The Pennsylvania State University
The Graduate School
The Huck Institute of Life Sciences

A SYSTEMS BIOLOGY APPROACH FOR THE STUDY OF BIOMOLECULES
AND BIONETWORKS

A Dissertation in
Integrative Biosciences
by
Gaurav Bhardwaj

© 2010 Gaurav Bhardwaj

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

May 2010
The dissertation of Gaurav Bhardwaj was reviewed and approved* by the following:

**Randen L. Patterson**  
Assistant Professor of Biology  
Dissertation Advisor  
Chair of Committee

**John E. Carlson**  
Professor of Molecular Genetics

**Ross C. Hardison**  
T. Ming Chu Professor of Biochemistry and Molecular Biology

**Edward C. Holmes**  
Professor of Biology

**Damian van Rossum**  
Special Member  
Assistant Professor, Research

**Peter Hudson**  
Willaman Professor of Biology  
Director, Huck Institutes of the Life Sciences

*Signatures are on file in the Graduate School*
ABSTRACT

We propose a systems biology approach to integrate non-kinetic data from interacting biological reactions into informative Bionetwork-Boolean models. We developed a descriptive and predictive bionetwork model for phospholipase C-coupled calcium signaling pathways, built with non-kinetic experimental information. Boolean models generated from these data yield oscillatory activity patterns for both the endoplasmic reticulum resident inositol-1,4,5-trisphosphate receptor (IP$_3$R) and the plasma-membrane resident canonical transient receptor potential channel 3 (TRPC3). Furthermore, knock-out simulations of the IP$_3$R, TRPC3, and multiple other proteins recapitulate experimentally derived results. The potential of this approach can be observed by its ability to predict previously undescribed cellular phenotypes. Indeed, our cellular analysis of DANGER1α confirms the counter-intuitive predictions from our Boolean models in two highly relevant cellular models. Based on these results, we theorize that with sufficient legacy knowledge, Boolean networks provide a robust method for predictive-modeling of any biological system.

A limiting factor to this bionetwork-boolean approach is the lack of information regarding structural, functional and evolutionary characteristics of individual network components. In most cases, this lack of information arises from inability of conventional homology detection programs to measure homology in highly divergent datasets. Further, inability to resolve deep node relationships is a major factor that stymies evolutionary studies of highly divergent/rapidly evolving protein families. To resolve the shortcomings of conventional homology detection programs, we propose a computational approach towards resolving homology between highly divergent familial proteins using phylogenetic profiles. Indeed,
phylogenetic profiles have been demonstrated as a method for simultaneous measurements of structure, function, and evolution.

Herein, we describe a MSA independent method to infer evolutionary relationships, and use this method to study rapidly evolving (Mab21-containing DANGER superfamily), highly divergent (Retroelements) and convergent (Haloacid Dehalogenase) benchmark superfamilies. We also compare the results obtained from our method (PHYRN) with other MSA dependent methods and show that PHYRN provides better evolutionary history recapitulation, and provides more robust measurements at deep nodes. Further, PHYRN also provides quantitative measures that can aid in identifying outgroups and convergent evolutionary events. Using Retroelements (RT) as a benchmark superfamily, we show that this approach can be scaled up efficiently to study mega-phylogenies with thousands of sequences. Taken together with PHYRN’s adaptability to any protein family, this method can serve as a good tool in resolving ambiguities in evolutionary studies of rapidly evolving/highly divergent protein families.
# TABLE OF CONTENTS

LIST OF FIGURES vii  
LIST OF TABLES ix  
ACKNOWLEDGEMENTS x  
Chapter 1: Introduction 1  
  1.1 Background 1  
  1.2 Motivations and Objectives 3  

Chapter 2: Exploring Phospholipase C-coupled Ca\(^{2+}\) Signaling Networks Using Boolean Modeling 6  
  2.1 Background and Motives 6  
  2.2 Results 10  
  2.2.1 Constructing the Bionetwork and Generating Boolean Rules 10  
  2.2.2 Synchronous and Asynchronous Simulations of the Serotonin Bionetwork 13  
  2.2.3 Experimental Validation of Network Simulations 16  
  2.2.4 Experimental Validation of Network Predictions 18  
  2.3 Materials and Methods 20  
  2.4 Discussion 23  
  2.5 Future Work 26  

Chapter 3: PHYlogenetic Reconstruc\(t\)ionN (PHYRN) and Resolving Evolutionary Origins of Retroelements 28  
  3.1 Background and Motives 28  
  3.2 Results 30  
  3.2.1 PHYRN Pipeline 30  
  3.2.2 Enriching and Amplifying Informative PSSMs 32  
  3.2.3 PHYRN Scoring Scheme 32  
  3.2.4 Phylogenetic Reconstruction of the RT superfamily 34  
  3.2.5 Phylogenetic Reconstruction of Rose/Seq-Gen Simulations 39
### 3.3 Materials and Methods

Chapter 4: Resolving Deep-node Evolutionary Relationships between Rapidly Evolving Proteins

- **4.1 Background and Motives**
- **4.2 Results**
  - 4.2.1 Phylogenetic Reconstruction (PHYRN)
  - 4.2.2 Statistical Support and Outgroup Identification
- **4.3 Materials and Methods**
- **4.4 Discussion**

### 3.5 Discussion

Chapter 5: Resolving Convergent Evolution Using PHYRN

- **5.1 Background and Motives**
- **5.2 Results**
  - 5.2.1 Single Lineage Assumption
  - 5.2.2 Pearson Clustering and identifying number of lineages
  - 5.2.3 Distinct HAD lineages
- **5.3 Methods**
- **5.4 Discussion**

Chapter 6: Conclusions and Future Perspectives

REFERENCE LIST
LIST OF FIGURES

Figure 2-1: Bionetwork of Serotonin Induced Calcium Signaling Pathway 9
Figure 2-2: Flowchart of Bionetwork-Boolean Modeling Workflow 11
Figure 2-3: Gβ3 binds and regulates IP₃R activity 14
Figure 2-4: Boolean Modeling Predicts Wild-Type (WT) and Knock Out (KO) phenotypes 15
Figure 2-5: Boolean Modeling accurately predicts in vivo DANGER KO phenotype. 19
Figure 3-1: PHYRN Concept and Work Flow 30
Figure 3-2: Enrichment and Amplification of Signal Source PSSMs 31
Figure 3-3: Phylogenetic profile based measurements of evolutionary distance 33
Figure 3-4: Towards comprehensive phylogenies 34
Figure 3-5: High Throughput Scaling Up of Retroelements 36
Figure 3-6: Scaling Up phylogenies improves the resolution 38
Figure 3-7: PHYRN recapitulates ‘true evolutionary history’ better than MUSCLE in simulated protein families 39
Figure 3-8: Deep Node Recapitulation of ‘true evolutionary history’ in mega-phylogenies 42
Figure 3-9: Deep Node Recapitulation in Simulated Datasets 43
Figure 4-1: Phylogenetic tree of DANGER derived using PHYRN and Conserved Domain Database (CDD) PSSM library 51
Figure 4-2: Resolving DANGER superfamily using PHYRN 52
Figure 4-3: Resolving Deep Nodes in DANGER Superfamily 55
Figure 4-4: Resolving evolution of DANGER Superfamily 56
Figure 4-5: PHYRN and Outgroup Identification 57
Figure 4-6: Evolution of DANGER Superfamily 60
Figure 5-1: HAD Superfamily phylogenetic tree based on single lineage assumption 65
Figure 5-2: Resolving Convergent Evolution using PHYRN and Hierarchical Clustering 66
Figure 5-3: Hierarchical clustering of PHYRN data reveals independent lineages in HAD Superfamily 68
Figure 5-4: Unrooted Neighbor Joining (NJ) tree of cNII nucleotidase derived HAD lineage

Figure 5-5: Unrooted Neighbor Joining (NJ) tree of 8KDO derived HAD lineage

Figure 5-6: Unrooted Neighbor Joining (NJ) tree of SPSC derived HAD lineage

Figure 5-7: Unrooted Neighbor Joining (NJ) tree of PSP/PATPase derived HAD lineage
LIST OF TABLES

Table 2-1: Boolean Rules for updating each node in network .................................. 12
Table 2-2: Knock out Phenotypes Predicted by Boolean Modeling and Literature Supporting the
Predictions .................................................................................................................. 17
I owe my deepest gratitude to my adviser, Dr. Randen L. Patterson and my co-adviser, Dr. Damian B. van Rossum, for their constant support and guidance. They provided encouragement, sound advice, and lots of great ideas throughout this work. Their good teaching and good company have helped me clear hurdles in research with great ease. Their passion for science is truly inspirational. Their enthusiasm for research has motivated me to continuously strive and work towards becoming a better scientist.

I would like to thank my thesis committee members, Dr. John C. Carlson, Dr. Ross Hardison, Dr. Edward C. Holmes for their efforts and guidance in making this work scientifically more enriching. I would also like to thank Dr. Reka Albert and Dr. Michael Axtell for creative dialogue and guidance.

I am extremely grateful to my labmates, Dr. Kyung Dae Ko, Yoojin Hong, Teresa Killick, and Dr. Sree Chintapalli for their generous support at every stage of this work. It has been a great pleasure to work with such skillful and supportive colleagues. I am indebted to Jason Holmes, Zhenhai Zhang, Sunando Roy, Gue Su Chang, Andre Wallace, Christine P. Wells, Nick Hartranft, Lindsay Kline, Evan Smith, Matthew Gotlin and Anand Padmanbha for providing help at some key stages of this work.

I am thankful to my friends, Shruti, Kaushal, Debleena, Charumani, Amit, Mayur, Saumya, Aditi, Vivek and Kokonad for their constant help and support.

Finally, I would like to thank my family, to whom I owe everything.
Chapter 1

Introduction

1.1 Background

In principle, amino acid sequence of a protein should determine its structural, functional and evolutionary characteristics, but in reality, there are still no reliable methods to these qualities purely from primary amino acid sequences. Further, evolutionary measurements are hampered by extreme divergence in many protein families (1-3). Although, purely based on permutations there are enormous amount of possibilities for sequence and structure, for a protein of given length. But, in reality, all proteins are formed by combinatorial arrangement of a limited number of domains, and motifs that can fold rapidly and reproducibly.

The ‘protein problem’ arises due to inability of conventional homology detection algorithms to infer relationships in high divergence range (~ <25% sequence identity, ‘twilight zone’) (4). Most homology detection algorithms either fail to detect homology if sequence identity falls below this threshold, or in some cases lack statistical support for their inferences (3;5;6). However, only a small number of residues are required for protein to fold correctly or to function. Indeed, in some cases, proteins with ~8% sequence identity between them, can perform same function and/or fold similarly (7-9). This is because, structurally or functionally important residues are evolutionarily more conserved, as compared to other regions in a protein. From a computational point of view, this merging of ‘signal’ from residues that are important in Structure, Function, & Evolution (S, F & E) characteristics, with ‘noise’ from not so important residues, leads to dilution of measurements, and thus, less reliable homology detection.
An important question for computational biologist is how can we sift through these S, F & E data spaces for relevant signals and reduce noise? One approach can be to develop tools that study structure, function and evolution individually. Indeed, this is the approach that has been utilized most so far, and has led to the development of tools such as i) Rosetta(10), MODELLER(11), Mustang(12) etc. for inferring structures, ii) Pfam(13), SMART(14), CDD(15), Interproscan(16) etc. for studying function, and iii) Multiple Sequence Alignment tools such as K-align(17), MUSCLE(18;19), CLUSTAL(20), combined with phylogenetic clustering algorithms (Neighbor-Joining(21), Minimum Evolution(22) etc.) for studying evolution. Although all these methods are very competent at measuring individual S, F & E characteristics, these algorithms fail to detect statistically significant homology in ‘twilight zone’ of sequence similarity.

An alternative approach lies in utilizing the combined information content of Structure, Function and Evolutionary data spaces in a way that, i) functional and evolutionary measurements can quantitatively inform structural modeling to derive atomic resolution protein structures; ii) structural and functional attributes can help inform evolutionary measurements; and iii) structure and evolution measurements can help infer the new functions by providing information about regulatory/active sites. Thus in theory, a simultaneous measurement of S, F & E, would improve the sensitivity and scope of these measurements.

Further, as a computational biologist, understanding S, F & E of individual proteins is not enough; another major challenge is to put this information in context of interacting proteins. Along with simultaneous measurements of S, F & E, it is important to take this information to the next level of interacting ‘protein circuits’. To achieve this aim, we can use structure, function and evolution characteristics as a linking parameter and build bionetworks that describe complex
cellular interactions.

Bionetworks when taken together with various kinetic and non-kinetic modeling approaches, present a new philosophy of studying biological reactions (23-25). Unlike the reductionist approach of studying individual components of a biochemical reaction, bionetworks represent a holistic approach of treating cell as a system. Bionetworks have already been used to study gene expression patterns, immune responses, signaling mechanisms, survival signaling in cancerous cells etc (23-34). Bionetwork modeling is a descriptive and predictive tool, which can work in collaboration with lab experiments in an iterative manner (i.e. where lab results drive the design of bionetwork, and network predictions drive further lab experimentation). Recently, networks have also been shown to aid in evolutionary measurements (35). Although in nascent stages, these studies demonstrate that networks, as well as simultaneous measurements of S, F & E, can work in collaborative manner.

1.2 Motivations and Objectives

This thesis is driven by two major objectives; 1) development of Bionetwork-Boolean model to describe oscillatory behavior of cellular calcium signaling pathways, 2) development of an MSA independent tool for studying structure, function and evolution (S, F & E).

While Bionetwork-Boolean models have previously been defined for various steady state phenotypes such as gene segmentation patterns in *Drosophila* (36), Abscisic acid signaling (28), immune responses (29), survival signaling in cancer (37), we have developed a bionetwork-boolean model for an oscillatory signaling pathway in this thesis. In, chapter 2, we present the design and development of a bionetwork for Phospholipase C-mediated Ca$^{2+}$ signaling pathway from literature and experimentation results. Further, we show that even in the absence of kinetic
data and extensive computational cost, informative network models can be developed. We present a tool for calcium signaling community that can be easily adapted to any cell-specific system due to its modular nature.

In Chapter 2, we discuss the development and design of a novel tool for studying evolution that is based on philosophy of measuring Structure, Function and Evolution (S, F & E) simultaneously. PHYlogenetic Reconstruction (PHYRN) is a Multiple Sequence Alignment (MSA) independent tool that utilizes the power of phylogenetic profiles for the measurement of homology in the “twilight zone” of sequence similarity. To test the sensitivity and specificity of this tool, we present our results for various simulated datasets in Chapter 3, and show that PHYRN recapitulates ‘true evolutionary history’ better than current ab initio MSA based tools (MSA). Furthermore, we also discuss a phylogenetic study on retroelements (RT) derived using PHYRN. RT superfamily is highly divergent, and its history has remained unresolved even after decades of research(38;39). Using PHYRN, we present the most recent and comprehensive study of RT evolution to date. Using this benchmark dataset, we also show that PHYRN can easily be scaled-up to study mega-phylogenies with thousands of sequences without compromising on quality of the results. Since RT superfamily has implications in early origins of life, we show that this MSA-independent tool can help evolutionary biologists to resolve questions that were previously considered intractable.

