Chapter 6

A Guide to Integrating Transcriptional Regulatory and Metabolic Networks Using PROM (Probabilistic Regulation of Metabolism)

Evangelos Simeonidis, Sriram Chandrasekaran, and Nathan D. Price

Abstract

The integration of transcriptional regulatory and metabolic networks is a crucial step in the process of predicting metabolic behaviors that emerge from either genetic or environmental changes. Here, we present a guide to PROM (*probabilistic regulation of metabolism*), an automated method for the construction and simulation of integrated metabolic and transcriptional regulatory networks that enables large-scale phenotypic predictions for a wide range of model organisms.

Key words: Systems biology, Metabolic networks, Transcriptional regulatory networks, Constraint-based modeling, Probabilistic regulation of metabolism, Microarray data

1. Introduction

In systems biology, we study the complexity of biological systems and try to comprehend and predict the way that the components of these systems interact. In the last two decades, advances in high-throughput experimental techniques and bioinformatics methodologies have produced an abundance of biological information. The successful management, integration, and utilization of these data are critical to enable systems biology approaches. One of the most fundamental processes necessary for life is metabolism, from which the cell harnesses energy from its food and builds the components necessary for growth and reproduction. Metabolism plays a central role in the functioning of an organism and is arguably the best understood cellular process. Therefore, systems biologists have taken an early interest in metabolic networks, their behavior, and their regulation.

Hal S. Alper (ed.), *Systems Metabolic Engineering: Methods and Protocols*, Methods in Molecular Biology, vol. 985,
DOI 10.1007/978-1-62703-299-5_6,

Metabolic networks display complicated structures and interactions, leading to nonlinear dynamic behaviors (1–3). The size and complexity of metabolic networks often limit our ability to test and analyze metabolism using more traditional simulation methods such as reaction kinetics, where the mechanisms or reactions and their regulation are modeled individually and in detail. Constraint-based modeling (4, 5) allows us to overcome such problems, because the only requirement is knowledge of the stoichiometry of the system in order to be able to accurately simulate the potential metabolic behavior of an organism. Over the years, a number of stoichiometry-based methodologies have been developed, with the most commonly used being flux balance analysis (FBA) (6). FBA identifies the optimal flux pattern of a network that would allow the system to achieve a particular objective, typically the maximization of biomass production.

FBA is a powerful method for predicting system behavior, but one of its drawbacks is that it ignores the often important effect of regulation. Metabolic networks are tightly controlled, in part, by intricate transcriptional networks – further increasing the complexity of the system. Being able to model this transcriptional regulation allows us to interpret the effect of mutations and environmental perturbations on functional metabolism, which in turn opens up the possibility of diagnosing metabolic disorders and identifying new drug targets.

In this chapter, we give an overview of PROM (probabilistic regulation of metabolism) (7), a method that utilizes probabilities to denote gene states and interactions between genes and transcription factors in order to enable straightforward integration of transcriptional and metabolic networks for modeling purposes. In the past, there have been relatively few efforts focused on the integration of metabolic and transcriptional regulatory networks (8, 9). PROM has shown improved results compared to previous approaches to integrate metabolism and transcriptional regulation such as regulatory FBA (RFBA) (8, 10) in studies published so far. Another benefit of PROM is that it estimates regulatory strengths automatically from high-throughput data, as opposed to the laborious manual process RFBA models are based on. Because PROM networks can be learned from high-throughput data, these models can be comprehensive, in contrast to the manually curated approaches that require extensive literature surveys. In addition, RFBA relies on Boolean logic, which has the drawback of only allowing two states for the regulated reactions: either fully active or completely inactive. PROM introduces probabilistic, soft constraints that can be automatically quantified from microarray data, thereby overcoming the limitations of RFBA (7).

2. Analysis Tools

2.1. Flux Balance Analysis

Mathematically, FBA is framed as a linear programming problem:

FBA Formulation
Maximize

$$Z = c_j v_j \tag{1}$$

subject to

$$\sum_j S_{ij} \cdot v_j = 0 \quad \forall i \tag{2}$$

$$v_j^L \le v_j \le v_j^U \quad \forall j, \tag{3}$$

where i indexes the set of metabolites, j indexes the set of reactions in the network, S_{ij} is the stoichiometric matrix, c_j is a vector that specifies which flux is being optimized (typically this is used for the maximization of growth), and v_j is the flux of reaction j. The objective function (1) is maximized over all possible steady-state fluxes satisfying certain stoichiometric constraints (2). In genome-scale metabolic models of microbial systems, a biomass-producing reaction is usually defined and used as the objective function. Upper and lower bounds are placed on the individual fluxes (v^U and v^L, respectively) (3). For irreversible reactions, $v^L = 0$. Specific bounds, based on enzyme capacity measurements or thermodynamic considerations, can be imposed on reactions; in the absence of any information, these rates are generally left unconstrained, i.e., $v^U = \infty$ and $v^L = -\infty$ for reversible reactions. To avoid unbounded solutions, i.e., Z reaching infinity, one rate (the input flux; typically the influx of glucose) needs to be fixed to a specific value, and all fluxes should be viewed as relative to the input flux.

