The Fourth International Conference on Computational Systems Biology (ISB2010)
Suzhou, China, September 9–11, 2010

Three Faces of Metabolites: Pathways, Localizations and Network Positions*

Jing Zhao[1], Petter Holme[2,3]

1 Department of Mathematics, Logistical Engineering University, Chongqing 400016, China
2 Department of Physics, Umeå University, 90187 Umeå, Sweden
3 Department of Energy Science, Sungkyunkwan University, Suwon 440–746, Korea

Abstract In this study, we investigate the relative organization of three (not independent) categorizations of metabolites—pathways, subcellular localizations and network clusters both qualitatively and quantitatively and further characterize the categories from topological point of view. The picture of the metabolism we obtain is that of peripheral modules, characterized both by being dense network clusters and localized to organelles, connected by a central, highly connected core. Pathways typically run through several network clusters and localizations, connecting them laterally. The significant overlap between organelles, pathways and network clusters suggest that, to some extent, the topology of metabolic networks could spell out the spatial isolations of cellular architectures and functional coherence of metabolic systems. Such systems level analysis of the correlation between different categorizations is helpful for understanding the influence of intracellular organization on the regulation of metabolic processes.

Keywords Metabolic network; Pathway; Subcellular localization; Topology; Modularity

1 Introduction

With the advent of databases attempting to record the entire biochemical reaction systems of different organisms, a host of research questions concerning system-wide organization of biochemical processes became accessible to researchers on metabolism. One way of characterizing the large-scale structure of the metabolism is to divide the metabolites into categories, capturing some roles or functions of the compound. These categories can be defined in different ways. Most conspicuous as building blocks of cellular organization are perhaps spatially isolated entities like organelles, so one way of classifying metabolites is to associate them to the organelles they are present in. Another way would be to take the full metabolic reaction system and try to find categories based on the network topology. Early studies of the topology of metabolic networks found e.g. a skewed degree distribution, hierarchical modular organization, and a well-defined core and a

* Corresponding authors. Email: zhaojanne@gmail.com; petter.holme@physics.umu.se

modular periphery (see Refs. [1] and [2] for reviews). Furthermore, this topological structure is, it has been argues, to some extent correlated with metabolic functionality [3-5].

A feature differentiating eukaryotic from prokaryotic cells is the presence of internal membrane-bound structures called organelles, such as nucleus, mitochondrion, and lysosome. The subcellular compartmentalization by these organelles aggregates enzymes and substrates into spatially isolated localizations, and can therefore regulate different metabolic processes. Some algorithms have been proposed to predict the subcellular localization of proteins or metabolic enzymes [6]. However, researchers have not studied the organizational features of these units quantitatively much before. Recently, this has been possible thanks to two databases that include information on subcellular localization [7, 8].

In this study we investigate how three ways of categorizing metabolites—into pathways, localization and network clusters—are interrelated, and what their relationship can tell us about the system-wide organization of metabolism. We used data from the BiGG database on the human metabolism [8], which includes the lists of catalyzed reactions, localizations of metabolites and annotated pathways. We constructed block models of metabolites to visualize the correlations between different categories and measured the overlap of these categories.

2 Methods

2.1 Data description

Our raw data was obtained from the BiGG [8] database of metabolic networks. This database includes a list of 3311 reactions occurring in eight subcellular compartments:Cytoplasm [c]; Extracellular Space [e]; Mitochondrion [m]; Golgi Apparatus [g]; Endoplasmic Reticulum [r]; Lysosome [l]; Peroxisome [x]; Nucleus [n]. The pathway annotations originated from the KEGG database [9] where reactions are labelled by the pathways as follows: Carbohydrate Metabolism(C); Energy Metabolism(E); Lipid Metabolism(L); Nucleotide Metabolism(N); Amino Acid Metabolism(A); Glycan Biosynthesis and Metabolism(G); Metabolism of Cofactors and Vitamins(V); Xenobiotics Biodegradation and Metabolism(X); Biosynthesis of Secondary Metabolites(S); Transport(T). We assign the pathways of a reaction to its participating metabolites. Every metabolite is thus associated with at least one pathway. When we analyze the network topological features of the nodes, we need each metabolite to belong to only one pathway. To achieve this, we added two pathway categories for metabolites in multiple pathways according to the following scheme:

- For metabolites belonging to two pathway categories including transport, we assign them to the other pathway category than transport.
- For metabolites that belong to at least two pathway categories not including transport, assign them to the Multiple Functions pathway (M).
- For metabolites that belong to at least three pathway categories including transport, assign them to the Multiple Functions and Transport pathway (MT).

