

# Systematic framework for metabolic networks with constrained biological information: analysis and applications

Zoran Nikoloski\*, Patrick May, Joachim Selbig

Max-Planck Institute for Molecular Plant Physiology  
University of Potsdam, Institute for Biochemistry and Biology  
Potsdam, Brandenburg, Germany  
\*E-mail: nikoloski@mpimp-golm.mpg.de

A *metabolic network* models the relationship among three different groups of constituting elements: reactions, enzymes, and metabolites, from a given set of metabolic pathways integrally related via gene-regulatory mechanisms. Recent pathway/genome databases inevitably require development of a different methodology for including detail in the network-centric approach to answer concrete biological questions [5]. We define six *metabolic models with constrained biological information* in which a portion of pathway-centric information is retained. In an attempt to identify the underlying rules which guide the integration of the constituents, much of the scientific effort has been focused thenceforth on metabolite-metabolite and enzyme-enzyme metabolic networks [6].

As simple optimization problems (*e.g.* shortest path, minimum cut) become NP-hard on hypergraphs [1], we define the *metabolite-enzyme-reaction (MER) network*, amenable to graph-theoretic analysis and related to metabolic kinetics [4], with node set partitioned into three sets of constituents. A directed link exists: from an enzyme  $u$  to a reaction  $v$  and vice versa if  $u$  catalyzes  $v$ ; from a metabolite  $u$  to a reaction  $v$  if  $u$  is a reactant in  $v$ ; and from a reaction  $u$  to a metabolite  $v$  if  $v$  is a product in  $u$ . Six possible networks can be obtained from a given MER network via graph transformations (Figure 1, a).

We conduct a comprehensive empirical study of the following network invariants for the obtained networks: degree distribution, degree-degree correlation, bipartite cores, strongly (weakly) connected components, diameter, and average path-length. The degree distribution is fitted to six different probability distribution using maximum likelihood estimation, of which Lognormal, Pareto, and Yule exhibits smallest value for the Akaike information criterion (Figure 2). The empirical evidence demonstrates that metabolic networks of *E. coli*, *S. cerevisiae*, and *Arabidopsis* are not scale-free, and exhibit different distribution of bipartite cores and average length than the existing power-law models [8, 7, 2] (Figure 1, b). These topological properties across the seven networks can be employed in a feature vector to efficiently compare metabolic networks across species [3].

As a consequence of the empirical analysis, we arrive at a growth model which mimics the invariants of metabolic networks with constrained biological information. Given a pool of metabolites and enzymes, an MER network is built by adding new reactions to the system as follows: (1) with a probability  $\alpha$ , add a new reaction with number of reactants chosen from distribution  $P_1$ , number of products chosen from distribution  $P_2$ , and an enzyme chosen

uniformly at random; and (2) with probability  $1 - \alpha$  select a reaction  $r$  uniformly at random, select a number  $c$  from a distribution  $P_3$ . If  $c$  is smaller than the number of reactants (products), choose  $c$  of the reactants (products) from  $r$  uniformly at random; otherwise, copy all reactants (products) and select the remaining uniformly at random. The empirical bipartite core distribution is used to validate the number shared metabolites in a set of reactions. Using arguments similar to those in [9, 10], we show that the model has degree distribution which follows the lognormal-Pareto, exhibits average path length bounded above by a constant independent on the number of reactions, and has the expected number of bipartite cores.