Sometimes, Multiple Sequence Alignment (MSA) algorithms fail to detect homology in rapidly evolving protein families. DANGER (Differentiation and Neural Growth Evolve Rapidly) is a rapidly evolving superfamily of proteins, which has roles in cellular differentiation and calcium physiology(40-42). Previous phylogenetic studies on DANGER superfamily that were based on MSA based algorithms, failed to infer this phylogeny with enough statistical
significance. Especially, the support at deep nodes was lacking. Also, there were ambiguities in identification of outgroup. In Chapter 4, we show that PHYRN can be used to resolve rapidly evolving benchmark phylogenies, and it outperforms other MSA based tools (DiAlign and MUSCLE) in deep node statistical support.

Another major limitation with MSA based phylogenetic tools is their inability to resolve convergent evolution. In Chapter 5, we present evidence that PHYRN also provides measures that can aid in outgroup identification and detection of evolutionary convergence. Using a benchmark protein family of Haloacid Dehalogenase (HAD)(43), we demonstrate that PHYRN can be used to identify number of multiple lineages in a protein superfamily and thereafter resolve these independent lineages. Overall, we show that PHYRN can be used to efficiently resolve some major hurdles that stymie phylogenetic studies, and its ability to infer homology relationships in twilight zone of sequence divergence (<25% sequence identity) can widen the scope and range of these studies.
Chapter 2

Exploring Phospholipase C-coupled Ca\textsuperscript{2+} Signaling Networks Using Boolean Modeling

In this chapter, we describe the construction of a descriptive and predictive bionetwork model for phospholipase C-coupled calcium signaling pathways, built with non-kinetic experimental information. Our Boolean model yields oscillatory activity patterns for both the endoplasmic reticulum resident inositol-1,4,5-trisphosphate receptor (IP\textsubscript{3}R) and the plasma-membrane resident canonical transient receptor potential channel 3 (TRPC3) (44). To further test the validity of these results, we show that randomization of the Boolean operators ablates oscillatory pattern formation. We also show that our Boolean model recapitulates the wild-type (WT) and knock-out (KO) phenotypes correctly, and our simulations results are well supported by experimental results in literature.

2.1 Background and Motives

Calcium signaling networks are comprised of multiple nodes including: (i) receptors and ligands, (ii) soluble second messengers, and (iii) selective/non-selective ion-channels. Within these signal transduction pathways, ligand-binding to either a G-protein coupled receptor (GPCR) or tyrosine-kinase receptor (RTK) in the plasma-membrane initiates numerous short-term and long-term cellular signals. These signals initiate cellular programs responsible for growth, development, secretion, and apoptosis (45;46). Subsequent to ligand binding, GPCRs activate the heterotrimeric G-protein complex (G\textsubscript{\alpha} and G\textsubscript{\beta\gamma}) which stimulates phospholipase C-
beta (PLCβ) mediated catalysis of phosphatidylinositol (4,5) bisphosphate (PIP₂) to the second messengers inositol-1,4,5-trisphosphate (IP₃) and diacylglycerol (DAG). Similarly, ligand activation of RTKs results in the direct activation of phospholipase C-gamma (PLCγ), and its catalysis of PIP₂. In this way, IP₃ is generated and activates inositol-1,4,5-trisphosphate receptors (IP₃R), a large conductance Ca²⁺ channel on membranes of Ca²⁺ containing stores. Active IP₃R releases Ca²⁺ into the cytosol, as well as other cellular compartments (e.g. nucleus, mitochondria, etc) (45;47). Recent studies also demonstrate that IP₃Rs are regulated by a large suite of proteins which can both positively and negatively influence their calcium conductance. Further, IP₃R is also integral to Ca²⁺ entry mechanisms (45;48), and is a known regulator of transient receptor potential channels (TRPs), a superfamily of ion-channels involved in sensory perception (48;49).

The output of these signals provides Ca²⁺ oscillations/waves which spatially regulate protein function to accomplish diverse physiological/pathophysiological processes, such as sensory perception (50), cell development and growth (51;52), hearing (53;54), taste (55), and fertility (56) as well as multiple diseases including mucolipidosis (57;58) and polycystic kidney disease (48;57;59). Although our understanding of many of the core components is well-developed for certain steps within IP₃R containing Ca²⁺ networks, our understanding of how these signals are integrated with temporal and spatial precision is severely lacking.

The Ca²⁺ signaling network, even within a single cell, may contain hundreds if not thousands of different nodes (proteins, small molecules, lipids, ions, etc). Pioneers of network analyses have attempted to model Ca²⁺ signaling networks using kinetic data from biochemical experimentation (32;33;60-62). One such study by Bhalla and Iyengar (30) demonstrated that by building a library of small signaling network modules (e.g. PLC, MAP kinase, PKC, PLA₂, etc),
they could build a network which could simulate multiple functional outputs such as intracellular $\text{Ca}^{2+}$ concentration, kinase activity, phosphatase activity, etc. Further, a number of non-intuitive results were derived that could be recapitulated in the laboratory (30). Kinetic models built from experimental data can also successfully predict the effects of protein knock-down on cellular $\text{Ca}^{2+}$ dynamics (33). Maurya and Subramaniam demonstrated that protein knock-down does not have a linear correlation to loss of function (e.g. 18% knock-down of $\text{G} \beta \gamma$ equals a 50% loss of activity, which is also equivalent to a 60% knock-down of receptor protein)(32;33). Taken together, these results highlight the utility of kinetic network modeling.

Network analyses using quantitative kinetic data are powerful and predictive (32;33;63;64); however, the differential equations required to process these types of data require detailed information and enormous computational power. Furthermore, kinetic data for individual steps in the network is often lacking. Thus, generating larger scale network simulations is sometimes not feasible. Discrete network modeling (e.g. Boolean) may provide a solution to this problem, as it was demonstrated that in many networks kinetic parameters are not essential in describing the overall dynamics (27;65). Moreover, Boolean models characterize nodes with two qualitative states and only use three operators (AND, OR, NOT); thus, significantly reducing the computational power needed to run Boolean simulations as compared to kinetic network models (23;66).

In this chapter, we investigate whether Boolean modeling can be used to study PLC-coupled $\text{Ca}^{2+}$ signaling pathways. Our results demonstrate that (i) $\text{Ca}^{2+}$ networks can be constructed from legacy knowledge, (ii) experimental data from these multiple sources can be used to define network directionality and Boolean update rules for each node in network, and (iii) network modeling based on these Boolean rules provide a descriptive and predictive model
of PLC-mediated \( \text{Ca}^{2+} \) signaling. In addition, we experimentally validate a previously uncharacterized node (DANGER1a) in the network simulations. Taken together, we suggest that Boolean modeling provides an effective method for creating receptor specific and/or cell-specific signaling network models.

Figure 2-1: Bionetwork of Serotonin Induced Calcium Signaling Pathway
Calcium channels have been shown in green color, calcium release and entry are blue, and all other signaling components have been shown in yellow. Solid arrows represent positive interaction between nodes, and dashed arrows represent inhibition. "*" denotes hydrolysis of phosphatidylinositol bisphosphate.

Full names of abbreviations used in the network are: 5-HTR, 5-hydroxytryptamine receptor; PLC\( \beta/\gamma \), Phospholipase C Beta/Gamma; PIP\( _2 \), phosphatidylinositol bisphosphate; DAG, Diacylglycerol; IP\( _3 \), Inositol trisphosphate; PKA, Protein Kinase A; PKC, Protein kinase C; SERCA, Sarco/Endoplasmic Reticulum \( \text{Ca}^{2+} \)-ATPase; IP\( _3 \)R, Inositol 1,4,5-Trisphosphate Receptor; TRPC3, Transient receptor potential cation channel, subfamily C, member 3; CaBP1, Calcium Binding Protein1; ATP\( _{cyt} \), cytoplasmic ATP in vicinity to IP\( _3 \)R channels; CAM, Calmodulin; CamkII, \( \text{Ca}^{2+} \)/calmodulin-dependent protein kinase II
2.2 Results

2.2.1 Constructing the Bionetwork and Generating Boolean Rules

We chose to build a model for PLC-coupled Ca\(^{2+}\) signaling pathways due to the significant literature base for this receptor (67). Nodes within the network were chosen based on two criteria: (1) they have known activity(s) within these pathways and (2) they are known to have causal interactions, and not just associative interactions with other nodes in the network (see Methods for a complete description). Figure 2-1 depicts the bionetwork generated using this scheme. Although this is a subset of the known nodes/integrators for phospholipase C-mediated signaling, bionetworks are often resilient, even in the absence of a significant number of nodes (28). In addition, nodes were chosen that have, in general, considerable experimental support(45;47). Furthermore, as all GPCRs and RTKs flow through either PLC\(\beta\) or PLC\(\gamma\), our network results should be applicable to numerous receptor types (e.g. muscarinic, serotonergic, B-cell receptors, etc). From this topology, we observe that the IP\(_3\)R and the canonical TRP channel 3 (TRPC3) form hubs (highly-connected nodes) within the network. This implies that these channels are highly regulated, which is reasonable as both of these proteins flux the signaling ion, Ca\(^{2+}\).

After constructing the network, we defined our Boolean rules for updating each node in terms of logic operators (AND, OR, NOT) (Figure 2-2). In this method, nodes can have only two states, ON or OFF (see Table 1). ON can represent states such as ACTIVE (e.g. kinases, phosphatases, lipases etc.), OPEN (e.g. channels) and/or BINDING (e.g protein-protein or protein-lipid interactions). Conversely, OFF can represent INACTIVE, CLOSED and/or NON-BINDING states.
Bionetwork is developed based on current knowledge and rules for updating each node defined in form of logical operators (AND/OR/NOT). For synchronous modeling, each node is updated at each time step. For asynchronous modeling, random number of nodes are updated at each time step. Boolean states of each node in each time step is noted down in each update method and analyzed for attracting behavior. (see Methods for details)

To define the Boolean update rules for each node, directionality information between nodes, as provided by legacy knowledge, is converted to logical operator driven statements. Thus, the future state of each node is dependent upon the present state of its interacting nodes. We created these logic operators for all nodes in the network using either legacy knowledge or our own experimental data in defining rules (see Table 1 for rules used). For example, we used our own experimental results to define update rules for Heterotrimeric G-protein β3 (Gβ3). Studies from Zeng et al. (68) demonstrate that heterotrimeric G-proteins have the ability to stimulate IP₃R activity directly, although direct binding has not been determined. We tested the hypothesis that Gβ3 and IP₃R bind directly. Yeast two-hybrid analysis reveals that the third and fourth WD40 repeats of Gβ3 are critical for binding (Figure 2-3a). We next examined the direct binding of the IP₃R to full-length Gβ3 or a mutant Gβ3 containing deletions in the third and fourth WD40 repeats (aa166–207) (Figure 2-3b). Whereas full-length Gβ3 binds to IP₃R, binding is abolished by deletion of the two WD40 repeats.
To explore an influence of Gβ3 on IP₃R function, we first evaluated the influences of 50 nM Gβ3 on [³H]IP₃ binding to IP₃R in rat brain membranes (Figure 2-3c). Gβ3 increases the potency of unlabeled IP₃ in competing for [³H]IP₃ binding ≈3 fold. We next evaluated the influence of 50 nM Gβ3 on IP₃-induced Ca²⁺ release from microsomal membrane preparations of COS7 cells (Figure 2-3d). We do not observe activation of IP₃R by Gβ3 alone, even with high concentration of Gβ3 (10 mM) (data not shown). These observations suggest that Gβ3 acts by enhancing the affinity of IP₃ for IP₃R rather than by altering the channel itself, which would lead to a greater maximal release. Taken together, the activity results are in accord with Zeng et al. although we do not observe direct activation of the IP₃R by Gβ3 in the absence of IP₃.

<table>
<thead>
<tr>
<th>Node</th>
<th>Boolean Regulatory Rule</th>
</tr>
</thead>
<tbody>
<tr>
<td>HTR</td>
<td>HTR’ = serotonin</td>
</tr>
<tr>
<td>PLCβ</td>
<td>PLCβ’ = HTR</td>
</tr>
<tr>
<td>cAMP</td>
<td>cAMP’ = HTR</td>
</tr>
<tr>
<td>DAG</td>
<td>DAG’ = PIP₃ and (PLCβ or PLCγ)</td>
</tr>
<tr>
<td>PLCγ</td>
<td>PLCγ’ = HTR</td>
</tr>
<tr>
<td>IP₃</td>
<td>IP₃’ = PIP₃ and (PLCβ or PLCγ)</td>
</tr>
<tr>
<td>PKC</td>
<td>PKC’ = {DAG and (Ca²⁺ or Ca²⁺⁺ ER(Low) or Ca²⁺⁺ ER(High))}</td>
</tr>
<tr>
<td>TRPC3</td>
<td>TRPC3’ = {(not HOMER) and (IP₃R Low and PLCγ and CAM)} or (DAG and not PKC)</td>
</tr>
<tr>
<td>HOMER</td>
<td>HOMER’ = not IP₃R Low</td>
</tr>
<tr>
<td>Ca²⁺⁺ ER(High)</td>
<td>Ca²⁺⁺ ER(High) = IP₃R High</td>
</tr>
<tr>
<td>CAM</td>
<td>CAM’ = (Ca²⁺ or Ca²⁺⁺ ER(Low) or Ca²⁺⁺ ER(High))</td>
</tr>
<tr>
<td>SERCA</td>
<td>SERCA’ = (Ca²⁺ or Ca²⁺⁺ ER(Low) or Ca²⁺⁺ ER(High))</td>
</tr>
<tr>
<td>CaBP1</td>
<td>CaBP1’ = Ca²⁺⁺ ER(High) and DANGER</td>
</tr>
<tr>
<td>ATP_cyt</td>
<td>ATP_cyt’ = not SERCA</td>
</tr>
<tr>
<td>CamKII</td>
<td>CamKII’ = CAM</td>
</tr>
<tr>
<td>IP₃R_Low</td>
<td>IP₃R_Low’ = ((IP₃ and DANGER and (PKC or PKA or CamKII or Gβ3)) and not (IRBIT or CaBP1 or ATP_cyt))</td>
</tr>
<tr>
<td>IP₃R_High</td>
<td>IP₃R_High’ = ((IP₃ and DANGER and PKC and PKA and Gβ3)) and not (IRBIT or CaBP1 or ATP_cyt)</td>
</tr>
<tr>
<td>DANGER</td>
<td>DANGER’ = HTR</td>
</tr>
<tr>
<td>Gβ3</td>
<td>Gβ3’ = HTR</td>
</tr>
<tr>
<td>IRBIT</td>
<td>IRBIT’ = CamKII</td>
</tr>
<tr>
<td>Ca²⁺</td>
<td>Ca²⁺’ = TRPC3</td>
</tr>
</tbody>
</table>

Table 2-1: Boolean Rules for updating each node in networks
Further, we generated primary B-cell lines from human patients which were either WT or homozygous for the C825T polymorphism (69). This is a naturally occurring polymorphism, which results in a frameshift mutation leading to the deletion of amino acids 166-207 in Gβ3. In these cells we measured Ca\(^{2+}\) release in response to B-cell receptor stimulation with anti-IgM which stimulates the RTK B-cell receptor(70-72). The area under the curve from these measurements were quantified from numerous replicates and compared (see Methods for complete description). Although we do observe that some of these preparations have alterations in Ca\(^{2+}\) signaling, our results from multiple genetic backgrounds provided no statistical difference between WT and mutant preparations (Figure 2-3e).