2.2. Flux Variability Analysis

Flux variability analysis (FVA) (11) is used to determine the range of allowable fluxes in the optimal solutions of a constraint-based analysis problem. Using FVA, we can determine the minimum and maximum possible flux through a reaction for a given optimal growth rate. After solving the FBA formulation above and identifying the optimal growth rate v_g^*, the following algorithm is used to determine the variability of fluxes in the network:

FVA Algorithm

For $r = 1$ to R

Minimize

$$Z = v_r \tag{4}$$

subject to

$$v_{growth} \ge v_g^* \tag{5}$$

$$\sum_j S_{ij} \cdot v_j = 0 \quad \forall i \tag{6}$$

$$v_j^L \leq v_j \leq v_j^U \quad \forall j \tag{7}$$

then
maximize

$$Z = v_r \tag{8}$$

subject to (5), (6), (7)
end,

where R is the total number of reactions j in the reconstructed network.

2.3. Kolmogorov–Smirnov Test

The Kolmogorov–Smirnov test (12) is a nonparametric test for the equality of continuous, one-dimensional probability distributions that can be used to compare a sample with a reference probability distribution or to compare two separate samples. The Kolmogorov–Smirnov test is used to check how much two of our expression profiles (described in the methods section) differ when compared to each other. The null hypothesis is that the two datasets are from the same distribution, whereas the alternative hypothesis is that they are from different continuous distributions. The Kolmogorov–Smirnov test has the advantage of making no assumption about the distribution of data. The method is used to select only those pairs of transcription factors and targets for which the target's expression changes significantly with respect to the transcription factor state.

3. Methods

In order to build an integrated model of a metabolic and transcriptional regulatory network for an organism (see Note 1 and Fig. 1), the following components are needed:

1. *The genome-scale reconstruction of the metabolic network of the organism* (13). The creation of metabolic reconstructions is often a laborious, painstaking process. Researchers either manually collect the necessary stoichiometric information from the literature, or the network is downloaded from organism-specific databases when available, with subsequent annotation and improvement of the data to make the model functional and in agreement with experimental data. Over the last 10 years, the metabolic network reconstructions of several organisms have been published and are publicly accessible. The simulation of the metabolic network within the PROM method is performed

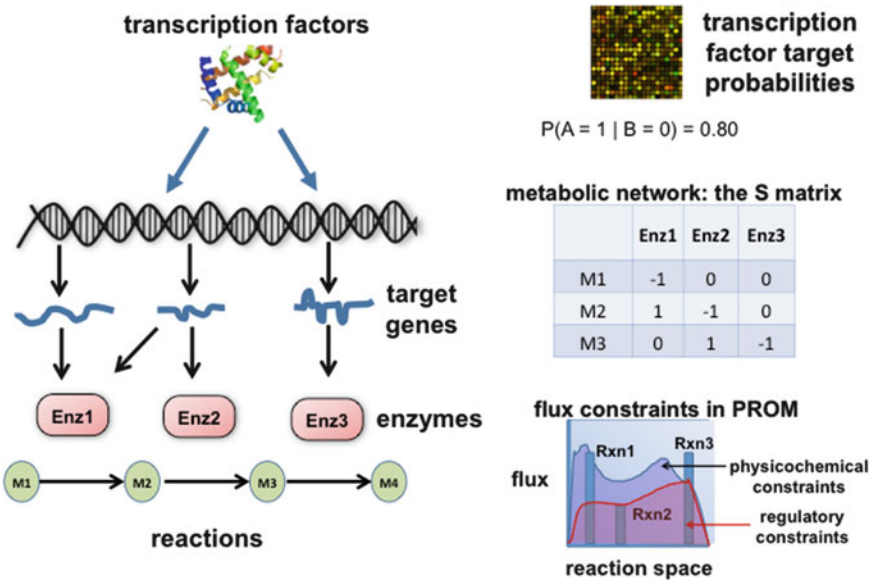

Fig. 1. The PROM method.

using FBA subject to additional constraints and a penalty function (described below) (Subheading 2.1).

