ELSEVIER

Methods of biological network inference for reverse engineering cancer chemoresistance mechanisms

Paola Lecca

Centre for Integrative Biology – University of Trento, Laboratory of Computational Oncology, via Sommarive 14, 38123 Trento, Italy

We review recent Bayesian network inference methodologies we developed to infer genetic and metabolic pathways associated to oncological drug chemoresistance. Bayesian inference is supported by a rigorous and widely accepted mathematical formalization of predictive analytics. It is an inherently integrative approach allowing the incorporation of prior knowledge and constraints. Moreover, it is recommended to treat noisy data, and large amount of data whose dynamics laws are mostly unknown. We focus on variational Bayesian methods for the inference of stochastic reaction processes and we present a compendium of the recent results of inference of gene and metabolic networks presiding at the development of pancreas cancer resistance to gemcitabine.

Introduction

Development of resistance in chemotherapeutic agents is a big concern in cancer patients. Chemoresistance is due to several mechanisms that decrease drug cytotoxicity. The majority of these mechanisms are still poorly understood. The current attempts to reduce chemotherapy resistance are based on assumptions about the various candidate mechanisms. The inability to identify and understand the molecular mechanisms associated to the pathways entailing with the resistance is a reason of the low clinical outcomes of the current strategies to mitigate chemotherapy resistance. Recently, some studies showed that genome expression may predict response to drugs and patient outcome [1–3]. Moreover, the study of genes that influence drug activity and toxicity, known as pharmacogenomics, could offer the possibility of tailoring therapy to the specific profile of individual patients and tumors. A pharmacogenetic approach can thus potentially increase response rates and survival outcome while decreasing toxicity and overall treatment costs. Nevertheless, the current pharmacogenomic studies describe genes determining the sensitivity and resistance to chemotherapy drugs, but, at the best of our knowledge, these studies did not address the correlations among these genes. With regard to drug metabolism, experimental studies devoted to the direct measurements of concentration profiles of metabolites and metabolizing enzymes, as well as a plethora of computational procedure supporting the identification of drug targets, have led to the creation of reliable, even if incomplete, knowledge of the network describing the biotransformations and the mechanism of action of drugs. What is still lacking in the panorama of network pharmacology [4] is an *integrative* methodology able to (i) infer with a general mathematical proceeding both gene network and metabolic network, (ii) simulate them, and (iii) assemble them into a larger network showing the correlations between genes and metabolism. The use of such a result is of immediate understanding. In fact, integrative network inference methods approaches could support and even guide the experiments devoted to the identification of mechanisms of action of oncological drugs. Integrative computational inference procedures are expressions of the recent paradigm of *algorithmic systems biology* [5], that propels a computational approach to infer new biological knowledge, and dissect the complexity and understand the time evolution of biological networks. Computational integrative network inference procedures, when applied to unexplored biochemical processes, can also act as generators of hypotheses about the association of genes to metabolizing enzymes and thus can help in saving time and reducing the cost of the wet activities. Many software platforms and applications are facing this challenge. One of the recent large-scale projects in this direction is Ingenuity [6]. Ingenuity systems are web-based application for analyzing and interpreting the biological meaning of huge amount of genomics data. In particular Ingenuity Knowledge Base

E-mail addresses: paola.lecca@unitn.it, pllcc023@gmail.com.

1359-6446/06/$ - see front matter http://dx.doi.org/10.1016/j.drudis.2013.10.026

contains biological and chemical interactions and functional annotations created from millions of individually modelled relationships between proteins, genes, complexes, cells, tissue, drugs, and diseases. Furthermore, Ingenuity systems offer also analytical tools for data statistics, data modelling and model inference.

The goal of network inference is to deduce topological and – in case of availability of appropriate data – also the outative causal relationships among the components of a biological system from a corpus of data describing such a system [7–10]. Namely, the topological and causal structures of a biological process are often represented by network and visualized as a graph. The nodes of the graph/network represent the components of the biological systems (molecules, metabolites, proteins, etc.). An edge connecting two nodes indicates the existence of an interaction between the biological entities represented by these nodes; the orientation of an edge denotes the cause-effect relationship between two nodes or between two interactions or pathways. The interpretation of the edge orientation depends on the type of network. There are no formal syntax and formal semantics for the specification of a biological systems as a network, however there are some very common symbols familiar to and accepted by the majority of biologists and modelers. For instance, in gene networks, oriented edges are often accompanied by a sign (+ or −) that indicate activation or inhibition respectively, while in metabolic or protein–protein networks, oriented edges represent directions of chemical reactions. Dashed or dotted arrows with differently shaped tips, such as → or ⊣ (or accompanied by a sign + or −) are frequently used to indicate activation and inhibition of biochemical reactions and pathways. Many cartoons depicting biological networks use this notation, we refer the reader to some of the most known networks and pathways databases, such as Biocarta [11], KEGG Pathways Database [12], REACTOME [13], NCBI BioSystems [14], and Pathway Commons [15].

