1
RECONSTRUCTING METABOLIC NETWORKS – A REVIEW

Choon Yee Wen
Mohd Saberi Mohamad

1.1 INTRODUCTION

Genome-scale metabolic networks reconstruction from different organisms have become popular in the recent year (Feist *et al.*, 2009). Reconstructions of the metabolic networks are found to be very useful in health, environmental and energy issues (Chandran *et al.*, 2008). The information from biochemical databases which contain thousands of metabolites and chemical reactions involved in the small molecule metabolism (SMM) can be represented as a graph. These data are representing how the molecules are converted into each other in the SMM network. In order to infer putative metabolic pathways some path finding algorithms are applied in the graph. However, the results are disappointing (Croes *et al.*, 2005).

The first section explains the model development in metabolic networks generally and methods of metabolic pathway analysis will be discussed next. After discussing how the approach works, some computational approaches will be discussed in the next section. Boolean network and Bayesian network are the most common algorithms which are often used for inferring network will be discussed.

The next section will explain about the current challenges and existing methods. Lastly, the review will be concluded with the importance of reconstructing a metabolic network.

1.2 MODEL DEVELOPMENT

There are many approaches in metabolic modeling, but the fundamental requirement for all of them is a stoichiometric matrix based on a reconstructed metabolic network (Terzer *et al.*, 2009). Stoichiometric approaches aim to find pathways in which compounds nodes satisfy a variety of different, but biochemically meaningful, stoichiometric constraints (Planes and Beasley, 2008).

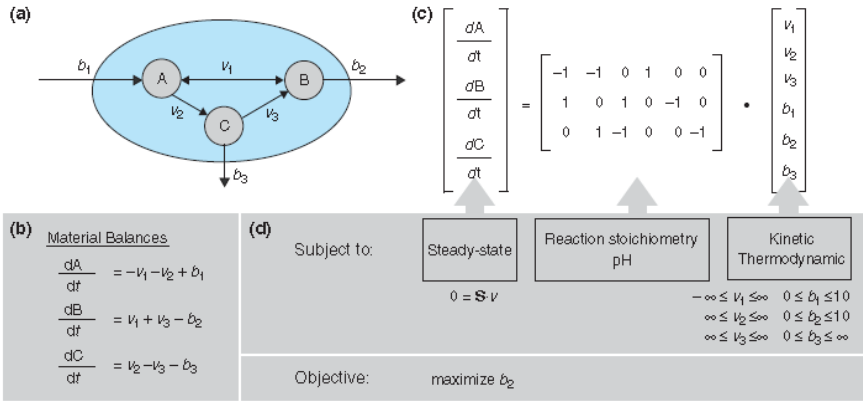

Figure 1.1 (a) A small reaction network consisting of three metabolites (A, B, and C), three transport reactions, and three enzymatic reactions is constructed. *vi* indicates the flux through reaction *i* and *bj* represents the flux through transport protein *j*. (b) Material balance equations are shown for each metabolite. (c) a stoichiometric matrix is populated according to Eq. 1. (d) Assumptions, constraints, and an objective are listed for the system (Terzer *et al*., 2009).

Figure 1.1 shows that each column of the stoichiometric matrix represents a chemical or transportation reaction, with non-zero values that identify the metabolites which participate in the reaction as well as the stoichiometric coefficients that correspond to each metabolite. Besides, the matrix also contains directionally, in the matrix substrate and product metabolites have the negative and positive coefficients respectively. The stoichiometric matrix

can also be thought of as the list of reactions in which a given metabolite participated by considering the matrix rows instead of the columns. This is very important for defining mass balances for each metabolite in the network (Terzer *et al*., 2009). Biological networks can be analyzed by using this matrix representation as it permits the use of analytical methods from linear algebra (Chandran *et al*., 2008).

The mass balances are expressed by a system of differential equations written for all metabolite concentrations **c** as follow

$$\frac{d\mathbf{c}(t)}{dt} = \mathbf{S}.\mathbf{v}(t) \qquad (1)$$

where
c = concentration
t = time
S = stoichiometric matrix
v*(t)* = vector of reaction rates

Metabolism operates much faster than regulatory or cell division events. It is reasonable to assume that where the metabolite concentrations do not change, the metabolic dynamics have reached a *quasi-* or *pseudo-steady state*. This assumption leads to the *metabolite balancing equation*

$$\mathbf{S} \cdot \mathbf{v}(t) = 0 \qquad (2)$$

where
S = stoichiometric matrix
v*(t)* = vector of reaction rates

Equation (2) is a homogenous system of linear equations. Each metabolite is consumed in the same quantity as it is produced and is required in this equation. An optimization step is needed to find the optimal **v** due to multiple solutions for this linear equation. **v** is optimized for a particular objective in flux balance analysis, such as maximizing protein or ATP production, under the constraint that $\mathbf{S.v}(t) = 0$, thus providing the steady state flux values for all the reactions in that system (Chandran *et al*., 2008).

