Pathway Semantics: An Algebraic Data Driven Algorithm to Generate Hypotheses about Molecular Patterns Underlying Disease Progression

Mary Frances McGuire
University of Texas Health Science Center at Houston

Follow this and additional works at: http://digitalcommons.library.tmc.edu/uthshis_dissertations

Part of the Medicine and Health Sciences Commons

Recommended Citation
http://digitalcommons.library.tmc.edu/uthshis_dissertations/18
PATHWAY SEMANTICS:

AN ALGEBRAIC DATA DRIVEN ALGORITHM TO GENERATE HYPOTHESES
ABOUT MOLECULAR PATTERNS UNDERLYING DISEASE PROGRESSION

A

DISSERTATION

Presented to the Faculty of
The University of Texas
School of Biomedical Informatics
at Houston
in Partial Fulfillment
of the Requirements

for the Degree of

Doctor of Philosophy

by

Mary Frances McGuire

Committee Members:

M. Sriram Iyengar, Ph.D.¹
Jack W. Smith, M.D., Ph.D.¹
Robert E. Brown, M.D.²
David W. Mercer, M.D., FACS.³

¹ The University of Texas School of Biomedical Informatics at Houston
² The University of Texas Medical School at Houston
³ The University of Nebraska Medical Center

Copyright © 2011 Mary Frances McGuire. All rights reserved.
Dedication

In memory of my parents, Margaret J. and Walter B. McGuire,

who encouraged me to trust in God and follow my dreams
Acknowledgements

I owe my deepest gratitude to those who have made this achievement possible through support, encouragement, faith, patience, knowledge and love. First, I thank my longstanding doctoral committee at the University of Texas at Houston. M. Sriram Iyengar PhD, my primary mentor at the School of Biomedical Informatics, introduced me to the exciting field of systems and computational biology, encouraged me to experiment with new ways of looking at challenging clinical problems, and treated me as kindly as a family member. Dean Jack W. Smith MD, PhD supported this research in innumerable ways, including database access. David W. Mercer MD piqued my interest in shock / trauma progression with an extensive data set on multiple organ failure. Because statistical analysis of this data was all too limiting, I set out on a quest to find other mathematical approaches that eventually coalesced into the Pathway Semantics Algorithm (PSA). Robert E. Brown MD painstakingly taught me about signaling pathways, helping me to overcome the twin hurdles of molecular notation and pronunciation. Although not on my committee, Keri Csencsits Smith PhD encouraged me to apply PSA to her hemophilia study. She performed the additional assays that validated PSA experimentally by confirming the predominant presence of a predicted molecule in the treated group.

I thank the faculty, staff, and my colleagues at UT Health – too many to enumerate – for their instruction, collaboration and help in many ways. In addition to those at my school, I thank the people at the UT Health Medical School Departments of Surgery and Pathology/Laboratory Science, UT Health Graduate School of Biological Sciences, UT
Health School of Nursing, UT Health School of Public Health and MD Anderson Cancer Center. Outside of the university, I thank the staff of the Keck Center for Interdisciplinary Bioscience Training – Gulf Coast Consortia for their encouragement and biologically-oriented lecture series, where I learned even more about molecular notation and pronunciation; I thank my Keck mentor, Devika Subramanian PhD at Rice University, for her insights into the crossroads of computer science and biology. I thank John Svirbely MD for his numerous emails about the latest publications and ideas about the molecular underpinnings of disease progression. I thank the organizers of, and participants in, the yearly interdisciplinary International Conferences on Complexity in Acute Illness (ICCAI) for their encouragement year after year; it’s rare finding a like-minded group of supportive mathematicians, engineers, biological scientists, industry folks and clinicians all working towards improved critical care. Finally, on the academic side, I acknowledge my undergraduate and graduate professors at the Columbia University School of Engineering, New York, for laying the educational groundwork that supports my current endeavors. In particular, I thank my master’s thesis mentor, Philip Wolfe PhD, now an IBM Research Yorktown Heights emeritus, for his emphasis on simplicity in theory construction for practical applications.

I gratefully acknowledge the following funding support:

• National Institutes of Health training fellowship from the Keck Center for Interdisciplinary Bioscience Training of the Gulf Coast Consortia NLM Grant 5-T15-LM07093-16 (2008–2010)
• Harvey S. Rosenberg, Endowed Chair in Pathology for the Morphoproteomics Initiative, University of Texas Medical School at Houston (2007–2010)

• UT Health Innovation for Cancer Prevention Research Fellowship, the University of Texas School of Public Health – Cancer Prevention and Research Institute of Texas, Grant RP101503 (2011–2012)

• HIMSS South Texas Chapter Travel Scholarship (2007)

• Doris Ross Competitive Academic Scholarship from the University of Texas School of Health Information Sciences (now Biomedical Informatics) at Houston (2007)

• Travel grant from the Eighth International Conference on Systems Biology (2007)

• Travel scholarship from the Committee on the Status of Women, University of Texas Health Sciences Center at Houston (2008)

• For data: The shock / trauma study was supported by the National Institutes of Health GM-38529, GM-08792, CReFF UCRC #M01RR002558 to David W. Mercer MD, FACS; I acknowledge the participation of the Memorial Hermann Hospital STICU, Houston in that study.

• For data: The hemophilia study was supported by a National Hemophilia Foundation Career Development Award to Keri Csencsits Smith PhD.

I acknowledge the support of my family and friends that has kept me sane during this time. I thank the Beck family in Houston for their generosity in providing me with a “home away from home” during my first two years of PhD studies, and my new Houston area friends for including me in holiday and social events when I was away from my
Calgary home. Finally, I thank my husband, Paul M. Beck for enduring our lengthy separations with humor, love, and wisdom as well as making my academic work possible. Paul provided computer system administration support, programming consultation, graphic design for posters, and a sounding board for my research ideas. He was instantly available to solve technical problems – or just to say words of encouragement – that kept me going towards completion of this research work.

Thank y’all!
PATHWAY SEMANTICS:
AN ALGEBRAIC DATA DRIVEN ALGORITHM TO GENERATE HYPOTHESES
ABOUT MOLECULAR PATTERNS UNDERLYING DISEASE PROGRESSION

Mary Frances McGuire, PhD
The University of Texas School of Biomedical Informatics at Houston, 2011

Primary Advisor: M. Sriram Iyengar, PhD

ABSTRACT
The overarching goal of the Pathway Semantics Algorithm (PSA) is to improve the in silico identification of clinically useful hypotheses about molecular patterns in disease progression. By framing biomedical questions within a variety of matrix representations, PSA has the flexibility to analyze combined quantitative and qualitative data over a wide range of stratifications. The resulting hypothetical answers can then move to in vitro and in vivo verification, research assay optimization, clinical validation, and commercialization. Herein PSA is shown to generate novel hypotheses about the significant biological pathways in two disease domains: shock / trauma and hemophilia A, and validated experimentally in the latter. The PSA matrix algebra approach identified differential molecular patterns in biological networks over time and outcome that would not be easily found through direct assays, literature or database searches.
In this dissertation, Chapter 1 provides a broad overview of the background and motivation for the study, followed by Chapter 2 with a literature review of relevant
computational methods. Chapters 3 and 4 describe PSA for node and edge analysis respectively, and apply the method to disease progression in shock / trauma. Chapter 5 demonstrates the application of PSA to hemophilia A and the validation with experimental results. The work is summarized in Chapter 6, followed by extensive references and an Appendix with additional material.

Disclaimer: The content is solely the responsibility of the author and does not necessarily represent the official views of any employer, funding agency or institution.

Conflict of Interest: none declared.
# Table of Contents

List of Illustrations ........................................................................................................... xi

List of Tables ...................................................................................................................... xii

Chapter 1 Introduction ....................................................................................................... 1
  Background and Motivation .............................................................................................. 1
  Problem Statement .......................................................................................................... 6
  Study Purpose, Scope and Deliverables .......................................................................... 8
  Study Significance .......................................................................................................... 8
  Theoretical / conceptual framework .............................................................................. 11
  Summary ......................................................................................................................... 13

Chapter 2 Literature Survey ............................................................................................. 15
  Systems Analysis and Systems Biology ......................................................................... 17
  Systems approaches in trauma ........................................................................................ 18
  Data for systems approaches in trauma .......................................................................... 19
  Computational Methods ................................................................................................. 22
  Basic probabilistic approaches ...................................................................................... 23
  Basic deterministic approaches .................................................................................... 24
  Graphical approaches .................................................................................................... 26
  Symbolic approaches ..................................................................................................... 29
  Artificial Neural Networks (ANN) and Agent-Based Models (ABM) ............................ 30
  Applications .................................................................................................................. 31
  Organism level .............................................................................................................. 33
  Organism level: Process models .................................................................................... 33
  Organism level: Predictive models ............................................................................... 36
  Organ level ..................................................................................................................... 38
  Tissue level ..................................................................................................................... 40
  Cellular/molecular levels ............................................................................................... 41
  Summary ......................................................................................................................... 45

Chapter 3 Pathway Semantics Algorithm: Node Analysis ........................................... 46
  The challenge of pathway analysis for shock / trauma .................................................. 48
  Using algebra to understand biological pathways in disease progression .................. 50
  Algorithm ....................................................................................................................... 53
  Application ..................................................................................................................... 56
  Laboratory materials and methods .............................................................................. 57
  Results ............................................................................................................................ 68
  Discussion ..................................................................................................................... 74
  Summary ......................................................................................................................... 78

Chapter 4 Pathway Semantics Algorithm: Edge Analysis .......................................... 79
  Background .................................................................................................................... 81
  Additional definitions ................................................................................................. 83
  Pathway Semantics Algorithm for Edges .................................................................... 84
  Application ..................................................................................................................... 87
  Results ............................................................................................................................ 91
  EM Matrices .................................................................................................................. 93
List of Illustrations

Figure 2-1 Computational methods integrate translational research .................................. 22
Figure 2-2 Computational analysis at different levels within an organism ......................... 32
Figure 3-1: Goal of Pathway Semantics Algorithm .......................................................... 52
Figure 3-2: PSA flow diagram for shock / trauma study node analysis .............................. 55
Figure 3-3: Biological networks differ by disease state .................................................... 63
Figure 3-4: Temporal Dependency Matrices (TDMs) example ........................................ 65
Figure 4-1: Graph example ................................................................................................. 87
Figure 4-2: Evoked biological pathways in MOF at hours 6–10 ........................................ 92
Figure 4-3: Section of EM matrix for network shown in Figure 4-2 ................................. 93
Figure 4-4: Edge interactions over time, outcome, and functional relationship ............... 94
Figure 4-5: Unique edge interaction counts ...................................................................... 96
Figure 4-6: Crosstalk in functional relationships across time and outcome ..................... 97
Figure 5-1: PBS Day 7 ......................................................................................................... 109
Figure 5-2: CFA/I Day 7 .................................................................................................. 110
Figure 5-3: PBS Day 14 .................................................................................................... 111
Figure 5-4: CFA/I Day 14 ............................................................................................... 112
Figure 5-5: Unique evoked pathway molecules on Day 7 ............................................... 114
Figure 5-6: Validation of predicted IL-25 at day 7 .......................................................... 116

All IPA network graphs ©2000-2011 Ingenuity Systems, Inc. All rights reserved.
List of Tables

Table 3-1: Cytokines in the Bio-Plex Human Cytokine 27-Plex Panel............................ 57
Table 3-2: Significance Sets of cytokines over time .......................................................... 61
Table 3-3: Molecular patterns of multiple organ failure...................................................... 69
Table 4-1: Incidence matrix for Figure 4-1........................................................................ 87
Table 4-2: Eleven cytokines were used for PSA edge analysis ........................................... 89
Table 5-1: Cytokines measured in Hemophilia A study.................................................... 107
Table 5-2: Evoked pathway molecule count...................................................................... 113
Chapter 1 Introduction

Background and Motivation

Clinical diagnosis, prognosis, and therapies for disease progression utilize established phenotypes and syndrome characterizations identified by the clinician based on patient data such as demographics, physiological observations over time, lab, radiology and microscopy reports, and transfusion and drug orders. Recent advances in the collection and assessment of molecular data offer the opportunity to use this data to increase understanding of disease progression, and, in the near future, to add this information to assist clinicians in patient care.

There are two major approaches to molecular pattern discovery. Quantitative methods produce lists of molecules that differentiate disease states based on biofluid or tissue analysis. Mass spectrometry is the most common technique for unbiased discovery where all molecular components within the capability of the equipment and its algorithms are identified (Rifai, Gillette, & Carr, 2006). Microarray immunoassay techniques are more sensitive and specific in molecular identification, but they only measure the concentrations of a predetermined panel of molecules (Jastrow et al., 2009). Both techniques have benefits and drawbacks for clinical usage, particularly in the analysis of human serum (Hoofnagle & Wener, 2009). Qualitative methods produce lists of molecules, molecular interactions, and biological pathways gathered from published literature and databases; techniques include manual and automated text mining and network analyses to uncover disease associations (Yang, Adelstein, & Kassis, 2009; Yang, Pospisil, Iyer, Adelstein, & Kassis, 2008; Yanliang, Yang, & Min, 2009). Factors
relevant to quality molecular pattern discovery include minimization of restrictions on initial molecule identification; measurement of biofluids likely to have molecules with significant sensitivity and specificity to a disease; identification of patterns of molecules and molecular interactions based on the measured biofluids; use of bioinformatics to connect identified molecules with published literature and databases; and support of computationally tractable algorithms for discovery (Good et al., 2007; Lescuyer, Hochstrasser, & Rabilloud, 2007; Rifai et al., 2006).

**Abundance of quantitative molecular data.** Rapid advances in lab technologies have made it easier and cheaper to measure vast quantities of molecular data in human blood and tissue. For example, multiplex molecular technologies now perform simultaneous measurements of millions of biological entities within the same assay. Not only is this faster than single measurements such as blots, performing the assays at the same time limits the environmental and operational variability that influences assay interpretation. The general public is now aware of the potential of molecular profiling. Early adopters can now contact the Biophysical Corporation in Austin, Texas for a $3,400 “biophysical” assay of more than 250 blood molecules considered to be “diagnostic biomarkers” associated with diseases or conditions (Biophysical, 2009).

Microarrays are a well-known multiplex technology that has been at the leading edge of biomedicine since the 1990s. The term microarray was originally used for the miniature DNA microarrays that measure thousands of specific DNA sequences from biological samples on glass slide chips or nano-well arrays. DNA microarray technology has been applied in many areas such as gene expression profiling to find which genes
change expression in response to disease and to compare genomes in different organisms (Khan et al., 1999; Tintle et al., 2008).

Next generation methods are manipulating and creating even more data: Illumina sequencing technology processes massively parallel sequencing of millions of DNA fragments (Illumina, 2009). The ability to simultaneously measure proteins, protein-protein interactions and protein-DNA interactions in tissue and biofluids is increasing. Multiple expression levels of cytokines – signaling proteins that play important roles in cell physiology and pathology - can now be simultaneously determined using enzyme-linked immunosorbent assay (ELISA)-based protein array technology. High density protein microarrays can now profile over 8000 proteins (Invitrogen, 2009).

Automated techniques are speeding up the processing of molecular data while decreasing the expense. Immunohistochemistry (IHC) localizes proteins in cells of tissue sections to identify molecules associated with cellular pathways and functions. IHC is used extensively to evaluate cancerous tumors, and has been traditionally scored by fluorescent in-situ hybridization (FISH), a sensitive but slow, complex and costly cytogenetic technique. New scoring methods, such as the automated cellular imaging system ACIS™, are now proving to be faster, less expensive, reliable and as accurate as FISH (Tawfik et al., 2006).

The technical ability to measure this deluge of quantitative molecular data is now about to move from the bioscience lab to clinical research for use in personalized medicine. For example, an integrated blood barcode assay chip is now in development for bedside use in clinical trials. It measures a large number of serum proteins within a few
minutes of a small sample (Fan et al., 2008). The challenge of new technologies is in understanding the biomedical significance of the huge amount of data generated (Hu, Coombes, Morris, & Baggerly, 2005; Rogers & Cambrosio, 2007).

**Abundance of qualitative pathway knowledge resources.** In addition to the quantitative data now measurable, there is an ever-increasing amount of qualitative data about molecular interactions. From the mid 1980s through 2008, PubMed gained 30,000 articles on signaling pathways and the advent of microarray technologies in the late 1990s spurred on signaling pathway research. More than 10,000 of the 30,000 PubMed signaling pathway articles relate to human cancer (NLM, 2008a), 347 articles to trauma, and 26 to multiple organ failure (NLM, 2008b). The under-representation in the latter two categories may be due to the fact that cancer studies examine molecular pathways in excised tissue, whereas trauma studies measure initiator signaling molecules in biofluids and infer the signaling pathways on an experiment by experiment basis.

Biological researchers, both experimental and theoretical, have organized the patterns of molecular interactions into spatio-temporal networks of pathways representative of cellular functions such as regulation of gene expression, metabolism and signaling (Slonim, 2002). Biological networks include measured molecules and interactions plus molecules and interactions inferred by computation or by similarities among organisms such as *Saccharomyces cerevisiae, Escherichia coli, Rattus norvegicus, Mus musculus, and Homo sapiens.*

Signal transduction pathways represent the cascades of reactions within or between cells that mediate cellular processes. They are the primary communication channels
within the organism to regulate physiology (Pawson & Nash, 2000) and they are usually drawn as networks with molecules as nodes and molecular interactions as edges. Once a network is in graphical form, it can be analyzed using graph theoretical methods from computer science, matrix algebra, and other mathematical constructs.

Signaling networks are inferred experimentally and/or computationally (Eungdamrong & Iyengar, 2004) from temporal and spatial patterns of molecules associated with specific intracellular functions such as apoptosis (Cho, Shin, Lee, & Wolkenhauer, 2003), gene activation and cell growth (Bhalla & Iyengar, 1999; Levchenko, 2003) or systemic functions such as inflammatory response (Calvano et al., 2005; E. Lin, Calvano, & Lowry, 2000; Salomao et al., 2008).

As technology advances, the ever-updated pathway information is disseminated through the internet, and there are numerous web-based commercial and academic knowledge bases of biological pathways. In 2001, only 18 pathway websites were active (Wixon, 2001). Today PathGuide.org references more than 290 pathway resources categorized by availability, data access methods, tools, organisms, network category, and contents. More than 30 million molecular interactions are accessible (PathGuide, 2010). The Pathway Database section of KEGG (Kyoto Encyclopedia of Genes and Genomes) includes networks relating to metabolism, genetic and environmental information processing, cellular processes, human diseases and drug development (KEGG, 2009). The Signal Transduction Knowledge Environment (STKE) lists 49 canonical pathways with 1084 component molecules and 33 specific pathways (specific to a particular organism,
tissue, or cell type) with 718 components (STKE, 2009). There is an abundance of accessible resources on biological pathways.

**Problem Statement**

The general problem is that there has been a dearth of methods that support data-driven, molecular based, clinical research in disease progression. First, patient molecular data is limited. There are usually few patients in a prospective observational non-cancer clinical study of disease progression. Measuring molecular patterns in bio-fluids or tissues is not a standard procedure in clinical care research, even in an Intensive Care Unit (ICU). This results in small sample sizes of patient data limited to non-parametric statistical analysis. Secondly, analysis methods for molecular patient data are uncommon, and time-oriented algorithms are even scarcer, as will be shown in Chapter 2.

The question is how to use the limited temporal molecular patient data to add more information that is available, but not easily accessible, from biochemical resources in literature and databases, and then, how to analyze that combined, larger data set over time and other stratifications to analyze disease progression. The challenge with resources, such as databases of molecular interactions, is analyzing the spatiotemporal interplay of uncountable numbers of molecular interactions within each cell and across the 100 trillion ($10^{14}$) cells in the human body.

The first step would be for research to uncover “gold standards” of molecular patterns associated with disease progression, followed by development of protocols and assays for bedside use.
The specific problem addressed by this study was how to connect quantitative spatiotemporal bioassay data of signaling molecules called cytokines with qualitative biological pathway information in order to uncover likely molecular patterns associated with systemic responses in disease progression.

Signaling pathways are initiated by biological entities outside the cell and they control cellular functions including proliferation, apoptosis, and necrosis. Signaling pathway initiators include cytokines, hormones and growth factors in the blood, lymph or interstitial tissue and biomechanical stimuli like tissue strain (Knobloch, Madhavan, Nam, Agarwal, & Agarwal, 2008; Lucitti et al., 2007; Morrow et al., 2007). When signaling molecules travel through the body and meet matching receptors on a cell, the functions of the respective signaling pathways in that cell are activated, modulated, or inhibited.