Based on these experimental findings we defined the Boolean rule for the Gβ3 node (Figure 2-1 and Table 1). Specifically, in our rules Gβ3 enhances the activity of the IP\(_3\)R although it is non-essential. Indeed, the IP\(_3\)R is expressed in all cells(45;73-75) whereas Gβ3 is not (76).

2.2.2 Synchronous and Asynchronous Simulations of the Serotonin Bionetwork

Following, we performed two types of Boolean modeling. Synchronous Boolean modeling assumes that the time required for each biological reaction is comparable and thus a unit time-step can be defined (Figure 2-2). Therefore, when performing synchronous modeling we simultaneously update all node states according to the rules (Table 1) at each time-step. Conversely, asynchronous Boolean modeling assumes that each biological reaction has distinct durations. Since there is not enough quantitative information available to estimate the reaction durations, we randomly shuffle the order of nodes at each time-step (Figure 2-2). Thus, this approach samples differential time intervals for each biological reaction in network. The results from these simulations were assayed for their attracting behavior (i.e. repetitive behavior).
Figure 2-3: Gβ3 binds and regulates IP₃R activity

a) Yeast two-hybrid analysis shows that Gβ3 binds to IP₃R through its 3rd and 4th WD repeats (aa 166-207). Fragments of Gβ3 lacking in region 166-207 do not show any binding.

b) Pull down experiments in HEK293 lysates with GST tagged full length and mutated Gβ3 (Δ166-207) were performed. Full length Gβ3 shows a direct binding to IP₃R while there is no binding between mutant Gβ3 and IP₃R.

c) In [³H]IP₃ treated Rat brain membranes, Gβ3 (50nM) increased IP₃ binding to IP₃R.

d) In microsomal membrane of COS7 cells, Gβ3 (50 nM) leads to maximum calcium release at 30 nM IP₃.

e) Quantification of Ca²⁺ mobilization in human B cells lines from either Wild Type (WT) or C825T homozygous mutants stimulated with 10 mM anti-human IgM. Area under individual calcium traces was measured and it shows that there is no statistical difference in Ca²⁺ release and entry between WT and C825T mutants (Area under the curve, n= average 396 cells from 16 replicates across 4 different cell lines for WT and n=average 220 cells from 10 replicates across 3 different cell lines for C825T mutants, Error Bars: Std. Dev.)
Figure 2-4: Boolean Modeling Predicts Wild-Type (WT) and Knock Out (KO) phenotypes
A) Plots of Boolean state vs. Time steps from synchronous modeling recapitulate the oscillatory pattern of IP₃R and TRPC3 in Wild Type cells. Both the channels show a distinct oscillatory pattern with fixed time periods. B) Boolean modeling with IP₃R node knocked out (Control) shows that IP₃R Boolean state remains fixed at 0 (Off) and TRPC3 oscillatory pattern changes predicting a reduced Ca²⁺ entry. C) Asynchronous update method replicates the oscillatory pattern of IP₃R and TRPC3 but random update of nodes leads to varying time periods. D) Rewiring of AND & OR in logical rules fails to provide oscillations in the model. E) Boolean modeling of Ca²⁺ network after knocking out Gβ3 node show no change in oscillatory pattern of IP₃R and TRPC3 Boolean states as compared to Wild Type (WT)

Results from our simulations can be plotted as a function of the Boolean state versus the time-step. Figure 2-4A depicts the behavior for the IP₃R and TRPC3 nodes in the network. We observe that both proteins have oscillatory behavior, in agreement with experimental data (77). Interestingly, these proteins display different oscillatory patterns in terms of their time-steps. As a control, we knocked-out the IP₃R from the bionetwork and observe that the IP₃R has no activity; interestingly, TRPC3 is still oscillatory, but has an altered pattern of oscillations (Figure 2-4B). This may be expected as TRPC3 can be activated in an IP₃R-independent manner, although this has often been described in over-expression experiments (49). In our asynchronous
model (Figure 2-4C), we still observe an oscillatory pattern, although both IP_3R and TRPC3 oscillatory patterns no longer exhibit a fixed time period. This demonstrates that this network is robust in generating oscillatory patterns. Moreover, since we are trying to find a correlation between change in oscillatory time periods and change in cellular Ca^{2+} levels, asynchronous Boolean modeling poses a major limitation. Specifically, asynchronous modeling alters the rate at which nodes are updated; thus, the time-periods measurements between oscillations would be corrupted. In order to characterize the oscillations of the system we need a non-random model; therefore, for the purposes of this manuscript we chose to explore synchronous models. Importantly, in control experiments, we randomized the Boolean rules for all nodes by changing all ANDs to OR and vice versa, and observe no oscillatory patterns in our synchronous models (Figure 2-4D). This data confirms the oscillatory patterns observed under both synchronous and asynchronous simulations are not random in nature.

2.2.3 Experimental Validation of Network Simulations

To determine whether our synchronous models are informative, we knocked-out all of the nodes in the network one at a time and recorded the results (Table 2). Several of these simulations are corroborated by experimental literature. For example, simulated deletion of Homer results in increased Ca^{2+} entry through TRPC3 without altering Ca^{2+} release through IP_3Rs, as is observed experimentally (78;79). Another example is IRBIT, which is an inhibitor of the IP_3R. Its absence in the network predicts increased Ca^{2+} release and entry, in accord with the studies of Ando et al (80;81). Knock-out of PKC from our network predicts that Ca^{2+} release is abolished, while Ca^{2+} entry is substantially enhanced. Indeed, the work of Venkatachalam et al. clearly demonstrates that pharmacological inhibition of PKC validates our prediction (49).
<table>
<thead>
<tr>
<th>Knock Node(s)</th>
<th>Oscillation Pattern</th>
<th>Predicted Phenotype</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>DANGER</td>
<td>1(ON)-6(OFF)</td>
<td>IP$_3$R remain open for shorter interval and TRPC3 oscillation change to single type oscillations</td>
<td>Experimentally validated</td>
</tr>
<tr>
<td>IP$_3$R</td>
<td>0</td>
<td>3(ON)-3(OFF)</td>
<td>Calcium release stops and TRP channels show only DAG mediated/Basal oscillations.</td>
</tr>
<tr>
<td>GB3</td>
<td>2(ON)-6(OFF)</td>
<td>3(ON)-3(OFF)-1(ON)-1(OFF)</td>
<td>No change as compared to wild type</td>
</tr>
<tr>
<td>CamKII</td>
<td>3(ON)-5(OFF)</td>
<td>5(ON)-3(OFF)</td>
<td>Calcium release and entry both increase as channel remain open for longer time and closing time decreases.</td>
</tr>
<tr>
<td>PKC</td>
<td>0</td>
<td>1</td>
<td>Calcium release goes off completely and TRP channels remain constantly ‘ON’</td>
</tr>
<tr>
<td>ATP$_{c31}$</td>
<td>3(ON)-6(OFF)</td>
<td>3(ON)-3(OFF)-1(ON)-1(OFF)-2(ON)</td>
<td>Calcium release increases, as IP$_3$R remain open for longer times. TRPC also remain open longer as compared to wild type.</td>
</tr>
<tr>
<td>CaBP1</td>
<td>2(ON)-6(OFF)</td>
<td>3(ON)-3(OFF)-1(ON)-1(OFF)</td>
<td>No change as compared to wild type</td>
</tr>
<tr>
<td>IP3</td>
<td>0</td>
<td>3(ON)-3(OFF)</td>
<td>Same as IP$_3$R KO</td>
</tr>
<tr>
<td>SERCA</td>
<td>3(ON)-6(OFF)</td>
<td>3(ON)-3(OFF)-1(ON)-1(OFF)-2(ON)</td>
<td>Similar to ATP-vicinity KO</td>
</tr>
<tr>
<td>DAG</td>
<td>4(ON)-6(OFF)</td>
<td>2(ON)-8(OFF)</td>
<td>Calcium release affected, as channel remain open for longer time but there is no high concentration release. Entry decreases, as TRPC remains closed for longer times.</td>
</tr>
<tr>
<td>PKA, cAMP</td>
<td>2(ON)-6(OFF)</td>
<td>3(ON)-3(OFF)-1(ON)-1(OFF)</td>
<td>Calcium release decreases as channel remain open for same time as wild type but no high concentration release events. Entry remains normal.</td>
</tr>
<tr>
<td>Ligand, Receptor, PLC G, PLC B</td>
<td>0</td>
<td>0</td>
<td>No oscillations in network as it remains all ‘OFF’</td>
</tr>
<tr>
<td>IRBIT</td>
<td>3(ON)-5(OFF)</td>
<td>5(ON)-3(OFF)</td>
<td>Release and entry both increase</td>
</tr>
<tr>
<td>Calmodulin</td>
<td>3(ON)-4(OFF)</td>
<td>3(ON)-4(OFF)</td>
<td>Release increases but entry decreases</td>
</tr>
<tr>
<td>HOMER</td>
<td>2(ON)-6(OFF)</td>
<td>5(ON)-3(OFF)</td>
<td>Release remains unaffected while calcium entry increases</td>
</tr>
</tbody>
</table>
Network simulations with Gβ3 knocked-out from the network, exhibit no change in calcium mobilization (Figure 2-4E). Based on in vitro data, one might predict that removal of Gβ3 from the network would result in a decrease in Ca\(^{2+}\) release. Although, our Boolean update rule for Gβ3 was based on in vitro data, but when taken together with essentiality information, which was derived from differential and compensatory cellular expression, the rule recapitulates the in vivo phenotype (Figure 2-3e).

2.2.4 Experimental Validation of Network Predictions

To extend on these findings, we examined the role of DANGER1α; a member of the newly-discovered DANGER developmental superfamily (41). This protein was discovered through a yeast-2-hybrid screen with the IP\(_3\)R, and was demonstrated to regulate IP\(_3\)R activity and neuronal development (41;42;100). Specifically, DANGER1α decreases the open-probability of the IP\(_3\)R at high Ca\(^{2+}\)-concentrations (>300 nM). This in vitro data was used to construct the Boolean rules for DANGER1α. Specifically, we defined two different activation states for IP\(_3\)R release, HIGH and LOW. Under HIGH release conditions, DANGER1α acts as an inhibitor to IP\(_3\)R activity, while, under LOW release conditions, it plays a non-essential, non-inhibitory role. As DANGER1α is inhibitor in our in vitro analyses, one would expect that knock-out of DANGER1α would result in increased Ca\(^{2+}\) release. However, as shown in Figure 2-5a, our dynamic model predicts that knock-out of DANGER1α would result in an overall decrease of Ca\(^{2+}\) mobilization.
Figure 2-5: Boolean Modeling accurately predicts in vivo DANGER KO phenotype.

a) Boolean Modeling of our network with DANGER1a node knocked out predicts that an altered oscillatory pattern and a reduced Ca$^{2+}$ mobilization. b) (top) Southern analysis of embryonic stem cells generated by Ozgene©. Cells from lane A57 were used to make DANGER1A mice. (bottom) PCR analysis for DANGER1a transcript in Wild Type (WT), Heterozygous (HET) and DANGER1a knock out (KO) mice. c) Quantification of Ca$^{2+}$ mobilization in spinal cord neurons from WT, HET and DANGER1a KO mice stimulated with 1 mM serotonin (% of WT, area under curve, n = total number of cells from 3 independent experiments, Error bars: Std dev) d) Representative calcium traces from one complete imaging experiment.

To reconcile these results we turned to DANGER1a knock-out mice. Shown in Figure 2-5b, the Cre/Flox DANGER1a knock-out mouse was properly constructed as shown by southern analysis and PCR (Figure 2-5b, see Methods for complete description). From embryonic mice, we cultured primary spinal cord neurons and measured intracellular Ca$^{2+}$ transients in response to serotonin, using the fluorescent Ca$^{2+}$ indicator, Fura-2AM (see Methods). These experiments reveal that spinal cord neurons from mice, either heterozygous or homozygous for DANGER1a knock-out, have a decrease in overall Ca$^{2+}$ mobilization when stimulated with serotonin (Figure
Surprisingly, these results recapitulate the predictions from our Boolean model, not the intuitive expectations that could be projected from the in vitro data alone.

2.3 Materials and Methods

Bionetwork and Boolean Modeling

Nodes and edges in bionetwork were chosen after a comprehensive literature search for components that are part of Serotonin Ca$^{2+}$ pathway and/or are known to interact with any of the other components of the network. Some of the interactions are very well established based on experimental data but in cases where conflicting results existed, nodes and edges were defined based on relevance and accuracy of experimental support provided in literature. Further, only nodes that have causal effects on the system were included. Associative interactions, where directionality information was not present in current literature could not be included. For a list of references used in generating the network, see Supplemental References.

The model in Figure 2-1 was drawn with yED graph editor (http://www.yworks.com/en/index.html). The Boolean modeling was performed with a custom Python script (http://www.python.org). Update rules for each node were defined based on legacy knowledge and experimental data obtained prior to this modeling. For synchronous modeling, all the nodes were updated for their respective rules at all the time steps. For asynchronous modeling, a random update order was selected at each step from N! possible permutations (N = no. of nodes). As shown in Chaves et al and Li et al (28;101), this method provides a random differential time to each reaction/interaction event. Boolean states of each node at each step (for 50 time steps in case of synchronous) were recorded and analyzed for repeating patterns (Attracting behavior).
For random rewiring of AND and OR in logical rules, all the ANDs in the logical rules were changed to OR and vice versa. Synchronous update method was used after rewiring to get an equivalent of randomization in this network study.

**Calcium Imaging**

Calcium imaging was performed as previously described (102). Briefly, neurons were loaded with 1mM Fura-2-AM for 5 minutes. B-cells were loaded with 2mM Fura-2-AM for 25 minutes. Neuronal experiments were conducted at 37°C and Kreb’s solution was supplemented with 2mM CaCl₂ and 1mM glycine. B-cell experiments were conducted at room temperature as previously described (70). HEK-293 and A7r5 cells were imaged exactly as in Patterson et al (70).