2. *A regulatory network structure, which consists of a list of transcription factors, the targets of these transcription factors, and their interactions* (14). These transcriptional regulatory networks have generally been constructed based on high-throughput protein–DNA interaction data and/or statistical inference of functional relationships from genomic and transcriptomic data (15–19).
3. *A collection of gene expression data measured under different conditions, which will allow the observation of various phenotypes for the organism under study.* Ideally, the microarray data that are chosen represent a diverse number of conditions under which gene expression has been measured (see Note 2).

For the purpose of constructing the integrated metabolic– regulatory network, it is important that the PROM method takes advantage of the abundance of high-throughput data that is currently available for most organisms. From these data, the transcriptional regulatory network of the organism can be quantified statistically, similar to the probabilistic Boolean networks of Shmulevich et al. (20). From the gene expression data that are available for the organism in question, only those that involve the expression of metabolic genes are retained (see Note 3). The data are then normalized and screened for false positives using the Kolmogorov–Smirnov statistic (Subheading 2.3), and only significant interactions, defined by $P < 0.05$, are kept in the model.

PROM introduces probabilities to represent gene states and interactions between a gene and a transcription factor. For example, the probability of gene A being active when the regulating transcription factor B is not active is represented by $P(A = 1|B = 0)$, whereas the probability of both gene and transcription factor being active by $P(A = 1|B = 1)$ (see Note 3). Information from the microarray data is then used to assign values to the relationship between transcription factor and target gene. To determine the relationship between a transcription factor and its target, we first binarize the data to represent either an ON or OFF state for all the genes. We can then model the relationship between the transcription factor and the target based on the following formula:

$$P(A = 1|B = 0) = \frac{N(A = 1|B = 0)}{N(B = 0)} \tag{9}$$

where N is the number of times the event is observed. For example, if in 80% of the samples we find the gene to be on when the transcription factor is off, then the probability $P(A = 1|B = 0) = 0.8$. For transcription factors that affect more than one gene, we need to calculate this relationship for all its target genes. The fluxes of the reactions controlled by these genes can then be constrained using this information. In our example, the flux through the reaction regulated by gene A when its corresponding regulator B is turned off would be

$$P \cdot v_A^L \leq v_A \leq P \cdot v_A^U \tag{10}$$

Generally, the bounds on the fluxes of reactions in our network are redefined from having an upper bound of v_j^U to an upper bound of $P \cdot v_j^U$, where P is the probability of the gene being active under the specific phenotype. For reversible reactions, the same applies to lower bounds v_j^L (see Note 5).

Estimates for reaction bounds v_j^L, v_j^U are obtained by running the FVA algorithm (Subheading 2.2) on the unregulated metabolic model or by utilizing literature or other kinds of prior knowledge.

Unlike thermodynamic or environmental constraints that cannot be violated, we want regulatory constraints to be soft constraints to compensate for the lower confidence level and inherent uncertainty that comes from the experimentation techniques, our (lack of) understanding of the gene–transcription factor interactions and noise in the measurements. The algorithm needs to be able to exceed regulatory constraints to maximize growth if necessary but with a penalty to avoid this happening regularly. Following this procedure, we arrive at the following final formulation for the PROM model, which satisfies most or all of the regulatory constraints:

PROM Formulation
Maximize

$$Z = c_j v_j + \sum_j (\kappa_j \cdot \alpha_j + \kappa_j \cdot \beta_j) \tag{11}$$

subject to

$$\sum_j S_{ij} \cdot v_j = 0 \quad \forall i \tag{12}$$

$$P \cdot v_j^L - \alpha_j \leq v_j \leq P \cdot v_j^U + \beta_j \quad \forall j \tag{13}$$

$$\alpha_j, \beta_j \geq 0 \quad \forall j, \tag{14}$$

where $P \cdot v_j^L$ and $P \cdot v_j^U$ are the transcriptional regulation bounds, α_j and β_j are positive variables that allow deviation from those bounds, and κ_j is the cost for such deviations. The term $(\kappa_j \cdot \alpha_j + \kappa_j \cdot \beta_j)$ represents the penalty for exceeding an upper or a lower bound. The higher the value of κ, the greater the constraint on the system based on transcriptional regulation. For values of κ significantly greater than 1, the regulatory constraints become hard, and for values less than 0.1, they become insignificant. Typically, a κ value of 1 is chosen for all simulations as it represents a trade-off between the two extremes.