Because the measurable signaling initiators of the pathways appear before the inferred signaling pathway functions are executed, hypotheses can be generated to discover what this “advance notice” means – such as what are the underlying biological mechanisms or how treatment might direct the body system to mitigate pathways that go out-of-control. For example, initiators of pathway cellular functions can make signaling pathways compete or cooperate to destroy malformed cells, or suppress that destruction, resulting in tumors that may be cancerous. In trauma, the initiators activate a pro-inflammatory systemic response across many pathways to help the body fight immediate injury; however, if the “turn off” set of signals is not received in time by the pathways in the cells, the person can die. Since the initiators occur in the body ahead of the response, and initiating signaling molecules such as cytokines can be measured in serum, there
exists an “advance warning system.” Early knowledge of which pathways are active or inactive over time within a specific biological context can assist clinical decisions. This is of particular use in trauma where time is of the essence.

**Study Purpose, Scope and Deliverables**

The purpose of this study was to develop and document a computational method to support the research hypotheses that cellular functions are the foundation for physiological mechanisms, and that systemic patterns of measured molecules could be associated with larger biological pathways amenable to time-based analysis of disease progression.

The method was named the Pathway Semantics Algorithm (PSA) because it analyzes the meaning of biological pathways within clinical contexts. PSA addresses the need for a computationally tractable methodology to connect quantitative molecular data with qualitative pathways. PSA was designed to support translational systems biology and enable dimensionality reduction of data, statistically guided pathway selection, comparison of pathways across clinical conditions and over time via matrix algebra, and generation of clinical/biological hypotheses for wet lab testing or clinical trials. PSA was applied to analyze systemic responses in two biomedical domains, and validated through literature, expert opinion, and experiment.

**Study Significance**

The capability exists today for generating large amounts of quantitative bioassay data for patients in a variety of clinical conditions such as critical care and cancer. Simultaneously, availability of qualitative information on biological pathways is rapidly
increasing. However, there is a dearth of methods that can effectively connect these two
categories of data with a view to achieving a deeper understanding of the relationship
between evoked pathways and concentrations of initiating bio-molecules. This
information can yield valuable insights into underlying mechanisms of disease
progression and help formulate therapy.

The Pathway Semantics Algorithm, the subject of this research, is a step in this
integrative direction, following the viewpoint expressed by Simon Rosenfeld of the
National Cancer Institute that the first small steps toward translational systems biology
models “should be in the direction of integration rather than towards further elaboration
of individual processes and their in-depth mathematical modeling. In fact, the mass of the
knowledge currently available is so monstrously huge that it may have already passed the
point of being manageable. There is a serious risk of completely losing this knowledge
for any practical purpose unless decisive steps towards integration are
undertaken.”(Rosenfeld & Kapetanovic, 2008)

In chapter 2, the literature review confirmed that there are a limited number of
computational methods available for translational clinical research. One of the challenges
is the development of new methods based on theoretically sound and computationally
tractable techniques that can be scaled to an organism level.

The potential significance of this study is that it adds a novel methodology called the
Pathway Semantics Algorithm (PSA) that addresses these needs. In contrast to other
graph/matrix systems biology tools, PSA compares networks of patient data-driven
biological pathways over time or other stratifications at the organism level. Temporal
analysis is important and the time factor has been recognized as important in the specificity of signaling pathways (De Meyts et al., 1995). However, time-based pathway models at the molecular level generally consist of sets of differential equations with assumed kinetic constants that are computationally intractable at the organism level. PSA overcomes this limitation.

The methodology appears to be repeatable, generalizable, scalable and extendable. PSA expands patient data by incorporating biological pathway data into the mix, and then facilitates analysis of that data over time and other stratifications using numerical linear algebra to generate useful hypotheses that answer biological questions. PSA algebraically post-processes evoked pathway networks to reveal changing molecular patterns not easily observed in the static text and graphical formats output by biological pathway generation programs such as Ingenuity Pathway Analysis (www.ingenuity.com). The algebraic post-processing changes the data representation; this is important because the data representation space is one of the four inter-related problem spaces in scientific discovery, along with the hypothesis space, the experiment space, and the experimental paradigm. Changes in data representation uncover regularities and invariants, facilitate categorization, and suggest alternative search strategies key to scientific discovery (Schunn & Klahr, 1995).

PSA differs from graphical analysis since it does not start with predetermined graphs of canonical pathways. PSA starts with clinical data upon which biological pathway networks are constructed based on most likely interactions even if they are not part of canonical pathways. Finally, PSA uses mathematical algorithms for matrix representation
and computation that are readily available and can be implemented in a wide variety of software. A key advantage of PSA is that it narrows down the potential hypotheses for further investigation, thus reducing costly lab and clinical research efforts.

PSA can benefit clinical practice because it shows how the fundamental mathematics of numerical linear algebra can be used in a novel manner for comparative analysis of biological pathways in disease progression. Simple computations can be performed using spreadsheet calculations familiar to clinicians; more complex work for specific clinical contexts can be developed into software. Finally, PSA can benefit future research because it is a fundamental method that is adaptable to a wide range of studies on disease progression or comparison.

**Theoretical / conceptual framework**

The theoretical substruction map (Dulock & Holzemer, 1991; Hinshaw, 1979; McQuiston & Campbell, 1997; Trego, 2009; Wolf & Heinzer, 1999) for the research project is given in the Appendix, illustrating the linkages from the source study fields used through to the data analysis performed. The framework is as follows.

**Study fields.** The project draws on work in the fields of systems biology and clinical research.

**Theories.** From systems biology come theories of molecular interactions and algebraic analysis of networks. From clinical research come theories of disease progression.

**Models.** Models used include network representation of biological interactions, matrix representation of a network, and models of inflammatory and immune response.
Concepts. The major concepts included are molecular interactions, directed graphs, molecules, matrix algebra, changing patterns of molecules over time, and measures of disease progression.

Aims. The specific aims of the project were to:

• develop the PSA method for node analysis and edge analysis,
• apply PSA to a study of multiple organ failure in shock / trauma and to a study of immune response in hemophilia A, and
• validate PSA through literature search, expert opinion, and, if feasible, laboratory experiment.

Research questions. The research questions for each application were:

• Which molecules differentiate disease progression?
• Which molecular interactions differentiate disease progression?

Variables. Variables included measured molecules, inferred molecules, inferred molecular interactions, time, clinical outcomes or treatment effects.

Operational definitions.

• Measured molecules: cytokines
• Inferred molecules: genes, proteins, chemicals
• Inferred molecular interactions: molecule to molecule interactions
• Time: time in periods of hours or elapsed time in days
• Outcome: multiple organ failure or not
• Treatment effect: PBS versus CFA/I
**Data analysis.** Cytokines were measured in pg/ml. Inferred molecules and molecular interactions were obtained from a pathway database based on the cytokine measures. Time was measured in hours or days. The analysis was done on the stratification of outcome or treatment effect.

**Summary**

To date there appears to be no generalizable, computable systems-level methods that utilize spatiotemporal bioassay data to answer biomedical questions arising from comparative analysis of biological pathways. However, there is a need to connect bioassay data with pathway information within specific biomedical contexts and to facilitate comparison of biological pathways by time, outcome, molecular location and/or cell cycle phase. If these needs could be met, clinical research could start utilizing the wealth of constantly updated biological pathway information on a regular basis, and generate baseline hypotheses for mechanisms and recommended therapy – and keep a few steps ahead of the molecular data deluge that is about to impact clinical medicine.

The overarching goal of the Pathway Semantics Algorithm is to improve the *in silico* identification of clinically useful hypotheses about molecular patterns in disease progression. By framing biomedical questions within a variety of matrix representations, PSA has the flexibility to analyze combined quantitative and qualitative data over a wide range of stratifications. The resulting hypothetical answers can then move to *in vitro* and *in vivo* verification, research assay optimization, clinical validation, and commercialization. Herein PSA is shown to generate novel hypotheses about the significant biological pathways in two disease domains: shock / trauma and hemophilia.
A, and validated experimentally in the latter. The PSA matrix algebra approach identified differential molecular patterns in biological networks over time and outcome that would not be easily found through direct assays, literature or database searches.

The next chapter is a literature review of relevant computational methods. Chapters 3 and 4 describe PSA for node and edge analysis respectively, and apply the method to disease progression in shock / trauma. Chapter 5 demonstrates the application of PSA to hemophilia A and the validation with experimental results. The work is summarized in Chapter 6, followed by extensive references and an Appendix with additional material.
Chapter 2 Literature Survey

This chapter presents a survey of computational tools, techniques, and resources from systems biology that can be used to increase understanding of the physiological mechanisms in disease progression. The intention is to show data processing and analysis methods that may be adaptable from the molecular and cellular levels to investigations at the tissue, organ, and whole body system levels, highlighting common processes and procedures that may be useful for investigative studies and personalized medicine. The scope of this chapter is limited to resources for potential applications in shock trauma and critical care; however, the underlying methods may be useful in a wide range of translational clinical research in disease progression.

It is my contention that systems methods can assist in the identification of measurable characteristics in critically injured trauma patients that point to underlying disrupted biological mechanisms that may be amenable to treatment with a resultant increase in survival (Karvunidis, Mares, Thongboonkerd, & Matejovic, 2009; Neugebauer & Tjardes, 2004; Polpitiya, McDunn, Burykin, Ghosh, & Cobb, 2009; Vodovotz, Csete, Bartels, Chang, & An, 2008). The challenge is that systems biology analyzes physiology from the “bottom-up”, modeling molecules, organelles, and biological pathways within cells whereas clinical medicine treats a patient from the “top-down”, evaluating the whole body, based on observable measures from biofluids, tissues, and organs. The data from both extremes vary considerably over scales of time and space. For example, data can be baseline, measured at specific intervals, or measured continuously. Data can be numbers, words or patterns describing serum protein concentrations or heart rates.
There are vast databases with biochemical data and computational models that can offer insights into disease progression, if they can be linked to specific patterns of patient characteristics and if clinicians were enabled to judge clinically relevant factors that arose from systems biology. Specific treatments based on an individual’s underlying physiology, in addition to phenotype (Butte, 2008; Hofer et al., 2009; Salluh & Bozza, 2008), are important in the development of personalized therapies. Clearly a two-stage translational approach is required: first, clinical researchers need to identify “gold standard” data patterns that include measurable biochemical data associated with prognosis, diagnosis or treatment. Secondly, protocols and parameters that incorporate this information must be developed for clinical use. In both stages, computational methods are needed: first for discovery of likely biochemical patterns and disease associations, and secondly, to provide patient data-driven reports that include systems information to assist clinicians in assessing and directing patient care.

Because ICUs are already prepared to monitor and collect massive amounts of temporal physiological and clinical data, they are a likely candidate location for studies and applications of translational systems biology. In this chapter, I present a selection of recent approaches and their application to research in biological processes in trauma and critical care, such as inflammatory, immune and injury responses. The emphasis is on computational methods that can be used for data-driven systems analysis of disease progression.

In the following section, I give a short overview of systems analysis and systems biology, approaches used in trauma research, and data available. In Section 3
computational methods are discussed, followed by Section 4 with selected applications. The literature review is summarized in Section 5.

**Systems Analysis and Systems Biology**

Systems analysis is “a method of describing and understanding complex interactions among large numbers of processes or components in a generalized way. The focus is on identifying the fundamental units of a system and defining how they interact rather than the internal processes of each unit” (Aber & Melillo, 2001; Yourdon, 1988). Systems analysis can be performed to generate or test hypotheses about the systems behavior within specific assumptions and constraints. Analysis techniques may be qualitative or quantitative, static or dynamic, stochastic or deterministic, or combinations. The fundamental units (components or processes) may be nested within a hierarchy or overlapping.

Systems biology, a subcategory of computational biology, is defined by the National Library of Medicine (NLM) as the “comprehensive, methodical analysis of complex biological systems by monitoring responses to perturbations of biological processes and using the large scale, computerized collection and analysis of the data to develop and test models of biological systems” (NLM, 2009). From its beginnings, systems biology aimed at building mathematical frameworks with some predictive abilities based on systematic organization of genomic and proteomic data (Aggarwal & Lee, 2003). Since then, the scope of systems biology has expanded and spawned a number of related offshoots such as translational research (NIH, 2009), translational systems biology (G. An, Faeder, & Vodovotz, 2008), translational bioinformatics (AMIA, 2006), and systems
Although analysis goals, abstraction levels, and scales vary widely, the fundamental units under study are usually molecules, cells, tissues, organs, and organisms within a hierarchical framework with modular control elements or related biological processes (G. C. An, 2010; M S Iyengar, Brown, & McGuire, 2007; Lauffenburger, 2000; Malhotra et al., 2008).

**Systems approaches in trauma**

Trauma refers to serious bodily injury, which, if of sufficient magnitude, may be accompanied by initiation of the systemic inflammatory response. Causes of trauma include penetrating injuries from gunshots and stab wounds, blunt injuries, such as those sustained during automotive accidents, and burns. In addition to direct tissue damage, trauma can result in injury to remote organs due to disruptions in normal physiology and underlying protective biological mechanisms. These remote injuries can be rapid in onset and potentially fatal if allowed to proceed unabated. Moreover, trauma is the leading cause of mortality in the US among individuals under 45 years of age, and the cause of death for 74% of all deaths for people ages 15-24 (Heron, Hoyert, Xu, Scott, & Tejada-Vera, 2008). In critically ill patients, normal biological processes are disrupted but the associated pathophysiology is incompletely understood (Deitch, 1992; Maier et al., 2007; Wan et al., 2008).

Within the past decade, a number of systems approaches for analysis of trauma and critical illness have been developed (Buchman, 2009; Vodovotz et al., 2007). Computational methods have been used to increase understanding of systemic functions...
such as inflammation and immune response (G. An, 2001; Dong, Foteinou, Calvano, Lowry, & Androulakis, 2010; N. Y. Li et al., 2008; Ta'asan & Gandlin, 2009; Vasilescu, Buttenschoen, Olteanu, & Flondor, 2007; Vodovotz et al., 2009) and the effects of drug dosing (Yamamura et al., 2004), as well as organ specific issues such as heart rate complexity (Cancio et al., 2008; Riordan, Norris, Jenkins, & Morris, 2009) and acute lung injury (Ware et al., 2009). Multiscale computational models of angiogenesis, from the molecular to the organ system levels, have been integrated to improve predictive capabilities (Qutub, Mac Gabhann, Karagiannis, Vempati, & Popel, 2009). At the molecular/cellular systems level, there are numerous computational approaches in systems biology used to study biological mechanisms such as signaling (Rangamani & Iyengar, 2008), metabolism (Palsson, Joshi, & Ozturk, 1987), and protein interactions (C. Y. Lin et al., 2008) that underlie disease progression. With the advent of new technologies that make it feasible—and soon cost-effective—to capture patient’s molecular data such as mRNA expression or serum protein concentrations, translational clinical research can benefit from using computational approaches beyond classical statistical inference. A systems-wide analysis of data from the molecular to the organism level can help design evidence-based personalized therapies.

Data for systems approaches in trauma

The complex and often rapid progressions of shock trauma and critical illness provide a vast quantity of patient data that can be collected and evaluated through real-time monitoring in an intensive care unit (ICU) on a continuous, hourly, or daily basis. Intensive care units collect one item of documented clinical information per patient each
minute (Manor-Shulman, Beyene, Frndova, & Parshuram, 2008). In addition to monitored data, patient data includes transfusion and drug orders, microscopy, radiology and laboratory reports, nursing and clinician notes, and patient demographics. As bedside biofluid measurement devices move from prototype to practicality (Fan et al., 2008; Sorger, 2008), a patient’s molecular data such as serum proteins can also be collected in time to be of use in the ICU; currently, turnaround time for molecular assays is not practical for other than research use. The challenge today is to understand the meaning of all this data in terms of disease progression, and develop data-driven protocols that will be in place when the technology is available. For example, research has shown that specific patterns of cytokine molecules over time are associated with trauma progression (Jastrow et al., 2009; Maier et al., 2007; Roumen et al., 1993). Cascade patterns of molecular interactions, such as those triggered by cytokines, have been identified as biological pathways – spatiotemporal networks representative of cellular functions that regulate gene expression, metabolism and signaling (Slonim, 2002). Because cytokines drive signaling in biological pathways, adding cytokine data may provide insight into the underlying biological mechanisms.

There are an ever-increasing number of databases with information about biological pathways. In 2001, only 18 pathway websites were active (Wixon, 2001). Today, PathGuide.org references more than 290 pathway resources categorized by availability, data access methods, tools, organisms, network category, and contents. More than 30 million molecular interactions are accessible via the Internet (PathGuide, 2010). The Pathway Database section of KEGG (Kyoto Encyclopedia of Genes and Genomes)
includes networks relating to metabolism, genetic and environmental information processing, cellular processes, human diseases and drug development (KEGG, 2009). More than 1400 curated, experimentally determined, metabolic pathways and enzyme data for microbial, plant, and vertebrate metabolism are available from the freely accessible MetaCyc database (Caspi et al., 2010). There are commercial and publicly available databases of molecular interactions (Tarcea et al., 2009), biological pathways (Elliott et al., 2008; Viswanathan et al., 2008; Wixon, 2001), and genomic correlates (L. T. Sam et al., 2009). The Signal Transduction Knowledge Environment (STKE) lists 49 canonical pathways with 1084 component molecules and 33 pathways specific to a particular organism, tissue, or cell type with 718 components (STKE, 2009). PubMed lists more than 250,000 articles with content about signal transduction pathways; the earliest articles are from 1947 – before systems biology as such existed (Baumgardt, 1947; Berliner, 1947; Monnier & Boehm, 1947).

The question is how to integrate all this data? One approach is to use computational methods from systems biology to connect patient data with data from basic science resources in biology, chemistry and physics to develop research models that can transition to data-driven clinical practice. See Figure 2-1. Even with computational approaches, the data translations and transformations among levels from molecule to organism and vice-versa are far from seamless. Most applications cobble together several methods to achieve their research goals. In the next section, I review some of the major computational methods that have been used to analyze biological processes related to trauma and critical illness; this is followed by Section 4, giving details of several
applications. The intent is to inspire creative use of methods and data in the investigation of trauma and critical illness, with the goal of improving patient care.

![Diagram of computational methods integrating translational research]

**Computational Methods**

The biological processes in shock trauma and critical illness are complex and unstable. There are simultaneous and rapid changes of biological pathways across and within the entire body. Extracellular and intracellular signaling modulates systems-wide mechanisms such as inflammatory response (Levi, Keller, Van Gorp, & Ten Cate, 2003; Pillay, Hietbrink, Koenderman, & Leenen, 2007; Rezende-Neto et al., 2002), sepsis, hemorrhagic shock, and resuscitation from hemorrhagic shock (Rittirsch, Flierl, & Ward,
The choice of systems analysis and computational methods depends on several factors:

- The systems level(s) under study, from molecule to organism;
- The available data; and
- Which biological processes are under study, within what context, and for what goals.

Hypotheses about disease progression can be generated computationally in many ways: from data-driven model-free discovery to the perturbation of *in silico* models of biological processes. This section is an overview of common computational methods in use plus some general considerations for data; selected applications related to trauma and critical care will be shown in Section 4. Here I first present basic probabilistic and deterministic approaches that utilize a wide variety of fundamental tools and techniques that can be used individually, combined, or in combination with other methods. This is followed by a selection of more specialized methods.

**Basic probabilistic approaches**

*Classical Statistical Inference* incorporates no prior information and assumes independent variables; the approach is used at all systems levels and underlies the primary tools, such as Student’s *t* test, used for static analysis of injury response where there is sufficient data. In contrast, *Bayesian Statistical Inference* incorporates prior information and handles interdependent variables. The Bayesian “conditional probability” approach is becoming more and more widely used in genetic data analysis (Beaumont & Rannala, 2004), clinical research (Moyé, 2008) and diagnostic
medicine; complex Bayesian analyses are usually performed using Markov Chain Monte Carlo (MCMC) computational methods (Broemeling, 2007). MCMC methods use Monte Carlo random sampling to produce a Markov Chain with state transitions that converge to an invariant distribution. A Markov Chain is the simplest autonomous form of a discrete-time probabilistic state-transition Markov model where the system state is observable.


**Basic deterministic approaches**

Deterministic approaches depend on initial states and chosen parameters. *Differential equations* are the primary methods of deterministic dynamic analysis, and are mostly used at the molecular and cellular levels because they are computationally intensive at higher levels. For example, modeling one NFκB signaling pathway in one cell activated by one signaling TNF-α molecule requires 18 nonlinear differential equations, with 33 independent variables and 16 dependent variables in a simplified reaction kinetics model (Cho et al., 2003); scaling this method directly to the organism level is computationally intractable. Ordinary differential equations (ODEs) model dynamic changes in items, such as protein concentrations, over one independent variable whereas partial differential equations model simultaneous changes over two or more independent variables. Explicit equations are used, usually with equilibria or other constraint assumptions. In addition to experimental data, the equations require data for estimated
biochemical kinetic parameters, which are usually inferred from published results. Differential equations can be solved using standard mathematical software available as open source or commercial software such as MATLAB (MathWorks, 2010) and Mathematica (Wolfram, 2010).