**Cell Culture**

Cells were cultured using the protocol from Alldred et al (103). Briefly, spinal cord neurons were generated from DANGER1A deficient embryonic day 14.5 embryos generated by crossing of DANGER1A⁺⁻ mice on a 129SvJ inbred background. Spinal cords were collected in PBS containing 5.5 mM glucose and treated with papain (0.5 mg/ml) and DNase I (10 µg/ml) (both from Sigma, St. Louis, MO) in PBS containing 1 mg/ml bovine serum albumin (Fraction V, Sigma) and 10 mM glucose for 15 min at room temperature. The cells were triturated with a fire-polished Pasteur pipette and plated on poly-L-lysine-coated glass coverslips (22 x 22 mm) at 4 x 10⁴ cells per square centimeter in modified Eagle medium (MEM) (Invitrogen) containing 10% v/v fetal bovine serum (FBS) (Invitrogen) in an atmosphere of 10% CO₂. After 60 min, the medium was replaced with fresh MEM containing 10% v/v FBS. The genotype of cultures was determined using PCR as follows. Tail biopsies (3 mm) were incubated for 30 min at 55°C in 25 µl lysis buffer (200 mM NaCl, 5 mM EDTA, 0.2% SDS, 100 mM Tris-HCl, pH 8.5), and 1% of
the supernatant was used for PCR using standard conditions with the primer 5'-ACAGAACCAATCGGTGTG-3' combined with either 5'-CTCGTTTCTCGGGAATCGT-3' to amplify the wild-type DANGER1A locus or the primer 5'-GCGCTTCCAAGGCTGTGAA-3' to amplify the mutant DANGER1A locus. The 24-hr-old cultures that exhibited the desired genotypes were turned upside down onto a glial feeder layer in a Petri dish containing Neurobasal-A supplemented with B27 (Invitrogen), in an atmosphere of 10% CO₂. Feeder cells were prepared from cortices of newborn rat pups as described (Banker and Goslin, 1998). Neuron cultures were maintained without medium change for 18 d in vitro (DIV) and then transferred into new Petri dishes containing Neurobasal A/B27 supplemented with 1 µM 6-cyano-7-nitroquinoxaline-2,3-dione (CNQX) and 100 µM 2-amino-5-phosphonovaleric acid (Sigma), with the cells facing up. Neurons were processed for calcium imaging at 10 DIV.

**Microsomal Assays**

Microsomal assays were performed as previously described (104;105).

**In vitro Binding assays**

*In vitro* assays were performed exactly as previously described (104;106)

**Yeast-2-Hybrid Assays**

Yeast-2-Hybrid assays were performed exactly as previously described (104;106).

**Materials**

Anti-myc antibody, 5-HT, Anti-human IgM, carbachol, and all buffer components were obtained from Sigma-Aldrich (St. Louis, MO). Yeast-2-Hybrid reagents were obtained from CLONETECH (Palo Alto, CA). All cell culture reagents and Fura-2-AM were obtained from Invitrogen. ³²P and ⁴⁵Ca²⁺ were obtained from Amersham Biosciences. DANGER1a knock-out mice were generated by Ozgene© (Australia).
2.4 Discussion

In this study, we successfully used qualitative data from various experimental studies to develop a Boolean dynamic model for the PLC-coupled Ca\textsuperscript{2+} signaling pathway. A number of implications can be derived from this work: (i) descriptive and predictive models can be built in the absence of kinetic data for multiple receptors (muscarinic, serotonergic, B-cell receptor), (ii) these models can be used to predict mutant/knock-out phenotypes which can be validated experimentally, and (iii) in some cases, these models can be used to anticipate non-intuitive biological effects. From these preliminary studies, we propose that refining Boolean models of Ca\textsuperscript{2+} signaling will likely provide a rich resource for experimental biologists.

By gathering and organizing current knowledge about PLC-coupled Ca\textsuperscript{2+} signaling into a bionetwork, the biological information is organized into local structures. This allows the interactions between components of the pathway to become more accessible and easier to understand. Indeed, network topology can identify the various feedback and feed-forward loops, as well as the most connected nodes (i.e. hubs) in the system. Feedback and feed-forward loops represent the regulatory control in the signaling pathway, while hubs signify the most regulated components in the network. In our bionetwork, both the IP\textsubscript{3}R and TRPC3 emerge as hubs. This is not surprising as both IP\textsubscript{3}R and TRPC3 are heavily regulated calcium channels (45;48;107). We also observe that our simulations accord with many observations from the literature (Table 2). Importantly, our bionetwork also approximates the oscillatory behavior observed during cellular Ca\textsuperscript{2+} signaling, as well as the experimental results obtained for Gβ3 and DANGER1a in serotonin-induced signaling pathways (Figure 2-4 and 2-5). Although similar types of results have been obtained from differential equation based models of calcium signaling (32;33), we
propose that Boolean modeling represents an independent and powerful tool for examining signaling pathways when detailed kinetic data does not exist. Due to the need for extensive computational power, kinetic models are often not feasible; thus Boolean bionetwork modeling has great potential for measuring \( Ca^{2+} \) signaling pathways that lack kinetic data.

Of particular interest is the observation that our Boolean model creates oscillatory patterns for both IP\(_3\)R and TRPC3 channel opening in wild type and various mutant/knock-out conditions (Figure 2-4A-C,E). This oscillatory pattern is not random and/or an artifact of our approach; when the logical rules for ‘AND’/‘OR’ are rewired, the network fails to generate any oscillations (Figure 2-4D). The accuracy of our network is further evident from comparisons with experimental results. Changes in oscillatory frequency of calcium channel activity observed in our Boolean analysis correlate with experimental measurements for changes in \( Ca^{2+} \) mobilization in various mutant/knock-out cell types (Figure 2-3,2-4, 2-5 and Table 2). The ability of this bionetwork to consistently replicate WT and mutant phenotypes suggests that the data used to construct the rules is accurate and that general trends in \( Ca^{2+} \) mobilization can be modeled.

Another interesting finding is that in some cases when \textit{in vitro} and \textit{in vivo} results conflict, Boolean modeling can predict results for the \textit{in vivo} phenotype using \textit{in vitro} rules. In our study, network modeling predicted phenotypes that were counter-intuitive to \textit{in vitro} results for DANGER1a. However, biochemical experiments are consistent with the predicted network phenotype. The ability of networks to predict counter-intuitive \textit{in vivo} phenotypes for mutant/knock-out cells may not always hold true, as we have observed it only for one node in our network. Nevertheless, this idea has appeal. It is not unprecedented that biochemical interactions under cellular conditions, and in presence of other interacting biochemical
components can behave differently from *in vitro* interactions. Given this, modeling studies should also use a holistic approach, and study cell as a system and not just as reductionist version of individual interactions. Our network illustrates that this systems biological perspective for studying cells as a network of cross-interacting biological reactions, rather than as individual interactions, may provide a more accurate depiction of cellular processes.

To date, Boolean analysis has been utilized to study gene regulation patterns (108-110), Abscisic acid signaling pathway (28), immune responses (29), cholesterol biosynthesis (31), and survival signaling in T-cells (111). Taken together with our analysis, we theorize that Boolean network can be used to study any biological pathways, providing clues for further experimental work. After experimental verification, this new information can feed back into the network, thereby improving the model.

A limiting factor to the Boolean modeling approach is that it depends on information obtained from different sources. Thus, the quality and accuracy of information used to define logical rules can significantly impact the network results. Likewise bionetworks can be incomplete, as in the case of this study. There are many other calcium networks described (but not modeled) which are bigger and more complex (e.g. Berridge model (112)), but in this paper we have focused only on a smaller network that forms backbone of all GPCR and RTK calcium networks. However, novel nodes and edges will be added in future, as our understanding of these pathways gets better. It is hard to say how much of the results obtained from this study might change after addition of novel nodes and edges in the future. Nonetheless, these experimental data can be used to iteratively define the rules for our Boolean modeling, with the resulting predictions paving the way for further experimental research.
The network presented here can be accessed online, and codes for creating Boolean network models are available at BooleanNet (http://code.google.com/p/booleannet/). We encourage the Ca^{2+}-signaling community to provide additional rules, which have been experimentally validated, to be incorporated into these networks. Currently, our bionetwork is very specific; however, the modular nature of Boolean networks makes it very easy to modify and/or increase the network complexity. Thus, we envision that development of this resource will likely lead to the construction of cell-specific Ca^{2+} signaling networks. Further, we believe that this wiki approach will allow for an accurate and detailed expansion of the Ca^{2+} signaling network, which in turn will improve its predictive power.

2.5 Future Work

In the future, our work will also focus on developing a more complex model by including nodes and connections already described in Berridge model (113) and by including other phenotypic characteristics into the bionetwork. For example, DANGER1a, TRPCs, and IP_{3}R are also involved in neural growth (41;48;114-116). Therefore, a neuronal development network can be integrated into the Ca^{2+} signaling network. In theory, this should allow for the simultaneous decoding of these two highly interconnected networks. Since calcium mobilization is not an ON/OFF process, we will also explore higher order discrete modeling approaches towards better simulations. We will also test hybrids of discrete/continuous models to better understand calcium mobilization; particularly in cases where some kinetic data is available. In addition to modeling signaling pathways, these networks can aid in the identification of drug targets based on network connectivity (i.e. some nodes serve as better drug targets than others (117;118)). We propose
that in the long-term, our refined Boolean modeling may lead to increased capacity to create predictive stage models likely to identify targets for therapeutic intervention.
Chapter 3

PHYlogenetic Reconstruction (PHYRN) and Resolving Evolutionary Origins of Retroelements

The inability to resolve deep node relationships is a major factor that stymies evolutionary studies of highly divergent/rapidly evolving protein families. In this chapter, we propose a Multiple Sequence Alignment (MSA) independent method to infer evolutionary relationships, and use this method to study biological protein families (retroelements) and simulated datasets (ROSE and Seq-Gen). We also describe new approach to enrich and amplify informative phylogenetic profiles by generating PSSMs from homologous regions of query sequences, and thus propose a paradigm shift in approach to study evolution of protein families that are in ‘twilight-zone’(<25% identity) of sequence similarity. Further, we also show that this approach can be scaled-up to study mega-phylogenies with thousands of sequences in a computationally efficient manner.

3.1 Background and Motives

Phylogenetic profiles have been suggested by us and others as a unified framework for measuring structural, functional, and evolutionary characteristics of proteins/protein-families(119-124). Under this paradigm, a protein is defined as a vector where each entry quantifies the alignments of a query sequence with a PSSM(119;121;125). Proteins within a phylogenetic profile can be defined in an N(query) by M(PSSM) matrix. In the case of evolutionary measurements, we previously demonstrated that phylogenetic profiles built in this
manner can be used to construct phylogenetic trees using Euclidian distance measurements(119:120).

Indeed, using reverse transcriptases (RT) as our benchmark dataset, we demonstrated that phylogenetic profiles perform well even at extreme levels of divergence (i.e. “twilight zone of sequence similarity”)(119). In this previous study we showed how pre-existing PSSMs, obtained from the Conserved Domain Database (CDD,(15)), could be utilized to construct an informative M-dimension. When we generated trees with the entire CDD we obtained a tree with perfect monophyly. Despite the perfect monophyly, the statistical support at most deep nodes was lacking. Interestingly, when we analyzed the alignments from the phylogenetic profiles, we determined that the most frequently occurring PSSM alignments were the 16 RT domain-containing profiles present in CDD. When trees were constructed using only these 16 RT PSSMs for our M-dimension, we still observed significant monophyly that is well above random(119).

Based on these results, it is reasonable to consider that expanding only the informative profiles within our knowledge base will improve the robustness of phylogenetic profile-based measurements, in addition to improving computational performance. In this study, we present data supporting this supposition. Further, we present a pipeline for enriching and amplifying informative PSSMs as well as an algorithmic improvement that drastically reduces computational expense. As evidence for these theories we analyzed biological (RTs) and simulated (ROSE,(126)) datasets. When compared with other multiple-sequence alignment (MSA) methods, PHYRN reliably recapitulates true evolutionary history in simulated datasets, and provides deep-node measurements with robust statistical support.
3.2 Results

3.2.1 PHYRN Pipeline

PHYRN begins by (i-ii) defining and extracting the domain specific region among the query sequences. (iii) Domain specific regions are then used to create PSSM library using PSI-BLAST. (iv-v) Positive alignments are then calculated between queries and PSSM library using rpsBLAST, and encoded as a PHYRN product score (%identity X %coverage) matrix. (vi) Product score matrix is converted to a Euclidean distance matrix by calculating Euclidean distance between each query pair. vii) Phylogenetic trees are then graphed using Neighbor-Joining (NJ) or Minimum Evolution (ME) as available in MEGA.

The algorithm begins by compiling a set of protein queries belonging to the same protein family/superfamily (Figure 3-1). We use CDD (15) and other approaches to define conserved domains present in most members of this superfamily (e.g. RT domain in reverse transcriptases). From this subset of knowledgebase PSSMs we utilize pairwise comparisons to define boundaries of homology. These homologous protein fragments are then utilized to construct a
database/library of query-based PSSMs using PSI-BLAST (6-iterations, e-value threshold= $10^{-6}$).(127) In this manner, a query set of 100 sequences can make a library of at least 100 PSSMs. We then use rpsBLAST to obtain pairwise alignments between full-length queries and the query-specific PSSM library. This alignment information (% identity, % coverage) is then encoded into phylogenetic profile matrix. Following, we calculate the Euclidian distance between each query. The results from these calculations can then be plotted as a phylogenetic tree using a variety of tree-building algorithms (e.g. Neighbor-Joining(21), Maximum Likelihood(128), Minimum Evolution(129), etc, see Methods for complete description of PHYRN). In the following sections we will highlight the methods for creating informative PSSM libraries and the scoring schemes used.

Figure 3-2: Enrichment and Amplification of Signal Source PSSMs
(a) Family/Superfamily specific PSSMs can be identified using NCBI Conserved Domain Database. (b) rpsBLAST can then be used to identify overlapping alignments between individual queries and family/superfamily specific knowledgebase PSSMs. Overlapping alignments are then used to define domain specific region as described in methods. (c) Domain specific regions as identified are then used to generate PSSMs using PSI-BLAST and NCBI non-redundant (nr) database.
3.2.2 Enriching and Amplifying Informative PSSMs

Although CDD provides a comprehensive resource for conserved domains, in all cases the number of PSSMs for any given domain is relatively small (>100). We have previously demonstrated that quality of phylogenetic profile measurements is proportional to the size and variety of the domain-specific PSSM library\(^{(119;121;122)}\). This increase in information content is exemplified in Figure 3-2. When the silkworm Non-LTR AAA92147.1 is analyzed with NCBI CDD, only one RT specific PSSM alignment is returned (Figure 3-2a). When the same query is analyzed using PHYRN and the 16 RT PSSMs from CDD, we observe 3 overlapping alignments from 3 different PSSMs (~19% of the library, Figure 3-2b). These overlapping alignments define the boundary from which to generate PSSMs using PSI-BLAST. This same approach was used for 100 RT query sequences; post-expansion, we obtain 102 RT-specific PSSMs. When we reanalyzed the silkworm sequence with the amplified library, ~56% of the PSSM library returns a result within the homologous region.