2.2 Network reconstruction

In this study, all of the reactions in BiGG database were used to reconstruct human metabolic network we study. In this network, one node is a metabolite in a specific subcellular compartment and there is an edge between two metabolites if they occur in the same reaction and one is the product and the other a substrate. For example, according to BiGG, glucose–6–phosphate is localized to both the compartments c and r giving two nodes in our network.

2.3 Metabolite clustering by network topology

To achieve the network clusters we use the simulated annealing algorithm to find the maximum modularity metric of the network[10]. A network cluster identified by this algorithm is a region of the network that is more strongly coupled within than it to other clusters. The general philosophy of this method is to maximize a measure of modularity of partitions of a network [11]. By allowing some disorder the algorithm avoids getting stuck in local minima. Although modularity maximization algorithms suffer from its resolution limit[12] and other measures aimed to conquer this limitation have been proposed [13], earlier studies have suggested that network clusters identified by modularity maximization algorithms are good candidates of biological modules [4, 10]. We compared the results with a more specialized algorithm [14], but the simulated annealing algorithm could find partitions with larger modularity than this method.

2.4 Block-model network of categories

Block modelling is a general way of structuring and simplifying large-scale organization commonly used in social science [15]. In this methodology one construct a network of categories of nodes that can be linked in various ways. In our plots of the block–model networks, the size of node is proportional to the number of entities that belong to the category, and the width of a link is proportional to the number of links between the two categories. These higher-order networks can be analyzed with general network methods. In this study, we constructed three types of block models with respect to pathways, subcellular localizations and network clusters.

2.5 Matching between different categorizations

We used overlap score to measure the similarity between different categorizations[16]. Consider two categorizations X and Y (for example be subcellular localization and pathways) and assume each metabolite is associated with a subset of the categories of X and Y. Let $\phi_X(x)$ denote the fraction of metabolites in category $x \in X$, and define $\phi_Y(y)$ correspondingly. Let $\phi_{XY}(x,y)$ denote the joint frequency of x and y, i.e. the fraction of vertices that are categorized both as $x \in X$ and $y \in Y$. In a random distribution of functions the expectation value of $\phi_{XY}(x,y)$ is $\phi_X(x)\phi_Y(y)$, but if the categories of different categorizations are overlapping, then some $\phi_{XY}(x,y)$, the ones that overlap, will be larger than $\phi_X(x)\phi_Y(y)$, while for the others $\phi_{XY}(x,y)$ will be lower than $\phi_X(x)\phi_Y(y)$. Thus, both overlapping and not overlapping categories will contribute to $|\phi_{XY}(x,y) - \phi_X(x)\phi_Y(y)|$ and a

prototypical overlap score is

$$\mu = \sum_{x \in X} \sum_{y \in Y} \left| \phi_{XY}(x,y) - \phi_X(x)\,\phi_Y(y) \right|$$

This quantity is, however, affected by finite sizes, meaning that it can be used to compare different systems of the same sizes, not as an overlap measure *per se*. To get a stand-alone overlap measure, we normalized the μ-value against those of perfect overlaps in a system of the same sizes and define overlap score of categorizations *X* and *Y* as follows:

$$\nu_{XY} = \frac{\mu_{XY}}{\max(\mu_{XX}, \mu_{YY})}$$

The value of ν is between 0 and 1 with 1 indicating a perfect match. We generated 1000 pairs of random clusters of the metabolic network, in which the cluster sizes are the same as those of *X* and *Y*, respectively. Then we use z-score [17] of ν to quantify if the overlap score of two categories *X* and *Y* is larger or smaller than expected.

3 Results and discussion

3.1 Relation between the categorizations—pathways, subcellular localizations and network clusters

From the reaction system, we derive a human metabolic network which consists of 2771 nodes and 9451 edges. The simulated-annealing algorithm was applied to decompose this network. 9 *network clusters* were identified with modularity value 0.676. We compared the results with those of the edge-betweenness based algorithm [14], which generated 10 network clusters with modularity value 0.674. Most network clusters from the two algorithms are strongly overlapping. Our second class is the *localization* (or *subcellular compartment*) of the metabolites, i.e. where in the cell a metabolite is occurring in a substantial amount. BiGG defines in total eight categories of this categorization. Our third class is the annotated *pathways* of the BiGG database. BiGG uses the ten pathways from the KEGG database [9]. Network clusters, localization and pathways are different ways of categorizing the metabolites, representing different traits of the metabolites; we will henceforth call them just *categorizations*.

