effect on the stability of the network. Such nodes can be identified by an analysis of their betweenness value. Enzymes that connect pathways do have relatively high betweenness value. This reflects the central

role that such enzymes play in relaying metabolites from one enzymatic reaction to another. The differences in the amino acids of the encoded protein (nonsynonymous changes) and some, because of the degeneracy of the genetic code, leave the amino acid unchanged (synonymous or silent changes). Counting up the number of each gives us a measure of the amount of change of the sequence. Chiaoquan Qi has shown that a negative correlation between K_a/K_s *(*non-synonymous/synonymous substitutions) and betweenness of an enzyme, providing clear evidence that high-betweenness enzymes evolve slowly [7] .

In Table 2 are listed top 10 enzymes with high betweenness values. These are highly central in the network as most of the reactions are catalyzed by these enzymes. These high betweenness enzymes belong to transferases that help in transfer of a functional group, oxidoreductases that catalyzes the transfer of electrons from one molecule, and hydrolases which hydrolysis a chemical bond. Theoretically these enzymes are more important as all the pathways involve reactions that help in transferring functional groups by breaking bonds.

Table 2. Top Ten Enzymes with High Betweeness values in the Network

Enzyme	Enzyme Name	Betweenness Values	Enzyme Class
2.3.3.8	ATP citrate synthase	48571.2	Transferase
2.4.2.17	ATP phosphoribosyltransferase	47192.5	Transferase
2.3.3.8	ATP citrate synthase	46350.1	Transferase
1.2.1.41	Glutamate-5-semialdehyde dehydrogenase	34695.7	Oxidoreductases
3.6.1.1	Hydrolases	22370	Hydrolases
3.3.1.1	Adenosylhomocysteinase	18708.4	Hydrolases
3.1.3.16	Phosphoprotein phosphatase	18333.2	Hydrolases
1.11.1.6	Catalase	18316.4	Oxidoreductases
1.7.7.1	Oxidoreductases	16426.6	Oxidoreductases

3.3 Enzymes with degree one

In the enzyme centric network of *A. thaliana*, there are 136 nodes with degree one. These enzymes are catalysis just a single reaction. These are referred to as choke points. A “chokepoint reaction” is a reaction that either uniquely consumes a specific substrate or uniquely produces a specific product. Enzymes associated with high damage are involved in the production of compounds of small connectivity that connect important parts of the metabolism. Inactivation of choke points may lead to an organism's failure to produce or consume particular metabolites which could cause serious problems for fitness or survival of the organism. We listed ten degree one enzymes in the *Arabidopsis* enzyme network is shown in Table 3. We show only top ten of them for convenience and comprehensiveness.

Table 3. List of Degree One Enzymes in *Arabidopsis thaliana* Enzyme Network

Enzyme ID	Enzyme Name
1.10.2.2	ubiquinol---cytochrome-c reductase
1.1.1.1	alcohol dehydrogenase
1.15.1.1	superoxide dismutase
1.1.5.3	glycerol-3-phosphate dehydrogenase
1.2.1.27	methylmalonate-semialdehyde dehydrogenase
1.2.4.1	Pyruvate dehydrogenase
1.2.4.4	3-methyl-2-oxobutanoate dehydrogenase
1.3.3.3	coproporphyrinogen oxidase
1.3.7.4	phytochromobilin:ferredoxin Oxidoreductases
1.3.99.3	acyl-CoA dehydrogenase

3.4 Enzyme Evolution

The earliest enzymes were probably weakly catalytic and multifunctional and specific new enzymes should have evolved through gene duplication, mutation and divergence. As enzymatic pathways became more complicated, new enzymatic function could have been generated by recruitment of individual enzymes from the same or different pathways. The age of metabolic enzymes and the evolution of their metabolism with phylogenetic analysis are interesting. Since the network is scale free, it follows preferential attachment of nodes. The preferential attachment means that new nodes attach to a growing network by connecting to nodes with existing high connectivity. Nodes with high connectivity are therefore often those that have been in the network for a very long time [8]. Thus, enzymes appearing in the early stages of evolution tend to be found more frequently in different organisms (e.g. those involved in glycolysis) and that much of the metabolism of current species is based on the products of those enzymes.