In short, the process of building stoichiometric matrices involves gathering a collection of genomic, biochemical, and physiological data from the primary literature as well as databases. A list of chemical and transport reactions with their metabolite participants for a given cell is synthesized by using the available information. Charge of each metabolite should be checked to make sure that the chemical reaction is balanced. A reconstructed network model is only as good as it is identical to the stoichiometric matrix, the amount and quality of the experimental evidence that support the inclusion of a reaction in the matrix can be very different. Therefore, careful curation and continual updates of the matrix is crucial (Terzer *et al.*, 2009).

The stoichiometric matrix can be annotated by including further important information about either the reactions or the metabolites. The reversibility of each reaction and the cellular compartment in which each reaction occurs are the most common matrix annotations. Reaction rates measured or estimated concentration ratios or to reflect the experimental setup are bound as a consequence of kinetic constants. Reaction directions can be defined by simply setting $v_{min} = 0$ or $v_{max} = 0$ for forward or backward irreversible reactions as well as the upper and lower limits can apply to fluxes of individual reactions ($v_{min} \leq v \leq v_{max}$). More detailed information about reaction kinetics might be included in additional matrix annotation (Terzer *et al*., 2009).

1.3 METHODS OF METABOLIC PATHWAY ANALYSIS

The two approaches which can be used for metabolic pathway analysis are constraint-based and graph-theoretical path finding methods (Pitkänen *et al.*, 2009). In constraint-based methods, the pathway is inferred where the intermediate metabolites are balanced in *pseudo-steady* state. In a steady-state, the net production of each intermediate metabolite is zero. Pathways satisfying this constraint can be branching, in general consisting of one or more linear paths enabling the production of the target metabolite from sources (Pitkänen *et al.*, 2009).

In graph-theoretical methods, a number of shortest paths leading from the source to the target metabolite are discovered. These methods only deal with linear, non-branching pathways. Therefore, graph-theoretical methods are always restricted to one source to one target metabolite. Results from graph-theoretical path finding and steady-state pathway analyses are correlated. Normally, many alternative pathways tend to be generated by the graph-theoretical approaches, thus there is a need to filter and rank the pathways with some realistic criteria to produce significant results (Pitkänen *et al.*, 2009).

For constraint-based methods, minimize or maximize the flux value for each reaction. The following will be the results produced by constraint-based methods.

(a) If minimal and maximal values are zero, the reaction is a zero flux reaction means it cannot have a flux value other than zero. It can be removed if no model corrections are made, without affecting the outcome of subsequent simulations.

(b) If minimal or maximal value is zero and the reaction is reversible, we have an unsatisfied reversibility. Either the reversibility constraint is too lax or another component is

missing, disabling the operation in one direction. Tightening this constraint might lead to better simulation performance.

(c) If the minimal and maximal values are non-zero and have equal sign, the reaction is essential. Deletion of the reaction is predicted to be lethal.

For reactions not of type (c), set the bounds to zero. If biomass cannot be produced, the reaction is essential and the removal of the reaction is lethal (Terzer *et al.*, 2009).

For graph-theoretical methods, *k*-shortest path algorithm associated with the 'shortest' possible paths in the network. Studies believe that the path-finding approaches are better improved than the stoichiometric approaches for the following reasons.

(1) It is a well-known problem of finding *k*-shortest paths from a source to a target in graph theory and it is computationally manageable for genome-scale metabolic networks.

(2) The requirement for EFMs and EPs which contributes to the problem of defining which compounds are internal and which external is avoided.

(3) Instead of computing all possible paths, computing *k*-shortest paths depending on the suitable distance metric appears to be a quite logical concept since not all the computed paths are biologically significance. Furthermore, since the number of *k* is

always restricted to a very small number, analysis of the metabolic paths is simplified.

1.4 BOOLEAN NETWORKS

Boolean network is a simple deterministic model of regulatory networks (Markowetz and Spang, 2007). Kauffman (Kauffman, 1969) was the first to introduce Boolean network model. A gene expression is simplified with two levels in these models which are ON and OFF. A Boolean network G(V,*F*) is defined by a set of node V = {x_1, … , x_n} and set of Boolean functions $F = \{f_1, \ldots, f_n\}$. A Boolean function $f_i(x_1, \ldots, x_k)$, where $i = \{1, \ldots, n\}$, with k specified input nodes (indegree) is assigned to node x_i. F is the regulation nodes. At time $t - 1$ given the values of the node (V), the Boolean function are used to update these values at time t (Kim *et al*., 2007).