*Matrix algebra* can be applied from molecular to organism levels. Stoichiometric matrices are used for flux-balance analysis (FBA) of metabolic biochemical reaction networks (Palsson et al., 1987; Schilling & Palsson, 1998) to stochastically simulate chemical kinetics. Unlike differential equation approaches, FBA does not require reaction rate kinetic parameters or metabolite concentration data. Instead, the key assumptions are that the system is homeostatic with a balanced system of energy production and consumption and that the metabolites are “well stirred” so that Gillespie’s Algorithm can be used (Gillespie, 2008). This steady-state approximation of cellular dynamics can offer insights into multiscale snapshots of disease progression. Matrix algebra formalisms have been used to study signaling and regulatory pathways using extreme pathway analysis, an adaptation of the stoichiometric approach used for metabolic analysis (Gianchandani, Papin, Price, Joyce, & Palsson, 2006; Papin & Palsson, 2004) and to generate signaling networks from sparse time series of observed data (Allen et al., 2007). The latter computational algebra approach has potential for analysis of signaling pathways in disease progression.

*Matrix decomposition* methods are the basis for a wide variety of factor and component analyses in data mining and graphical analyses (Skillicorn, 2007; Sun, Xie, Zhang, & Faloutsos, 2008). In addition to techniques such as singular value
decomposition (SVD), new matrix approaches are evolving such as the graph-decorrelation algorithm (GraDe) that performs detailed temporal analyses on large-scale biological data using knowledge-based matrix factorization. In a recent time-course microarray experiment of mouse hepatocytes, GraDe provided a detailed separation of the time-dependent responses to IL-6 stimulation compared to standard methods (Kowarsch et al., 2010).

Matrix algebra can be performed using software as simple as spreadsheets; more complex calculations use software such as MATLAB or Mathematica. Code for Gillespie’s Algorithm is available for R (http://cran.r-project.org/web/packages/GillespieSSA/index.html) and for Mathematica (http://demonstrations.wolfram.com/DeterministicVersusStochasticChemicalKinetics/).

**Graphical approaches**

Cascades of molecular interactions can be represented as directed graphs in order to use computational methods from graph theory to explore pathways within the graph. The analysis is usually at the molecular and cellular levels, although the methods can be adapted for higher levels. Biological pathways can be abstracted as network graphs with nodes representing molecules and edges being molecular interactions (Alon, 2007; Ma'ayan, 2008). Patterns of molecules, such as serum cytokines, have been associated with disease progression in trauma, and graph theory methods offer ways to analyze this data. *Graph theory* is supported by extensive computational methods from mathematics and computer science that are used for analysis of static and dynamic systems ranging
from computer systems to social networks. Structural properties of the graph can be measured in many ways such as counting the number of nodes and edges, number of edges per node or nodes per edge, identifying primary hubs and sub-network motifs. Computational methods are usually analysis specific. For example, the web-based Hub Objects Analyzer (Hubba-Hubba, http://hub.iis.sinica.edu.tw/Hubba/index.php) identifies essential hubs in a protein interaction network by using a combination of software including databases, graph generators, and topology calculators(C. Y. Lin et al., 2008).

A **Bayesian network** is a probabilistic graphical model constructed as a directed acyclic graph (DAG) with nodes representing variables and edges representing the conditional dependencies between the nodes. Bayesian networks are used for process modeling and diagnostic reasoning(Darwiche, 2009; Koller & Friedman, 2009). One class of Bayesian networks is based on **Hidden Markov Models (HMMs)** - Markov Chains with hidden rather than visible states but with visible state-dependent outputs. HMMs can be used to uncover an optimal sequence of state transitions. One limitation of Bayesian networks is that they must satisfy the local Markov property - each node is conditionally independent of its non-descendents(Russell & Norvig, 2009); as a result, graphs with cycles are not supported. This limits modeling of biological pathways to small sections without loops. Recently, an extension to Bayesian network models, called **Generalized Bayesian Networks (GBN)**, has been proposed that can model cyclic networks for use in translational systems biology(Sachs, Itani et al., 2009). There are a number of software packages for Bayesian networks including the Python library Pebl (http://code.google.com/p/pebl-project/) and Hugin (www.hugin.com).
Petri Net methods have been used to analyze Bayesian networks where the nodes are molecules and the edges represent the dependencies of the interactions between the nodes. Petri nets perform qualitative, stochastic and continuous analysis of small biochemical networks by modeling token-based transitions, such as reactions, between “places” such as proteins. Dynamic modeling is performed by incorporating differential equations to assign rate functions to transitions (Heiner, Donaldson, & Gilbert, 2010).

Petri Net Toolboxes are available for MATLAB and Mathematica, and systems biology-oriented Petri Net software called Snoopy is freely available (http://www-dssz.informatik.tu-cottbus.de/snoopy.html; www.informatik.uni-hamburg.de/TGI/PetriNets).

Finally, Spectral Graph Theory incorporates both graph theory and matrix algebra to examine a network in terms of the eigenvalues and eigenvectors (spectrum) of the adjacency matrix mapped from the network graph (Cvetković, Doob, Sachs, & Torgasev, 1988). This method has been used to compare basic metabolic networks at the systems level in three organisms Mycobacterium tuberculosis, Mycobacterium leprae and Escherichia coli. The results found that the most highly connected biochemical reactions in an organism are not necessarily those most central to the organism’s metabolism, suggesting that hubs present in mycobacterial networks that are absent in the human metabolome may be potential drug targets (Verkhedkar, Raman, Chandra, & Vishveshwara, 2007).

Pathway databases use graph theory with published biological pathway data and proprietary computational network analysis algorithms to generate specific biological
pathways. For example, Biobase (BIOBASE GmbH, Germany; www.biobase-international.com) has a data analysis system called ExPlain, and the Ingenuity Knowledge Base (Ingenuity Systems, US; www.ingenuity.com) supports Ingenuity Pathway Analysis. Advantages of these combined database / algorithm systems are that the pathway / molecular interaction data are kept up-to-date, and that the algorithm is specifically designed to work well with the database to uncover the pathways associated with the input data. Although this is advantageous for the general user, it must be noted that the underlying computational methods are not amenable to modification because they are usually based on proprietary algorithms with limited documentation. In addition, access to commercial web-based pathway databases and their analysis software is by paid subscription.

Symbolic approaches

Symbolic logic is a formal qualitative modeling approach used to answer questions at various levels of abstraction. The questions usually focus on a specific intracellular function such as signaling and a model is created based on system states and rules for state changes. Symbolic models can be analyzed or run as simulations; models can be formally checked and verified. A wide variety of computational methods for symbolic systems biology have been developed (M. S. Iyengar, 2010). There are several implementations of rule-based modeling for signaling networks (Hlavacek et al., 2006) such as Pathway Logic, a symbolic rewriting logic based on pi-calculus (Knapp et al., 2005; Talcott, 2006).
Artificial Neural Networks (ANN) and Agent-Based Models (ABM)

ANN and ABM methods are used with organism-level data. However, both are computationally intensive and may require specialized software along with multiprocessor hardware. The artificial neural network (ANN) is a computational method to uncover nonlinear patterns based on input data. Several ANN models are generated from the training input data, and the one with the best fit between predicted and observed values is considered the optimal ANN model to be used for the actual data and predictions. Optimization techniques, such as the conjugate gradient decent method (Fletcher, 2000; Mary F. McGuire & Wolfe, 1973), may be used to optimize the model. Nonlinear ANN modeling has been shown to be comparable to linear logistic regression analyses when sample size is adequate. However, it has been shown that training sets for ANN need at least 800 observations to generate an adequate model – a sample size not usually found in ICU trauma / critical care studies (Clermont, Angus, DiRusso, Griffin, & Linde-Zwirble, 2001). Standard mathematical and statistical software including MATLAB, Mathematica, SPSS and SAS have built-in algorithms or add-on modules for neural network analysis and optimization.

Agent based simulation consists of an agent-based model (ABM) composed of autonomous fundamental units, or agents, defined at multiple scales or levels within a system and the rules that govern the state change interactions among them. The rules may be deterministic or stochastic. No explicit equations are used and there are no equilibria assumptions as in most models. The goal is to predict patterns of emergent behaviors that arise in complex systems from simple rules (G. C. An, 2010). The model must be verified.
and validated in some way; simulations must be run many times to uncover relevant patterns. Two open-source software packages for ABM development are NetLogo (Wilensky, 2010) and SeSAm (Würzburg, 2010). FLAME (Flexible Large-Scale Agent Modelling Environment, www.flame.ac.uk) (Coakley, 2007; Richmond, Walker, Coakley, & Romano, 2010) is a formal framework that allows a wide range of agent and non-agent models to work together within one simulation environment.

Applications

In this section, I describe computational approaches currently used, or that have the potential for use, in critical care and trauma-related research. The applications are organized by research goals at levels from the organism level down to the cellular/molecular level. A short paragraph summarizes the goal, processes and context for the example, followed by a list of the methods and data used. See Figure 2-2.
Figure 2-2 Computational analysis at different levels within an organism
Organism level

The organism level includes research performed at the molecular or cellular levels that investigates systemic processes such as inflammation, immune and injury responses. In the next two subsections, I show example of process models and predictive models.

Organism level: Process models

The inflammatory process is a normal physiological response in acute and chronic disease, and part of the immune response to infection. However, despite numerous computational models, much work still needs to be done to automate integration of these models with data across system levels with software usable by non-mathematicians (Vodovotz et al., 2009).

Abstraction. One of the earliest agent based models was developed by An (G. An, 2001) to create a simple abstraction to simulate the nonlinear behavior and dynamic structure of the inflammatory response. Although the model was based at the cellular level, the abstraction was used for inference of the systemic response at the organism level.

- Method: Agent Based Model using StarLogo software. (StarLogo is now available as open source OpenStarLogo at http://education.mit.edu/starlogo/)
- Data: Abstractions of three cell types used as agents: endothelial cells (with injury states), neutrophils, and circulating mononuclear cells, plus rules for agent interactions.

Challenge / Response. Endotoxin (LPS) and other challenges have long been used in shock trauma research (Foteinou, Calvano, Lowry, & Androulakis, 2008; Waage,
Brandtzaeg, Halstensen, Kierulf, & Espevik, 1989; Webster & Galley, 2009) to probe challenge/response relationships.

Dong (Dong et al., 2010) created an agent based simulation to model the host response to endotoxin using the molecular interactions involved in the NFκB signaling pathway, coupled with the spatial orientation of various inflammation specific molecules and cell populations such as macrophages and T-helper cells.

- **Method:** Agent Based Model using NetLogo software (http://ccl.northwestern.edu/netlogo/)
- **Data:** Gene expression data from human subjects injected with endotoxin or a placebo. Biological data for agents (macrophages, cells and molecules) and agent rules (interaction behavior and rates).

In contrast, Vasilescu (Vasilescu et al., 2007) developed an equation based model to evaluate whether endotoxin (LPS) tolerance is a component of the immune dysregulation in patients with trauma, severe acute pancreatitis, and diffuse peritonitis.

- **Method:** Differential equations
- **Data:** Endotoxin levels, TNF-α plasma levels, and TNF-α releasing capacity of the whole blood in patients with severe acute pancreatitis, diffuse peritonitis, and trauma.

Muller (Muller & Tjardes, 2003) found bistability in the early inflammatory response by using an in vitro model of IL-1 challenge to derive an equation based in vivo model. The in vitro model was first developed by challenging endothelial cells with IL-1; then, the expected value of IL-6 at a specific time under a specific challenge was derived. The
basic mechanism of the *in vitro* model was expanded to a whole animal IL-1 challenge that modeled *in vivo* multistate inflammatory response. Of interest was the outcome that a small challenge did not lead to a response; however, a challenge above a certain threshold completely activated the endothelial cells.

- **Method:** Differential equations
- **Data:** IL-1 challenge levels and resulting IL-6 production levels in endothelial cells over time scales of minutes, hours and days.

Guthke (Guthke, Moller, Hoffmann, Thies, & Topfer, 2005) generated plausible models of the gene regulatory networks involved in the human peripheral blood mononuclear cells’ immune response to an *Escherichia coli* infection challenge. The immune interaction networks were reconstructed by reverse engineering. First, a statistical cluster analysis of the scaled time profiles of the gene expression data was performed using the fuzzy C-means (FCM) algorithm (Bezdek, Keller, Krisnapuram, & Pal, 2005), and then expression profiles of the representative genes were used to drive three dynamic models of gene regulatory networks based on *linear differential equations*, systems of *linear algebraic equations*, or *heuristic search* strategies.

- **Method:** Statistics, differential equations, linear algebra, heuristic search
- **Data:** Log-ratios of the expression intensities of more than 18,000 genes in peripheral blood mononuclear cells at five time points before and after infection by heat-killed pathogenic *Escherichia coli*. 
Organism level: Predictive models

Probabilistic methods are used extensively in clinical research. Among the more common algorithms are the parametric Student’s t test for normally distributed quantitative variables, the non-parametric Mann-Whitney test for quantitative variables without a normal distribution, and chi-square tests for qualitative variables. For example, these methods were used by Pidcoke (Pidcoke et al., 2007) to demonstrate that the diurnal patterns of blood glucose and insulin requirements in burn ICU patients are similar to those in healthy subjects.

- Method: Means, frequency analysis, simple and cosine regressions, Student’s t test, Mann-Whitney test, chi-square test
- Data: From 156 burn patients: total body surface area burned, injury severity score, polytrauma, age, gender, inhalation injury, glucose level (hourly), insulin dose (hourly), outcomes.

Cohen (Cohen et al., 2010) used hierarchical clustering to identify patterns of patients’ changing physiological states that were predictive of infection, multiple organ failure and mortality. Clustering is a multidimensional analysis that identifies groups of similar variables, with the results displayed as a dendogram tree structure. Limitations are that a variable may belong to only one cluster group, and the number of clusters may influence the result.

- Method: Hierarchical clustering, linear discriminant analysis, correlations
- Data: 45 measures of physiological, clinical, and treatment data were collected every minute from each of 17 severely injured trauma patients. Data for the
cluster analysis: continuous heart monitor, ventilator, and microdialysis data over 24–72 hours (52,000 data points).

Using a classical statistical model, Ware (Ware et al., 2009) identified a combination of biologic and clinical markers in patient data that predicted acute lung injury and acute respiratory distress syndrome.

- Data: Retrospective study from NHLBI ARDS randomized controlled trial: patient baseline plasma measures of IL-6, IL-8, TNFR1, von Willebrand factor, surfactant protein D, sICAM-1, PAI-1, protein C plus baseline clinical variables such as age, cause of ALI/ARDS and APACHE III scores.

In contrast, Peelen (Peelen et al., 2010) constructed three Markov models based on clinical data to gain insights into the probabilistic state transitions in organ failure progression in successive days of ICU stay. Peelen’s models identified potential clinical patient states (number and type of failing organ systems) along with the probabilities that each state would be followed by another state, or persist over time.

- Method: Markov models with dimensionality reduction via additive logistic regression; implementation by hierarchical dynamic Bayesian networks; followed by stochastic simulations.
• Data: Temporal clinical data from a prospective study of severe sepsis patients. Patient data included SOFA scores per each of six organ systems plus total SOFA scores.

*Artificial neural networks* (ANNs) were constructed by Yamamura (Yamamura et al., 2004) to predict the plasma concentration of Arbekacin sulfate, an aminoglycoside, in burn patients and, from that prediction, identify patients whose Arbekacin sulfate antibiotic would be sub-therapeutic based on the patients’ physiological data. ANN results were superior to multivariate logistic regression analysis in classifying patients’ outcomes.

• Method: Three-layered ANN model (Statistica software, [www.statsoft.com](http://www.statsoft.com)). Conjugate gradient decent method for optimization during ANN training with training data. Leave-one-out cross-validation of predictive performance with test data. Multivariate logistic regression analysis (SPSS, JMP(SAS) and Statistica software).

• Data: Clinical physiological data from 30 burn patients, plus data for assessing burn severity. Training data for ANN model: dose, body mass index (BMI), parenteral fluid, creatinine concentration and burn severity parameters.

Organ level

**Heart.** Using *Multiscale Entropy* (MSE) to assess Heart Rate Complexity (HRC), Riordan (Riordan et al., 2009) found that early loss of HRC was predictive of mortality regardless of anatomic location, severity or mechanism of injury.

• Data: MSE; continuous physiological data from the first available 6 hours plus clinical data and demographics from 2718 trauma patients.

Although HRC seems to be a useful mortality predictor in trauma, most HRC measures require a traditional 800-beat data set. In an emergency situation, such as a battlefield, this large amount of data presents a monitoring challenge. Using three entropy measures with data sets as small as 100 beats to assess HRC, Batchinsky (Batchinsky et al., 2009) found that HRC was decreased in pre-hospital trauma patients who died.

• Method: HRC assessed by approximate entropy, sample entropy and similarity of distributions. Statistics (SAS): Student’s t test, Mann-Whitney U test, logistic regression, Receiver Operator Curves (ROCs), odds ratios, maximum likelihood, Pearson chi-square.

• Data: EKGs with 800 RRIs from 31 pre-hospital trauma patients, with data sets sampled at 800, 600, 400, 200, and 100-beat data sets.

Brain. Numerous computational models have been developed to increase understanding of traumatic brain injury resulting from blast survivability in war zones with the goal of improving design of personal protective equipment. Moore (D. F. Moore et al., 2009) used equation based models to study the effects of primary blasts on the central nervous system, and found that blast waves directly propagate into the brain and
that stresses develop in central nervous system tissues comparable to those significant in mTBI or concussive injuries in sports.


- **Data**: Peak blasts at two pressure levels; impact deceleration. High-resolution T1 MR images.

  **Tissue level**

Adra (Adra, Sun, MacNeil, Holcombe, & Smallwood, 2010) developed a multiscale 3D model of the human epidermis to explore the functions of TGF-β1, a potent growth factor, during epidermal wound healing. A computational virtual epidermis was created using an integrated agent/COPASI model, followed by investigation of several hypotheses, including the changes in epidermal wound healing associated with different wound sizes.

- **Method**: COPASI (COmplex PAthway SImulator) (Hoops et al., 2006) ordinary differential equations model for sub-cellular TGF-β1 functions, linked to a cellular agent based model of normal human keratinocytes (NHKs) in FLAME ([www.flame.ac.uk](http://www.flame.ac.uk)), linked to a multi-cellular layer with a mathematical solver that resolved physical issues.
• Data: Chemical reactions and coefficient factors of TGF-β1 expression and signaling; biological rules for the behavior of normal human keratinocytes (NHK) when subjected to injury signals.

Wound healing research has also been used to investigate inflammatory response at the tissue level. The pathogenesis of vocal fold scarring in humans is not well understood despite extensive experimental and clinical temporal data from animal studies. Li (N. Y. Li et al., 2008) developed an agent-based simulation to model the patient-specific vocal fold inflammation and wound healing following acute phonotrauma.

• Method: Agent Based Model using NetLogo software
• Data: IL-1β, IL-6, IL-8, TNF-α, matrix metalloproteinase (MMP)-8, and IL-10 from 4 samples of laryngeal secretions from 9 human subjects.

**Cellular/molecular levels**

Cellular and molecular level approaches offer novel avenues for investigative research into disease progression. Computational biology has developed a wide variety of methods to model cells, molecular interactions in the form of biological pathways, and molecules at varying levels of abstraction (Mary F. McGuire & Iyengar, 2010), along with extensive databases of results containing inferred and experimentally validated data. The challenge is to how to modify these methods and use these models and data for clinical insights into disease progression.

During the past ten years, signaling pathways have become the cornerstone of cancer research (Dy & Adjei, 2008). Signaling pathways are the primary multilevel, multiscale communication channels within the organism that regulate physiology (Pawson & Nash,
2000; Zhang, Ramesh, Uematsu, Akira, & Reeves, 2008); they clearly play important roles in disease progression. Cellular signaling pathways are initiated by extra-cellular biological entities including cytokine signaling molecules, hormones and growth factors in the blood, lymph or interstitial tissue, and biomechanical stimuli such as tissue strain (Knobloch et al., 2008; Lucitti et al., 2007; Morrow et al., 2007). Because signaling triggers can be measured noninvasively in biofluids such as serum, urine or saliva, they may be useful for monitoring the rapid disease progression found in trauma and critical care. Specific patterns of signaling molecules such as cytokines have been associated with mortality in septic shock (Waage et al., 1989), critical illness (Roche & Gussler, 1992), trauma (Roumen et al., 1993), and multiple organ failure (Jastrow et al., 2009). In trauma, the signaling molecules activate a pro-inflammatory systemic response (Oberholzer, Oberholzer, & Moldawer, 2000) across many pathways to help the body fight immediate injury; however, if the “turn off” set of signals is not received in time by the pathways in the cells – or a compensatory systemic response is too much or too little – death may ensue (Adib-Conquy & Cavaillon, 2009; E. E. Moore et al., 2005).

