Mathematical Biosciences 270 (2015) 156–168

Contents lists available at ScienceDirect

Mathematical Biosciences

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

From genome-scale data to models of infectious disease: A Bayesian network-based strategy to drive model development

Weiwei Yin[a,1], Jessica C. Kissinger[b], Alberto Moreno[c], Mary R. Galinski[c], Mark P. Styczynski[a,*]

[a] School of Chemical & Biomolecular Engineering, Georgia Institute of Technology, Atlanta, GA 30332-0100, USA
[b] Department of Genetics, Institute of Bioinformatics, Center for Tropical and Emerging Global Diseases, University of Georgia, Athens, GA, USA
[c] Division of Infectious Diseases, Emory Vaccine Center, Yerkes National Primate Research Center, Emory University School of Medicine, Emory University, Atlanta, GA, USA

ARTICLE INFO

Article history:
Available online 17 June 2015

Keywords:
Bayesian network inference
Large-scale data analysis
Model development
Infectious diseases
Malaria

ABSTRACT

High-throughput, genome-scale data present a unique opportunity to link host to pathogen on a molecular level. Forging such connections will help drive the development of mathematical models to better understand and predict both pathogen behavior and the epidemiology of infectious diseases, including malaria. However, the datasets that can aid in identifying these links and models are vast and not amenable to simple, reductionist, and univariate analyses. These datasets require data mining in order to identify the truly important measurements that best describe clinical and molecular observations. Moreover, these datasets typically have relatively few samples due to experimental limitations (particularly for human studies or *in vivo* animal experiments), making data mining extremely difficult. Here, after first providing a brief overview of common strategies for data reduction and identification of relationships between variables for inclusion in mathematical models, we present a new generalized strategy for performing these data reduction and relationship inference tasks. Our approach emphasizes the importance of robustness when using data to drive model development, particularly when using genome-scale, small-sample *in vivo* data. We identify the use of appropriate feature reduction combined with data permutations and subsampling strategies as being critical to enable increasingly robust results from network inference using high-dimensional, low-observation data.

Abbreviations

ANOVA	Analysis of variance.
ARACNE	Algorithm for reconstruction of accurate cellular networks.
BN	Bayesian network.
CLR	Context likelihood of relatedness.
DREAM	Dialogue for reverse engineering assessment of methods.
FDR	False discovery rate.
FOM	Figure of merit.
MI	Mutual information.
MMC	Modulated modularity clustering.
MRMR	Maximum relevance/minimum redundancy.
MRNET	Minimum redundancy networks.
PCA	Principal component analysis.
PLS-DA	Partial least squares discriminant analysis.
RFE	Recursive feature elimination.
SVM	Support vector machine.
TMI	Total mutual information.

1. Introduction

The proliferation of genome-scale experimental analysis techniques—proteomics, transcriptomics, metabolomics, and others—brings with it numerous challenges in data analysis. With many more measurements (or variables) than observations, it is complex (both conceptually and computationally) to identify those variables that are most important in determining the phenotype or outcome of a system, as well as how these variables interact with each other. The identification of these variables and interactions is a crucial step in most downstream work, whether the development of diagnostics or the detailed study and modeling of a biological system.

Modeling of infectious diseases is a particularly salient and important example of where addressing this challenge is critical. The mechanisms, presentation, and transmission of infectious disease are quite complex and depend on a number of factors including the host, the pathogen, the environment, and potentially even the vector.

* Corresponding author. Tel.: +1 404 894 2825.
E-mail addresses: wwyin@me.com (W. Yin), jkissing@uga.edu (J.C. Kissinger), camoren@emory.edu (A. Moreno), Mary.Galinski@emory.edu (M.R. Galinski), Mark.Styczynski@chbe.gatech.edu (M.P. Styczynski).
[1] Present address: Key Laboratory for Biomedical Engineering of Education Ministry, Department of Biomedical Engineering, Zhejiang University, Hangzhou, PR China

http://dx.doi.org/10.1016/j.mbs.2015.06.006
0025-5564/

Malaria, for example, is a disease caused by five different species of *Plasmodium*, and each of these pathogens can cause different clinical presentations and degrees of severity of the disease. *Plasmodium vivax* infections are typically characterized by fever spikes every 48 h, but the shorter life cycle of *P. knowlesi* typically manifests as a daily spike in fever [1]. Around 10% of *P. falciparum* infections result in severe malaria, with somewhere between 10% and 50% of those severe cases being fatal [2], while *P. vivax* infections may more rarely cause severe malaria [3]. Within the same species, specific strains of the parasite can be quite different, causing (for example) varying efficiencies of vector infection (and thus of disease transmission) [4] or strain-specific resistance to certain classes of drugs [5,6]. All of these parasite variations are on the backdrop of host variations, which can have a tremendous impact on the presentation of the disease even between different individuals infected by isogenic parasites. Complicating the situation even further is that all of the above factors (whether related to host, parasite, or vector) are only things already known to be key in the disease, with potentially numerous additional critical factors that we just have not discovered yet.