Kauffman (1993) has further improved the model system into Random Boolean network model. The introduction of probabilistic Boolean network by Shmulevich *et al*. (2002) has made Boolean network attracted much attention. There are many algorithms proposed for the inference of Boolean networks. Liang et al. (1998) introduced REVEAL algorithm for causal inference by using mutual information, which is the most essential and general measure of correlation. A Boolean network structure based on the consistency problem can be used to determine the consistency of an existence network with the observed data was constructed by Akutsu *et al*. (1999). The Best-Fit Extension problem (Boros *et al*., 1998) is used for the inference of probabilistic Boolean networks in the recent studies of Boolean network algorithm. Every node is given a chance to acquire different Boolean functions in probabilistic Boolean networks. Due to the probabilistic selection of Boolean functions the flexibility in the determination of the steady state of Boolean networks and

monitoring of the dynamical network behavior for gene perturbation or intervention is increased (Kim *et al.*, 2007).

Boolean network is more commonly used in gene regulatory networks. Boolean network offers several advantages in the estimation of gene regulatory networks. First, the Boolean network model is able to explain the dynamic behaviour of living system effectively. Realistic complex biological phenomena such as cellular state dynamics that exhibit switch-like behavior, stability, and hysteresis can be represented by the simplistic Boolean formalism (Huang, 1999). It also enables the modeling of non-linear relations in complex living systems (Thomas, 1991). Besides, Boolean algebra is an established science that a large set of algorithms is available for supervised learning in the binary domain, for example, logical analysis of data (Boros *et al*., 1997) and Boolean-based classification algorithms (Akutsu *et al.*, 2001). Lastly, the accuracy of classification can be improved by dichotomization to binary values and by reducing the noise level in experimental data the obtained models can be simplified (Kim *et al*., 2007).

However, the Boolean network has some drawbacks. Most of the Boolean network algorithms can only be used with a small number of genes and a low indegree value due to extremely high computing times to construct reliable network structures (Kim *et al*., 2007). In order to increase the efficiency of searching in solution space, these algorithms should be speed up through parallelization for higher indegree value (Liang *et al*., 1998). The consistency problem (Akutsu *et al*., 1999) works in time complexity as follow

$$O(2^{2^k} \cdot \binom{n}{k} \cdot m \cdot n \cdot poly(k)) \quad (3)$$

where
m = number of observed time points
n = the total number of genes
poly(k) = the time required to compare pair of examples respectively

The Best-Fit Extension problem (Boros et al., 1998) also works in time complexity in equation (3). The improved consistency algorithm and Best-Fit Extension problem still face an exponential increase in the computing time for the parameter n and k in time complexity as follow

$$O(\binom{n}{k} \cdot m \cdot n \cdot poly(k)) \quad (4)$$

where
m = number of observed time points
n = the total number of genes
poly(k) = the time required to compare pair of examples respectively

In the study of large-scale gene regulatory and gene interaction systems using Boolean networks the high computing times are a critical problem.

1.5 BAYESIAN NETWORK

A Bayesian network is a graphical representation of the dependency structure between the components of random vector **X**. The individual random variables are associated with the vertices of a directed acyclic graph (DAG) *G*, which describes the dependency

structure. Each node is described by a local probability distribution (LPD) and the joint distribution $p(\mathbf{x})$ over all nodes factors as

$$p(x) = \prod_{v \in V} p(x_v \mid x_{pa(v)}, \theta_v) \tag{5}$$

where

θ_v = parametrization of the local distribution

$x_{pa(v)}$ = vector of parent state denoting the activity levels of gene's regulators

The DAG structure involves in an ordering of the variables. The parents of each node are those variables that contribute to it independent of all other predecessors. The key property of Bayesian network is the factorization of the joint distribution. It enables the set of variables to be divided into families, which can be treated individually.

Bayesian networks read off independence statements from full conditional independence graph in directed Global Markov condition also known as d-separation (Pearl, 2000). There are three archetypical situations in d-separation which are chain, fork, and collider as shown in Figure 1.2.