It is clear from the above that with PROM, gene states can take values other than just 0 and 1 due to the use of probabilities, which allows us to distinguish between strong and weak regulators. Another benefit of PROM is that we can incorporate interactions for which we have strong evidence from the literature or experimentation. If we have an example of an interaction with high-confidence proof from the literature, then the user can assign a probability of 0 or 1 for that particular interaction, setting the corresponding gene to either fully active or completely inactive. The probabilities for the rest of the interactions can then be determined based on the microarray data, following the method described here. Additional interactions involving enzyme regulation by metabolites and proteins can also be modeled and included in the PROM algorithm. If no probability can be inferred, or no data are available for a specific interaction, we set the corresponding probability to $P = 1$.

By applying the algorithm as presented above, we can then run the resulting model in order to predict the effect of knockouts of transcription factors on the metabolic fluxes. The lethality of transcription factor knockouts is predicted in a similar way to that of Shlomi et al. (21); if the respective prediction of the mutated organism's maximal growth rate is less than 5% of the wild-type growth rate, it is considered lethal, whereas knockouts that display

a growth rate lower than the wild-type growth rate are considered suboptimal.

The PROM algorithm is available for download at the following address (see Note 6): http://price.systemsbiology.net/downloads_tmp.php

4. Notes

1. Integrated metabolic–regulatory methods can be built with the PROM method for any organism for which reconstructed metabolic network models; regulatory interaction data and a sufficient amount of microarray experiment data are available.
2. The purpose of using microarrays in PROM is to quantify the relationship between transcription factors and target genes. This can be done only if we study their relationship under as many conditions as possible. If we were to use microarrays from a single condition only, we would not be able to see any change in the expression of transcription factors and target genes, and therefore, we could not learn or quantify their relationship.
3. PROM predicts phenotypes based on a gene's effect on metabolism; it cannot determine correctly the phenotypes of genes with major nonmetabolic functions.
4. When using microarray data to estimate the necessary probabilities, gene expression values under a predefined low threshold are considered inactive, and the remaining values are considered active. PROM uses the 33rd percentile as a default threshold to determine gene activity or inactivity. Generally, thresholds from 0.2 to 0.4 have provided comparably accurate predictions for the systems we have tested. The PROM algorithm, as implemented in the downloadable code, produces a warning to the user if the threshold used is not sufficient to estimate the probabilities.
5. For cases in which the probability of interaction cannot be estimated by using microarray data because of unavailability of expression data for that specific target gene or transcription factor, or if the gene was active or inactive under all conditions, we usually set the probability to a default value of 1. A value of 1 implies that the flux bounds for the reaction are not adjusted, but remain the same as in an unregulated model, whereas if the probability is set to 0, the reaction is considered inactive.
6. Pointers for running the code:
   (a) While running the PROM code, ensure that you have expression data for all the genes and regulators in the interaction data. If a relatively small fraction of the data

has missing values, PROM can impute these missing values using the k-nearest neighbors algorithm; however, it cannot handle data for genes with no expression data.

(b) As PROM predicts phenotypes based on a gene's effect on metabolism, the regulatory interactions must be between regulators and target genes that are part of the metabolic model.

(c) The names or identifiers used in the regulatory interaction data must match the names or IDs used for gene expression data and the names of the genes in the metabolic model.

(d) By default, PROM computes the predicted knockout growth rates for all the transcription factors in the network, and the growth rate is outputted in alphabetical order of the transcription factors.

Acknowledgments

We acknowledge funding from the Grand Duchy of Luxembourg for ES and NDP, a NIH Howard Temin Pathway to Independence Award in Cancer Research, an NSF CAREER grant, and the Camille Dreyfus Teacher-Scholar Program for NDP, and a Howard Hughes Medical Institute Predoctoral Fellowship for SC.