From the categorizations, we constructed three block-model networks [15] with respect to pathway, subcellular-localization and network-cluster, in which the nodes correspond to the categories of the three categorizations respectively. To picture the overlap between the different categorizations, we plot a pie chart per node of the relative number of metabolites of different categories belonging to one categorization (c.f. the cartographic representation of Ref. [10]).

Figure 1 shows the linkages among network clusters. The structure of the block-model networks has the same type of a core-periphery organization observed in Ref. [3]. Clusters 1, 3, 4 and 5 are connected by many reactions, thus forming a core of the block-model network displayed in Fig. 1 and in the full metabolic

network.

Unlike the pathways and subcellular compartments, network clusters are identified without any prior biological background knowledge. As seen in Fig. 1-I and II there is an overlap between network clusters and both pathways and subcellular localizations. Three of the four core clusters are mixtures of metabolites in cytoplasm and extracellular space, reflecting the central role of cytoplasm in cellular metabolism, and many opportunities of material exchange between the cell and its surroundings. The other clusters, including the five more peripheral categories and one core category, are dominant by metabolites from a single organelle, respectively, suggesting a high extent of overlap between the network-cluster and localization categorizations. This feature implies that each organelle respectively defines a compact region in the metabolic network. As a relatively independent organelle, mitochondrion may have more complex functions and more interactions with the cytoplasm than the others. Projecting to topology, its corresponding network cluster, 3, is analoguously a core cluster. From functional point of view, the core clusters are multi-functional categories in which multiple pathways are almost evenly distributed, whereas the peripheral categories exhibit to own a major function, for instance, glycan biosynthesis and metabolism for 2 and 9, and lipid metabolism for 6 and 7.

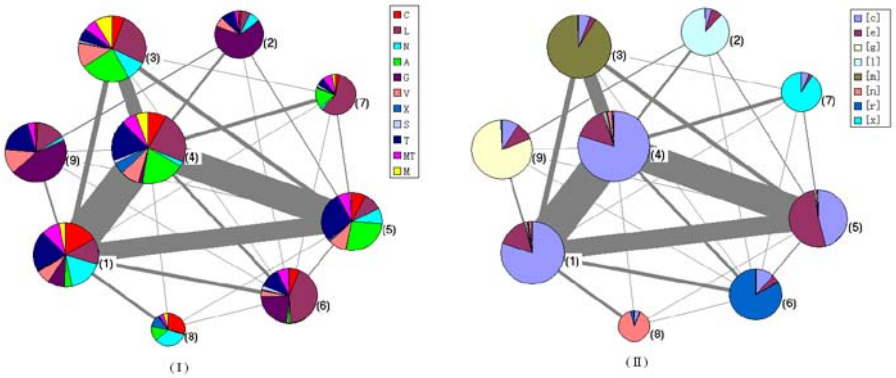

Figure 1- Cartographic representations of the block-model with respect to network clusters. Each circle represents a network cluster and is coloured according to the fractions of pathway (I) and localization (II) respectively, while the edges reflect the connections between clusters.

3.2 Quantitative difference between categories

In this section we apply an overlap score ν together with the z-score to quantitatively measure the overlap of the three categories compared with randomly assigned categories. As defined in the Method section, the value of ν is between 0 and 1, with 1 representing a perfect match. A larger ν indicates a higher extent of overlap between the two categories. To get a meaningful value of how much

different from expected the value is we use the z-score. The z-score measures how many multiples of standard deviation that the gap between the overlap score of the two categories and the average overlap score of the random category pairs. Since about 99.7% of values drawn from a normal distribution lie within three standard deviations away from the mean, a z-score larger than three means, with 99.7% certainty, a larger value of *f* than expected. Thus it can quantify if the overlap score of two categories is larger or smaller than expected.

For localization and network clusters we obtain $\nu = 0.72$, z =108.46; for network clusters and pathways we get $\nu = 0.37$, z = 45.94; and finally pathways versus localization gives $\nu = 0.42$, z = 54.69. The fairly large z-scores for all the categories mean that the categories are significantly correlated with each other. On the other hand, there are differences; so they do also measure different organizational traits. The subcellular localization and network clusters are overlapping most of the three pairs of functional categories. This effect seems to stem from the strong overlap between organelles and the peripheral network clusters (as seen in Fig.1-II), suggesting that topologically compact regions in the metabolic network correspond to physically and morphologically individual subcellular entities.