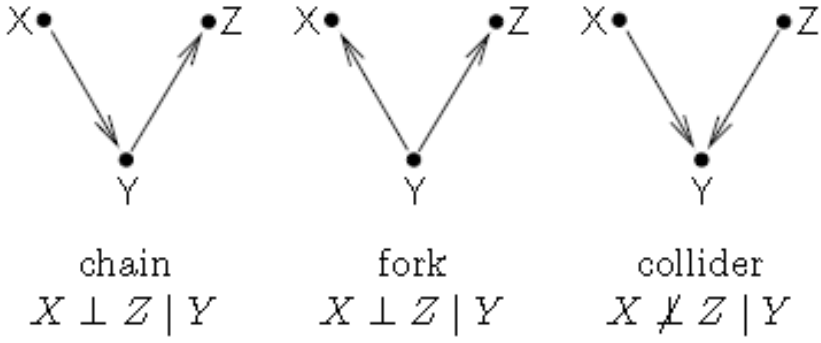

Figure 1.2 Conditional independence in directed graphs.

In a chain, X → Y → Z, in the middle node Y blocks the information flow between X and Z and thus it holds that X ⊥ Z | Y. In the fork, again it holds that X ⊥ Z | Y due to that X and Z are both regulated by Y, knowing the state of the regulator renders the regulatees conditionally independent. The last case is if X and Z are independent regulators with a common target Y, then the state of Y will provide the information about X and Z. Thus, in the collider X → Y ← Z the middle node Y "unblocks" the path between X and Z and that holds X ⊥̸ Z | Y.

Bayesian networks allow the highest resolution of correlation structure. However, they suffer from a major drawback that they are acyclic. The joint distribution cannot be decomposed with cycles, but biological networks are all known to contain feedback loops and cycles (Alberts *et al*., 2002). Gak-Viks *et al*. (2006) proposed the factor graph network model that is an extension of Bayesian network that includes cyclic structures. Another way to solve the cycle problem is by assuming that the system evolves over time, as show in Figure 1.3.

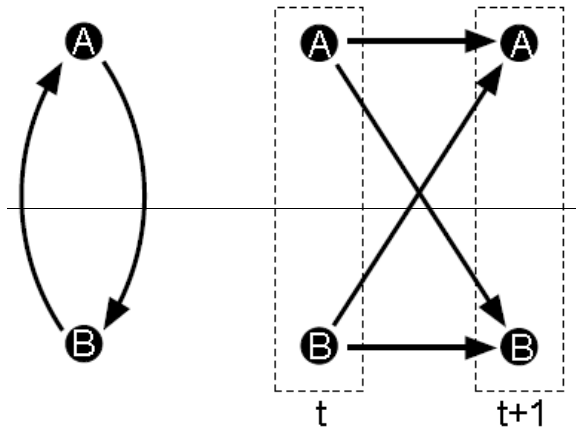

Figure 1.3 Cycles unroll over time.

With this assumption, the system is no longer model a static random vector **X,** but a time series **X**[1], …, **X**[T] by observing **X**

at T time points. Assume that $\mathbf{X}_v$ at time t+1 can only have parents at time t, and then cycles "unroll" and the resulting model is again acyclic and tractable, it is known as Dynamic Bayesian network (DBN).

Due to the complex modeling strategies which are estimating a large number of parameter in Bayesian network algorithms there is a limitation in determining an important network structure. Besides, there is another drawback of Bayesian network that is a long computation time for searching all potential network structures on genome-wide expression data (Kim *et al*., 2007).

1.6 CURRENT CHALLENGES

There are challenges in reconstructing metabolic networks. First, the developing computational approaches for fully automatic network reconstruction and reconciliation as the unknown reactions and the necessary validation of database entries are still resulted in time-intensive manual network curation.

One automated reconstruction method aims to identify fundamental reactions from an organism-wide database such that these reactions could allow growth of mutants that are experimentally viable, but predicted to be inviable by an existing stoichiometric model (Reed *et al.*, 2006). Only one experimental condition is considered at a time in this approach, it produces potentially large sets of candidate reactions, and it is computationally expensive because each condition and candidate reaction flux balance analysis (FBA) has to be performed. Optimization-based methods help in identifying gaps in metabolic network reconstructions, and the models are consolidated by introducing new reactions or by modifying the existing reaction. However, global changes in network structures or potential effects on the quality of model predictions are not taken into consideration by the existing methods (Satish *et al*., 2007). Therefore, the

available methods have limitations in automatically generating predictive network models.

Another challenge is the cellular optimality and design. The choice of a biologically meaningful objective function is crucial for FBA. It can be considered as the inverse problem of FBA in identifying the objective function or cellular design principles. Given some objective function where FBA finds an optimal flux vector, it is more challenging to infer the objective for an experimentally determined reference flux vector (Terzer *et al*., 2009).

