



Virginia Commonwealth University  
**VCU Scholars Compass**

---

Theses and Dissertations

Graduate School

---

2010

# A Novel Method to Detect Functional Subgraphs in Biomolecular Networks

Sterling Thomas

*Virginia Commonwealth University*

Follow this and additional works at: <http://scholarscompass.vcu.edu/etd>

 Part of the [Life Sciences Commons](#)

© The Author

---

Downloaded from

<http://scholarscompass.vcu.edu/etd/154>

This Dissertation is brought to you for free and open access by the Graduate School at VCU Scholars Compass. It has been accepted for inclusion in Theses and Dissertations by an authorized administrator of VCU Scholars Compass. For more information, please contact [libcompass@vcu.edu](mailto:libcompass@vcu.edu).

# **A Novel Method to Detect Functional Subgraphs in Biomolecular Networks**

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at Virginia Commonwealth University

By  
Sterling Wells Thomas  
Bachelor of Science, Old Dominion University 2006

Program Director: Robert M. Tombes, Ph.D.  
Dissertation Advisor: Danail Bonchev, Ph.D., Sc.D.

Virginia Commonwealth University  
Richmond Virginia  
December, 2010

## Acknowledgement

The author wishes to thank several people. I would like to thank my wife, Charisse for her patience, love and support over the last five years. I would like to thank my son, Coulter, for his patience and understanding while I was in graduate school. I would like to thank my parents for their love, support and guidance in preparing me for becoming a scientist. Finally, I would like to thank my advisor, Dr. Bonchev, for teaching me the skills I needed to complete this work, directing me through the process that has produced this dissertation, and for all his help and guidance over the past six years.

## Table of Contents

Chapter 1: Introduction.....	1
Significance.....	3
Innovation.....	7
<b>Chapter 2: Network Construction.....</b>	<b>9</b>
Biomolecular Databases:.....	10
Construction of Networks.....	11
Analysis of Optimized Network:.....	20
<b>Chapter 3: Network Simulation of Bcl2 Family Proteins.....</b>	<b>24</b>
Bcl2 Simulations:.....	26
<b>Chapter 4: Differential Analysis.....</b>	<b>35</b>
Summary.....	35
Analysis of Microarray Data from Lung Cancer Patients.....	35
<b>Chapter 5: Autism Networks.....</b>	<b>45</b>
Summary.....	47
Analysis of Protein Interaction Networks.....	47
Network comparison.....	49
<b>Chapter 6: Discussion and Future Work.....</b>	<b>51</b>
Summary.....	51

## Table of Figures:

Figure 1 - Visual model of ideal master/slave relationship described in the hypothesis. ....	2
Figure 2 – Venn diagram of PSI-MI Tags from Apoptosis Network (PSI standard introduced by HUPO) .....	12
Figure 3 - Largest apoptosis subnetwork from PSICQUIC search of apoptosis associated biomolecules. ....	13
Figure 4 - 3D clustering of key terms from PSICQUIC search using starlight visualization tool. ....	15
Figure 5 - Word cloud representation of interaction data from PSICQUIC search of apoptosis. This search was compared with the Starlight search to identify which experimental methods could be excluded to increase the quality of the apoptosis network. ....	16
Figure 6 – Visual representation of the optimized apoptosis network produced using PSICQUIC enabled tools. ....	20
Figure 7 - Log Distribution Histograms of Node Degrees from Apoptosis Network (Top is from the complete network resulting from the search, bottom is the optimized network with unknowns and low scoring biological evidence removed) .....	22
Figure 8 - 3D Log Plot of Centroid Value by Node Degree by Occurrence number in Apoptosis Network. (Top is from the complete network resulting from the search, bottom is the optimized network with unknowns and low scoring biological evidence removed).....	23
Figure 9 – First model of the Bcl-2 network used in cellular automata simulation of proteins associated with apoptosis.....	28
Figure 10 – Second model of the Bcl-2 network used in cellular automata simulation of proteins associated with apoptosis. This model expanded on the second model by including the Bcl-2 homodimer reported in co-immunoprecipitation studies. ....	28
Figure 11 – The third model: A different version of the second Bcl-2 model where the Bcl-2 homodimer was replaced with a Bcl-2, Bcl-X heterodimer. ....	29
Figure 12 – The third Bcl-2 model used in cellular automata simulation. This model combined the two versions of the second model where Bcl-2 can produce a homodimer, and a heterodimer with Bcl-X.....	29
Figure 13 – This expanded Bcl-2 network included interacting partners that were not known to be Bcl-2 family members, but were similar in structure and had been reported as interacting partners with known Bcl-2 family proteins. ....	31
Figure 14 – 3-Node, 2 Edge Linear Motif .....	34
Figure 15 – 4-Node 4-Edge Biparallel Motif.....	34
Figure 16 – Normalized expression values were plotted after being sorted (low to high) by different clinical features. Each color represents a different clinical feature use for sorting the expression values. ....	37
Figure 17 – This network was produced using Ingenuity Systems IPA network analysis tool. A list of proteins/genes was entered into the search tool and this network was produced. The list of proteins/genes was produced using the correlation analysis described in the text. ....	39
Figure 18 – 4-Node, 4-Edge Non-Symmetrical Non-Linear Motif .....	45
Figure 19 – 4-Node, 4-Edge Symmetrical Non-Linear Bi-Fan Motif .....	45
Figure 20 – 4-Node, 4-Edge Bi-parallel Motif .....	45
Figure 21 – 6-Node, 9-Edge Tri-parallel Motif .....	45
Figure 22 – 5-Node, 6-Edge Tri-parallel Motif .....	45

Figure 23 - Autism network produced using the NCBI search tool. This was the only search target and tool that was able to produce a network using search terms associated with autism...	48
Figure 24 – 4-Node, 4-Edge Non-Symmetrical Non-Linear Motif .....	53
Figure 25 – 4-Node, 4-Edge Symmetrical Non-Linear Bi-Fan Motif .....	53
Figure 26 – 4-Node, 4-Edge Bi-parallel Motif .....	53
Figure 27 – 6-Node, 9-Edge Tri-parallel Motif .....	53
Figure 28 – 5-Node, 6-Edge Tri-parallel Motif .....	53

## Table of Tables:

Table 1 – Survey of publicly available sources of biomolecular interactions used in network analysis.....	10
Table 2 – Occurrence of words describing the biological evidence of each interaction in the apoptosis network. ....	14
Table 3 – PSI-MI returned during the PSICQUIC search for interactions associated with apoptosis. ....	17
Table 4 – List of interacting partners include in the expanded Bcl-2 network (Figure 13).....	32
Table 5 – List of probabilities created for each pathway found in the expanded Bcl-2 network (Figure 13) .....	32
Table 6 - Ontology of chromatin genes identified as potential biological markers using the analysis described in the text. These markers could be used to discriminate between diseased and healthy tissue.....	39
Table 7 – Results of the motif search using NetMatch plugin with Cytoscape against the apoptosis network. All the ratios are low when compared to the mean network counts from all 1000 random networks.....	46
Table 8 - Results of the motif search using NetMatch plugin with Cytoscape against the autism network. All the ratios are high when compared to the mean network counts from all 1000 random networks.....	49
Table 9 - Results of the motif search using FANMOD against the autism network. All the ratios are high when compared to the mean network counts from all 1000 random networks. ....	50

## Abstract

### A NOVEL METHOD TO DETECT FUNCTIONAL SUBGRAPHS IN BIOMOLECULAR NETWORKS

Sterling Wells Thomas, Doctor of Philosophy

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at Virginia Commonwealth University

Virginia Commonwealth University, 2010

Danail Bonchev, Ph.D., Sc.D., Senior Fellow, Professor and Director of Research on Bioinformatics, Center for the Study of Biological Complexity

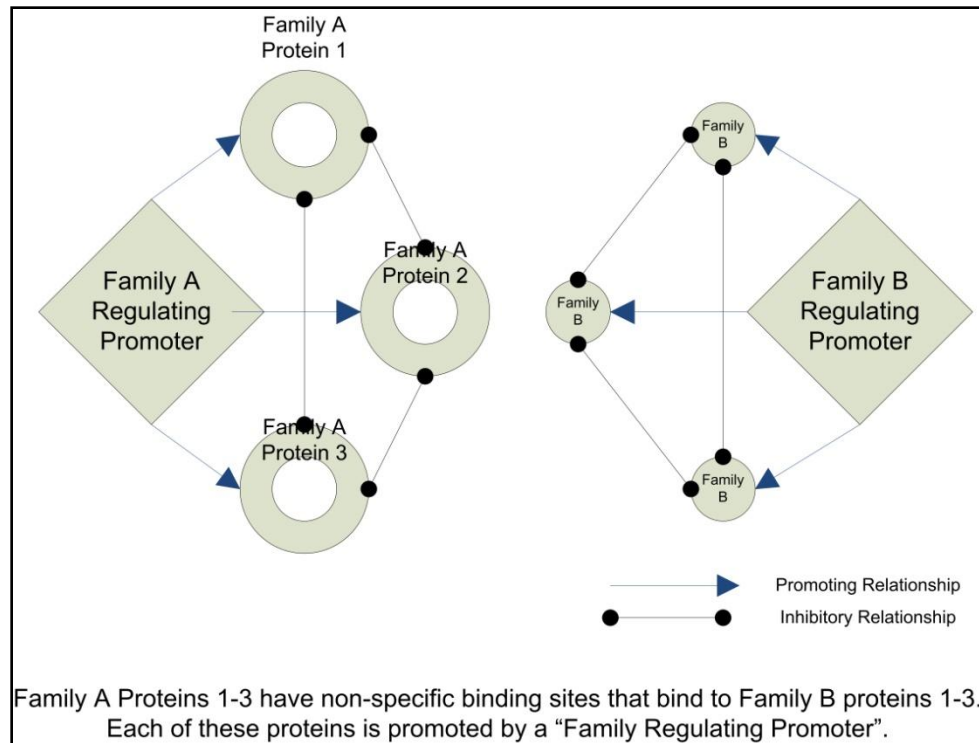
Several biomolecular pathways governing the control of cellular processes have been discovered over the last several years. Additionally, advances resulting from combining these pathways into networks have produced new insights into the complex behaviors observed in cell function assays. Unfortunately, identification of important sub-networks, or “motifs”, in these networks has been slower in development. This study focused on identifying important network motifs and their rate of occurrence in two different biomolecular networks. The two networks evaluated for this study represented both ends of the spectrum of interaction knowledge by comparing a well defined network (apoptosis) with and poorly studied network that was early in development (autism). This study identified several motifs that could be important in governing and controlling cellular processes in healthy and diseased cells. Additionally, this study revealed an inverse relationship when comparing the occurrence rate of these motifs in apoptosis and autism.

## Chapter 1: Introduction

Over the past decade scientists have discovered many genetic/molecular pathways responsible for the control of cellular processes. The purpose of this control is to activate or deactivate processes responsible for specific functions in the cell that are critical for cell survival, but not needed during the every step of the cell cycle. These cellular functions under control include cell cycle progression and division, DNA damage repair, apoptosis, synthesis and many more. Additionally, the controlling pathways for these functions have a significantly higher rate of mutation or deletion in solid malignant tumors. Resulting from the discovery of signaling pathways and their high rate of mutation and deletion, many new diagnostic and therapeutic studies are focused on testing and repairing these signaling networks. One of the major challenges of these studies is identifying all the networks formed by overlapping and redundant signaling pathways commonly found in *Homo sapiens*. During my preliminary studies it has been observed that biomolecular networks responsible for signaling and control of apoptosis include several closely related proteins termed *slave* that have similar functions but competitively suppress the activation of a controlling protein (termed *master*). The slave/master proteins thus form a redundant activation system that is resistant to minor mutations or deletions.

The hypothesis of this thesis is a master/slave arrangement in signaling and regulatory networks represents a specific network motif that can be used to identify other signaling and controlling networks as targets for diagnostics or therapeutics.





**Figure 1 - Visual model of ideal master/slave relationship described in the hypothesis.**

The following four stages of this study will aid in verifying motifs are influenced by both evolutionary pressures and functional pressures and whether this hypothesis holds true.

**Stage 1:** Construct a network of cellular gene/protein agents and potential molecular regulatory pathways specific for apoptosis including the integrated Bcl2 family of genes and proteins from interactions reported in BioGRID, Cancer Cell Map, Human Protein Reference Database, IntAct, MINT, NCI/Nature Pathway Interaction and Reactome databases. Network nodes and interactions will be identified using mRNA expression data from gene arrays based on whole-blood samples from individuals diagnosed with adenocarcinoma of the lung obtained from the Oncomine and Gene Expression Omnibus (GEO) databases.

**Stage 2:** Model and simulate Bcl2 family genes and protein networks to identify subgraphs using cellular automata driven agent-based modeling. Calculate topological measures of the discovered subgraphs to identify those that follow the hypothesized master and slave node model. Topological measures include

but are not limited to connectivity (Zagreb M2 and Betweenness), node degree, number of second neighbor nodes ( $k_1$  and  $k_2$ ), number of first and second neighbor edges ( $A(k)$ ). Develop a mathematical model describing the hypothesized relationship to be used in Stage 3 and 4.

**Stage 3:** Identify a differential network of cellular gene/protein agents and potential gene regulatory pathways specific for functional changes associated with the progression of adenocarcinoma of the lung. Network nodes and interactions will be identified using mRNA expression data from gene arrays based on whole-blood samples from individuals diagnosed with adenocarcinoma of the lung obtained from the Oncomine and Gene Expression Omnibus (GEO) databases. Identify networks subgraphs that follow the proposed hypothesis.

**Stage 4:** Identify a differential network of cellular gene/protein agents and potential gene regulatory pathways specific for functional changes associated with the progression of Autism Spectrum Disorder (ASD). Network nodes and interactions will be identified using mRNA expression data from gene arrays based on whole-blood samples from individuals diagnosed with ASD and obtained from the Autism Genome Resource Exchange (AGRE). Topological analysis will be limited to identification of subgraphs that follow the hypothesized functional structure from Aim 1.

Significance:

Lung Adenocarcinoma is the most common Non Small Cell Lung Cancer NSCLC and is the most common lung cancer among in life-long non-smokers. Lung cancer occurs in 62.5 of every 100,000 men and women in the United States, and is one of the most common occurring cancers (SEER). Lung Cancer commonly metastasizes and is found distally at diagnosis in 56% of patients. Lung Cancer detected distally at diagnosis has a 5-year survival rate only of 3.5% (SEER). Adenocarcinoma accounts for 40% of the tumors at diagnosis and

represents a significant challenge to the United States medical community. The average age at diagnosis is 71 years, and although occurrence has recently declined by 0.8%, is expected to increase with the aging “Baby-Boomer” generation.

The Genetic Epidemiology of Lung Cancer Consortia (<http://epi.grants.cancer.gov/Consortia/>) recently discovered potential familial lung cancer gene (Bailey-Wilson, Amos et al. 2004). Although many lung tumors are the result of environmental exposure (smoking), adenocarcinoma of the lung most commonly occurs in non-smokers and suggests a familial link. The study identified regions on chromosome 6 showing alterations in a significant number of first degree family members with lung tumors, and less significant alterations on chromosomes 12, 14, and 20. These are considered milestones in the oncogenesis of lung cancer and do not represent a single hit cause of the disease.

Recently, Drs. Bonchev and Kuznetsov (Kuznetsov, Thomas et al. 2008) developed a unique method to identify early detection biomarkers in Adenocarcinoma of the Lung. In this study Kuznetsov et al. identified a 5 gene biomarker that could be used to classify lung cancer using mRNA expression. The 5 gene marker set was revealed to represent cross-talk of 25 biomolecular pathways in lung cancer. Based on the network analysis, these 25 pathways have significant influence on the state of disease and control of cell growth and apoptosis. This type of study is a good representation of amount of biomarker research done in the lung cancer field. Based in the significant genetic and proteomic research done on lung cancer, it represents a good model for this study.

Autism Spectrum Disorders (ASD), which include Autism, Asperger’s Disorder, and Pervasive Developmental Disorder-Not Otherwise Specified (PDD-NOS), are severe, extensive neurodevelopmental diseases characterized by deficits in social and emotional interactions and communication, as well as the presence of stereotypic behavior and restricted interests. Recently, the Centers for Disease Control along with Health Resources and Services Administration (HRSA)

reported that one in 100 children is affected with an ASD, and declared the condition “an urgent public health issue” [United States Centers for Disease Control, 2009(2007)]. Clinicians are now faced with a large population of children and teenagers with ASD who require accurate and comprehensive assessment and effective treatments. Understanding the biological underpinnings, including genetic characterization, of this group of disorders will directly impact diagnosis and treatment development. In addition, costs associated with healthcare and with special education of children with autism are enormous and have far ranging consequences on society.

Twin and epidemiological studies have shown that ASD is the most frequently inherited psychiatric disease. Understanding the biological factors that influence the course and outcome of ASD is complicated by clinical heterogeneity in this group of disorders, by multiple gene etiologies that predispose an individual to the disease and by the innate complexity of dynamic gene/protein interactions. Identification of cellular protein biomarkers and gene regulatory pathways associated with ASD is essential to assess accurately the clinical prognosis and allow for the earliest possible intervention to correct behavioral anomalies associated with the disease. However, limited information of protein interaction networks and complexes involved in ASD is available.

Insel and Abrahams both recently noted that understanding affected proteins and their associated cellular pathways in ASD and other psychiatric disorders is a necessary step that will improve diagnostic tests as well as treatments [(Insel and Lehner 2007), (Abrahams and Geschwind 2008)]. Inroads relating to analysis of genes involved in this complex phenotype have recently been reported. Transcription profiling through the use of cDNA microarrays has permitted simultaneous analysis of thousands of genes whose expression is either increased or decreased in individuals with ASD [(Purcell, Jeon et al. 2001), (Yonan, Palmer et al. 2003), (Baron, Liu et al. 2006), (Baron, Tepper et al. 2006), (Hu, Frank et al. 2006), (Walker, Segal et al. 2006), (Nishimura, Martin et al. 2007), (Gregg, Lit et al. 2008)]. A bioinformatics analysis by Yonan et al. (Yonan, Palmer et al. 2003) of autism positional candidate genes using biological

databases identified 383 candidate genes predicted by a genomewide genetic linkage analysis of families that had two or more members diagnosed with ASD. In addition, peripheral blood and lymphoblastoid cell lines from autistic patients have been used successfully to identify autism-associated gene changes in peripheral cells (Yonan, Palmer et al. 2003; Baron, Liu et al. 2006; Baron, Tepper et al. 2006; Hu, Frank et al. 2006; Walker, Segal et al. 2006; Nishimura, Martin et al. 2007). Hu et al. (Hu, Frank et al. 2006) demonstrated that lymphoblastoid cell lines from monozygotic twins discordant for severity of autism have differential gene expression patterns as determined by microarray analysis. Furthermore, genes with the greatest degree of differential expression were those associated with nervous system development, structure and function. Many of these same genes mapped to chromosomal regions determined in earlier studies to be associated with autism candidate genes. A recent report by Gregg et al. (Gregg, Lit et al. 2008) also characterized gene expression differences in blood leukocytes of ASD patients using microarrays and found a small group of genes expressed predominately in natural killer cells, among others. Further, Nishimura et al. (Nishimura, Martin et al. 2007) utilized lymphoblastoid cell lines from autism patients and identified 68 genes that were dysregulated in common between autism with FMR1-FM and dup(15q). This study (Nishimura, Martin et al. 2007) also demonstrated increased expression of FMR1 interacting protein 1 (CYFIP1) in dup(15q) individuals, which suggests a putative link between FMR1-FM and dup(15q).

Recently, Wall et al. (2009)(Wall, Esteban et al. 2009) compared gene networks associated with autism to those of 432 additional neurological diseases with the intent of defining shared molecular mechanisms and identifying new genes that were important in autism. Sixty-six candidate autism genes were linked to one other disorder—these were referred to as multi-disorder gene set (MDAG). Gene networks and protein interactions with all MDAG genes were then defined, which extended the potential autism-specific genes to 334 candidates. Of these 334 genes, 154 genes had not previously been linked to autism, but were differentially expressed in autistic individuals. The authors concluded that at least

a fraction of these genes may act as “sub-components” of the autism gene network and will provide insights into current gaps of our knowledge of this disease. Along this line, Purcell et al. (Purcell, Jeon et al. 2001) used microarrays to study gene expression on post-mortem brain tissue from individuals with autism and matched controls without any symptoms of autism. The study demonstrated differential regulation of 30 genes in individuals with autism compared to matched controls. Overall, these results affirmed that autism results from the dysregulation of multiple genes, which is predicted to have a profound effect on brain development. Taken together, understanding the biological underpinnings, including genetic characterization and functional pathways, of this group of disorders will directly impact diagnosis and treatment development.

### **Innovation**

This proposal is innovative because it proposes a new method to determine functional groups in signaling and protein interaction networks. The new method combines network topology analysis using existing methods and equations with simulation for the purpose of classifying each node in the network as master or slave. Existing network analysis tools provide an ideal way to classify the nodes in a network as central or peripheral (eccentric). These analyses make this classification based entirely on the structure of network, assuming there are no edge weights. This method is ideal for networks of computers where each node shares basic characteristics. This method is also ideal for early analysis of biological networks (signaling and protein interaction) to determine what nodes are central assuming each node share based characteristics. The shortcoming of this method is that in biological networks each node does not always share basic characteristics. An example of this is demonstrated by the diversity of protein structure and functionality found in most protein interaction networks. This proposal addresses this shortcoming by combining the metrics described above with cellular automata simulation to provide an additional classification of master and slave. The master/slave relationship in the network topology might be considered by analogy with central

and peripheral nodes, but provides an additional level of information where there are multiple redundant pathways sharing the same or similar function.