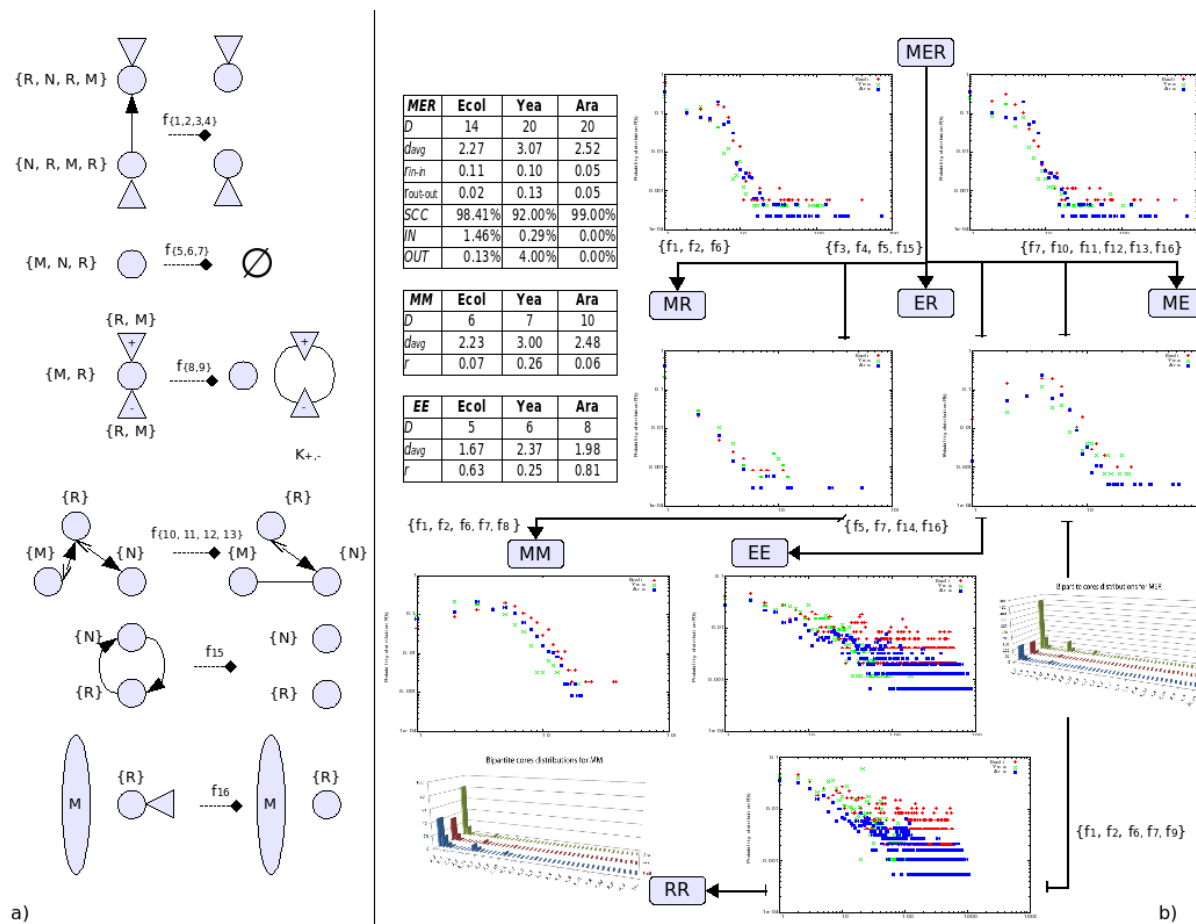


Figure 1: a) Graph transformations applied to MER—R (node is from the set of reactions), N (from the set of enzymes), and M (from the set of metabolites).  $f_1$  takes a directed edge from an enzyme node to a reaction node and removes it from MER. Metabolite-metabolite (MM) is obtained by applying  $\{f_1, f_2, f_6, f_7, f_8\}$  in an arbitrary order; for the other networks, the set of transformations are given on the arcs of the tree b) Summary of topological properties for the seven types of metabolic networks. The degree distributions for six networks are presented in a log-log scale next to the label of the network. The bipartite core distribution for partition sizes from (2,2) to (10, 10) is given for MER and MM networks. The average path length does not depend on the order and the size of the network, as shown in the legend.

AIC	num parameters	LN	PAR	YUL
		3	2	1
MER in-degree	Ecoli	-10466,22	-15047,96	-15049,96
	Yea	-10256,15	-10159,33	-10161,33
	Ara	-24475,58	-31770,20	-31772,20
MER out-degree	Ecoli	-10466,22	-15047,96	-15049,96
	Yea	-7280,02	-10159,33	-10161,33
	Ara	-24475,58	-31770,20	-31772,20
MM	Ecoli	-2995,69	-2580,26	-2582,26
	Yea	-2298,66	-2376,15	-2378,15
	Ara	-6530,06	-6169,68	-6171,68
EE	Ecoli	-4175,68	-1369,89	-1371,89
	Yea	-1974,23	-1297,66	-1299,66
	Ara	-8442,02	-3505,86	-3507,86

Figure 2: Akaike information criterion (AIC) for ranking models of degree distribution. The results demonstrate quantitatively that the degree distribution across the seven types of metabolic networks is very likely to be lognormal.

## References

- [1] G. Ausiello, R. Giaccio, G. Italiano, and U. Nanni. Optimal traversal of directed hypergraphs. Manuscript, 1997.
- [2] A.-L. Barabasi, R. Albert, and H. Jeong. Mean-field theory for scale-free random networks. *Physica A*, 272:173–189, 1999.
- [3] M. Dehmer and F. Emmert-Streib. Comparing large graphs efficiently by margins of feature vectors. *Applied Mathematics and Computation*, 188:1699–1710, 2007.
- [4] K. Gatermann and S. Hosten. Computational algebra for bifurcation theory. *Journal of Symbolic Computation*, 40(4–5):1180–1207, 2005.
- [5] S. Huang. Back to the biology in systems biology: What can we learn from biomolecular networks? *Briefings in Functional Genomics and Proteomics*, 2:279–297, 2004.
- [6] H. Jeong, B. Tomber, R. Albert, Z. Oltvai, and A.-L. Barabasi. The large-scale organization of metabolic networks. *Nature*, 407:651, 2000.
- [7] J. Kleinberg, R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. The web as a graph: measurements, models, and methods. In *Proceedings of the 5th International Conference on Combinatorics and Computing*, pages 1–17, 1999.
- [8] R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tomkins, and E. Upfal. Stochastic models for the web graph. In *Proceedings of the 41st Annual Symposium on Foundations of Computer Science*, pages 57–65, 2000.
- [9] W. J. Reed and M. Jorgensen. The double pareto-lognormal distribution: A new parametric model for size distributions. *Communications in Statistics: Theory and Methods*, 33(8):1733–1753, 2003.
- [10] N. C. Wormald. Differential equations for random processes and random graphs. *Annals of Applied Probability*, 5(4):1217–1235, 1995.