FEBS Letters 583 (2009) 3923–3930

FEBS Letters

journal homepage: www.FEBSLetters.org

Review

Systems analysis of cellular networks under uncertainty

Hans-Michael Kaltenbach, Sotiris Dimopoulos, Jörg Stelling *

Department of Biosystems Science and Engineering, ETH Zurich, 4058 Basel, Switzerland
Swiss Institute of Bioinformatics, ETH Zurich, 4058 Basel, Switzerland

ARTICLE INFO

Article history:
Received 15 September 2009
Revised 26 October 2009
Accepted 27 October 2009
Available online 29 October 2009

Edited by Johan Elf

Keywords:
Dynamic model
Uncertainty
Structural network analysis
Chemical reaction network theory
Monotone system
Ensemble modeling

ABSTRACT

Besides the often-quoted complexity of cellular networks, the prevalence of uncertainties about components, interactions, and their quantitative features provides a largely underestimated hallmark of current systems biology. This uncertainty impedes the development of mechanistic mathematical models to achieve a true systems-level understanding. However, there is increasing evidence that theoretical approaches from diverse scientific domains can extract relevant biological knowledge efficiently, even from poorly characterized biological systems. As a common denominator, the methods focus on structural, rather than more detailed, kinetic network properties. A deeper understanding, better scaling, and the ability to combine the approaches pose formidable challenges for future theory developments.

1. Introduction

It is a key concept of systems biology to iteratively combine large-scale experimental analysis with mathematical modeling in order to eventually elucidate how biological systems operate [1]. This concept, per se, is not new – similar approaches have been followed for decades in physiology and theoretical biology [2]. Unprecedented developments of high-throughput, large-scale experimental technologies – such as genomics, proteomics, metabolomics – however, have opened realistic opportunities for system-wide analyses in biology, at least at the cellular level. Mathematical modeling can employ this data to generate qualitative or quantitative predictions and, thereby, evaluate biological hypotheses, suggest new experiments for validation, and ultimately increase our systems-level understanding [3]. Transcriptional regulatory networks provide one example where this combination of experimentation and modeling has been particularly successful [4].

The complexity of cellular systems constitutes an obvious challenge for mathematical modeling in systems biology. For instance, it is unclear how detailed dynamic models of small-scale systems could eventually be scaled to entire cells, or joined with coarser-grained but large-scale models [4]. Uncertainty is another important and less appreciated factor that requires new theory development to increase the power of mathematical models as systems analysis tools. Probabilistic models cope with uncertainty by design, but they often do not respect first principles such as mass conservation, do not cover dynamic processes, and yield only limited insight into mechanistic detail. Hence, it is important to consider the effect of uncertainty on mechanistic models.

Uncertainties fall into two broad categories [5]. Aleatoric uncertainty stems from the inherent randomness in the behavior of the system under examination. In biology, for example, noise in gene expression induces uncertainty in the model output. Since the noise stems from physical principles, this uncertainty cannot be avoided and needs to be addressed by stochastic analysis. This highly active research area is summarized in recent reviews [6–8]. Here, we focus on the second type of uncertainty, epistemic uncertainty, which results from our lack of knowledge on the system. In current mechanistic mathematical models of biological systems, epistemic uncertainty is profoundly present due to practical limitations, such as lack of understanding of the underlying mechanisms, incomplete coverage and measurement errors in various modeling quantities and, most commonly, parameters derived from noisy or incoherent data sets. A recent study of coverage in shotgun proteomics illustrates such technical limitations quantitatively [9] for one type of data required for mathematical modeling.

For future perspectives, consider the example of budding yeast as a best-case scenario. *S. cerevisiae* is arguably the most intensely studied eukaryotic model organism of 'manageable' complexity.

* Corresponding author. Address: Department of Biosystems Science and Engineering, ETH Zurich, 4058 Basel, Switzerland. Fax: +41 61 38 73994.
E-mail address: joerg.stelling@bsse.ethz.ch (J. Stelling).

0014-5793/$36.00 © 2009 Federation of European Biochemical Societies. Published by Elsevier B.V. All rights reserved.
doi:10.1016/j.febslet.2009.10.074

Yet, ten years after sequencing the genome in 1996, roughly 1/6 of the organism's genes still remained un-annotated [10], and only very recently, it was possible to establish growth phenotypes for all individual genes using chemical genomics approaches [11]. Hence, ever accumulating perfect quantitative knowledge on biological systems for epistemic uncertainty to cease currently seems unrealistic.

Incomplete knowledge affects mathematical models based on first principles in different ways. Here, we consider ordinary differential equation (ODE) models derived from mass balances of n individual components in a biochemical network with r reactions. The general form of such an ODE system is:

$$\frac{d\mathbf{c}(t)}{dt} = \mathbf{N}\cdot\mathbf{v}(\mathbf{c}(t),\mathbf{u}(t),\mathbf{k}), \quad \mathbf{c}(t_0)=\mathbf{c}_0 \tag{1.1}$$

with the $n\times 1$ vector of time-dependent concentrations $\mathbf{c}(t)$, the $n\times r$ stoichiometric matrix $\mathbf{N}$, and the $r\times 1$ vector function of reaction rates – or fluxes – $\mathbf{v}(\cdot)$. The fluxes depend on the system state $\mathbf{c}(t)$, on potentially time-varying inputs $\mathbf{u}(t)$, and on kinetic parameters $\mathbf{k}$ such as affinity constants. Finally, $\mathbf{c}_0$ denotes the initial state of the system, for instance, absolute protein concentrations. Uncertainty enters in the form of unknown or poorly estimated parameter values $\mathbf{k}$ and initial conditions $\mathbf{c}_0$. Moreover, missing or incorrect reactions affect the model structures, namely stoichiometries $\mathbf{N}$ and reaction rate laws $\mathbf{v}(\cdot)$. The resulting nested model uncertainties are difficult to handle [12].