3.2.3 PHYRN Scoring Scheme

In order to encode alignment information between queries and our PSSM libraries, we utilize a product score (%identity \( \times \) %coverage) during our PHYRN analysis. Equation sets for %identity and %coverage are as defined in \(^{(119)}\). The algebraic derivation in Figure 3-3a demonstrates that our product score is equivalent to \((1-p\text{-distance}) \times \) gap weight. In data not shown, we determined that the gap weight is a negligible variable, and when removed does not alter our results. Figure 3-3b depicts the distribution of % gaps, % identity, and % coverage of alignments between 100 RT queries and 102 RT-specific PSSMs. Overall, 92% of alignments are less than 25% identity and the average percentage identity between 100 RT sequences is
21.8% (± 5.6% s.d.). Within a smaller subset of 88 RT sequences, 3644 pairs (95.2%) among 3828 possible pairs of these 88 sequences have less than 25% sequence identity. As a whole RT sequences reside in the “twilight zone” of sequence similarity, underscoring the reason why deducing evolutionary relationships within the RT family is extremely challenging. Furthermore, % gap and % coverage measurements have wide variances.

Figure 3.3: Phylogenetic profile based measurements of evolutionary distance.
(a) Algebraic derivation of p-distance from PHYRN product scoring scheme (%Identity x %coverage). (b) Distribution of 100 Retroelements by measurements of % Identity, % coverage, and %gap. (c) Unrooted phylogenetic tree of 100 full-length retroelements measured using 102 PSSMs generated from the RT domain. The pairwise distances among them were acquired based on Euclidean distance measurement in the 100 X 102 data matrix, and an unrooted phylogenetic tree was derived from the 100 X 102 distance matrix using a minimum evolution method. The tree is drawn to scale, with branch lengths in the same units as those of the Euclidean distances calculated from the data matrix. Bootstrap and jackknife (80% fraction of samples) values were obtained from 1,000 replicates and are reported as percentages.
Figure 3-4: Towards comprehensive phylogenies.
Linearized phylogenetic tree of 716 full-length retroelements measured using 846 PSSMs generated from the RT domain and rooted with retrointrons. The pairwise distances among them were acquired based on Euclidean distance measurement in the 716 X 846 data matrix, and an unrooted phylogenetic tree was derived from the 716 X 716 distance matrix using a neighbor-joining (NJ) method. The tree is drawn to scale, with branch lengths in the same units as those of the Euclidean distances calculated from the data matrix.

3.2.4 Phylogenetic Reconstruction of the RT superfamily

Despite the low-identity and high variance in % gaps and coverage, when a phylogenetic tree is constructed using Neighbor-Joining (21)(see Methods) algorithm with a phylogenetic profile matrix comprised of 100 RT query sequences and 102 RT-specific PSSMs, we obtain a
robust monophyletic tree with deep statistical support (bootstrap and jackknife, Figure 3-3c). In all aspects, this tree is superior to the tree constructed pre-expansion of the RT-specific PSSMs (Fig 3 in Chang et al (119)). Specifically, the Hepadnaviruses now form a clear monophyletic clade, and the Mt Plasmids now reside in the prokaryotic group as expected.

As previously mentioned, phylogenetic profile measurements improve with larger datasets(119;121;122). Therefore, we wondered whether our resolution could be increased by the inclusion of additional sequences across multiple taxa, thereby increasing the size of the PSSM library. We collected 716 full-length RT containing sequences from the literature(38;130-135) and PSI-blast aided searches of NCBI non-redundant database. These sequences were subsequently included in our RT-specific PSSM library as previously described. Figure 3-4 depicts a linearized phylogenetic tree of 716 full length retroelements measured using 846 PSSMs generated from the RT domain and rooted with retrointrons. The pairwise distances among them were acquired based on Euclidean distance measurement in the 716 × 846 data matrix, and an unrooted phylogenetic tree was derived from the 716 × 716 distance matrix using a neighbor-joining method. The tree is drawn to scale, with branch lengths in the same units as those of the Euclidean distances calculated from the data matrix.

Even at this size and level of divergence (~17% identity between groups), the PHYRN tree has robust monophyly. Within these monophyletic nodes, there are multiple subclades which are evident from this analysis. For example, we observe numerous subgroups of TY3 retroelements, and proper subclade groupings of retroviral RTs. Recently, Zimmerly et al. reported that there is a plethora of uncharacterized bacterial RTs(134). Our analysis is congruent with this proposal as we also observe novel clades of bacterial origin.
Figure 3-5: High Throughput Scaling Up of Retroelements

PHYRN based phylogenetic tree as obtained from 8000 Retroelements and 14000 RT-specific PSSMs. Tree was obtained after calculating Euclidean distances between each query using their PHYRN product score matrix. Tree inferred using Neighbor-joining method on N (query) X N (query) Euclidean distance matrix.

To further strengthen our measurements, we decided to collect all the RT-containing sequences from NCBI nr database and generate a phylogenetic tree of all these sequences. Using our PSI-BLAST aided searches of nr database, we were able to collect ~8000 sequences. After including these sequences in our PSSM library, we obtained a PSSM library of ~14000 RT-specific PSSMs. Further, during the course of these experiments, we determined that although effective, the pipeline as described is computationally expensive for large datasets. To overcome these limitations, we made the following changes to the pipeline. Specifically, we compiled all
PSSMs into a single database that can be used by standard rps-BLAST(15). This drastically reduced the number of operations and enabled us to use a sliding window of e-value thresholds to recover positive alignments. In data not shown, these changes afforded us a >600-fold increase in computational speed as well as improved speciation (see Methods for complete details). Tree obtained from PHYRN using 8000 RT queries X 14000 is shown in Figure 3-5. Although, using our new method of designing PSSM library, we were able to generate a tree with such mega-proportions, but the tree thus obtained lacks monophyly and shows significantly erroneous speciation pattern. We believe that this degradation of RT-tree on scale-up is due to inclusion of ‘noisy’ (false positive RT) sequences in our N-dimension, which subsequently corrupts our PSSM libraries as well.

To counter this problem of noisy sequences during scale-up, we decided to follow a step-wise approach to scaling up, where we filter for only the sequences that show high PHYRN product score with previous best RT-specific PSSM library. Using this approach, we collected positive alignments between 8000 query sequences and our previously described PSSM library of 846 RT-specific PSSMs. Following, we calculated PHYRN product score for all positive alignments, and selected queries with only top 0.1% of PHYRN product score. Thus, post PHYRN score filtration, we obtained 1590 query sequences. After including these sequences in PSSM library, we obtained a PSSM library with 1590 PSSMs. Using PHYRN and this dataset of 1590 RT sequences and 1590 RT-specific PSSMs, we obtained a phylogenetic tree using neighbor joining method. Figure 3-6 shows PHYRN tree obtained from this dataset and rooted with Group II introns. PHYRN tree shows robust monophyletic groups, and as evident there are multiple subclades within these monophyletic clades. Furthermore, similar to our 716 X 846 PHYRN tree, we observe multiple subclades of TY3 retroelements, bacterial RTs, and proper
subclade groupings of retroviral RTs. While these results are promising, the true evolutionary history of RT family is unknown and therefore we cannot fully evaluate the performance of PHRYN using this dataset. Under such circumstances, one way to validate PHYRN’s ability to infer ‘true evolutionary history’ is to use simulated datasets, where evolutionary history is known and can be recorded during generation of these datasets.

Figure 3-6: Scaling Up phylogenies improves the resolution.
Linearized phylogenetic tree of 1590 full-length retroelements measured with 1590 RT-specific PSSMs. The pairwise distance between them acquired from Euclidean distance between their phylogenetic profiles, and tree inferred using neighbor-joining method as available in MEGA. The tree is drawn to scale, with branch lengths in same units as those of Euclidean distances calculated from the data matrix.
Figure 3-7: PHYRN recapitulates ‘true evolutionary history’ better than MUSCLE in simulated protein families.
Consensus tree between original ROSE tree and tree generated using a) PHYRN and b) MUSCLE. Simulated protein family generated using ROSE, with an average distance of 550 (p distance ~0.83). Red circles mark the branch points (nodes) that are not recapitulated correctly. (no. of query sequences = 67)
3.2.5 Phylogenetic Reconstruction of Rose/Seq-Gen Simulations

Rose (Random Model of Sequence Evolution - Version 1.3; (http://bibiserv.techfak.uni-bielefeld.de/rose/) implements a probabilistic model for protein sequence evolution(126). In this simulation, sequences are created from a common ancestor to produce a dataset of known size, divergence, and history. In this artificial evolutionary process, the accurate history is recorded since the multiple sequence alignment is created simultaneously. This allows us to have perfect control over evolutionary rates, allowing us to test the efficacy of our approach.

Figure 3-7 provides the results from PHRYN and MUSCLE(19) using 67 sequences simulated for 17% identity by ROSE (see Methods for full description of PSSM generation). Whereas MUSCLE performs poorly on this dataset (Figure 3-7a), PHRYN recaptures 94% of the true evolutionary history and has only one deep node incorrectly identified (Figure 3-7b). In a second simulation, we maintained a similar level of divergence while increasing the size of the simulated dataset to 584 sequences. In this simulation, MUSCLE performance decreases significantly (no deep-nodes are correctly obtained, Figure 3-8a), while PHRYN performance is not degraded (Figure 3-8b). Further, we also counted the deep node recapitulation in simulated datasets. PHRYN recapitulates 85% of deep nodes in dataset with 67 sequences, while MUSCLE recapitulates only 11% of deep nodes correctly. Further, we also generated multiple replicates of datasets at extreme divergence range of p-distance~0.93 using Seq-Gen(136;137). Using same approach as used for ROSE datasets, we quantified recapitulation rates of PHRYN and MUSCLE for these datasets. We also generated multiple replicates (n = 25) of simulated protein families using SeqGen at 7 different ranges of divergence (40%-7% identity, Figure 3-9a). At lower levels of divergence, PHRYN marginally outperforms MUSCLE at recapitulating deep nodes; however, at higher divergences, PHRYN performs significantly better. Even at extreme
divergence range of p-distance \sim 0.93 (7\% identity, SeqGen scaling factor = 1), PHYRN recapitulates 51.7\% of deep nodes correctly. In the same data sets, MUSCLE recovers only 24.3\% of deep nodes correctly and Dialign fails to generate trees at all. Within 6 different ROSE data sets (n=25) including indels, PHYRN performs significantly better than MUSCLE at recapitulating deep nodes in divergent data sets (Figure 3-9b). For data sets generated using ROSE at an average distance of 600 PAM* (pdistance \sim 0.92, 8\% identity), PHYRN recapitulates 95\% of deep nodes correctly, while MUSCLE is only able to recapitulate 33\% of deep nodes. Overall, the 325 total simulations assessed here provide confidence that PHYRN reliably measures branching orders in highly divergent or rapidly evolving sequences.
Figure 3-8: Deep Node Recapitulation of ‘true evolutionary history’ in mega-phylogenies.
Consensus tree between original ROSE tree and tree generated using a) MUSCLE and b) PHYRN. Simulated protein family generated using ROSE, with an average distance of 550 (p distance ~0.82). Green circles mark the deep nodes that are recapitulated correctly in the consensus trees. (no. of query sequences = 584)
Figure 3-9: Deep Node Recapitulation in Simulated Datasets.

a) Graphical representation of %Deep Node recapitulation versus SeqGen scaling factor. Number of replicates for each bar = 25, Error bars = +/- S.E.M. *p-value < 0.01. Number of sequences in each data set =100, Length of sequences = 450.

b) Graphical representation of %Deep Node recapitulation vs. Average Rose Distance. Number of replicates for each bar = 25, Error bars = +/- S.E.M. *p-value < 0.01. Number of sequences in each data set = 100, Avg. Length of sequences = 450.

3.3 Materials and Methods

Sequence collection

The sequence set reported in (119) was used and additional sequences were curated from literature (38;130-135) and PSI-BLAST aided search of NCBI’s non-redundant (nr) database. We used Position Specific Iterative (PSI)-BLAST and our previously defined query and PSSM sequences to search for more RT-containing homologous sequences in NCBI nr database.

Defining RT specific boundaries and generating RT-specific PSSMs

Statistical analysis was performed on overlapping alignments generated between each query sequence and RT-containing profiles in NCBI’s Conserved Domain Database (CDD(15)). The ‘start’ position of RT-specific region was defined from the start of the most N-terminal alignment and ‘End’ position was defined from the end of the most C-terminal alignment (as
described in (119). The RT-specific regions defined in the aforementioned manner were then extracted from their full-length query sequences using a custom PERL script. We generated PSSMs with these regions using Position Specific Iterative (PSI)-BLAST with NCBI non-redundant (nr) database, 6 iterations and e=10^{-6} as our parameter settings(127). Using, this method, at least one PSSM is generated from every query sequence. Thus, we designed a PSSM library of 102 RT-specific from our query set of 100 RT sequences. Similarly, we designed a RT-specific PSSM library of 846 sequences from 716 query sequences. For computational purposes, these libraries were designed as an assembly of individual single-domain databases, where each database consists of a single RT-specific PSSM.

Sequence simulation using ROSE/Seq-Gen and generating PSSM libraries of simulated sequences

We used Rose v1.3(126) to generate protein families of simulated sequences. Nematostella DANGER D6 sequence was used as a root sequence, and default substitution parameters were used to generate datasets of with average length of 450 amino acids and average distance of 550 (p-distance ~0.83). We generated two different datasets; a simulated dataset of 67 sequences and another simulated dataset with 584 sequences. To generate multiple replicates of simulated datasets, we used Seq-Gen v 1.3.2(138) with PAM as default substitution matrix.

To generate PSSM libraries for simulated datasets, we used full-length sequences to generate PSSMs, as biological conserved domain boundaries cannot be defined in simulated sequences. Further, owing to the lack of biological homologues for these simulated sequences in non-redundant (nr) database, we added all the query sequences from a particular simulated dataset to nr database. Following, we used full-length simulated sequences to generate PSSMs
using PSI-BLAST with aforementioned, modified nr database, 6 iterations and e-value=$10^{-6}$ (127)

We also modified the PSSM library architecture to allow for increased computational speed in simulated libraries. Instead of organizing PSSM library as an assembly of individual single-domain databases, we changed library organization to have one single database comprised of all the PSSMs.

**PHYRN Phylogenetic trees**

*Reverse Transcriptase (RT) Trees*

To generate phylogenetic trees with RT sequences, modified rpsBLAST (GDDA-BLAST) (119) was used to recover positive alignments between each query and RT-specific PSSM library. Equation sets for calculating %identity and %coverage are defined in Chang et al. (119). But, unlike the composite score mentioned in Chang et al. (2008), we used PHYRN product score (%identity X % coverage) for each PSSM that provided an alignment without embedding. This composite score for positive alignments was then encoded as $N$ (no. of queries) X $M$ (no. of PSSMs) matrix of product scores. Following, Euclidean distance for each query was calculated based on the product score of each query and an $N \times N$ Euclidean distance matrix was generated. Phylogenetic trees were inferred using Neighbor Joining (NJ) and Minimum Evolution (ME) method, as available in MEGA (129).