The data used to infer biological networks concern time series and/or steady state of the abundance of the system's components such as genes, molecules, proteins, metabolites, enzymes. Namely, metabolic reactome data, measurements of the metabolites concentration and rate of drug influx and efflux, as well as gene expression data are the input of the advanced techniques of network inference applied to pharmacology. Once a network is inferred we dispose of a pharmacokinetics model deeply rooted in real *in vivo* or *in vitro* experimental data. The network is a model that can be specified in a mechanistic way and whose dynamic can be simulated through algorithms implementing the kinetics of biochemical interactions in case of metabolic network, or the dynamic of gene interaction in case of gene networks.

When applied to pharmacology and in particular to drug discovery and design process, network inference methodologies have to be necessarily *integrative* [16]. Namely, in order to be predictive and reliable they need to integrate different types of data, mainly gene expression levels and metabolic data, because the response to a pharmacological treatment is directed by the genetic profile of the patient and by its individual metabolic and pharmacokinetic rates. Furthermore, nowadays there is stronger and stronger demand by the biologists and modelers to integrate in the network inference methods also simulation algorithms. Alternating runs of network inference and network simulation allows to iteratively evaluate the effectiveness of the inference methods at recovering models of complex gene networks and complex pharmacokinetics in cases of a limited amount of data [17]. Furthermore, by simulating the inferred network it is possible to get the time behavior also of the unobserved system's components. In this way the simulated time series of all system components together with the inferred network used as an *a priori* knowledge, can be used for a further run of network inference, in an iteratively process. Finally, software tools for network inference equipped with procedure for the integration of different types of inferred network are highly demanded [18] by the community of modelers and biologists. The researchers employed in the identification of chemoresistance mechanisms are particularly interested in disposing of procedure and tools for data integration upstream and/or downstream of the network inference procedure.

While available sequence of gene expression data are rapidly provided by high-throughput experiments, our current knowledge of the interactions among genes responsible for the tumor chemoresistance and chemosensitivity is still poor. Furthermore, although recent advancements in experimental high-throughput metabolomics and proteomics allowed the collection of a huge amount data both on single proteins or metabolites concentration profiles and on metabolic pathways [19–22] what we currently know about the relationship between gene networks determining drug sensitivity and/or resistance and networks of drug transporters and metabolizing enzymes constitutes only a small fraction of what remains to be discovered. In situations in which there is a large amount of data, but little theory describing the source of those data (i.e. models or hypotheses), Bayesian inference provides a natural and principled way of including into an inferential probabilistic framework prior knowledge embedded in the observed data to deduce models underlying the observed data themselves. In Bayesian inference all forms of uncertainty are expressed in terms of probability. The Bayesian method starts with the formulation of a model (or hypothesis) that we suppose is adequate to describe the background information and the data. A prior probability is assigned to this model. Finally the Bayes' theorem is used to evaluate a posterior probability (i.e. a degree of belief) for the model in light of the available data [23]. It is worth noting that the Bayesian approach is not directly concerned with the creative process, how to generate new hypothesis or models. It is concerned mainly with assessing the extent to which models reproduce the available knowledge and data. Its use is then particularly suitable to a present context in which the high quality high-throughput data enable the formulation of *a priori* knowledge able to initialize and guide the inference procedure.

In this paper we focus on pancreatic cancer cell lines treated with gemcitabine. We review a general integrative network inference methods we developed in the last two years in close collaboration with experts in pharmacokinetics and pharmacodynamics (Veltkamp *et al.* [24]) and guided by the methods proposed by mathematicians and computer scientists (Samoilov *et al.* [25]) that can be used to infer both the network of interactions among genes responsible of the gemcitabine resistance and the network of gemcitabine metabolism in pancreatic cancer cells. The inferential framework combines variational Bayesian methods recently proposed by Lawrence *et al.* [26] with correlation-based methods presented by Lecca *et al.* [28,29].