Metabolic Engineering 12 (2010) 196–211

Contents lists available at ScienceDirect

Metabolic Engineering

journal homepage: www.elsevier.com/locate/ymben

Prediction of kinetic parameters from DNA-binding site sequences for modeling global transcription dynamics in *Escherichia coli*

Timo Hardiman [1], Hannes Meinhold, Johannes Hofmann, Jennifer C. Ewald [2], Martin Siemann-Herzberg, Matthias Reuss *

Institute of Biochemical Engineering, University of Stuttgart, Allmandring 31, 70569 Stuttgart, Germany

ARTICLE INFO

Article history:
Received 6 July 2009
Received in revised form
13 September 2009
Accepted 26 October 2009
Available online 4 November 2009

Keywords:
Cell volume
Central carbon metabolism
Cra
Fed-batch
Fructose 1,6-bis(phosphate)
fruR
Glucose-limitation
Gene regulatory network
Growth rate-dependent regulation
Position weight matrix
Promoter
Regulation
RNA polymerase
Transcription rate

ABSTRACT

The majority of dynamic gene regulatory network (GRN) models are comprised of only a few genes and do not take multiple transcription regulation into account. The models are conceived in this way in order to minimize the number of kinetic parameters. In this paper, we propose a new approach for predicting kinetic parameters from DNA-binding site sequences by correlating the protein–DNA-binding affinities with nucleotide sequence conservation. We present the dynamic modeling of the *cra* modulon transcription in *Escherichia coli* during glucose-limited fed-batch cultivation. The concentration of the Cra regulator protein inhibitor, fructose 1,6-bis(phosphate), decreases sharply, eventually leading to the repression of transcription. Total RNA concentration data indicate a strong regulation of transcription through the availability of RNA polymerase. A critical assessment of the results of the model simulations supports this finding. This new approach for the prediction of transcription dynamics may improve the metabolic engineering of gene regulation processes.

1. Introduction

The workflow of the dynamic modeling of genetic regulatory networks (GRNs) taking into account network interactions between DNA, RNA, proteins and small molecules reveals similar problems to those observed in the dynamic modeling of metabolic networks. In both cases, the initial step involves the overall network model architecture, since the qualitative aspects of network performance depend on network topology. Hecker et al. (2009) have reviewed the most important models for inferring the appropriate network architecture, including information theory models, Boolean and Bayesian networks (such as the approach for reconstructing large-scale Boolean networks proposed by Mehra et al., 2004), as well as differential and difference equations, to name just a few. As far as *E. coli* is concerned, the topology of many regulatory networks was previously studied on the basis of fundamental molecular biology studies which led to hierarchical concepts (Lengeler et al., 1999). The significance of the modular structure of the key regulation processes of the central carbon metabolism enzymes under glucose limitation has been confirmed by the results obtained from complementary microarray and metabolic flux analyzes during fed-batch cultivations with constant feeding rates (Hardiman et al., 2007; Lemuth et al., 2008).

The next step towards improving the understanding of network performance is dynamic models that include quantitative details of molecular interactions, e.g., reaction kinetics. As is the case for the dynamic modeling of metabolic networks, the difference can also be seen here between top-down computational models based on the canonical representations of kinetic properties and bottom-up approaches, which integrate mechanistic biological information about the individual molecular interactions. In other words, systems behavior is modeled by combining the individual reactions. For some metabolic networks, the two complementary approaches have been illustrated by Reuss et al. (2007).

* Corresponding author. Fax: +49 711 68565164.
E-mail address: reuss@ibvt.uni-stuttgart.de (M. Reuss).
[1] Current address: Development Anti-Infectives, Sandoz GmbH, Biochemiestrasse 10, A-6250 Kundl/Tirol, Austria
[2] Current address: Institute of Molecular Systems Biology, ETH Zurich, CH-8093 Zurich, Switzerland

1096-7176/$ - see front matter © 2009 Published by Elsevier Inc.
doi:10.1016/j.ymben.2009.10.006

Nomenclature

c_j	intracellular concentration of compound j [mol$(l_{cytosol})^{-1}$]
$c_{j,t}$	total intracellular concentration of compound j [mol$(l_{cytosol})^{-1}$]
X_j	specific intracellular concentration of compound j [mol(g dry weight)$^{-1}$]
ρ_j	intracellular mass concentration of compound j [g(lcytosol)$^{-1}$]
N_j	number of compound j [dimensionless]
r_i	rate of intracellular reaction i [mol $(l_{cytosol}\ s)^{-1}$]
V_{cell}	cell volume [$l_{cytosol}$]
v_X	specific cell volume [$l_{cytosol}$ (g dry weight)$^{-1}$]
k_j	individual rate constant for compound j [s^{-1}]
μ	specific growth rate [h^{-1}]
ω	fractional change of the specific cell volume [h^{-1}]
ϕ_k	probability of binding to DNA-binding site k [dimensionless]
Φ_k	probability of transcription with respect to binding of a protein to the DNA-binding site k [dimensionless]
α_k	spacer penalty k [dimensionless]
δ_k	enhancement factor k [dimensionless]
η_j	efficiency of transcription from promoter j [dimensionless]
a_j,b_j	parameters for linear approximation functions
N_A	Avogadro constant
K_k	equilibrium constant of reaction k, [M^{-1}] or [M^{-4}]
C_i^j	control coefficient for species i in response to changes in j [dimensionless]
$score_k$	specificity score for the nucleotide sequence k [dimensionless]
SpM	specificity matrix with the entries $a_{m,n}$ [dimensionless]