*Trees for ROSE simulated datasets*

Since PSSM library architecture was different in simulated datasets, we were able to use rpsBLAST for recording positive alignments between simulated sequences and their respective PSSM library. We used an e-value threshold of $10^{10}$ for selecting positive alignments, which were scored using PHYRN product scoring scheme (%identity X %coverage), and encoded as $N$
An (no. of queries) X M (no. of PSSMs) matrix of product scores. Euclidean distance was calculated based on the PHYRN product score of each query and an N X N Euclidean distance matrix was generated. Similar to method used for RT trees, phylogenetic trees were inferred using Neighbor Joining (NJ) and Minimum Evolution (ME) method, as available in MEGA(129).

**Generating phylogenetic trees using MUSCLE**

Optimal Multiple Sequence Alignment (MSA) for a given dataset was obtained using MUSCLE v3.6 (18;19). Phylogenetic trees for these optimal MSA were inferred using MEGA’s Neighbor-joining (NJ) and Minimum-evolution (ME) algorithm, with pairwise deletion and p-distance as default settings.

**Consensus trees with ROSE ‘true history’**

We used ‘consense’ program of PHYLIP v3.67 package (http://evolution.genetics.washington.edu/phylip.html) (139;140) to generate consensus trees between PHYRN and ROSE trees, as well as, between MUSCLE and ROSE. Recapitulation rate and percentages were then calculated from consensus tree newick files.

**Bootstrap and Jackknife**

We generated 3,000 random samples from our PHYRN M-dimension, using random number generator code from PHYLIP source code (http://evolution.genetics.washington.edu/phylip.html) (139;140). During resampling, same columns were allowed to be selected more than once. We then used Fitch program with default settings in PHYLIP 3.67 package to generate minimum-
evolution (ME) trees for each sample, followed by Consense program for generating a consensus trees of all samples by majority rule.

For jackknife resampling, we followed a similar approach to generate 1,000 random samples, however only 80% of original M-dimensional data was resampled each time. FITCH and Consense programs were then used in similar manner as used in bootstrap resampling.

3.5 Discussion

Our case-study of the RT superfamily and simulated datasets demonstrates that PHYRN is capable of inferring deep evolutionary relationships between highly divergent proteins. A number of implications can be derived from this study: (i) phylogenies built with PHYRN recapture more of the true evolutionary history and have robust statistical support; (ii) phylogenies built on pairwise alignments outperform conventional MSA methods and (iii) this method is scalable to thousands of sequences. This improved performance is due the improved information content contained in the PSSM libraries used in this study. We improved the efficacy of our PSSMs by: (i) limiting the PSSMs to homologous domains, (ii) optimizing the PSI-BLAST settings for their generation, and (iii) creating a pipeline that is sufficiently fast to handle large datasets.

Conversely, with respect to MSA dependent methods, increasing the number of query sequences makes it increasingly difficult to obtain an optimal multiple sequence alignment(141); in PHYRN, increasing number of query sequences also increases the dimensionality of the phylogenetic profile, thus increasing the alignment information space. This increase in information space leads to better, more robust measurements of relative rates. This ‘comprehensive survey’ approach, where more sequences are better, is in contrast to ‘random
walk’ approach of MSA dependent methods where increased sequences are a problem. Further, use of frequency tables in the phylogenetic profiles provides more informative measurements for calculating relative rates of evolution. This approach provides PHYRN with a potential to generate trees with thousands of sequences where the only theoretical limit is the available sequencing data. Indeed, when we expanded the RT tree from 100 to 1590 sequences comprising >14 groups we obtain a tree that is consistent yet higher resolution than previously reported RT studies (38;119;131;132).

As PHRYN is well suited to making measurements on large divergent datasets, we hypothesize this approach may be capable of solving a number of unanswered questions related to the ancient origins of life and speciation. Moreover, since PHYRN functions in the twilight zone of sequence similarity, this algorithm may have the ability to inform whether functionally or structurally similar proteins have a common ancestor or occurred via convergent evolution. In conclusion, our study provides strong evidence that, even in its nascent stage, PHYRN measurements can provide key insight into evolutionary relationships among distantly related and/or rapidly-evolving proteins.
Chapter 4

Resolving Deep-node Evolutionary Relationships between Rapidly Evolving Proteins

DANGER (Differentiation and Neural Growth Evolve Rapidly) is a recently discovered, highly divergent, (p-distance ~0.8) and rapidly evolving protein superfamily (41). Due to its divergent nature, previous Multiple Sequence Alignment (MSA) based phylogenetic studies of this family suffer from lack of statistical support in their measurements, especially at deep nodes (41). In this chapter, we use PHYRN to study this rapidly evolving benchmark superfamily. We also compare the results obtained from our method (PHYRN) with other MSA dependent methods and show that PHYRN provides more robust measurements at deep nodes. Further, by using various other PHYRN measurements we show that robust deep node relationships can also overcome ambiguities in outgroup identification. By using phylogenetic profiles generated from homologous Mab-21 regions of query sequences as our information source, we resolve many of the ambiguities in previously conducted phylogenetic studies of DANGER superfamily.

4.1 Background and Motives

There is a common frustration in evolutionary studies: obtaining statistical support for relationships between highly-divergent proteins from the "primordial ooze." Often the available data and taxa are too sparse and/or noisy to support straightforward phylogenetic analyses that would elucidate the origins of interesting proteins, and thus pathways and ecological phenotypes. Due to the information explosion in this genomic era, there is a great need for knowledge of deep evolutionary events. As with most efforts to push the development of analyses of large complex
problems, a successful tool will likely be recognized as useful in many other applications. Conventional MSA methods tend to get progressively worse with additional divergence\(^{(4;6)}\). This is due to the low information content of divergent sequences and the subsequent loss of informative points from which to measure. Theoretically, methods that could amplify the number of sequences measured and the signal contained in each sequence, would get progressively better with additional divergence\(^{(121)}\).

We recently demonstrated that phylogenetic profiles built using position-specific scoring matrices (PSSMs) of protein domains from NCBI’s Conserved Domain Database (CDD\(^{(15)}\)) have the capacity to infer evolutionary distances between highly-divergent proteins\(^{(119;120)}\). These results are not random, and provide a phylogenetic measure that is independent of multiple-sequence alignment (MSA) and/or hybrid-based methods (MSA coupled to pair-wise alignments). Although informative, these phylogenies lacked robust statistical support. Data from this study suggested that the “signal source” of our measurements is contained in PSSMs that are generated from homologous protein domains\(^{(119)}\). We reasoned that expanding only the informative profiles within our knowledge-base would improve the robustness of our measurements, as well as improve computational performance.

In the present manuscript, we challenged these suppositions in a new benchmark dataset composed of rapidly-evolving proteins. Until recently, the MAB21-domain containing superfamily (DANGER) escaped detection due to the extreme mutational rates found in these proteins\(^{(41)}\). This developmental superfamily is ubiquitously expressed and is linked to multiple physiological (\(\text{Ca}^{2+}\) signaling, cranio-facial development, neurite outgrowth\(^{(40-42;142-144)}\)) and pathophysiological (chromosomal breakage) processes\(^{(145;146)}\). Our earlier attempts utilized
Figure 4-1: Phylogenetic tree of DANGER derived using PHYRN and Conserved Domain Database (CDD) PSSM library.

Phylogenetic tree of 108 DANGER sequences, inferred using rpsBLAST and 24,280 PSSMs from NCBI CDD database. The pairwise distances among them were acquired based on Euclidean distance measurement in the 108 X 24,280 data matrix, and a phylogenetic tree was derived from the 108 X 108 distance matrix by using a neighbor-joining method. The tree is drawn to scale, with branch lengths in the same units as those of the Euclidean distances calculated from the data matrix. Two kinds of statistical estimation (Bootstrap and Jackknife) for tree branching were performed and are each branch is colored based on percentage values. Bootstrap and jackknife values were obtained from 3,000 replicates. Branches with bootstrap and jackknife support both above 80% are colored black, while branches with any of the bootstrap/jackknife value < 80% are colored red.
MSA-based methods to resolve the origins of this superfamily; however, these were stymied due to their extreme divergence. Indeed, significant genetic correlates were required to support monophyletic groups. Even upon rigorous genetic analyses and extensive manual-editing of these alignments, deep-node statistical support was unattainable (41). In the present study, we demonstrate that: (i) high-resolution phylogenies can be built using PSSMs generated from homologous protein domains, (ii) these measurements have robust statistical support and inform within-group and between-group relationships, (iii) these measures outperform both MSA (MUSCLE) and hybrid (Dialign) alignment algorithms, and (iv) these measurements are orders of magnitude faster than our previous approach.

Figure 4-2: Resolving DANGER superfamily using PHYRN
Mab-21 specific boundary region in DANGER sequences was identified, by collecting positive alignments between DANGER sequences and CDD Mab-21 PSSM (pfam06381) (see Methods). Mab21-specific regions are then used to create PSSM library using PSI-BLAST. Positive alignments are then calculated between queries and PSSM library using rpsBLAST, and encoded as PHYRN product score (%identity X %coverage) matrix. Product score matrix is converted to a Euclidean distance matrix by calculating Euclidean distance between each query pair. Phylogenetic trees are then graphed using Neighbor-Joining (NJ) or Minimum Evolution (ME) as available in MEGA.
4.2 Results

4.2.1 Phylogenetic Reconstruction (PHYRN)

Using our previous approach (119;120), we analyzed 108 DANGER sequences and generated a phylogenetic tree (Figure 4-1). Unlike the results we achieved for reverse-transcriptases (RT), we do not obtain monophyly or statistical support for this superfamily. In light of the results obtained using only RT-domain specific PSSMs, we developed a standardized method to generate MAB21-domain specific PSSMs (Figure 4-2, see Methods). The innovation of this approach is the isolation of the homologous region in each protein and the subsequent conversion of this region into a PSSM (Figure 3-2a-c). Following, pairwise alignments for each query sequence are then measured by rps-BLAST (15) against the PSSM library (Figure 4-2d). All results from rps-BLAST are represented as a product score (% identity X %coverage). As previously described, the derivatives of these measures are equivalent to [1-(p-distance)]. Thus, each query sequence is represented in a vector of non-negative numbers in M dimensions (M= # of PSSMs tested) (Figure 4-2e). Using the Euclidian distance between each query sequence, this data can be used to generate phylogenetic trees. Importantly, by creating “signal source” PSSMs, PHYRN: (i) does not require sequence embedding as our previous algorithm(119;120), and (ii) reduces our computational expense by >99%.

The phylogenetic trees for DANGER using PHYRN are displayed in Figure 4-3a and Figure 4-4a. These trees were built using 108 full-length query sequences and 112 MAB21-domain PSSMs. The pairwise distances among them were acquired based on Euclidean distance measurement in the 108 × 112 data matrix and a phylogenetic tree was derived from the 108 × 108 distance matrix using a neighbor-joining method(21) (see Methods). The tree is drawn to scale, with branch lengths in the same units as those of the Euclidean distances calculated from
the data matrix. For comparison, phylogenetic trees were generated using Dialign(147) (Figure 4-3b and Figure 4-4b) and MUSCLE(19) (Figure4-3c). We observe that PHYRN identifies 6 monophyletic clades (D1-D6), with reasonable speciation. For example, within the D6 clade, cnidaria assumes the outgroup position, followed by nematode, urochordates, arthropods, and chordates. However, a single sequence from sea-urchin emerges after arthropods, and thus appears to be misplaced; although, both Dialign(147) and MUSCLE(19) place this sequence in the same position. While Dialign also identifies 6 monophyletic clades (Figure 4-3b), MUSCLE separates sequences in the D2 clade (Figure 4-3c). The relative rates of cladistic evolution vary between all models.

4.2.2 Statistical Support and Outgroup Identification

To determine which of these topologies is most parsimonious, we conducted bootstrap (PHYRN, MUSCLE, Dialign) and jack-knife (PHYRN) statistical resampling (see Methods). We observe that PHYRN obtains support of >83% (bootstrap) and >88% (jack-knife) for all deep-nodes except for the placement of the D4 clade. Conversely, neither MUSCLE nor Dialign obtain significant results for any deep-node other than the D5/D6 clades, which are the most conserved clades in this study. In our previous evolutionary study of DANGER, we identified a single sequence from choanoflagellate which was used as the putative outgroup; although this sequence obtained no statistical support for this position(41). To test whether this sequence was indeed an outgroup, we identified 13 additional sequences from the NCBI and CAMERA databases with sequence similarity. When we incorporate these sequences into our analyses, PHYRN obtains a monophyletic topology; however, the choanoflagellate sequences form a distinct clade, with D3 as the nearest neighbor (Figure 4-5a). However, their inclusion drastically reduces the statistical support across the entire tree.
Figure 4-3: Resolving Deep Nodes in DANGER Superfamily
Compressed Neighbor-Joining (NJ) trees for 108 DANGER sequences generated using a) PHYRN b) DiAlign c) MUSCLE. Statistical support for PHYRN calculated using Bootstrap and Jackknife analysis, while for Dialign and MUSCLE trees, only bootstrap was used. Branches with bootstrap score <80% are colored red. PHYRN outperforms both DiAlign and MUSCLE in deep node statistical support.
Figure 4-4: Resolving evolution of DANGER Superfamily

Neighbor-Joining (NJ) trees for 108 DANGER sequences generated using a) PHYRN b) DiAlign. Statistical support for PHYRN calculated using Bootstrap and Jackknife analysis, while for Dialign only bootstrap was used. The blank marked “/” in the statistical support indicates that the clustering of the branching connection cannot be measured in a standardized fashion by the given resampling method (see Materials and Methods). Bootstrap and jackknife values were obtained from 3,000 replicates and are reported as percentages.
Due to the matrix generated by PHYRN, we can obtain additional quantitative measurements from this dataset. These include the group-wise distribution of the PSSMs used, as well as their information content (Figure 4-5b). In all cases, choanoflagellate PSSMs have the fewest alignments across all clades, and their sequences have the lowest information content (average product score, ± S.E.M). Further, the positions of the choanoflagellate sequences within the tree are suspect. This is due to the position of the D1 clade, which is vertebrate specific. In this scenario, multiple clades which contain ancient species (e.g. cnidarians, nematodes, arthropods, etc) would have evolved after D1. Thus, in order for this scenario to make sense, D1 proteins would have to be lost from all species prior to chordates, which is not parsimonious. Further, after exhaustive searches, we could not identify any DANGER sequences in species before choanoflagellate or between choanoflagellate and cnidaria. In addition, neither MUSCLE nor Dialign can generate a tree including these sequences, once again suggesting these sequences are not homologous to DANGER (data not shown).

