

# Automated analysis of information processing, kinetic independence and modular architecture in biochemical networks using MIDIA

Clive G. Bowsher

School of Mathematics, University of Bristol and Centre for Systems Biology Edinburgh, UK

Associate Editor: Martin Bishop

## ABSTRACT

**Motivation:** Understanding the encoding and propagation of information by biochemical reaction networks and the relationship of such information processing properties to modular network structure is of fundamental importance in the study of cell signalling and regulation. However, a rigorous, automated approach for general biochemical networks has not been available, and high-throughput analysis has therefore been out of reach.

**Results:** Modularization Identification by Dynamic Independence Algorithms (MIDIA) is a user-friendly, extensible R package that performs automated analysis of how information is processed by biochemical networks. An important component is the algorithm's ability to identify exact network decompositions based on both the mass action kinetics and informational properties of the network. These modularizations are visualized using a tree structure from which important dynamic conditional independence properties can be directly read. Only partial stoichiometric information needs to be used as input to MIDIA, and neither simulations nor knowledge of rate parameters are required. When applied to a signalling network, for example, the method identifies the routes and species involved in the sequential propagation of information between its multiple inputs and outputs. These routes correspond to the relevant paths in the tree structure and may be further visualized using the Input–Output Path Matrix tool. MIDIA remains computationally feasible for the largest network reconstructions currently available and is straightforward to use with models written in Systems Biology Markup Language (SBML).

**Availability:** The package is distributed under the GNU General Public License and is available, together with a link to browsable Supplementary Material, at <http://code.google.com/p/midia>. Further information is at [www.maths.bris.ac.uk/~macgb/Software.html](http://www.maths.bris.ac.uk/~macgb/Software.html).

**Contact:** C.Bowsher@bristol.ac.uk

**Supplementary information:** The Supplementary Material contains extensive description of the MIDIA package and is available at *Bioinformatics* online.

Received on October 21, 2010; revised on November 24, 2010; accepted on December 10, 2010

## 1 INTRODUCTION

Understanding how information is encoded and transferred by biochemical networks is of central importance in cellular and systems biology (Barkai and Shilo, 2007; Nurse, 2008). We recently introduced a rigorous new approach grounded in biochemical, mass

action kinetics and based on dynamic conditional independences between species trajectories (Bowsher, 2010, 2011). A species trajectory is the time course of the number of molecules of a particular type of biomolecule. By deriving dynamic conditional independences, we are able to identify the species trajectories that fully encode the relevant information and thus to trace the sequential process of information transfer through the network (Section 5.2.1 in Supplementary Material). Importantly, the approach is applicable to a wide class of stochastic dynamics (Bowsher, 2011).

A suitable foundation has thus been laid for automated and potentially high-throughput computational analysis of information processing, kinetic independence and modular architecture for a wide class of biochemical networks. Here, the necessary algorithms and software for such analyses are made available to the community in a user-friendly, extensible package called MIDIA (Modularization Identification by Dynamic Independence Algorithms). The package is written in the freely available R language and is therefore applicable to the majority of operating systems (including Windows, Mac and Linux). MIDIA is straightforward to use with models of intracellular dynamics written in Systems Biology Markup Language or SBML (Hucka *et al.*, 2003) and hence also with public repositories of systems biology models such as the BioModels Database (Le Novère *et al.*, 2006).

An important component of MIDIA is the ability to compute exact network decompositions based on dynamic independence properties of the modules (Section 5.2.3 in Supplementary Material). MIDIA identifies the species resulting in the overlap of modules and thus the important intermediaries in the network. Of all those species present in other modules, only those in a given module's overlap region or intersection are relevant for the instantaneous kinetics of its other species and (if stipulated) for the trajectory of the module. Noteworthy departures from previous approaches to module identification are that the decompositions are explicitly based on dynamic, informational properties, and that these decompositions are not partitions. Graphical, community detection-based methods (Guimerà and Amaral, 2005; Kashtan and Alon, 2005) were extended by Saez-Rodriguez *et al.* (2008) to kinetic (but solely deterministic) models of signalling networks by partitioning the species between modules so as to minimize inter-modular 'retroactivity'. Constraint-based module detection methods such as correlated reaction sets (or 'Co-sets', Papin *et al.*, 2004) assume steady-state dynamics in which concentrations are constant over time. We believe that this limits the scope of their applicability,

particularly in the context of the dynamics of gene regulatory and signalling networks.