Additionally, simulation combined with existing topology metrics being used to describe the functionality of ASD progression networks has never been done and is badly needed.

## Chapter 2: Network Construction

### Summary:

**This chapter is about construction of a network of cellular gene/protein agents and potential molecular regulatory pathways specific for apoptosis including the integrated Bcl2 family of genes and proteins from interactions reported in BioGRID, Cancer Cell Map, Human Protein Reference Database, IntAct, MINT, NCI/Nature Pathway Interaction and Reactome databases. Searching and extracting interaction data from databases represent the most common way of constructing biomolecular networks. This chapter describes the individual methods and techniques used to construct and maintain each of these databases.**

The current standard for creating visual representations of interacting molecules is to search curated databases of interacting molecules derived from peer-reviewed publications. The process of populating these databases is called “natural language searching”, a technology popularized with the rise of the Internet. Natural language searching includes two major processes: 1. Identifying and extracting interaction data from peer-reviewed publications and 2. Indexing based on an existing Molecular Interaction ontology of the Proteomics Standards Initiative (PSI-MI) (Montecchi-Palazzi, Kerrien et al. 2009). The resulting record represents a binary interaction with supporting data (experiment type and reference id). A third party technology (i.e. Cytoscape) can then search these binary records and create n-ary interaction map describing linkages between binary records that produce biological interaction/signaling pathways.



### Biomolecular Databases:

Biomolecular interaction databases store data submitted by experimentalists and extracted from peer-reviewed publications. The data extraction process is sometimes automated using optical character recognition software with a list of key words that describe interactions. When not automated, the extraction process is done manually. Subject matter experts (SME) read publications that describe biomolecular interactions. When enough evidence is described in the publications the SME will enter the interaction into the database they are curating for. Over time updated versions of the databases are released to the scientific community. Table 1 describes the data-capture technique and update schedule for each of the databases used for this work.

Table 1 – Survey of publicly available sources of biomolecular interactions used in network analysis.

<b>Database Name</b>	<b>Data Collection Method</b>	<b>Update Schedule</b>	<b>References</b>
BioGrid thebiogrid.org	Manually Populated and Curated	Monthly	(Stark, Breitkreutz et al.)
ChEMBL Ebi.ac.uk/chembl/	Automated	Periodic (unscheduled)	(Overington 2009)
DIP Dip-doe-mbi.ucla.edu	Manually Populated and Curated	Periodic (unscheduled)	(Sprinzak, Cokus et al. 2009)
INTACT Ebi.ac.uk/intact/	Manually Populated and Curated	Weekly	(Aranda, Achuthan et al.)
IRefIndex Irefindex.uio.no	Manually and Automated (Meta-DB)	Continuous	(Razick, Magklaras et al. 2008)

MatrixDB Matrixdb.idcp.fr	Manually and Automated (Meta-DB)	Six Months	(Chautard, Ballut et al. 2009)
MINT Mint.bio.uniroma2.id/mint/	Manually Populated and Curated	Yearly	(Ceol, Chatr Aryamontri et al.)
MPIDB Jcvi.org/mpidb/	Manually Populated and Curated	Yearly	(Goll, Rajagopala et al. 2008)
Reactome Reactome.org	Manually Populated and Curated	Continuous	(Croft, O'Kelly et al.)

### Construction of Networks

Using the newly released Cytocape version 2.7.0, an interaction database search was conducted using the PSICQUIC Universal Web Service Client. PSICQUIC is a Google sponsored implementation of the PSI-MI that provides programmatically accessible molecular interaction data through a standardized web service. The database schema, XML and web service client was to be installed locally at VCU to allow unpublished interaction data to be integrated into the database that was to be created by mRNA data. This became unnecessary and will be described in a later section. PSICQUIC provides access to BioGrid, ChEMBL, DIP, IntAct, IRefIndex, MatrixDB, MINT, MPIDB and Reactome databases. To identify records important for this study the keyword apoptosis was used. As of writing this dissertation, a keyword search of PSICQUIC produced 15396 records. The search tool produced individual networks from each of the data sources which were merged, and their overlap used as a basic network. A cursory analysis of the search terms was done to evaluate the validity of a semantic based analysis. Figure 2 describes the bases of this analysis where terms representing key experimental methods represent overlaps that can be used to improve the validity of a biomolecular network.

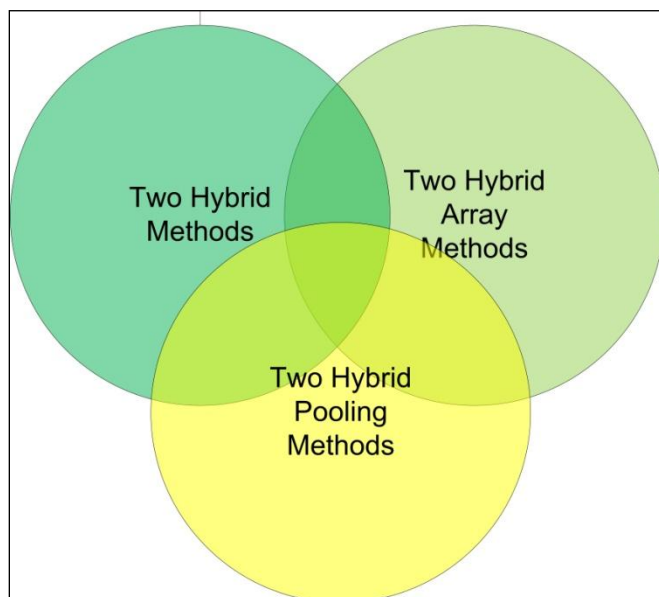


Figure 2 – Venn diagram of PSI-MI Tags from Apoptosis Network (PSI standard introduced by HUPO)

IntAct is a database of interactions between bio-molecules ranging from large protein complexes to small ligands and synthesized molecules. The search included one term, Apoptosis, and produced a network of 6393 nodes and 14614 interactions (as defined by PSI Molecular Interaction standard) (Figure 3 The nodes included 31 small molecules, 14 genbank sequences, 21 ensemble entities, 3 genbank nucleolus entities, 90 genbank proteins, 70 unique IntAct entities, 11 UniParc protein sequences, 6144 UniProt proteins. The network was not completely connected, but included 169 components where the largest connected network had 5724 nodes. The second largest network component had 43 nodes.

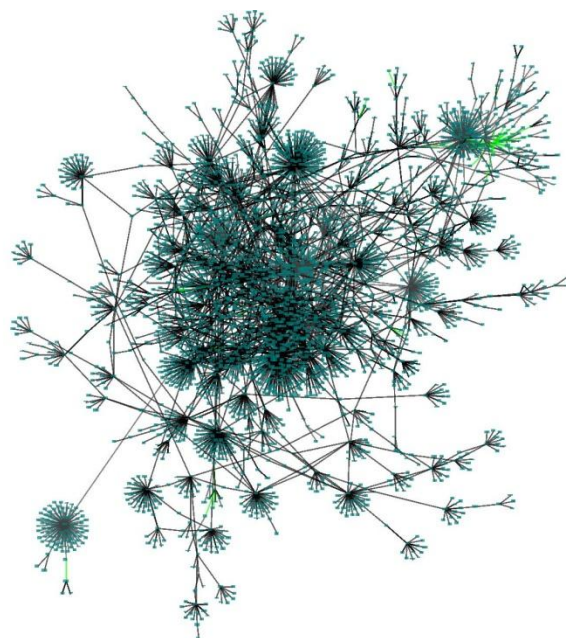


Figure 3 - Largest apoptosis subnetwork from PSICQUIC search of apoptosis associated biomolecules.

Search results from PSICQUIC connected databases (Figure 3) were studied for biological validity by analyzing the PSI-MI tags using Starlight Information Visualization System created by Pacific Northwest National Laboratory (<http://www.eurekalert.org/features/doe/2002-03/dnnl-ias061402.php>). PSI-MI tags describe the types of experiments that produced the interactions described in the above databases. Table 3 shows PSI-MI tags from the Apoptosis network. These terms are associated with different experimental protocols that share common methods. Starlight extracted the common methods from the PSI-MI tags and creates diagrams of overlap as seen in Figure 2. This analysis identified records that use unproven protocols which can be excluded, producing a higher confidence network.

Search results from Cytoscape were extracted in XML format (XGMML). This format included all data provided with the interaction record (source, experimental type, publication reference, etc. This XML was imported into Starlight's VIS system using the import wizard. Within the XML schema, edge attributes and their associated values were exported and visualized. Starlight identified 87,982 words, 219 of which were unique. Starlight was able to cluster

these words into 17 groups that were visually filtered into 10 groups (Table 2) that included words that described experimental techniques.

Table 2 – Occurrence of words describing the biological evidence of each interaction in the apoptosis network.

<b>Keyword/s</b>	<b>Occurrence in Edge Evidence</b>
<b>Physical, Association</b>	9011
<b>Bait, coimmunoprecipitation, anti</b>	3387
<b>Hybrid, approach, pooling</b>	3123
<b>Array, assay, microscopy, technology, peptide, chromatography, protein</b>	2123
<b>Tag, coimmunoprecipitation, anti</b>	1543
<b>Pull</b>	1534
<b>Purification, tandem, affinity</b>	1524
<b>Interaction, direct</b>	735
<b>Reaction, phosphorylation</b>	231
<b>Crystallography, x-ray, stelzl</b>	159

Starlight has the ability to visually depict the clusters in three dimensions (Figure 4). The three dimensions were: occurrence, number of shared XML records, and distance between records.

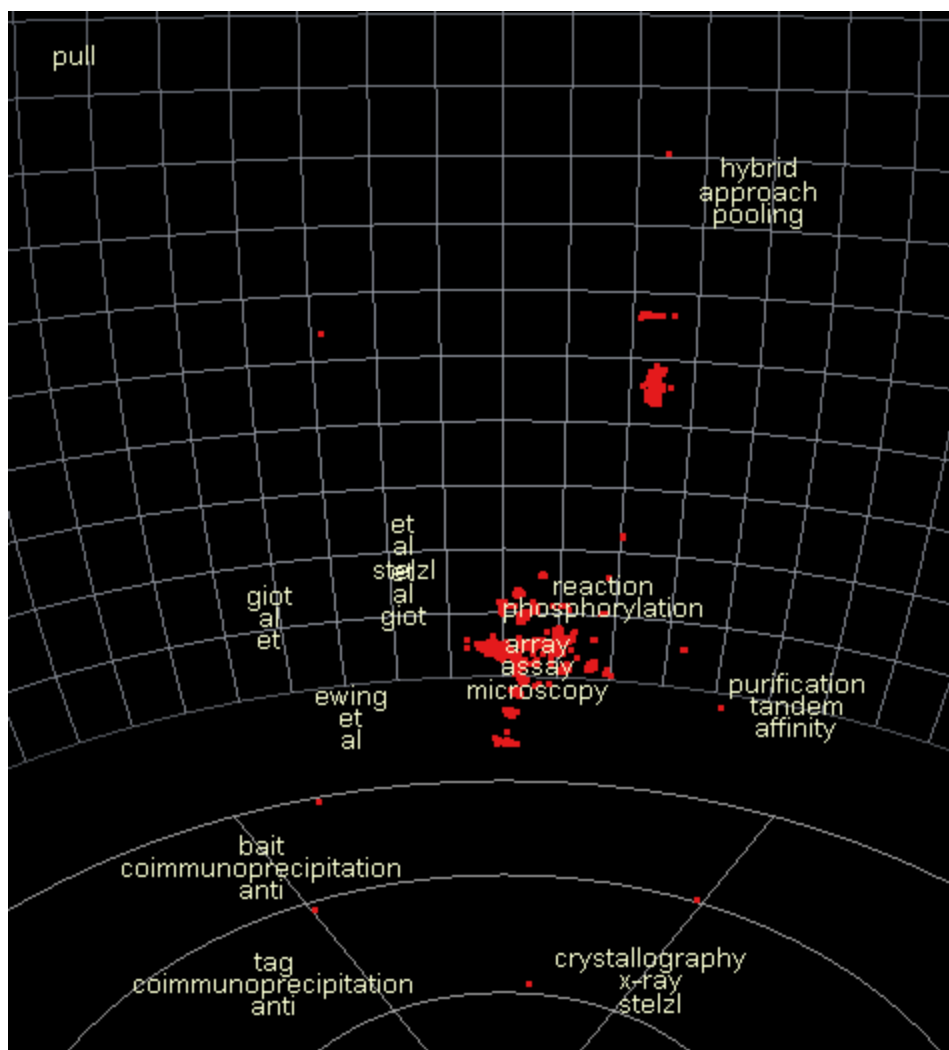


Figure 4 - 3D clustering of key terms from PSICQUIC search using starlight visualization tool.

Semantic modeling of data is becoming a more popular method of filtering search results from web search tools (Nelson, Avraham et al.). Semantic modeling is the bases for Starlight VIS, but represents only one type of analysis. To verify the results a word map was generated of the same terms used in Starlight.

In a word cloud the font size is determined by the occurrence of the word in the source data (Kaser 2007). The algorithm for determining font size is:

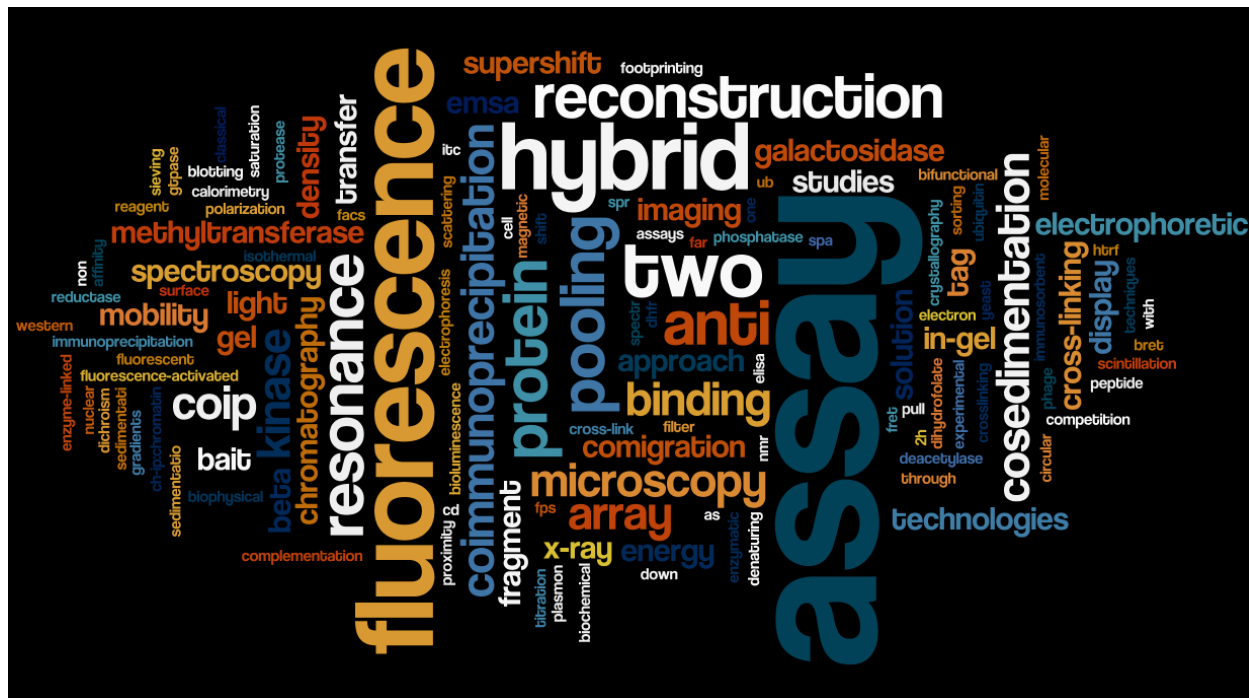


Figure 5 - Word cloud representation of interaction data from PSICQUIC search of apoptosis. This search was compared with the Starlight search to identify which experimental methods could be excluded to increase the quality of the apoptosis network.

Although the word cloud represents a different analysis of using semantic grouping, the outcome is similar. From the analysis done in Starlight, combined with the word cloud analysis, a more accurate network was prepared (Figure 6).

The number of protein coding genes is estimated to be around 23,000 (Stein 2004). This could yield 50,000 to 100,000 thousand proteins in human cells. 6400 proteins would represent between 6 and 13% of the entire proteome. It is possible that apoptosis includes a high percentage of the proteome, it is more likely this network includes inaccuracies. To remove some of these inaccuracies, the PSI-MI codes for the experiments for each interaction were verified above. 76 Uniquely coded experiments were used to create the

Apoptosis network of 5714 nodes. 12 of the 76 were not associated with a protocol or were associated with a poorly represented protocol. Removing the records associated with these 12 missing protocols produced a network of 3414 nodes with 5890 edges (Figure 6).

Table 3 – PSI-MI returned during the PSICQUIC search for interactions associated with apoptosis.

ID	Term ( <b>Type of experiment used to verify interaction</b> )
<b>MI:004</b>	chromatography: affinity chromatography technologies
<b>MI:006</b>	anti bait coip: anti bait coimmunoprecipitation
<b>MI:007</b>	anti tag coip: anti tag coimmunoprecipitation
<b>MI:0010</b>	beta galactosidase: beta galactosidase complementation
<b>MI:0012</b>	bret: bioluminescence resonance energy transfer
<b>MI:0013</b>	Biophysical
<b>MI:0016</b>	cd: circular dichroism
<b>MI:0017</b>	fluorescence spectr: classical fluorescence spectroscopy
<b>MI:0018</b>	two hybrid
<b>MI:0019</b>	coip: coimmunoprecipitation
<b>MI:0027</b>	Cosedimentation
<b>MI:0028</b>	solution sedimentati: cosedimentation in solution
<b>MI:0029</b>	density sedimentatio: cosedimentation through density gradients
<b>MI:0030</b>	cross-link: cross-linking studies
<b>MI:0031</b>	protein crosslinking: protein cross-linking with a bifunctional reagent
<b>MI:0040</b>	electron microscopy
<b>MI:0045</b>	Experimental
<b>MI:0047</b>	far western blotting
<b>MI:0049</b>	filter binding
<b>MI:0051</b>	fluorescence: fluorescence technologies
<b>MI:0053</b>	fps: fluorescence polarization spectroscopy
<b>MI:0054</b>	facs: fluorescence-activated cell sorting
<b>MI:0055</b>	fret: fluorescent resonance energy transfer



---

<b>MI:0065</b>	itc: isothermal titration calorimetry
<b>MI:0067</b>	light scattering
<b>MI:0071</b>	molecular sieving
<b>MI:0077</b>	nmr: nuclear magnetic resonance
<b>MI:0081</b>	peptide array
<b>MI:0084</b>	phage display
<b>MI:0089</b>	protein array
<b>MI:0091</b>	Not Found
<b>MI:0096</b>	pull down
<b>MI:0099</b>	spa: scintillation proximity assay
<b>MI:0107</b>	spr: surface plasmon resonance
<b>MI:0111</b>	dhfr reconstruction: dihydrofolate reductase reconstruction
<b>MI:0112</b>	ub reconstruction: ubiquitin reconstruction
<b>MI:0114</b>	x-ray: x-ray crystallography
<b>MI:0115</b>	yeast display
<b>MI:0231</b>	Not Found
<b>MI:0276</b>	Not Found
<b>MI:0363</b>	Not Found
<b>MI:0364</b>	Not Found
<b>MI:0397</b>	two hybrid array
<b>MI:0398</b>	two hybrid pooling: two hybrid pooling approach
<b>MI:0399</b>	2h fragment pooling: two hybrid fragment pooling approach
<b>MI:0401</b>	Biochemical
<b>MI:0402</b>	ch-ip:chromatin immunoprecipitation assays
<b>MI:0404</b>	comigration in gel: comigration in non denaturing gel electrophoresis
<b>MI:0405</b>	competition binding
<b>MI:0406</b>	deacetylase assay
<b>MI:0411</b>	elisa: enzyme-linked immunosorbent assay
<b>MI:0412</b>	emsa supershift: electrophoretic mobility supershift assay
<b>MI:0413</b>	emsa: electrophoretic mobility shift assay
<b>MI:0415</b>	enzymatic studies

---

<b>MI:0416</b>	fluorescence imaging: fluorescence microscopy
<b>MI:0417</b>	Footprinting
<b>MI:0419</b>	gtpase assay
<b>MI:0423</b>	in-gel kinase assay: in-gel kinase assay
<b>MI:0424</b>	protein kinase assay
<b>MI:0426</b>	light microscopy
<b>MI:0428</b>	imaging techniques
<b>MI:0423</b>	one hybrid
<b>MI:0434</b>	phosphatase assay
<b>MI:0435</b>	protease assay
<b>MI:0440</b>	saturation binding
<b>MI:0510</b>	Htrf
<b>MI:0515</b>	methyltransferase as: methyltransferase assay
<b>MI:0663</b>	Not Found
<b>MI:0676</b>	Not Found
<b>MI:0678</b>	Not Found
<b>MI:0728</b>	Not Found
<b>MI:0729</b>	Not Found
<b>MI:0826</b>	Not Found
<b>MI:0841</b>	Not Found
<b>MI:0921</b>	Not Found

