BIOINFORMATICS APPLICATIONS NOTE Vol. 27 no. 19 2011, pages 2761–2762
doi:10.1093/bioinformatics/btr448

Systems biology

Advance Access publication July 29, 2011

JMassBalance: mass-balanced randomization and analysis of metabolic networks

Georg Basler[1,*] and Zoran Nikoloski[1,2]

[1]University of Potsdam, Institute for Biochemistry and Biology and [2]Max Planck Institute for Molecular Plant Physiology, D-14476 Potsdam, Germany

Associate Editor: Mario Albrecht

ABSTRACT

Summary: Analysis of biological networks requires assessing the statistical significance of network-based predictions by using a realistic null model. However, the existing network null model, *switch randomization*, is unsuitable for metabolic networks, as it does not include physical constraints and generates unrealistic reactions. We present *JMassBalance*, a tool for mass-balanced randomization and analysis of metabolic networks. The tool allows efficient generation of large sets of randomized networks under the physical constraint of mass balance. In addition, various structural properties of the original and randomized networks can be calculated, facilitating the identification of the salient properties of metabolic networks with a biologically meaningful null model.

Availability and Implementation: *JMassBalance* is implemented in Java and freely available on the web at http://mathbiol.mpimp-golm.mpg.de/massbalance/.

Contact: basler@mpimp-golm.mpg.de

Received on May 30, 2011; revised on July 26, 2011; accepted on July 27, 2011

1 INTRODUCTION

Network-based studies of biological systems attempt to relate topological properties to biological function. The first step in drawing this connection involves determining the network properties which do not arise by chance. To this end, a network null model can be used to assess the statistical significance of network properties.

The common approach for determining the statistical significance of a given property is to determine a P-value based on the following procedure: (i) determine the chosen property from an investigated biological network; (ii) sample a large number of random networks under biologically meaningful constraints; and (iii) estimate the mean and variance of the property from the simulated networks to calculate a z-score (with the corresponding P-value) under the assumption of normal distribution.

Clearly, the significance of a network property strongly depends on the null model. The commonly used method, *switch randomization* (Guimerà *et al.*, 2007; Milo *et al.*, 2002; Sales-Pardo *et al.*, 2007), does not account for physical constraints, and thus generates unrealistic biochemical reactions (see Basler *et al.*, 2011, for an example). Thus, it is questionable whether the significance determined by this generic randomization scheme helps to elucidate the relation between network properties and biological functions.

*To whom correspondence should be addressed.

Motivated by the lack of a biologically meaningful null model for metabolic networks, we developed a method for randomizing metabolic networks under the constraint of mass balance, and analyzed its computational complexity and uniformity of sampling (Basler *et al.*, 2011). Here, we present a tool which can be run via a graphical user interface (GUI) or from the command line, and implements mass-balanced randomization of metabolic networks provided in one of three standard data formats: (i) BioCyc (http://www.biocyc.org); (ii) Systems Biology Markup Language (SBML, http://sbml.org); or (iii) a customizable text file format.

2 METHOD

A metabolic network is represented as a weighted directed bipartite graph $G=(V_c \cup V_r, E)$, where V_c is the set of compound nodes, V_r the set of reaction nodes, and $E \subseteq (V_c \times V_r) \cup (V_r \times V_c)$ is the set of weighted, directed edges denoting stoichiometric substrate-reaction and product–reaction relationships. For example, an edge (c,r) specifies that compound c is a substrate of reaction r, while the stoichiometric coefficient $s_{c,r}$ of c in r is represented as the weight of (c,r).

A compound node is uniquely represented by a name, a compartment and a mass vector, $m_c \in \mathbb{N}^n$, i.e. the vector representation of the compound c over n chemical elements. For instance, when considering the six most abundant elements in biological systems: carbon (C), hydrogen (H), nitrogen (N), oxygen (O), phosphorus (P) and sulfur (S), then the mass vector of water is $m_{H_2O}=(0,2,0,1,0,0)\cdot(C,H,N,O,P,S)^T$. The set of considered chemical elements can be specified in a configuration file (see Reference Manual, available online at http://mathbiol.mpimp-golm.mpg.de/massbalance/).

For a reaction r, r_{in} denotes the set of substrates, and r_{out} the set of products. A reaction node is uniquely represented by a name and its direction: reversible reactions are represented by one reaction node for each direction, r^+ and r^-, where $r^+_{in}=r^-_{out}$ and $r^+_{out}=r^-_{in}$. A reaction is *mass balanced*, i.e. chemically feasible with respect to the conservation of mass, if the sum of its substrate atoms equals the sum of its product atoms:

$$\sum_{c \in r_{in}} s_{c,r} \cdot m_c = \sum_{k \in r_{out}} s_{k,r} \cdot m_k. \quad (1)$$

The randomization procedure consists of a pre-calculation step, which classifies the compounds from the network according to their chemical sum formula (see Basler *et al.*, 2011), followed by the actual randomization. The pre-calculation is executed only once for all subsequent randomizations of the same network, and renders the method applicable to large networks. A network is randomized by replacing the substrates and products of randomly chosen reactions by compounds from within the same network, and choosing their stoichiometric coefficients, such that Equation (1) is satisfied (Fig. 1). The polynomial-time algorithm generates randomized networks uniformly at random and clearly outperforms switch randomization (see Basler *et al.*, 2011, Supplementary Table S1).

