Pre-symptomatic Infectious Disease Diagnostics Using Dynamic Bayesian Networks to Learn and Recognize Temporal Genomic/Proteomic Biosignatures

Kenneth L. Drake, CTO, Seralogix, LLC

Hypothesis

Host-pathogen studies indicate that exposed individuals display unique peripheral blood mononuclear cell (PBMC) mRNA expressions and serum protein temporal patterns (biosignature) to pathogenic agents prior to the onset of full illness[1]. It is our hypothesis that we can exploit the unique temporal patterns for early detection and even the identification of the infectious agents (including biowarfare agents and other diseases of high public health concern). New computational tools are required to automatically learn and recognize the unique biosignature patterns from the volumes of complex, time-course genomic/ proteomic PBMC expression studies.

Our goal is to implement a framework of modeling and computational tools to validate the hypothesis that each pathogen produces a unique biosignature that can be used for pre-symptomatic diagnosis and therapeutic management.

Biosignatures

Biosignatures are composed of the temporal change in a few hundred to thousands of biomarkers (genes and proteins) and physiologic markers. The biomarkers are measured using either gene microarrays, 2-D gel/mass spec, or multiplex immunoassays as shown in Figure 1. All of these technologies enable our ability to measure the host-pathogen response, but making practical use of the data remains elusive because of our lack of analytical software tools.

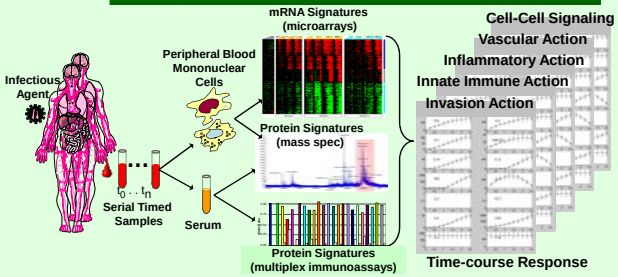

Figure 1. The primary analytical methods employed to measure the temporal biosignature response to pathogen challenge.

Computational Methods

Our computational tools are based on the probabilistic power of dynamic Bayesian networks (DBNs)[2] which are utilized to learn, model and recognize the dynamic pattern-of-change of mRNA and proteins ("biosignature") of the host-pathogen innate immune and inflammatory responses. A unique feature of our approach is the inclusion of "time" combined with prior quantitative and qualitative knowledge that is key to the recognition accuracy between different pathogenic agents. Our methodology represents a sound statistically based approach for model generation and Bayesian inference for biosignature pattern recognition. The methodology has two data flow paths (Figure 2). The top path labeled "Dynamic Bayesian Net Model Learning" generates the models from experimental training datasets. This path combines prior biological knowledge with experimental data to produce a family of biosignature models that are uniquely associated with each pathogen. The bottom path labeled "Pathogen Biosignature Pattern Recognition" is a process that leads to the comparison of an unidentified biosignature (partial or complete test data case) for pattern match to a library of DBN pathogen models.

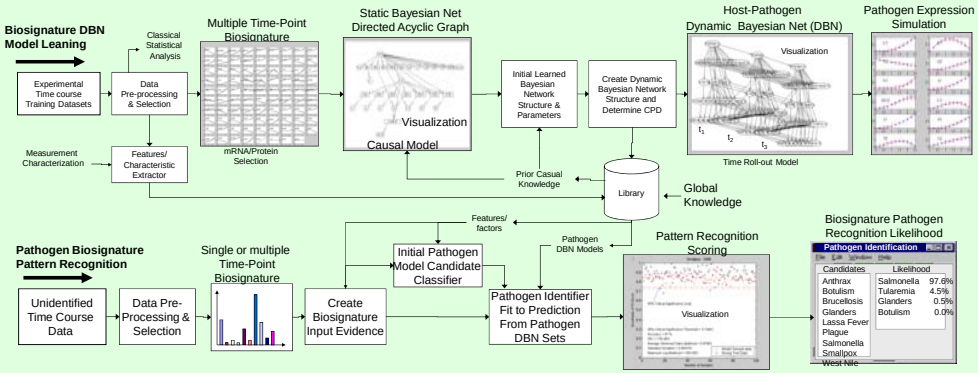

Figure 2. Biosignature DBN computational methodology for model learning and pattern recognition