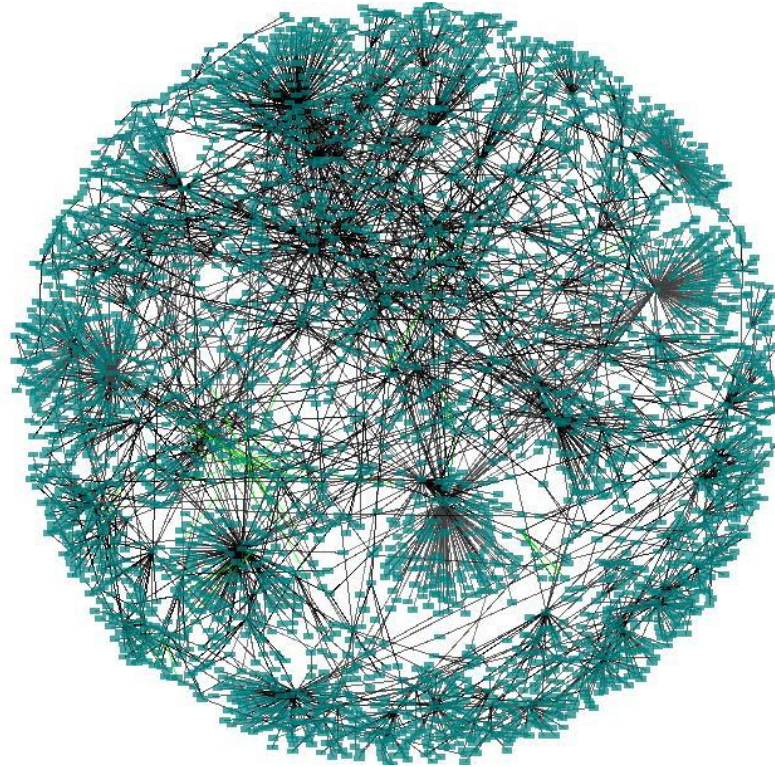


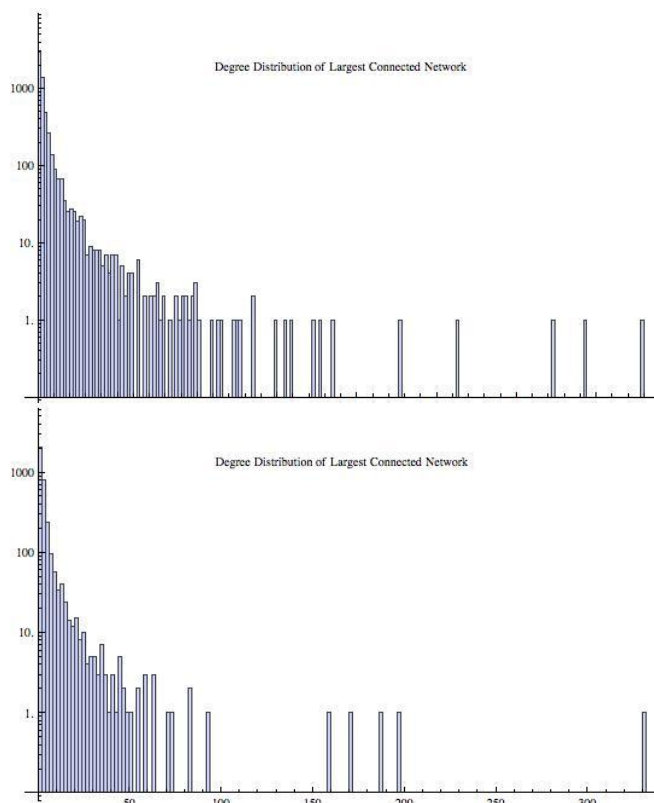
Figure 6 – Visual representation of the optimized apoptosis network produced using PSICQUIC enabled tools.

### **Analysis of Optimized Network**

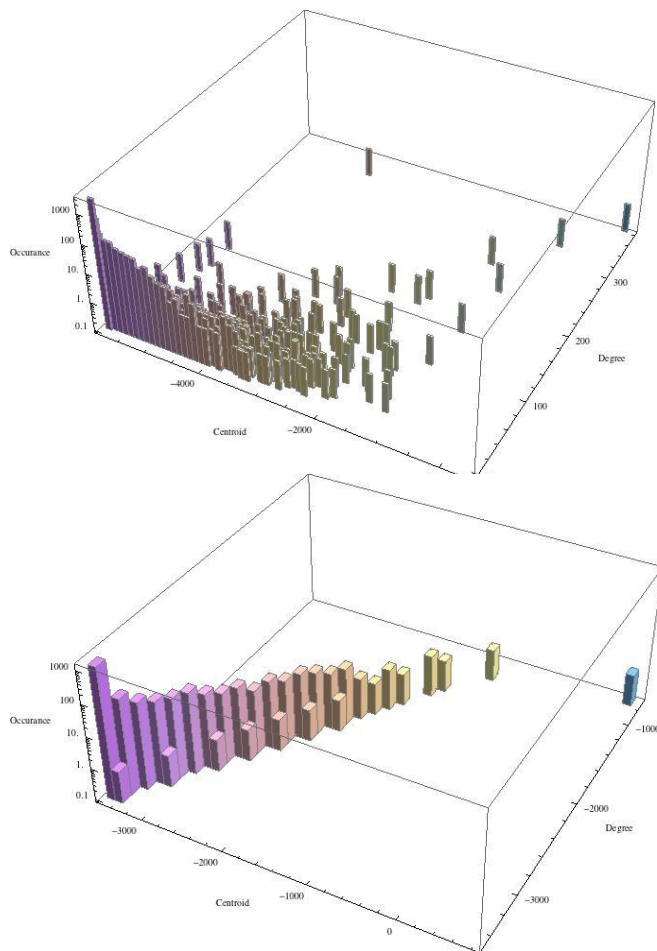
In order to predict the potential modules the largest network was analyzed for centroid and degree. Centroid is calculated by focusing on two nodes ( $w,v$ ) and calculating the number of nodes that are closer to each of the nodes ( $w$ ), compared to the other node in the pair ( $v$ ) (Scardoni, Petterlini et al. 2009). This evaluation is then repeated exhaustively for an entire network. This produces a centrality value where the higher the centroid index, the closer the node is to all other nodes in the network. Degree is the total number of interactions of each node.

To estimate the functional modules described in this document (Figure 1), I calculated the centroid score and node degree for the largest connected network of the Apoptosis network described above. Node degree provided a general understanding of the connectedness of the entire network of 5724 nodes.

Figure 7 shows the range of degrees or connections of the over 5724 nodes. The range of degrees spans 378 to one, with one representing the most common degree (A biomolecule with degree 378 is unrealistic and demonstrates the need to improve the “biological validity” of the data). The protein with the highest degree is Q9Y4K3 or E3 ubiquitin lygase, critical for Map Kinase activity. The majority are terminal nodes with one connection that may connect to other networks not included in the search, or final products. Only 20 of the 5724 nodes had a degree of over 100, and without the centroid score most could not be interpreted as central. In Figure 8 the centroid score is combined with node degree to provide a clearer picture of centrality and control of the network. Several of the proteins with high degrees have relatively low centroid values. P25694 had a relatively high degree of 126, but a very low centroid score of -5281. This protein is responsible for scaffold deconstruction, a function that is not involved in apoptosis activation. P35438 also had a high degree and a low centroid score. The protein is key to magnesium sensitivity and little to do with initiating apoptosis. Both of these proteins represent the limitation of using node degree to determine key proteins in functional specific networks. Although centroid provides a good measure of centrality, it cannot identify modules. The average centroid score is -5444, but the end of the second largest quartile is -5480. The heavy distribution of low scoring nodes demonstrated the low sensitivity of the centroid score for functional modules, even when the module is central to the network.



**Figure 7 - Log Distribution Histograms of Node Degrees from Apoptosis Network (Top is from the complete network resulting from the search, bottom is the optimized network with unknowns and low scoring biological evidence removed)**



**Figure 8 - 3D Log Plot of Centroid Value by Node Degree by Occurrence number in Apoptosis Network. (Top is from the complete network resulting from the search, bottom is the optimized network with unknowns and low scoring biological evidence removed)**

Functional modules of signaling networks with related outcomes have a common structure of a “master” component that is highly connected and activates a series of redundant effected components, which competitively inhibit each other but lead to similar outcomes. This property of the “master” component, or “Regulating Promoter” in Figure 1, will be centrally placed within the network. To determine whether existing centrality algorithms could predict these modules I searched the IntAct database for all records associated with Apoptosis and performed a centrality analysis using degree and centroid score.

## Chapter 3: Network Simulation of Bcl2 Family Proteins

### Summary:

**Bcl2 is a family of proteins responsible for apoptosis and cell cycle pathways related to apoptosis. Bcl2 represents an ideal model of network motifs that control large cellular functions. To identify motifs represented in the Bcl2 family pathways I simulated Bcl2 family protein networks to using cellular automata driven agent-based modeling. The outcome was list of potential motifs that fit the behavior described in my hypothesis and a deeper understanding of why this these motifs demonstrate very complex behavior.**

### Cellular Automata analysis

Cellular Automata was introduced by John von Neumann and Stanislaw Ulam in the 1960's (Kier, Seybold et al. 2010). Von Neumann proposed two-dimensional CA systems represented on a grid could change states following rules derived from the state of the neighboring cells (Neumann and Burks 1966). Von Neumann's early CA systems allowed for up-to 29 different states and provided limitless potential outcomes which made them become cumbersome to work with. In the early 1980's, Stephen Wolfram (Wolfram 1983) developed one-dimensional CA systems where the state of a cell was determined from the state of the cells during the previous iteration of the simulation. As iterations continued, one-dimensional CA's produced a finite number of outcomes. Rules based on these outcomes have been used to develop a new theory for computational analysis (Wolfram 2002).

Two-dimensional simulations have been adapted for simulations used in molecular interaction and biology (Bonchev, Thomas et al. ; Kier 2008) . In these simulations, the grid introduced by von Neumann represented interstitial space, and each square was a molecule. The rules von Neumann used to change the

state of his cells were modified to closely follow molecular mechanics. Simulations resulting from these modified von Neumann systems were shown to predict the boiling point and melting point of several compounds, and predict the diffusion rate of one mixture through another (Kier, Seybold et al. 2010).

Following completion of the interaction searches and network topology analysis, simulation using two dimensional CA systems was used to predict the behavior of the interaction network(Strogatz 2001; Kier, Bonchev et al. 2005). The rules were based on the maximal number of interactions for each molecule represented in the network coupled with structural analysis of bonding regions identified from 3D structures (from the protein databank). Structural analysis included evaluation of multiple isoforms, and mutated versions of proteins for lipophobicity and electrostatic potential using the Adaptive Poisson-Boltzmann Solver with Amber force-fields as described by Baker et al. (Baker, Sept et al. 2001). These properties were mapped, color-coded and visualized for evaluation using PyMOL. Areas with high lipophobicity and high electrostatic potential were considered possible binding sights for the simulation but provided no additional rule not already extracted from the literature searches.

Where graphical representations did not exist, generalizations were employed based on biochemical characteristics of the predicted solvent accessible surface combined with interaction reported in IntAct. Multiple generalizations were based on most probable conformation providing a range of bonding properties to be included in simulations were multiple confirmations were possible. Simulations were done on a 64 bit multiple node supercomputing cluster at the CSBC. CA models representing protein networks were comprised of four-sided shapes (cells). The cells were variegated allowing up to 4 unique protein-protein interactions per representative protein. Rules of motion were based on probability and govern movement, joining and breaking of cells. Movement, joining and breaking opportunities will be universal and have equal probability. The only variations in behavior resulted from rules of interaction defined above. Movement was applied randomly to each cell representing an asynchronous model (as seen in natural processes). Types of bonds and



complexes were tracked and represented the output of the simulation. Iterations begin with a random distribution of cells and executed until a steady state was identified (no change in binding for multiple iterations).

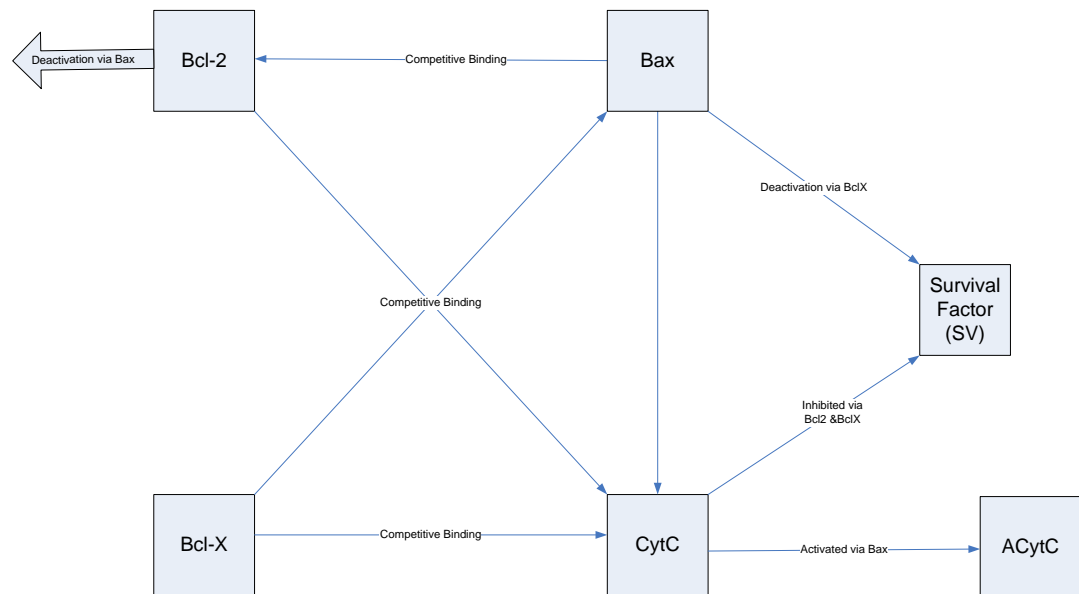
A new set of simulations was conducted where each rule will be exhaustively changed and removed and run to reach a new steady state. As described below, not all rules needed to be changed to obtain the complete behavior of the network. Many rules represented repetitive or redundant functions and only needed to be simulated by a single rule change. Changes or loss of steady state was tracked and evaluated for impact on the overall network. The results were molecular bonding probabilities (strengths) that separate equilibrium states or other simulation behaviors. Gene/protein representations that lead to large changes in the simulations were considered potential regulatory pathways with a “master” classification. Remaining genes/proteins were ranked by predicted impact on simulation behavior with a “slave” classification. Completion of these tasks resulted in network motifs that address the validity of the hypothesis.

#### Bcl2 Simulations:

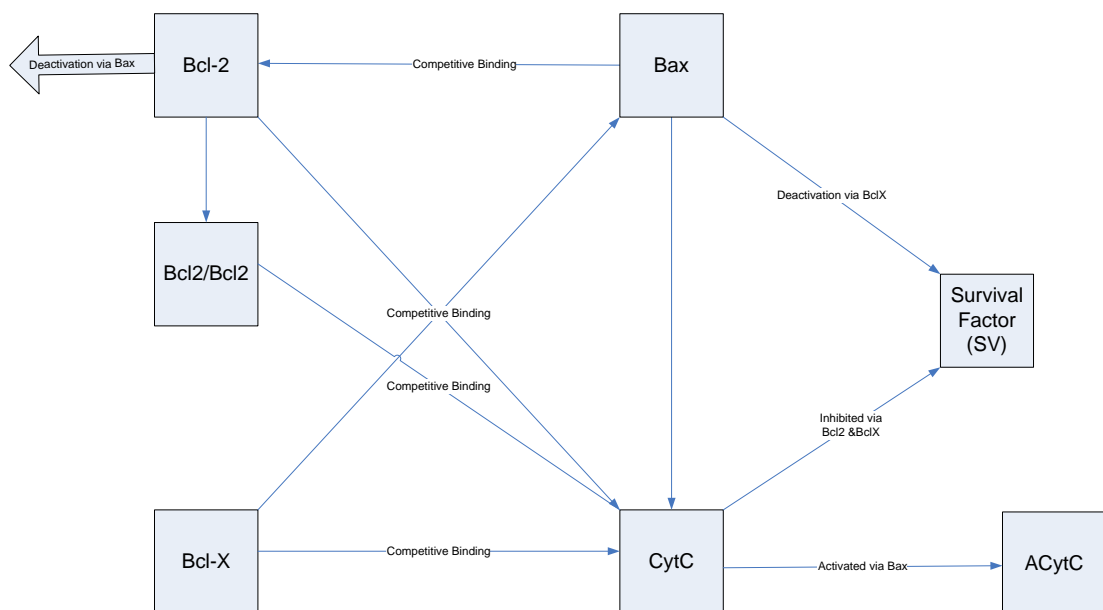
Two Bcl2 networks were simulated. The first includes four agents {Bcl2, Bcl2L10 (BclX), Bax, Cytochrome C (CytC)}. These four agents may represent multiple states (activation) and interaction. Four models were created that included each of these agents. The variations between the models were based on disagreements in the literature about the interaction of BclX with Bcl2. The use of diverse but similar models also provided a useful method to identify higher-level behaviors that were used to establish methods for analysis.

The first model in this network included no interaction between BclX and Bcl2 (Figure 9). The model included changes in the states of Bax, Bcl2 and CytC. Bax changed states from activated Bax to survival factor (SV), to represent inhibition of Bax’s apoptosis promotion properties by BclX (Hoffmann and Valencia 2004). Bcl2 changed states from activated Bcl2 to deactivated Bcl2. This represents the inhibition of Bcl2’s apoptosis inhibition properties by Bax

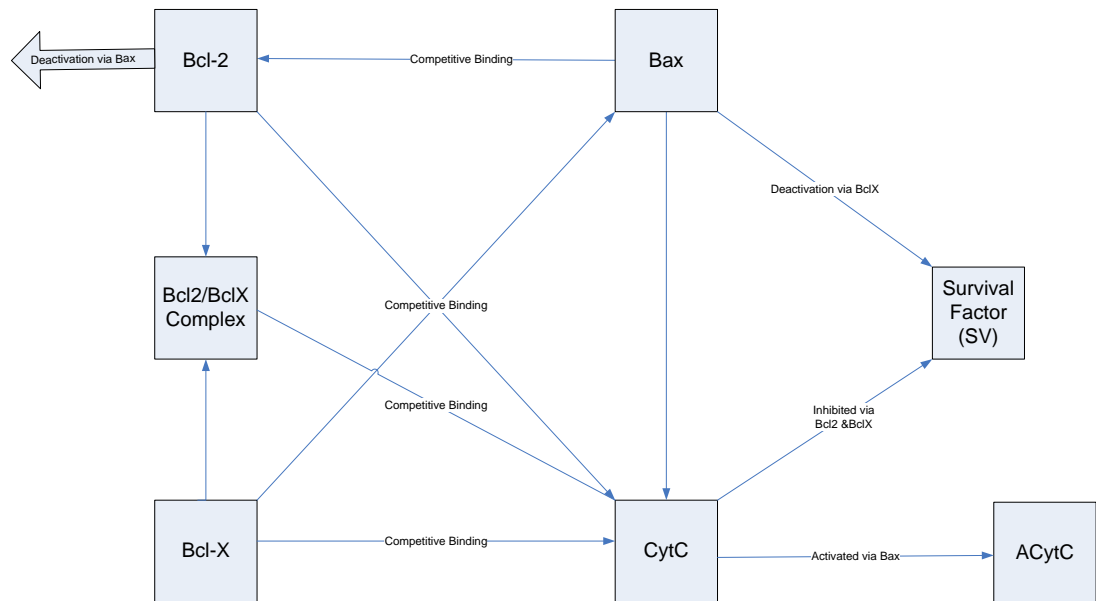
.CytC changed state from inactive to active CytC (AcytC), associated with the release of cytochrome C from pores in the mitochondria promoted by Bax. CytC also changed to SV to represent direct inhibition of the release of cytochrome C from the mitochondrial pores by BclX and Bcl2 . This network represents five separate pathways. The second model (Figure 10) includes each of the interactions represented in Figure 9 but adds one change state. Bcl2 changed to Bc2 homodimer (B2B2) to represent homodimerization reported in the literature (Figure 10)(Zhang, Szustakowski et al. 2009). The third model included each of the interactions represented in the first model (Figure 9), but added two changed states (Figure 11). Bcl2 changed to Bcl2 bonded to BclX (B2BX). BclX changed to B2BX. Each of these interactions had been reported in the literature (Cheng, Wei et al. 2001; Jourdan, Reme et al. 2009). The fourth model combined each of the previous three to create a more complete interaction map (Figure 12). This model includes eight individual pathways with seven changed states.