Molecular signaling profiles may be one of the keys to personalized medicine; however, there is still much work to be done to make them clinically relevant. Sachs (Sachs, Itani et al., 2009) created an algorithm to extend acyclic Bayesian network theory to permit loops, or cycles, in networks. The resulting Generalized Bayesian Network (GBN) is a Bayesian network model of nodes representing molecular data, augmented by state nodes and edges representing the statistical dependencies among the nodes. As a proof of principle, Sachs used GBN to characterize disease states and patient-
specific signaling profiles in human follicular lymphoma tumors following B-cell antigen receptor stimulation (Sachs, Gentles et al., 2009). The results showed differences in comparably diagnosed patients that might influence individual prognosis and therapy.

- **Method**: Generalized Bayesian Networks (GBN): Bayesian Networks model of the phospho-protein signaling pathway augmented with state nodes for patient and disease
- **Data**: Flow cytometry measures of six phospho-protein levels (SYK, ERK, p38, CBL, SFK, BTK) before and after B-cell antigen receptor signaling. Patient state and disease state.

Another research question is how signaling events trigger cellular responses. Signaling pathways that lead to apoptosis are of particular interest in disease progression. Using a data-driven computational technique called Model-Breakpoint Analysis, Janes (Janes, Reinhardt, & Yaffe, 2008) found that the dynamic range of the molecular signals had a greater influence on predicting cytokine-induced apoptotic cellular response than either basal or maximally inducible signal strength; results were validated experimentally. The results suggest that changes in dynamic range, due to subtle molecular amino-acid changes from disease mutations, could lead to pathophysiology.

- **Method**: Partial least-squares regression, principal component analysis, equation-based model breakpoint analysis.
- **Data**: Model of cytokine-induced apoptosis based on 7,980 measurements of molecular signals that are activated by combinations of the death stimulus, tumor
necrosis factor (TNF), together with a survival stimuli of epidermal growth factor (EGF) or insulin.

Petri nets can represent signal transduction networks at varying abstraction levels using graphs with molecules for nodes, edges for transitions, and tokens generated by the transitions. Petri net models can be constructed from limited knowledge of the pathway behavior, with ambiguities resolved through subsequent model validation. Qualitative Petri net models can be extended to quantitative stochastic or continuous Petri net models by the addition of rate equations (Heiner et al., 2010). Simulations of stochastic models can be run dynamically, and ordinary differential equation solvers can run static deterministic analyses of continuous Petri net models. Heiner (Heiner, Koch, & Will, 2004) developed and validated a qualitative Petri net model of apoptotic pathways, using formal computer science methods to represent pathway structure and behavior. Although not directly linked to clinical data, Heiner’s Petri net process model could be perturbed to gain insights into apoptosis not easily seen in other representations.

• Method: Step-wise incremental Petri net modeling with repeated analyses. Linear algebra using the incidence matrix and transition vector from the network graph.

• Data: A published schematic overview of apoptosis, comprising both extrinsic and intrinsic pathways, induced by DNA damaging and Fas signals, resulting in DNA fragmentation combined with a Fas-induced MAPK (mitogen-activated protein kinase) pathway and TNFR-1 receptor-induced pathways. Apoptosis inhibitors were not taken into account.
Summary

Research in trauma and critical illness is especially challenging because the effects of the original insult can be widespread across the entire body, affecting multiple organ systems. Disease progression is typically rapid, measured in hours and sometimes in minutes. At present there are growing capabilities to collect vast amounts of temporally indexed quantitative and qualitative data at multiple levels, from concentrations of biomolecules to sophisticated imaging modalities. These capabilities have the potential to support translational “bedside to bench and back” research leading to personalized therapies. However, the current resources to integrate and interpret patient-specific data within the context of acute illness are still limited and new computational approaches are needed (Zenker, Rubin, & Clermont, 2007).

In this chapter I have presented a selection of computational approaches and data sources primarily from systems biology that can be useful for translational clinical research in disease progression. Clearly, due to the vast scope and complexity of human pathophysiology, no one methodology can be a magic bullet. I believe that judicious selection, adaptation, and application of techniques such as these can yield valuable insights into the underlying mechanisms of disease progression and help formulate effective and personalized therapies. In the next two chapters I present the Pathway Semantics Algorithm, a novel computational method that uses matrix algebra to bridge biology and medicine for the translational analysis of disease progression.
Chapter 3 Pathway Semantics Algorithm: Node Analysis

Rapid advances in lab technologies have made it easier and cheaper to measure vast quantities of molecular data in biofluids and tissue, resulting in exponential growth in the amount of quantitative and qualitative data available about molecular pathway interactions. In 2001, only 18 pathway websites were active (Wixon, 2001). By 2010, more than 205 million molecular interactions were accessible via the Internet (PathGuide, 2010). There are commercial and publicly available databases of molecular interactions (Ingenuity, 2010; Tarcea et al., 2009), biological pathways (Caspi et al., 2008; Elliott et al., 2008; Visvanathan et al., 2008; Wixon, 2001), and genomic correlates (L. T. Sam et al., 2009).

Increasingly detailed molecular data from patients can be measured in a timely manner. For example, Luminex’s xMAP technology can measure multiplex analysis of up to 500 unique analytes in a single test well, generating up to 48,000 data points in less than one hour (Luminex, 2010). Recently, the technical ability to efficiently measure vast quantities of patients’ molecular data has moved from the bioscience laboratory to the patient’s bedside with the advent of lab-on-a-chip sensor technologies (Jokerst & McDevitt, 2010; Mark, Haeberle, Roth, von Stetten, & Zengerle, 2010) that can support personalized medicine.

This deluge of molecular data creates opportunities for translational biomedical research that connects patients’ disease states and molecular data with existing pathway databases. However there is a dearth of algorithms and computationally tractable methods that facilitate analysis of bedside-to-bench-and-back information. Researchers have
called for improved discovery processes (Rifai et al., 2006) and methods for process quality assessment (Tuglus & van der Laan, 2008). Plausible and meaningful hypotheses must be derived from a deluge of quantitative and qualitative experimental data that are spread over a variety of experimental paradigms such as clinical outcome, time, cell cycle phase, or molecular localization. In addition to analytical methods, there is a need for ways to uncover new findings that lead to interesting hypotheses.

Studies of scientific discovery have demonstrated that most new findings arise from data-driven hypotheses generated from unexpected observations rather than from verification of pre-determined hypotheses based on theories (Klahr & Simon, 1999). In a bedside-to-bench approach, discovery is driven by patient data collected at the bedside. Mechanisms or therapies are confirmed later at the lab bench. Data-driven, evidence-based molecular patterns are a fundamental component of personalized medicine research; the molecular patterns can be used to identify drug targets or candidate biomarkers.

Notable diagnostic successes based on the molecular patterns found in patient data include the validation of 14-3-3 proteins found in cerebrospinal fluid (CSF) as diagnostic of transmissible spongiform encephalopathies (Hsich, Kenney, Gibbs, Lee, & Harrington, 1996) and the validation of a panel of 18 urinary molecules that discriminate antibody-associated vasculitis from other renal diseases (Haubitz et al., 2009). Overall, my goal is to advance diagnostic, prognostic and therapeutic knowledge and increase understanding of the biological mechanisms underlying disease progression.
The challenge of pathway analysis for shock / trauma

In shock / trauma, cytokine signaling molecules activate a pro-inflammatory systemic response (Oberholzer et al., 2000) across many biological pathways to fight immediate injury; however, if the compensatory systemic response is too much or too little, death or morbidity may ensue (Adib-Conquy & Cavaillon, 2009; E. E. Moore et al., 2005). Specific cytokines have been associated with mortality in septic shock (Waage et al., 1989), critical illness (Roche & Gussler, 1992), and trauma (Roumen et al., 1993). Cytokine activity patterns change rapidly within the first 24 hours of trauma, and sets of cytokines significantly associated with one outcome at a specific time from insult may not be associated with any outcome at another time (Jastrow et al., 2009).

Today, immunoassay methods provide highly accurate measurements of cytokine levels. This measurable cytokine data has great potential for developing therapies that minimize the occurrence of possibly preventable syndromes associated with trauma such as multiple organ failure. However, it is difficult to draw inferences about the meaning of likely molecular patterns without efficient algorithms and techniques. Algorithms that select and prioritize molecular patterns for further investigation are beneficial because they can limit the exploration space for *in vitro* and *in vivo* hypothesis testing, minimize risks and costs of experimentation and provide evidence-based information for clinical trials research.

Current approaches to molecular pattern identification in disease include the use of high throughput measurement techniques such as mass spectrometry and microarray immunoassays and qualitative methods such as text mining and graphical analysis. Mass
spectrometry is the most common technique for “unbiased” discovery where all protein and peptide components of tissues and biofluids are identified within the capability of the equipment. Microarray immunoassays are more sensitive and specific; they measure the concentrations of pre-determined analytes using immunological reactions. Both assay methods have benefits and drawbacks for clinical usage (Hoofnagle & Wener, 2009). Text mining algorithms search published literature for information about molecular function and disease associations while graphical analysis uses algorithms from computer science to identify subgraph motifs in canonical pathway networks of molecular interactions found in diseases. Network-based graphical analysis using gene expression patterns has been shown to generate novel hypotheses about the classification of breast cancer metastasis, including the finding that some gene associations can only be detected using network rather than conventional analysis (Chuang, Lee, Liu, Lee, & Ideker, 2007).


There are drawbacks to the current approaches. The most significant molecular interactions associated with the disease may appear in a non-canonical pathway (W. X. Li, 2008) that text mining and in silico modeling may overlook. Although ordinary differential equations (ODEs) can provide time-based analysis of biological pathways, they usually model a small group of canonical pathways within a single cell and are not easily computable at the organism level. For example, an ODE model of one NFκB signaling pathway in one cell activated by one TNF-α signaling molecule uses 18
nonlinear differential equations, with 33 independent variables and 16 dependent variables in a simplified reaction kinetics model (Cho et al., 2003). Hence, additional analytical techniques are needed to overcome the research bias towards canonical pathway associations.

In summary, to develop deeper insights into the mechanisms of disease progression and to improve treatment, it is useful to examine in detail the biological pathways that are activated over time, resulting in differential outcomes. Because patterns of cytokines have been shown to change rapidly in trauma, it is likely that their associated biological pathways offer clues to the underlying pathophysiology, and perhaps even, what is about to happen next. The challenge of pathway analysis for shock / trauma is to develop computationally tractable bedside-to-bench methods that can infer the most likely biological pathways activated by cytokine signaling, provide an analytical framework to examine those pathways in terms of biomedical questions, support inquiries about the relationships among the clinical states and the underlying biological progressions, and suggest hypotheses as to how treatment might influence physiology to mitigate pathways that go out-of-control.

**Using algebra to understand biological pathways in disease progression**

In disease progression, a specific molecular pattern can be associated with a certain outcome only within a certain time period (Jastrow et al., 2009). The use of a mathematical representation that enables scalable computation can uncover hidden molecular patterns associated with disease progression. In particular, representing data in matrix form can be a way to perform powerful and tractable computations that analyze
and compare changes in molecular patterns over patient outcome, time and other stratifications.

Matrices are extensible and computable in n-dimensions; they provide a theoretically sound structure that can be used for biomedical analysis. Matrices can frame biomedical questions in a way that can uncover relationships between phenomena and hypotheses through algebra. In addition, matrices facilitate analysis over multiple stratifications such as time, outcome, and disease states even with constraints such as a small sample size. Although the process of converting bio-assay data into matrix representation and using matrix algebra to answer questions of biomedical interest is non-trivial, once the framework is set up, a multitude of analyses can be performed.
In this chapter, I present the Pathway Semantics Algorithm (PSA) that first converts the bioassay data to matrix representations and then performs matrix algebra to generate clinically useful hypotheses that answer biomedical questions. See Figure 3-1. When applied to a trauma research study on time-based cytokine patterns related to multiple organ failure (MOF) (Jastrow et al., 2009), PSA revealed novel patterns – beyond those of the cytokines – in the evoked biological pathways that differentiated the outcomes of MOF or non-MOF. The algorithm differs from use of standalone pathway analyses, such as those performed in Ingenuity Pathway Analysis (Ingenuity® Systems, www.ingenuity.com), because PSA preprocesses the data before input to Ingenuity

Figure 3-1: Goal of Pathway Semantics Algorithm

The overarching goal of the Pathway Semantics Algorithm (PSA) is to efficiently generate clinically useful hypotheses about disease progression using matrix algebra to integrate quantitative and qualitative data.
Pathway Analysis (IPA), tailoring data to the biological and clinical questions under study.

Herein, the next section outlines the algorithm, followed by a demonstration of PSA based on serum cytokine protein data from Jastrow’s prospective observational study at a Level I trauma center (Jastrow et al., 2009). The next section presents the results followed by a discussion of the matrix algebra approach along with considerations for its application. The chapter ends with a summary.

**Algorithm**

PSA first processes the input data to generate biological pathways (Steps 1-2) and then maps the results to matrices constructed to answer the biomedical questions under study (Steps 3-4). If biological pathways are already available, for example, from morphoproteomic tissue analysis (R. E. Brown, 2005), only Steps 3 and 4 need be performed.

**Dimensionality Reduction.** This process selects characteristic subsets of the measured molecules. The assayed molecules are assembled into *Significance Sets* of those molecules that statistically differentiate the disease states over the stratifications under study, such as outcome, time period of measurement, cell cycle phase observed, or a combination of stratifications. The statistical analysis is utilized as a simple factor analysis, or feature extraction tool, to identify significant molecules.

**Pathway Generation.** The Significance Set for each stratification group plus the statistically observed average values (means or medians as appropriate) for each molecule in the group are input to a pathway generation algorithm that expands each set
to include its likely neighboring molecules, based on published literature and pathway databases. A network diagram is then created of the biological pathways showing the interactions among the molecules for each stratification group.

**Convert Network Diagrams to Matrices.** Matrix representations, suitable for the biomedical questions under study, are created from the network diagrams. The molecules, or nodes, in the network diagram are mapped to a node matrix (or vector) of molecules over the disease states; the molecular interactions, or edges, in each network diagram are converted to a matrix (or vector) of molecular interactions. In the simplest form, the node matrix has 1 in a row/column cell if the row molecule (or molecular interaction) is present in the column disease state; 0 otherwise.

**Matrix Analysis.** Algebra is used to compare the matrices to identify differential patterns of molecules and molecular interactions of biomedical significance over outcome, time and other stratifications.

Figure 3-2 illustrates a PSA flow diagram for node analysis as applied to the trauma study.
Figure 3-2: PSA flow diagram for shock / trauma study node analysis
Application

Trauma is the leading cause of mortality in the US among individuals below 45 years of age. In 2006, 156,000 deaths occurred due to trauma; trauma is the cause of 74% of all deaths for people ages 15-24 (Heron et al., 2008). Multiple organ failure (MOF) – also known as Multiple Organ Dysfunction Syndrome – is one of the most potentially preventable syndromes arising from trauma, yet its pathophysiology is not well understood (Deitch, 1992; Maier et al., 2007). The syndrome is unique in that the organs that fail are not necessarily injured from the trauma and that late MOF may arise days to weeks after the initial incident. MOF continues to be a leading cause of morbidity in patients who survive the initial trauma (Stewart, 2007; Watson et al., 2009).

As previously reported, non-parametric statistical analysis showed that certain cytokine patterns within the first 24 hours from trauma were associated with the outcome of multiple organ failure before other symptoms were visible (Jastrow et al., 2009). Cytokines are small proteins released by stimulated macrophages, monocytes, T cells, and other cells; they bind to specific receptors to induce a wide variety of local and systemic responses particularly within the innate and adaptive immune systems (Janeway, Travers, Walport, & Shlomchik, 2004).

PSA used de-identified patient data from the Jastrow study, extracted from the UTHSC-H Trauma Research Database with the approval of the Committee for the Protection of Human Subjects (Institutional Review Board / IRB) of the UTHSC-H (HSC-SHIS-09-0237). The data included serum cytokine measurements, collection times, and MOF outcomes for 48 patients from an IRB approved prospective observational
trauma study conducted in the Shock / trauma Intensive Care Unit (STICU) at Memorial Hermann Hospital, a Level I trauma center in Houston, Texas from January through December 2005.

Laboratory materials and methods

Twenty-seven cytokines were measured by Bio-Plex immunoassay. See Table 3-1. For detailed study methods please see Jastrow (Jastrow et al., 2009).

Table 3-1: Cytokines in the Bio-Plex Human Cytokine 27-Plex Panel

<table>
<thead>
<tr>
<th>Gene Name</th>
<th>UNIPROT ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eotaxin</td>
<td>CCL11</td>
</tr>
<tr>
<td>FGF Basic</td>
<td>FGF2</td>
</tr>
<tr>
<td>G-CSF</td>
<td>CSF3</td>
</tr>
<tr>
<td>GM-CSF</td>
<td>CSF2</td>
</tr>
<tr>
<td>IFN-γ</td>
<td>IFNG</td>
</tr>
<tr>
<td>IL-1β</td>
<td>IL1B</td>
</tr>
<tr>
<td>IL-1ra</td>
<td>IL1RN</td>
</tr>
<tr>
<td>IL-2</td>
<td>IL2</td>
</tr>
<tr>
<td>IL-4</td>
<td>IL4</td>
</tr>
<tr>
<td>IL-5</td>
<td>IL5</td>
</tr>
<tr>
<td>IL-6</td>
<td>IL6</td>
</tr>
<tr>
<td>IL-7</td>
<td>IL7</td>
</tr>
<tr>
<td>IL-8</td>
<td>IL8</td>
</tr>
<tr>
<td>IL-9</td>
<td>IL9</td>
</tr>
<tr>
<td>IL-10</td>
<td>IL10</td>
</tr>
<tr>
<td>IL12 (p70)</td>
<td>IL12A/B</td>
</tr>
<tr>
<td>IL-13</td>
<td>IL13</td>
</tr>
<tr>
<td>IL-15</td>
<td>IL15</td>
</tr>
</tbody>
</table>
Cytokine measurement was performed according to the manufacturer’s instructions using a Bio-Plex multiplexed suspension immunoassay (171-A11127 Bio-Plex Human Cytokine 27-Plex Panel, 1 x 96-well, Bio-Rad Laboratories, Hercules, CA). The characteristics of the assay varied for each cytokine. According to the manufacturer, the assay working ranges had lower limits of 1.4 pg/ml for MIP-1α to 92.6 for IFN-γ, and upper limits ranging from 836 pg/ml for MIP-1α to 95,484 pg/mL for TNF-a. Assay sensitivity (Limits of Detection, LOD) ranged from 0.6 pg/ml for IL-1β to 6.4 pg/ml for IFN-γ. Intra-assay coefficient of variation (%CV) ranged from 5% for IL-10 and IL-15 to 15% for IFN-γ. Inter-assay %CV ranged from 4% for IL-8 to 11% for IL-6 and Eotaxin(Zhou, Ma, Fedynyshyn, Tan, & Wang, 2009). There is no record that the manufacturer’s specifications and %CV’s were confirmed during the lab assays.

**Data preparation.** To normalize the data to time from injury, measurement times were adjusted to the estimated time from trauma insult, including transport time by land
or air, as well as hospital time before the start of the resuscitation protocol. If transport time was not given, the average time for that mode of transportation (by land or helicopter) was used. This was done in order to preserve biological relationships over time so that the cytokine pattern activities “lined up” for analysis.

Of the planned 12,960 measurements, 1,107 were missing; 2,057 were “low” and 74 “high”, due to readings outside the immunoassay range. Rather than discard the low and high measurements, or reduce the actual measurements to ordinal values, the low and high readings were converted to numerical values based on the range of each cytokine value. High was replaced by 150% of the maximum value of that cytokine; low by 50% of the minimum value of that cytokine. This approach was taken because the non-parametric statistical analysis was itself ordinal - based on rank - to differentiate outcomes. For example, [5, 2, 7, low, 9] was replaced by [5, 2, 7, 1, 9]. All five data points were retained and the rank order would be the same. Missing data were treated as such.

Because the measured molecules were signaling molecules, the number of molecules available to trigger biological pathways was considered more important than their total mass. Therefore the cytokine data were converted from pg/ml units to SI units before input to the software that generated the most likely biological pathways based on relative concentrations of molecules.

The data were grouped over stratifications to facilitate discrete analysis. This preserved the original data without making the continuity assumption that the concentrations of the cytokine molecules varied smoothly between measurement times.
The cytokine data were partitioned for analysis purposes into 6 groups by time periods: hours 2–6, 6–10, 10–14, 14–18, 18–22 and 22–24. The four-hour time period was chosen because that was the scheduled time between clinical measurements.

For clarity and simplicity, the mathematical representation used was limited to vectors over time in the form of two-dimensional matrices.