Training and Testing Datasets

We created synthesized host-pathogen datasets by combining previous animal experiment immune response data, augmented with published data and predictions resulting in nine training datasets. We used time course data from SEB, LPS, Salmonella, Tularemia, and anthrax. We simulated the time-course response pattern and variance of each biomarker variable as a function of time, biomarker causal relations, pathogen type, host weight, pathogen dose and host genotype (resistant or susceptible). This provided us with a means to generate data with known causal and temporal relationships that closely mimicked actual immune response. We can perturb the patterns by modifying noise, genotype, weight, and pathogen dose to evaluate the effects of variation on biosignature pattern recognition performance. Our baseline noise factor assigned to each variable produced a standard deviation from ±20% to ±30% of the mean values for each variable across each time-point.

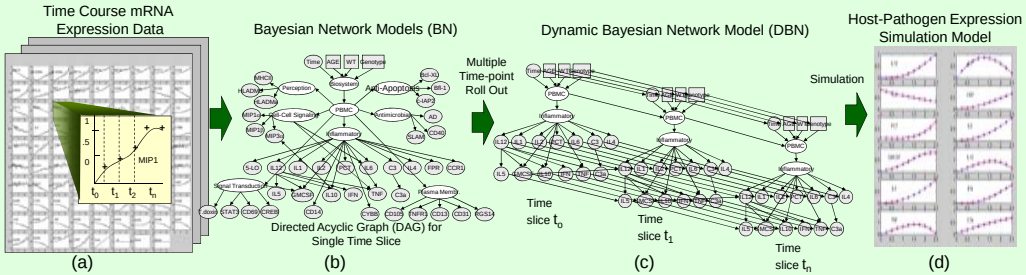

Figure 3. Illustrates the use of training data sets (a) to create the model parameters for the Bayesian network (b) that is rolled out for three time slices (c). The subsequent time slices in (c) show a dependence on the prior time slice variable values. Note also that prior time slice variables can be parents (dependencies) on other variables representing up or down gene regulation for the child variable. Each node represents an observed or hidden variable (gene/protein, etc). The variables can be discrete or continuous and include physiologic factors such as sex, weight, genotype, etc. DBN model time-course simulation versus training data is shown in (c).

Results

Our bioinformatics prototype demonstrated proof-of-concept that our computational approach learned correct DBN models from time-course data comprised of several hundred mRNA, protein and physiologic response factors mimicking the innate immune response to several different pathogens. The prototype performed remarkably well in both the representation of the time-course biosignatures as DBN models and for correctly identifying an unknown biosignature to the correct infectious agent with better than 98% accuracy and correct time elapse from initial exposure (Table 1).

Table 1: Results of Biosignature Recognition Evaluation

DBN_{PGEN} Models	True Test Cases	False Test Cases	True Time-slice Identification %		Biosignature Recognition Rates			
					Single Time-slice		Two Time Slices	
			Exact	± 1 time point.	Positive %	False Positive %	Positive %	False Positive %
Salmonella	20	20	10%	90%	100%	5%	80%	11%
Tularemia	20	20	33%	41%	100%	8%	100%	0%
Glanders	20	20	25%	58%	100%	25%	83%	0%
Smallpox	20	20	25%	60%	100%	5%	80%	5%
Brucellosis	20	20	16%	75%	100%	0%	100%	0%
Anthrax	20	20	50%	33%	100%	0%	100%	0%
Lassa	20	20	16%	42%	100%	0%	100%	0%
Plague	20	20	50%	41%	100%	25%	100%	50%
West Nile	20	20	42%	8%	58%	0%	58%	0%

Conclusion

We believe that our DBN based computational tools will be important for:

- deciphering the cellular signaling pathways and mechanisms of virulence and toxicity of pathogens and toxins
- creating new diagnostics for real-time, pre-symptomatic pathogenic identification
- understanding disease progression to aid in creating new intervening drugs and therapeutic strategies.

[1] M. Jett, et.al.: "Identification of Changes in Gene Expression Induced by Toxic Agents: Implications for Therapy and Rapid Diagnosis," NATO Conference: Operational Issues in Chemical and Biological Defense Human Factors in Medicine Panel, Proceeding May 2001.

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

Acknowledgement: Funding provided by National Institute of Allergies and Infectious Diseases Grant No. R43 AI055061-01

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

Seralogix
8329 N. MoPac Expwy
Austin, TX 78759