3.3 Network structure of the categories of the three categorizations

From the early observations of broad and skewed degree distributions [18], we first examine the average degree in the network for each category of the three categorizations, i.e. the average interactions within the categories, see Figure 2-I–III. For categorization according to different pathways, the hub metabolites are gathered in the categories “Multiple Functions and Transport” (MT) and “Multiple Functions” (M), especially in MT. These categories constitute only 6% and 4% of the metabolites respectively but are essential for keeping the network connected [18]. In contrast, the metabolites appearing in T (the pure transport category) have the lowest extent of metabolic interactions with others. One explanation of this is that MT does not include metabolites localized in extracellular space, whereas most metabolites of T are localized in extracellular space (66%) and cytoplasm (25%). The MT metabolites seem more involved in moving metabolites across intracellular membrane boundaries than the T metabolites that is more specialized in transport across cell walls. The corresponding study for the different subcellular compartments (seen in Fig. 2-II) indicates that metabolites in cytoplasm and mitochondrion have more interactions with others, and those in extracellular space have the lowest average degree. In Fig. 2-III we see that for the network-cluster categorization the high-degree nodes are primarily located in the core clusters, meaning that the cores themselves are more highly connected than the peripheries.

The node degree measures the local importance of a node. To get a more global view about the position of metabolites in the network, we also measure the betweenness centrality. The betweenness of a node is proportional to the number of shortest paths between pairs of nodes. Assuming metabolic processes preferably occur via short paths, betweenness should be a better indicator than degree for global centrality and importance. In particular, the betweenness is high for nodes connecting different network clusters. Figure 2-IV through VI shows the average

betweenness for the different categorizations—pathways, subcellular compartments and network clusters, respectively. Metabolites in the MT pathway, localizations c and g, and network cluster 4 have high betweenness. One example showing that the global information of betweenness is more informative than the degree is that the Golgi apparatus is the organelle with highest betweenness, which is consistent with its central function of packaging macro molecules for secretion. For the network clusters, as can be guessed from Fig. 1, cluster 4 is the one with highest average betweenness of the metabolites. The significant higher betweenness of MT metabolites than those of both M and T ones suggests the obviously different roles these three clusters of metabolites play in the metabolic flux — metabolites involved in multiple pathways including transport (MT) are the most important media for material and information flow in the metabolic system.

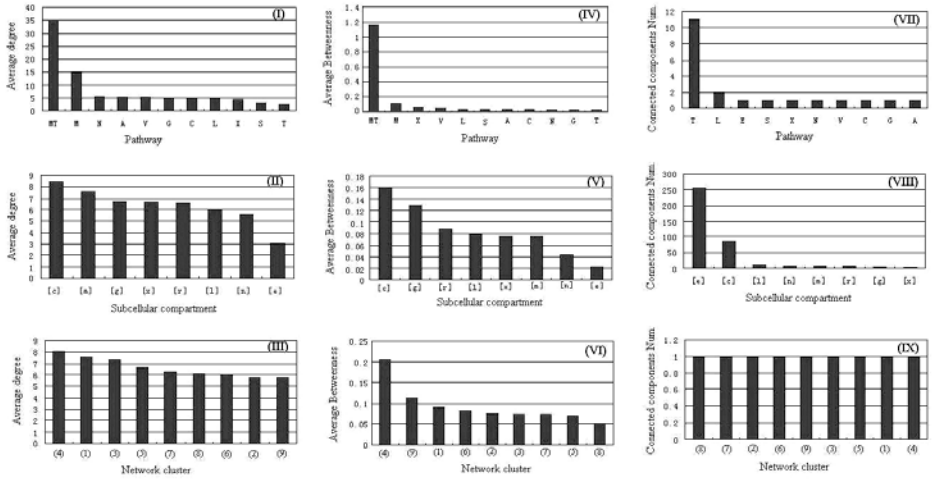

Figure 2- Network structures of the categories.

In Fig. 2-VII–IX, we investigate the number of connected components of the subnetworks of nodes of the same category. For the pathway categorization (Fig. 4-VII), we note that subnetwork defined by the transportation pathway is broken into over 10 isolated clusters (note that this pathway is also the sparsest, see Fig 4-I, sparse networks are naturally more prone to be disconnected). This highlights a difference between pathways and network clusters—pathways need not be independent units (an often quoted definition of "module") by construction. The energy metabolism pathway is connected (and also the second most densely linked pathway). Many metabolites of this pathway are small molecules normally labelled as carriers for transferring electrons or certain functional groups, such as ATP, NADH and H_2O. Such "currency metabolites" have often many more links than regular metabolites, explaining the density and connectedness of this pathway. Figure 2-VIII shows that the extracellular space compartment has a very fragmented network, even more than the transport pathway. The other non-organelle compartment, cytoplasm, is also disconnected. The organelle compartments, on the

other hand, are connected. In sum, metabolites localized to the extracellular space and cytoplasm act as links between the more independent metabolic subnetworks of the organelles. Since our network clustering algorithm is designed to find densely connected regions it is no wonder that the network clusters are all connected (Fig. 2-IX).