**Step 1: Dimensionality Reduction**

Significance Sets $S_{i=1,6}$ of molecules $c_{i=1,6; a=1,A}$ that statistically differentiated the K outcomes $q_{k=1,K}$ over time periods $x_{i=1,6}$ were created based on the non-parametric Mann–Whitney–Wilcoxon (MWW) test executed in each of 6 time periods within the first 24 hours from insult. Outcomes were $q_1 = \text{MOF}$ (multiple organ failure) or $q_2 = \text{NMOF}$ (non-multiple organ failure). Time periods from insult were $x_{i=1,6} = 2–6, 6–10, 10–14, 14–18, 18–22$ and 22–24. The Significance Sets $S_1, S_2$ and $S_6$ contained the names of 10 of the 27 measured cytokines; $S_3$ and $S_5$ contained 14 cytokines; and $S_4$ had 15 cytokines. The names of the cytokines differed in each $S_i$. For example, $S_1$ contained: $c_{1,1} = \text{Eotaxin}; c_{1,2} = \text{G-CSF}; c_{1,3} = \text{GM-CSF}; c_{1,4} = \text{IFN-}\gamma; c_{1,5} = \text{IL-1ra}; c_{1,6} = \text{IL-6}; c_{1,7} = \text{IL-8}; c_{1,8} = \text{IP-10}; c_{1,9} = \text{MCP-1}$ and $c_{1,10} = \text{MIP-1}\beta$. See Table 3-2. Dimensionality reduction was achieved by selecting for further analysis only the group of cytokine molecules identified as statistically significant outcome differentiators in each time period – a basic factor analysis.
Table 3-2: Significance Sets of cytokines over time

<table>
<thead>
<tr>
<th></th>
<th>S₁</th>
<th>S₂</th>
<th>S₃</th>
<th>S₄</th>
<th>S₅</th>
<th>S₆</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eotaxin</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>G-CSF</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>GM-CSF</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>IFN-γ</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>IL-1ra</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>IL-5</td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IL-6</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>IL-7</td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>IL-8</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>IL-9</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IL-10</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IL-13</td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IP-10</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>MCP-1</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>MIP-1β</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>RANTES</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>TNF-α</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
</tbody>
</table>

Sᵢ contains the names of the molecules in the Significance Set in time period xᵢ. X represents the median values vᵢ,a,k for each outcome. Note that the significant molecules in Sᵢ differ by time period, reflecting the dynamic nature of the cytokine signaling patterns in shock / trauma progression.
For consistency with the original non-parametric statistical analysis, the statistical weight $v_{i,a,k}$ was set equal to the median concentrations in pg/ml of each assayed molecule $c_{i,a}$ in $S_i$ for each outcome $q_k$. Table 2 was expanded to 2 tables, one for MOF and one for non-MOF, with the median values for each outcome in place of the “X’s”.

**Step 2: Pathway Generation**

Ingenuity Pathways Analysis (IPA) was used to find the likely neighboring molecules because the software provides a literature and pathway database search along with a pathway generation algorithm that utilizes weighted lists of molecules (Ingenuity® Systems, [www.ingenuity.com](http://www.ingenuity.com)). The algorithm breaks “ties” about which neighbors to add to an evoked network based on the relative weightings of the input molecules ([Ingenuity, 2005](http://www.ingenuity.com)). Because the analytes were signaling molecules, the relative numbers of molecular signals, rather than the relative weights of the molecules, generate more representative biological pathways ([M. F. McGuire, Iyengar, & Mercer, 2007](http://www.ingenuity.com)). Therefore, two additional data modifications were performed. First, the units for the median values $v_{i,a,k}$ were converted from concentrations in pg/ml to $v'_{i,a,k}$, the number of molecules per liter (pmol/L) based on the mass of the cytokine in kDa as reported in UniProt ([www.uniprot.org](http://www.uniprot.org)). Second, certain cytokines must be present in multiples or have multiple receptors to send signals. Therefore the $v'_{i,a,k}$ were further adjusted to $v''_{i,a,k}$ by how many molecules were required for one signal. The adjusted calculation details are given in the Appendix.

An IPA data template was prepared for each $S_i$ with the assayed molecule weightings $v''_{i,a,k}$ (intensities) for both outcomes $q_k$ in time period $x_i$ and the molecule’s
“Gene/Protein ID”. The molecule was identified by its UniProt Knowledgebase (UniProtKB) Accession Number, based on the best match for human (subunit A or chain A). Each v_i,a,k was entered as an “Observation/Expression k”, with k=1 for MOF and k=2 for non-MOF. The 6 datasets generated 12 time-stamped network groups with one to three 35-molecule networks in each group (the default 35-molecule limit is adjustable.) Each group was exported as a text list of molecules (network nodes) and as a graphic image of molecular interactions (network edges) The molecules lists were combined to identify 193 unique subject molecules. See Figure 3-3.

Figure 3-3: Biological networks differ by disease state

Although the same Significance Set of cytokines was used for generation of both pathways in a time period, the cytokine median values for each outcome were different, resulting in different networks: MOF on the left, non-MOF on the right, at hours 10–14 from trauma. (See Appendix for network details not visible at this scale.)
Visual inspection shows differences in the biological networks evoked by different disease states. Shown in Figure 3-3 are the networks for multiple organ failure (left) and non-multiple organ failure (right) based on patient cytokine data at hours 10–14 from trauma. Both networks were evoked from the same set of molecules, $S_3$, with different median concentrations $\nu^{\prime\prime}_{3,a,k}$ for each outcome $q_{k=1,2}$. See the Appendix for all 12 graphs that generated the 193 unique subject molecules.

**Step 3: Convert Network Diagrams to Matrices**

A summary list $T_r$ of 193 unique subject molecule names $m_r$ was assembled from the 12 groups and entered into column 1 of two temporal dependency matrices $TDM_{MOF}(m_r, x_i)$ and $TDM_{NMOF}(m_r, x_i)$, with the headers for columns 2–7 set as the time periods $x_i$ and 1 or 0 in row/column cells $z$ denoting the presence or absence of the molecule as depicted in the example matrices in Figure 3-4. $TDM_1$ (above), $TDM_2$ (below), show 6 molecules $m_r$ over 3 time periods $x_i$ in 2 outcomes $q_k$. To identify molecular patterns by outcome and over time, a summary list $m$ was compiled of the names of the molecules present in any of the biological networks evoked from the assayed molecules. Then a temporal dependency matrix (TDM) matrix was constructed for each outcome $q_k$, with the molecule names $m_r$ as the first column and the time periods $x_i$ as the headers across the remaining columns. If the molecule was present in the time period in the outcome, a 1 was placed in the row/column cell $z_{kri}$; otherwise 0. The rationale behind this process was to facilitate computational comparisons over time and outcome using matrix algebra and logic.
Matrix algebra was then used to compare the TDMs over disease state stratifications to elucidate disease progression and explore questions of biological significance.

Figure 3-4: Temporal Dependency Matrices (TDMs) example

TDM$_1$ (above), TDM$_2$ (below), show 6 molecules $m_i$ over 3 time periods $x_i$ in 2 outcomes $q_k$. 
Step 4: Matrix Analysis

The mapping of pathways to matrices enabled a wide variety of computational analyses using pathway molecules (nodes) and their interactions (edges) to uncover hidden network patterns. Two examples of using node analysis follow.

**Example 1 Node Analysis.** Identify molecules $m_i$ that appear at least once in both outcomes in the same time period $x_i$ and at least once in either outcome in a different time period.

**Background:** Danger-associated molecular patterns (DAMP) in the systemic inflammatory response syndrome (SIRS) and sepsis induce the production of pro and anti-inflammatory mediators by pattern-recognition receptors (PRR). A dysfunctional acute inflammatory response may lead to MOF (Bianchi, 2007; Castellheim, Brekke, Espevik, Harboe, & Mollnes, 2009).

**Biomedical questions:** In this study, are there molecules that are “time-shifted” in different outcomes? Is a molecular interaction continuing past its “normal” innate response?

**Hypothesis:** If the identified molecules appear in both outcomes at different times, then additional research may show how to modulate those molecules to minimize negative outcomes.

Let $Z_{MOF} =$

\[
\begin{bmatrix}
Z_{11} & Z_{12} & \cdots & Z_{1m} \\
Z_{21} & Z_{22} & \cdots & Z_{2m} \\
\vdots & \vdots & \ddots & \vdots \\
Z_{R1} & Z_{R2} & \cdots & Z_{Rm}
\end{bmatrix}
\]
Let $Z_{\text{NMOF}} =
\begin{bmatrix}
Z_{11}' & Z_{12}' & \cdots & Z_{1I}' \\
Z_{21}' & Z_{22}' & \cdots & Z_{2I}' \\
\vdots & \vdots & \ddots & \vdots \\
Z_{RI}' & Z_{R2}' & \cdots & Z_{RI}'
\end{bmatrix}$

Let $Z^\prime = Z_{\text{MOF}} + Z_{\text{NMOF}}^\prime$

The cells $z_{\text{ir}}^\prime$ of the resulting matrix $Z^\prime$ have a 2 if the molecule $m_i$ was present in both outcomes in time period $x_i$, a 1 if it was present in one outcome or the other, and 0 if it was not present in either. A molecule $m_i$ was selected if there was at least one 2 and one 1 in its row. Using these criteria, four molecules were identified that appeared at least once in both outcomes in the same time period and at least once in either outcome in a different time period: CIITA, HIRA, IG9, and KSR2.

**Example 2 Node Analysis.** Identify molecules that appeared only in one outcome or the other in more than one time period.

Background: Cytokine patterns are associated with different trauma outcomes (Jastrow et al., 2009; Maier et al., 2007).

Biomedical question: Are there molecules in the pathways triggered by the measured cytokines that are associated only with one outcome in at least 2 of the 6 time periods under study?

Hypothesis: Molecules that meet these criteria may reveal underlying mechanisms that have not yet been associated with specific clinical outcomes.
Let $\text{MOF\_SELECT}(m_i) = 1$, 
if $\left( \sum_{i,l,j} z_{ri} > 1 \right) \land \left( \sum_{i,l,j} z'_{ri} = 0 \right)$; else 0  
(1)

Based on these criteria, four molecules were identified as appearing only in MOF: Egfr-Erb2, IFI6, MRAS and NOD1; no molecules appeared solely in NMOF.

**Results**

The matrix analysis in Step 4 identified eight molecules from the 193 molecules evoked by the assayed cytokines whose patterns at different times differentiated outcomes. Literature searches were performed on each molecule to ascertain associations with multiple organ failure or other shock syndromes. IG9(Calderon et al., 2000) was deleted because the molecule’s identification was withdrawn (T. M. Calderon, personal communication). See Table 3-3.
Of the seven molecules that differentiated outcomes of MOF or non-MOF, only three have been previously been associated with shock / trauma: CIITA, EGFR and NOD1. The citations were retrieved from PubMed on February 6, 2010 based on a search for the molecule name and the MeSH term “shock,” which includes the following syndrome categories: Multiple Organ Failure, Cardiogenic Shock, Hemorrhagic Shock, Surgical Shock, Traumatic Shock, and Systemic Inflammatory Response Syndrome (including Septic Shock). Following are short descriptions of the seven molecules: CIITA, EGFR, HIRA, IFI6, KSR2, MRAS, and NOD1, and what they suggest for shock / trauma progression.

### Table 3-3: Molecular patterns of multiple organ failure

<table>
<thead>
<tr>
<th></th>
<th>2–6</th>
<th>6–10</th>
<th>10–14</th>
<th>14–18</th>
<th>18–22</th>
<th>22–24</th>
</tr>
</thead>
<tbody>
<tr>
<td>CIITA</td>
<td>M, N</td>
<td></td>
<td></td>
<td>M</td>
<td>M</td>
<td></td>
</tr>
<tr>
<td>EGFR</td>
<td>M*</td>
<td></td>
<td></td>
<td>M*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HIRA</td>
<td>N</td>
<td></td>
<td></td>
<td>M, N</td>
<td>M</td>
<td></td>
</tr>
<tr>
<td>IFI6</td>
<td></td>
<td></td>
<td></td>
<td>M*</td>
<td>M*</td>
<td></td>
</tr>
<tr>
<td>KSR2</td>
<td>M, N</td>
<td>M</td>
<td></td>
<td></td>
<td></td>
<td>M</td>
</tr>
<tr>
<td>MRAS</td>
<td>M*</td>
<td>M*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NOD1</td>
<td>M*</td>
<td></td>
<td></td>
<td>M*</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Key** M: appears in MOF, N: appears in non-MOF, M*: appears only in MOF. The header row shows the time in hours from trauma.
Molecule 1: CIITA

EntrezGene, UNIPROT ID: EG 4261, P33076

PubMed search: ("shock"[MeSH Terms] ) AND CIITA[All Fields]

Known in Shock: Yes (Le Tulzo et al., 2004; Pachot et al., 2005; Pangault et al., 2006; Wilson et al., 2008)

Known Functions: CIITA is up-regulated by PPARγ in vascular smooth muscle cells, which enhances IFNγ-mediated transcription and rescues the TGFβ antagonism (Kong, Fang, Fang, Li, & Xu, 2009). CIITA directly inhibits viral replication and spreading; CIITA triggers antigen presentation to CD4+ T cells leading to an adaptive immune response (Tosi, Bozzo, & Accolla, 2009). Enteral glutamine decreases infectious complications in trauma by protecting the gut. Glutamine administered to the post-ischemic gut has been correlated with transcriptional activation of PPARγ. There is also smooth muscle in the gut; therefore CIITA may be up-regulated due to the PPARγ activated by the administration of enteral glutamine, which has been shown to be safe during active shock resuscitation (Santora & Kozar, 2009), (McQuiggen et al., 2008).

Molecule 2: EGFR

EntrezGene, UNIPROT ID: EG 1956, P00533

PubMed search: ("shock"[MeSH Terms]) AND "EGF receptor"[All Fields]

Known in Shock: Yes (Miettinen et al., 1995; Sanchez, Viladrich, Ramirez, & Soley, 2007; Viladrich, Sanchez, Soley, & Ramirez, 2008)

Known Functions: Transactivation of EGFR and ErbB2 protects intestinal epithelial cells from TNF-induced apoptosis (Yamaoka et al., 2008). EGF is a potential therapeutic
agent for the treatment of sepsis (Clark, Clark, Hotchkiss, Buchman, & Coopersmith, 2008).

**Molecule 3: HIRA**

EntrezGene, UNIPROT ID: EG 7290, P54198

PubMed search: ("shock"[MeSH Terms]) AND HIRA[All Fields]

Known in Shock: No

Known Functions: HIRA promotes replication-independent nucleosome assembly (Eitoku, Sato, Senda, & Horikoshi, 2008).

**Molecule 4: IFI6**

EntrezGene, UNIPROT ID: EG 2537, P09912

PubMed search: ("shock"[MeSH Terms]) AND IFI6[All Fields]

Known in Shock: No

Known Functions: IFI6 is believed to play a critical role in the regulation of apoptosis, or programmed cell death and is a marker for interferon beta (IFNB) activity (Serrano-Fernandez et al., 2009).

**Molecule 5: KSR2**

EntrezGene, UNIPROT ID: EG 283455, Q6VAB6

PubMed search: ("shock"[MeSH Terms]) AND KSR2[All Fields]

Known in Shock: No

Known Functions: KSR2 regulates insulin sensitivity and glucose (Costanzo-Garvey et al., 2009). Hyperglycemia associated with insulin resistance is common in critically ill patients (Van Den Berghe et al., 2001). KSR2 inhibits MAP3K8 (Cot, Tpl2) kinase
activity and signaling (Channavajhala et al., 2003). Inhibition of MAP3K8 in primary human cell types can decrease the production of TNF alpha and other pro-inflammatory mediators such as MAP3K3-mediated IL-8 (EG 283455) during inflammatory events (Hall et al., 2007).

**Molecule 6: MRAS**

EntrezGene, UNIPROT ID: EG 22808, O14807

PubMed search: ("shock"[MeSH Terms]) AND MRAS[All Fields]

Known in Shock: No

Known Functions: MRAS is involved with adhesion signaling, inducing lymphocyte function-associated antigen 1 (LFA-1)-mediated cell aggregation (Yoshikawa et al., 2007).

**Molecule 7: NOD1**

EntrezGene, UNIPROT ID: EG 10392, Q9Y239

PubMed search: ("shock"[MeSH Terms]) AND NOD1[All Fields]

Known in Shock: Yes (Cartwright et al., 2007; Chin et al., 2002; Kim et al., 2008; Nembrini et al., 2009).

Known Functions: Activation of NOD1 has been shown to induce septic shock and multiple organ injury (Cartwright et al., 2007). NOD1 protects the intestine from inflammation-induced tumorigenesis (Chen, Shaw, Redondo, & Nunez, 2008). NOD1 is involved in the direct killing of *Helicobacter pylori* bacteria in the stomach and duodenum by epithelial cells (Grubman et al., 2009). Commensal bacteria promote immune homeostasis via the innate immune receptor NOD1 (Chen & Nunez, 2009).
CIITA, NOD1 and EGFR have been previously associated with shock / trauma. They maintain intestinal epithelial cell homeostasis during immune and inflammatory responses and appear in MOF pathways in this study. This is consistent with previous findings that pathophysiology of the gut (epithelium, mucosal immune system, and the commensal bacteria) contributes to critical illness (Clark & Coopersmith, 2007) and to multiple organ failure (Hassoun et al., 2001).

Although four molecules - HIRA, IFI6, KSR2, and MRAS - have not yet been associated with shock / trauma, their biological functions seem to be consistent with trauma progression. MRAS appears in hours 2–10 solely in MOF; it is implicated in the regulation of integrin-mediated leukocyte adhesion in inflammatory and immune responses (Yoshikawa et al., 2007). IFI6 appears in hours 14–22 solely in MOF; it regulates apoptosis, suggesting that programmed cell death is essential to MOF (Serrano-Fernandez et al., 2009). HIRA is observed in non-MOF in the first hours, and later in MOF. It promotes nucleosome assembly (Eitoku et al., 2008). This may indicate either the activation of gene transcription or silencing, with different timings associated with different outcomes. Likewise, KSR2 is associated with both outcomes early on, but appears solely in MOF in hours 22–24. It regulates insulin sensitivity (Salluh & Bozza, 2008) and, through inhibition of MAP3K8, decreases pro-inflammatory mediators (Channavajhala et al., 2003), (Hall et al., 2007). Hence, the presence of KSR2 may reflect the up-regulation of pathways in an attempt to modulate the inflammatory response after injury. This may be an underlying mechanism related to the fact that
insulin resistance and hyperglycemia are common in non-diabetic critically ill patients (Van Den Berghe et al., 2001).

**Discussion**

In this application, PSA identified and qualified 7 molecules in patterns across time of the progression of multiple organ failure; of these, only 3 had been previously associated with any shock / trauma syndrome. A literature search confirmed that the molecules’ biological functions were consistent with the current understanding of MOF. PSA also highlighted the dynamic nature of trauma response, indicating that molecular patterns are specific to certain time periods from insult. PSA uncovered novel molecular patterns in shock / trauma using an unbiased data-driven approach that integrated what was known about the patient and what was known about molecular interactions. The appearance of these patterns made sense within the disease context, and suggested hypothetical answers to the biomedical questions about which molecules differentiated patient outcomes. All 7 of these molecules were in the evoked biological pathways over time and were not measured directly. Instead, they were inferred from published literature documenting molecular interactions.

Although these results provide insights into potential hypotheses that may be useful in trauma, the quality of this algorithm depends on quality data from assays, literature and biological pathway databases as well as the statistical and network algorithms used. Following are some key considerations:
**Quality of the patient data and the assay method.** In the MOF application, 8.5% of the data was missing. Only one assay method was used, and, its working ranges and limits of detection (LOD) varied depending on the cytokine being assayed.

**Quantity of the patient data.** Only 11 of the 48 patients had outcomes of multiple organ failure; however, there were several thousand cytokine measurements taken on a regular time basis. Because the time periods were based on time from trauma, the number of measurements differed in each time period, with the fewest being in the first time period 2–6 due to patient travel time and the time of protocol entry. In comparison, this sample contained more cytokine data than found in the Trauma Related Database (TRDB) of the multi-center, multi-year *Inflammation and the Host Response to Injury Large Scale Collaborative Program*. As of 2008, the TRDB contained only 80 trauma subjects with cytokine data sampled irregularly ([www.gluegrant.org](http://www.gluegrant.org)).

**Dimensionality reduction through Significance Sets.** Dimensionality reduction, or limiting the number of variables under consideration, was performed to reduce false positives, noise and redundancy in the input data and to reduce the computational burden in subsequent steps. The trade-off was loss of pattern information.

**Choice of statistical analysis used to identify Significance Sets.** In this exploratory analysis, I identified six time-based Significance Sets using the Mann–Whitney–Wilcoxon (MWW) test on two independent samples (MOF or NMOF) over 27 observed molecules in each time period. MWW was selected because more sophisticated techniques rely on normality, a condition not satisfied in these data sets. In this exploratory analysis, I chose to identify six Significance Sets rather than one Significance
Set from a repeated measures test in order to yield more a detailed understanding of disease progression. With my focus on inclusiveness for hypothesis generation, I tolerated the 5% false positive rate in the Significance Sets and the assumption of independence of the observed molecules. However, if enough data are available, multivariate methods such as MANOVA could be applied to account for correlations among the observations. Note that the statistical analysis is being used to judge the significance of a variable (e.g. a cytokine in a time period), not the significance of a value (e.g. an observation of a patient’s cytokine in a time period.) Given a larger sample size with a normal distribution, exploratory factor analysis methods could be used to identify the Significance Sets.