Due to the extensive use of differential and difference equations for inferring the network architecture, the two complementary approaches are more difficult to tell apart in the case of GRNs. The models used in these dynamic representations are either linear (Chen et al., 1999) or implement non-linearity through polynomial functions (Sakamoto and Iba, 2001), S-systems models (Vilela et al., 2008; Voit, 2008), generic sigmoidal functions (Haixin et al., 2007; Weaver et al., 1999) and neural network frameworks (Vohradsky, 2001a, 2001b). There is an inherent conflict between model quality and complexity, which affects the reliability of parameter estimation in higher-connected networks. Thus, robust identification of model structure and parameters will frequently encounter problems in terms of discriminating between different models (cf. also Hadlich et al., 2009). Furthermore, the estimated parameters cannot be interpreted in a physically and biologically meaningful way using such approaches.

In contrast, bottom-up models based on well understood molecular interactions enable the integration of mechanistically meaningful parameters. The use of such deterministic models for describing the regulation of gene expression has a long tradition. The majority of these models focus on the *lac* operon and the phenomena of diauxic growth of *E. coli* (Harder and Roels, 1982; Kremling et al., 2001; Kremling and Gilles, 2001; Lee and Bailey, 1984a, 1984b; Roels, 1978; Schmid et al., 2004; Sevilla et al., 2007; Wong et al., 1997) (Table S1 of the supplementary material). The mathematics of all these models is similar and rests largely upon the concept suggested by Yagil and Yagil (1971). Applying the operon model of Jacob and Monod (1961), Yagil and Yagil estimated the probability of transcription initiation of a gene that is regulated by a repressor or activator protein by assuming equilibrium reactions of the regulator protein (R) and the effector molecule (E) or the DNA binding site ($DNAbs$):

$$R+E \overset{k_1}{\rightleftharpoons} RE, \tag{1}$$

$$R+DNAbs \overset{K_2}{\rightleftharpoons} R.DNAbs, \tag{2}$$

$$\varphi = \frac{c_{R.DNAbs}}{c_{DNAbs,t}} = \frac{K_2\, c_R(K_1, c_E, c_{R,t})}{1+K_2\, c_R(K_1, c_E, c_{R,t})}. \tag{3}$$

ϕ is the fraction of DNAs that are bound by R and that can be used for calculating the probability of transcription (Yagil, 1975; Yagil and Yagil, 1971). c is the concentration of the particular model component, where c_R describes the free regulator protein, and $c_{R,t}$ the total regulator protein. The major progress that resulted from Yagil and Yagil's approach (1971) is that no more than two equilibrium constants (K_1,K_2) were needed for the mechanistic modeling of gene regulation. K_2 is the binding constant of the protein–DNA reaction. However, microbial transcription units usually include many DNA-binding sites and multiple promoters. For example, the *lac* operon has 11 DNA-binding sites, 5 of which are targeted by the Crp activator protein or the LacI repressor protein, and 5 promoters. When applying the concept of Yagil and Yagil (1971) to complex GRNs, the high number of individual DNA-binding constants (K_2) will again restrict detailed mechanistic modeling. It is therefore not surprising that in previous studies, many of the regulatory interactions were not taken into account (Table S1, supplementary material). Under defined environmental conditions, the pooling (lumping together) of multiple promoters and regulatory sites seems reasonable in order to reduce the number of individual DNA-binding constants (K_2) (Table S1). In contrast, for dynamic GRNs that aim at disentangling all potential regulatory interactions at the same time as achieving a quantitative understanding, the high number of individual DNA-binding constants (K_2) may become a critical limitation. Therefore, an approach is needed to reduce the number of parameters using *a priori* information (Hecker et al., 2009). One possibility is to estimate the relevant DNA-binding constants directly from data of surface plasmon resonance experiments (Majka and Speck, 2007), for example.

The prediction of parameters from conserved DNA-binding site sequences involves an entirely different and straightforward approach. The approach of Mulligan et al. (1984), Schneider et al. (1986), Berg and von Hippel (1987) (reviewed by Stormo, 1990) was to use the frequency of each nucleotide at each position of the DNA-binding sites for calculating binding constants. The frequency data are embedded in a position weight matrix, PWM (also referred to as position-specific scoring matrix, PSSM) (Wasserman and Sandelin, 2004). However, this approach has never been applied to the dynamic modeling of GRNs.

For fed-batch experiments at constant feeding rates, which is the focus of the present paper, an additional factor is of particular importance – growth rate-dependent regulation. In previous models, growth rate-dependent regulation was usually addressed by simple linear approximation functions, under the assumption that (i) the transcription rate is directly proportional to the specific growth rate (μ) and (ii) that in kinetic terms this regulation is the same for all genes (Gondo et al., 1978; Harder