This implies that evolutionarily early enzymes tend to have more connectivity to other enzymes or metabolites. The high degree enzymes in the enzyme network should therefore be highly conserved and should belong to primitive class of enzymes. In Molecular Ancestry Network database (MANET) a method the evolutionary scale of enzymes is computed using information in the Structural Classification of Proteins (SCOP) database[9]. MANET traces evolution of protein architecture in bimolecular networks by linking information in the Structural Classification of Proteins (SCOP), the Kyoto Encyclopedia of Genes and Genomes (KEGG), and phylogenetic reconstructions depicting the evolution of protein fold architecture. In MANET, the phylogenetic tree drawn for all the enzymes in KEGG and the phylogenetic distance is defined as the ancestry value. Table 4 shows the list

of top ten high degree enzymes and their ancestory value predicted by MANET from a scale of (0-1), where the values closer to zero means they are older enzymes. It also shows the enzyme fold class and the PDB id of the enzyme structure. Most of the enzymes with high degree have an ancestory value closer to zero and belong to ancient fold class, proving that they are highly conserved and are older enzymes.

Table 4. Enzyme Ancestry value for the Top Ten Degree Enzymes in the Enzyme Network

Enzyme	Degree	Ancestry Value	Fold class	PDB
1.11.1.6 Catalase	113	0.0314	c.23.16.3	1SY7
2.3.3.8 ATP citrate synthase	111	0.0188	c.1.12	NA
3.5.1.5 Urease	106	0.0188	c.1.9.2	4UBP
4.2.1.52 Dihydrodipicolinate synthase	102	0.0188	c.1.10.1	1S5W
1.2.1.41 Glutamate-5-semialdehyde dehydrogenase	101	0.213	c.82.1.1	1VLU
4.1.1.31 Phosphoenolpyruvate carboxylase	101	0.0188	c.1.12.3	1QB4
4.2.1.11 Phosphopyruvate hydratase	101	0.0188	c.1.11.1	7ENL
1.9.3.1 Cytochrome-c oxidase	101	0.088	a.118.11.1	2OCC
3.1.3.16 Phosphoprotein phosphatase	101	0.088	a.118.8.1	1A17
4.2.1.9 Dihydroxy-acid dehydratase	100	0.0188	c.37.1	NA
3.6.1.1 inorganic diphosphatase	93	0.044	b.40.5.1	8PRK
2.4.2.17 ATP phosphoribosyltransferase	67	0.0125	d.58.5.3	1Q1K
3.3.1.1 Adenosylhomocysteinase	66	0.025	c.2.1.4	1V8B

4 Conclusion

Here we have reconstructed and analyzed the enzyme centric network of *Arabidopsis thaliana* using information available in the KEGG metabolic pathway data. The scale-free behavior of the enzyme network of *A. thaliana* suggests that suggests that during evolution new nodes tend to attach preferentially to a few ancient nodes. This possibly indicates that enzymes (nodes) with very high-degree are likely to be very ancient, which is further confirmed by the analysis of high degree nodes. Our preliminary analysis of betweenness centrality measure suggests that apart from high-degree nodes, nodes having high betweenness values are also very crucial for the stability of the network, and these are typically enzymes catalyzing reactions that connect pathways, or are involved in catalyzing important reactions such as those transferases that help in transfer of a functional group, oxidoreductases that catalyzes the transfer of electrons from one molecule, etc.

5 References

[1] C. R. Yang, An enzyme-centric approach for modelling non-linear biological complexity, *BMC Syst Biol*, vol. 2, p. 70, (2008).

[2] W. d. Nooy, *et al.*, *Exploratory social network analysis with Pajek*. New York: Cambridge University Press, (2005).

[3] ftp://ftp.genome.jp/pub/kegg/ligand/enzyme/enzyme, Kyoto Encyclopedia of Genes and Genomes

[4] H. Jeong, *et al.*, The large-scale organization of metabolic networks, *Nature*, vol. 407, pp. 651-4, Oct 5 (2000).

[5] S. A. Rahman and D. Schomburg, Observing local and global properties of metabolic pathways: 'load points' and 'choke points' in the metabolic networks, *Bioinformatics*, vol. 22, pp. 1767-74, Jul 15 (2006).

[6] A. L. Barabasi and E. Bonabeau, Scale-free networks, *Sci Am*, vol. 288, pp. 60-9, May (2003).

[7] Chiaoquan Qi, Genes encoding hub and bottleneck enzymes of the Arabidopsis metabolic network preferentially retain homeologs through whole genome duplication , *BMC Evolutionary Biology*, **10**:145, (2010).

[8] B. S. Hartley, Evolution of enzyme structure, *Proc R Soc Lond B Biol Sci*, vol. 205, pp. 443-52, Sep 21 (1979).

[9] H. S. Kim, *et al.*, MANET: tracing evolution of protein architecture in metabolic networks, *BMC Bioinformatics*, vol. 7, p. 351, (2006).