To cope with the combination of complexity and uncertainty in biological systems, it is important to realize that mathematical models need to be adapted to the phenomena of interest as well as to the scientific questions they are intended to answer. Or, as a lesson from other complex systems: "Don't model bulldozers with quarks" [13]. In particular, it is not always necessary to specify ODE-based models completely. Using advanced computational methods for exploiting the existing knowledge to the largest possible extent provides a pragmatic approach to gain biological knowledge and to reduce the experimental efforts. However, it poses important theory challenges. Here, we discuss methods that can potentially cope with increasing levels of uncertainty; they originate from different long-term theory developments that start to converge.

2. Structural network analysis

Network structures, especially the stoichiometries of biochemical reactions, are relatively well-characterized and therefore suitable starting points to analyze the largely unknown relationships between structure, function and control in complex cellular networks.

2.1. Concepts

Horn and Jackson were the first to study the effects of stoichiometric coupling in (chemical) reaction networks on their behavior [14]. They realized that – even without knowing parameter values $\mathbf{k}$ and component concentrations $\mathbf{c}$ – the stoichiometry imposes important constraints on network fluxes. Considering only steady states and neglecting the dynamics, Eq. (1.1) simplifies to:

$$\frac{d\mathbf{c}}{dt} = \mathbf{0} \;\Rightarrow\; \mathbf{N}\cdot\mathbf{v} = \mathbf{0} \tag{1.2}$$

Now, the space of feasible fluxes $\mathbf{v}$ is only determined by properties of the stoichiometric matrix $\mathbf{N}$ and potentially other constraints on $\mathbf{v}$ such as reaction reversibilities and capacities as illustrated in Fig. 1. The approach rapidly gained importance for biology after Palsson and colleagues proposed the flux balance analysis (FBA) method in 1992. FBA determines a specific flux distribution in a metabolic network by additionally considering optimality of, for example, biomass production [15]. The method and its extensions have found numerous applications in biology to analyze specific properties and behaviors of genome-scale metabolic networks; see [16,17] for recent reviews. However, it might be even more rewarding to characterize *all* possible behaviors of a network, which is the domain of metabolic pathway analysis.

2.2. Metabolic pathways

Formal metabolic pathways – as opposed to conceptual 'glycolysis', 'TCA cycle', etc. – have well-defined mathematical structures. Two such concepts are elementary flux modes (EFMs) and extreme pathways (EPs); they both originate from Clarke's early work on convex analysis of stoichiometric networks [18,19]. Since EFMs and EPs are closely related, we only consider EFMs in the following. Importantly, EFMs correspond to minimal (that is, non-decomposable) subnetworks that can operate a network at steady state while fulfilling all constraints that are imposed by reaction stoichiometries and reversibilities. They open up a *constructive* inroad for characterizing network behavior: all feasible flux distributions, and only those, are obtained by non-negative (convex) combinations of EFMs (see [20] for technical details). Hence, pathway analysis, in principle, allows one to *comprehensively* investigate the space of all states of (metabolic) networks that are meaningful for the cell.

Early applications of EFM analysis considered small-scale networks such as parts of canonical monosaccharide metabolism, thereby recovering text book pathways as well as proposing new network operation modes [21]. The first larger-scale analysis of *Escherichia coli* central metabolism showed the potential of EFM analysis for characterizing phenotypes and robustness of metabolic networks [22]. Primarily due to increased capabilities for computing EFMs, applications are now becoming possible for genome-scale networks. Recent studies indicated remarkable variability of fluxes within a narrow region of growth optimality [23], and helped uncover optimal, but rather counter-intuitive modes of network operation [24].

Importantly, large-scale analyses showed that focusing on 'core metabolism', while neglecting the peripheral network, can be very misleading [23,25]. This implies important challenges for EFM-based network analysis. First, it is not yet possible to analyze networks larger than ~300 reactions because of the high computational demands for determining EFMs. High-performance computing approaches and parallelization are options for further scaling. In addition, already now, EFM computations result in up to 10^7 pathways for large networks. Even simple statistical analysis such as determining the distribution of path lengths thus becomes non-trivial, and more advanced methods for clustering and other analyses at this scale are needed [23].

2.3. Network structure and regulation

While structural network analysis traditionally focuses on metabolic networks, several theories and computational methods have been proposed to infer (transcriptional) regulatory network structures and behavior from metabolic network structures. This is different from designing hybrid models that abstractly represent *known* genetic control mechanisms, for instance, via Boolean logic to more accurately predict the metabolic phenotype [26]. Reverse-engineering of the control network, at least in microorganisms, appears feasible because the control structures are sparse – only few metabolites directly control each regulator and vice-versa – and hierarchical [27]. In addition, there is increasing evidence that metabolic networks operate in a limited number of dominant functional modes in steady-state. Additional assumptions on optimal