Figure 4-5: PHYRN and Outgroup Identification
Compressed phylogenetic tree of DANGER superfamily, including Monosiga sequences derived using PHYRN. Statistical support calculated using Bootstrap and Jackknife analysis. Bootstrap and jackknife values were obtained from 3,000 replicates and are reported as percentages. b) Group-Wise Distribution of PHYRN product score (%identity X %coverage) distribution. Errors bars represent Standard error in measurements.
Thus, the question arises: which DANGER clade is the oldest? In our quantitative statistics, we observe that PSSMs from the D6 clade have the highest group-wise distribution and D6 sequences have the highest information content (Figure 4-5b). Further, in the unrooted tree, the D6 clade has the longest branch-length. Taken together, D6 is the most logical outgroup of the superfamily based on (i) statistical support, (ii) information content, and (iii) speciation.

4.3 Materials and Methods

Sequence collection- The sequence set reported in (41) was used and additional sequences were curated from NCBI, Community Cyberinfrastructure for Advanced Marine Microbial Ecology Research and Analysis (CAMERA) and Department of Energy Joint Genome Institute (JGI) database. A total of 123 sequences were curated from these sources.

Defining MAB21 specific boundaries and generating MAB21 specific PSSMs- Statistical analysis was performed on alignments generated between each query sequence and CDD MAB21-containing profiles. As, mentioned in Chapter 3, the ‘start’ position of MAB21-specific region was defined from the start of the most N-terminal alignment and ‘End’ position was defined from the end of the most C-terminal alignment (as described in (119)). The MAB21-specific regions defined in the aforementioned manner were then extracted from their full length query sequences using a custom PERL script. We generated PSSMs with these regions using Position Specific Iterative (PSI)-BLAST(127) with NCBI non-redundant (nr) database, and 6 iterations and e=10^-6 as our parameter settings.

PHYRN Phylogenetic trees- For the generation of a DANGER tree with CDD PSSMs, each full length query was screened with 24,280 CDD PSSMs. But, unlike the composite score mentioned in Chang et al. (2008), we used PHYRN product score (%identity X % coverage) for each PSSM
that provided an alignment without embedding. Thus, an N (no. of queries) X M (no. of profiles) matrix of product score values was generated. Euclidean distance was calculated based on the product score of each query and an N X N Euclidean distance matrix was generated. Phylogenetic trees were inferred using Neighbor Joining (NJ) and Minimum Evolution (ME) method, as available in MEGA(129). For the generation of DANGER trees containing monosiga sequences, each query was screened against 126 MAB21-specific PSSMs, and the Euclidean distance matrix was generated as above. In trees lacking monosiga sequences, their respective PSSMs were removed from the PSSM library and processed as above.

**Bootstrap and Jackknife-** We generated 3,000 random samples from our PHYRN M-dimension, using random number generator code from PHYLIP source code (http://evolution.genetics.washington.edu/phylip.html). During resampling, same columns were allowed to be selected more than once. We then used Fitch program with default settings in PHYLIP 3.67 package to generate minimum-evolution (ME) trees for each sample, followed by Consense program for generating a consensus trees of all samples by majority rule. For jackknife resampling, we followed a similar approach to generate 3,000 random samples, however only 80% of original M-dimensional data was resampled each time. FITCH and Consense programs were then used in similar manner as used in bootstrap resampling.

**4.4 Discussion**

Our case-study on the DANGER superfamily demonstrates that PHYRN is capable of inferring deep evolutionary relationships in rapidly-evolving proteins. A number of implications can be derived from this study: (i) phylogenies built with PHYRN can have robust statistical support; (ii) phylogenies built on pairwise alignments outperform conventional MSA and hybrid
methods; (iii) quantitative measurements of the information content in PHYRN can aid in determination of homologous sequences and outgroups. Furthermore, due to the decreased computational expense afforded by PHYRN, this method is scalable to thousands of sequences.

Based on our statistical analysis of PHYRN measurements, if we root with D6 as an outgroup, we see a ‘simple to complex’ evolutionary pattern for DANGER superfamily as vertebrate specific clade D1 attains a position farthest away from the outgroup in the tree. Even within monophyletic clades, we see appearance of simpler organisms before more complex organisms. In D6 clade, cnidarians are the first ones to show DANGER followed by nematodes, Arthropods and then chordates. Also, appearance of chordate specific D4 and D5 clades before cnidarians sequences containing groups like D3 and D2 suggests a possible loss of those genes in earlier species.

Figure 4-6: Evolution of DANGER Superfamily
Putative scenario for evolution of DANGER superfamily inferred using PHYRN. Earliest DANGER appeared in Cnidarian organisms and then evolved into 6 different clades. Chordate specific group, D1 attains furthest position from the root of tree, D6 group.
We studied a rapidly evolving developmental superfamily with PHYRN and demonstrate that robust deep node relationships can be inferred using PHYRN. We compared the results obtained with Dialign(147) and MUSCLE(19). Although all these methods are able to predict the monophyletic groups to a good degree but only PHYRN provided a good statistical support for deep nodes. PHYRN proposes a paradigm shift in approach to studying deep node relationships between rapidly evolving/highly divergent groups. While in MSA dependent methods, increasing the number of query sequences makes it increasingly difficult to obtain an optimal multiple sequence alignment, in PHYRN, increasing number of query sequences, also increases no. of phylogenetic profiles, thus increasing the alignment information space. This increase in information space leads to better, more robust measurements of relative rates.

Overall, we show with the DANGER benchmark set that PHYRN can be used to resolve deep-node relationships in rapidly evolving families. Due to modular nature of its implementation libraries, PHYRN can be trained on any homologous domain, and thus it can serve as a good tool to study other intractable problems in evolution. Taken together with its ‘scale-up’ abilities, we propose that PHYRN can be used to infer deep node evolution of rapidly evolving/highly divergent protein families.
Chapter 5

Resolving Convergent Evolution Using PHYRN

Haloacid Dehalogenase (HAD) superfamily is a highly divergent protein superfamily, which has members in all three super-kingdoms of life. Members of this superfamily act on wide variety of substrates, such as, nucleotides, proteins and sugars. Earlier studies relying on structural, sequence and phyletic patterns have shown that HAD superfamily evolved convergently, and there are at least 5 different lineages in extant members of HAD superfamily(43). Although, a lot is known about the functional and structural aspects of this superfamily, phylogenetic studies of this superfamily have been hampered by the convergent evolutionary nature of this superfamily. Conventional MSA dependent phylogenetic algorithms often fail to detect convergence, and so these studies have to be supplemented by other structural and sequence analysis to detect convergence in any protein family(43). In this chapter, we show a novel method relying on PHYRN and hierarchical clustering, which can be used to infer convergent evolution in structurally homologous superfamilies.

5.1 Background and Motives

Members of Haloacid Dehalogenase (HAD) superfamily are involved in catalysis of carbon or phosphoryl group transfer(148). Since phosphoryl or carbon transfer is an important step in multiple biochemical reactions, members of HAD family evolved to act on variety of substrates, ranging from nucleotides to sugars. Indeed, this superfamily contains 33 different groups, which include enzymes like ATPase, phosphonatses, dehalogenase, and phosphoesterases(149;43). Due to the wide applicability and essentiality of this reaction step,
members of HAD superfamily have evolved in all three kingdoms of life. Although essentiality of different members of HAD superfamily varies extensively across different proteomes and kingdoms.

Structurally, members of HAD superfamily contain a Rossmanoid fold with inserts called caps within this Rossmanoid fold. These caps can belong to various groups, such as, C1 caps that are part of a flap like region over the active site, while C2 caps form a cover over this flap region(43). These distinctive modifications in caps actually provide varying degrees of steric properties to members of HAD for substrate binding(150-152). In fact, diversification of the cap region is major factor contributing to wide range of substrate specificities of HAD members(153). Although, there are many X-ray crystallographic structures and sequences available for members of HAD superfamily, there is still a lack of extensive phylogenetic study on the HAD superfamily.

Earlier attempts of an extensive phylogenetic study for this superfamily have been stymied by extreme divergence and also potential convergence during evolution of HAD superfamily. Studies that rely on combining structural, sequence, and/or phyletic patterns have extrapolated at least five different lineages in HAD superfamily. The five distinct lineages among extant HAD members have been extrapolated to evolve from 1) a MDP and CTD precursor 2) a NagD sequence 3) a Cof sequence 4) a PATPase sequence 5) a helical C1 cap containing sequence(43).

In this chapter we have described a novel method that uses PHYRN and hierarchical clustering, to identify number of distinct lineages in HAD superfamily. Further, we have also performed an MSA independent phylogenetic study on different lineages of HAD superfamily.
5.2 Results

5.2.1 Single Lineage Assumption

Amino acid sequences belonging to HAD superfamily were collected from Burroughs et al (43) and NCBI non-redundant (nr) database, and as a null hypothesis we assumed that HAD family did not evolve in a convergent fashion and thus, all extant sequences of HAD superfamily belong to a single lineage. After using the PHYRN approach mentioned in earlier chapters, we defined a single phylogenetic tree with sequences from all groups of HAD superfamily. As shown in figure 5-1, phylogenetic tree based on ‘single lineage assumption’ shows a significant lack of monophyly. Even after intensive pruning efforts, groups like dehr and Sdt1p split up into multiple groups. Further, speciation pattern shows many anomalies in this tree. In some cases, a close examination of species information of each group presents highly unlikely evolutionary scenarios that require loss of certain HAD groups from many genera, followed my multiple reappearances in other closely related genera. Statistical analysis on this tree also shows that tree is not robust (data not shown). As we have seen in our earlier studies with DANGER and Reverse Transcriptase families that addition of unrelated sequences in the trees, can add ‘noise’ to our PHYRN measurements and lead to results that are similar to the ones seen here. Thus, the reason for abnormal speciation pattern and lack of robustness among this phylogenetic tree based on ‘single lineage assumption’, can be that all HAD sequences do not belong in one tree. So based on 1) lack of monophyly 2) anomalies in speciation pattern, and 3) lack of robustness in phylogenetic tree constructed with all the HAD sequences taken together as one lineage (Figure 5-1), we believe that HAD superfamily did not evolve from a single ancestral sequence.
Figure 5-1: HAD Superfamily phylogenetic tree based on single lineage assumption. Neighbor-Joining tree of HAD Superfamily generated using PHYRN and based on the single lineage assumption. Tree has been rooted with PATPase group. Tree shows significant lack of monophyly and displays lot of anomalies in speciation pattern.
Figure 5-2: Resolving Convergent Evolution using PHYRN and Hierarchical Clustering.  
i) Composite score from alignments between HAD sequences and HAD specific PSSMs as obtained using PHYRN are recorded in composite score matrix.  
ii) composite score matrix is further compressed into group-wise averages and an average composite score matrix listing groupwise averages is generated.  
iii) Cluster 3.0 is used to calculate correlation values (Pearson’s Correlation Coefficients) between each group based on their average composite scores.  
iv) Java TreeView is then used to represent these correlation scores in graphical tree format.

5.2.2 Pearson Clustering and identifying number of lineages

If HAD superfamily did not evolve as a single lineage, then the question arises; How many different lineages are present in extant groups of HAD superfamily? In order to find an answer to this question, we decided to use a hybrid between our PHYRN methodology and hierarchical clustering. Using PHYRN approach as described in chapter 3, we collected pair-wise alignments between full-length HAD sequences and HAD specific PSSMs. A composite score (% identity X % coverage) matrix for alignment between each HAD sequence and HAD profile
was obtained (figure 5-2i). Since all the query sequences belong to one of the 33 individual HAD sequence groups, therefore all the PSSMs generated from these sequences can also be divided into 33 PSSM groups respectively. We calculated the average composite scores for alignments between each query group vs each PSSM group. These average group-specific composite scores were used to construct an average composite score matrix (figure 5-2ii). We then used Cluster 3.0(154) to calculate Pearson correlation coefficients for this average composite score matrix (Figure 5-2iii). Further, Java TreeView(155) was used to infer a dendrogram from the correlation coefficients calculated in earlier step (figure 5-2iv). The dendrogram thus obtained is shown in Figure 5-3. Based on this PHYRN based clustering approach, we believe there are 4 different lineages in HAD superfamily (branch points marked in red circles). In this approach, we are assuming that all the different clusters in our dendrogram evolved from a different ancestor. Thus, number of distinct clusters is equal to the number of different lineages present in HAD superfamily. If this assumption holds true, then phylogenetic trees of individual clusters should be significantly better and robust, than the ‘single-lineage’ HAD phylogenetic tree (Figure 5-1).

5.2.3 Distinct HAD lineages

Using PHYRN, we constructed individual lineage specific phylogenetic trees, after sorting sequences and PSSMs for each of 4 clusters inferred from Pearson clustering. Individual trees thus obtained for each cluster are better than HAD tree that contained all the sequences because they are 1) more robust and 2) provide better speciation pattern. Our clustering approach suggests 4 different points of origin for HAD lineages; 1) representative of cNII nucleotidase 2) a 8KDO representative 3) a SPSC representative 4) a PATPase/PSP representative.
Figure 5-3: Hierarchical clustering of PHYRN data reveals independent lineages in HAD Superfamily. PHYRN and Pearson's clustering approach shows 4 different clusters (lineages) in the HAD superfamily. Branch points marking individual lineages are labelled with red circles.

Figure 5-4 shows a HAD lineage tree that evolved from a representative of cNII nucleotidase group. With the exception of 38K group, all other groups form distinct monophyletic clades. Archeal 38K sequences form a distinct clade, but it is possible that these archeal 38K sequences were progenitors of other groups like RNApol and deoxyribonucleotidase, alongwith giving rise to 38K in higher organisms. cNII nucleotidase also
emerges as a natural root in the tree i.e. it has the longest branch length. Further, rooting with cNII also presents an evolutionary scenario, where cNII-nucleotidase is the oldest group in this lineage. As cNII is a universal group with sequences from all three kingdoms of life, it is potentially the oldest as it has had the most time to evolve. This universal group during the course of evolution then gave rise to other bacteria-specific and universal groups. Further, earlier studies on HAD families also believe that groups that used nucleotides as substrates are evolutionarily older. Thus our analysis on this lineage corroborates these studies (43).

Figure 5-4: Unrooted Neighbor Joining (NJ) tree of cNII nucleotidase derived HAD lineage. Tree generated using PHYRN methodology shows good monophyly, with the exception of 38K group where archaeal members form a separate clade. (n=340 sequences)
Figure 5-5 shows another phylogenetic lineage that putatively evolved from a single representative of 8KDO group. This phylogenetic tree shows good monophyletic distribution of sequences, with all groups forming separate clades. It is interesting to note that phage specific PNKPs form a separate clade. It is not unprecedented in HAD superfamily that, bacterial, archeal and viral sequences of a particular group form separate clades. Earlier studies by Burroughs et al. (43) have shown that bacterial, archeal, and viral members in certain groups have distinctive structural and functional attributes than other members belonging to same group. Thus, phage specific PNKP group may indeed be structurally and functionally different from other PNKP sequences.