## 2 RESULTS

Use of the majority of the functionality of MIDIA requires no knowledge of the R language beyond that needed to call the `MIDIA(.)` function. The function's input variables are shown immediately below.

```
> MIDIA(mS,uG,mR,mP,MODE,OutputSelectn,Fraternise,
PLOT,InfoProc,Granularity,MaxIter,PlotJTree)
```

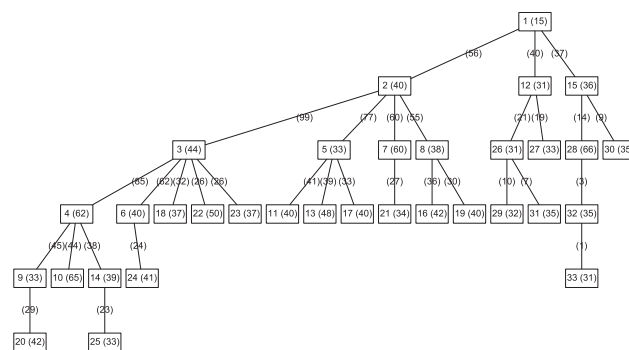
### 2.1 Inputs

To analyse a biochemical reaction network, MIDIA requires specification of that network in terms of its *reactant*, *product* and *quasi-stoichiometric* matrices. The first two matrices simply identify the reactants and products of each reaction. The third identifies two groups of species for each reaction: those overall consumed by the reaction and those overall produced by it. Input of a full (as opposed to quasi-) stoichiometric matrix is not necessary but, when available, refines the resultant analysis. The MIDIA function `mRmSmPfromSBML("SBMLmodel.xml")` returns the required input matrices for a network written in SBML using the `rsbml` package (an R binding for `libsbml` available within Bioconductor's `biocLite`).

The full functionality of MIDIA, specifically the analysis of information processing properties, requires that a *reaction* network is specified (in the manner just described). Nevertheless, for all other types of biomolecular network that may be represented as a graph, MIDIA provides a generic approach for exact, computationally efficient decomposition into possibly overlapping modules. Protein interaction and gene regulatory networks, for example, can thus be modularized by directly inputting the undirected version of the graph corresponding to the network (using the variable `uG`). The extent of coarse-graining employed in a modularization may be controlled using the variable `Granularity`. A full description of all input variables is given in Sections 3 and 4 in Supplementary Material.

### 2.2 Outputs

Several outputs may be computed and plotted by MIDIA when analyzing a biochemical reaction network. The Kinetic Independence Graph (KIG, Section 5.2.2 in Supplementary Material), is a fundamental, graphical description of the network's mass action kinetics and is needed for all subsequent outputs. For each node (i.e. species), this directed graph displays those other nodes whose copy number influences the instantaneous, stochastic kinetics of that species (Bowsher, 2010, 2011). There are two basic types of modularization that serve as MIDIA outputs (Section 5.2.3 in Supplementary Material)—those based on instantaneous or local kinetics alone ( $\mathcal{T}_{MI}$ ), and those based also on the dynamic conditional independence of modules ( $\mathcal{T}_M$ ). Both modularization types are represented and plotted by MIDIA as (junction) trees. The second type,  $\mathcal{T}_M$ , is computed from the first and is the one that enables the analysis of information processing properties. Finally, the Input–Output Path Matrix (Section 5.2.4 in Supplementary Material) is a powerful tool—complementary to the tree structure  $\mathcal{T}_M$ —for visualizing ‘routes’ of information propagation between specified input–output pairs or between pairs