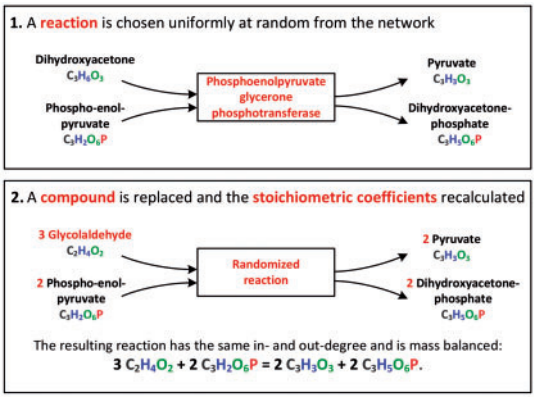

Fig. 1. Mass-balanced substitution of a substrate. A large number of substitutions is applied in order to obtain fully randomized networks.

3 APPLICATION

JMassBalance is written in Java and comes with all required libraries. Hence, an installation is not required, and it can be used on any operating system with installed Java (http://www.oracle.com).

The randomization procedure accepts network files in BioCyc, SBML, or a customizable text format. Additional optional parameters allow specifying whether unbalanced reactions in the original network should be fixed, whether compartments should be considered, the randomization depth and probability, and the number of randomized networks to generate. All calculations can easily be parallelized by executing the program multiple times with different network indices (see online Reference Manual). Switch randomization is also implemented, and can be applied to compare the results of the two null models.

In addition to randomization, the following structural properties can be calculated for the original and randomized networks, respectively, which allows to determine their statistical significance in a biologically meaningful context:

- Average path length: the average number of reactions on the shortest path between two compounds.
- Clustering coefficient: average fraction of mutually connected neighbors of a node in the corresponding (unipartite) metabolite–metabolite network.
- Assortativity: correlation coefficient of the in-/out-degree of a node and the average in-/out-degree of its predecessors/successors in the corresponding (unipartite) metabolite–metabolite network.
- n-cycles: the number of directed cycles of length n in the corresponding (unipartite) metabolite–metabolite network.
- Path: test whether the given compounds constitute a path.
- Connectedness: test whether the given compounds are connected via paths.
- Transition degree: the number of possible mass-balanced substitutions.
- Local essentiality: the ratio of successor reactions affected by the knockout of a reaction.
- Reaction centrality: the ratio of reactions globally affected by the knockout of a reaction.
- Knockout set: the set of reactions globally affected by the knockout of a given reaction.
- Degree distribution: the compound degree distribution.
- Weight distribution: the distribution of edge weights.
- Scope size distribution (Handorf *et al.*, 2005): the distribution of the number of compounds producible from a random set of seed compounds of the given size.
- Distribution of $\Delta_r^0 G$ (Mavrovouniotis, 1991): the distribution of the standard Gibbs free energy change of reactions.

The randomized networks may be printed as stoichiometric matrices or as text files, thus enabling subsequent investigations, such as constraint-based analysis (Feist *et al.*, 2010).

4 CONCLUSION

JMassBalance is a flexible and efficient tool for assessing the significance of metabolic network properties through a biologically meaningful null model. It can be used to determine the salient structural properties of metabolic networks and to identify new properties, which are statistically significant and independent of basic physical constraints. Thus, we believe the tool is useful for the initial analysis of reconstructed metabolic networks, as well as subsequent network-based research.

Funding: German Federal Ministry of Education and Research (grant number 0313924).

Conflict of Interest: none declared.

REFERENCES

Basler,G. *et al.* (2011) Mass-balanced randomization of metabolic networks. *Bioinformatics*, **27**, 1397–1403.

Feist,A.M. *et al.* (2010) Model-driven evaluation of the production potential for growth-coupled products of escherichia coli. *Metab Eng.*, **12**, 173–186.

Guimerà,R. *et al.* (2007) Classes of complex networks defined by role-to-role connectivity profiles. *Nature Physics*, **3**, 63–69.

Handorf,T. *et al.* (2005) Expanding metabolic networks: Scopes of compounds, robustness, and evolution. *J. Mol. Evol.*, **61**, 498–512.

Mavrovouniotis,M. (1991) Estimation of standard gibbs energy changes of biotransformations. *J. Biol. Chem.*, **266**, 14440–14445.

Milo,R. *et al.* (2002) Network motifs: simple building blocks of complex networks. *Science*, **298**, 824–827.

Sales-Pardo,M. *et al.* (2007) Extracting the hierarchical organization of complex systems. *Proc. Natl Acad. Sci. USA*, **104**, 15224–15229.