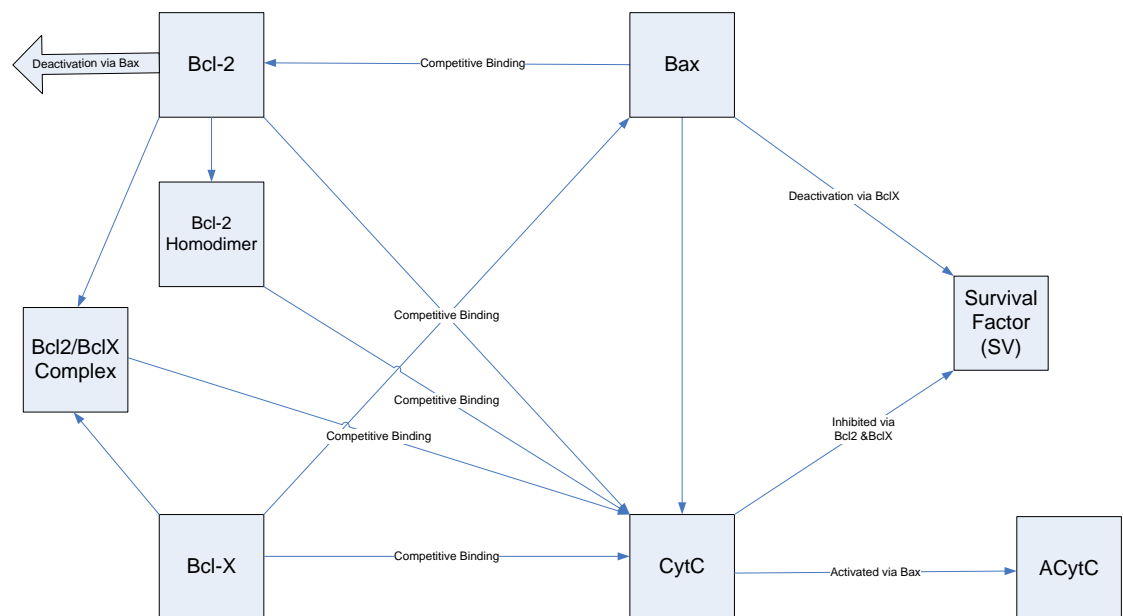
**Figure 9 – First model of the Bcl-2 network used in cellular automata simulation of proteins associated with apoptosis.**



**Figure 10 – Second model of the Bcl-2 network used in cellular automata simulation of proteins associated with apoptosis. This model expanded on the second model by including the Bcl-2 homodimer reported in co-immunoprecipitation studies.**



**Figure 11 – The third model: A different version of the second Bcl-2 model where the Bcl-2 homodimer was replaced with a Bcl-2, Bcl-X heterodimer.**



**Figure 12 – The third Bcl-2 model used in cellular automata simulation. This model combined the two versions of the second model where Bcl-2 can produce a homodimer, and a heterodimer with Bcl-X**

When simulated using CA the results of the first model showed convergence of the SV (representing survival) and AcyC (representing apoptosis) concentrations between the bonding probabilities of 0.5 to 0.6 of BclX interacting with Bax (inhibition of Bax). This result was used to identify other convergence points in the other three models. New convergence points only emerged in the third (Figure 11) and fourth (Figure 12) models. The new convergence points in model three occurred from the interaction between BclX and B2BX at probabilities 0.01 to 0.02 and 0.07 to 0.08. In the fourth model convergence points emerged at the interactions of BclX and CytC at probabilities of 0.72 to 0.7; the interaction between BclX and Bax at probabilities of 0.42 to 0.4; and the interaction between BclX and B2BX at probabilities of 0.015 to 0.01. The resulting narrow ranges of existence of the survival state demonstrate the significant challenges for drug discovery and clinical treatment of lung cancer.

The Bcl2 family-related apoptotic pathway is highly redundant of pro-survival and pro-apoptotic proteins, which provides a unique network for modeling and simulation. The second Bcl2 network analyzed here was designed to uncover the impact of this redundancy on the Bcl2 protein, which has limited spatial accessibility. The model included the pro-survival proteins Bcl2 and BclX (Bcl2L1) and the pro-apoptotic BID and BIM (Bcl2L11) (Xue, Chiu et al. 2003; Kerrien, Alam-Faruque et al. 2007; Jourdan, Reme et al. 2009; Renouf, Wood-Baker et al. 2009). We modeled the limited Bcl2 accessibility by altering the dissociation constant from 1 to 0. Also included were non-Bcl2 proteins that were reported to bind to Bcl2 family proteins in tumor formation. All interactions included in the model had been reported in the IntAct database (Kerrien, Alam-Faruque et al. 2007) to be related to cancer or apoptosis. The completed model included 11 proteins with a total of 14 interactions (Figure 13). Two of the 11 proteins were isoforms of included proteins that were reported in different pathways from the wild-type. To simulate the model each protein was represented in a CA simulation by a single variegated cell. A variegated cell has

more than one rule of behavior for each of its sides. The CA engine determines whether a variegated cell can bind by evaluating the rule of the specific side of the square cell that is neighboring another cell. In the case where a protein has one binding side the variegated cell would have two rules, one binding and three non-binding. The single binding side would be able to bind other proteins and the three non-binding sides would not be able to bind with any other protein. The ability to bind was determined by normalized probabilities (the sum of each cell's binding probabilities, describing interactions with other cells, cannot exceed one). Of 14 interactions, four had published dissociation constants that were used as breaking probabilities between two cells in the CA simulation. Where no disassociation constant was reported, a breaking probability of zero was entered. Each summation was run for approximately 10,000 iterations to reach a steady state, and repeated 50 times.



**Figure 13 – This expanded Bcl-2 network included interacting partners that were not known to be Bcl-2 family members, but were similar in structure and had been reported as interacting partners with known Bcl-2 family proteins.**

The CA input files included, among others, the list of all pairwise interactions and their probabilities (Table 4).

**Table 4 – List of interacting partners include in the expanded Bcl-2 network (Figure 13)**

<b>Bcl2</b>	<b>Bcl2</b>
<b>Bcl2</b>	<b>BclX:iso</b>
<b>Bcl2</b>	<b>NLRP1</b>
<b>Bcl2</b>	<b>BIM</b>
<b>Bcl2</b>	<b>PPP1CA</b>
<b>Bcl2</b>	<b>BIM:iso</b>
<b>Bcl2</b>	<b>BID</b>
<b>BclX</b>	<b>BID</b>
<b>BclX</b>	<b>NLRP1</b>
<b>BclX</b>	<b>PPP1CA</b>
<b>BclX</b>	<b>BIM</b>
<b>BclX</b>	<b>HRK</b>
<b>BclX:iso</b>	<b>SIVA1</b>
<b>PPP1CA</b>	<b>BAD</b>

The model probabilities were normalized to a unit (Table 5).

**Table 5 – List of probabilities created for each pathway found in the expanded Bcl-2 network (Figure 13)**

<b>BIM:iso-Bcl2 + Bcl2-Bcl2 + BID-Bcl2 + BclX:iso-Bcl2 + NLRP1-Bcl2 + BIM-Bcl2 + PPP1CA-Bcl2</b>	<b>1</b>
<b>Bcl2-BclX:iso + SIVA1-Bclx:iso</b>	<b>1</b>
<b>BAD-PPP1CA + BclX-PPP1CA + Bcl2-PPP1CA</b>	<b>1</b>
<b>HRK-BclX + BID-BclX + NLRP1-BclX + BIM-BclX + PPP1CA-BclX</b>	<b>1</b>
<b>BID-Bcl2 + BID-BclX</b>	<b>1</b>
<b>BIM-Bcl2 + BIM-BclX</b>	<b>1</b>
<b>NLRP1-Bcl2 + NLRP1-BclX</b>	<b>1:</b>

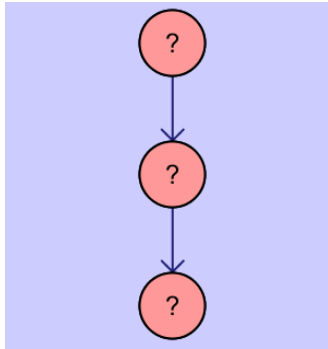
Bcl2 had a dissociation constant of one, which was included in the control simulation. Three other simulations were then run representing the absence of the redundant BIM path (see Figure 13), the loss of Bcl2's ability to dissociate and the combination of the two. The terminal path including the BIM isoform was monitored to evaluate the result of the changes. The loss of Bcl2's ability to dissociate from its binding partners and the loss of BIM resulted in a 20% reduction of the BIM isoform complex. This is believed to be the result of sequestered free Bcl2.

One may conclude that the highly redundant Bcl2 protein family network provides an ideal model for studying the effects of damaged redundant network paths in cancer, yielding alternative therapeutic approaches. The early results of this simulation provided a link to the effects of a lost redundant path to the sequestering of free Bcl2. We predict that the effects of a lost path, as a result of carcinogenesis, could be overcome by altering the dissociation of Bcl2.

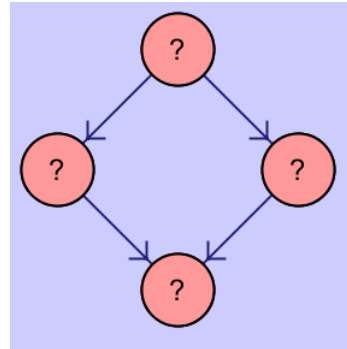
### **Resulting Motifs**

In addition to the analysis of the importance of concentration in the Bcl2 family network previously described, two motifs were identified in the second Bcl2 network (Figure 14). More network motifs would exist if the Bcl2 network was analyzed in context of the larger apoptosis network, discussed in the next chapter. The first identified network motif was a three-node two-edge directed one (Figure 15), representing a basic signaling cascade. This motif occurred 40 times with a ratio of 1.05 (ratio = (observed in network)/(mean occurrence in random networks)) with a p-value < 0.01.





**Figure 14 – 3-Node, 2 Edge Linear Motif**



**Figure 15 – 4-Node 4-Edge Biparallel Motif**

The second motif of four nodes and four edges occurred twice with a ratio of 2 and p-value <0.01. Further analysis of these motifs and the three additional networks that were dependent on the rest of the apoptosis network are discussed in the next chapter.

## Chapter 4: Differential Analysis

### Summary:

Recently, new methods to develop de novo networks from patient data have been developed. To identify a differential network of cellular gene/protein agents and potential gene regulatory pathways specific for functional changes associated with apoptosis I compared these methods to earlier described database searching methods. Although these methods are useful in evaluating networks associated with well understood functions in small populations, they can produce overly specific networks. For this reason the motif analysis only included networks from the databases described earlier.

### **Analysis of Microarray Data from Lung Cancer Patients**

mRNA microarray expression data from individuals diagnosed with Adenocarcinoma of the lung were obtained from the Oncomine database, the NCBI GEO database and EBI's ArrayExpress database. Experiments that share platforms, methods and patient characteristics including mRNA from adjacent tissue were merged to provide a dataset of interaction data from at least 200 patients with adenocarcinoma of the lung.

The first steps of the Network Builder algorithm involves extracting random directional vectors for each gene in Gene Ontology (GO) lists (Ashburner, Ball et al. 2000; Bard 2003). As stated in the Introduction, correlation-based methods show the most promise while maintaining the complex systems properties of biological networks. Current tools that offer "de-novo" network creation only calculate a pair-wise correlation. The tool used in this analysis correlated groups of probes represented by vectors. Network Builder, the tool used in this study, integrates a unique 2-D clustering with maximization of non-linear correlations that maintains dynamic behavior when creating a de-novo network from

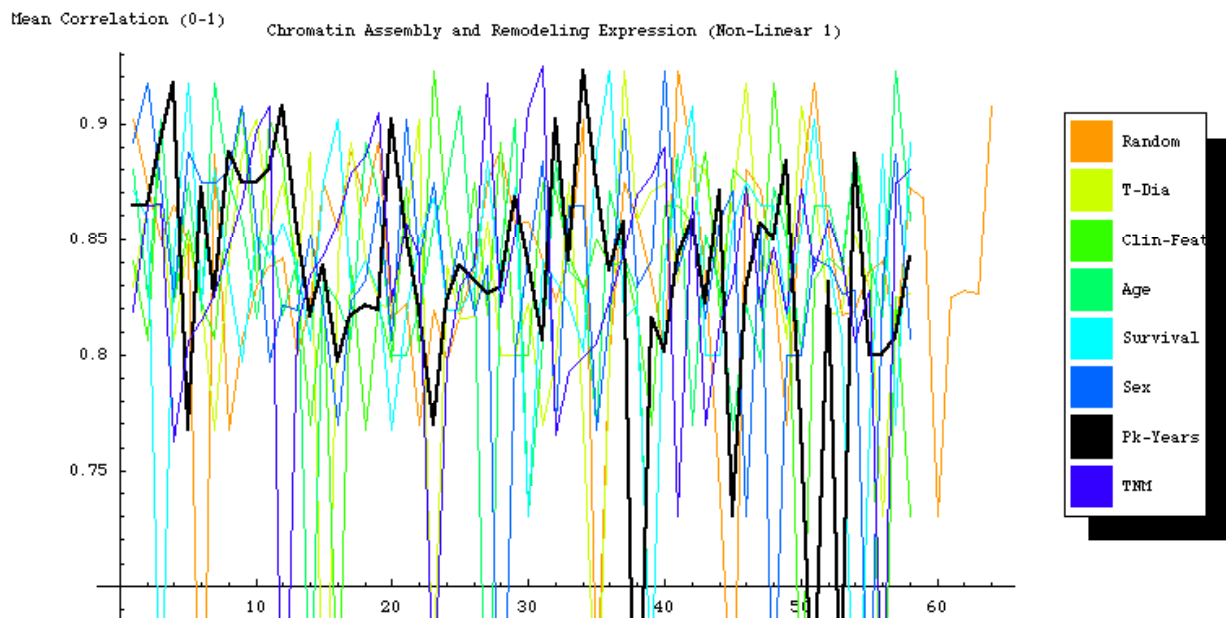
experimental data. Recent studies by Zhang et al. (Zhang, Ji et al. 2007) have described the strength of multi-gene correlation analysis compared to pair-wise. This tool expanded on Zhang's work by moving beyond three gene interactions. This study utilized expression vectors representing 3 to 50 interactions, which are evaluated using a combined clustering and correlation algorithm. The algorithm aligned each vector individually on a two-dimensional grid. Clusters were identified using GAP statistic. Each pair of two genes that presents more than one significant cluster was recorded as having a potential interaction. The algorithm then organized networks for the entire pool of 1000 genes. Sub-networks were created based on first neighbors for each gene starting with the gene represented with the highest node degree (potential interactions). Random vectors were extracted from each partition and non-linear correlation coefficients (Rho) was calculated between groups of vectors from genes in a sub-network between lung tumor samples and adjacent normal tissue. This value represented a random base correlation. In contrast, partial regression uses a linear-based reduction method that cannot account for compounding dynamic behaviors commonly seen in networks.

New vectors were extracted from each partition after the data was redistributed based on clinical features (i.e., stage of tumor, age, etc) that are believed to vary in the patient pool. Spearman Rank Correlation Coefficient (Rho) was derived for each sub-network with each directional clinical-based distribution. Spearman correlation produces a coefficient between -1 and 1, where as the distance from 0 (on a number line) increases, the statistical dependence between two vectors increases. Clinical distributions with the largest rho (exceeding random base correlation) derived from these groups of vectors was selected and used in the following analysis representing directional base correlation. Sub-networks were reduced gene by gene with new values for rho calculated and comparisons done between patients with varying clinical characteristics.

When rho for each group of directional vectors reached statistical significance (using distance from 0 and within a small range), subnetworks were then

combined to produce an accurate interaction network representing chromatin associated pathways. Ontology terms were used to partition the merged network into a subnetwork of apoptosis pathways associated with adenocarcinoma of the lung and chromatin. This network will then be tested against networks described earlier. Each network will be studied to identify topological limitations, i.e., the network diameter, node and edge degree as reported (Milo, Shen-Orr et al. 2002; Dorogovtsev, Goltsev et al. 2003; Rives and Galitski 2003). These values described the maximum number of potential interactions.

Subnetworks were then searched in the resulting network. It was expected that subnetworks will match and rho values will be evaluated to see whether they indicate a new unpublished interaction. Whenever the rho value was too small, the interaction was labeled as a false positive and removed from the sub network and a comparison was done. After this optimization step was complete, ontology terms were added to each sub network to see if the sub network represents a functional group. The chromatin example below provides an example of this de-novo network building approach.

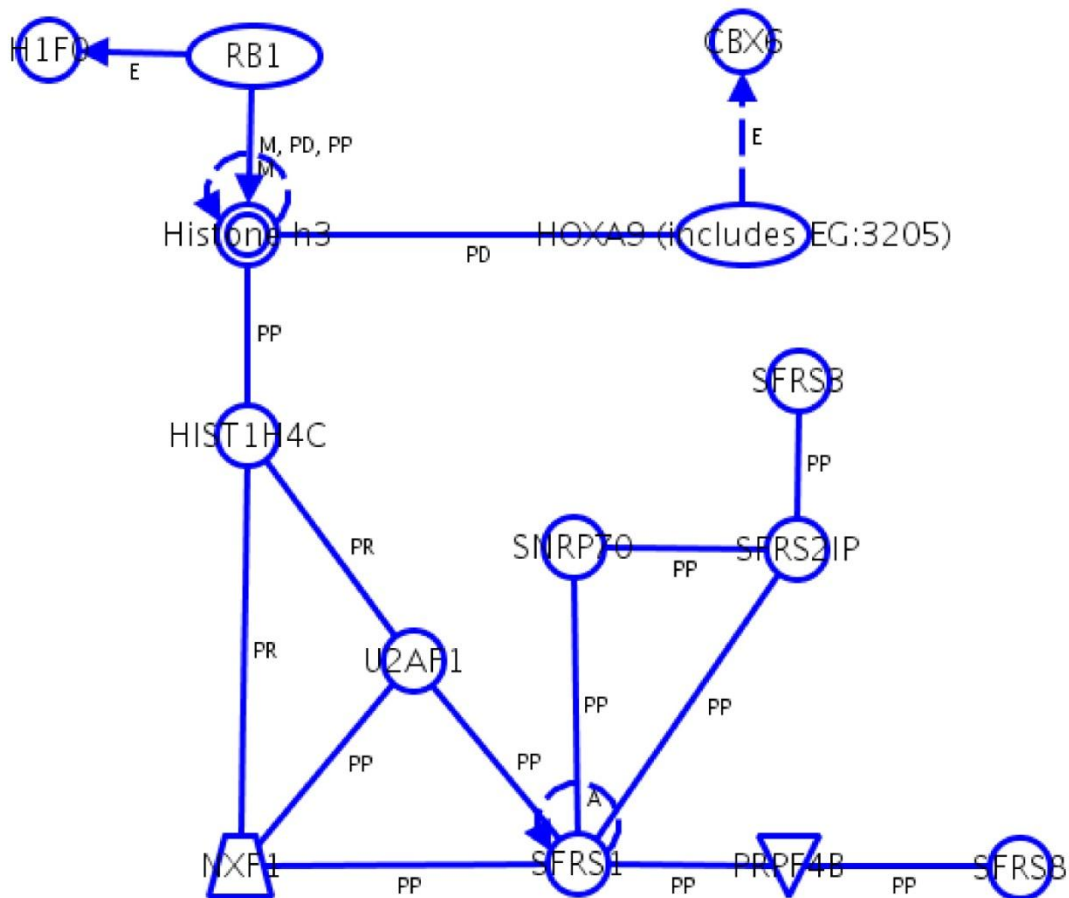


**Figure 16 – Normalized expression values were plotted after being sorted (low to high) by different clinical features. Each color represents a different clinical feature used for sorting the expression values.**

**Chromatin (This work was done in collaboration with Dr. Vladimir Kunetsov from the Bioinformatics Institute of Singapore):**

An initial search of the Gene Ontology DB produced a list of 134 genes associated with chromatin (Ashburner, Ball et al. 2000). This list included genes that are associated with chromatin accumulation, assembly and remodeling. Modeling and simulating the entire group of 134 genes, as described earlier, was difficult. An alternative analysis of expression profiles using non-linear correlation coefficients and a proprietary clustering algorithm was employed. Expression data of 71 patients with adenocarcinoma of the lung compared were to 16 healthy samples (Rhodes, Kalyana-Sundaram et al. 2007). Expression data was captured using an Affymetrix u95a platform. Results of the analysis identified pack-years followed by TNM staging as providing the most predictable coefficients (Figure 16). This analysis also identified 14 genes as potential core genes (Figure 17 – Created by Dr. Danail Bonchev).

## Core



© 2000–2007 Ingenuity Systems, Inc. All rights reserved.