**Fig. 1.** Kinetic Modularization of YEASTNET, the consensus reconstruction of yeast metabolism. The modularization,  $\mathcal{T}_M$ , based on mass action kinetics. The species in each rectangle and in its associated edges in the tree constitute a single module of the decomposition. Each edge thus represents the overlap of 2 particular modules. The species in the edges, 13% of all network species, are identified by MIDIA as important kinetic intermediaries. Modules are uniquely numbered and numbers of species are shown in parentheses. The variables `MODE` and `Granularity` were set to `COARSEGRAIN` and 30, respectively. No rectangle in the tree has species in common with an edge or another rectangle. The comparable analysis using the version of YEASTNET without multiple reaction compartments (data not shown) finds 12% of all network species in the edges.

of modules of particular interest. The various output types and their interpretation are discussed, together with illustrative examples and sample code, in Sections 2 and 5.2 in Supplementary Material (see also Sections 3 and 4 in Supplementary Material).

## 3 DISCUSSION

In order to illustrate the ability of the MIDIA software to analyse large network reconstructions, it is applied here to YEASTNET (v4.02, see [www.comp-sys-bio.org/yeastnet](http://www.comp-sys-bio.org/yeastnet)), the consensus reconstruction of yeast metabolism that builds on the one published by Herrgård *et al.* (2008). For simplicity, attention is focused on separate analysis of the largest, single connected component, consisting of 1730 reactions and 1520 species. Figure 1 shows the modularization based on instantaneous kinetics,  $\mathcal{T}_{MI}$ , returned by MIDIA (run time of 2.1 h on a 2.5 GHz, 3.5 GB RAM laptop machine). The analysis identifies, e.g. the particular 202 of the 1520 species (13%) solely responsible for mediating kinetic effects between the module rectangles that themselves contain the vast majority of the network species (see caption also).

MIDIA is a powerful, user-friendly and extensible software tool for computational analysis of information processing, dynamic independence and modular architecture in biochemical networks, including reaction networks that exhibit a wide range of stochastic dynamics. A particular strength is that uncertainty about the network structure and kinetics is accommodated in the following respects: no knowledge of rate parameters is required; partial stoichiometric information alone needs to be used as input (exact stoichiometries are not needed); and the false inclusion of reactions in the network specification invalidates neither the information processing analysis nor the computed modularizations.

## ACKNOWLEDGEMENT

The author is grateful to Peter Swain for helpful encouragement.

**Funding:** The research was funded jointly by the MRC and EPSRC, UK.

**Conflict of Interest:** none declared.

## REFERENCES

- Barkai,N. and Shilo,B.-Z. (2007) Variability and robustness in biomolecular systems. *Mol. Cell*, **28**, 755–760.
- Bowsher,C.G. (2010) Stochastic kinetic models: Dynamic independence, modularity and graphs. *Ann. Stat.*, **38**, 2242–2281.
- Bowsher,C.G. (2011) Information processing by biochemical networks: a dynamic approach. *J. Roy. Soc. Interface*, **8**, 186–200.
- Guimerà,R. and Amaral,L.A.N. (2005) Functional cartography of complex metabolic networks. *Nature*, **433**, 895–900.
- Herrgård,M.J. *et al.* (2008) A consensus yeast metabolic network reconstruction obtained from a community approach to systems biology. *Nat. Biotechnol.*, **26**, 1155–1160.
- Hucka,M. *et al.* (2003) The systems biology markup language (SBML): A medium for representation and exchange of biochemical network models. *Bioinformatics*, **9**, 524–531.
- Kashtan,N. and Alon,U. (2005) Spontaneous evolution of modularity and network motifs. *Proc. Nat. Acad. Sci.*, **102**, 13773–13778.
- Le Novère,N. *et al.* (2006) Biomodels database: a free, centralized database of curated, published, quantitative kinetic models of biochemical and cellular systems. *Nucleic Acids Res.*, **34** (Suppl. 1), D689–D691.
- Nurse,P. (2008) Life, logic and information. *Nature*, **425**, 424–426.
- Papin,J.A. *et al.* (2004) Hierarchical thinking in network biology: the unbiased modularization of biochemical networks. *Trends Biochem. Sci.*, **29**, 641–647.
- Saez-Rodriguez,J. *et al.* (2008) Automatic decomposition of kinetic models of signaling networks minimizing the retroactivity among modules. *Bioinformatics*, **24**, 213–219.