Deciphering and Modeling the Genomic and Proteomic Regulatory Pathways of the Innate Immune System using Dynamic Bayesian Networks

Kenneth L. Drake, CTO, Seralogix, LLC

Abstract

Seralogix is developing new computational tools and methods for the identification, analysis, and modeling of the mechanisms and pathways associated with the host-pathogen innate immune and inflammatory responses to infectious diseases - including biowarfare agents. Our core computational tool is based on the statistical power of dynamic Bayesian networks (DBNs), which is utilized to learn and model the complex dynamic pattern-of-change of DNA, mRNA, proteins and metabolites, which we refer to as the "temporal biosignature" of the host-pathogen response. DBNs are based on sound probabilistic methods that allow us to combine prior knowledge with time-course empirical data for deciphering host-pathogen biosignatures with a biological system perspective. Because of the complex multi-dimensional data resulting from genomic and proteomic investigations, new computational tools, with built in intelligence, are required to serve the investigative needs of the 21st century.

We hypothesize that our DBN methodology should substantially improve the statistical significance for inferring innate and inflammatory pathways from less experimental data while also confronting noisy, hidden, and/or missing data points.

Deciphering Host-Pathogen Mechanism

Understanding the mechanisms of the immune and inflammatory response is one of the most important problems in contemporary biology. Extensive studies by the researchers on this project and others have been done both experimentally as well as by modeling to understand the immune response system. However, typically only limited portions (sub-components) of the immune system have been studied, leaving major gaps in our understanding of the fuller system behavior and time-course of the host-pathogen response, mostly due to lack of technology and computational tools. Simultaneously analyzing patterns of intracellular gene and protein expressions and intercellular signaling protein during the early onset (incubation) of infection and its initial progression holds great promise for deciphering infectious disease causal relationships. The lack of computational tools limits our ability to describe the regulatory and mediator pathways and key interdependences that may explain the observed differences between pathogens.

Innate Immune Modeling Using DBN

We represent the innate immune and inflammatory response and their underlying mechanisms and pathways as a Dynamic Bayesian Networks (DBN)[1]. We have prototyped DBN models from prior knowledge and synthesized time-course data for both bacterial and viral type innate immune and inflammatory responses. We capture and visualize prior known causal relationships by using graphical modeling techniques, resulting in a directed acyclic graph (DAG) as illustrated in Figure 1. Here, we have assumed a simple pathogen-host immune response model and empirical data as shown on the (Fig 1(a)) that translates to the static Bayesian network DAG structure (Fig 1(b)). The gene/protein biomarkers and other information are the random variables. The yellow shaded nodes in our DAG are the observed/measured biomarker nodes. The yellow nodes represent measured gene/protein expression levels and physiologic factors. The white nodes are hidden variables used to capture certain simplification assumptions and other unmeasured model influences. These nodes accumulate the influences of hidden processes such as genetic translation pathways, individual cellular protein secretions, body protein clearance and measurement variability. The arcs in the model represent causal relationships.

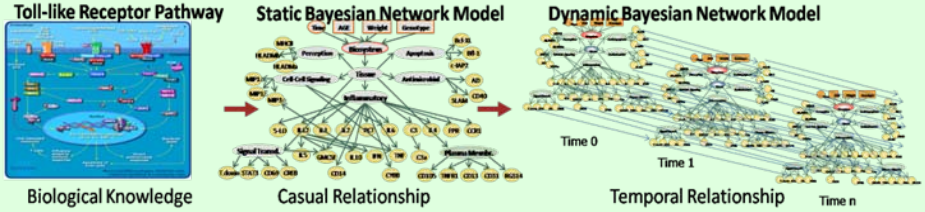

Figure 1. The translation from prior biological knowledge and structure learning from gene/protein expression data (a) to a directed acyclic graph (DAG). The DAG represents a simplification of our work and shows the static Bayesian network model (b) rolled out for only three time-points as the dynamic Bayesian network (c). The DBN model parameters are learned from empirical data as well as from prior data. Different DAGs will be required to model host-pathogen responses because of the differences in the pathogen invasion and evasion mechanisms.

[1] Dynamic Bayesian Networks: Representation, Inference and Learning, UC Berkeley, Computer Science Division. July 2002.

Regulatory Pathway and Mechanistic Discovery

Our modeling approach is accomplished with a hypothesis driven host-pathogen time-course experimental design and methodology. Our modeling/computational tools support time-course studies that are key to the deciphering of regulatory pathways and mechanisms. We are developing new algorithms that allow us to apply automated Bayesian network structure learning based on prior pathway knowledge and the correlation of changes that occur between control groups, experimental groups of different genotypes (resistant vs. susceptible) and the measured change in gene/protein expression during disease progression (Figure 2).

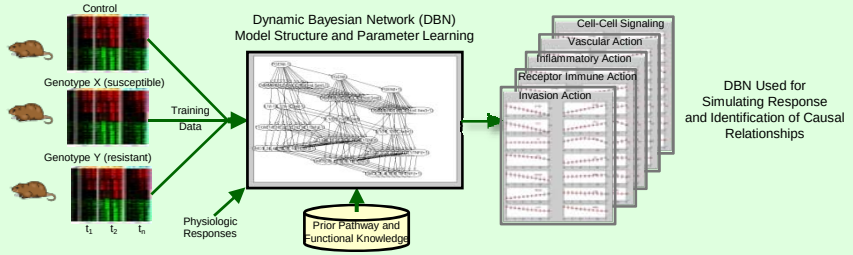

Figure 2. Dynamic Bayesian Network Model (DBN) generated from time-course host-pathogen reponse studies allow the combination of control and experimental data with prior biological functional and pathway knowledge to infer the causal differences between the groups. Further, our DBN can link genomic response to physiologic response, all in a single model.

Results

We are in the early stage of software development and have only preliminary data to report. Our initial focus has been on validating our computational approach for learning the correct dynamic Bayesian network model from synthesized training data that was generated from a process with known causal relation-ships. We developed nine different sets of training data representing different pathogens with only subtle changes in the simulated mRNA/protein host-pathogen response data. Each dataset creates a DBN with its own unique parameters resulting in very accurate representation of the underlying empirical data. Table 1 shows the results of our initial evaluation where the majority of the training cases resulted in close to 100% accuracy. We intentionally induce strong non-linearities in the datasets for Glanders and Lassa pathogens which lowered our model representation accuracy. We can correct for non-linearity by using other probability distribution functions within our learning algorithm. We developed our existing models reported herein using Gaussian distributions.

Table 1: Evaluation of DBN Modeling Correctness

DBN Models	Training Cases	Test Cases	Model Correctness p=.05
Salmonella	25	100	99%
Tularemia	25	100	99%
Glanders	25	100	75%
Smallpox	25	100	100%
Brucellosis	25	100	100%
Anthrax	25	100	100%
Lassa	25	100	85%
Plague	25	100	100%
West Nile	25	100	98%

Conclusion

We believe that our tools in conjunction with emerging, quantitative genomic and proteomic array technologies will be essential in supporting new *in vitro* and *in vivo* investigations that combine time, system level analysis, and prior signaling pathway knowledge to support:

- new drug and vaccine developments; and
- new pre-symptomatic diagnostics and therapeutic management approaches important for 21st century healthcare and the realization of personalized medicine.

Acknowledgement: Funding provided by National Institute of Allergies and Infectious Diseases Grant No. R43 AI058362-01

Contact Information: Kenneth Drake, Ph.D. , drake@seralogix.com, Phone: 512-533-2056

Seralogix
8329 N. MoPac Expwy
Austin, TX 78759