With such complex dependence on different aspects of the host and pathogen, varying phenotypic effects after infection, and the general uncertainty about what all of the controlling factors in disease progression are, creation of mathematical models of infectious diseases is obviously quite difficult. One critical question, even if by virtue of its primacy in the process of developing models, is: what variables should be included in the model? As suggested above, a wealth of proteomics, metabolomics, and other measurements can provide the data that can help to build an effective model, but sifting through the large volume of measurements that do not correlate to the phenomenon of interest to identify the ones that do is a monumental task requiring appropriate statistical treatment. Even supposing that the right variables to include in a mathematical model could be identified, the notion of how to include these variables is the next significant task. One may know that a specific gene is important in a process, but that does not sufficiently inform the mathematical model. Does that gene affect only one other molecule (variable) in the system? If so, which variable does it affect? If it affects multiple variables, how many does it affect, and which ones? And then beyond this, if one knows which interactions to include, there is still the open question of the appropriate functional form to represent this interaction.

Thus, the ability to use modern high-throughput, high-information content, genome-scale data effectively will be essential in developing models for infectious diseases. These models may be on a molecular scale, indicating transcriptional or other regulation within a pathogen or indicating the interactions between host and pathogen biology. They may also be on a much larger scale, capturing epidemiological dynamics as a function of key aspects of the hosts, pathogens, and vectors. In either case, it is crucial that the methods used to identify the variables and relationships to be included in the model are as robust and accurate as possible. This is particularly relevant for cases where taking a large number of samples is not feasible, particularly *in vivo* clinical studies involving infected humans in need of treatment and non-human primate experiments where both cohort size and sampling frequency are limited on ethical grounds. One must extract as much information as possible in as reliable a way as possible with a comparatively small number of observations.

Here, we will first briefly review some of the common approaches to whittle genome-scale data into candidate knowledge for inclusion in mathematical models. We will then focus on one specific and promising approach to achieve this goal, Bayesian networks, and address some of the difficulties inherent in using this approach. We consider this task particularly in the context of systems where we expect to have a fairly small number of observations with significant biological variability. We present a unique approach that uses clustering to reduce the dimensionality of a dataset, concatenation of clustered genes to increase the effective number of observations, and permutation and cross-validation analysis to ensure that the results of network inference are trustworthy for the purposes of modeling and not disproportionately influenced by random variation. Taken together, this represents an efficient and reasonable approach to drive the generation or improvement of mathematical models in infectious disease research.

2. Background

Multivariate dimensional reduction, classification, and visualization approaches are often used as first-line analyses in the interpretation of high-variable, low-observation genome-scale datasets. Methods include principal component analysis (PCA), partial least squares discriminant analysis (PLS-DA), and numerous variations thereof that reduce the original variable space to a few composite variables [7,8]. In the ideal case, samples from the same group are close to each other in the reduced feature space, and the weight of the original variables in these key composite variables can be used to drive further downstream interpretation and analysis, including ontological analyses like enrichment analysis. Other methods for group classification tasks (e.g., support vector machines or artificial neural networks) can create classifiers capable of separating two sample classes, though with potentially increased difficulty in interpreting the biological meaning of the mathematical representations in the inferred classifier.

What such classification and dimensional reduction approaches largely do not permit is the ability to discover new interactions between variables. Much richness in biological systems is driven by the complexities of regulation, which manifests itself by the apparent correlation of biomolecules across time or experimental conditions. The ability to identify interactions between variables is a valuable tool to learn more about the molecular level of novel or understudied complex biological systems, and in particular it is valuable for knowing what variables should be included in mathematical models of such systems.

As discussed in Section 1, the process of reducing genome-scale data to a form that can be integrated into mathematical models can be broadly divided into three steps: feature selection, identification of candidate interactions between features, and mathematical formulation of those candidate interactions.

2.1. Feature selection

Feature selection is the process of reducing a larger set of variables to a subset for use in model construction or further analysis. This is due to the expectation (quite appropriate for genome-scale data) that a significant fraction of the measured variables are either not relevant to the task at hand or are redundant. The latter is a particularly troubling problem for the construction of mathematical models, as the inclusion of redundant variables will greatly increase the computational complexity of estimating parameters in the mathematical model, and may in fact prevent many parameters from being identifiable. Feature selection methods may be independent filters, they may be search-and-score approaches that select subsets of features and assess the accuracy of the model derived from those features ("wrapper" methods), or they may be directly embedded into (and specific to) the model development [9]. A set of common methods to perform this task is described herein; while representative, this list is by no means exhaustive.

Common embedded methods include recursive feature elimination (RFE) and Lasso (the least absolute shrinkage and selection operator). RFE is an embedded approach often used in the development of support vector machines (SVMs), a powerful tool for classification problems [10–12]. In this approach, variables deemed unimportant in early model development are considered candidates for elimination from the model, with progressively more parsimonious models