Quality of the biological pathway knowledge base and the algorithm used to evoke biological pathways based on assay measurements. PSA used the commercial product Ingenuity Pathway Analysis. IPA is well accepted in the biological sciences community as seen in several hundred references in PubMed (www.ncbi.nlm.nih.gov/pubmed). I chose to use IPA because it is capable of using concentration data to generate pathway networks, and has the flexibility to generate biological networks of any size incorporating the closest interaction neighbors to the input data. To minimize the effects of noise in the data, median values were used as input to IPA. The default size of 35 nodes per network was used in this study, with 1 to 3 networks generated for each outcome in each time period. Each network group was combined before matrix analysis, resulting in up to 105 nodes connected by direct and indirect molecular interaction edges per time period per outcome.
Changing nature of the biological pathway knowledge base. IPA’s knowledge base is constantly updated based on current findings. It is likely that somewhat different networks will be generated from the same input data each time IPA is run, with the newer, more relevant, knowledge base information added to the old. The earlier information may have been rescinded or simply only visible in a network larger than 35 nodes.

Biological scope of the generated network. If the biological scope is limited to certain species or disease states, the generated network will reflect only current knowledge with the result that potential molecular interactions in other species and disease states may be overlooked. Since the goal of applying PSA to MOF was to uncover hypotheses about potential molecular patterns underlying trauma, it was preferable to run the IPA network generation algorithm without constraints, with the understanding that some of the molecular patterns identified may need to be verified in human shock / trauma progression.

Utility of the molecular patterns. The identified molecules may be difficult to assay clinically due to their primary presence in tissue rather than biofluids, low concentrations, or lack of existing assays. However, the molecular patterns may be useful for \textit{in vitro} and \textit{in vivo} verification of the underlying biological mechanisms that may elicit more clinically useful information.

Resource requirements to implement PSA. Published data for time-based analysis of biofluids and tissues in disease progression may not be readily available although access to biological pathway algorithms and data ranges from free open source to
commercial products. This presents opportunities for research studies to collect more data in areas such as trauma and critical care where rapid changes are seen and rapid response to changing patient condition is required.

Summary

This chapter demonstrates that PSA is useful in generating novel hypotheses about the significant molecules, or pathway nodes, in the changing biological pathways within the first 24 hours of shock / trauma. The PSA matrix algebra approach identified differential molecular patterns in biological networks over time and outcome that would not be easily found through direct assays, literature or database searches.

The biological questions of interest in this application of PSA were capable of being answered by simple matrix constructions. Insights into more intricate biological questions, such as the influence of crosstalk in disease progression, require more complex matrix algebra. In the next chapter, PSA is applied to molecular interaction, or “edge”, analysis of the evoked pathways in shock / trauma.
Chapter 4 Pathway Semantics Algorithm: Edge Analysis

In recent years, advances in technology have made it possible to measure a wide variety of molecules and molecular interactions in cell lines, bio-fluids and tissues. The availability of these data has opened new avenues of biomedical research, and challenged the scientific community to uncover the meaning of molecular data in contexts ranging from cell signaling pathways to phenotype/genotype associations to personalized medicine (Weng, Bhalla, & Iyengar, 1999). Molecular interactions offer a rich source of information that should be examined in detail to further understand their roles in disease progression and outcomes. Computer scientists, mathematicians, physicists, and industrial engineers are joining biologists and medical researchers to develop new quantitative and qualitative analytical methods to answer questions about underlying biological mechanisms and therapeutic efficacies. Algorithms driven by patient data that incorporate knowledge bases of molecular patterns are of particular interest because of their potential for hypothesis generation in personalized diagnosis, prognosis, and therapies. Such algorithms are the focus of my research.

The Pathway Semantics Algorithm (PSA) described in Chapter 3 generates the most likely biological pathways evoked from patients’ molecular data over stratifications such as time and/or outcome, and then converts the pathway graphs to matrices of various formats depending on the biomedical questions being studied. In the pathway graph and the transformed matrix, there are two major types of entities: nodes that correspond to specific bio-molecules and edges that correspond to the interactions among the molecules. The transformation of graphs to matrices enables the application of powerful
techniques from matrix algebra to develop mathematical comparison methods, analyses, and metrics leading to useful insights into disease progression across time and clinical outcomes.

For example, in the PSA node analysis of Chapter 3, I focused on the molecular components of pathway graphs and developed a matrix format called a Temporal Dependency Matrix that was instrumental in revealing novel molecular patterns evoked from patient data over time in shock / trauma, where disease progression is rapid yet not clinically visible. The computational results predicted seven molecules, based on input from the original assays, associated with the biological mechanisms underlying multiple organ failure; only three had been previously recognized as associated with any shock / trauma syndrome. In this chapter I turn my attention to the edges of pathway graphs, corresponding to interactions between molecules including genes, RNAs, proteins, or chemicals. I applied matrix methods to investigate patterns of molecular interactions across time and across clinical outcomes in terms of four functional relationship categories: activation, expression, transcription and inhibition. Applying graph theory and linear algebra, I found that the interaction patterns of relationship sub-graphs changed rapidly within the first 24 hours of trauma insult, and that these patterns differed across clinical outcomes of multiple organ failure (MOF) and non-multiple organ failure (non-MOF). In addition, I developed a numerical metric of crosstalk in molecular pathways called XTALK. In contrast to current practice that merely classifies a network in strictly binary fashion as having crosstalk or not, XTALK quantifies crosstalk among molecular interactions from 0% to 100%, thereby leading to a deeper, fine-grained understanding of
crosstalk and its variation due to disease progression. These methods were applied to the same shock / trauma data set used in the previous chapter. Results obtained suggest that a diagnosis, prognosis or therapy based on molecular interaction mechanisms may be most effective within a certain time period and for a certain functional relationship.

The following sections present background information and definitions relating to molecular interactions and mathematical notation, followed by a description of the application of the Pathway Semantics Algorithm to analysis of molecular interactions in the first 24 hours of trauma progression, the results and a discussion of their meaning, concluding with my plans for future work.

**Background**

At a sub-cellular level, molecular interactions can be analyzed using the rules of biochemistry when they are represented as sets of differential equations. However, due to computational complexity and lack of interaction parameter rate data, this approach is not suitable for larger comparative analyses. Instead, molecular interactions, such as protein-protein or gene-protein interactions, are commonly combined into biological pathway networks represented as graphs, where the node, or vertex, is the molecule and the edge is the interaction. This representation facilitates the use of qualitative and quantitative methods derived from graph theory and algebra because the same biological pathway network graph can be mapped to a matrix in different ways, allowing for a choice of mathematical methods appropriate to the biomedical question under study.

Recently, interest has shifted from analysis of nodes, or vertices, in biological pathway networks, to examining edges, or links between the nodes(Ahn, Bagrow, &
Lehmann, 2010; Evans & Lambiotte, 2009). This parallels the current research into “link communities” in social networks, where one person may be connected to several overlapping communities of home, work, and interests. In both social and biological networks, the edges are directional, showing the influence from one node (a person or molecule) upon another in a multi-directional cascade.

Biological link communities also overlap; a molecule may participate in several different interaction categories simultaneously with the same target molecule, or inversely, several interactions may occur simultaneously with different molecules to achieve the same target function. This latter property has been defined as degeneracy – the ability of structurally different elements to perform the same function or yield the same output; in contrast, redundancy requires identical elements to perform the same function (Edelman & Gally, 2001; Tononi, Sporns, & Edelman, 1999). Degeneracy is a key property underlying the robustness of complex adaptive biological systems, such as the immune system (Macia & Sole, 2009; Tieri et al., 2010; Whitacre, 2010).

Crosstalk in biological pathways can be defined as consisting of the redundant signaling messages sent over degenerate edges that achieve the same biological function. This is consistent with Bruni’s definition that crosstalk exists when edges are functionally compatible to, or dependent, on other edges (Bruni, 2007). Crosstalk relates to how pathways determine functional specificity, how ubiquitous messengers transmit specific information, and how similar messages crosslink within the system while undesired signals are minimized. Quantifying crosstalk in patient data-driven biological pathways can give insights into the relative robustness of different biological functions and suggest
timing and approaches for therapies directed at pathway modulation. For simplicity, this study measured crosstalk in one molecular interaction function at a time in each pathway; cascades of “mixed-function” molecular interactions that overall would result in execution of the same target function were not considered.

**Additional definitions**

Notation and definitions used correspond to those used by Ingenuity Pathways Analysis (IPA) (Ingenuity, 2010). The term node is used rather than vertex.

A *molecule* is any gene, RNA, protein or chemical. A molecule is represented by a node on the directed graph of a biological pathway.

A *relationship* is a functional interaction from one molecule to another. A relationship is represented by an edge on the directed graph of a biological pathway. A directed graph, in mathematical terminology, has specific properties that can be exploited computationally. IPA designates relationships as direct or indirect, in a different sense of the word “direct”. A *direct relationship* is a direct physical contact interaction between the two molecules. It is represented by a solid line edge. An *indirect relationship* is an interaction that does not require physical contact but is explicitly documented in the literature. It is represented by a dotted line edge. A *relationship graph* is a directed graph whose edges are in the same relationship category. Molecules or edges are called *invariant* when they are the same in different stratifications. For example, edges are invariant over all time in one outcome if they do not change over all time periods for that outcome; alternatively, edges are invariant over outcome if they are the same in both outcomes in one time period or more as specified.
Let \( B(E,N) \) be a directed graph with \( E \) edges and \( N \) nodes that represents a biological pathway with relationship interactions as edges and molecules as nodes. Then \( A \) is a relationship sub-graph of \( B \) with \( A \subseteq B \) when \( \forall E \) in \( A \) are in the same relationship category.

**Pathway Semantics Algorithm for Edges**

**Step 1 and Step 2.** The Pathway Semantics Algorithm (PSA) first processes the input data to generate biological pathways (Steps 1-2) and then maps the results to matrices constructed to answer the biomedical questions under study (Steps 3-4). For details of Steps 1 and 2 of Pathway Semantics, please see Chapter 3. For this edge analysis, relationship sub-graphs were extracted from each pathway for each selected molecular interaction relationship within each outcome and time period. The sub-graphs were represented as cyclic digraphs (directed graphs with cycles). Each directed edge, or arc, of a sub-graph was a one-way interaction relationship from one molecule to another. The sub-graphs could also contain loops, or cycles because feedback, feed forward, and self-loops occurred in molecular interactions. This necessitated the use of incidence matrices for computation and limited graph metrics to those for cyclic digraphs.

**Step 3. Map graphs to matrices.**

In Step 3, the pathway networks were mapped to matrices. Each relationship sub-graph was mapped to an incidence matrix, called an Edge-Molecule (EM) matrix, where each row represented a *from*-to edge, and each column represented a molecule, with doubles for self-loops. A -1 was placed in the *from* molecule column, a +1 in the *to*
column and 0 otherwise. All molecules evoked in the study were placed in the column header row.

Definition. The incidence matrix $M = [m_{ij}]$ of a directed graph $B = B(E,N)$ is a $E \times N'$ matrix, $M(E,N')$ where $E =$ number of edges and $N' =$ number of nodes (with duplicate nodes for self-loops) such that $m_{ij} = -1$ if edge $i$ leaves node $j$, $+1$ if edge $i$ enters node $j$, 0 otherwise (Bondy & Murty, 2008).

**Step 4. Compare biological pathways using matrices**

In Step 4, algebraic comparisons were performed across stratifications. First, a descriptive analysis was performed to count the number of edges in each relationship in each outcome over time and to identify edges that were unchanged over time and outcome. Linear algebra was then used to calculate XTALK, the crosstalk for each relationship, time period, and outcome. With the XTALK measure, relationship subgraphs could be analyzed to uncover which functional relationships have the most or the least crosstalk in different outcomes and how crosstalk changes over stratifications such as time. The XTALK measure is based on the calculation of matrix rank:

**Definition.** The rank $R$ of a matrix $M$ is the maximal number of its linearly independent columns or rows (Birkhoff & MacLane, 1953). Rank can be calculated using Gaussian elimination or singular value decomposition.

If rank $R$ is greater than or equal to $E$, the number of edges (rows), then all the edges act independently. The percentage, or ratio, of independent edges $= \frac{R}{E}$, and the ratio of dependent edges is $1 - \frac{R}{E}$.
I propose the biological interpretation that the maximum number of independent molecular interactions (edges) required for a molecular function is the same as the rank of the incidence matrix constructed from the functional relationship sub-graph, and that a measure of crosstalk for that function can be based on the percentage of dependent edges.

Definition. The XTALK ratio of a directed graph \( B = B(E,N) \) with incidence matrix \( M(E,N') \) is defined as \( 1 - \frac{\text{rank} \ (M(E,N'))}{E} \).

If XTALK = 0%, then all edges act independently for a particular function. The XTALK measure includes normalization by the total number of edges in a graph to allow comparisons of crosstalk over time and outcome.

To illustrate the graph mapping to the incidence matrix, see Figure 4-1 and Table 4-1, representing a network with 3 edges and 3 nodes. The calculated rank of the incidence matrix for the graph is 2. This means that 2 edges are independent and 1 edge is dependent. It can be seen that the path along the edge_A_C is a combination of edge_A_B followed by edge_B_C. The network shows the property of degeneracy: the target function can be achieved by edgeA_C or by the edgeA_B followed by the edgeB_C. With rank \( R = 2 \), and the number of edges = 3, XTALK = \( 1-\frac{2}{3} = 33\% \), suggesting there exists one-third crosstalk in the biological functional relationship represented by the graph.
On completion of Step 4, the results were reviewed in the light of published literature and expert opinion to generate targeted hypotheses about the molecular mechanisms of disease progression that may be verified clinically or in the lab.

**Application**

As mentioned in Chapter 3, trauma is the cause of 74% of all deaths for people ages 15-24 (Heron et al., 2008). Disease progression in shock trauma is rapid and deadly; patients who survive the initial trauma may suffer morbidity from potentially preventable syndromes such as multiple organ failure (MOF) (Stewart, 2007; Watson et al., 2009). The pathophysiology underlying MOF is still unclear (Deitch, 1992; Maier et al., 2007). Patterns of signaling molecules called cytokines (Janeway et al., 2004) have been associated with patient outcomes in trauma and critical care for some time (Hranjec et al., 2010; Jastrow et al., 2009; Roumen et al., 1995; Visser, Pillay, Koenderman, & Leenen, 2008; Vodovozt, 2010) and analysis of the biological pathways evoked from cytokines may offer insights into disease progression.
De-identified patient data from the Jastrow study (Jastrow et al., 2009) were extracted from the UTHSC-H Trauma Research Database with the approval of the Committee for the Protection of Human Subjects (Institutional Review Board / IRB) of the UTHSC-H (HSC-SHIS-09-0237). The data included serum cytokine measurements, collection times, and MOF outcomes for 48 patients from an IRB approved prospective observational trauma study conducted in the shock / trauma Intensive Care Unit (STICU) at Memorial Hermann Hospital, a Level I trauma center in Houston, Texas from January through December 2005. Twenty-seven cytokines were measured by the Bio-Plex Human Cytokine 27-Plex Panel.
Although all 27 cytokines were used in the previous PSA molecule (node) study, only 11 were used in this edge analysis due to export limitations of the pathway generation software (Ingenuity Pathway Analysis) and the fact that all edges had to be manually transcribed. The 11 cytokines were chosen by the shock trauma clinicians as those most likely related to multiple organ failure. See Table 4-2. In addition, the analysis was limited to three time periods: hours 6–10, 10–14, and 22–24 hours from trauma; two outcomes: multiple organ failure (MOF) and non-multiple organ failure (non-MOF); and

<table>
<thead>
<tr>
<th>Gene Name</th>
<th>UNIPROT ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eotaxin</td>
<td>CCL11</td>
</tr>
<tr>
<td>G-CSF</td>
<td>CSF3</td>
</tr>
<tr>
<td>GM-CSF</td>
<td>CSF2</td>
</tr>
<tr>
<td>IFN-γ</td>
<td>IFNG</td>
</tr>
<tr>
<td>IL-1β</td>
<td>IL1B</td>
</tr>
<tr>
<td>IL-1ra</td>
<td>IL1RN</td>
</tr>
<tr>
<td>IL-2</td>
<td>IL2</td>
</tr>
<tr>
<td>IL-6</td>
<td>IL6</td>
</tr>
<tr>
<td>IL-8</td>
<td>IL8</td>
</tr>
<tr>
<td>IL-10</td>
<td>IL10</td>
</tr>
<tr>
<td>TNF-α</td>
<td>TNF</td>
</tr>
</tbody>
</table>
four relationship categories of molecular interactions: activation, expression (including metabolism and synthesis for chemicals), inhibition and transcription, for a total of 24 pathway network graphs. The relationships are defined by Ingenuity Systems (Ingenuity Systems, personal communication) as follows:

- **Activation**: includes activation events such as activation, activity, stimulation, reactivation, and specific activity.
- **Inhibition**: includes inhibition events such as inhibition, desensitization, inactivation, repression and autoinhibition.
- **Expression**: includes expression events such as expression, upregulation, downregulation, translation, production, microRNA targeting, and induction.
- **Transcription**: includes transcription events such as including transcription, germline transcription, transactivation, and transrepression

Both direct and indirect interactions were used in the edge analysis. Direct interactions required that two molecules make direct physical contact with each other and included chemical modifications, such as phosphorylations, if there was evidence that the two factors involved interacted directly rather than through an intermediary. Although indirect interactions did not require physical contact between the two molecules, the interactions had to be explicitly stated in the literature – not inferred (Ingenuity, 2010). The edge analysis was performed for each relationship, time, and outcome using the EM matrices, generating edge counts and crosstalk as defined in Section 3. The IPA pathway network graphs for this analysis were generated from September to November 2008.
Results

Steps 1 and 2 of PSA evoked 24 IPA network graphs of the most likely biological pathways. Note that multiple graphs for an outcome in a time period were consolidated into a single graph. A total of 1,264 edges were examined among the 132 molecules evoked in the first 24 hours from insult; each edge was identified by its outcome, time period, “FROM” molecule, “TO” molecule, and molecular interaction relationship category. See Figure 4-2 for one graph; all are shown in the Appendix.
Figure 4-2: Evoked biological pathways in MOF at hours 6–10

The “expression” interaction edges are highlighted in gray. White nodes are the evoked “nearest neighbor” molecules; shaded nodes are the input molecules in the Significance Set. The darker the shade, the higher the measured median value.
EM Matrices

In Step 3, the edges were mapped into twenty-four EM matrices by outcome, time period, and category. Duplicate molecule columns were added to the 132 molecule columns in all EM matrices to include the 12 molecules that had self-loop interaction edges. This was required to enable the from-to edge. The 12 molecules with self-loop feedback were: CCL11, CCNA1, Cyclin A, Cyclin E, IL6, TNF, IFNG, IL1, IL10, Hsp70, RARB, and MYBL2. The final number of columns in each EM matrix was 144, with the number of row edges changing according to the interaction type and the time period. Figure 4-3 shows a portion of the EM matrix for the Figure 4-2 graph.

![Figure 4-3: Section of EM matrix for network shown in Figure 4-2](image)
**Edge Counts**

**Dominant functions.** Based on the Step 4 edge count, the most interactions per time period were in the activation function category, except in hours 22–24 for non-MOF when activation interactions were fewer than expression interactions. Inhibition and transcription interactions were most active in hours 10–14. See Figure 4-4.

![Figure 4-4: Edge interactions over time, outcome, and functional relationship](image)

**Invariant interactions across all time.** No inhibition interactions were consistently present during the three time periods. Two transcription interactions were present in both MOF and non-MOF: PDGF BB, which is involved in the transcription of CSF2 (GM-
CSF) and IL1 (IL-1β), which increases transcription of IL8. PDGF BB is a platelet-derived growth factor homodimer that causes mitosis in cells of mesenchymal origin; here it affects the transcription of CSF2, which encodes a cytokine that controls the production, differentiation, and function of granulocytes and macrophages. IL1 is a cytokine produced by activated macrophages that mediates the inflammatory response, in this case by increasing transcription of IL8, a chemokine that functions as a neutrophil polymorphonuclear cell (PMN) chemoattractant. It is also a potent angiogenic factor.

**Unique interactions in each time period.** Although the majority of molecular interactions were similar in each time period over both outcomes, distinct differences were revealed by a count of the edges unique to MOF or non-MOF. See Figure 4-5. In hours 6 to 10 from trauma, there were twice as many unique activation interactions in non-MOF than MOF; whereas by hours 10–14, MOF surpassed non-MOF with a greater number of unique interactions in all categories. In hours 22–24, MOF had twice as many unique activation edges than non-MOF, although both had the same number of unique expression edges. There were few unique inhibition or transcription interactions. Overall, there were more interactions that appeared solely in MOF than in non-MOF. Another point of interest is that IL6 was involved in ~50% of the unique expression interactions in both outcomes in the first 6–10 hours, while IFNG became dominant in hours 10–14.
Crosstalk

XTALK, a measure of crosstalk based on the dependency between the functional edges as calculated by matrix rank, ranged from 0% to a high of 71%, and changed over time. See Figure 4-6.