References

1. Lazebnik Y (2002) Can a biologist fix a radio? Or, what I learned while studying apoptosis. Cancer Cell 2(3):179–182
2. Mendes P, Kell D (1998) Non-linear optimization of biochemical pathways: applications to metabolic engineering and parameter estimation. Bioinformatics 14(10):869–883. doi:10.1093/bioinformatics/14.10.869
3. Szallasi Z, Stelling J, Periwal V (2006) System modeling in cellular biology: from concepts to nuts and bolts, 1st edn. The MIT Press, Boston
4. Covert MW, Famili I, Palsson BO (2003) Identifying constraints that govern cell behavior: a key to converting conceptual to computational models in biology? Biotechnol Bioeng 84 (7):763–772. doi:Doi 10.1002/Bit.10849
5. Price ND, Reed JL, Palsson BO (2004) Genome-scale models of microbial cells: evaluating the consequences of constraints. Nat Rev Microbiol 2(11):886–897. doi:Doi 10.1038/Nrmicro1023
6. Kauffman KJ, Prakash P, Edwards JS (2003) Advances in flux balance analysis. Curr Opin Biotechnol 14(5):491–496. doi:DOI 10.1016/j.copbio.2003.08.001
7. Chandrasekaran S, Price ND (2010) Probabilistic integrative modeling of genome-scale metabolic and regulatory networks in *Escherichia coli* and *Mycobacterium tuberculosis*. Proc Natl Acad Sci U S A 107(41):17845–17850. doi:DOI 10.1073/pnas.1005139107
8. Covert MW, Knight EM, Reed JL, Herrgard MJ, Palsson BO (2004) Integrating high-throughput and computational data elucidates bacterial networks. Nature 429(6987):92–96
9. Herrgard MJ, Lee BS, Portnoy V, Palsson BO (2006) Integrated analysis of regulatory and metabolic networks reveals novel regulatory mechanisms in *Saccharomyces cerevisiae*. Genome Res 16(5):627–635. doi:Doi 10.1101/Gr.4083206

10. Covert MW, Schilling CH, Palsson B (2001) Regulation of gene expression in flux balance models of metabolism. J Theor Biol 213 (1):73–88. doi:DOI 10.1006/jtbi.2001.2405
11. Mahadevan R, Schilling CH (2003) The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. Metab Eng 5(4):264–276. doi:DOI 10.1016/j.ymben.2003.09.002
12. Young IT (1977) Proof without prejudice—Use of Kolmogorov-Smirnov test for analysis of histograms from flow systems and other sources. J Histochem Cytochem 25 (7):935–941
13. Feist AM, Herrgard MJ, Thiele I, Reed JL, Palsson BO (2009) Reconstruction of biochemical networks in microorganisms. Nat Rev Microbiol 7(2):129–143. doi:Doi 10.1038/Nrmicro1949
14. Bansal M, Belcastro V, Ambesi-Impiombato A, di Bernardo D (2007) How to infer gene networks from expression profiles. Mol Syst Biol 3:122. doi:Artn 78. doi: Doi 10.1038/Msb4100120
15. Davidson EH, Rast JP, Oliveri P, Ransick A, Calestani C, Yuh CH, Minokawa T, Amore G, Hinman V, Arenas-Mena C, Otim O, Brown CT, Livi CB, Lee PY, Revilla R, Rust AG, Pan ZJ, Schilstra MJ, Clarke PJC, Arnone MI, Rowen L, Cameron RA, McClay DR, Hood L, Bolouri H (2002) A genomic regulatory network for development. Science 295 (5560):1669–1678
16. Schlitt T, Brazma A (2007) Current approaches to gene regulatory network modelling. BMC Bioinformatics 8. doi:Artn S9. doi: Doi 10.1186/1471-2105-8-S6-S9
17. Shen-Orr SS, Milo R, Mangan S, Alon U (2002) Network motifs in the transcriptional regulation network of *Escherichia coli*. Nat Genet 31(1):64–68. doi:Doi 10.1038/Ng881
18. Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, Gerber GK, Hannett NM, Harbison CT, Thompson CM, Simon I, Zeitlinger J, Jennings EG, Murray HL, Gordon DB, Ren B, Wyrick JJ, Tagne JB, Volkert TL, Fraenkel E, Gifford DK, Young RA (2002) Transcriptional regulatory networks in *Saccharomyces cerevisiae*. Science 298(5594):799–804
19. Chandrasekaran S, Ament SA, Eddy JA, Rodriguez-Zas SL, Schatz BR, Price ND, Robinson GE (2011) Behavior-specific changes in transcriptional modules lead to distinct and predictable neurogenomic states. Proc Natl Acad Sci U S A 108(44):18020–18025. doi: DOI 10.1073/pnas.1114093108
20. Shmulevich I, Dougherty ER, Kim S, Zhang W (2002) Probabilistic Boolean networks: a rule-based uncertainty model for gene regulatory networks. Bioinformatics 18(2):261–274
21. Shlomi T, Cabili MN, Herrgard MJ, Palsson BO, Ruppin E (2008) Network-based prediction of human tissue-specific metabolism. Nat Biotechnol 26(9):1003–1010. doi:Doi 10.1038/Nbt.1487