Figure 5-5: Unrooted Neighbor Joining (NJ) tree of 8KDO derived HAD lineage. Phylogenetic tree was generated using PHYRN with lineage specific sequences (n = 273) and lineage specific PSSMs. Individual groups show good monophyly but phage PNKP sequences form a separate clade.
Figure 5-6: Unrooted Neighbor Joining (NJ) tree of SPSC derived HAD lineage.
PHYRN derived phylogenetic tree of a HAD lineage which putatively evolved from a representative of SPSC group. Archeal Cof(s) form a separate clade and SPP group splits into plant specific and bacteria specific groups. (No. of sequences = 363)

Another putative lineage in HAD superfamily evolved from a single representative of SPSC group. An unrooted neighbor-joining tree for this particular phylogenetic lineage as generated using PHYRN is shown in Figure 5-6. This tree also shows good monophyly, with the exception of some archaea specific and bacteria specific groups, which form separate clades. As shown in figure 5-6, an archea specific group of Cof was a common ancestor for M3P and other Cof sequences. Also, there are some plant specific groups very close to the root of the tree. It is possible that plants may have lost some of the other members of HAD lineage over a period of time. In this phylogenetic tree SPSC groups shows longest branch length and thus is an ideal
candidate for the root position. If we root with SPSC group, which contains members from bacteria, fungi and plant kingdoms, then we get an evolutionary scenario in which a group with wide divergence gave rise to few phylum specific and other universal groups. Another possible scenario is that, as stated by Burroughs et al. (43), Cof group is indeed the root of this group, and not the SPSC group. A more extensive analysis of speciation pattern and PHYRN information content can help us infer the true outgroup of this lineage.

Figure 5-7: Unrooted Neighbor Joining (NJ) tree of PSP/PATPase derived HAD lineage. PHYRN derived phylogenetic tree of a HAD lineage which putatively evolved from a representative of PSP/PATPase group. dehr and Sdt1p groups split into different sequential groups. (No. of sequences = 843 sequences)
An unrooted Neighbor-Joining tree of a putative HAD lineage that evolved from a single representative of PSP/PATPase group is shown in Figure 5-7. This tree shows good monophyletic distribution of groups, alongwith multiple subgroups in some clades. Groups like dehr, Sdt1p, and ZR25 show proper subclade groupings. Further, we also observe previously unknown groups like ‘dehr/PSP like’ which may represent the ancestors of other groups like dehr, PPh and BPMG. Overall, the tree is very robust and shows good speciation pattern. Even in the unrooted tree, universal groups like PATPase attain the root position, giving further rise to a few phyla-specific, and mostly universal groups.

5.3 Methods

*Sequence collection and PHYRN*

Haloacid Dehalogenase (HAD) query sequences were collected from (43) and NCBI nr database, using expansion approach as defined in chapter 3. PHYRN was used as described in chapters 3 and 4.

*Clustering*

Cluster 3.0 ([http://bonsai.ims.u-tokyo.ac.jp/~mdehoon/software/cluster](http://bonsai.ims.u-tokyo.ac.jp/~mdehoon/software/cluster))(154), was used for hierarchical clustering of group-specific average PHYRN score matrix. Following, Java TreeView ([http://jtreeview.sourceforge.net/](http://jtreeview.sourceforge.net/)) (155) was used to infer a dendrogram from the clustered data.

5.4 Discussion

In this chapter, we describe an approach to study convergent evolution in divergent protein superfamilies using PHYRN. As a case study, we have used HAD superfamily of proteins. HAD
superfamily was chosen, because of its divergent nature, wide substrate range and convergent nature of its origin. A number of implications can be drawn from this study: 1) PHYRN and Hierarchical Clustering can be used to identify distinct lineages in a superfamily, and strongly suggests that 2) convergent superfolds of divergent ancestry can be identified.

Earlier studies on HAD superfamily have shown that it is hard to resolve the HAD phylogeny by relying exclusively on Multiple Sequence Alignment (MSA) based methods. Besides the divergent nature of this family that stymie phylogenetic studies of this family, Burroughs et al. also extrapolated occurrence of convergent evolution for this family. Existence of multiple lineages in extant HAD sequences, makes it even more difficult to resolve this superfamily using conventional MSA dependent phylogenetic methods. We have described a novel approach to study convergent evolution which relies on our previously described PHYRN methodology to quantitatively measure evolutionary signals from HAD phylogenetic profiles, followed by Pearson’s clustering approach to cluster these evolutionary signal information into distinct lineages. Individual lineages thus obtained can then be resolved for their evolutionary rates using PHYRN.

Using HAD superfamily as a case study, we first rejected the null hypothesis that HAD phylogeny evolved as a single lineage. Using our previously described PHYRN methodology, we constructed a single phylogenetic tree containing all the sequences of HAD superfamily. Although we were able to generate such a tree (Figure 5-1), the tree showed a significant lack of robustness and lot of anomalies in the speciation pattern. Based on these criteria, we rejected the notion that HAD superfamily evolved as a single lineage. To calculate the number of distinct lineages in HAD suerfamily, we followed a hierarchical clustering methodology as shown in figure 5-2. Individual HAD groups were clustered using Pearson’s hierarchical clustering based
on average group-wise PHYRN composite score. Based on this method, we predicted 4 different lineages in extant HAD superfamily (figure 5-3). To further analyze these independent lineages, we studied these lineages as separate phylogenetic trees.

As shown in figure 5-4, 5-5, 5-6 and 5-7, the trees constructed for these 4 independent lineages are statistically robust. Further, the speciation pattern in individual trees shows a commonly seen simple to complex evolutionary pattern. Natural root in these trees is always either a universal group, which has members in all three kingdoms of life or a bacterial specific group, while the groups furthest groups away from root group are either mostly higher organism specific or universal group. Unlike earlier studies, bacterial/archeal members of groups, in most cases, form monophyletic clades with other members of their group in our study. Although in few groups, we do find that archeal members of a group do form separate clades, but they may represent evolutionarily older progenitor groups. To better understand some of these archeal and bacteria specific groups, we will try to include more of their homologous sequences in our future studies. Also, another round of Pearson clustering within individual lineages can shed more light on occurrence of any more lineages. In the future, we wish to extend this study to a comprehensive evolutionary survey of HAD superfamily, by including all the available HAD sequences in our analysis.

We propose that this hybrid methodology of PHYRN and hierarchical clustering can be used to study protein families that evolved concurrently on multiple occasions. Taken together, with ability of PHYRN to resolve rapidly evolving and/or highly divergent superfamilies, this method can be used as phylogenetic tool to resolve problems concerning early origins of life.
Chapter 6

Conclusions and Future Perspectives

Collective aim of systems biologists is to construct a comprehensive, high-resolution kinetic bionetwork model, which includes flux information for each known biochemical reaction and their components. The biggest roadblock to achieving this aim is significant lack of kinetic data for many biochemical pathways. An approach to overcome this problem is to build discrete models that rely on non-kinetic information, and then iteratively increase the complexity of such models to achieve a coarse grained kinetic model that can be easily enriched with flux information of individual reaction steps when they get available. Based on this approach, many groundbreaking studies have used boolean-bionetwork modeling approach to model complex pathways(27-29;31;36). These boolean-bionetwork models have modeled steady state processes. In this thesis, we have shown a bionetwork-boolean model for an oscillatory biological process. Our model of Phospholipase-C mediated Ca\textsuperscript{2+} signaling pathway correctly replicates oscillatory pattern of calcium channels, IP3R and TRPC3(44). We also showed that this model correctly replicates wild type and mutant behavior of other nodes in the network, and there is extensive support in literature for these network simulations. Further, as seen for the DANGER node, our model recapitulates \textit{in vivo} phenotype of its knock-out, even when rules were framed from \textit{in vitro} data. Although we have seen this for only one node in our network model, this idea holds promise. This effect may represent the power of philosophy involved in these networks, which treats cells as a system of interacting biochemical reactions, instead of individual reactions working independently.

In future, we plan to take this model from being just a backbone model of calcium
signaling pathways to a more comprehensive representation of calcium signaling pathways. We wish to apply our approach of boolean modeling to networks as described in Berridge et al.(47). Further, we also plan to increase complexity at the modeling stage. Instead of using boolean approach, we will try to use higher order discrete modeling approaches to model biological behavior more closely. We also propose that if we use evolutionary rates as linking parameters in bionetworks, then we can define evolutionary networks, which can aid current phylogenetic analysis in multiple ways. Taken together, we believe, bionetworks can be a great tool for biologists and can help lab research extensively by compiling vast amount of information into easily comprehensible information circuits, as well as providing directionality to laboratory research through their predictions.

Besides, developing bionetworks from functional data spaces, we also propose a novel approach of studying structure, function and evolution data spaces simultaneously to solve the ‘protein problem’. Using this approach, we show that we can infer homology better than other conventional methods, especially in ‘twilight zone’ of sequence similarity. We describe development of a novel phylogenetic tool (PHYRN) that utilizes the power of phylogenetic profiles, and outperforms current MSA-based methods at recapitulating ‘true evolutionary history’ in simulated datasets. Further, it provides robust statistical support at deep nodes in highly divergent superfamilies such retroelements and the DANGER. Using these datasets, we show that PHYRN approach can be scaled up easily to study phylogenies with thousands of sequences. We also present a roadmap to enrich and amplify for informative sequences and phylogenetic profiles, and use this family/superfamily-specific evolutionary data spaces to construct comprehensive phylogenetic trees. Using this approach of scaling up, we propose to construct largest, most resolved phylogenetic tree of retroelements. Since RT have implications
in early evolution of life(156), a tree of such proportions and divergence will help us answer questions about early evolution of life, that were previously considered intractable. We also propose to include RNA dependent RNA polymerases (RdRp) in our analysis of retroelements, to explore whether these proteins have a common evolutionary ancestor, which has been proposed in multiple studies(39;157;158).

Our study of the DANGER superfamily demonstrates that PHYRN provides better deep node statistical support as compared to other MSA dependent phylogenetic tools. Further, PHYRN provides quantitative measures to calculate distribution of conservation signals among phylogenetic profiles. Such analysis can be utilized to identify sequence groups that do not belong in the same tree, and thus this analysis can be beneficial in outgroup identification, as well as the study of convergent evolution. To validate whether PHYRN can be used to study convergent evolution, we defined a hybrid approach of PHYRN and hierarchical clustering. In this approach, we first define number of distinct lineages in a superfamily and subsequently resolve individual lineages into phylogenetic trees using PHYRN. Using Haloacid Dehalogenase (HAD) as a benchmark family for convergent evolution, we inferred that there are 4 different lineages in extant HAD sequence groups. In the future, we propose to scale up DANGER and HAD phylogenies for under-represented sequence groups and thus improve the resolution of these phylogenetic trees.

Since structure, function and evolution (S, F, & E) data spaces also govern the essentiality of different cellular components, we propose to modify PHYRN’s approach of phylogenetic profiles, and use a similar approach to define essentiality profiles. We believe that construction of such essentiality profile libraries will provide measures to inform essentiality of proteins in different genomes in a high throughput manner. A tool based on these criteria will prove
enormously helpful in identification of important biological targets for therapeutic intervention.

Overall, in PHYRN we describe a tool that provides 1) better recapitulation of true evolutionary history 2) better deep node statistical support 3) quantitative measures to identify outgroup and distinct lineages, and a ‘comprehensive survey approach’ for evolutionary data spaces, that is radically different but significantly better than earlier methods. Taken together, we believe that PHYRN can be a good MSA independent phylogenetic tool which is likely to provide unprecedented success in exploration of complex evolutionary questions, which have remained unanswered and/or unresolved so far, due to high divergence, rapid evolutionary rates, data size and/or convergence.
Reference List


(43) Burroughs AM, Allen KN, Dunaway-Mariano D, Aravind L. Evolutionary genomics of the HAD superfamily: understanding the structural adaptations and catalytic diversity.


(51) Li HS, Xu XZ, Montell C. Activation of a TRPC3-dependent cation current through the neurotrophin BDNF. Neuron 1999 Sep;24(1):261-73.


(55) Zhang Y, Hoon MA, Chandrashekar J, Mueller KL, Cook B, Wu D, et al. Coding of sweet, bitter, and umami tastes: different receptor cells sharing similar signaling pathways. Cell 2003 Feb 7;112(3):293-301.


(63) von Dassow G. The segment polarity network is a robust developmental module. Nature 2000;406(6792):188.


(84) Ferris CD, Huganir RL, Snyder SH. Calcium flux mediated by purified inositol 1,4,5-
trisphosphate receptor in reconstituted lipid vesicles is allosterically regulated by

(85) Mak DO, McBride S, Foskett JK. ATP regulation of type 1 inositol 1,4,5-trisphosphate
receptor channel gating by allosteric tuning of Ca(2+) activation. J Biol Chem 1999

(86) Haynes LP, Tepikin AV, Burgoyne RD. Calcium binding protein 1 is an inhibitor of
agonist-evoked, inositol 1,4,5-trisphosphate-mediated calcium signalling. J Biol Chem
2003 Oct 21;M309617200.

(87) Uchiyama T, Yoshikawa F, Hishida A, Furuichi T, Mikoshiba K. A Novel
Recombinant Hyperaffinity Inositol 1,4,5-Trisphosphate (IP3) Absorbent Traps IP3,
Resulting in Specific Inhibition of IP3-mediated Calcium Signaling. J Biol Chem 2002

(88) Bruce JI, Shuttleworth TJ, Giovannucci DR, Yule DI. Phosphorylation of inositol 1,4,5-
trisphosphate receptors in parotid acinar cells. A mechanism for the synergistic effects

(89) Straub SV, Giovannucci DR, Bruce JI, Yule DI. A role for phosphorylation of inositol
1,4,5-trisphosphate receptors in defining calcium signals induced by Peptide agonists in

(90) Grimaldi M, Favit A, Alkon DL. cAMP-induced cytoskeleton rearrangement increases
calcium transients through the enhancement of capacitative calcium entry. J Biol Chem
1999 Nov 19;274(47):33557-64.

(91) Bartlett PJ, Young KW, Nahorski SR, Challiss RA. Single cell analysis and temporal
profiling of agonist-mediated inositol 1,4,5-trisphosphate, Ca2+, diacylglycerol, and
10;280(23):21837-46.

(92) Kim UH, Fink D, Jr., Kim HS, Park DJ, Contreras ML, Guroff G, et al. Nerve growth
factor stimulates phosphorylation of phospholipase C-gamma in PC12 cells. J Biol

(93) Acharya JK, Jalink K, Hardy RW, Hartenstein V, Zucker CS. InsP3 receptor is essential

(94) Rebecchi MJ, Pentyala SN. Structure, function, and control of phosphoinositide-

(95) Rhee SG. Regulation of phosphoinositide-specific phospholipase C. Annu Rev
Biochem 2001;70:281-312.


VITA
GAURAV BHARDWAJ

Education

**PhD in Cell and Developmental Biology (IBIOS program) **  May 2010
Pennsylvania State University, State College, Pennsylvania, USA

**B. Tech in Biotechnology **  2002- 2006
G. G. S Indraprastha University, Delhi, India
Thesis: “Partial Isolation of Ubiquitin Receptor/Binding Protein from Human Peripheral Blood Mononuclear Cells”

Publications


