

# A subnet algorithm for genome scale metabolic networks

Wynand S. Verwoerd<sup>1</sup>  
verwoerw@lincoln.ac.nz

<sup>1</sup> Centre for Advanced Computational Solutions, AGLS Division, Lincoln University, Ellesmere Junction Road, P.O Box 84, Lincoln 7647, New Zealand

## Abstract

The concept of a coherent metabolic subnetwork is introduced. A computational algorithm is described to find this from the genome scale metabolic network, based on probabilistic random walks and progressive removal of nodes that link subnets and represent biochemically plausible external compounds. An interactive software implementation of the procedure is demonstrated by application to extract the flavonoid subnetwork for the model plant *Arabidopsis*.

**Keywords:** biochemical reaction network, flux balance analysis, subnet splitting, stoichiometry matrix

## 1 Introduction

The metabolic state of a homeostatic cell is specified by the fluxes through all its biochemical reactions. These reactions are interconnected in a complex network as compounds produced by one reaction is consumed by another, so that the fluxes are constrained by the stoichiometries of the reactions. Mathematically the network is a hypergraph with nodes representing compounds, and reactions are multiply connected edges. Nodes on the periphery of the network represent either inputs (e.g nutrients) or outputs (e.g. biomass) and are not subject to stoichiometry constraints. Characterising these as external compounds, they are considered to be buffered in reservoirs. By contrast, for internal nodes production and consumption need to balance to maintain homeostasis and are hence described as internal compounds.

Sophisticated methods based on matrix algebra have been developed to analyse the effects of stoichiometric constraints on the metabolic states accessible to a cell, and are commonly referred to as Flux Balance Analysis (FBA). Implementations of these methods are readily available in software packages such as CellNetAnalyzer [2].

In principle FBA is best applied to the full genome scale metabolic network, and this has been done for microorganisms such as *E. Coli* [3] with networks containing a few hundred nodes. However, the larger network for a complex organism such as *Arabidopsis* makes the results harder to interpret. Also, metabolic studies often focus on one functional aspect such as production of secondary metabolites. It is then desirable to extract a subnetwork, that retains the full biochemical context relevant to the target functionality without carrying unnecessary baggage.

The extraction is visualized as cutting its links at the nodes that connected it to the full network. Nodes, internal in the full net, thus become external in the subnet. This implies discarding their stoichiometry constraints, instead creating reservoirs. There are many specific compounds for which this can be justified biochemically. For example, the continual recycling of carrier molecules such as ATP and ADP can plausibly be represented by unlimited reservoirs. In this work a subnet for which all node cuts can be similarly justified is referred to as a coherent subnetwork.

The problem of dividing a network into segments is a familiar one in graph theory. Many algorithms have been devised to solve this “clustering” problem, for example the highly efficient and elegant Markov Clustering (MCL) [1]. However, cluster methods isolate highly connected node subsets, whereas the subnets of interest here need not be highly connected. Instead they are defined by having external nodes that are acceptable by criteria beyond just the network structure.

## 2 Method and Results

The algorithm presented here shares some of the conceptual framework of MCL [2]. In particular, it uses random walks (expressed by a probability matrix) to explore network connections by probability flow (“expansion”), but does not apply the “inflation” step crucial to MCL. Instead, the strategy is to recognize that cutting an internal node stops probability flow through it, which is implemented by removing it from the probability matrix. Successful division into subnets corresponds to producing a non-overlapping block structure in the probability matrix. Discovering blocks requires optimal reordering of matrix rows and columns. Generally this yields only imperfect blocking, but appropriate analysis allows candidate compounds that will decouple blocks when made external, to be identified. Removing these and iterating the process allows a network to be progressively divided into smaller subnets to the desired level of granularity. The main stages of the process are as follows:

1. **DAG calculation.** Matrix multiplication of separate probability matrices for compound-reaction and reaction-compound random walks is used to reduce the metabolic hypergraph to a compounds-only simple graph. Potentiating this matrix to convergence (“expansion”) gives the probabilities for a so-called DAG (Directed Acyclic Graph) that identifies some nodes as sinks and others as sources of the random walk, but with no discernible association into blocks.
2. **Group the DAG.** The best possible blocking is achieved by grouping similar rows and columns together. Rows are first compared using the Sokal-Sneath vector similarity measure and used to set up a hierarchical row clustering structure. Rows are then reordered according to the leaf structure of the corresponding dendrogram plot. The same procedure is applied independently to the columns of the DAG. The number of distinct groups derived from the hierarchical structure is controlled by setting a cutoff dissimilarity that distinguishes groups.
3. **Minimize block discrepancies.** In a perfectly blocked matrix, all non-zero elements in a row of the DAG belong to the same column group and vice versa. Departures from this, as well as discrepancies between column and row groups, indicate imperfect blocking. To quantify this, a blocking matrix is calculated such that fractions different from 0 and 1 reflect the amount of probability that has leaked out of row or column groups. The group cutoff is dynamically adjusted to minimize discrepancies.
4. **Identify new externals.** In a grayscale visualisation of the blocking matrix, areas where row and column grouping agree on block identification appear black, contradictions medium grey and leakage associated with coupling between blocks appear lighter grey. All light grey cells are collected and the minimal set of rows and/or columns to cover these cells determined by solving a suitable integer linear programming problem. The compounds associated with the minimal set are interactively presented to the user for approval as externals. Approved externals are removed from the DAG and stages 1 – 4 repeated, until blocking to the desired level of granularity has been achieved.
5. **Reconstitute subnetworks.** Each block in the final DAG contains all internal compounds of a subnet. All reactions and associated external compounds are restored to each subnet from the full network stoichiometry matrix and saved as a separate subnet stoichiometry matrix for independent use in FBA. Applied to the full 1527x1394 stoichiometry matrix for the model plant *Arabidopsis Thaliana*, Aracyc V4.0 [3] this procedure results in the separation of 33 subnets, among which the identical flavonoid subnet that was obtained by a tedious and ad hoc manual procedure in a previous study.

## References

- [1] Enright, A.J., Van Dongen, S. and Ouzounis, C.A., An efficient algorithm for large-scale detection of protein families, *Nucleic Acid Research*, 30(7):1575-1584, 2002.
- [2] Klamt, S., Stelling, J., Ginkel, M. and Gilles, E.D., Fluxanalyzer: Exploring structure, pathways and flux distributions on interactive flux maps, *Bioinformatics*, 19(2):261-269, 2003.
- [3] Rhee, S.Y., et al, The Arabidopsis Info Resource (TAIR), *Nucleic Acid Research*, 31(1):224-228, 2003
- [4] Reed, J.L., Vo, T.D., Schilling, H. and Palsson, B.O., An expanded genome-scale model of Escherichia Coli K12., *Genome Biology*, 4:R54.1-R54.12, 2003.