**Activation.** Activation crosstalk was calculated at ~69% in hours 6–10, staying steady to 71% at hours 10–14, and decreasing in hours 22–24 to 45% in MOF and 32% in non-MOF.
**Expression.** In hours 6–10, expression edge crosstalk was 51% in MOF and 46% in non-MOF. This increased in hours 10–14 with MOF rising to 62% and non-MOF to 54%. Crosstalk then decreased in hours 22–24 to 27% in MOF and 31% in non-MOF.

**Inhibition.** There was no crosstalk in inhibition interactions in hours 6–10 and 22–24; however, crosstalk increased to 17% in MOF and 20% in non-MOF in hours 10–14.

**Transcription.** 9% transcription crosstalk was calculated in both outcomes in hours 6–10, rising to ~21% in hours 10–14, then decreasing to 0% by hours 22–24.

---

**Figure 4-6:** Crosstalk in functional relationships across time and outcome

Percentages calculated using the XTALK measure.
Interaction Summary

**Activation.** In hours 6–10, there were twice as many unique activation edges in non-MOF compared to MOF; however the reverse was the case in the later time periods. This may imply that in non-MOF, a large number of favorable molecular interactions were underway early, so fewer unique activations were needed as the pathways approached a favorable outcome of non-MOF. The percentage of activation crosstalk was about the same in hours 6–10 and 10–14 in both outcomes, decreasing only in hours 22–24.

**Expression.** By hours 10–14, MOF had more than three times the number of unique expression edges than non-MOF, implying higher energy consumption in MOF metabolism than in non-MOF at this time. The percentage of expression crosstalk was slightly lower in non-MOF than MOF in the first two time periods, changing to slightly higher by the end.

**Inhibition.** Unique inhibition interactions appeared solely in MOF in the last two time periods. Crosstalk appeared in both outcomes only during hours 10–14; it was slightly higher in non-MOF. Again, this suggests an attempt to damp down molecular interactions in both outcomes starting in hours 10–14 that was continued in hours 22–24 by additional unique inhibitory interactions in MOF.

**Transcription.** Unique transcription interactions appeared in both outcomes in hours 10–14, with the majority in MOF. Crosstalk in transcription interactions increased initially, and disappeared in both outcomes by hours 22–24 when only two transcription interactions occurred in each outcome.
Discussion

Today it is generally accepted that there is a need to develop computational, data-driven algorithms to exploit the vast quantity of molecular information available in knowledge bases in order to advance systems biology and to improve patient care (Aristotelis, 2006; Wenyuan Li, Xu, & Zhou, 2010; Ruths, Nakhleh, Iyengar, Reddy, & Ram, 2006; Tipney et al., 2009; Veliz-Cuba, Jarrah, & Laubenbacher, 2010). Due to several successes (Kanehisa, Goto, Furumichi, Tanabe, & Hirakawa, 2009; Sachs, Gentles et al., 2009; L. Sam, Liu, Li, Friedman, & Lussier, 2007), in silico hypotheses generators are no longer denigrated as “fishing expeditions” (Brent & Lok, 2005).

The Pathway Semantics Algorithm (PSA) is an initial in silico data integration and analysis step towards formulating hypotheses about disease progression for personalized diagnosis, prognosis, and therapies that can be validated in the laboratory and in the clinic. PSA is based on a novel, flexible approach that uses graph theory and numerical algebra to computationally compare non-canonical biological pathways evoked from patient data over time.

PSA was applied to a prospective observational study of shock / trauma, a research area where patient data is sparse and difficult to obtain even at a Level I trauma center; randomized controlled trials are not an option. By using patients’ molecular cytokine data to evoke non-canonical biological pathways from the Ingenuity Pathway Knowledge Base, PSA expanded the existing information to include the most likely molecules and molecular interactions evoked by the patients’ cytokines. With the expanded information set, and its representation as pathway graphs, PSA was able to use computational tools
and algorithms from graph theory and numerical algebra to compare patterns of molecules and molecular interactions over different stratifications. In particular, PSA was able to analyze patterns over time – an absolute necessity for clinicians who treat disease as it unfolds (Shahar, 2000). This feature shows the potential of PSA to support temporal reasoning in medical decision-making and support systems.

Overall response to trauma insult

The results from these analyses suggest that molecular interaction activity – and the nature of that activity – changed dramatically within the first 24 hours of trauma. In both outcomes, the number of interactions peaked during hours 10–14 from insult, lessening to about half of the initial activity by hours 22–24; this may be due to the effects of interventions during the first 24 hours combined with the innate systemic response. There were core sets of molecular interactions that were invariant over outcomes in each time period plus unique interactions only in one outcome or the other. This suggests a primary molecular response to the injury that was modulated by the unique interactions edges towards favorable or unfavorable outcomes. MOF had fewer unique interactions early in response, but by hours 10–14, MOF had almost three times as many unique edges as non-MOF – perhaps an excessive number.

Changes in the gene regulation process

Multiple organ failure has been characterized as an adaptive, multilevel time-based stress response with marked changes in gene expression (Adib-Conquy & Cavaillon, 2009; Cobb, Buchman, Karl, & Hotchkiss, 2000; Warren et al., 2009). I believe that this is the first study to quantify the changing aspects of gene expression in MOF over time.
By examining edge interactions \textit{in silico}, changes in functional relationships and their crosstalk over time and outcome were revealed.

Molecules must be activated before they can be transcribed and then expressed, and inhibition can halt any step in the gene regulation process. It is known that cells respond quickly to stress by altering their metabolism; they can induce apoptosis or cell-cycle arrest and alter nuclear pathways for DNA repair (Boulon, Westman, Hutten, Boisvert, & Lamond, 2010). Activation interactions dominated the initial response in both outcomes through hours 10–14, showing the immediate cellular response to stress. Expression was higher in MOF, suggesting a higher metabolic load on the system. Inhibition and transcription interactions were a small proportion of the overall count.

\textbf{Crosstalk changes over time and outcome}

For demonstration purposes, I performed a simple analysis that did not include interaction cascades of different functions in order to focus on a “black box” of four dominant functions. Even with this limitation, differences were observed across time and outcome. This is important because it suggests that a diagnosis, prognosis or therapy based on molecular data might only be valid within a certain time period and for a certain functional relationship, due to the degeneracy in the biological network. For example, because there appear to be few inhibition relationships and little or no inhibition crosstalk in initial trauma, it may be worth exploring increasing inhibition interactions early on in order to limit the excessive unique expression interactions in MOF in hours 10–14. Crosstalk decreased over time in the first 24 hours from trauma, suggesting that therapies should consider time from insult as well as which interaction functions they are targeting.
in order to be effective. This also suggests that trauma therapies may have to be administered in a particular sequence, similar to certain cancer therapies.

**Study considerations and limitations**

The quality of the PSA results depends on the quality of the patient data, the clinical study protocol, the assay method, the choice of statistical analysis, and the accuracy of the biological pathway networks generated by the Ingenuity Pathway Algorithm from its knowledge base. Some key considerations for this edge analysis are as follows:

**Age of pathway data.** IPA generated the biological networks evoked in this study in Fall, 2008. Since that time, there have been extensive, continuous updates to the IPA knowledge base. It is not possible to access older versions of the knowledge base (Ingenuity Systems, personal communication) nor is it possible to export interaction data in other than graphical formats, resulting in extensive manual transcription before computation can be done. Therefore, this chapter is intended as a demonstration of the algorithm, and the actual biomedical results may differ somewhat based on current research. My assumption is that the evoked biological networks will primarily be the same, with the difference that new discoveries may bring new “closest neighbors” into the graph, pushing out existing molecules past the default 35 node limit per graph. This can be addressed by generating new graphs with larger node limits. In addition, the relationships between molecules may be augmented with new relationships or reclassified to related relationships. However, as with published research, older information about relationships is rarely deleted.
Incomplete pathway data. Some functional relationships may be more highly represented in the Ingenuity Pathways Knowledge Base than others due to the type of experiments performed in the published research, rather than the reality of the true proportion of those relationships in nature. This was addressed in the crosstalk calculation by normalizing XTALK by the number of edges in each relationship sub-graph to facilitate comparison across stratifications.

Linearity assumption. Using matrix rank as a basis for the XTALK measure implies that the edges are related in a linear manner – that is, each edge can be represented as a combination of nodes with coefficients of -1, 0, or 1. This can be considered to be a linear approximation to a non-linear function, computed by taking the first term in the representative Taylor series.

Conclusions and Future Work

The Pathway Semantics Algorithm identified different patterns of molecular interactions over time, outcomes, and functional relationships in biological networks that would not be easily found through direct assays, literature or database searches. The differences in the number of edges, the number of unique edges, and the XTALK ratios show the utility of evaluating a molecular interaction not just as a connection between two molecules, but as a directed interaction from one molecule to another that may carry out one or many specific functions (Wu, Zhang, Yu, & Ouyang, 2009). The crosstalk measure XTALK provided a novel perspective on the changing functional interaction relationships in disease progression; the results supported the existence of the property of degeneracy in biological networks.
Chapter 5 Experimental Results / Findings

Introduction

The Pathway Semantics Algorithm was validated by experimental findings in the disease domain of hemophilia A in an ongoing study of the cellular immune response to factor VIII (FVIII), an essential blood-clotting factor. The study is being conducted by Keri Csencsits Smith, PhD, Assistant Professor, Department of Pathology and Laboratory Medicine at the University of Texas Medical School at Houston.

Background

Hemophilia A is a disease characterized by a deficiency in FVIII clotting activity that results in bleeding episodes. The disease is managed by intravenous infusions of FVIII after the onset of bleeding, and, in severe cases, by prophylactic infusions of FVIII concentrate several times during a week to prevent spontaneous bleeding (Brower & Thompson, 1993). However, as many as one-third of hemophilia A patients develop inhibitor antibodies to FVIII, and these patients can no longer receive infusion therapy (Zakarija et al., 2011). Alternative therapies to manage hemophilia A patients with FVIII inhibitors are expensive and not often successful.

Several causes, including genetics and inflammatory response, have been proposed that may influence the production of inhibitor antibodies. Dr. Csencsits Smith’s project focuses on cytokine signaling. Because specific cytokines mediate T cell help for B cell antibody production (Bray et al., 1993), modifying cytokine response may prevent or reduce inhibitor formation. Recently, T cells that secrete IL-17 (known as T_{h}17 cells) have been linked with the development of inflammatory disease (Korn, Bettelli, Oukka, &
Kuchroo, 2009). A compound called colonization factor antigen I (CFA/I), derived from human enterotoxigenic Escherichia coli, has been shown to decrease T\textsubscript{h}17 response and increase the number of T cells that can regulate the immune response (Jun et al., 2005; Kochetkova, Trunkle, Callis, & Pascual, 2008; Ochoa-Reparaz et al., 2008).

The Pathway Semantics Algorithm (PSA) was used to assess how CFA/I treatment modified the immune response in a mouse model by examination of the pathway networks evoked from cytokine profiles over time. The hypothesis was that CFA/I would reduce the formation of inhibitor antibodies by decreasing T\textsubscript{h}17 response. PSA was then validated in the laboratory by confirmation of the presence of a predicted molecule from mouse splenocyte culture supernatant.

Materials and Methods

Laboratory. FVIII-deficient mice (B6;129S4-F8tm1Kaz/J; Jackson Laboratories) (Bi et al., 1995) were intravenously immunized with 2 µg recombinant human FVIII (Kogenate®; Bayer) weekly for 4 weeks. This protocol resulted in the production of anti-FVIII antibodies with inhibitory capacity (Qian, Borovok, Bi, Kazazian, & Hoyer, 1999; Reipert, Ahmad, Turecek, & Schwarz, 2000). Production of inhibitor antibodies was confirmed by Bethesda assay. Beginning 3 days before the initial FVIII immunization, mice received either 100 µg of colonization factor antigen I (CFA/I) isolated from human enterotoxigenic Escherichia coli (E. coli) fimbriae or control sterile phosphate-buffered saline (PBS) via intranasal instillation. Mice were euthanized on days 7, 14 and 28 post-primary FVIII injection, and splenocytes were isolated and cultured for 72 hours in PRMI-1640 media + 10% fetal bovine serim (FBS). Tissue culture
supernatants were harvested and used as directed in a MILLIPLEX™ Mouse Cytokine/Chemokine 22-plex Panel (www.millipore.com/catalogue/item/mpxmcyto70kpmx22) and results analyzed by Luminex 100. 22 cytokines were measured in pg/ml for each treatment for each day, resulting in 6 cytokine profiles.

All procedures incorporated in the studies were reviewed and approved by the University of Texas Medical School. The investigation was guided by the Veterinarian Director of the Animal Resource Facility and associates with regard to handling, analgesics and euthanasia for these studies. The Animal Resource Facility is fully accredited by the American Association for the Accreditation of Laboratory Animal Care for the use of warm-blooded animals in research, training or other activities sponsored by grants, awards or contracts.

**In silico.** The Pathway Semantics Algorithm used the mouse (*Mus musculus*) cytokine data for a node analysis of the biological pathways evoked by the treatments of PBS or CFA/I at two time points: day 7 and day 14. There was insufficient data for day 28 at the time of the PSA analysis. Each mouse cytokine was identified by its gene name. Its UNIPROT Accession ID was used to obtain the molecular weight, which was adjusted by the number of receptors required for signal transduction. See Table 5-1. Method details for node analysis have been previously detailed in Chapter 2.
Table 5-1: Cytokines measured in Hemophilia A study

<table>
<thead>
<tr>
<th>Gene Name</th>
<th>UNIPROT Accession ID</th>
<th>UNIPROT MW of the unprocessed precursor</th>
<th># receptors reqd for signal</th>
<th>Weighted value for signal transduction</th>
<th>divide pg/ml by __ to get pmol/L</th>
</tr>
</thead>
<tbody>
<tr>
<td>G-CSF</td>
<td>Csf3 P09920</td>
<td>22,421</td>
<td>1.0</td>
<td>22421</td>
<td>22.4</td>
</tr>
<tr>
<td>GM-CSF</td>
<td>Csf2 P01587</td>
<td>16,091</td>
<td>1.0</td>
<td>16091</td>
<td>16.1</td>
</tr>
<tr>
<td>IFN-γ</td>
<td>Ifng P01580</td>
<td>17,907</td>
<td>2.0</td>
<td>35814</td>
<td>35.8</td>
</tr>
<tr>
<td>IL-1a</td>
<td>Iil1a P01582</td>
<td>31,023</td>
<td>1.0</td>
<td>30931</td>
<td>30.9</td>
</tr>
<tr>
<td>IL-1b</td>
<td>Iil1b P10749</td>
<td>30,931</td>
<td>1.0</td>
<td>30931</td>
<td>30.9</td>
</tr>
<tr>
<td>IL-2</td>
<td>Iil2 P04351</td>
<td>19,400</td>
<td>1.0</td>
<td>19400</td>
<td>19.4</td>
</tr>
<tr>
<td>IL-4</td>
<td>Iil4 P07750</td>
<td>15,834</td>
<td>1.0</td>
<td>15834</td>
<td>15.8</td>
</tr>
<tr>
<td>IL-5</td>
<td>Iil5 P04401</td>
<td>15,410</td>
<td>2.0</td>
<td>30820</td>
<td>30.8</td>
</tr>
<tr>
<td>IL-6</td>
<td>Iil6 P08505</td>
<td>24,384</td>
<td>1.0</td>
<td>24384</td>
<td>24.4</td>
</tr>
<tr>
<td>IL-7</td>
<td>Iil7 P10168</td>
<td>17,727</td>
<td>1.0</td>
<td>17727</td>
<td>17.7</td>
</tr>
<tr>
<td>IL-9</td>
<td>Iil9 P15247</td>
<td>16,075</td>
<td>1.0</td>
<td>16075</td>
<td>16.1</td>
</tr>
<tr>
<td>IL-10</td>
<td>Iil10 P18893</td>
<td>20,673</td>
<td>2.0</td>
<td>41346</td>
<td>41.3</td>
</tr>
<tr>
<td>IL-12 p35</td>
<td>Iil12a P43431</td>
<td>24,179</td>
<td>2.0</td>
<td>48358</td>
<td>48.4</td>
</tr>
<tr>
<td>IL-13</td>
<td>Iil13 P20109</td>
<td>14,108</td>
<td>1.0</td>
<td>14108</td>
<td>14.1</td>
</tr>
<tr>
<td>IL-15</td>
<td>Iil15 P48346</td>
<td>18,593</td>
<td>1.0</td>
<td>18593</td>
<td>18.6</td>
</tr>
<tr>
<td>IL-17</td>
<td>Ctra8 Q62386</td>
<td>17,490</td>
<td>2.0</td>
<td>34980</td>
<td>35.0</td>
</tr>
<tr>
<td>IP-10</td>
<td>Cxcl10 P17515</td>
<td>10,789</td>
<td>1.0</td>
<td>10789</td>
<td>10.8</td>
</tr>
<tr>
<td>KC</td>
<td>Cxcl1 P12850</td>
<td>10,254</td>
<td>1.0</td>
<td>10254</td>
<td>10.3</td>
</tr>
<tr>
<td>MCP-1</td>
<td>Ccl2 P10148</td>
<td>16,326</td>
<td>1.0</td>
<td>16326</td>
<td>16.3</td>
</tr>
<tr>
<td>MCP-1a</td>
<td>Ccl3 P10855</td>
<td>10,345</td>
<td>1.0</td>
<td>10345</td>
<td>10.3</td>
</tr>
<tr>
<td>RANTES</td>
<td>Ccl5 P30882</td>
<td>10,071</td>
<td>1.0</td>
<td>10071</td>
<td>10.1</td>
</tr>
<tr>
<td>TNF-α</td>
<td>Tnf P06804</td>
<td>25,896</td>
<td>3.0</td>
<td>77688</td>
<td>77.7</td>
</tr>
</tbody>
</table>

22 mouse (*Mus musculus*) cytokines were measured in pg/ml. For pathway generation, the median values of the *Significance Sets* were converted from pg/ml to a signal-transduction weighted pmol/L, which more accurately represented the signal strength.
Step 1. Dimensionality Reduction. Statistical analysis was performed using the non-parametric Mann–Whitney–Wilcoxon (MWW) test (www.spss.com) to identify the Significance Sets of molecules that differentiated the treatments on both days.

Step 2. Ingenuity Pathways Analysis (IPA) was used to find the likely neighboring pathways molecules (Ingenuity® Systems, www.ingenuity.com). Although the inputs were identified as mouse cytokines, the pathway generation algorithm was not limited by species in order to include information from other species’ pathways that may not have yet been discovered in mouse pathways.

Step 3. The IPA network diagrams were converted to temporal dependency matrices (TDMs), one for each day and each treatment.

Step 4. The TDMs were compared to identify molecules that differentiated the treatment of PBS or CFA/I.

Laboratory. Based on the PSA results, an immunoassay test would be performed on mouse supernatant for at least one of the predicted molecules at one time point to validate the PSA. The molecule(s) tested would depend on availability of the appropriate laboratory kit(s) to Dr. Csencsits Smith.

Results

PSA identified the Significance Sets of measured cytokines that statistically differentiated treatment of PBS (control) or CFA/I (p<.05, Mann Whitney) on day 7 as IL-5, RANTES (CCL5), and TNF-α; on day 14 as G-CSF, IL-1a, IL-12, IL-13, IP-10 (CXCL10). The median pmol/L values of the significant cytokines were entered into IPA for both
treatments on both days to generate the most likely biological pathways. Four network graphs were generated as shown in Figure 5-1, Figure 5-2, Figure 5-3, and Figure 5-4.

**Figure 5-1: PBS Day 7**

PBS is the “control” group. The molecules used for pathway generation are shaded: IL-5, RANTES (CCL5), and TNF-α (TNF).
CFA/I is the treatment by a novel tolerogenic protein. Both networks were generated based on the same *Significance Set* of 3 molecules; however, very different networks resulted due to the difference in the median values between treatments. In PBS, IL-5 had the lowest concentration whereas in CFA/I, TNF-α had the lowest concentration.
The molecules used for pathway generation on day 14 were G-CSF, IL-1a, IL-12, IL-13, IP-10 (CXCL10).
There was little difference in the evoked networks for day 14 although both networks were generated based on a larger Significance Set of 5 molecules than in day 7.
Lists of the evoked molecules were assembled and counted; there were a total of 79 molecules. See Table 5-2.