4 Conclusions

To understand a large system such as the metabolism one need to simplify and categorize its components. In this paper we have investigated three ways of doing this—grouping metabolites according to pathways, localization, and network clusters respectively. From topological point of view, all the three categories are globally organized in a modular core-periphery pattern. Specifically, the compartmentalization is clearly organized into a core of extracellular media and cytoplasm, and a periphery of organelles; while the core and periphery modules of the pathway interactions correspond to housekeeping and advanced specific functions respectively. There are peripheral network clusters overlapping almost completely with the organelle categories or being dominated by one major category of pathways. Qualitative and quantitative analysis shows that the three categories are significantly correlated with each other, suggesting the interrelationship between biochemical specific functions, spatial isolations in cells, and topological compact regions in metabolic networks. Our results suggest that the spatial organizations and functional coherence of cellular metabolic systems have been imprinted in the topology of metabolic network. Therefore, though the traditional classification of metabolites into pathways or organelles has provided abundant information to biologists, our study about their correlation according to topology could shed light on the underlying structures supporting the metabolic function of cells, and thus could provide a basis for further metabolic modelling.

Acknowledges

JZ is supported by grant from National Natural Science Foundation of China (10971227). PH acknowledges support from the Swedish Foundation for Strategic Research, the Swedish Research Council and the WCU program through NRF Korea funded by MEST R31–2008–000–10029–0.

References

[1] Zhao J, Yu H, Luo J et al. Complex networks theory for analyzing metabolic networks. Chinese Science Bulletin, 2006, 51:1529-1537.
[2] Lacroix V, Cottret L, Thébault P et al. An introduction to metabolic networks and their structural analysis. IEEE / ACM Transactions on Computational Biology and Bioinformatics 2008, 5:594-617.
[3] Zhao J, Ding G-H, Tao L et al. Modular co-evolution of metabolic networks. BMC Bioinformatics, 2007, 8:311.
[4] Vitkup D, Kharchenko P, Wagner A. Influence of metabolic network structure and function on enzyme evolution. Genome Biology, 2006, 7:R39.
[5] Zhao J, Yu H, Luo J-H et al. Hierarchical modularity of nested bow-ties in metabolic

networks. BMC Bioinformatics, 2006, 7:386.
[6] Mintz-Oron S, Aharoni A, Ruppin E et al. Network-based prediction of metabolic enzymes' subcellular localization. Bioinformatics, 2009, 25:i247-1252.
[7] Duarte Nc, Herrgard Mj, Palsson Bo. Reconstruction and Validation of Saccharomyces cerevisiae iND750, a Fully Compartmentalized Genome-Scale Metabolic Model. Genome Res, 2004, 14:1298-1309.
[8] Duarte Nc, Becker Sa, Jamshidi N et al. Global reconstruction of the human metabolic network based on genomic and bibliomic data. Proc Natl Acad Sci USA, 2007, 104:1777-1782.
[9] Kanehisa M, Goto S. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res, 2000, 28:27-30.
[10] Guimera R, Amaral Lan. Functional cartography of complex metabolic networks. Nature, 2005, 433:895-900.
[11] Newman M, Girvan M. Finding and evaluating community structure in networks. Phys Rev E, 2004, 69:026113.
[12] Fortunato S, Barthelemy M. Resolution limit in community detection. Proc Natl Acad Sci USA, 2007, 104:36-41.
[13] Li Z, Zhang S, Wang R-S et al. Quantitative function for community detection. Phys Rev E, 2008, 77:e036109.
[14] Newman M, Girvan M. Finding and evaluating community structure in networks. Phys Rev E, 2004, 69:026113.
[15] Wasserman S, Faust K: Social Network Analysis: Methods and Applications. Cambridge: Cambridge University Press; 1994.
[16] Holme P. Model validation of simple-graph representations of metabolism. J R Soc Interface, 2009, 6:1027-1034.
[17] Maslov S, Sneppen K, Zaliznyak A. Detection of topological patterns in complex networks: correlation profile of the internet. Physica A: Statistical and Theoretical Physics, 2004, 333:529-540.
[18] Jeong H, Tombor B, Albert R et al. The large-scale organization of metabolic networks. Nature, 2000, 407:651-654.