**Figure 17 – This network was produced using Ingenuity Systems IPA network analysis tool. A list of proteins/genes was entered into the search tool and this network was produced. The list of proteins/genes was produced using the correlation analysis described in the text.**

**Table 6 - Ontology of chromatin genes identified as potential biological markers using the analysis described in the text. These markers could be used to discriminate between diseased and healthy tissue.**

H2AFV	Histone Construct
SFRS3	Splicing Factor/RNA Process
RB1	Transcription Regulator
HistH3	Histone Construct
CBX6	Transcription Regulator
HOXA9	Transcription Regulator

Hist1H4C	Histone Construct
NXF1	Transporter
U2AF1	Unknown
SFRS1	Splicing Factor/RNA Process
SNRP70	Splicing Regulator/RNA
CASP11	Splicing Regulator/RNA
PRPF4B	Kinase
SFRS8	Splicing Factor/RNA Process

We then used additional mRNA expression data for individuals diagnosed with adenocarcinoma of the lung from Oncomine (Rhodes, Kalyana-Sundaram et al. 2007) and from Genome Expression Omnibus (GEO, 2007) (Barrett, Troup et al. 2007). These were two very diverse datasets based on different platforms and even patients of different ethnicity, selected as a test for the classification reproducibility. The Oncomine U95A platform of expression data (12,600 probes) involved a classification study of patients with lung malignancies (Bhattacharjee, Richards et al. 2001) including 62 adenocarcinomas and 17 normal samples. The 62 adenocarcinomas were selected based on agreement between assessments of two independent pathologists. Samples where one report did not indicate pure adenocarcinoma were excluded, and the same was done with the data for patients with secondary metastasis of a different morphology. This produced a dataset of pure adenocarcinomas with no metastasis, tumor sizes of 1 to 8 cm, and all stages. The Bhattacharjee et al. expression data (Bhattacharjee, Richards et al. 2001) were thus partitioned into five subsets, four of which containing 15 or 16 tumor samples each randomly selected from the pool of 62 samples, and one set of expression data with 17 normal samples. The GEO U133A platform (22,284 probes) contained expression data from 27 adenocarcinoma patients with accompanying data from adjacent normal tissue (Su, Chang et al. 2007). Expression data was thus presented in two sets of 27 samples each (normal and diseased ones).

Our novel strategy limited strongly the starting gene list by focusing on certain hypothesis as to which biological processes and molecular functions are of importance in carcinogenesis. One can thus identify proteins that are strongly affected in the different stages of lung adenocarcinoma. Recent studies on the mechanisms of lung cancer pay a considerable attention on the chromatin structure changes, and histone as the chief protein component of chromatin (Cameron, Bachman et al. 1999; Reisman, Sciarrotta et al. 2003; Sasaki, Moriyama et al. 2004; Gibbons 2005; Medina, Carretero et al. 2005). Proteins related to DNA modification, MAP Kinases activity, and transcription were also considered of importance. This choice enabled the compiling of two lists of proteins. The first one with 98 genes was based on Gene Ontology Database [GO] (Ashburner, Ball et al. 2000; Bard 2003) classifications provided by Affymetrix Inc. The second list with 43 genes was a merge of Panther (Thomas, Campbell et al. 2003) and David (Sherman, Huang da et al. 2007) ontology and GO terms, of the four categories described above. Thus, our search started with a combined list of 141 genes.

For each of the two gene expression datasets (Su et al. data (U133A)(Su, Chang et al. 2007) GEO ID GSE7670; Bhattacharjee data (U95A)(Bhattacharjee, Richards et al. 2001), provided by Oncomine (add reference) we first calculated the Spearman non-parametric correlation coefficients  $\rho$  between pairs of hybridization signals of all probesets of expressed genes presented on a microarray for all patients included (complete cohort, normal and diseased). Gene expression was defined as the normalized signal for each probeset ID associated with a gene, as provided by Affymetrix Corporation. The number of probeset IDs per gene ranged from one to eight. The results of the pair-wise analyses did not provide clear results, due to large groups of predicted interactions with the same correlation coefficients. In addition, the correlation values were significant but not highly correlated. This result agrees with the recent conclusions of Zhang et al. (Zhang, Ji et al. 2007) that pair-wise correlation could be “a poor predictor of any molecular interaction associated with signaling and control of cellular function”. Instead, Zhang used correlation



between triples of genes. In this study, we applied a more sophisticated correlation analysis which extends the correlation to larger groups of 3 to 100 genes and includes a clustering step (Thomas 2008). We employed two-dimensional clustering using the so-called gap statistic to define a cluster of probesets. No pre-filtering procedure was used for removing noise signals and outlier patients. A cluster is defined as a group of expression signals where the number of members is greater than one, and the expression values used for input are from two different probesets representing two different genes (for example, a cluster including expression signals for probeset A and probeset B with a cohort of 20 samples would consist of 40 unpaired expression signals). The Gap statistics uses the output of traditional clustering algorithms but determines the difference between an internal dispersion of the variables within the predicted cluster and a null distribution. The distance between two nonnormalized expression signals was determined as Euclidean distance. The two dimensional aspect of our clustering approach refers to the bivariate plot graph in which the clustering is determined. The graph x coordinate stands for the values for all expression signals for probeset of gene A across all samples, while y is the same for a probeset of gene B across the same samples. So, if an individual sample is denoted by k, the point on the graph representing that sample is denoted as  $k_{xy}$  and the number of such points equals the total number of samples. Distance is then determined between each  $k_{xy}$  and  $k_{x'y'}$  and compared to the null distribution created by the Gap Statistics. If the sum of distances within the entire subset is less than that in the null distribution, the subset is reported as a cluster, as defined above. If the clustering of two expression signals (probes A and B) produced two or more clusters, then an interaction between the genes presented by the two expression signals from the probes was recorded as a predicted interaction. Since this clustering step was done exhaustively, each gene in the gene lists had several other genes identified as potential interactions. The Spearman correlation coefficient  $\rho$  was then determined for each list of potential interactions. All lists of each probeset represented the hybridization signal of transcripts on microarray were then

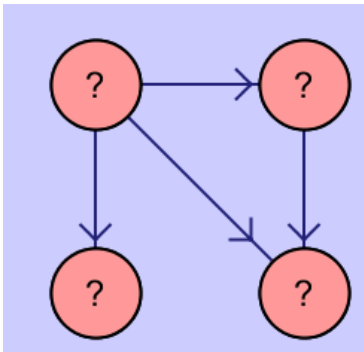
combined and the number of predicted interactions for each probesets was counted. The count was used as a predicted node degree in a network representation of highly correlating genes, the nodes and edges in which represent the genes and their interactions, only correlation coefficients identified as significant ( $p < 0.05$ ) were used.

Supervised optimization was performed using Spearman's  $\rho$  to compare potential interaction lists resulting from clustering. Optimization was considered complete when  $\rho$  reaches statistical significance ( $p < 0.001$ ). The optimization process included removing the probe with the lowest cluster number, and reevaluating correlation. This was repeated for each probeset represented in the potential interaction lists. When the correlation coefficient increased, the probeset was permanently removed, and vice versa, when the correlation coefficient decreased the probeset was returned to the potential interaction list. After completing the optimization, network maps were built. Each probeset ID was replaced by a gene symbol and when multiple probe IDs existed for a single gene correlation coefficients were compared. If correlation coefficients were the same, the duplicated probe ID was dropped. If the coefficients were not the same, a study of the hybridization was done in the form of a literature search and the less reliable probe ID was dropped. This last step was not necessary, due to the normalization using Bioconductor ([www.bioconductor.org](http://www.bioconductor.org)) and similar tools done prior to analysis (Zhang, Szustakowski et al. 2009). The count of genes in the optimized list represents the final node degree in the network. The analysis of the close neighborhood of this network makes possible the prediction of other potential markers among the genes that are closely connected to the preliminary selected candidate genes. For all predicted interactions determined by the correlation analysis we performed a literature search and Pathways Studio® (Nikitin, Egorov et al. 2003) search for information on existing regulatory, binding and other interactions or other relationships. This final step was used to verify potential biological implications of the constructed networks. Each network was also analyzed to identify topological limitations, such as network diameter, node

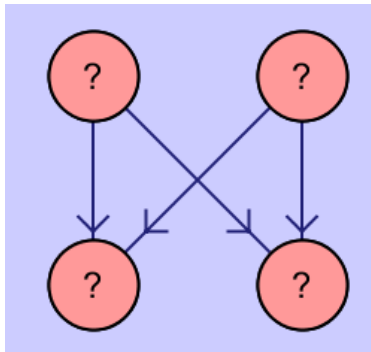
and edge degree, centrality, etc. (Milo, Shen-Orr et al. 2002; Dorogovtsev, Goltsev et al. 2003; Rives and Galitski 2003).

### **Resulting Motifs**

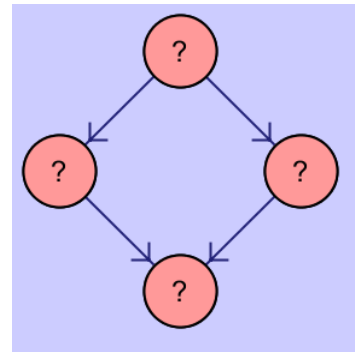
As a result of the high level specificity of the chromatin analysis I determined that the network resulting from patient data was likely to not have enough information to identify important controlling motifs. The network used during the analysis was from the reduced, high confidence, PSICQUIC network shown in Figure 6. To identify motifs I used the NetMatch Cytoscape Plugin produced by the Bader lab at the University of Toronto (Ferro, Giugno et al. 2007). Five motifs were identified as possible candidates to address the hypothesis.



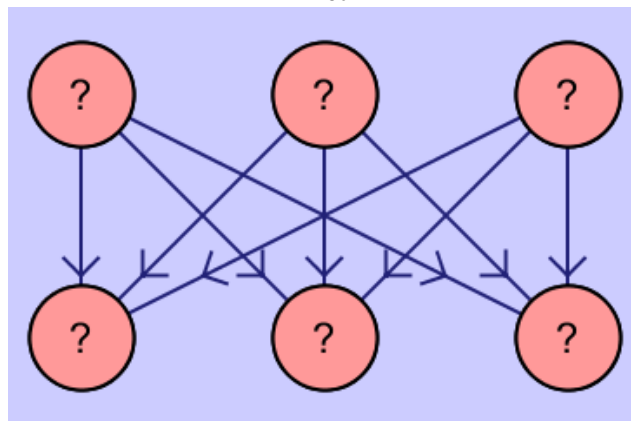
**Figure 18 – 4-Node, 4-Edge  
Non-Symmetrical Non-Linear  
Motif**



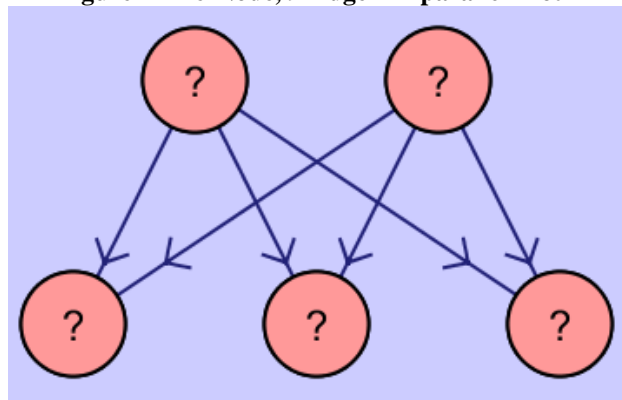
**Figure 19 – 4-Node, 4-Edge  
Symmetrical Non-Linear Bi-Fan  
Motif**



**Figure 20 – 4-Node, 4-Edge Bi-  
parallel Motif**



**Figure 21 – 6-Node, 9-Edge Tri-parallel Motif**



**Figure 22 – 5-Node, 6-Edge Tri-parallel Motif**

A selected subgraph is only important and is called “motif” if it occurs at a rate different from that in random networks. To test this I generated 1000 random networks with the same average node degree, diameter, number of nodes and number of edges to compare the frequency of occurrence. Table 8 contains the results of this analysis.

**Table 7 – Results of the motif search using NetMatch plugin with Cytoscape against the apoptosis network. All the ratios are low when compared to the mean network counts from all 1000 random networks.**

Figure	18	19	20	21	22
Standard Deviation	4004.16	990.74	285.68	1369.7	33,890
Mean from Random Net.	38,217	8561	19,643	4689	36,935
Count in Network	3954	5732	706	2628	39084
Standard Error	400	99	9.03	144.37	3572.4
z-stat	-85.57	-28.55	-2096	-14.27	0.587
p-value	<0.01%	<0.01%	<0.01%	<0.01%	27.8%
Ratio	0.103	0.669	0.036	0.560	N/A

FANMOD (Wernicke and Rasche 2006) is another tool commonly used for the identification and analysis of motifs in networks. FANMOD has strict formatting requirements that limit ability of this tool to read biomolecular networks generated from database searches (described earlier). The FANMOD input file must include only integers, representing vertices, and have one edge per row. Additionally the software can only read windows encoded Unicode 8-bit text files, even when the program is installed on a Linux or Unix machine. Due to these limitations, I was only able to analyze the autism network described in Chapter 5. The methods used to convert the network to be analyzed in FANMOD and the results are described in that chapter.

## Chapter 5: Autism Networks

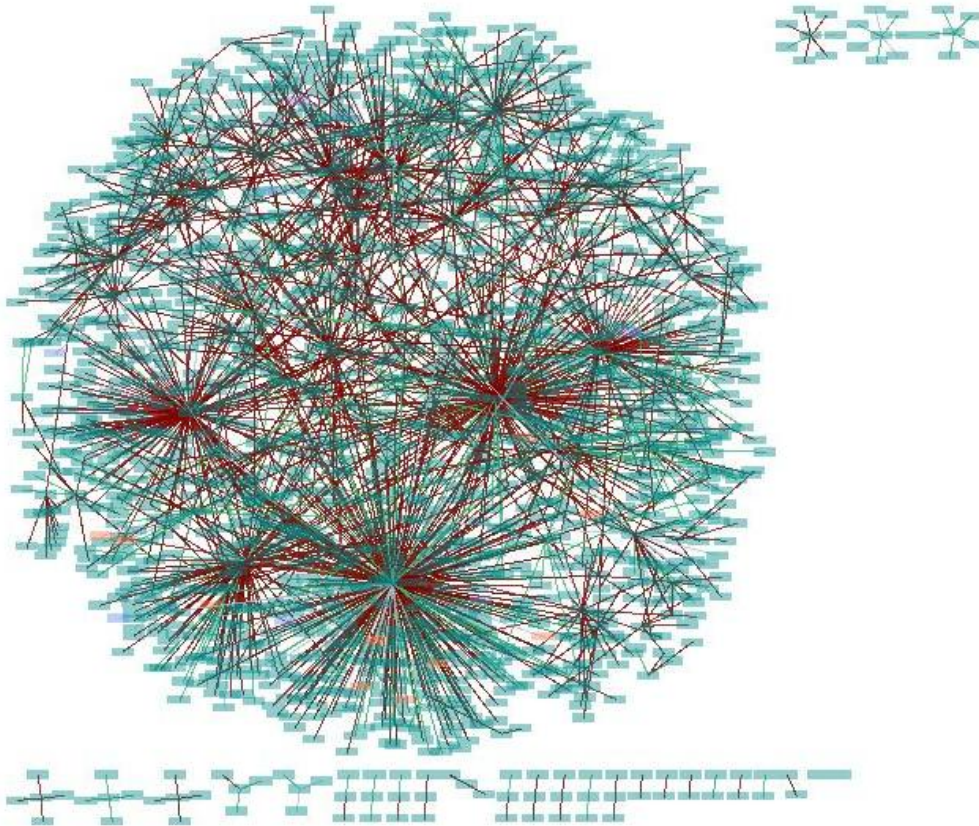
### **Summary:**

**To identify whether the motifs found in apoptosis were limited to only networks that are well studied I created a network of interactions reported in Autism Spectrum Disorder (ASD). From this network we identified the same motifs as those found in apoptosis, including increased significance in the larger motif. Interestingly, in the apoptosis network we found motif counts significantly lower than those in random networks, while in Autism we found motifs in consistently higher abundance than found in random networks.**

### **Analysis of Protein Interaction Networks**

To generate a protein interaction network associated with apoptosis, I used the PSICQUIC search tools built into Cytoscape. This tool was unable to identify any interactions associated with Autism. The PSICQUIC tool is designed to use an XML based standard that has a limited ontology. As a result the user can get a network from a meta-database search with better specificity than searching each database individually. This removes the redundancy and reduces the false hits. Unfortunately, this also removes the diversity of terms available. Since most interactions are associated with cancer, without a cancer related term, it is difficult to get a larger representative network from PSICQUIC.

NCBI provides an additional search tool built into Cytoscape (Sayers, Barrett et al.). This tool allows Cytoscape users to search with the Entrez utility across NCBI's databases. By using this tool, I was able to generate a network of biomolecular interactions in Autism with 1621 nodes and 3997 edges (Figure 23).



**Figure 23 - Autism network produced using the NCBI search tool. This was the only search target and tool that was able to produce a network using search terms associated with autism.**

In addition to NetMatch, FANMOD was used to search for and analyze motifs in this Autism network. NCBI uses integers as identifiers in networks generated by the Entrez tool. FANMOD is sensitive to gaps (missing integers) in the set of integers representing nodes, and cannot accept integers larger than 65,500. The Autism network has 1621 nodes with, integers ranging from 100 to 37,547,124. To compress the range of integers, I factored the numbers on a log scale ranging from 4 to 4000. The result was a set of integers that fit into the range of FANMOD. The factoring included a rounding step where factoring did not produce an integer. This step did result in the loss of several nodes, which may have significant impact on the results. This was made clear by the statistics of the Apoptosis network reported by FANMOD (number of nodes = 65107, number of edges = 1888). The error in node count is a programming bug where FANMOD reports the highest integer as the node count for the network. The loss

of half the edges could be a combination of multiple records (edges) for the same interaction, and loss resulting from the factoring and rounding of the integers.

### Resulting Motifs

As described in the previous chapter, 1000 randomized autism networks were generated to determine how far the occurrence count of the motifs deviated from that of random networks using the NetMatch Cytoscape plugin. The autism network included all five motifs (figures 18 - 22) which deviated significantly from the average 1000 random networks (Table 9).

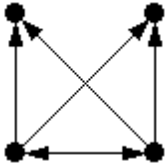
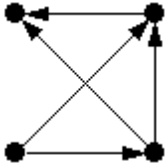
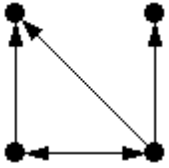
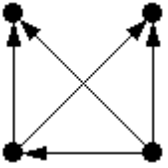
**Table 8 - Results of the motif search using NetMatch plugin with Cytoscape against the autism network. All the ratios are high when compared to the mean network counts from all 1000 random networks.**

Figure	18	19	20	21	22
Standard Deviation	10,402	696.56	31.87	465.5	399.04
Mean from Random Net.	15,510	4760	50.24	561	482.5
Count in Network	25158	9836	272	2412	2412
Standard Error	1040	73.42	1.01	46.6	39.9
z-stat	9.27	69.12	220	39.75	48.35
p-value	<0.01%	<0.01%	<0.01	<0.01	<0.01
Ratio	1.62	2.07	5.41	4.30	5.00

The results obtained from FANMOD (Table 9) did not include the same motifs found using the NetMatch plug-in (Table 8), although, the motifs that were found existed in the same high ratios as discovered using NetMatch.



**Table 9 - Results of the motif search using FANMOD against the autism network. All the ratios are high when compared to the mean network counts from all 1000 random networks.**

Motif				
Standard Deviation	6.38 e-6	2.67 e-6	2.36 e-4	4.43 e-5
<b>z-score</b>	5.39	4.4594	4.06	3.54
<b>p-value</b>	0.005	0.004	0.005	> 0.01
<b>Ratio</b>	9.43	4.67	4.61	2.4

Of the two tools used to study motifs, NetMatch is likely more reliable because it uses the networks in the original formatting without any modification. FANMOD is a more robust tool, because it has the ability to exhaustively search for motifs of a defined size. Unfortunately, the limited input formats make it less reliable for biomolecular networks created using interaction databases, even when the resulting network uses integers as identifiers for nodes. Accounting for the qualities and limitations of each program, high ratios were exhibited in all motifs found in the Autism network. Since it was found using both programs, it is difficult to define it as an artifact, but it is more likely a descriptive characteristic of the known Autism network, and could be useful in understanding the depth of existing knowledge about signaling and regulation associated with Autism. Due to the uncertainty associated with the FANMOD results, the following chapter will only discuss the NetMatch results. Further analysis with FANMOD should be performed for additional validation, but only in combination with a more robust input file conversion tool.

## **Chapter 6: Discussion and Future Work**

### **Summary:**

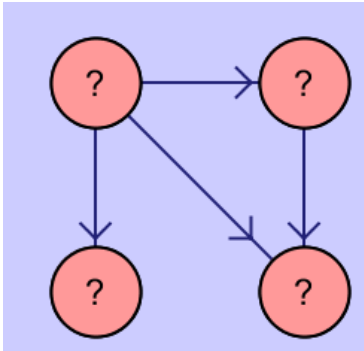
**The data resulting from this study has provided some support for the hypothesis that a master slave arrangement shown in a motif represents a key functional group of the network. The results between apoptosis and autism were inconsistent, but within each disease type they were very consistent. This suggests the ratio of occurrence is sensitive to the network completeness and could be exploited as a measure for this completeness, in addition to providing an indication of the presence of functional groups, affected by disease.**

### **Discussion**

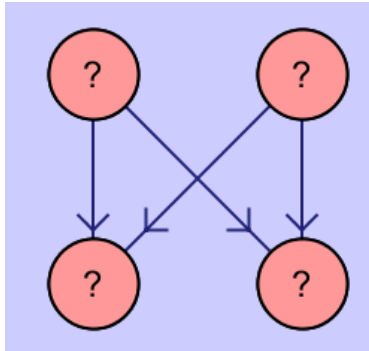
The purpose of these studies was to discover if there is an optimized sub-network (motif) structure that could be used to identify functional groups. Optimization, in this context, can have at least two meanings. If the result of the optimized network is to accelerate the flow of information, then the optimized sub-network should have a higher frequency in biological networks. If acceleration of information flow is a secondary result of the optimization, and the primary result is to control the expression of pathways connected to the motif, then the sub-network should have a lower frequency in biological networks. I expected the second meaning of optimization to include biomolecular signaling networks associated with control. Although the design of this study was heavily influenced by these definitions, they are not exhaustive and need further investigation prior to being considered theory.