<table>
<thead>
<tr>
<th>Table 5-2: Evoked pathway molecule count</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td>-------------------------------</td>
</tr>
<tr>
<td>PBS only</td>
</tr>
<tr>
<td>CFA/I only</td>
</tr>
<tr>
<td>PBS &amp; CFA/I</td>
</tr>
</tbody>
</table>

79 molecules were evoked in the generated pathways for both outcomes on both days. Of these, 36 appeared solely in one treatment.

The unique day 7 molecules are shown in Figure 5-5. Day 14 evoked LGAL59 and PKCa/b for PBS and CCL6 & LBP under CFA/I treatment. Of particular interest on day 7 is the fact that PBS evokes defensin and TLR-2/TLR-4 and pro-inflammatory cytokines of the IL-1 family. CFA/I evokes GT1b ganglioside that may decrease expression of certain T-cell antigens; HLX, a transcription factor which may regulate the T_{H1} pathway, and pro-inflammatory cytokines of the TNF family plus IL-25 (IL-17E), a cytokine favoring the T_{H2}-type immune response. By day 14, even though there were more cytokines in the Significance Set than on day 7, the evoked networks only differed by two molecules each, suggesting that the effects of the CFA/I treatment were time limited.
Because of the study hypothesis that T cells that secrete IL-17 have been linked with the development of inflammatory disease, and the fact that IL-25 (IL-17E) – a cytokine that favors a response opposite to that of IL-17 (IL-17A-F) – was evoked in the CFA/I treatment on day 7, a validation test was run to compare the presence of IL-25 in the day 7 PBS and CFA/I supernatants.

**Laboratory Method and Results.** The immunoassay used was the R&D Systems Mouse IL-17E (IL-25) Duoset ELISA kit (www.rndsystems.com/pdf/DY1399.pdf). Mouse spleens were harvested at the indicated time, and lymphocytes isolated by mechanical dissociation followed by water lysis to remove red blood cells. CD4+ T cells were enriched using an EasySep Mouse CD4+ T cell enrichment kit (www.stemcell.com/en/Products/Popular-Product-Lines/EasySep.aspx), and 1 x 10^6 T
cells/ml were cultured for 72 hours in RPMI-1640 media (ATCC) supplemented with 10% fetal calf serum (FCS, Atlanta Biologicals) in the presence of 3 U/ml recombinant IL-2 (R&D Systems), 10 microg/ml FVIII (Kogenate FS, Bayer Pharmaceuticals), and 5 x 10^5/ml mitomycin c -treated feeder cells. Supernatants were harvested and stored at -80 degrees C before use in the ELISA.

Test results showed that the concentration of IL-25 was almost twice as high in the CFA/I treatment as in the control PBS treatment. See Figure 5-6. This suggests that the CFA/I treatment alters the immune signaling networks involved in the production and prevention of anti-FVIII antibodies in a mouse model of hemophilia A.
Discussion

The Pathway Semantics Algorithm used cytokine data from a mouse model of the disease hemophilia A to generate hypotheses about the progressive changes in the immune response in anti-factor VIII antibody production. The experimental finding of the predicted IL-25 molecule in the CFA/I treated group supports the study hypothesis that CFA/I modulates the T_{H17} response. Although IL-25 (IL-17E) has significant sequence homology to the IL-17 family (Lee et al., 2001), the molecule has very different properties. The laboratory result that IL-25 decreased in both the PBS control and the
CFA/I treatment to a level close to cell culture controls by day 28 is consistent with the *in silico* prediction that by day 14 CFA/I treatment effects would diminish. This is of interest because, although there were more measured molecules that statistically differentiated the treatments on day 14 than day 7, the evoked biological pathways on day 14 were almost identical. Based on this result, Dr. Csencsits Smith is planning experiments on the timing for repeated CFA/I treatments to manage inhibitor production because the peak response seemed to be at 7 days based on PSA.

Although these results are promising, there is much work to be done. In addition to the considerations given in Chapter 2 about the Pathway Semantics Algorithm and its reliance on external data, more laboratory validation is required. The presence of the one predicted molecule that supports the study hypothesis is a beginning. What PSA brings is the ability to use multi-cytokine panels of measured data to infer non-measured molecules and interactions. For the Laboratory Scientist, the Pathway Semantics Algorithm narrows the scope for exploration, thus reducing labor and costs associated with lab research and potentially increasing the chance for novel discovery.
Chapter 6 Conclusions and Suggestions for Future Research

Conclusions

Problem Statement. There is a need to connect bioassay data with pathway information within specific biomedical contexts and to facilitate comparison of biological pathways by time, outcome, and other stratifications. If these needs could be met, clinical research could start utilizing the wealth of constantly updated pathway information on a regular basis to generate baseline hypotheses for biological mechanisms, diagnosis, prognosis and therapy.

Problem Assessment. Based on an extensive literature survey of computational methods, there appears to be no generalizable, computable systems-level methods that utilized spatiotemporal bioassay data to answer biomedical questions arising from comparative analysis of biological pathways.

Research Goal. To address this need, the overarching aim of my dissertation project was to develop a computationally tractable and mathematically sound algorithm that enabled hypotheses generation about mechanisms of disease progression based on quantitative and qualitative data from molecular and clinical sources.

Research Deliverable. I developed a novel method called the Pathway Semantics Algorithm (PSA) that used matrix algebra to bridge the gap between biological and clinical resources by improving the in silico identification of clinically useful hypotheses about molecular patterns in disease progression. By framing biomedical questions within a variety of matrix representations, PSA has the flexibility to analyze combined quantitative and qualitative data over a wide range of stratifications. The resulting
hypothetical answers can then move to *in vitro* and *in vivo* verification, research assay optimization, clinical validation, and commercialization.

**Application Findings.** PSA identified differential temporal patterns of molecules and molecular interactions in pathways evoked from measured cytokines in two disease domains: shock / trauma and hemophilia A. The PSA results were validated by literature and expert opinion, and also by experiment for hemophilia A. In the node analysis described in Chapter 3, PSA identified and qualified 7 molecules in patterns across time of the progression of multiple organ failure; of these, only 3 had been previously associated with any shock / trauma syndrome. In the edge analysis described in Chapter 4, PSA identified different patterns of molecular interactions over time, outcomes, and functional relationships in biological networks evoked during the progression of multiple organ failure. Differences in the number of edges and unique edges for each stratification were uncovered, showing the changing temporal patterns of the molecular functions of activation, expression, inhibition, and transcription during the first 24 hours of trauma. In addition, the novel matrix-based measure XTALK gave quantitative insights into functional crosstalk in disease progression; the calculated results supported the existence of the property of degeneracy in biological networks.

**Algorithm Generalizations.** PSA confirmed that computationally tractable matrix algebra can integrate disparate quantitative and qualitative data for analysis over many stratifications, even where data is sparse, such as in shock / trauma studies. In addition, matrix algebra can facilitate computational time-based analysis that is usually thought of as requiring differential equations or complex time series. By discretizing the data into...
time periods related to clinical measurement intervals, molecular data analysis can be synchronized with the ongoing clinical care.

**Application Generalizations.** The Pathway Semantics Algorithm has been designed as a general method that can be applicable to a wide range of translational biomedical research. For example, stratifications over time, outcome, and interaction function were demonstrated; others could be used, such as cell cycle or sub-cellular molecular location if the data is available. The data requirements depend on the pathway generation software used; for IPA, a wide variety of data types are supported.

**Algorithm Limitations.** The Pathway Semantics Algorithm is built from validated quantitative components such as statistics and algebra, and the pathways come from qualitative peer-reviewed biomedical literature through the Ingenuity Knowledge Base. An extensive review of the literature, including the material in Chapter 2, showed that there are no standard methods for connecting bioassay data with pathway information within specific biomedical contexts or for comparing biological pathways over time and other stratifications. In addition, there seem to be no standard approaches for assessment of such methods and no established procedure to validate hypotheses derived from pathway models. It can be argued is that the gold standard is verification of hypotheses *in vivo*. Such an extended study is clearly beyond the scope of this research project. Instead, an informal content validity assessment of PSA by literature and expert opinion has been performed during a four-year period, during which time it has not been discredited; summary details follow in the validation section. Recently, PSA was confirmed experimentally in the hemophilia A study. Due to the network export limitations of the
IPA software and the resulting extensive manual labor currently involved, PSA has yet to be formally documented for estimates of reliability and validity (Higgins & Straub, 2006); it is planned to do this once PSA is automated and application results updated.

**Application Limitations.** Although these results provide insights into potential hypotheses that may be useful in disease progression, the quality of the PSA algorithm depends on quality data from assays, literature and biological pathway databases as well as the statistical and network algorithms used. Due to the small sample sizes in the disease domains under study, it was not feasible to perform a cross-validation of the results. In cross-validation, the input data is partitioned into training and test sets. The results would be analyzed to see if different hypotheses arise (Gutierrez-Osuna, 2009; W. Li, Arena, Sussman, & Mazumdar, 2003). In addition, the Pathway Semantics Algorithm was not formally assessed for reliability through a test-retest procedure (Nunnally, 1978) due to lack of automation; it is planned to do this in future applications.

The limitations of the PSA node analysis were detailed in chapters 3 and 4.

Summarizing, the results rely on:

- Quality of the patient data and the assay method
- Quantity of the patient data
- Dimensionality reduction through Significance Sets
- Choice of statistical analysis used to identify Significance Sets
- Quality of the biological pathway knowledge base and the algorithm used to evoke biological pathways based on assay measurements
- Changing nature of the biological pathway knowledge base
• Biological scope of the generated network
• Utility of the molecular patterns
• Resource requirements to implement PSA

Additional limitations resulted from the edge analysis described in Chapter 4:
• Age of pathway data
• Incomplete pathway data
• Linearity assumption

**Validation of Algorithm and Application Results.** The Pathway Semantics Algorithm was first presented for public critique in 2007. Since then, PSA, its applications and results have been informally reviewed and validated for content by multi-domain experts at conferences in systems and computational biology, trauma and critical care, operations research, pathology and numerical algebra. Presentations have been as follows; two awards have been received:

*Signaling Pathways in Multiple Organ Failure* (Poster).

• The Eighth International Conference on Systems Biology. October 1–6, 2007.
• Research Day, University of Texas Health Science Center at Houston. October 19, 2007.

*Temporal Analysis of Signaling Pathways in Multiple Organ Failure* (Presentation).

• 7th International Conference on Complexity in Acute Illness/International Shock Conference, Cologne, Germany, July 2008.
• 7th International Conference on Pathways, Networks, and Systems Medicine; Corfu, Greece. Aegean Conferences 2009.


Measurement Units May Impact Results of Pathway Analysis (Abstract)


Measuring Crosstalk in Biological Pathways (Poster)

• 2009 Keck Center Annual Research Conference, Houston, TX, October 2009.

• Research Day, University of Texas Health Science Center at Houston. November 20, 2009. (Poster Award: 2nd Place Student Clinical and Translational Research)

• Annual Meeting of the Association of Clinical Scientists, San Antonio, TX May 2010 (Presentation).


• INFORMS Annual Meeting 2010: Austin, TX, Nov 7–10, 2010. (Finalist Student Competition).

Uncovering Immune Signaling Networks Involved in Anti-FVIII Antibody Production in a Mouse Model of Hemophilia A Via Computational Analysis. (Poster)

• 2010 Keck Center Annual Research Conference, Houston, TX, October 2010.

123
Implications and Expected Contributions to Science. The use of matrix representation and numerical algebra, as used in the Pathway Semantics Algorithm (PSA), offers a way to computationally integrate and analyze qualitative and quantitative approaches for improved hypothesis generation about disease progression. PSA identifies and compares likely temporal molecular patterns in biological pathways derived from patients’ data, an important benefit that supports personalized medicine and that may reduce the costs of disease research by narrowing the scope towards more likely hypotheses.

This dissertation research project advances science in two areas: informatics and translational biomedicine. Informatics is advanced by the demonstration that Pathway Semantics facilitates the integration of qualitative and quantitative data, and along with statistical and mathematical processing, strengthens the information content upon which hypotheses are generated. The methodology appears to be repeatable, generalizable, scalable and extendable. Translational biomedicine is advanced because Pathway Semantics is designed to overcome the knowledge barrier between the ability to measure abundant quantitative molecular data and the ability to connect that data with qualitative biological pathway mechanisms that evoke hypotheses of clinical and biological significance.

Suggestions for Future Research

Algorithm. Additional software development is required to automate the Pathway Semantics Algorithm for node and edge analysis in general. As a composite method, PSA now relies on commercial software processing components for statistical analysis,
algebraic calculations, and biological network generation. The first two are generally available as open source links; however, network generation algorithms and pathway data are usually proprietary. Current intellectual property restrictions by Ingenuity Systems have required extensive workarounds to utilize the data in the Ingenuity Knowledge Base in a manner acceptable under the license provisions and also useful for PSA; it is expected that meetings with Ingenuity Systems will continue on how to resolve this. Meanwhile, other sources of biological pathway data and network generators are under review. Once PSA is automated, future enhancements include support of data perturbation analysis combined with support for probabilistic scenarios. In perturbation, or sensitivity, analysis, the input data is changed incrementally and the stability of the output hypotheses is assessed (Geard, Willadsen, & Wiles, 2005). For probabilistic scenarios, the generated hypotheses can be strengthened by Bayesian inference, a form of inductive reasoning commonly used in medical decision-making to assess the value of a diagnostic test (Sox, Blatt, Higgins, & Marton, 1988). As an example, instead of using median cytokine values to generate representative signaling pathways, a researcher could use 75% values and see how the evoked pathways change. Combined with a sensitivity analysis of biological pathways evoked from bioassay data, PSA could then be used for “rule-in”/”rule-out” guidelines for clinical interpretation of molecular signaling data. 

**Application.** I plan to continue analysis of disease progression in shock /trauma with re-runs of all cytokine data collected in the original study. This will update the generated biological pathways with the latest research, and add additional patient data measurements beyond the first 24 hours already studied. The hypothesis is that biological
pathways evoked in the first 24 hours from trauma differ from those in later times due to changes in systemic response. The hemophilia A study continues, with the intent of using patients’ cytokine data as they become available.

In a very different way, I plan to apply PSA to analysis of cancer progression within the context of molecular tumor profiling. In chronic conditions such as cancer, the time course is slower and the focus is on the molecules within the signaling pathways in tumors that may be targeted by drugs. Instead of time, the component molecules in the signaling pathways need to be evaluated by their cellular location, such as cytoplasm or nucleus, and their cell cycle phase to determine cellular dysfunctions. Measurements from molecular tumor profiling (using MorphoproteomicSM analysis by immunohistochemistry(Tan, 2008)) include expression of signaling molecules in compartment locations and analytes relating to cell cycle phase. This approach is used for patients with tumors that have no established protocol or that have not responded to conventional therapy(Robert E. Brown, Lun, Prichard, Blasick, & Zhang, 2004). Here there is a need for an analysis of signaling pathways in tumor profiles that compares molecular expression across specific locations and cell cycle phase for a better understanding of the pathways that differentiate cancers. This will assist in the choice of therapy for hard-to-manage cases.

**Summary**

The fundamental contribution of this research project is the methodology called the Pathway Semantics Algorithm (PSA) as one solution towards the goals of connecting bioassay data with pathway information within specific biomedical contexts and
facilitating comparison of biological pathways by time, outcome, and other stratifications to further diagnosis, prognosis and treatment for disease progression. This is a novel approach to meet the need for both hypothesis and data driven strategies for result analysis and interpretation of clinically derived data (Ghazal, 2008); this work advances both informatics and translational biomedicine.


international conference on formal methods for the design of computer, communication, and software systems.,


NLM. (2008b). (MESH: ("Multiple organ failure"[All Fields] OR "mods"[All Fields] OR "multiple organ dysfunction syndrome"[All Fields]) AND ("signaling pathways"[All Fields] OR "signaling pathway"[AllFields]) (Publication., from National Library of Medicine:


Viladrich, M., Sanchez, O., Soley, M., & Ramirez, I. (2008). Alterations in liver parenchyma after sialoadenectomy in mice: contribution of neutrophils and


PATHWAY SEMANTICS:
AN ALGEBRAIC DATA DRIVEN ALGORITHM TO GENERATE HYPOTHESES
ABOUT MOLECULAR PATTERNS UNDERLYING DISEASE PROGRESSION
Mary Frances McGuire, PhD
The University of Texas School of Biomedical Informatics at Houston, 2011
Primary Advisor: M. Sriram Iyengar, PhD

Note: Due to the detailed images, this Appendix is best viewed as a pdf file.
Chapter 1

Theoretical / conceptual framework
Chapter 3

Adjusted weightings for IPA: pg/ml to pmol/L

Note: calculations are only for the 17 cytokines in the Significance Sets $S_1$

<table>
<thead>
<tr>
<th>Gene Name</th>
<th>UNIPROT ID</th>
<th>MW of the unprocessed precursor</th>
<th># receptors reqd for signal</th>
<th>Weighted value for signal transduction</th>
<th>divide pg/ml by ___ to get pmol/L</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eotaxin</td>
<td>CCL11</td>
<td>P51671</td>
<td>10732</td>
<td>1.0</td>
<td>10732</td>
</tr>
<tr>
<td>G-CSF</td>
<td>CSF3</td>
<td>P09919</td>
<td>22293</td>
<td>2.0</td>
<td>44586</td>
</tr>
<tr>
<td>GM-CSF</td>
<td>CSF2</td>
<td>P04141</td>
<td>16295</td>
<td>1.0</td>
<td>16295</td>
</tr>
<tr>
<td>IFN-γ</td>
<td>IFNG</td>
<td>P01579</td>
<td>19348</td>
<td>0.5</td>
<td>9674</td>
</tr>
<tr>
<td>IL-1ra</td>
<td>IL1RN</td>
<td>P18510</td>
<td>20055</td>
<td>1.0</td>
<td>20055</td>
</tr>
<tr>
<td>IL-5</td>
<td>IL5</td>
<td>P05113</td>
<td>15238</td>
<td>1.0</td>
<td>15238</td>
</tr>
<tr>
<td>IL-6</td>
<td>IL6</td>
<td>P05231</td>
<td>23718</td>
<td>1.0</td>
<td>23718</td>
</tr>
<tr>
<td>IL-7</td>
<td>IL7</td>
<td>P13232</td>
<td>20187</td>
<td>1.0</td>
<td>20187</td>
</tr>
<tr>
<td>IL-8</td>
<td>IL8</td>
<td>P10145</td>
<td>11098</td>
<td>1.0</td>
<td>11098</td>
</tr>
<tr>
<td>IL-9</td>
<td>IL9</td>
<td>P15248</td>
<td>15909</td>
<td>1.0</td>
<td>15909</td>
</tr>
<tr>
<td>IL-10</td>
<td>IL10</td>
<td>P22301</td>
<td>20517</td>
<td>1.0</td>
<td>20517</td>
</tr>
<tr>
<td>IL-13</td>
<td>IL13</td>
<td>P35225</td>
<td>14319</td>
<td>1.0</td>
<td>14319</td>
</tr>
<tr>
<td>IP-10</td>
<td>CXCL10</td>
<td>P02778</td>
<td>10881</td>
<td>1.0</td>
<td>10881</td>
</tr>
<tr>
<td>MCP-1</td>
<td>CCL2</td>
<td>P13500</td>
<td>11025</td>
<td>1.0</td>
<td>11025</td>
</tr>
<tr>
<td>MIP-1β</td>
<td>CCL4</td>
<td>P13236</td>
<td>10212</td>
<td>1.0</td>
<td>10212</td>
</tr>
<tr>
<td>RANTES</td>
<td>CCL5</td>
<td>P13501</td>
<td>9990</td>
<td>1.0</td>
<td>9990</td>
</tr>
</tbody>
</table>
Biological Networks generated from IPA in September 2009 (based on 27 cytokines)

Graphic images were exported on January 17, 2010. Measured input molecules in pink to red; the darker the color, the higher the concentration. Solid lines are direct interactions; dotted lines, indirect interactions. Here line color has no meaning; it resulted from merging several networks.
<table>
<thead>
<tr>
<th>3-9 MOF hours 18-22</th>
</tr>
</thead>
<tbody>
<tr>
<td>3-10 NMOF hours 18-22</td>
</tr>
</tbody>
</table>
Chapter 4

Biological Networks generated from IPA in July 2007 (based on 11 cytokines)

Graphic images were exported in January, 2011. Measured input molecules in pink to red; the darker the color, the higher the concentration. Solid lines are direct interactions; dotted lines, indirect interactions. Here, orange lines highlight the interactions involved in the functional relationship under examination, such as activation.
4-7 MOF hours 6-10,

Transcription

4-8 NMOF hours 6-10,

NONE

Transcription
4-9 MOF hours 10-14, Activation

4-10 NMOF hours 10-14,
Expression

4-11 MOF hours 10-14,

Expression

4-12 NMOF hours 10-14,
Expression

4-19 MOF hours 22-24,

Expression

4-20 NMOF hours 22-24,
Inhibition

4-21 MOF hours 22-24,

Inhibition

4-22 NMOF hours 22-24,
4-23 MOF hours 22-24,

Transcription

4-24 NMOF hours 22-24,

Transcription