1.7 CONCLUSION

Reconstructing metabolic networks which fall under the field of synthetic biology although in its infancy, it has great potential. With the association of well-characterized biology parts, computer-aided design tools, mathematical modeling and efficient methods for constructing or synthesizing the sequence parts, synthetic biology enables the cells to function like devices similar to the electronic or mechanical devices. This capability can resolve problems from health issues to environmental issues, cells can be designed to combat cancer, produce drugs, detect pollutants, catalyze reactions, or produce environmentally safer fuel. From here, synthetic biology can be foreseen to be an important field in the future.

Acknowledgements

I would like to thank for the supports and assistance of Dr Md Saberi Mohamad, Prof Safaai Deris and members of AIBIG and MBI. This work has been funded by Zamalah/Institutional Scholarship provided by Universiti Teknologi Malaysia and the Ministry of Higher Education of Malaysia.

References

Akutsu, T., Miyano, S. and Kuhara, S. 1999. "Identification of genetic networks from small number of gene expression patterns under the Boolean network model." *Pacific Symposium on Biocomputing*, 4:17-28.

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K. and Walter, P. 2002. "Molecular Biology of the Cell." *Garland Science*, 4th edition, New York.

Boros, E., Ibaraki, T., Makino, K. 1998. "Error-Free and Best-Fit Extension of partially defined Boolean functions." *Information and Computation*, 140:254-283.

Chandran*, D. Copeland, W.B., Sleight, S.C. and Sauro, H.M. 2008. *"Mathematical modeling and synthetic biology."* Elsevier, **5**(2):299-309.

Croes, D., Couche, F., Wodak, S.J. and van Helden, J. 2005. "Metabolic PathFinding: inferring relevant pathways in biochemical networks." *Nucleic Acids Research*, 33:326-330.

Feist, A.M., Herrgård, M.J., Thiele, I., Reed, J.L., Palsson, B.O. 2009. "Reconstruction of biochemical networks in microorganisms." *Nat. Rev. Microbiol.,* **7**2:129-143.

Gat-Viks, I., Tanay, A., Rajiman, D. and Shamir, R. 2006. *"A probabilistic methodology for integrating knowledge and experiments on biological networks."* J Comput Biol, 132:165-181

Huang, S. 1999. "Gene expression profiling, genetic networks and cellular states: An integrating concept of tumorigenesis and drug discovery." *Journal of Molecular Medicine*, 77:469-480.

Kauffman, S.A. 1969. "Metabolic stability and epigenesist in randomly constructed genetic nets." *Journal of Theoretical Biology*, 9:3273-3297.

Kauffman, S.A 1993. "The origins of Order: Self-organization and Selection in Evolution." *Oxford University Press*, New York, United States.

Kim, H., Lee, J.K. and Park, T. 2007. "Boolean networks using chi-square test for inferring large-scale gene regulatory networks." *BMC Bioinformatics*, 8:37

Liang, S., Furhman, S. and Somogyi, R. 1998. "REVEAL, A general reverse engineering algorithm for inference of genetic network architectures." *Pacific Symposium on Biocomputing*, 3:18-29.

Markowetz, F. and Spang, R. 2007. "Inferring cellular networks – a review." *BMC Bioinformatics*, **8**(Suppl 6)**:**S5

Pearl, J. 2000. "Causality: Models, Reasoning and Inference." *Cambridge University Press*, Cambridge.

Pitkänen, E., Jouhten, P., and Rousu, J. 2009. "Inferring branching pathways in genome-scale metabolic networks." *BMC Systems Biology*, 3:103

Planes, F.J. and Beasley, J.E. 2008. "A critical examination of stoichiometric and path-finding approaches to metabolic pathways." *Briefing in Bioinformatics*, 95:422-436

Reed, J.L., et al. 2006. "Systems approach to refining genome annotation." *Proc. Natl. Acad. Sci.* USA, 103:17480-17484.

Satish, K.V., Dasika, M.S. and Maranas, C.D. 2007. "Optimization based automated curation of metabolic reconstructions." *BMC Bioinformatics*, 8:212.

Shmulevich, I., Dougherty, E.R., Seungchan, K. and Zhang, W. 2002. "Probabilistic Boolean networks: A rule-based uncertainty model for gene regulatory networks." *Bioinformatics*, 18:261-274.

Terzer, M., Maynard, N.D., Covert, M.W. and Stelling, J. 2009. "Genome-scale metabolic networks." *Wiley Interdiscip Rev. Syst. Biol. Med.*, **1**(3):285-297.

Thomas, R. 1991. "Regulatory networks seen as asynchronous automata: a logical description." *Journal of Theoretical Biology*, 153:1-23.