This study was limited to the optimization associated with pathway control seen in the Apoptosis in Lung Cancer network and the Autism network. This is because these two networks had the highest confidence, while being significantly

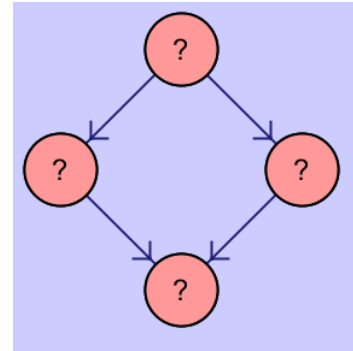
diverse and of large enough size. After normalization by use of the semantic analysis of the PSI-MI tags, the Apoptosis network had 3400 nodes representing proteins, genes, ligands and small molecules. This network also had 5857 edges. Compared to the original network of 5724 nodes and 13810 edges, there was a reduction of 41% by node and a 58% reduction by edge from the semantic optimization. This optimization likely influenced the occurrence of the sub-networks by removing false positives. In each of the four node networks, tested in the CA simulation of the Bcl2 family proteins, the occurrence in the apoptosis network was lower than the null hypothesis (mean occurrence in simulated random networks), the opposite occurred in the Autism network. The three 4-node sub-networks are displayed in figures 18-21 (reproduced below).



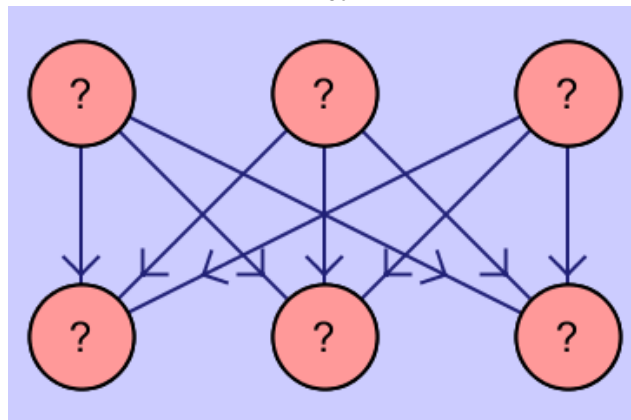
**Figure 24 – 4-Node, 4-Edge Non-Symmetrical Non-Linear Motif**



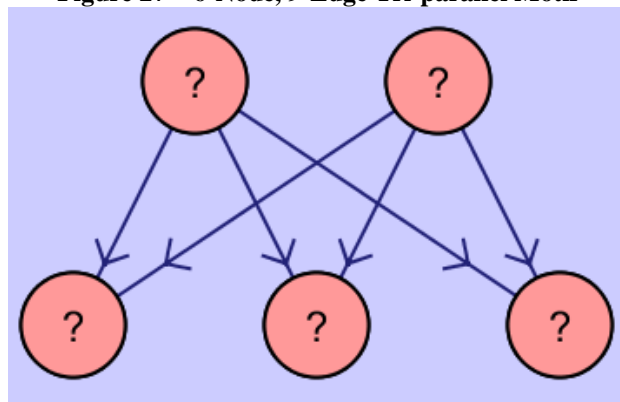
**Figure 25 – 4-Node, 4-Edge Symmetrical Non-Linear Bi-Fan Motif**



**Figure 26 – 4-Node, 4-Edge Bi-parallel Motif**



**Figure 27 – 6-Node, 9-Edge Tri-parallel Motif**



**Figure 28 – 5-Node, 6-Edge Tri-parallel Motif**

The motif in Figure 18 occurred at a ratio of 0.10 (ratio = (observed in network)/(mean occurrence in random networks)) in the Apoptosis network, versus 1.62 in the Autism network. This pattern continues through the motif in Figure 20 where the apoptosis network had a ratio of 0.67, vs. 2.07 in the Autism network. It also occurs in the motif in Figure 19 where the apoptosis network had

a ratio of 0.04, compared to 5.41 in the Autism network. Additionally this continued in the 6-node, 9-edge motif in Figure 21. The motif in Figure 22 could not be compared, because it was not significant in the apoptosis network, but the pattern remained in the Autism one.

The ratio of the motifs found in apoptosis (Table 7), were consistently low in apoptosis and consistently high in the autism network. The consistency of these results suggests there are fundamental differences in the apoptosis signaling and autism networks. Apoptosis has been studied in detail for decades, producing a wealth of interaction data available in the databases used in this study. Autism has only recently received attention and little of the pathogenicity of autism is understood. The apoptosis network needed to be normalized to remove the noise of repeat or irrelevant data that was produced by a basic search that yielded more proteins than could occur naturally. Autism required multiple search attempts revealing a single data-source of uncurated metadata. A possible reason for this inconsistency is the apoptosis signaling network is focused on one group of functions, those of cell death through apoptosis, and it is well understood and heavily studied. The autism network is likely a mix of functional groups and is still very incomplete. In apoptosis, these motifs do represent regions of functional control and therefore occur at a lower ratio when compared to random networks. In Autism these motifs are over represented because Autism is associated with a mix of functional groups and the network is incomplete. Additionally, in autism these motifs are occurring at a higher ratio because their role in controlling signaling and pathway selection/function increases the likelihood they are observed during the onset of autism.

Based on this result, the ratio of motifs could be used to identify whether a larger network represents more than one functional group, possibly as a measure of biomolecular network completeness. Although this is not evidence the hypothesis holds true, this study suggests the ratio of these motifs area a reporter for the presence or absence of functional groups in larger biomolecular networks.

**Future work**

In order to understand the entire role of each of these motifs that hint at function, a study of the expression of their associated genes and proteins needs to be done. This would require the creation of a software based tool that will extract expression data from a data-source, calculate the average across all related disease types and stages, and then compare the ratio of occurrence of the motif with the expression of each gene and protein represented in the network. Cytoscape's NetMatch would be an excellent tool because it has the ability to report the identity of each motif, even with over-represented interactions (where more than one edge is present between two nodes). The result of this study would be a list of network motifs and genes/proteins that would increase the likelihood for the network to be complete, and whether it includes few or many functional groups. This would be a significant expansion of the use of network motifs in studying biomolecular networks and their control of cell function and disease.

## Bibliography

## Bibliography

Abrahams, B. S. and D. H. Geschwind (2008). "Advances in autism genetics: on the threshold of a new neurobiology." Nat Rev Genet.

Aranda, B., P. Achuthan, et al. "The IntAct molecular interaction database in 2010." Nucleic Acids Res **38**(Database issue): D525-31.

IntAct is an open-source, open data molecular interaction database and toolkit. Data is abstracted from the literature or from direct data depositions by expert curators following a deep annotation model providing a high level of detail. As of September 2009, IntAct contains over 200,000 curated binary interaction evidences. In response to the growing data volume and user requests, IntAct now provides a two-tiered view of the interaction data. The search interface allows the user to iteratively develop complex queries, exploiting the detailed annotation with hierarchical controlled vocabularies. Results are provided at any stage in a simplified, tabular view. Specialized views then allows 'zooming in' on the full annotation of interactions, interactors and their properties. IntAct source code and data are freely available at <http://www.ebi.ac.uk/intact>.

Ashburner, M., C. A. Ball, et al. (2000). "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium." Nat Genet **25**(1): 25-9.

Bailey-Wilson, J. E., C. I. Amos, et al. (2004). "A major lung cancer susceptibility locus maps to chromosome 6q23-25." Am J Hum Genet **75**(3): 460-74.

Lung cancer is a major cause of death in the United States and other countries. The risk of lung cancer is greatly increased by cigarette smoking and by certain occupational exposures, but familial factors also clearly play a major role. To identify susceptibility genes for familial lung cancer, we conducted a genomewide linkage analysis of 52 extended pedigrees ascertained through probands with lung cancer who had several first-degree relatives with the same disease. Multipoint linkage analysis, under a simple autosomal dominant model, of all 52 families with three or more individuals affected by lung, throat, or laryngeal cancer, yielded a maximum heterogeneity LOD score (HLOD) of 2.79 at 155 cM on chromosome 6q (marker D6S2436). A subset of 38 pedigrees with four or more affected individuals yielded a multipoint HLOD of 3.47 at 155 cM. Analysis of a further subset of 23 multigenerational pedigrees with five or more affected individuals yielded a multipoint HLOD score of 4.26 at the same position. The 14 families with only three affected relatives yielded negative LOD scores in this region. A predivided samples test for heterogeneity comparing the LOD scores from the 23 multigenerational families with those from the remaining families was significant ( $P=.007$ ). The 1-HLOD multipoint support interval from the multigenerational families extends from C6S1848 at 146 cM to 164 cM near D6S1035, overlapping a genomic region that is deleted in sporadic lung cancers as well as numerous other cancer types. Parametric linkage and variance-components analysis that incorporated effects of age and personal smoking also supported linkage in this region, but with somewhat diminished support. These



- results localize a major susceptibility locus influencing lung cancer risk to 6q23-25.
- Baker, N. A., D. Sept, et al. (2001). "Electrostatics of nanosystems: application to microtubules and the ribosome." Proc Natl Acad Sci U S A **98**(18): 10037-41.
- Bard, J. (2003). "Ontologies: Formalising biological knowledge for bioinformatics." Bioessays **25**(5): 501-6.
- Baron, C. A., S. Y. Liu, et al. (2006). "Utilization of lymphoblastoid cell lines as a system for the molecular modeling of autism." J Autism Dev Disord **36**(8): 973-82.
- Baron, C. A., C. G. Tepper, et al. (2006). "Genomic and functional profiling of duplicated chromosome 15 cell lines reveal regulatory alterations in UBE3A-associated ubiquitin-proteasome pathway processes." Hum Mol Genet **15**(6): 853-69.
- Barrett, T., D. B. Troup, et al. (2007). "NCBI GEO: mining tens of millions of expression profiles--database and tools update." Nucleic Acids Res **35**(Database issue): D760-5.
- The Gene Expression Omnibus (GEO) repository at the National Center for Biotechnology Information (NCBI) archives and freely disseminates microarray and other forms of high-throughput data generated by the scientific community. The database has a minimum information about a microarray experiment (MIAME)-compliant infrastructure that captures fully annotated raw and processed data. Several data deposit options and formats are supported, including web forms, spreadsheets, XML and Simple Omnibus Format in Text (SOFT). In addition to data storage, a collection of user-friendly web-based interfaces and applications are available to help users effectively explore, visualize and download the thousands of experiments and tens of millions of gene expression patterns stored in GEO. This paper provides a summary of the GEO database structure and user facilities, and describes recent enhancements to database design, performance, submission format options, data query and retrieval utilities. GEO is accessible at <http://www.ncbi.nlm.nih.gov/geo/>
- Bhattacharjee, A., W. G. Richards, et al. (2001). "Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses." Proc Natl Acad Sci U S A **98**(24): 13790-5.
- We have generated a molecular taxonomy of lung carcinoma, the leading cause of cancer death in the United States and worldwide. Using oligonucleotide microarrays, we analyzed mRNA expression levels corresponding to 12,600 transcript sequences in 186 lung tumor samples, including 139 adenocarcinomas resected from the lung. Hierarchical and probabilistic clustering of expression data defined distinct subclasses of lung adenocarcinoma. Among these were tumors with high relative expression of neuroendocrine genes and of type II pneumocyte genes, respectively. Retrospective analysis revealed a less favorable outcome for the adenocarcinomas with neuroendocrine gene expression. The diagnostic potential of expression profiling is emphasized by its ability to

discriminate primary lung adenocarcinomas from metastases of extra-pulmonary origin. These results suggest that integration of expression profile data with clinical parameters could aid in diagnosis of lung cancer patients.

Bonchev, D., S. Thomas, et al. "Cellular automata modelling of biomolecular networks dynamics." *SAR QSAR Environ Res* **21**(1): 77-102.

The modelling of biological systems dynamics is traditionally performed by ordinary differential equations (ODEs). When dealing with intracellular networks of genes, proteins and metabolites, however, this approach is hindered by network complexity and the lack of experimental kinetic parameters. This opened the field for other modelling techniques, such as cellular automata (CA) and agent-based modelling (ABM). This article reviews this emerging field of studies on network dynamics in molecular biology. The basics of the CA technique are discussed along with an extensive list of related software and websites. The application of CA to networks of biochemical reactions is exemplified in detail by the case studies of the mitogen-activated protein kinase (MAPK) signalling pathway, the FAS-ligand (FASL)-induced and Bcl-2-related apoptosis. The potential of the CA method to model basic pathways patterns, to identify ways to control pathway dynamics and to help in generating strategies to fight with cancer is demonstrated. The different line of CA applications presented includes the search for the best-performing network motifs, an analysis of importance for effective intracellular signalling and pathway cross-talk.

Cameron, E. E., K. E. Bachman, et al. (1999). "Synergy of demethylation and histone deacetylase inhibition in the re-expression of genes silenced in cancer." *Nat Genet* **21**(1): 103-7.

Densely methylated DNA associates with transcriptionally repressive chromatin characterized by the presence of underacetylated histones. Recently, these two epigenetic processes have been dynamically linked. The methyl-CpG-binding protein MeCP2 appears to reside in a complex with histone deacetylase activity. MeCP2 can mediate formation of transcriptionally repressive chromatin on methylated promoter templates *in vitro*, and this process can be reversed by trichostatin A (TSA), a specific inhibitor of histone deacetylase. Little is known, however, about the relative roles of methylation and histone deacetylase activity in the stable inhibition of transcription on densely methylated endogenous promoters, such as those for silenced alleles of imprinted genes, genes on the female inactive X chromosome and tumour-suppressor genes inactivated in cancer cells. We show here that the hypermethylated genes *MLH1*, *TIMP3* (*TIMP3*), *CDKN2B* (*INK4B*, p15) and *CDKN2A* (*INK4*, p16) cannot be transcriptionally reactivated with TSA alone in tumour cells in which we have shown that TSA alone can upregulate the expression of non-methylated genes. Following minimal demethylation and slight gene reactivation in the presence of low dose 5-aza-2'-deoxycytidine (5Aza-dC), however, TSA treatment results in robust re-expression of each gene. TSA does not contribute to demethylation of the genes, and none of the treatments alter the chromatin structure associated with the hypermethylated promoters. Thus, although DNA methylation and histone

deacetylation appear to act as synergistic layers for the silencing of genes in cancer, dense CpG island methylation is dominant for the stable maintenance of a silent state at these loci.

Ceol, A., A. Chatr Aryamontri, et al. "MINT, the molecular interaction database: 2009 update." Nucleic Acids Res **38**(Database issue): D532-9.

MINT (<http://mint.bio.uniroma2.it/mint>) is a public repository for molecular interactions reported in peer-reviewed journals. Since its last report, MINT has grown considerably in size and evolved in scope to meet the requirements of its users. The main changes include a more precise definition of the curation policy and the development of an enhanced and user-friendly interface to facilitate the analysis of the ever-growing interaction dataset. MINT has adopted the PSI-MI standards for the annotation and for the representation of molecular interactions and is a member of the IMEx consortium.

Chautard, E., L. Ballut, et al. (2009). "MatrixDB, a database focused on extracellular protein-protein and protein-carbohydrate interactions." Bioinformatics **25**(5): 690-1.

SUMMARY: MatrixDB (<http://matrixdb.ibcp.fr>) is a database reporting mammalian protein-protein and protein-carbohydrate interactions involving extracellular molecules. It takes into account the full interaction repertoire of the extracellular matrix involving full-length molecules, fragments and multimers. The current version of MatrixDB contains 1972 interactions corresponding to 4412 experiments and involving 259 extracellular biomolecules.

AVAILABILITY: MatrixDB is freely available at <http://matrixdb.ibcp.fr>

Cheng, E. H.-Y. A., M. C. Wei, et al. (2001). "Bcl-2, Bcl-X Sequester BH3 Domain-Only Molecules Preventing BAX- and BAK Mediated Mitochondrial Apoptosis." Molecular Cell **8**(3): 705-711.

Croft, D., G. O'Kelly, et al. "Reactome: a database of reactions, pathways and biological processes." Nucleic Acids Res.

Reactome (<http://www.reactome.org>) is a collaboration among groups at the Ontario Institute for Cancer Research, Cold Spring Harbor Laboratory, New York University School of Medicine and The European Bioinformatics Institute, to develop an open source curated bioinformatics database of human pathways and reactions. Recently, we developed a new web site with improved tools for pathway browsing and data analysis. The Pathway Browser is an Systems Biology Graphical Notation (SBGN)-based visualization system that supports zooming, scrolling and event highlighting. It exploits PSIQUIC web services to overlay our curated pathways with molecular interaction data from the Reactome Functional Interaction Network and external interaction databases such as IntAct, BioGRID, ChEMBL, iRefIndex, MINT and STRING. Our Pathway and Expression Analysis tools enable ID mapping, pathway assignment and overrepresentation analysis of user-supplied data sets. To support pathway annotation and analysis in other species, we continue to make orthology-based inferences of pathways in non-human species, applying Ensembl Compara to identify orthologs of curated

human proteins in each of 20 other species. The resulting inferred pathway sets can be browsed and analyzed with our Species Comparison tool. Collaborations are also underway to create manually curated data sets on the Reactome framework for chicken, *Drosophila* and rice.

Dorogovtsev, S. N., A. V. Goltsev, et al. (2003). "Spectra of complex networks." Phys Rev E Stat Nonlin Soft Matter Phys **68**(4 Pt 2): 046109.

We propose a general approach to the description of spectra of complex networks. For the spectra of networks with uncorrelated vertices (and a local treelike structure), exact equations are derived. These equations are generalized to the case of networks with correlations between neighboring vertices. The tail of the density of eigenvalues  $\rho(\lambda)$  at large  $|\lambda|$  is related to the behavior of the vertex degree distribution  $P(k)$  at large  $k$ . In particular, as  $P(k)$  approximately  $k^{-\gamma}$ ,  $\rho(\lambda)$  approximately  $|\lambda|^{-(1-2\gamma)}$ . We propose a simple approximation, which enables us to calculate spectra of various graphs analytically. We analyze spectra of various complex networks and discuss the role of vertices of low degree. We show that spectra of locally treelike random graphs may serve as a starting point in the analysis of spectral properties of real-world networks, e.g., of the Internet.

Ferro, A., R. Giugno, et al. (2007). "NetMatch: a Cytoscape plugin for searching biological networks." Bioinformatics **23**(7): 910-2.

NetMatch is a Cytoscape plugin which allows searching biological networks for subcomponents matching a given query. Queries may be approximate in the sense that certain parts of the subgraph-query may be left unspecified. To make the query creation process easy, a drawing tool is provided. Cytoscape is a bioinformatics software platform for the visualization and analysis of biological networks. AVAILABILITY: The full package, a tutorial and associated examples are available at the following web sites:

<http://alpha.dmi.unict.it/~ctnyu/netmatch.html>,

<http://baderlab.org/Software/NetMatch>.

Gibbons, R. J. (2005). "Histone modifying and chromatin remodelling enzymes in cancer and dysplastic syndromes." Hum Mol Genet **14 Spec No 1**: R85-92.

Inactivation of tumour suppressor genes is central to the development of cancer. Although this inactivation was once considered to be secondary to intragenic mutations, it is now clear that silencing of these genes often occurs by epigenetic means. Hypermethylation of CpG islands associated with the tumour suppressor genes was the first manifestation of this phenomenon to be described. It is apparent, however, that this is one of a host of chromatin modifications which characterize gene silencing. Although we know little about what determines which loci are affected, our understanding of the nature of the epigenetic marks and how they are established has blossomed. There is no compelling evidence that cancer ever develops by purely epigenetic means, but it is apparent that perturbations in the apparatus which establish the epigenome may contribute to the development of cancer. This review will focus on the role of two classes of

chromatin remodelling enzymes, those that alter histones by the addition or removal of acetyl and methyl groups and those of the SWI/SNF family of proteins that change the topology of the nucleosome and its DNA strand via the hydrolysis of ATP, and we shall examine the consequence of mutations in, or mis-expression of, these factors. In some cases, mutations in these factors appear to play a direct role in cancer development. However, their general role as important intermediaries involved in regulating gene expression makes them attractive therapeutic targets. In exciting developments, it has been shown that inhibition of these factors leads to the reversal of tumour suppressor gene silencing and the inhibition of cancer cell growth.

Goll, J., S. V. Rajagopala, et al. (2008). "MPIDB: the microbial protein interaction database." Bioinformatics **24**(15): 1743-4.

SUMMARY: The microbial protein interaction database (MPIDB) aims to collect and provide all known physical microbial interactions. Currently, 22,530 experimentally determined interactions among proteins of 191 bacterial species/strains can be browsed and downloaded. These microbial interactions have been manually curated from the literature or imported from other databases (IntAct, DIP, BIND, MINT) and are linked to 24,060 experimental evidences (PubMed ID, PSI-MI methods). In contrast to these databases, interactions in MPIDB are further supported by 8150 additional evidences based on interaction conservation, co-purification and 3D domain contacts (iPfam, 3did).

AVAILABILITY: <http://www.jcvi.org/mpidb/>

Gregg, J. P., L. Lit, et al. (2008). "Gene expression changes in children with autism." Genomics **91**(1): 22-9.

Hoffmann, R. and A. Valencia (2004). "A gene network for navigating the literature." Nat Genet **36**(7): 664.

Hu, V. W., B. C. Frank, et al. (2006). "Gene expression profiling of lymphoblastoid cell lines from monozygotic twins discordant in severity of autism reveals differential regulation of neurologically relevant genes." BMC Genomics **7**: 118.

Insel, T. R. and T. Lehner (2007). "A new era in psychiatric genetics?" Biol Psychiatry **61**(9): 1017-8.

Jourdan, M., T. Reme, et al. (2009). "Gene expression of anti- and pro-apoptotic proteins in malignant and normal plasma cells." Br J Haematol **145**(1): 45-58.

The survival of malignant plasma cells is a key event in disease occurrence, progression and chemoresistance. Using DNA-microarrays, we analysed the expression of genes coding for 58 proteins linked with extrinsic and intrinsic apoptotic pathways, caspases and inhibitor of apoptosis proteins. We considered six memory B cells (MBC), seven plasmablasts (PPC), seven bone marrow plasma cells (BMPC) and purified myeloma cells (MMC) from 92 newly-diagnosed patients. Forty out of the 58 probe sets enabled the separation of MBC,

PPC and BMPC in three homogeneous clusters, characterized by an elevated expression of TNFRSF10A, TNFRSF10B, BCL2A1, CASP8, CASP9 and PMAIP1 genes for MBC, of FAS, FADD, AIFM1, BIRC5, CASP2, CASP3 and CASP6 for PPC and of BCL2, MCL1, BID, BIRC3 and XIAP for BMPC. Thus, B cell differentiation was associated with change of expression of pro-apoptotic and anti-apoptotic genes. Regarding MMC, the major finding was TRAIL upregulation that might be counteracted by a high osteoprotegerin production by BM stromal cells and a decreased expression of FAS, APAF1 and BNIP3 compared to normal BMPC. Out of the 40 genes, CASP2 and BIRC5 expression in MMC had adverse prognosis in two independent series of previously-untreated patients.

Kaser, O., Lamire, D. (2007). Tag-Cloud Drawing: Algorithms for Cloud Visualization. Tagging and Metadata for Social Information Organization (WWW 2007).

Kerrien, S., Y. Alam-Faruque, et al. (2007). "IntAct--open source resource for molecular interaction data." Nucleic Acids Res **35**(Database issue): D561-5.

IntAct is an open source database and software suite for modeling, storing and analyzing molecular interaction data. The data available in the database originates entirely from published literature and is manually annotated by expert biologists to a high level of detail, including experimental methods, conditions and interacting domains. The database features over 126,000 binary interactions extracted from over 2100 scientific publications and makes extensive use of controlled vocabularies. The web site provides tools allowing users to search, visualize and download data from the repository. IntAct supports and encourages local installations as well as direct data submission and curation collaborations. IntAct source code and data are freely available from <http://www.ebi.ac.uk/intact>.

Kier, L. B. (2008). "A review of recent studies relating ligand diffusion, general anesthesia, and sleep." AANA J **76**(2): 109-12.

This review article presents 3 theories related to ligand diffusion, general anesthesia and sleep. The first theory describes the diffusion of molecules across a protein surface to a receptor. It is based on the effect of the amino acid side chains on the protein surface on the structure of bulk water nearby. This influence creates pathways, called chreodes, through the water near the protein surface, permitting a rapid diffusion of molecules to the receptors. A second theory involving the role of chreodes presents a mechanism of action of nonspecific anesthetic agents. These agents interrupt the diffusion of neurotransmitter molecules to their receptors, bringing on the anesthetic effects. Finally, building on the similarities of anesthesia and sleep, a theory is presented proposing that an external agent influences sleep in a way similar to that of the nonspecific anesthetic molecules. This external agent is proposed to be elemental nitrogen. Several observations are presented to support this mechanism.

Kier, L. B., D. Bonchev, et al. (2005). "Modeling biochemical networks: a cellular-automata approach." Chem Biodivers **2**(2): 233-43.

The potential of the cellular-automata (CA) method for modeling biological networks is demonstrated for the mitogen-activated protein kinase (MAPK) signaling cascade. The models derived reproduced the high signal amplification through the cascade and the deviation of the cascade enzymes from the Michaelis-Menten kinetics, evidencing cooperativity effects. The patterns of pathway change upon varying substrate concentrations and enzyme efficiencies were identified and used to show the ways for controlling pathway processes. Guidance in the selection of enzyme inhibition targets with minimum side effects is one outcome of the study.

Kier, L. B., P. G. Seybold, et al. (2010). Cellular Automata Modeling of Chemical Systems, Springer.

Kuznetsov, V., S. Thomas, et al. (2008). "Data-driven Networking Reveals 5-Genes Signature for Early Detection of Lung Cancer." 2008 International Conference on BioMedical Engineering and Informatics 1: 413-417.

Medina, P. P., J. Carretero, et al. (2005). "Transcriptional targets of the chromatin-remodelling factor SMARCA4/BRG1 in lung cancer cells." Hum Mol Genet 14(7): 973-82.

BRG1, also called SMARCA4, is the catalytic subunit of the SWI/SNF chromatin-remodelling complex and influences transcriptional regulation by disrupting histone-DNA contacts in an ATP-dependent manner. BRG1 and other members of the SWI/SNF complex become inactivated in tumours, implying a role in cancer development. To understand the contribution of BRG1 to lung tumourigenesis, we restored BRG1 in H1299 lung cancer cells and used cDNA microarray analysis to identify changes in gene expression. Forty-three transcripts became activated, whereas two were repressed. Chromatin immunoprecipitation of resulting candidate genes revealed that the CYP3A4 and ZNF185 promoters recruited BRG1 and that recruitment to the CYP3A4 promoter was specific to this gene and did not involve the CYP3A5 or CYP3A7 family members. Moreover, specifically BRG1 but not its homologue BRM was recruited to the CYP3A4 and ZNF185 promoters. To explore their potential relevance in lung tumours, levels of CYP3A4 and ZNF185 transcripts were evaluated in seven additional lung cancer cell lines. CYP3A4 was undetectable in any of the lung cancer cells tested, and only the CYP3A5 family member was present in the A549 and Calu-3 cells. In contrast, the amount of ZNF185 transcript clearly varied among lung cancer cell lines and severely reduced levels were observed in BRG1-deficient cells, except those of A427. We extended these observations to 27 lung primary tumours using real-time RT-PCR (TaqMan) and observed that four (15%) and 14 (52%) of them had BRG1 and ZNF185 downregulation, respectively, when compared with normal lung. No significant correlation between reduced levels of BRG1 and ZNF185 was observed, indicating that additional mechanisms to BRG1 inactivation may contribute to the loss of ZNF185 expression in lung tumours. In conclusion, our results provide evidence that transcriptional activation of ZNF185 and CYP3A4 is mediated by direct association of BRG1 with their promoters and



also indicate that a decreased level of ZNF185 is a common feature of lung tumours and may be of biological relevance in lung carcinogenesis.

Milo, R., S. Shen-Orr, et al. (2002). "Network motifs: simple building blocks of complex networks." *Science* **298**(5594): 824-7.

Complex networks are studied across many fields of science. To uncover their structural design principles, we defined "network motifs," patterns of interconnections occurring in complex networks at numbers that are significantly higher than those in randomized networks. We found such motifs in networks from biochemistry, neurobiology, ecology, and engineering. The motifs shared by ecological food webs were distinct from the motifs shared by the genetic networks of *Escherichia coli* and *Saccharomyces cerevisiae* or from those found in the World Wide Web. Similar motifs were found in networks that perform information processing, even though they describe elements as different as biomolecules within a cell and synaptic connections between neurons in *Caenorhabditis elegans*. Motifs may thus define universal classes of networks. This approach may uncover the basic building blocks of most networks.

Montecchi-Palazzi, L., S. Kerrien, et al. (2009). "The PSI semantic validator: a framework to check MIAPE compliance of proteomics data." *Proteomics* **9**(22): 5112-9.

The Human Proteome Organization's Proteomics Standards Initiative (PSI) promotes the development of exchange standards to improve data integration and interoperability. PSI specifies the suitable level of detail required when reporting a proteomics experiment (via the Minimum Information About a Proteomics Experiment), and provides extensible markup language (XML) exchange formats and dedicated controlled vocabularies (CVs) that must be combined to generate a standard compliant document. The framework presented here tackles the issue of checking that experimental data reported using a specific format, CVs and public bio-ontologies (e.g. Gene Ontology, NCBI taxonomy) are compliant with the Minimum Information About a Proteomics Experiment recommendations. The semantic validator not only checks the XML syntax but it also enforces rules regarding the use of an ontology class or CV terms by checking that the terms exist in the resource and that they are used in the correct location of a document. Moreover, this framework is extremely fast, even on sizable data files, and flexible, as it can be adapted to any standard by customizing the parameters it requires: an XML Schema Definition, one or more CVs or ontologies, and a mapping file describing in a formal way how the semantic resources and the format are interrelated. As such, the validator provides a general solution to the common problem in data exchange: how to validate the correct usage of a data standard beyond simple XML Schema Definition validation. The framework source code and its various applications can be found at <http://psidev.info/validator>.

Nelson, R. T., S. Avraham, et al. "Applications and methods utilizing the Simple Semantic Web Architecture and Protocol (SSWAP) for bioinformatics resource discovery and disparate data and service integration." *BioData Min* **3**(1): 3.



**ABSTRACT: BACKGROUND:** Scientific data integration and computational service discovery are challenges for the bioinformatic community. This process is made more difficult by the separate and independent construction of biological databases, which makes the exchange of data between information resources difficult and labor intensive. A recently described semantic web protocol, the Simple Semantic Web Architecture and Protocol (SSWAP; pronounced "swap") offers the ability to describe data and services in a semantically meaningful way. We report how three major information resources (Gramene, SoyBase and the Legume Information System [LIS]) used SSWAP to semantically describe selected data and web services. **METHODS:** We selected high-priority Quantitative Trait Locus (QTL), genomic mapping, trait, phenotypic, and sequence data and associated services such as BLAST for publication, data retrieval, and service invocation via semantic web services. Data and services were mapped to concepts and categories as implemented in legacy and de novo community ontologies. We used SSWAP to express these offerings in OWL Web Ontology Language (OWL), Resource Description Framework (RDF) and eXtensible Markup Language (XML) documents, which are appropriate for their semantic discovery and retrieval. We implemented SSWAP services to respond to web queries and return data. These services are registered with the SSWAP Discovery Server and are available for semantic discovery at <http://sswap.info>. **RESULTS:** A total of ten services delivering QTL information from Gramene were created. From SoyBase, we created six services delivering information about soybean QTLs, and seven services delivering genetic locus information. For LIS we constructed three services, two of which allow the retrieval of DNA and RNA FASTA sequences with the third service providing nucleic acid sequence comparison capability (BLAST). **CONCLUSIONS:** The need for semantic integration technologies has preceded available solutions. We report the feasibility of mapping high priority data from local, independent, idiosyncratic data schemas to common shared concepts as implemented in web-accessible ontologies. These mappings are then amenable for use in semantic web services. Our implementation of approximately two dozen services means that biological data at three large information resources (Gramene, SoyBase, and LIS) is available for programmatic access, semantic searching, and enhanced interaction between the separate missions of these resources.

Neumann, J. V. and A. W. Burks (1966). Theory of Self-Reproducing Automata. Urbana and London, University of Illinois Press.

Nikitin, A., S. Egorov, et al. (2003). "Pathway studio--the analysis and navigation of molecular networks." Bioinformatics **19**(16): 2155-7.

**SUMMARY:** PathwayAssist is a software application developed for navigation and analysis of biological pathways, gene regulation networks and protein interaction maps. It comes with the built-in natural language processing module MedScan and the comprehensive database describing more than 100 000 events of regulation, interaction and modification between proteins, cell processes and small molecules. **AVAILABILITY:** PathwayAssist is available for commercial

licensing from Ariadne Genomics, Inc. The light version with limited functionality will be available for free for academic users at [www.ariadnegenomics.com/downloads/](http://www.ariadnegenomics.com/downloads/).

Nishimura, Y., C. L. Martin, et al. (2007). "Genome-wide expression profiling of lymphoblastoid cell lines distinguishes different forms of autism and reveals shared pathways." *Hum Mol Genet* **16**(14): 1682-98.

Overington, J. (2009). "ChEMBL. An interview with John Overington, team leader, chemogenomics at the European Bioinformatics Institute Outstation of the European Molecular Biology Laboratory (EMBL-EBI). Interview by Wendy A. Warr." *J Comput Aided Mol Des* **23**(4): 195-8.

Purcell, A. E., O. H. Jeon, et al. (2001). "Postmortem brain abnormalities of the glutamate neurotransmitter system in autism." *Neurology* **57**(9): 1618-28.

Razick, S., G. Magklaras, et al. (2008). "iRefIndex: a consolidated protein interaction database with provenance." *BMC Bioinformatics* **9**: 405.

**BACKGROUND:** Interaction data for a given protein may be spread across multiple databases. We set out to create a unifying index that would facilitate searching for these data and that would group together redundant interaction data while recording the methods used to perform this grouping. **RESULTS:** We present a method to generate a key for a protein interaction record and a key for each participant protein. These keys may be generated by anyone using only the primary sequence of the proteins, their taxonomy identifiers and the Secure Hash Algorithm. Two interaction records will have identical keys if they refer to the same set of identical protein sequences and taxonomy identifiers. We define records with identical keys as a redundant group. Our method required that we map protein database references found in interaction records to current protein sequence records. Operations performed during this mapping are described by a mapping score that may provide valuable feedback to source interaction databases on problematic references that are malformed, deprecated, ambiguous or unfound. Keys for protein participants allow for retrieval of interaction information independent of the protein references used in the original records.

**CONCLUSION:** We have applied our method to protein interaction records from BIND, BioGrid, DIP, HPRD, IntAct, MINT, MPact, MPPI and OPHID. The resulting interaction reference index is provided in PSI-MITAB 2.5 format at <http://irefindex.uio.no>. This index may form the basis of alternative redundant groupings based on gene identifiers or near sequence identity groupings.

Reisman, D. N., J. Sciarrotta, et al. (2003). "Loss of BRG1/BRM in human lung cancer cell lines and primary lung cancers: correlation with poor prognosis." *Cancer Res* **63**(3): 560-6.

A role for the SWI/SNF complex in tumorigenesis based on its requirement for retinoblastoma induced growth arrest and p53-mediated transcription and the appearance of tumors in SWI/SNF-deficient mice. In addition, Western blot data

have shown that the SWI/SNF ATPase subunits cell, BRG1 and BRM (BRG1/BRM), are lost in approximately 30% of human non-small lung cancer cell lines. To determine whether loss of expression of these proteins occurs in primary tumors, we examined their expression in 41 primary lung adenocarcinomas and 19 primary lung squamous carcinomas by immunohistochemistry. These analyses showed that 10% of tumors show a concomitant loss of BRG1 and BRM expression. Moreover, patients with BRG1/BRM-negative carcinomas, independent of stage, have a statistically significant decrease in survival compared with patients with BRG1/BRM. This report provides supportive evidence that BRG1 and BRM act as tumor suppressor proteins and implicates a role for their loss in the development of non-small cell lung cancers.

Renouf, D. J., R. Wood-Baker, et al. (2009). "BCL-2 expression is prognostic for improved survival in non-small cell lung cancer." *J Thorac Oncol* **4**(4): 486-91.

**OBJECTIVE:** We used a large patient population to identify immunohistochemical biomarkers to enable improved prognostication in patients with non-small cell lung carcinoma (NSCLC). **METHODS:** A tissue microarray was constructed using duplicate 0.6 mm cores of formalin-fixed paraffin-embedded tissue blocks from 609 patients with NSCLC. Immunohistochemical was used to detect 11 biomarkers including epidermal growth factor receptor, Her2, Her3, p53, p63, bcl-1, bcl-2, Thyroid transcription factor, carcinoembryonic antigen, chromogranin, and synaptophysin. A clinical database was generated prospectively at the time of tissue collection. Survival outcomes were obtained from a Provincial Cancer Registry database. Univariate and multivariate analyses were performed to look for a relationship between biomarker expression, smoking history, and survival. **RESULTS:** Survival data for 535 cases were available. As of June 2005, 429 patients (80%) had died; of these 286 (54%) died of lung cancer, 117 (22%) died of other known causes, and for 26 (5%) the cause of death was not available. Univariate analysis revealed that bcl-2 ( $p = 0.007$ ) was the only biomarker prognostic for improved overall survival (OS). bcl-2 ( $p = 0.021$ ) and p63 ( $p = 0.025$ ) were both found to be prognostic for improved disease-specific survival (DSS). Multivariate analysis (using age and biomarker expression) revealed that bcl-2 expression is prognostic for improved OS ( $p = 0.005$ ) and DSS ( $p = 0.021$ ). **CONCLUSIONS:** Our results suggest that bcl-2 expression is prognostic for improved OS and DSS in NSCLC. Testing for bcl-2 expression in a prospective study will help to determine its clinical relevance in prognostication.

Rhodes, D. R., S. Kalyana-Sundaram, et al. (2007). "Oncomine 3.0: genes, pathways, and networks in a collection of 18,000 cancer gene expression profiles." *Neoplasia* **9**(2): 166-80.

DNA microarrays have been widely applied to cancer transcriptome analysis; however, the majority of such data are not easily accessible or comparable. Furthermore, several important analytic approaches have been applied to microarray analysis; however, their application is often limited. To overcome these limitations, we have developed Oncomine, a bioinformatics initiative aimed

at collecting, standardizing, analyzing, and delivering cancer transcriptome data to the biomedical research community. Our analysis has identified the genes, pathways, and networks deregulated across 18,000 cancer gene expression microarrays, spanning the majority of cancer types and subtypes. Here, we provide an update on the initiative, describe the database and analysis modules, and highlight several notable observations. Results from this comprehensive analysis are available at <http://www.oncomine.org>.

Rives, A. W. and T. Galitski (2003). "Modular organization of cellular networks." Proc Natl Acad Sci U S A **100**(3): 1128-33.

Sasaki, H., S. Moriyama, et al. (2004). "Histone deacetylase 1 mRNA expression in lung cancer." Lung Cancer **46**(2): 171-8.

Histone deacetylases (HDACs) play a crucial role in tumorigenesis, however, the expression status of HDACs in lung cancer tissues has not been reported. We have investigated that HDAC 1 mRNA levels and other clinico-pathological data, including MTA 1 mRNA expression in lung cancer. The study included 102 lung cancer cases. The HDAC1 mRNA levels were quantified by real time reverse transcription-polymerase chain reaction (RT-PCR) using LightCycler (Roche Molecular Biochemicals, Mannheim, Germany). The HDAC1/GAPDH mRNA levels were not significantly different in tumor tissues from lung cancer (30.654 +/- 33.047) and adjacent non-malignant lung tissues (18.953 +/- 56.176 , P = 0.1827). No significant difference in HDAC1/GAPDH mRNA levels was found among age, gender, and lymph node metastasis. The HDAC1/GAPDH mRNA levels were significantly higher in stage III or IV lung cancer (50.929 +/- 120.433) than in stage I lung cancer (11.430 +/- 25.611, P = 0.0472). HDAC1/GAPDH mRNA levels were significantly higher in T3 or T4 lung carcinoma (54.326 +/- 127.018) than in T1 or T2 lung cancers (14.790 +/- 48.670, P = 0.1601). HDAC1/GAPDH mRNA levels were correlated with MTA1/GAPDH mRNA levels ( $y = 0.0106x + 2.5827$  , P = 0.0352 ). HDAC1/GAPDH mRNA levels were also correlated with HDAC1 protein (P = 0.0484) expression by immunohistochemistry. Using the LightCycler RT-PCR assay, the HDAC1 gene expression might correlate with progression of lung cancers. However, further studies are needed to confirm the impact of HDAC1 for the molecular target of the lung cancer.

Sayers, E. W., T. Barrett, et al. "Database resources of the National Center for Biotechnology Information." Nucleic Acids Res **38**(Database issue): D5-16.

In addition to maintaining the GenBank nucleic acid sequence database, the National Center for Biotechnology Information (NCBI) provides analysis and retrieval resources for the data in GenBank and other biological data made available through the NCBI web site. NCBI resources include Entrez, the Entrez Programming Utilities, MyNCBI, PubMed, PubMed Central, Entrez Gene, the NCBI Taxonomy Browser, BLAST, BLAST Link (BLink), Electronic PCR, OrfFinder, Spidey, Splign, Reference Sequence, UniGene, HomoloGene, ProtEST, dbMHC, dbSNP, Cancer Chromosomes, Entrez Genomes and related

tools, the Map Viewer, Model Maker, Evidence Viewer, Trace Archive, Sequence Read Archive, Retroviral Genotyping Tools, HIV-1/Human Protein Interaction Database, Gene Expression Omnibus, Entrez Probe, GENSAT, Online Mendelian Inheritance in Man, Online Mendelian Inheritance in Animals, the Molecular Modeling Database, the Conserved Domain Database, the Conserved Domain Architecture Retrieval Tool, Biosystems, Peptidome, Protein Clusters and the PubChem suite of small molecule databases. Augmenting many of the web applications are custom implementations of the BLAST program optimized to search specialized data sets. All these resources can be accessed through the NCBI home page at [www.ncbi.nlm.nih.gov](http://www.ncbi.nlm.nih.gov).

Scardoni, G., M. Petterlini, et al. (2009). "Analyzing biological network parameters with CentiScaPe." *Bioinformatics* **25**(21): 2857-9.

**SUMMARY:** The increasing availability of large network datasets along with the progresses in experimental high-throughput technologies have prompted the need for tools allowing easy integration of experimental data with data derived from network computational analysis. In order to enrich experimental data with network topological parameters, we have developed the Cytoscape plug-in CentiScaPe. The plug-in computes several network centrality parameters and allows the user to analyze existing relationships between experimental data provided by the users and node centrality values computed by the plug-in. CentiScaPe allows identifying network nodes that are relevant from both experimental and topological viewpoints. CentiScaPe also provides a Boolean logic-based tool that allows easy characterization of nodes whose topological relevance depends on more than one centrality. Finally, different graphic outputs and the included description of biological significance for each computed centrality facilitate the analysis by the end users not expert in graph theory, thus allowing easy node categorization and experimental prioritization.

**AVAILABILITY:** CentiScaPe can be downloaded via the Cytoscape web site: [http://chianti.ucsd.edu/cyto\\_web/plugins/index.php](http://chianti.ucsd.edu/cyto_web/plugins/index.php). Tutorial, centrality descriptions and example data are available at: <http://profs.sci.univr.it/approximatelyscardoni/centiscape/centiscapepage.php> **CONTACT:** giovanni.scardoni@gmail.com **SUPPLEMENTARY INFORMATION:** Supplementary data are available at Bioinformatics online.

Sherman, B. T., W. Huang da, et al. (2007). "DAVID Knowledgebase: a gene-centered database integrating heterogeneous gene annotation resources to facilitate high-throughput gene functional analysis." *BMC Bioinformatics* **8**: 426.

**BACKGROUND:** Due to the complex and distributed nature of biological research, our current biological knowledge is spread over many redundant annotation databases maintained by many independent groups. Analysts usually need to visit many of these bioinformatics databases in order to integrate comprehensive annotation information for their genes, which becomes one of the bottlenecks, particularly for the analytic task associated with a large gene list. Thus, a highly centralized and ready-to-use gene-annotation knowledgebase is in demand for high throughput gene functional analysis. **DESCRIPTION:** The

DAVID Knowledgebase is built around the DAVID Gene Concept, a single-linkage method to agglomerate tens of millions of gene/protein identifiers from a variety of public genomic resources into DAVID gene clusters. The grouping of such identifiers improves the cross-reference capability, particularly across NCBI and UniProt systems, enabling more than 40 publicly available functional annotation sources to be comprehensively integrated and centralized by the DAVID gene clusters. The simple, pair-wise, text format files which make up the DAVID Knowledgebase are freely downloadable for various data analysis uses. In addition, a well organized web interface allows users to query different types of heterogeneous annotations in a high-throughput manner. **CONCLUSION:** The DAVID Knowledgebase is designed to facilitate high throughput gene functional analysis. For a given gene list, it not only provides the quick accessibility to a wide range of heterogeneous annotation data in a centralized location, but also enriches the level of biological information for an individual gene. Moreover, the entire DAVID Knowledgebase is freely downloadable or searchable at <http://david.abcc.ncifcrf.gov/knowledgebase/>.

Sprinzak, E., S. J. Cokus, et al. (2009). "Detecting coordinated regulation of multi-protein complexes using logic analysis of gene expression." *BMC Syst Biol* **3**: 115.

**BACKGROUND:** Many of the functional units in cells are multi-protein complexes such as RNA polymerase, the ribosome, and the proteasome. For such units to work together, one might expect a high level of regulation to enable co-appearance or repression of sets of complexes at the required time. However, this type of coordinated regulation between whole complexes is difficult to detect by existing methods for analyzing mRNA co-expression. We propose a new methodology that is able to detect such higher order relationships. **RESULTS:** We detect coordinated regulation of multiple protein complexes using logic analysis of gene expression data. Specifically, we identify gene triplets composed of genes whose expression profiles are found to be related by various types of logic functions. In order to focus on complexes, we associate the members of a gene triplet with the distinct protein complexes to which they belong. In this way, we identify complexes related by specific kinds of regulatory relationships. For example, we may find that the transcription of complex C is increased only if the transcription of both complex A AND complex B is repressed. We identify hundreds of examples of coordinated regulation among complexes under various stress conditions. Many of these examples involve the ribosome. Some of our examples have been previously identified in the literature, while others are novel. One notable example is the relationship between the transcription of the ribosome, RNA polymerase and mannosyltransferase II, which is involved in N-linked glycan processing in the Golgi. **CONCLUSIONS:** The analysis proposed here focuses on relationships among triplets of genes that are not evident when genes are examined in a pairwise fashion as in typical clustering methods. By grouping gene triplets, we are able to decipher coordinated regulation among sets of three complexes. Moreover, using all triplets that involve coordinated regulation with the ribosome, we derive a large network involving this essential cellular complex. In this network we find that all multi-protein complexes that belong to the same



functional class are regulated in the same direction as a group (either induced or repressed).

Stark, C., B. J. Breitkreutz, et al. "The BioGRID Interaction Database: 2011 update." Nucleic Acids Res.

The Biological General Repository for Interaction Datasets (BioGRID) is a public database that archives and disseminates genetic and protein interaction data from model organisms and humans (<http://www.thebiogrid.org>). BioGRID currently holds 347 966 interactions (170 162 genetic, 177 804 protein) curated from both high-throughput data sets and individual focused studies, as derived from over 23 000 publications in the primary literature. Complete coverage of the entire literature is maintained for budding yeast (*Saccharomyces cerevisiae*), fission yeast (*Schizosaccharomyces pombe*) and thale cress (*Arabidopsis thaliana*), and efforts to expand curation across multiple metazoan species are underway. The BioGRID houses 48 831 human protein interactions that have been curated from 10 247 publications. Current curation drives are focused on particular areas of biology to enable insights into conserved networks and pathways that are relevant to human health. The BioGRID 3.0 web interface contains new search and display features that enable rapid queries across multiple data types and sources. An automated Interaction Management System (IMS) is used to prioritize, coordinate and track curation across international sites and projects. BioGRID provides interaction data to several model organism databases, resources such as Entrez-Gene and other interaction meta-databases. The entire BioGRID 3.0 data collection may be downloaded in multiple file formats, including PSI MI XML. Source code for BioGRID 3.0 is freely available without any restrictions.

Stein, L. D. (2004). "Human genome: end of the beginning." Nature **431**(7011): 915-6.

Strogatz, S. H. (2001). "Exploring complex networks." Nature **410**(6825): 268-76.

Su, L. J., C. W. Chang, et al. (2007). "Selection of DDX5 as a novel internal control for Q-RT-PCR from microarray data using a block bootstrap re-sampling scheme." BMC Genomics **8**: 140.

Thomas, P. D., M. J. Campbell, et al. (2003). "PANTHER: a library of protein families and subfamilies indexed by function." Genome Res **13**(9): 2129-41.

In the genomic era, one of the fundamental goals is to characterize the function of proteins on a large scale. We describe a method, PANTHER, for relating protein sequence relationships to function relationships in a robust and accurate way. PANTHER is composed of two main components: the PANTHER library (PANTHER/LIB) and the PANTHER index (PANTHER/X). PANTHER/LIB is a collection of "books," each representing a protein family as a multiple sequence alignment, a Hidden Markov Model (HMM), and a family tree. Functional divergence within the family is represented by dividing the tree into subtrees based on shared function, and by subtree HMMs. PANTHER/X is an abbreviated ontology for summarizing and navigating molecular functions and biological

processes associated with the families and subfamilies. We apply PANTHER to three areas of active research. First, we report the size and sequence diversity of the families and subfamilies, characterizing the relationship between sequence divergence and functional divergence across a wide range of protein families. Second, we use the PANTHER/X ontology to give a high-level representation of gene function across the human and mouse genomes. Third, we use the family HMMs to rank missense single nucleotide polymorphisms (SNPs), on a database-wide scale, according to their likelihood of affecting protein function.

Thomas, S. W. (2008). "Network Builder 1.-" USPTO Application.

Walker, S. J., J. Segal, et al. (2006). "Cultured lymphocytes from autistic children and non-autistic siblings up-regulate heat shock protein RNA in response to thimerosal challenge." Neurotoxicology **27**(5): 685-92.

Wall, D. P., F. J. Esteban, et al. (2009). "Comparative analysis of neurological disorders focuses genome-wide search for autism genes." Genomics **93**(2): 120-9.

The behaviors of autism overlap with a diverse array of other neurological disorders, suggesting common molecular mechanisms. We conducted a large comparative analysis of the network of genes linked to autism with those of 432 other neurological diseases to circumscribe a multi-disorder subcomponent of autism. We leveraged the biological process and interaction properties of these multi-disorder autism genes to overcome the across-the-board multiple hypothesis corrections that a purely data-driven approach requires. Using prior knowledge of biological process, we identified 154 genes not previously linked to autism of which 42% were significantly differentially expressed in autistic individuals. Then, using prior knowledge from interaction networks of disorders related to autism, we uncovered 334 new genes that interact with published autism genes, of which 87% were significantly differentially regulated in autistic individuals. Our analysis provided a novel picture of autism from the perspective of related neurological disorders and suggested a model by which prior knowledge of interaction networks can inform and focus genome-scale studies of complex neurological disorders.

Wernicke, S. and F. Rasche (2006). "FANMOD: a tool for fast network motif detection." Bioinformatics **22**(9): 1152-3.

**SUMMARY:** Motifs are small connected subnetworks that a network displays in significantly higher frequencies than would be expected for a random network. They have recently gathered much attention as a concept to uncover structural design principles of complex biological networks. FANMOD is a tool for fast network motif detection; it relies on recently developed algorithms to improve the efficiency of network motif detection by some orders of magnitude over existing tools. This facilitates the detection of larger motifs in bigger networks than previously possible. Additional benefits of FANMOD are the ability to analyze colored networks, a graphical user interface and the ability to export results to a



variety of machine- and human-readable file formats including comma-separated values and HTML.

Wolfram, S. (1983). "Statistical mechanics of cellular automata." Reviews of Modern Physics **55**(3): 601-644.

Wolfram, S. (2002). A New Kind of Science, Wolfram Media, Inc.

Xue, L. Y., S. M. Chiu, et al. (2003). "Photodamage to multiple Bcl-xL isoforms by photodynamic therapy with the phthalocyanine photosensitizer Pc 4." Oncogene **22**(58): 9197-204.

The antiapoptotic oncoprotein Bcl-2 is now a recognized phototarget of photodynamic therapy (PDT) with the phthalocyanine Pc 4 and with other mitochondrion-targeting photosensitizers. Photodamage, observed on Western blots as the loss of the native 26-kDa Bcl-2 protein, is PDT dose dependent and occurs in multiple cell lines, in the cold, and immediately upon photoirradiation. In our initial study, no photochemical damage was observed to Bcl-xL, in spite of its similarity in size, sequence, location and function to Bcl-2. The original study used a commercial anti-Bcl-xS/L antibody. We have revisited this issue by examining Western blots developed using one of three epitope-specific anti-Bcl-xL antibodies from commercial sources, a polyclonal antibody generated to the entire protein, as well as the antibody used previously. All five Bcl-xL antibodies recognized bacterially expressed Bcl-xL, but not Bcl-2, whereas an anti-Bcl-2 antibody recognized Bcl-2 and not Bcl-xL. All five Bcl-xL antibodies recognized at least one protein migrating at approximately 30 kDa; two of the antibodies recognized an additional band, migrating at approximately 33 or approximately 24 kDa. We now observe Pc 4-PDT-induced photodamage to all Bcl-xL-related proteins, except the 33-kDa species, in several human cancer cell lines. The results indicate that, in addition to the expected quantitative differences that may reflect exposure of individual epitopes, the antibodies also detect proteins of different apparent molecular weights that may be distinct isoforms or post-translationally modified forms of Bcl-xL. No evidence for PDT-induced phosphorylation or degradation was observed. Bcl-xL localized to mitochondria was considerably more sensitive to photodamage than was Bcl-xL in the cytosol, indicating that as previously found for Bcl-2, Bcl-xL must be membrane localized to be photosensitive.

Yonan, A. L., A. A. Palmer, et al. (2003). "Bioinformatic analysis of autism positional candidate genes using biological databases and computational gene network prediction." Genes Brain Behav **2**(5): 303-20.

Zhang, J., Y. Ji, et al. (2007). "Extracting three-way gene interactions from microarray data." Bioinformatics **23**(21): 2903-9.

Zhang, Y., J. Szustakowski, et al. (2009). "Bioinformatics analysis of microarray data." Methods Mol Biol **573**: 259-84.

Gene expression profiling provides unprecedented opportunities to study patterns of gene expression regulation, for example, in diseases or developmental processes. Bioinformatics analysis plays an important part of processing the information embedded in large-scale expression profiling studies and for laying the foundation for biological interpretation. Over the past years, numerous tools have emerged for microarray data analysis. One of the most popular platforms is Bioconductor, an open source and open development software project for the analysis and comprehension of genomic data, based on the R programming language. In this chapter, we use Bioconductor analysis packages on a heart development dataset to demonstrate the workflow of microarray data analysis from annotation, normalization, expression index calculation, and diagnostic plots to pathway analysis, leading to a meaningful visualization and interpretation of the data.

Zhang, Y., J. Szustakowski, et al. (2009). "Bioinformatics Analysis of Microarray Data." Cardiovascular Genomics, Methods in Molecular Biology **573**: 259-284.

## Vita

### **Sterling Wells Thomas**

#### A. Personal History

Home Address:  
2523 Pembroke Court  
Woodbridge, Virginia 22192  
(804) 677-3177  
sthomas@vcu.edu

Birth Date:  
May 9, 1975

Married – Charisse Thomas  
Children: Coulter (6)

Citizenship: USA

#### B. Educational History

Old Dominion University, Norfolk, Virginia  
BS: Health Sciences 2006 (Cum Laude)  
CT: Cytotechnology 2006 (Cum Laude)

Virginia Commonwealth University, Richmond, Virginia  
PhD Candidate: Integrated Life Sciences (Bioinformatics) Current

#### C. Technical Knowledge

Programming Languages: Java, m (Mathematica)  
Scripting Languages: Perl (bio-perl), python (bio-python), NetworkX (python), shell  
Operating Systems: Linux (RH, Fedora, Ubuntu, Suse), Unix (Solaris), OS X, Windows  
Server, Windows Desktop  
High Performance Computing: Sun Grid Engine, PBS Gridworks, Mathematica Grid  
Computing  
XML etc: SBML, XGMML

#### D. Research Positions

Senior Scientific Staff, Noblis Inc. 2010 – Current

- Develop integration strategies and data standards for a large national infectious disease program.
- Apply systems biology techniques, specifically biological network models, to address challenges in infectious disease research

Director of Bioinformatics, Calibrant Biosystems Inc. 2009 – Current

- I designed and deployed network analysis methods to Calibrant's existing biomarker discovery pipeline. My design was based on open source low cost solutions in addition to creating new software and scripts.
- Maintain and update existing software and IT hardware infrastructure.

Research Assistant to Danail Bonchev, PhD – Bioinformatics Computational Core Laboratory 2006 – Current

- I designed algorithms, supporting software and databases used to identify potential markers and associated pathways for adenocarcinoma of the lung and abnormal human embryonic stem cells.
- Identified new grant opportunities, developed projects and applications for submission.

Bioinformatics and Bioengineering Summer Institute Fellow, Virginia Commonwealth University Center for the Study of Biological Complexity 2005 – 2006 (NIH/NSF Funded)

- I created a mathematical model based on the interactions of Bcl2, BclX(L), Bax, and Cytochrome C.
- Member of the Bioinformatics Computational Core Laboratory (BCCL) – Virginia Commonwealth University
- Member of Human Diagnostics Research Laboratory – Old Dominion University

Research Assistant, Intercet Ltd. – Full Time position. 2001 – 2005

Duties:

- I assisted in creating a mathematical model for the cell cycle of cells involved in Adenocarcinoma of the colon.
- I assisted in reviewing a model for the dispersion of asbestos over grinder. The model included the diverse behavior of free and imbedded asbestos fibers and the ratio of buoyant fibers and fibers weighted by a foreign substance.
- I assisted in editing the 2004 version of the Emergency Response Procedure's drafted by the American Industrial Hygiene Association. I was responsible for reviews new compound levels and references.

## E. Business Positions

CEO/CIO/Board Member – Midwest Proteomics, Richmond, Virginia 2006-Current

- Midwest Proteomics is a small business founded to use modern signaling network based technologies in drug design.

- My responsibilities included developing and executing strategy for private and public investment, developing new markets by developing a position strategy and implementing the strategy through new bioinformatics products and services.
- Inventor Network Builder: Unsupervised signaling network discovery engine/plugin for cytoscape.

CSO/Cofounder – Canswers Inc. McLean, Virginia 1999-2003

- Canswers was a small business founded to develop and market software based diagnostic tools to CLIA approved pathology laboratories.
- I was responsible for the development of the software, market research for the business development team and presentation of our marketing strategy to board members and potential investors.
- Canwers raise approximately \$2M USD through strategic partnerships and private investment.

#### F. Awards and Honors

May 2008 – Recipient of Phi Kappa Phi Scholarship

May 2008 – Best Paper, IEEE International Conference on Biomedical engineering and Informatics

April 2008 - Recipient of the “Excellence Award in Molecular Medicine” for research presented at Molecular Medicine: Applying Current Technologies and Emerging Technologies Symposium

December 2008 – Invited to participate on the Virginia Biotechnology Park Leadership Council

May 2006 – Lindsay Rettie Research Award, Old Dominion University, Norfolk Virginia

April 2006 – Distinguished Student Research Award, EVMS – ODU - NSU Research Exposition, Norfolk Virginia

April 2006 – Honorable Mention American Association for Cancer Research Student Session, AACR National Meeting, Washington D.C.

#### G. Membership in Professional Associations

Phi Kappa Phi Honor Society

Golden Key Honor Society

American Association for Cancer Research  
Student member

Society for Industrial and Applied Mathematics

American Chemical Society  
Member of Washington Section  
Member of Biological Chemistry Division

#### H. Community Activities

Grant Writer, American Red Cross (Prince William County, VA)

#### I. Publications

Bonchev, D. Thomas, SW. Apte, A. Kier, L. Cellular Automata Modeling of Bimolecular Networks Dynamics. ACCEPTED 2009, Ref in Print.

Kuznetsov, V., Thomas, SW., Bonchev, D. Data-driven Networking Reveals 5-Gene Signature for Early Detection of Lung Cancer, Institute of Electrical and Electronics Engineers – Biomedical Engineering and Informatics 2008, pp.413-417

Thomas, SW. et al. A novel method to simulate complex networks of apoptosis in adenocarcinoma of the lung. Proc Amer Assoc Cancer Res 2006;47:1560.

Thomas, SW. Computers and Cancer. The World and I. March 2004 pp 132-139

Thomas RD. et al. Molecular Profiling and Computer modeling in Early Detection and Treatment of Cancer. Society of Toxicology Annual Meeting 2004.

#### J. Past Presentations

March 2008- Emerging Technology in Bio-IT., VCU Executive CIO symposium (Invited Speaker)

November 2007 – Mid-Atlantic Biotechnology Conference – Bethesda, Maryland

June 2007 – Summit on Systems Biology. Richmond VA

November 2006 – American Society of Cytopathology, Toronto

August 2005 – Cancer and the Bcl2 Family, VCU Bioinformatics and Bioengineering Summer Institute closing symposium, Richmond

April 3, 2006 – American Association for Cancer Research National Meeting  
Washington, D.C.

March 31, 2006 – Colonial Academic Alliance Undergraduate Research Conference  
James Madison University, Virginia

March 17, 2006 – National Association of Bioinformatics and Bioengineering Summer  
Institutes (Invited Speaker)  
Bethesda, Maryland

#### K. Patents

EP1198195 - Software Patent: Human Cancer Virtual Simulation System  
60/934,415 – Provisional Patent Application: Method of Evaluating Pharmacological  
Activity

#### L. Projects Underway

Bcl2 Family Proteins and Cancer: I've developed a protein network model that is simulated using cellular automata using data obtained using Affymetrix gene chips for Adenocarcinoma of the Lung. This technology is now being tested for as a possible method for lung cancer.

#### M. Professional Interests

My research interests include the molecular biology of carcinogenesis and signaling pathways. I'm specifically interested in the mechanics at the cellular and biochemical levels and the use of mathematical models for simulation.

#### N. References

David Lohr, Virginia Biosciences Development Center – (804) 828-7084  
Dennis Peffley, PhD, JD – Kronos Laboratories – (877) 576-6675  
Patricia Hentosh, ScD – Old Dominion University (757) 683-3611  
Dr. Danail Bonchev, PhD. – VCU CSBC (804) 827-0026